Analytic Technical Assistance and Development
U.S. Department of Education
August 2017

Multi-armed RCTs: A design-based framework

Peter Z. Schochet
Mathematica Policy Research, Inc.

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S Department of Education

NCEE 2017-4027

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased, large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

August 2017

This report was prepared for the Institute of Education Sciences (IES) by Decision Information Resources, Inc. under Contract ED-IES-12-C-0057, Analytic Technical Assistance and Development. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Schochet, P.Z. (2017) *Multi-armed RCTs: A design-based framework* (2017–4027). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development.

Contents

Tables

Figures

(This page left intentionally blank)

Topics on design-based causal inference for multi-armed RCTs

Design-based methods have recently been developed as a way to analyze data from impact evaluations of interventions, programs, and policies (Imbens and Rubin, 2015; Schochet, 2015, 2016). The estimators are derived using the building blocks of experimental designs with minimal assumptions, and are unbiased and normally distributed in large samples with simple variance estimators. The methods apply to randomized controlled trials (RCTs) and quasi-experimental designs (QEDs) with comparison groups for a wide range of designs used in social policy research. The methods have important advantages over traditional model-based impact estimation methods, such as hierarchical linear model (HLM) and robust cluster standard error (RCSE) methods, and perform well in simulations (Schochet, 2016). The free *RCT-YES* software (www.rct-yes.com) estimates and reports impacts using these design-based methods.

The literature on design-based methods has focused on RCTs with a *single* treatment and a *single* control group. This theory, however, has not been formally extended to designs with *multiple* research groups. This is an important gap in the literature because multi-armed RCTs can simultaneously examine the effects of multiple interventions in a single study, thereby increasing the amount that researchers and policymakers can learn from impact evaluations. In social policy research, these designs are particularly relevant for interventions that are relatively easy to implement—for example, an education RCT testing several texting initiatives to improve student engagement and achievement. Relatedly, multi-armed designs are useful for rapid-cycle or opportunistic experiments aimed at continuous program improvement, for example, using behavioral-based interventions and encouragement designs.

Multi-armed RCT designs have been used in education research in a variety of contexts. For instance, they have been used to test the effects of different forms of teacher-to-parent communication on student outcomes (Kraft and Rogers, 2014) and the effects of text messaging and peer mentoring on college enrollment rates among high school graduates (Castleman and Page, 2015). Multi-armed RCTs have also been used in larger studies to test the effects of competing math curricula (Agodini et al., 2009) and reading curricula (James-Burdumy et al., 2009). They have also been used internationally, for example, in Honduras to examine the effects of various data-driven assessment tools to improve teaching practices and student outcomes (Toledo et al., 2015). These studies were all able to test a range of policy-relevant interventions at once without sacrificing statistical rigor.

This report discusses several key topics for estimating average treatment effects (ATEs) for multi-armed designs. The report is geared toward methodologists with a strong background in statistical theory and a good knowledge of design-based concepts for the single treatment-control group (two-group) design. The report builds on Schochet (2016), referencing key results and formulas to avoid repetition, and serves as a supplement to that report. The focus is on RCTs, although key concepts apply also to QEDs with comparison groups.

The report is in three sections. Section 1 discusses how design-based ATE estimators for the two-group design need to be modified for the multi-armed design when comparing pairs of research groups to each other. Section 2 discusses multiple comparison adjustments when conducting hypothesis tests across pairwise contrasts to identify the most effective interventions. Finally, Section 3 shows that the assumptions required to identify and estimate the complier average causal effect (CACE) parameter using an instrumental variable (IV) framework become much more complex in the multi-armed context, and may not be possible in some cases. While Sections 2 and 3 are germane to multi-armed designs regardless of the impact estimation methods used for the analysis, these sections emphasize approaches that align with the non-parametric underpinnings of the design-based framework.

1. Estimators for average treatment effects (ATEs)

An important consideration for multi-armed RCTs is the pairwise contrasts of interest to best address the study research questions. For instance, researchers may be interested in all pairwise comparisons across the research groups, pairwise comparisons with the control group, or pairwise comparisons with the best of the other treatments (Hsu, 1996, Westfall et al., 1999). This section considers design-based ATE estimators for each pairwise contrast in isolation. For example, for a design with three treatment groups (T1, T2, and T3) and a control group (C), the methods apply to each possible pairwise contrast (for example, T1-T2, T1-T3, T1-C, T2-C, and so on) as well as to contrasts formed by combining groups (for example, comparing the combined T1 and T2 groups to the C group). Thus, the methods apply to the multi-armed context regardless of the full set of contrasts of interest. Because multi-armed RCTs involve multiple hypothesis testing across contrasts, the inflation of Type 1 errors for each individual test is of concern. We hold off on a discussion of multiple comparison adjustments until Section 2.

We focus on the designs presented in Schochet (2016) defined by two key features: (i) clustering and (ii) blocking. *Non-clustered designs* are those where the unit of analysis aligns with the unit of randomization (such as analyzing student-level data with student-level random assignment), whereas *clustered designs* are those where the unit of analysis is nested within the unit of random assignment (such as analyzing student-level data with school- or teacher-level random assignment). For *non-blocked designs*, randomization is conducted within a single population, whereas for *blocked designs*, randomization is conducted separately within distinct sub-populations (such as school districts or grades). In combination, these two design features cover most RCTs in the social policy field. We consider design-based estimators for both the finite-population (FP) model where impacts are assumed to pertain to the study sample only, and the super-population (SP) model where impacts are assumed to generalize to an infinite population (which may be vaguely defined). We also consider estimators for models with and without baseline covariates and weights.

As formalized mathematically in this section, key components of the design-based theory for the two-group design apply also to the multi-armed context. However, two simple modifications are required:

1. Under the FP model, ATE estimators for each pairwise contrast pertain to the *entire* randomized sample, not just to the two groups being compared. Thus, variance estimators for the FP model for the two-group design need to be adjusted slightly to reflect the broader inference population.
2. For similar reasons, analysis weights for each pairwise comparison need to be scaled to reflect the size of the full randomized sample for each block and subgroup.

In this section, we do not consider statistical power considerations for multi-armed designs, but note here that for several reasons, these designs could require larger samples to produce precise impact estimates than for the two-group design. First, for multi-armed designs, the sample is split across more research groups. Second, we might expect impacts to be smaller when contrasting variants of

a treatment than when comparing a treatment to a control (status quo) condition. Finally, larger sample sizes might be required to compensate for multiple comparison adjustments when conducting hypothesis tests across the pairwise contrasts. These factors could be mitigated somewhat for rapid-cycle, multi-armed RCTs that focus on mediating or proximal outcomes (such as teacher knowledge) where we might expect intervention effects to be larger than for more distal outcomes (such as student test scores).

In what follows, we focus on *changes* to the ATE estimators as we move from the two- to multi-group design. Thus, we often reference Schochet (2016) for background results and pertinent details that are not repeated here. The approach is based on the Neyman-Rubin-Holland potential outcomes framework that underlies experiments (Holland, 1986; Neyman, 1923, Rubin, 1974, 1977).

a. Non-clustered designs

Consider the simplest RCT design where n students from a single population are randomly assigned to one of K distinct research groups ($K \geq 3$). The research groups could include a control (business-as-usual) group, but do not have to. Each treatment group is offered a different intervention or combination of interventions (for example, for two interventions T1 and T2, the four research groups for a full factorial design could be defined by the receipt of both T1 and T2, T1 only, T2 only, or neither). We assume the same design is used to select each research group.

We do not consider orthogonal fractional factorial designs where the interventions consist of components that could each be turned on or off to form different service packages, and where the research groups include only a subset of all possible treatment combinations (see Box et al. 2005; Wu and Hamada, 2009). Our focus is on examining pairwise contrasts between distinct research groups, whereas factorial designs are structured to estimate main and interaction effects by comparing *combinations* of research groups to each other. Under factorial designs, main and two-way interaction effects become confounded with higher-order interaction effects (and even with each other in some designs with small numbers of tested combinations), and the nature of the confounding depends on the adopted parameterization of the factorial design. This confounding complicates the assumptions required to develop design-based impact estimators, and these assumptions are likely to vary based on the structure of the factorial design. We do not address this topic here, but it is an interesting area for future research.

We assume the sample contains $n_k = np_k$ individuals in research group k $(k = 1,2,...,K)$, where p_k is the research group sampling rate $(0 < p_k < 1; \sum_{k=1}^{K} p_k = 1)$ and K is finite. Let $Y_i(k)$ be the potential outcome for individual i in research group k, and let $T_i(k)$ be the research group indicator variable that equals 1 if an individual is assigned to group k, and 0 otherwise $(\sum_{k=1}^{K} T_i(k) = 1)$. More succinctly, let $Q_i = k$ for individuals assigned to group k.

Design-based estimators in the multi-armed context rely on several assumptions that generalize the corresponding assumptions for the two-group design (see Schochet, 2016, pp 28-29):

- The stable unit treatment value assumption (SUTVA) (Rubin, 1986) which has two parts. First, for any two random assignment vectors $\mathbf{Q}$ and $\mathbf{Q}'$, if $Q_i = Q_i'$ for individual i, then $Y_i(\mathbf{Q}) = Y_i(\mathbf{Q}')$. This means that the potential outcomes of an individual depend only on that person's research assignment and not on the assignments of other individuals. It also implies that the potential outcomes for a particular treatment are independent of the number and nature of the other treatments. The second SUTVA condition is that an individual offered a particular treatment does not receive different forms of the treatment either in isolation or in combination with other treatments. For example, if interest lies in estimating interaction effects for treatments T1 and T2, the same version of T1 must be delivered to the group receiving T1 only and the group receiving both T1 and T2, and similarly for T2; otherwise, the T1-T2 intervention should be considered a different intervention (T3) rather than one that supplements one treatment with the other.

- Independence between research status and potential outcomes, $Q_i \perp\!\!\!\perp (Y_i(1), Y_i(2), ..., Y_i(K))$, which is ensured by randomization for RCTs and is assumed to hold conditional on baseline covariates for QEDs with comparison groups.

- A positive probability of assignment to each research group for each individual.

- Finite first and second moments for all potential outcome distributions.

FP model

Under the FP model, the n individuals participating in the study are assumed to define the population universe, and potential outcomes are assumed to be fixed for the study. In this setting, the ATE parameter of interest for comparing interventions (research groups) k and k' is

$$(1) \quad \beta_{nclus,FP}(k,k') = \sum_{i=1}^{n} (Y_i(k) - Y_i(k'))/n = \bar{Y}(k) - \bar{Y}(k').$$

Importantly, this parameter pertains to the full sample of n individuals, not just to the $(n_k + n_{k'})$ individuals randomized to the contrasted groups. Thus, for multi-armed designs, ATE estimators for each pairwise contrast generalize beyond the estimation sample to the full randomized sample. Accordingly, the approach for developing estimators in Schochet (2016) for the SP model with an infinite population apply also to the multi-armed context, except that the variances need to be adjusted to reflect the finite rather than infinite inference population.

To demonstrate these adjustments, consider first the data generating process for the observed outcome y_i :

$$(2) \quad y_i = \sum_{k=1}^{K} T_i(k) Y_i(k).$$

This relation states we can observe $Y_i(k)$ if an individual is randomly assigned to research group k, but not the person's potential outcomes in other research conditions. In this expression, y_i is random because the $T_i(k)$ indicators are random (the potential outcomes are assumed to be fixed).

Consider the simple differences-in-means estimator for $\beta_{nclus,FP}(k,k')$ calculated using the sample randomized to conditions k and k' :

$$(3) \quad \hat{\beta}_{nclus,FP}(k,k') = \bar{y}(k) - \bar{y}(k') = \frac{1}{n_k} \sum_{i:T_i(k)=1}^{n_k} y_i - \frac{1}{n_{k'}} \sum_{i:T_i(k')=1}^{n_{k'}} y_i.$$

To show that this estimator is unbiased, we use the law of iterated expectations. First, we calculate the expectation of $\hat{\beta}_{nclus,FP}(k,k')$ with respect to the distribution (R) of all possible randomizations to groups k or k', *conditional* on the $(n_k + n_{k'})$ students assigned to the two groups and their fixed potential outcomes. Second, we average over random draws of $(n_k + n_{k'})$ individuals from the population (I) of n individuals in the study. Mathematically, this approach can be expressed as

$$(4) \quad E_{RI}(\hat{\beta}_{nclus,FP}(k,k')) = E_I(E_R(\hat{\beta}_{nclus,FP}(k,k') \mid \{Y_i(k), Y_i(k')\} : i \ni T_i(k) + T_i(k') = 1; n_k, n_{k'})).$$

We know from Schochet (2016) that the simple differences-in-means estimator for the two-group design is unbiased for the FP model. Thus, the interior conditional expectation in (4) equals $(\bar{Y}^*(k) - \bar{Y}^*(k'))$, where $\bar{Y}^*(k)$ and $\bar{Y}^*(k')$ are mean potential outcomes for those in the estimation sample. Thus, (4) reduces to

$$(5) \quad E_I(\bar{Y}^*(k) - \bar{Y}^*(k')) = \frac{1}{(n_k + n_{k'})} \sum_{i=1}^{n} E_I(Z_i(k,k')[\bar{Y}^*(k) - \bar{Y}^*(k')]) = \bar{Y}(k) - \bar{Y}(k'),$$

where $Z_i(k,k')$ equals 1 for individuals randomized to group k or k' and 0 otherwise. The last equality holds because (i) $Z_i(k,k')$ is independent of $\bar{Y}^*(k)$ and $\bar{Y}^*(k')$ due to randomization, (ii) $E_I(Z_i(k,k')) = (n_k + n_{k'})/n$ and (iii) $E_I(\bar{Y}^*(k) - \bar{Y}^*(k')) = \bar{Y}(k) - \bar{Y}(k')$. This proves that $\hat{\beta}_{nclus,FP}(k,k')$ is unbiased.

We can use a similar conditioning approach to calculate the variance of $\hat{\beta}_{nclus,FP}(k,k')$ using the law of total variance where, to simplify notation, we do not display the conditioning set:

$$(6)\quad Var_{RI}(\hat{\beta}_{nclus,FP}(k,k')) = E_I(Var_R(\hat{\beta}_{nclus,FP}(k,k'))) + Var_I(E_R(\hat{\beta}_{nclus,FP}(k,k'))).$$

Using variance results for the FP model for the two-group design (Schochet, 2016, Equation (5.6)), we have that

$$(7)\quad E_I(Var_R(\hat{\beta}_{nclus,FP}(k,k'))) = \frac{\sigma_I^2(k)}{n_k} + \frac{\sigma_I^2(k')}{n_{k'}} - \frac{\sigma_{\tau I}^2(k,k')}{n_k + n_{k'}},$$

where $\sigma_I^2(k) = \sum_{i=1}^{n}(Y_i(k) - \overline{Y}(k))^2/(n-1)$ is the variance of $Y_i(k)$ across the entire randomized sample, similarly for $\sigma_I^2(k')$, and $\sigma_{\tau I}^2(k,k') = \sum_{i=1}^{n}([Y_i(k) - \overline{Y}(k)] - [Y_i(k') - \overline{Y}(k')])^2/(n-1)$ is the variance of individual-level treatment effects for the contrasted groups. We hereafter refer to the final term in (7) as the "FP heterogeneity term."

Similarly, because the differences-in-means estimator is unbiased for the FP model, we have that

$$(8)\quad Var_I(E_R(\hat{\beta}_{nclus,FP}(k,k'))) = Var_I(\overline{Y}^*(k) - \overline{Y}^*(k')) = \frac{Var_I(Y_i^*(k) - Y_i^*(k'))}{n_k + n_{k'}} = (1-f)\frac{\sigma_{\tau I}^2(k,k')}{n_k + n_{k'}},$$

where $(1-f)$ is the finite population correction (*fpc*) with $f = (n_k + n_{k'})/n$. Intuitively, the fpc accounts for the sampling of individuals (and their associated treatment effects) from the full randomized sample. Finally, collecting terms in (7) and (8), we find that the variance in (6) is

$$(9)\quad Var_{RI}(\hat{\beta}_{nclus,FP}(k,k')) = \frac{\sigma_I^2(k)}{n_k} + \frac{\sigma_I^2(k')}{n_{k'}} - \frac{\sigma_{\tau I}^2(k,k')}{n}.$$

The critical difference then between the variance expression for the multi-armed RCT and the two-group RCT is that the FP heterogeneity term contains the divisor n rather than $(n_k + n_{k'})$. Thus, the variance increases as we add more research groups due to the FP heterogeneity term. Stated differently, the variance increases as the size of the inference population outside the estimation sample becomes larger.

Unbiased estimates for $\sigma_I^2(k)$ can be obtained using the sample variance, $s_I^2(k) = \sum_{i:T_i(k)=1}^{n_k}(y_i - \overline{y}(k))^2/(n_k - 1)$ and similarly for $\sigma_I^2(k')$. The FP heterogeneity term, $\sigma_{\tau I}^2(k,k')$, is not identifiable because it is not possible to observe an individual in both research

conditions. However, following Schochet (2016), we can bound this term by noting that $\sigma^2_{\tau I}(k,k') \geq (\sigma_I(k) - \sigma_I(k'))^2$, which yields the following conservative variance estimator:

$$(10) \quad \hat{Var}_{RI}(\hat{\beta}_{nclus,FP}(k,k')) = \frac{s^2_I(k)}{n_k} + \frac{s^2_I(k')}{n_{k'}} - \frac{(s_I(k) - s_I(k'))^2}{n} \ .$$

The estimator $\hat{\beta}_{nclus,FP}(k,k')$ is asymptotically normal (Schochet, 2016, Lemmas 5.1 and 5.2). Thus, hypothesis testing can be conducted using *z*-tests or *t*-tests with $(n_k + n_{k'} - 2)$ degrees of freedom (multiple comparisons adjustments are discussed in Section 2). Similar results apply for estimating impacts for subgroups defined by baseline characteristics (such as for males and females).

For models that include weights (w_i) to adjust for data nonresponse or other design factors, an asymptotic variance estimator for the weighted differences-in-means estimator is as follows:

$$(11) \quad \hat{AsyVar}(\hat{\beta}_{nclus,FP,W}(k,k')) = \frac{s^2_{IW}(k)}{\bar{w}^2_k n_k} + \frac{s^2_{IW}(k')}{\bar{w}^2_{k'} n_{k'}} - \frac{f_W[(s_{IW}(k)/\bar{w}_k) - (s_{IW}(k')/\bar{w}_{k'})]^2}{n_k + n_{k'}},$$

where $\bar{w}_k = (\sum_{i:T_i(k)=1}^{n_k} w_i / n_k)$ and $\bar{w}_{k'} = (\sum_{i:T_i(k')=1}^{n_{k'}} w_i / n_{k'})$ are average weights for the two groups, $s^2_{IW}(k) = [\sum_{i:T_i(k)=1}^{n_k} w_i^2 (y_i - \bar{y}_W(k))^2 / (n_k - 1)]$ is the weighted sample variance, and similarly for $s^2_{IW}(k')$. The fpc term, f_W, could be set to $(n_k + n_{k'})/n$, but other options exist depending on the nature of the weights.

SP model with an infinite super-population

Model-based ATE estimation methods, such as ordinary least squares (OLS) and HLM, typically assume a SP framework with an infinite super-population. For this framework, the parameter of interest for each pairwise contrast is the mean effect in the infinite super-population, $\beta_{nclus,SP}(k,k') = E_I(Y_i(k) - Y_i(k'))$.

In this setting, ATE estimators for the multi-armed design are *identical* to those for the two-group design. The simple differences-in-means estimator for each contrast is unbiased and asymptotically normal, with the same variance estimator as in (10) except that the FP heterogeneity term disappears because n is infinite. These results hold regardless of the number of research groups. Intuitively, each research group is a random sample from the infinite super-population, so the potential outcomes across research groups are uncorrelated (which differs from the FP framework).

The finding that ATE estimators under the SP model are the same for the two-group and multi-armed RCT extends to blocked and clustered designs. This occurs because the SP framework for these designs assumes random sampling of study blocks, clusters, and individuals from infinite

populations so the FP heterogeneity terms do not apply. This same finding regarding the absence of the FP heterogeneity terms for the SP model also applies to models that include baseline covariates. Thus, we focus only on the FP model for the remainder of this section.

Models with baseline covariates

Researchers analyzing RCT data often include baseline covariates in the estimation models to increase precision and to adjust for random imbalances between the research groups. For multi-armed designs, *separate* regression models can be estimated for each pairwise contrast. In this case, the statistical properties of the multiple regression estimator for the two-group design apply to each pairwise model. The only difference is that the FP heterogeneity term for the variance estimator contains the divisor n rather than $(n_k + n_{k'})$.

To demonstrate these results more formally, using the estimation sample assigned to conditions k and k', consider an OLS regression of y_i on the vector, $\tilde{\mathbf{z}}_\mathbf{i} = (1 \;\; \tilde{T}_i \;\; \tilde{\mathbf{x}}_\mathbf{i})$, with associated parameters $(\beta_0 \;\; \beta_{nclus,FP}(k,k') \;\; \boldsymbol{\gamma})$. In this regression, $\tilde{T}_i = T_i - p_k^*$ is the centered treatment assignment indicator variable, T_i equals 1 if the individual is assigned to condition k, and 0 if assigned to condition k'; $p_k^* = n_k/(n_k + n_{k'})$ is the sampling proportion to condition k among the estimation sample; and $\tilde{\mathbf{x}}_\mathbf{i} = \mathbf{x}_\mathbf{i} - \overline{\mathbf{x}}$ are centered covariates. Applying Lemmas 5.3 and 5.4 in Schochet (2106) to the multi-armed context, we find then that the multiple regression estimator for the FP model, $\hat{\beta}_{nclus,MR,FP}(k,k')$, is asymptotically normal with asymptotic mean $\beta_{nclus,FP}(k,k')$ and the following conservative variance estimator:

$$(12) \quad \hat{AsyVar}_{RI}(\hat{\beta}_{nclus,MR,FP}(k,k')) = \frac{MSE(k)}{n_k} + \frac{MSE(k')}{n_{k'}} - \frac{[\sqrt{MSE(k)} - \sqrt{MSE(k')}]^2}{n}, \text{ where}$$

$$MSE(k) = \frac{1}{n_k - vp_k^* - 1} \sum_{i:T_i(k)=1}^{n_k} (y_i - \hat{\beta}_0 - \hat{\beta}_{nclus,MR,FP}(k,k')(1-p_k^*) - \tilde{\mathbf{x}}_\mathbf{i}\hat{\boldsymbol{\gamma}})^2 \text{ and}$$

$$MSE(k') = \frac{1}{n_{k'} - v(1-p_k^*) - 1} \sum_{i:T_i(k')=1}^{n_{k'}} (y_i - \hat{\beta}_0 + \hat{\beta}_{nclus,MR,FP}(k,k')p_k^* - \tilde{\mathbf{x}}_\mathbf{i}\hat{\boldsymbol{\gamma}})^2.$$

In this expression, the $MSE(.)$ terms are regression mean square errors for the two research groups, respectively; v is the number of baseline covariates (assumed to be the same across pairwise contrasts); and $\hat{\beta}_0$, $\hat{\beta}_{nclus,MR,FP}(k,k')$, and $\hat{\boldsymbol{\gamma}}$ are parameter estimates. Hypothesis tests can be conducted using *t*-tests with $(n_k + n_{k'} - v - 2)$ degrees of freedom.

Another approach is to estimate a single model using the entire randomized sample, where y_i is regressed on an intercept, the $\tilde{T}_i(k) = T_i(k) - p_k$ variables (leaving out one), and the covariates, $\tilde{\mathbf{x}}_\mathbf{i}$, centered using covariate means for the whole sample. This specification would reduce the number of degrees of freedom for hypothesis testing, and produce a uniform set of regression-adjusted group means across all pairwise contrasts, facilitating the reporting of the impact results. However, the design-based theory for this specification becomes complex due to the negative covariances between the $\tilde{T}_i(k)$ and $\tilde{T}_i(k')$ regressors (equal to $-p_k p_{k'}$), leading to sub-matrices with elements, $E_I(\tilde{T}_i(k)\tilde{T}_i(k'))$, that need to be inverted to examine the asymptotic properties of the multiple regression estimator. An additional complication is that the ATE parameter estimates for the $\tilde{T}_i(k)$ variables are correlated in finite samples through their shared correlations with the covariates. Thus, we do not pursue this approach further here. Rather, we adopt an approach where a separate regression model is estimated for each pairwise contrast.

Finally, if a regression model with covariates is used to estimate ATEs for subgroups (for example, for boys and girls) by including subgroup-by-treatment interaction terms in the model, the analysis weights should be scaled to reflect the size of the subgroups across all randomized research groups, not just for the two groups being compared. For example, suppose there are three research groups, where the first research group has 100 boys and 150 girls, the second research group has 200 boys and 180 girls, and the third research group has 300 boys and 250 girls. Suppose also that no analysis weights are specified so that each individual is to receive equal weight in the analysis. In this case, when comparing the first two research groups, weights should be constructed equal to (600/300) for boys and (580/330) for girls so that the impact results can generalize to all three research groups. The models should then be estimated using weighted least squares. These weights are not necessary for subgroup analyses that omit baseline covariates.

Extensions to blocked designs

Blocked designs occur when random assignment is conducted separately within distinct study subpopulations (such as sites or grades). Consider a blocked design with K research groups in each block. Let the subscript "b" indicate blocks $(b = 1, 2, ..., h)$, and let S_{ib} be a block indicator variable that equals 1 if individual i is in block b, and 0 for individuals in other blocks. Sampling rates to the research groups (that is, the p_{kb} probabilities) could differ both within and across blocks.

In the multi-armed FP context, the ATE parameter for block b is

$$(13) \quad \beta_{nclus,b,FP}(k,k') = \bar{Y}_b(k) - \bar{Y}_b(k') = \sum_{i:S_{ib}=1}^{n_b} (Y_{ib}(k) - Y_{ib}(k'))/n_b,$$

which is calculated over the full randomized sample of n_b individuals in the block. The ATE parameter across all blocks can then be expressed as

$$(14) \quad \beta_{nclus,blocked,FP}(k,k') = \sum_{b=1}^{h} w_b \beta_{nclus,b,FP}(k,k') / \sum_{b=1}^{h} w_b,$$

which is a weighted average of the block-specific ATEs with weights w_b.

Importantly, w_b in (14) pertains to *all* randomized individuals in the block, not just to those assigned to the contrasted groups. Thus, if blocks are weighted by their total sample size, we would set $w_b = n_b$ rather than $w_b = (n_{kb} + n_{k'b})$, as would be the case for the two-group design. As an example, consider a design with two blocks and three research groups, where Block 1 has 10 individuals per research group, Block 2 has 20 individuals per research group, and the pairwise contrast of interest is for research groups *k* and *k'*. In this case, the analysis weights should be calculated based on $w_1 = n_1 = 30$ and $w_2 = n_2 = 60$, instead of $w_1 = (n_{k1} + n_{k'1}) = 20$ and $w_2 = (n_{k2} + n_{k'2}) = 40$. Note, however, that in this scenario, these two approaches will yield the same weights, with Block 1 receiving one-third of the weight and Block 2 receiving two-thirds. More generally, the weights for the two-group and multi-armed designs will only differ in practice if the combined sampling rate (the p_{kb} probabilities) for the contrasted research groups differs across blocks (that is, if $p_{k1} + p_{k'1} \neq p_{k2} + p_{k'2}$). This could occur, for instance, if individuals in Block 1 have an equal one-third probability of being assigned to each of the three research groups, while individuals in Block 2 have a 20 percent chance of being assigned to research group *k*, and a 40 percent chance of being assigned to the other two research groups.

ATE estimators for the parameter in (14) for the multi-armed design are the same as for the two-group design (see Schochet, 2016; Chapter 6). The only differences are that the block-specific weights for each pairwise contrast pertain to the full randomized sample, and the FP heterogeneity terms in the variance estimators have divisors n_b rather than $(n_{kb} + n_{k'b})$. For example, the simple differences-in-means estimator for a particular pairwise contrast is

$$(15) \quad \hat{\beta}_{nclus,blocked,FP}(k,k') = \sum_{b=1}^{h} w_b (\bar{y}_b(k) - \bar{y}_b(k')) / \sum_{b=1}^{h} w_b,$$

where $\bar{y}_b(k)$ and $\bar{y}_b(k')$ are block-specific mean outcomes. This estimator is unbiased and asymptotically normal with the following variance:

$$(16) \quad Var_{RI}(\hat{\beta}_{nclus,blocked,FP}(k,k')) = \sum_{b=1}^{h} w_b^2 Var_{RI}(\hat{\beta}_{nclus,b,FP}(k,k')) / (\sum_{b=1}^{h} w_b)^2,$$

which can be estimated by applying (10) separately for each block. Similarly, weighted regression-adjusted estimators can be obtained using Equations (6.9) and (6.16) in Schochet (2016) with the same modifications to the weights and FP heterogeneity term as above that reflect sample sizes for the full randomized sample in the block, not just block sample sizes for the pairwise contrast of interest. For subgroup analyses, the weights should be scaled by subgroup.

b. Clustered designs

The above theory extends directly to clustered designs where groups (such as schools or classrooms) rather than individuals (such as students) are randomly assigned to the research groups, and where outcome data are collected on individuals (such as students). In this section, we show that the same issues apply to clustered designs as non-clustered designs as we move from the two-group to multi-armed context. Our focus is on the FP model, because ATE estimators for clustered designs under the SP model do not change in the multi-armed setting. Because parallel issues exist for clustered and non-clustered designs, we provide less detail in this section than before.

For the analysis, we use similar notation as for the non-clustered design with the addition of the subscript "j" to indicate clusters. For instance, for the non-blocked design, $Y_{ij}(k)$ is the potential outcome for individual i in cluster j in research condition k; y_{ij} is the observed outcome; $T_j(k)$ is the research status indicator variable that equals 1 if cluster j is randomly assigned to group k, and 0 otherwise; and $Q_j = k$ for clusters assigned to group k. We assume that the sample contains m clusters with $m_k = mp_k$ clusters assigned to group k, where p_k is the research group sampling rate. It is assumed that cluster j has n_j individuals.

Similar to the non-clustered design, we rely on several assumptions for the clustered design: (i) SUTVA, which in three parts states that for any two random assignment vectors $\mathbf{Q}$ and $\mathbf{Q}'$, if $Q_j = Q'_j$ for cluster j, then $Y_{ij}(\mathbf{Q}) = Y_{ij}(\mathbf{Q}')$, that there are not different forms of the same treatment, and that potential outcomes for a given treatment do not depend on the number or nature of the other treatments; (ii) randomization, defined as the independence between cluster-level research status and potential outcomes, $Q_j \perp\!\!\!\perp (Y_{ij}(1), Y_{ij}(2), \ldots, Y_{ij}(K))$; (iii) positive assignment probabilities for each cluster to each research group; and (iv) finite first and second moments for the potential outcome distributions.

In the multi-armed setting for clustered designs, the ATE parameter for comparing interventions k and k' in finite samples is as follows:

$$(17) \quad \beta_{clus,FP}(k,k') = \sum_{j=1}^{m} w_j(\bar{Y}_j(k) - \bar{Y}_j(k')) / \sum_{j=1}^{m} w_j,$$

where w_j is the fixed cluster weight (for example, $w_j = 1$ or $w_j = n_j$) and $\overline{Y}_j(k) = (\sum_{i=1}^{n_j} Y_{ij}(k)/n_j)$ and $\overline{Y}_j(k') = (\sum_{i=1}^{n_j} Y_{ij}(k')/n_j)$ are mean, cluster-level potential outcomes in conditions k and k', respectively. This ATE parameter is a weighted average of treatment contrasts across all m clusters in the sample, not just the $(m_k + m_{k'})$ clusters randomized to the two research groups. Note that (17) can also be expressed as a weighted average of student-level treatment effects.

To develop estimators for the ATE parameter in (17), note first that for clustered designs, the design-based methods developed in Schochet (2016) average the individual data to the cluster level. Thus, the data generating process for the observed mean outcome for cluster j can be expressed as

$$(18) \quad \overline{y}_j = \sum_{k=1}^{K} T_j(k)\overline{Y}_j(k),$$

where $\overline{y}_j = (\sum_{i=1}^{n_j} y_{ij}/n_j)$.

Consider the weighted simple differences-in-means estimator using the aggregated data:

$$(19) \quad \hat{\beta}_{clus,FP}(k,k') = \overline{\overline{y}}_W(k) - \overline{\overline{y}}_W(k')$$
$$= \frac{\sum_{j:T_j(k)=1}^{m_k} w_j \overline{y}_j}{\sum_{j:T_j(k)=1}^{m_k} w_j} - \frac{\sum_{j:T_j(k')=1}^{m_{k'}} w_j \overline{y}_j}{\sum_{j:T_j(k')=1}^{m_{k'}} w_j} = \frac{\sum_{j=1}^{m} w_j T_j(k)\overline{Y}_j(k)}{\sum_{j=1}^{m} w_j T_j(k)} - \frac{\sum_{j=1}^{m} w_j T_j(k')\overline{Y}_j(k')}{\sum_{j=1}^{m} w_j T_j(k')},$$

where the third equality holds using (18). To examine the statistical properties of this estimator in the multi-armed context, we build on the properties of this estimator for the two-group design. Schochet (2016; Chapter 7) shows for the two-group design that in finite samples the estimator in (19) is biased for the parameter in (17) if the weights differ across clusters (and cluster-level ATEs are heterogeneous), because the denominators in (19) will depend on the particular allocation of clusters to the two research groups. However, as m gets large, $\hat{\beta}_{clus,FP}(k,k')$ is a consistent estimator of the following asymptotic ATE parameter:

$$(20) \quad E_{FP}[w_j(\overline{Y}_j(k) - \overline{Y}_j(k'))] / E_{FP}(w_j),$$

where $E_{FP}(.)$ signifies expectations (assumed to be fixed, nonnegative real numbers) over an increasing sequence of finite populations. Schochet (2016) shows also that $\hat{\beta}_{clus,FP}(k,k')$ is asymptotically normal with variance

$$(21)\quad AsyVar_R(\hat{\beta}_{clus,FP}(k,k')) = \frac{1}{E_{FP}(w_j)^2}\left[\frac{\overline{S}^2_W(k)}{m_k} + \frac{\overline{S}^2_W(k')}{m_{k'}} - \frac{\overline{S}^2_{\tau W}(k,k')}{m_k + m_{k'}}\right],$$

where $\overline{S}^2_W(k) = \lim \sum_{j=1}^{m} w_j^2(\overline{Y}_j(k) - \overline{\overline{Y}}_W(k))^2/(m-1)$ pertains to intervention k, similarly for $\overline{S}^2_W(k')$, and $\overline{S}^2_{\tau W}(k,k') = \lim \sum_{j=1}^{m} w_j^2\{(\overline{Y}_j(k) - \overline{\overline{Y}}_W(k)) - (\overline{Y}_j(k') - \overline{\overline{Y}}_W(k'))\}^2/(m-1)$ is the FP heterogeneity term. Accordingly, a consistent (upper-bound) variance estimator for (21) is

$$(22)\quad \hat{AsyVar}_R(\hat{\beta}_{clus,FP}(k,k') = \frac{s^2_W(k)}{\overline{w}^2_k m_k} + \frac{s^2_W(k')}{\overline{w}^2_{k'} m_{k'}} - \frac{1}{(m_k + m_{k'})}\left(\frac{s_W(k)}{\overline{w}_k} - \frac{s_W(k')}{\overline{w}_{k'}}\right)^2,$$

where

$$s^2_W(k) = \frac{1}{m_k - 1}\sum_{j:T_j(k)=1}^{m_k} w_j^2(\overline{y}_j - \overline{\overline{y}}_W(k))^2, \quad s^2_W(k') = \frac{1}{m_{k'} - 1}\sum_{j:T_j(k')=1}^{m_{k'}} w_j^2(\overline{y}_j - \overline{\overline{y}}_W(k'))^2,$$

$$\overline{w}_k = \frac{1}{m_k}\sum_{j:T_j(k)=1}^{m_k} w_j, \quad \overline{w}_{k'} = \frac{1}{m_{k'}}\sum_{j:T_j(k')=1}^{m_{k'}} w_j.$$

We can now extend these results to the multi-armed design using the same conditioning arguments as discussed in the previous section for the non-clustered design. First, we can show that the estimator in (19) is a consistent estimator for the ATE parameter in (20) using the law of iterated expectations, where, for arbitrarily large m, we first calculate expectations with respect to the randomization distribution (R),, conditional on the $(m_k + m_{k'})$ clusters assigned to the two research groups and their fixed potential outcomes, and then average over random draws of $(m_k + m_{k'})$ clusters from the population of m clusters. Thus, the only difference from the two-group design is that the increasing sequence of finite populations pertains to the entire randomized sample, not just to the clusters randomized to the two groups being compared.

Similarly, for large m, we can use the law of total variance to examine the asymptotic variance of $\hat{\beta}_{clus,FP}(k,k')$ in the multi-armed context. Following the same arguments as for the non-clustered design, we find similar variance expressions as in (21) and (22) above for the two-group design. The only difference is that the denominators in the FP heterogeneity terms contain m rather than $(m_k + m_{k'})$. The estimator in (19) is asymptotically normal.

Similar adjustments apply to variance expressions for regression-adjusted estimators, where separate weighted regression models with baseline covariates can be estimated for each pairwise contrast using

cluster-level means. The variance formulas in (7.21) and (7.22) in Schochet (2016) still apply except that the FP heterogeneity terms are divided by m rather than $(m_k + m_{k'})$.

Finally, no new issues arise for blocked, clustered designs. The formulas in Schochet (2016; Chapter 8) still apply in the multi-armed context, except that the block-specific weights for each pairwise contrast now pertain to the full randomized sample of clusters, and the FP heterogeneity terms in the variance estimators are now divided by m_b (the total number of clusters in the block) rather than $(m_{kb} + m_{k'b})$ (the number of clusters in the block randomized to the two contrasted groups). For subgroup analyses, the weights should be modified for each subgroup separately.

2. Multiple comparisons adjustments

In the analysis of data for multi-armed designs, it is good research practice to adjust for the inflation of Type 1 errors when conducting multiple hypothesis tests across pairwise contrasts (Hsu, 1996; Schochet, 2009; Westfall et al., 1999). For example, if there are four research groups, there are six possible pairwise contrasts for each outcome variable. Thus, applying a 5 percent significance level $(\alpha = .05)$ for each test for a single outcome variable, if all null hypotheses are true (that is, there are no differences between the interventions), the chance of finding at least one spurious impact is about 26 percent (assuming independent tests). This false positive error rate becomes even larger if we increase the number of hypothesis tests by adding more research groups and outcome variables. Thus, without accounting for the multiple comparisons being conducted, users of the study findings may draw incorrect conclusions regarding the most effective interventions. Clearly, the importance of adjusting for inflated Type 1 errors and how they should be balanced against Type 2 errors will depend on how the study findings will be used for policy.

There are a host of methods that have been developed to adjust for the multiple comparisons problem (Hsu, 1996 and Westfall et al., 1999 are excellent references). Most standard statistical packages such as SAS, R, and Stata implement many of these procedures. The purpose of this section is to briefly summarize these approaches, with an eye towards how they align with the non-parametric model that underlies the design-based framework. Our focus is on adjustment methods for testing all pairwise comparisons across the research groups (the most common testing strategy), although many of the same issues apply to other testing strategies as well (for example, pairwise comparisons relative to the control condition or to the best of the other treatments).[1]

a. Joint tests of no differences across interventions

When analyzing data from multi-armed RCTs, it is common practice to test the joint null hypothesis of no differences in mean outcomes across the tested interventions. In the design-based context,

[1] This section does not address multiple comparisons adjustment methods for selecting covariates to obtain regression-adjusted impact estimates or for selecting preferred model specifications for QED analyses.

because the impact estimators are asymptotically normal, the joint test for a single outcome can be conducted using the following chi-squared statistic:

$$(23) \quad ChiSq = \hat{\boldsymbol{\lambda}}' \hat{\boldsymbol{\Phi}}_{\lambda}^{-1} \hat{\boldsymbol{\lambda}},$$

where $\hat{\boldsymbol{\lambda}}$ is a $(K-1)x1$ vector of estimated contrasts for each of the $(K-1)$ interventions relative to the reference intervention (for example, intervention K), and $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\lambda}}$ is the associated $(K-1)x(K-1)$ variance-covariance matrix with the estimated variances from Section 1 along the diagonals and the estimated variance of the mean outcome for the reference intervention in the off-diagonal cells.[2] This chi-squared statistic has $(K-1)$ degrees of freedom. This same approach can be used for conducting a joint test for a particular subgroup (such as females) or for assessing baseline equivalency across the research groups for a particular baseline covariate.

Note that we use a chi-squared statistic rather than an *F*-statistic because under the design-based framework, variances are allowed to differ across research conditions. However, an approximate *F*-statistic can be computed by dividing (23) by $(K-1)$, which has numerator degrees of freedom equal to $(K-1)$ and denominator degrees of freedom that will depend on the sample size, the number of covariates, and the number of blocks.

With multiple outcomes, the research question of interest for the joint test is: Are there any differences between the interventions for any outcome? In this case, the test statistic in (23) for the non-clustered design can be applied by (i) stacking the $\hat{\boldsymbol{\lambda}}$ vectors for each of the Q outcomes and (ii) stacking the $\hat{\boldsymbol{\Phi}}_{\boldsymbol{\lambda}}$ matrices for each outcome along the diagonal blocks of a larger matrix, $\hat{\boldsymbol{\Phi}}_{\lambda}^{*}$, where the off-diagonal blocks contain the covariances, $Cov(\bar{y}_{kq} - \bar{y}_{Kq}, \bar{y}_{k'q'} - \bar{y}_{Kq'})$, where $\bar{y}_{kq}$ is the mean value for outcome variable q for research group k (assuming research group K is the reference group). These covariances take the form $[Cov(\bar{y}_{kq}, \bar{y}_{kq'}) + Cov(\bar{y}_{Kq}, \bar{y}_{Kq'})]$ if $k = k'$, and $Cov(\bar{y}_{Kq}, \bar{y}_{Kq'})$ if $k \neq k'$, and can be estimated using parallel expressions as for the variances of the estimated contrasts presented earlier. For instance, for the simple differences-in-means estimator, we can estimate $Cov(\bar{y}_{kq}, \bar{y}_{kq'})$ using

$$(24) \quad \hat{Cov}(\bar{y}_{kq}, \bar{y}_{kq'}) = \frac{1}{n_k(n_k - 1)} \sum_{i:T_i(k)=1}^{n_k} (y_{iq} - \bar{y}_{kq})(y_{iq'} - \bar{y}_{kq'}),$$

[2] For regression-adjusted estimates, the variances of the mean outcome for the reference intervention will differ across the pairwise contrasts because a separate regression model is estimated for each contrast. One possible approach is to use a weighted average of these variances, where the weights are proportional to the sample size for each pairwise contrast.

and similarly for the other covariances. The resulting chi-squared statistic has $Q(K-1)$ degrees of freedom. Similar methods can be used for regression-adjusted estimators and clustered designs.

The same approach can be used for subgroup analyses to test the null hypothesis of no differences in intervention effects across subgroup levels (for example, by gender). To illustrate, assume a single outcome variable, and let the G subgroup levels be indexed using the subscript g, so that $\bar{y}_{kg}$ is the mean outcome for subgroup g in research group k. The $\hat{\boldsymbol{\lambda}}$ vector is now of dimension $(K-1)(G-1)x1$ with elements of the form $[(\bar{y}_{kg}-\bar{y}_{Kg})-(\bar{y}_{kG}-\bar{y}_{KG})]$, where subgroup G is assumed to be the reference subgroup (and research group K remains the reference research group). The associated variance-covariance matrix, $\hat{\boldsymbol{\Phi}}_{\lambda}$, has diagonal elements of the form $[Var(\bar{y}_{kg}-\bar{y}_{Kg})+Var(\bar{y}_{kG}-\bar{y}_{KG})]$, which can be estimated using the design-based variance estimators for each subgroup, and has off-diagonal elements of the form $[Var(\bar{y}_{Kg})+Var(\bar{y}_{KG})]$, which can also be estimated using the design-based estimators. The resulting chi-squared statistic has $(K-1)(G-1)$ degrees of freedom. This approach can be generalized to incorporate multiple outcomes (not shown).

b. Assessing which pairwise contrasts differ

The joint tests discussed in the previous section can efficiently address the question: Do any of the tested contrasts differ? However, they do not address the question: Which contrasts differ? This issue requires adjusting the individual inferences to control the chances of finding a spurious impact when the entire family of inferences is considered.

The familywise error rate (FWER) defined by Tukey (1953) has traditionally been the focus of research in the multiple comparisons area. The FWER is the probability that at least one null hypothesis will be rejected when all null hypotheses are true (it equals $1-(1-\alpha)^N$ for N independent tests). The most well-known method is the Bonferroni procedure, which sets the significance level for individual tests at (α/N). The Bonferroni procedure controls the FWER when all null hypotheses are true or when some are true and some are false (that is, it provides strong control of the FWER). The Bonferroni method applies to both continuous and discrete data, controls the FWER when the tests are correlated, and provides adjusted confidence bounds by using α/N rather than α in the calculations. Many modified and sometimes more powerful versions of the Bonferroni method have been developed that provide strong control of the FWER, such as the Šidák (1967), Scheffé (1959), Holm (1979), Hochberg (1988), and Rom (1990) methods.

The Bonferroni-type methods align with the design-based framework in that they do not require assumptions on the distributions of the potential outcomes or the model structure. However, by ignoring the correlational structure across tests, and in particular the repetition of samples across contrasts in the multi-armed context, the Bonferroni-type methods yield conservative bounds on

Type I error and, hence, sacrifice statistical power. For example, in a design with two treatment groups (T1 and T2) and a control group (C), there are three possible pairwise contrasts (T1-T2, T1-C, and T2-C). Each research group enters two of these contrasts, yielding test statistics that are correlated. Ignoring these correlations can reduce precision, as demonstrated later in this section using simulations.

Similar issues pertain to the Benjamini and Hochberg (1995) adjustment method that controls the false discovery rate (FDR)—the expected fraction of significant test statistics that are false discoveries—instead of the FWER. This approach—which is similar to the Bonferroni-type methods in that it is based only on *p*-values for each test in isolation—can lead to power gains if there are many contrasts that truly differ. However, the FDR uses a preponderance-of-evidence standard that allows for extra false positives if many contrasts are found to be statistically significant. Thus, the FDR evidence standard is less stringent than the FWER standard, and may not always be appropriate for multi-armed RCTs that require a high bar of evidence to identify the most effective treatments.

The Tukey-Kramer (Tukey 1953, Kramer 1956) method is a common approach used in multi-armed RCTs for controlling the FWER if the family of tests consists of all pairwise comparisons across the research groups; the Dunnett (1955) method is a parallel approach if comparisons with a single control group are of primary interest. These methods are more powerful than the Bonferroni-type methods because they directly account for the dependence across test statistics due to the repetition of the research group samples across contrasts.

The Tukey-Kramer approach calculates *t*-test cutoff values, $q_{1-\alpha,K,df}$, using the studentized range distribution—the distribution, under the null hypothesis, of the maximum *t*-statistic across the pairwise contrasts. This approach assumes that potential outcomes are independent and normally distributed and have constant variances. The degrees of freedom for this test is $df = (n - K)$ for the simple differences-in-means estimator and non-clustered design (and similarly for regression-adjusted estimators and clustered designs). Hayter (1984) proved that the FWER for this method will be less than α if sample sizes differ across the research groups and will equal α for balanced designs.

Table 1 provides simulation results to assess precision gains using the Tukey-Kramer method relative to the Bonferroni method. The simulations assume (i) a single confirmatory outcome variable; (ii) $\alpha = .05$ and a two-tailed test; (iii) 3 to 10 research groups (K); and (iv) 10 or 50 randomized units per research group. To generate *t*-distribution critical values, we used a significance level of $\alpha^* = \alpha / N$ for the Bonferroni method where $N = K(K-1)/2$ is the number of pairwise contrasts, and the probmc function in SAS for the Tukey-Kramer method.[3] For both methods, the table displays: (i) *t*-distribution critical values for individual tests to determine statistical significance; (ii)

[3] We divided the cutoff value that the probmc function generated by $\sqrt{2}$ to adjust for the way SAS scales the statistic.

associated *p*-value cutoffs (α^* values); and (iii) ratios of original to adjusted significance levels, $FACTOR = \alpha / \alpha^*$, which are useful for power calculations when designing multi-armed RCTs.

Table 1. Cutoff values for significance testing using the Bonferroni and Tukey-Kramer methods

Number of Research Groups	*t*-distribution cutoff values		*p*-value cutoff values (α*)		*FACTOR* = (.05/α*)	
	Bonferroni	Tukey-Kramer	Bonferroni	Tukey-Kramer	Bonferroni	Tukey-Kramer
10 randomized units per research group						
3	2.55	2.48	0.0167	0.0197	3	2.54
4	2.78	2.70	0.0083	0.0107	6	4.67
5	2.94	2.85	0.0050	0.0068	10	7.38
6	3.06	2.98	0.0033	0.0047	15	10.67
7	3.16	3.07	0.0024	0.0034	21	14.50
8	3.23	3.16	0.0018	0.0026	28	18.88
9	3.30	3.23	0.0014	0.0021	36	23.76
10	3.36	3.30	0.0011	0.0017	45	29.11
50 randomized units per research group						
3	2.42	2.37	0.0167	0.0192	3	2.60
4	2.67	2.59	0.0083	0.0103	6	4.86
5	2.83	2.75	0.0050	0.0064	10	7.76
6	2.96	2.87	0.0033	0.0044	15	11.31
7	3.06	2.97	0.0024	0.0032	21	15.48
8	3.14	3.05	0.0018	0.0025	28	20.27
9	3.22	3.12	0.0014	0.0019	36	25.69
10	3.27	3.18	0.0011	0.0016	45	31.73

The results suggest there are some precision gains to using the Tukey-Kramer method. Using the *t*-distribution cutoff values, the Tukey-Kramer method will yield confidence intervals that are about 2 to 3 percent smaller than the Bonferroni intervals for the considered sample size combinations. The *FACTOR* values for the Tukey-Kramer method are smaller than for the Bonferroni method, which similarly indicates that there are some precision gains to using the Tukey-Kramer method.

It is important to note that the Tukey-Kramer procedure does not fully align with design-based theory, because it assumes the same variances across the research groups, whereas the design-based estimators allow for different variances (to account for heterogeneity of treatment effects). Several procedures have been developed to modify the Tukey-Kramer approach when population variances differ. For example, for the non-clustered design, Games and Howell (1976) suggest using the Tukey-Kramer cutoff values with degrees of freedom equal to the integer value of

$$(25) \quad df = \left[\frac{s_I^2(k)}{n_k} + \frac{s_I^2(k')}{n_{k'}} \right]^2 / \left(\frac{1}{n_k - 1} \left[\frac{s_I^2(k)}{n_k} \right]^2 + \frac{1}{n_{k'} - 1} \left[\frac{s_I^2(k')}{n_{k'}} \right]^2 \right),$$

which can be easily adapted to the clustered design.[4] Similarly, Dunnett's C method (1980) uses a precision-weighted average of *t*-statistic cutoff values across the two contrasted groups:

$$(26) \quad q = \left[q_{1-\alpha, K, (n_k - 1)} \frac{s_I^2(k)}{n_k} + q_{1-\alpha, K, (n_{k'} - 1)} \frac{s_I^2(k')}{n_{k'}} \right] / \left[\frac{s_I^2(k)}{n_k} + \frac{s_I^2(k')}{n_{k'}} \right].$$

These methods have been shown to perform well in simulations (Olejnik and Lee, 1990).

Bootstrap and permutation resampling methods are alternative, more computer-intensive methods that provide strong control of the FWER (see, for example, Westfall and Young 1993). These methods incorporate distributional and correlational structures across tests, so they tend to be less conservative than the Bonferroni-type methods and, hence, may have more power. Furthermore, they are applicable in many testing situations (for example, for testing impacts on medians or quartiles rather than means), and have the advantage that the normality assumption can be relaxed. Although these methods assume homogenous variances across research groups, simulations have shown that they perform well except under extreme heteroscedasticity (Westfall and Young, 1993).

Finally, the above methods can be generalized to simultaneously adjust for multiplicity across multiple research groups and multiple outcome variables. For instance, the Bonferroni-type and Benjamini-Hochberg methods can be applied by stacking *p*-values across the full set of contrasts. One approach that has not been addressed in the literature (to the author's knowledge) is to combine the Bonferroni and Tukey-Kramer-like procedures by (i) calculating $\alpha^* = \alpha / Q,$ where Q is the number of outcome variables and (ii) using α^* rather than α in the Tukey-Kramer procedure. The resampling methods can also be used in this context and can lead to power gains by accounting for correlations in test statistics across both research groups and outcome variables. Similar methods can be used for conducting hypothesis tests across multiple levels of a population subgroup.

[4]Equation (25) can include the FP heterogeneity term in both the numerator (using Equation (10)) and in the denominator by subtracting the term $([s_I(k) - s_I(k')]^2 / n)^2 / (n-1)$.

In sum, it is good research practice to adjust for multiple comparisons in multi-armed RCTs when conducting significance tests across pairwise contrasts. This issue holds regardless of the methods used to estimate intervention effects. From a design-based perspective, the Bonferroni-type and Benjamini-Hochberg methods align well with the non-parametric design-based methods in that they do not rely on assumptions regarding distributions of the potential outcomes and the model structure. The Tukey-Kramer and related methods require more assumptions (for example, normally distributed outcomes), but are widely used and can yield some power gains. The resampling methods are perhaps the best alternative for design-based analyses because they require minimal assumptions (although they do assume equal group variances) and account for correlations across the test statistics. However, more research needs to be conducted to assess whether the resampling methods apply to the full range of design-based estimators for the finite- and super-population frameworks.

3. Identification and Estimation of the complier average causal effect (CACE) parameter

For RCTs, the CACE parameter pertains to average treatment effects for compliers in the study population—those who would *receive* intervention services offered to their research group but not interventions slated for other groups (Angrist, Imbens, and Rubin, 1996; Bloom, 1984). If we conceptualize compliance decisions as dichotomous, Angrist et al. (1996) derived the assumptions required to identify the CACE parameter for the single treatment-control, non-clustered RCT using an instrumental variable (IV) framework. Schochet and Chiang (2011) generalized these findings to the clustered design where compliance decisions can be made by both clusters (for instance, school staff) as well as individuals within clusters (for instance, students within schools).

In this section, we show that the identifying assumptions for the CACE parameter for the single treatment-control group design do not easily generalize to the multi-armed context (Cheng and Small, 2006; Long, Little, and Lin, 2010; Blackwell, 2016). Our main purpose is to provide motivation on why CACE estimation may not always be possible for multi-armed RCTs, cite the relevant literature, and provide suggestions on additional model assumptions that could be invoked to estimate or at least bound the CACE parameters in the multi-armed context.

We adopt an IV framework because it aligns well with the non-parametric underpinnings of the design-based framework; we do not consider other estimation approaches, such as Bayesian principal stratification (Frangakis and Rubin, 2002; Long, Little, and Lin, 2010) and structural nested mean models (Robins, 1994, Tan 2010) that rely on alternative, parametric assumptions to identify CACE effects. To simplify the presentation, our focus is on CACE identification for non-clustered designs using the super-population framework; parallel issues arise for clustered designs and the finite-population model. Finally, we consider designs where individuals are randomized to multiple treatment groups at the same time, and not designs where treatments are randomized sequentially (Blackwell, 2016).

a. Notation and base assumptions

We assume an RCT with K research groups and K^* treatment groups (note that $K^* = K$ if the study has no control group, and $K^* = K - 1$ otherwise). Let $D_i(k) = D_i(k, Q_i)$ denote an indicator variable that equals 1 if individual i would *receive* intervention k if assigned to research group Q_i ($k = 1, ..., K^*; Q_i = q, ..., K^*$, where $q = 1$ if there is no control group, and $q = 0$ otherwise). This notation allows for crossovers (non-compliers) in one research group to receive interventions slated for other research groups for whatever reason. Let $Y_i(Q_i, D_i(1), D_i(2), ..., D_i(K^*))$ denote the individual's potential outcome for a given value of $(Q_i, \mathbf{D_i})$, where $\mathbf{D_i}$ is a vector containing the $D_i(k)$ indicators.

For each value of Q_i, there are 2^{K^*} possible $\mathbf{D_i}$ combinations. Thus, there are 2^{K^*K} possible population subgroups based on their compliance decisions across the K research conditions. For example, if the study contains two treatment groups and a control group, then $K = 3$ and $K^* = 2$ leading to 64 (2^6) possible compliance subgroups. As we shall see, these large number of cells makes it difficult to identify the CACE parameters for multi-armed RCTs.

Angrist et al. (1996) specified three key assumptions for identifying the CACE parameter for the single treatment-control RCT that we generalize to the multi-armed setting (in addition to the assumptions required to identify the ATE parameter discussed in Section 1):

1. **SUTVA:** $D_i(k)$ and $Y_i(Q_i, \mathbf{D_i})$ are unrelated to the research status of other individuals, and $Y_i(Q_i, \mathbf{D_i})$ is unrelated to the intervention receipt status of other individuals. This allows us to express $Y_i(Q_i, \mathbf{D_i})$ in terms of Q_i and $\mathbf{D_i}$ rather than the vector of treatment and service receipt statuses of all individuals.
2. **Monotonicity:** $D_i(k, k) \geq D_i(k, k')$, which means that individuals are at least as likely to receive intervention k if assigned to research group k $(Q_i = k)$ than if assigned to another research group $(Q_i = k')$. We also assume a positive rate of receipt of intervention k for those offered intervention k.
3. **Exclusion restriction:** $Y_i(k, \mathbf{D_i}) = Y_i(k', \mathbf{D_i})$, which means that the outcome for an individual who receives a particular set of intervention services would be the same regardless of the assigned research condition. For example, if an individual would receive Treatment 1 if assigned to either Research Group 1 or 2, then the potential outcomes of that person are the same in Research Groups 1 and 2.

b. Identification and Estimation

For the single treatment-control group design, where Q_i equals 1 for the treatment group and 0 for the control group, the three assumptions from above identify the CACE parameter, $E_I(Y_i(1,1)-Y_i(0,0))$. This can be seen by expressing the overall ATE parameter, $E_I(Y_i(1)-Y_i(0))$, as a weighted average of ATEs for four mutually exclusive compliance subgroups: (i) *compliers* who would receive treatment services only in the treatment condition $[D_i(1,1)=1$ and $D_i(1,0)=0]$; (ii) *never-takers* who would never receive treatment services $[D_i(1,1)=0$ and $D_i(1,0)=0]$; (iii) *always-takers* who would always receive treatment services $[D_i(1,1)=1$ and $D_i(1,0)=1]$; and (iv) *defiers* who would receive treatment services only in the control condition $[D_i(1,1)=0$ and $D_i(1,0)=1]$.

Monotonicity implies that there are no defiers and the exclusion restriction implies that ATEs are zero for never-takers and always-takers. Thus, the overall ATE parameter can be expressed as a function of the ATE for compliers only. Stated differently, the CACE parameter, $E_I(Y_i(1,1)-Y_i(0,0))$ can be identified from the data using $E_I(Y_i(1)-Y_i(0))/p_{CL}$, where p_{CL} is the proportion of compliers in the study population (which can be shown to be the difference between intervention receipt rates in the treatment and control conditions). Schochet (2016) discusses design-based estimation of this CACE parameter.

The situation becomes more complex in the multi-armed context. Consider first a simple two-group design with two treatment groups (Research Groups 1 and 2) but no control group $(K=K^*=2)$. In this case, the overall ATE parameter can be expressed as a weighted average of ATEs for 16 possible compliance groups, because in each research group, individuals can receive Treatment 1 only (labeled "T1"), Treatment 2 only ("T2"), both treatments ("T12"), or neither treatment ("T0"). To simplify notation, define the couplet (a,b), where "a" and "b" represent a person's compliance decisions if assigned to Research Groups 1 and 2, respectively. For example, (T0,T2) represents those who would receive no intervention services if assigned to Research Group 1 and would receive Treatment 2 if assigned to Research Group 2.

Consider a simplified scenario where there are no crossovers, so that those in Research Group 1 cannot receive Treatment 2 and vice versa. In this case, as shown in Figure 1, the overall ATE parameter, $E_I(Y_i(2)-Y_i(1))$, can be expressed as a weighted average of ATEs for four compliance subgroups: (T1,T2), (T1,T0), (T0,T2), and (T0,T0). Note that the condition of no crossovers is more stringent than the monotonicity condition (which leaves 9 complier subgroups rather than 4), because the absence of crossovers removes compliance subgroups such as (T1,T1) and (T1,T12) that are not excluded by monotonicity.

Figure 1: Depiction of compliance decisions for an RCT with two treatment groups and no crossovers

		Receive Treatment 2 if assigned to Research Group 2		
		Yes	**No**	**% Receiving Treatment 1**
Receive Treatment 1 if assigned to Research Group 1	**Yes**	(T1,T2)	(T1,T0)	p_1
	No	(T0,T2)	(T0,T0)	$(1-p_1)$
	% Receiving Treatment 2	p_2	$(1-p_2)$	1

Note: $p_1>0$ is the proportion of the study population who would receive Treatment 1 if assigned to Research Group 1 and similarly for $p_2>0$.

The first question to address is: What is the CACE parameter of interest? The (T1,T2) ATE parameter is probably of most interest because it represents the relative effect of receiving Treatment 1 compared to Treatment 2 for compliers in both research groups (always-compliers). However, interest might also lie in the (T1,T0) parameter that measures the effects of receiving Treatment 1 relative to the status quo condition for Research Group 1 compliers.

In either case, these CACE parameters are not identified without additional assumptions. While the exclusion restriction implies that the ATE for the (T0,T0) subgroup is zero, this restriction does not apply to the three remaining compliance subgroups. Further complicating identification, we cannot even identify the proportions of the study population in each compliance subgroup. Cheng and Small (1996) provide bounds on both the subpopulation proportions and ATE parameters assuming binary outcomes (for a design that also includes a control group). However, with more than two treatment groups, bounds become more difficult to calculate and are likely to be non-informative (wide). In addition, if the study contains crossovers, the number of compliance subgroups increases, leading to intractable identification conditions and bounds.

It is possible to make some headway under the two-treatment design in settings where it is realistic to assume an additional monotonicity condition where one treatment is always preferred to the other. Under this condition, a person who takes up the offer of the less preferred treatment will always take up the offer of the more preferred treatment. For instance, if Treatment 2 is the preferred intervention, the monotonicity condition implies that $D_i(2,2) \geq D_i(1,1)$. In this case, there are no individuals in the (T1,T0) subgroup, and treatment take-up rates will always be larger in Research Group 2 than in Research Group 1 (that is, $p_2 > p_1$; see Figure 1 above).

With this additional monotonicity condition, it is possible to identify the proportions of the population in each compliance subgroup: p_1 for the (T1,T2) subgroup, $(p_2 - p_1)$ for the (T0,T2)

subgroup, and $(1-p_2)$ for the (T0,T0) subgroup. Furthermore, because the exclusion restriction implies that the ATE for the (T0,T0) group is zero, we can express the full-sample ATE as a weighted average of ATEs for only two compliance subgroups: (T1,T2) and (T0,T2):

$$(27) \quad ATE = p_1 ATE(T1,T2) + (p_2 - p_1) ATE(T0,T2),$$

where, to simplify the presentation, we use the notation $ATE = E_I(Y_i(2) - Y_i(1))$, which is the overall average treatment effect; $ATE(T1,T2) = E_I(Y_i(2,0,1) - Y_i(1,1,0))$, which compares the effects of Treatment 2 to 1 for the always-compliers; and $ATE(T0,T2) = E_I(Y_i(2,0,1) - Y_i(1,0,0))$, which compares the effects of Treatment 2 to no treatment for compliers in Group 2 but not in Group 1.

If we divide the relation in (27) by p_2, the CACE parameter (ATE / p_2) can be identified as a weighted average of ATEs for the (T1,T2) and (T0,T2) subgroups, with weights (p_1 / p_2) and $(p_2 - p_1)/p_2$, respectively. This parameter measures the effects of receiving Treatment 2 for Research Group 2 compliers, but is somewhat difficult to interpret because the counterfactual condition includes both those who received Treatment 1 and those who did not. Perhaps a better justification for this CACE parameter is that it serves as a *lower* bound on the $ATE(T1,T2)$ parameter if we assume that $ATE(T0,T2) \geq ATE(T1,T2)$, which is plausible since we might expect intervention effects to be larger when comparing an intervention to the status quo condition than to another intervention (the author has not seen this bound discussed in the literature). Design-based estimators and standard errors for the (ATE / p_2) parameter can be obtained using results in Schochet (2016) for the single treatment-control group design.

This bounding approach, can be generalized to settings with additional treatment groups if there is an assumed monotone ordering of treatment preferences, there are no crossovers, and we invoke additional exclusion restrictions described below. For example, suppose we added a third treatment group (Research Group 3) which is preferred to Treatment 2, which in turn is preferred to Treatment 1, and extend the notation from Figure 1 to the three-group design. We then have four possible compliance subgroups: (T1,T2,T3), (T0,T2,T3), (T0,T0,T3), and (T0,T0,T0) with respective subpopulation proportions, p_1, $(p_2 - p_1)$, $(p_3 - p_2)$, and $(1 - p_3)$. Contrasting Treatments 2 and 3, we can obtain a lower bound on the CACE parameter for the (T1,T2,T3) subgroup (the always-compliers) using $(ATE_{2,3} / p_3)$, where $ATE_{2,3} = E_I(Y_i(2) - Y_i(3))$ is the full-sample ATE for the Treatment 2-3 contrast. However, we need to invoke two additional assumptions. First, we must invoke an extra exclusion restriction:

$$(28) \quad ATE_{2,3}(T1,T2,T3) = ATE_{2,3}(T0,T2,T3),$$

which states that the Treatment 2-3 contrast for Research Group 2 and 3 compliers is the same for those who would take up the offer of Treatment 1 if given the chance and those who would not. Second, we must assume that $ATE_{2,3}(T0,T0,T3) \geq ATE_{2,3}(T1,T2,T3)$. Parallel arguments and assumptions can be used to obtain lower bounds for the Treatment 1-3 contrast $(ATE_{1,3} / p_3)$ and Treatment 1-2 contrast $(ATE_{1,2} / p_2)$ for the always-compliers.

Multi-armed RCTs with control groups allow for additional possibilities to identify CACE parameters without the added monotonicity condition, assuming no crossovers. To fix concepts, consider a design with a control group (Research Group 0) and two treatment groups (Research Groups 1 and 2). In this setting, $K = 3$ and $K^* = 2$, so the overall ATE parameter can be expressed as a weighted average of ATEs for 64 possible compliance subgroups. Define the triplet (a,b,c), where "a", "b", and "c" represent a sample member's compliance decisions if assigned to Research Groups 0, 1, and 2, respectively. Assuming no crossovers, the structure of Figure 1 still holds except that the four compliance subgroups now include the triplets (T0,T1,T2), (T0,T1,T0), (T0,T0,T2), and (T0,T0,T0) rather than the couplets. These four subgroups exist in each research group.

With three research groups, there are three possible pairwise contrasts of interest comparing each treatment group to each other and to the control group. Thus, additional CACE parameters emerge across the contrasts. When contrasting the two treatment groups, the same issues with CACE identification arise as for the two-treatment design discussed above. However, the situation becomes more tractable when contrasting each treatment group to the control group.

Consider comparing the first treatment group (Research Group 1) to the control group (Research Group 0). Using Figure 1, the full-sample ATE for this contrast, $ATE_{1,0}$, can be expressed as a weighted average of treatment-control contrasts for four complier subgroups:

$$(29) \quad ATE_{1,0} = \pi_1 ATE_{1,0}(T0,T1,T2) + \pi_2 ATE_{1,0}(T0,T1,T0) + \pi_3 ATE_{1,0}(T0,T0,T2) + \pi_4 ATE_{1,0}(T0,T0,T0),$$

where $\pi_1,...,\pi_4$ are unobserved subgroup proportions. The exclusion restriction removes the final two terms from (29). Furthermore, we have from Figure 1 that $p_1 = \pi_1 + \pi_2$. Suppose we invoke the following additional exclusion restriction similar to (28):

$$(30) \quad ATE_{1,0}(T0,T1,T2) = ATE_{1,0}(T0,T1,T0),$$

which states that the average treatment effect for those receiving Treatment 1 is the same for those who would take up the offer of Treatment 2 if given the chance and those who would not. In this case, the CACE parameter for Research Group 1 compliers can be identified as $CACE_{1,0} = (ATE_{1,0} / p_1)$. Note that the assumption in (30) is not needed to identify $CACE_{1,0}$ if we

instead invoke the monotonicity condition that Treatment 2 is preferred to Treatment 1 (but not vice versa). A parallel argument identifies the $CACE_{2,0} = (ATE_{2,0} / p_2)$ parameter when comparing the second treatment group (Research Group 2) to the control group; for that contrast, the parallel exclusion restriction to (30) is $ATE_{2,0}(T0,T1,T2) = ATE_{2,0}(T0,T0,T2)$. This same approach generalizes to designs with more than two treatment groups.

The above analysis suggests that in multi-armed RCTs with control groups, we can identify well-defined CACE parameters by simply comparing each treatment group to the control group, and can estimate them using the IV estimators found in the literature for the single treatment-control design. This approach, however, relies on the assumptions of no crossovers (or a small number of crossovers who can safely be ignored in the analysis) and additional exclusion restrictions of the form shown in (30). The multiple comparisons methods discussed in the previous section can be employed for significance testing across the multiple CACE estimators.

An important potential limitation of the above approach, however, is that the $CACE_{k,0}$ parameters comparing each treatment group to the control group refer to *different* complier subpopulations (Long, Little and Lin, 2011). For example, in the above scenario, the $CACE_{1,0}$ parameter refers to those in the (T0,T1,T2) and (T0,T1,T0) subgroups (Research Group 0 and 1 compliers), whereas the $CACE_{2,0}$ parameter refers to the those in the (T0,T1,T2) and (T0,T0,T2) subgroups (Research Group 0 and 2 compliers). These subgroups do not fully overlap.

If it is reasonable to assume that treatment-control differences are homogeneous across compliance subgroups (perhaps, based on the similarity of ATE impact findings across baseline subgroups or other ancillary information), the $CACE_{1,0}$ and $CACE_{2,0}$ parameters could be compared. In this case, the $CACE_{1,2}$ parameter contrasting Treatments 1 and 2 can be estimated using $\hat{CACE}_{1,2} = \hat{CACE}_{1,0} - \hat{CACE}_{2,0}$, where we use the "^" symbol to denote estimators. Note that this parameter pertains to all compliance subgroups excluding the never-takers. A variance estimator for $\hat{CACE}_{1,2}$ can be obtained using the expression $\hat{Var}(\hat{CACE}_{1,0}) + \hat{Var}(\hat{CACE}_{2,0}) - 2\hat{Var}(\bar{y}(0)/(\hat{p}_1\hat{p}_2))$, where $\bar{y}(0)$ is the control group mean that enters both $\hat{CACE}_{1,0}$ and $\hat{CACE}_{2,0}$. These variance components can be estimated, for example, using the asymptotic variance formulas for CACE analyses based on Taylor series approximations presented in Schochet (2016; Chapter 5, Section l).

In sum, estimating CACE parameters in the multi-armed context is more complex than for the two-group, treatment-control design. Even in a relatively simple RCT with only two treatment groups, the CACE parameter for the always-compliers—the group typically of most interest—is not identified, even in the absence of crossovers. These effects can be bounded using linear programming methods (for bounded outcome variables), but these methods are complex and do not easily generalize to designs with more than two treatment groups. If it is reasonable to assume a monotone ordering of

treatment preferences, we can obtain sharper bounds using the linear programming approach, and can calculate reasonable lower bounds on CACE parameters for the always-compliers.

If the study includes a control group, the simplest approach is to first estimate CACE parameters by contrasting each treatment group to the control group and then to compare these CACE estimates. However, this approach is feasible only in the absence of crossovers and requires additional exclusion restrictions. Furthermore, because each of these CACE estimators pertains to different subpopulations, it may not be possible to compare them unless it is realistic to assume homogeneity of treatment-control contrasts across subpopulations.

There are no easy solutions for dealing with non-compliance in multi-armed RCTs. Clearly, the credibility of imposing specific and sometimes untestable assumptions to identify CACE parameters will depend on the study context. Researchers should be aware that there may be instances where it is not possible to estimate CACE effects for studies with multiple treatment groups. Developing simple design-based solutions to adjust for non-compliance in multi-armed RCTs is a ripe area for future research.

References

Agodini, R., B. Harris, S. Atkins-Burnett, S. Heaviside, T. Novak, R. Murphy (2009). *Achievement effects of four early elementary school math curricula: Findings from first graders in 39 schools*. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*, 444–472.

Benjamini, Y. & Y. Hochberg (1995). Controlling the false discovery rate: A new and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B,* 57, 1289-1300.

Blackwell, M. (2016). Instrumental variable methods for conditional effects and causal interaction in voter mobilization experiments. *Institute for Quantitative Social Sciences Working Paper*: Harvard University.

Bloom, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* 8(2), 225-246.

Box, G. E., J. S. Hunter, & W. G. Hunter, W. G. (2005). *Statistics for experiments: Design, innovation, and discovery* (2nd ed.). New York, NY: Wiley.

Castleman, B.L. and L.C. Page (2015). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic Behavior & Organization*, 115, 144-160.

Cheng, J. & D. Small (2006). Bounds on causal effects in three-arm trials with non-compliance. *Journal of the Royal Statistical Society, Series B,* 68(5), 815-837.

Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association,* 50, 1096-1121.

Dunnett, C.W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association,* 75, 796-800.

Games, P.A. & J. F. Howell (1976), Pairwise multiple comparison procedures with unequal Ns and/or variances: A monte carlo study. *Journal of Education Statistics,* 1, 113-125.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Statistics,* 12(1), 61-75.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika,* 75, 800-802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics,* 6, 65-70.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association,* 81(396), 945–960.

Hsu, J.C. (1996). *Multiple comparisons: theory and methods.* London: Chapman and Hall.

Imbens, G. & Rubin, D. (2015). Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge, UK: Cambridge University Press.

James-Burdumy, S. et al. (2009). *Effectiveness of selected supplemental reading comprehension interventions*. Washington, DC: U.S. Department of Education Institute of Education Sciences.

Kramer, C.Y. (1956). Extension of the multiple range test to group means with unequal numbers of replications. *Biometrics,* 12, 307-310.

Kraft, M.A. (2014). *The underutilized potential of teacher-to-parent communication: Evidence from a field experiment*. Cambridge MA: Harvard Kennedy School Working Paper RWP14-049.

Long, Qi, R. Little, & X. Lin (2010). Estimating causal effects in trials involving multiple-treatment arms subject to non-compliance: A Bayesian framework. *Journal of the Royal Statistical Society Series C,* 59(3), 513-531.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles. Section 9, Translated in *Statistical Science*, 1990: 5(4).

Olejnik, S. & J. Lee (1990). Multiple comparison procedures when population variances differ. *Paper presented at the annual meeting of the American Education Research Association.*

Robins, J.M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics,* 23(8), 2379-2412.

Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika,* 77, 663-665.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Education Psychology,* 66, 688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Education Statistics, 2*(1), 1–26.

Scheffé, H. (1959). *The Analysis of Variance.* New York: John Wiley & Sons.

Schochet, P. Z. (2009). An approach for addressing the multiple testing problem in social policy impact evaluations. *Evaluation Review, 33*(6), 539–567.

Schochet, P. Z. & H. Chiang (2011). Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Educational and Behavioral Statistics,* 36, 307-345.

Schochet, P. Z. (2016 Second Edition; 2015 First Edition). *Statistical theory for the RCT-YES software: Design-based causal inference for RCTs* (NCEE 2015–4011). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/pubs/20154011/pdf/20154011.pdf.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association,* 62, 626-633.

Tan, Z. (2010). Marginal and nested structural models using instrumental variables. *Journal of the American Statistical Association,* 105, 157-169.

Toledo, C., Humpage-Liuzzi, S., Murray, N., & Glazerman S. (2015). *Data-driven instruction in Honduras: An impact evaluation of the EducAccion Promising Reading intervention*. AEA RCT Registry: https://www.socialscienceregistry.org/trials/780

Tukey, J.W. (1953). The problem of multiple comparisons. In *Mimeographed Notes.* Princeton, NJ: Princeton University.

Westfall, P.H, Y. Lin, Y., & S. Young (1990). Resampling-based multiple testing. In *Proceedings of the Fifteenth Annual SAS Users Group International.* Cary, NC: SAS Institute, Inc., 1359-1364.

Westfall, P.H., R. Tobias, D. Rom, R. Wolfinger, & Y. Hochberg (1999). *Multiple Comparisons and Multiple Tests Using SAS.* Cary, NC: SAS Institute, Inc.

Westfall, P.H. & S.S. Young (1993). Resampling-based multiple testing: Examples and methods for p-value adjustment. New York: John Wiley & Sons.

Wu, C. F. J., & M. S. Hamada (2009). *Experiments: Planning, analysis and parameter design optimization* (2nd ed.). Hoboken, NJ: John Wiley and Sons, Inc.