How Three-Arm Random Assignment within Sites Can Improve Non-Experimental Cross-Site Estimates of the Relationship between Program Characteristics and Impact

by
Laura R. Peck
Shawn R. Moulton

REVISED DRAFT for APPAM
October 26, 2018

Abstract
This paper considers a new method, called Cross-Site Attributional Model Improved by Calibration to Within-Site Individual Randomization Findings (abbreviated as “CAMIC”), which seeks to reduce bias in analyses of the contributions of a program’s various design, implementation and contextual characteristics to its overall impacts. It requires a multi-site experiment where some (or all) sites randomize individuals to one of three arms: a standard treatment group, an enhanced treatment group (that receives the standard treatment plus a program “enhancement”), or a control group (that has no access to the program). A recent evaluation—that of the Health Profession Opportunity Grants (HPOG) program—provides a practical example of this design and its potential for both methodological and substantive learning. We conclude that the promise for CAMIC lies in situations where the correlations between the selected program enhancement and alternative program characteristics of interest are relatively high, implying that producing an experimental estimate of the enhancement can reduce bias in the estimation of other non-randomized program characteristics.

Acknowledgements
Funding for the work that underlies this paper came from the U.S. Department of Health and Human Services (HHS), Administration for Children and Families (ACF), Office of Planning, Research, and Evaluation (OPRE). The authors are grateful input on this draft from Larry Buron and Eleanor Harvill at Abt Associates, and from Nicole Constance, Hilary Forster, and Amelia Popham at ACF/OPRE. Eleanor Harvill and Stephen Bell are co-authors on a prior iteration of this work, and we are grateful to their input on developing, testing, and advancing the CAMIC method. The views expressed in this paper do not necessarily reflect the views or policies of the HHS/ACF/OPRE.
How Three-Arm Random Assignment within Sites Can Improve Non-Experimental Cross-Site Estimates of the Relationship between Program Characteristics and Impact

Randomized experiments provide researchers with a powerful method for understanding the effectiveness of programs. However, once we understand the magnitude and sign of program impacts, the next question is why the program did or did not have its intended effect. A growing body of research has used data from multi-site experiments to investigate what it is about a particular program that determines the magnitude of its impacts. This paper proposes how multi-armed, multi-site experiments—which are increasing in number (cite ER special issue)—can build on that literature with experimental estimates while potentially improving non-experimental impact estimation.

The paper proceeds as follows: first, we discuss the empirical challenge involved with determining how program characteristics contribute to impact magnitude. Then, we discuss two common approaches that prior researchers have used to address this challenge. Next, we explain how multi-site, multi-armed experiments can eliminate bias in estimates of the effect of randomized program components and, in turn, potentially reduce bias in estimates of the effect of non-randomized program characteristics, via the use of a method we call Cross-Site Attributional Model Improved by Calibration to Within-Site Individual Randomization Findings (abbreviated as “CAMIC”). Finally, we offer some evidence on how the CAMIC method performs in simulations, and discuss implications for evaluation research and program practice.

The Empirical Challenge

Cross-site estimates of the relationship between site-specific program characteristics and impact magnitude may be subject to omitted variable bias if there are unmeasured or omitted
site-level variables that are correlated with the program characteristics included in the model and that influence treatment impact magnitude (Moulton, Bell, & Peck, 2014). Researchers often face this empirical challenge because the program components themselves (such as a given educational program’s curriculum or related support service) are rarely randomized to the locations within which evaluations take place (we refer to these locations as “sites”); and implementation features (such as the dynamism of a program manager or office culture) cannot be randomized. Instead, each site chooses its own configuration of program components to adopt, each possesses its own set of implementation features, and each operates within its own context with its own target populations.

The planners’ or managers’ own traits may associate with the decisions of both what program components to offer and how to implement in practice as well as the magnitude of the program’s impacts. For example, the local program manager’s enthusiasm and leadership might be associated with both the choice to offer access to peer support groups as part of their job training; and those same traits may also lead the program to have larger or smaller impacts. In this situation, the manager’s traits are associated both with what is offered and with impact magnitude. If the researcher fails to control for these traits (which are often difficult to measure), then cross-site non-experimental estimates of the relationship between program design and impact magnitude will be biased.

Likewise, the local environment—including economic conditions and policy context—might be associated with both the program components that sites adopt and the routes by which program participants can benefit from it, such as the nature and extent of jobs available in the local economy. This type of contextual influence also may be difficult to measure and control for in the analysis.
These and other such scenarios would bias estimates of the influence of program components and implementation features when they do not vary randomly across sites. It is possible to randomize individuals to gain access to program components: for example, a lottery can be used in a job training program to decide which participants are offered internships. That said, it generally is not possible to randomize implementation features, either among individuals or across sites. Moreover, although it is possible to control for some contextual factors, it is not possible to control for (or even know) all of them. Nevertheless, researchers, policy makers and program managers are interested in understanding the influence of both program components and implementation features on impact magnitude, and multi-site experiments can facilitate that learning.

Methods for Analyzing the Contribution of Program Characteristics to Impacts

Two methods have commonly been used in recent years as analytic means for assessing the contribution of program characteristics to impacts in the context of experimentally-designed evaluations: (1) a multi-level modelling approach, and (2) an instrumental variables approach. Both of these use an experimental evaluation design, essentially estimating the relationship between the site-level program characteristics and experimental site-level program impacts. This section describes briefly each of these two approaches and then describes how multi-site, multi-armed experiments provide a way to eliminate bias in estimates of randomized program enhancements. Capitalizing on the multi-site, multi-armed design, the next section will discuss the CAMIC approach, a distinct kind of “within-study-comparison” (WSC) opportunity to improve the non-experimental analysis by leveraging experimental results.
Method 1: Multi-level Modeling of the Contribution of Program Characteristics to Impacts

To inform how program characteristics contribute to impact magnitude, a number of studies have taken advantage of the naturally occurring cross-site variation in the specific services offered by programs (program components) and in how these services are delivered (implementation features), while controlling for other contextual influences (e.g., Bloom, Hill, & Riccio, 2001, 2003, 2005; Dorsett & Robins, 2013; Godfrey & Yoshikawa, 2012; Greenberg & Robins, 2011). These analyses commonly use multi-level modeling based on Bryk and Raudenbush (2001) to compute non-experimental estimates of the relationship between site-specific program characteristics and experimentally estimated impact magnitude. This method of relating program characteristics to impact requires the assumption that the model is not subject to omitted variable bias in that there are no site-level variables that are both correlated with the program characteristics included in the model and influence treatment impact magnitude (e.g., Moulton, Bell, & Peck, 2014).

This kind of analysis originated in the work of Greenberg, Meyer, and Wiseman (1993), who were the first to categorize the questions of interest as being about “prying the lid from the black box.” That is, impact analyses treat “the program” as a “black box,” an unknown space within which varied activities would take place. The black box nature of the characteristics surrounding program implementation begs the questions: which of those activities, and which implementation and contextual factors, are responsible for any observed program impacts? Placing a multi-site evaluation in a multi-level framework provides a means for answering those “black box” questions, at least correlationally (Greenberg, Meyer, & Wiseman, 1994).
Method 2: Using Site-by-Treatment Interactions in an Instrumental Variables Analysis

More recently, similarly motivated analyses have used each of the site-by-treatment interactions that exist in a multi-site experiment as instruments for estimating the contribution of site-level characteristics (e.g., Bos & Granger, 2000; Kling, Liebman, & Katz, 2007; Magnuson & McGroder, 2002). One key instrumental variables assumption is the exclusion restriction. In this context, the exclusion restriction means that the effect of the program on the outcome must be mediated—and mediated alone—by the program characteristic of interest. However, programs are commonly multifaceted—especially those subject to these “prying the lid from the black box” sorts of questions—which makes it difficult to separate the effects of one program characteristic from the effects of another characteristic using instrumental variables (Gennetian, Bos, & Morris, 2002). An additional challenge of using instrumental variables is that, if the impacts of the treatment on the program characteristic do not vary significantly across sites, then the use of multiple instruments may lead to substantially decreased precision and increased finite sample bias (Reardon et al., 2013).

Multi-Site, Multi-Armed Experiment’s Distinctive Method

Going a step further, the impact evaluation of the first round of Health Profession Opportunity Grants (HPOG 1.0) was designed to eliminate bias in estimates of the contribution of specific program characteristics (Peck et al., 2018). The impact evaluation assesses whether providing access to career pathways training for healthcare occupations improves participant outcomes overall while also examining what about the multi-faceted intervention associates with impacts. To do this, individuals were randomized to the HPOG treatment group or to a control group that did not have access to HPOG-funded services. Importantly, in some sites there were
two treatment groups, where one treatment group had access to HPOG while the second treatment group had access to HPOG enhanced with one of three selected program components: (1) emergency assistance for specific needs, (2) noncash incentives designed to encourage desirable program outputs and outcomes, and (3) facilitated peer support groups. This design enabled the researchers to produce experimental estimates of the contribution of peer support, emergency assistance, and noncash incentives to the HPOG Program’s impact.

As described further below, the evaluation has both experimental and non-experimental evidence on the relative effectiveness of each of these program components. As a result, HPOG involves a special kind of WSC that provides an opportunity for additional methodological learning. The goal of other WSC studies (beginning with LaLonde, 1986) has been to learn which approach to measuring program impact, subject to selection and other sources of bias using observational data, best replicates an experimental finding for the same impact quantity. A series of such studies has begun to point evaluators toward the conditions that yield more-reliable non-experimental findings than other options do (see Cook et al., 2008; Glazerman et al., 2003).

**CAMIC Method**

The CAMIC method was conceived in the HPOG 1.0 Impact Study as a possible way of improving on common practice by employing a WSC-style approach, thereby reducing bias in non-experimental estimates of the contribution of selected program characteristics to overall
impacts (Bell et al., 2017). It does so by identifying the most useful measures to include as covariates in the statistical model.

As depicted by Exhibit 1, consider the experimental evaluation design, where two sets of sites exist, with respective designs as follows:

- **Design in Sites that Randomize Individuals to an Enhanced Treatment Group.** Individuals in a subset of sites are randomly assigned to one of three arms: a standard treatment group, an enhanced treatment group (that receives the standard treatment plus an enhancement component), or a control group (that has no access to the program).\(^1\) For example, HPOG’s emergency assistance enhancement added access to emergency funds to meet needs stemming from imminent eviction from housing, utility shutoff, vehicle repair needs, childcare needs, etc. to the standard intervention (Peck et al., 2018).

- **Design in Sites that Do Not Randomize Individuals to an Enhanced Treatment Group.** Individuals in another subset of sites are randomly assigned into one of (at least) two arms: a standard treatment group (where the treatment does not include the program component that was—in other sites—randomized to as an enhancement) and a control group.\(^2\)

---

\(^1\) For example, HPOG’s emergency assistance enhancement added access to emergency funds to meet needs stemming from imminent eviction from housing, utility shutoff, vehicle repair needs, childcare needs, etc. to the standard intervention.

\(^2\) The CAMIC methods allows for additional treatment arms in these other sites (e.g., a third treatment arm that receives the standard treatment plus some other enhancement).
Given this design, the CAMIC method has the potential to reduce bias in cross-site non-experimental estimates of program components and implementation features by completing the following steps.

**Step 1:** Experimentally estimate the impact of the program enhancement (e.g., the emergency assistance program enhancement) using the sample limited to sites that randomize individuals to an enhanced treatment group using three-armed random assignment (as depicted by the top panel of Exhibit 1). An experimental estimate of the impact of the program enhancement can be computed in these sites by comparing mean outcomes for individuals randomly assigned to the enhanced treatment group with mean outcomes for individuals randomly assigned to the standard treatment group.

More specifically, we can produce an experimental estimate of the enhancement in these sites by estimating the following model:
\[ Y_{ji} = \alpha_0 + \beta_0 T_{Eji} + \pi_e E_{ji} + \{v_j + \mu_j T_{Eji} + \omega_j E_{ji} + \varepsilon_{ji}\} \] (Eq. 1)

In this equation, \( Y \) is the outcome of interest for individual \( i \) in site \( j \), \( T_{Eji} \) is an overall treatment indicator which denotes whether the individual was assigned to either the standard treatment or enhanced treatment, and \( E_{ji} \) is an indicator for whether the individual was assigned to the enhanced treatment group. The model also includes a series of error terms. See Exhibit 2 for definitions of the terms included in these equations and all models presented throughout the manuscript.

Estimating Equation (1) through linear regression, we obtain—among other things—an estimate \( \hat{\pi}_e^X \) of \( \pi_e \) straight from the experiment, based on purely random variation in which individuals receive a program that includes the enhancement component \( e \) and which individuals receive a program that does not. The \( X \) superscript denotes the unbiased experimental nature of this estimate.

**Step 2:** In this step, we calculate several non-experimental estimates of the contribution of the program enhancement to impact magnitude using cross-site variation in whether the enhancement is included in a given site’s program (this follows “Method 1,” from above). To produce this estimate, the sample is limited to (1) the control and enhanced treatment arms from sites that conducted three-armed random assignment using the enhancement and (2) the control and standard treatment arms from sites that did not randomize to the enhancement.3 We use

---

3 Sample members randomly assigned to the standard treatment arm from sites that conducted three-armed random assignment are excluded from this analysis, which ensures that the effect of the program enhancement is estimated as if it were not randomized as an enhancement. A benefit if this site-selection strategy is that the same set of enhancement-implementing sites is used to produce both the experimental estimate of the enhancement in Step 1 and the non-experimental estimate of the enhancement in Step 2. This ensures that the differences in these two estimates of the program enhancement’s effect is not confounded by differences between the sites that implemented the enhancement.
variations of the following model to produce non-experimental estimates of the contribution of
the enhancement component to impact magnitude:

\[ Y_{ji} = \alpha_0 + \beta_0 \cdot TE_{ji} + \pi_e \cdot P_{ej} \cdot TE_{ji} + \sum_m \pi_m \cdot P_{mj} \cdot TE_{ji} + \sum_n \varphi_n \cdot X_{nj} \cdot TE_{ji} + \{ v_j + \mu_j \cdot TE_{ji} + \varepsilon_{ji} \} \]  

(Eq. 2)

In this equation, \( Y \) is the outcome of interest for individual \( i \) in site \( j \) and \( TE_{ji} \) is an overall
treatment indicator which denotes whether the individual was assigned to either the standard
treatment or enhanced treatment. \( P_{ej} \) is an indicator for whether the enhancement component
(e.g., emergency assistance) was included as part of the program in site \( j \) (i.e., \( P_{ej} \) equals 1 in
three-arm sites and equals 0 in two-arm sites), \( P_{mj} \) captures all other key program characteristics
of interest which the researcher wants to ensure are included in the model subject to degrees of
freedom limitations (e.g., the researchers may be interested in the extent to which tuition is
covered by the program affects impact magnitude, even though tuition assistance is not offered
as a randomized enhancement), and \( X_{nj} \) captures all other site-level covariates that could be
included in the model to reduce bias (e.g., the unemployment rate). These site-level program
characteristics are all multiplied by the treatment indicator, and the resulting interaction terms
capture the influence of a given site-level measure on the magnitude of the program’s impact.
The model also includes a series of error terms, as defined in Exhibit 2.
**Exhibit 2. Definition of Model Terms**

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome and Covariates</strong></td>
<td></td>
</tr>
<tr>
<td>$Y_{ji}$</td>
<td>The outcome measure for individual $i$ from site $j$</td>
</tr>
<tr>
<td>$TE_{ji}$</td>
<td>The treatment group indicator (1 for those individuals assigned to the standard treatment or enhanced treatment group, and 0 for those individuals assigned to the control group; this is labelled “TE” for the combination of standard “treatment” and “enhanced” treatment)</td>
</tr>
<tr>
<td>$E_{ji}$</td>
<td>The enhanced treatment group indicator (1 for those individuals assigned to the enhanced treatment group, and 0 otherwise; this is labelled “E” for “enhanced” treatment)</td>
</tr>
<tr>
<td>$P_{ej}$</td>
<td>An indicator for whether the enhancement component was included as part of the program in site $j$. In this example, $P_{ej} = 1$ in three-arm sites and $= 0$ in two-arm sites.</td>
</tr>
<tr>
<td>$P_{mj}$</td>
<td>Program characteristic $m$ for site $j$, $m = 1, \ldots, M$ (these are labelled “P” for “program”)</td>
</tr>
<tr>
<td>$X_{nj}$</td>
<td>Additional site-level covariate $n$ for site $j$, $n = 1, \ldots, N$. These site-level covariates are included in the model to reduce bias in other site-level measures of interest, and may include program characteristics of secondary interest and local context measures (e.g., the unemployment rate in the area where the site is located).</td>
</tr>
<tr>
<td><strong>Model Coefficients</strong></td>
<td></td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>The grand mean control group outcome</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>The grand mean impact of the standard treatment</td>
</tr>
<tr>
<td>$\pi_e$</td>
<td>The grand mean impact of being offered the enhanced intervention inclusive of component $e$, rather than the standard intervention without $e$</td>
</tr>
<tr>
<td>$\pi_m$</td>
<td>The influence of program component $m$ on impact magnitude, $m = 1, \ldots, M$</td>
</tr>
<tr>
<td>$\varphi_g$</td>
<td>The influence of bias-reducing covariate $g$ on impact magnitude, $g = 1, \ldots, G$</td>
</tr>
<tr>
<td><strong>Error Terms</strong></td>
<td></td>
</tr>
<tr>
<td>$\varepsilon_{ji}$</td>
<td>A random component of the outcome for each individual</td>
</tr>
<tr>
<td>$\mu_j$</td>
<td>A random component of the standard intervention impact for each site</td>
</tr>
<tr>
<td>$\nu_j$</td>
<td>A random component of the mean outcome for each site</td>
</tr>
<tr>
<td>$\omega_j$</td>
<td>A random component of the enhanced intervention’s incremental impact for each site</td>
</tr>
</tbody>
</table>

Estimating Equation (2) through linear regression, we obtain a non-experimental estimate of the influence of the enhancement component on impact magnitude $\hat{\pi}_e^N$ of $\pi_e$ using cross-site variation in whether the enhancement is part of the program in a given site. The $N$ superscript denotes the non-experimental nature of this estimate which may be subject to bias. Additionally, we obtain non-experimental estimates of the influence of other key program characteristics of interest $\hat{\pi}_1^N, \ldots, \hat{\pi}_M^N$.

Bias in the estimates of the influence of the enhancement component $P_{ej}$ and the influence of other key program characteristics $P_{mj}$ will arise when these program characteristics
are correlated with omitted site-level factors that also influence impacts. For example, in the case of the HPOG impact evaluation these factors—often omitted because they are unobserved in the data—may include aspects of the program that administrators choose (e.g., an unmeasured program component or implementation feature), aspects of the program context that are beyond their control (e.g., the local unemployment rate), and other unobserved factors that influence both the program components and implementation features and the impacts of the program (e.g., the raw talent of the program leadership). For all of these types of unobserved factors, if the program characteristics of interest are correlated with unobserved factors, the estimated influence of those characteristics will reflect the influence of the unobserved factors (Bell et al., 2017).

Importantly, in this step we produce several different estimates of $\hat{\pi}_e^N$ (and $\hat{\pi}_1^N, ..., \hat{\pi}_M^N$) by varying the set of bias-reducing site-level covariates $X_{nj}$ included in the model, given degrees of freedom limitations.  

**Step 3:** The next step in implementing the CAMIC method is to identify the version of the Equation (2) model from Step 2 that minimizes the difference between the experimental estimate of the program enhancement’s effect computed in Step 1 and non-experimental estimate of the program enhancement’s effect computed in Step 2 (using the best “similarity” metric available from WSC scholarship$^5$), where bias is measured subject to sampling variability as $|\hat{\pi}_e^N - \hat{\pi}_e^X|$.

---

4 Note that adding a covariate as an impact moderator to an attributional model may actually increase omitted variable bias, even if the added variable is highly correlated with omitted confounders (Steiner & Kim, 2015). One goal of the CAMIC method is to avoid this mistake. This perverse result can arise from two phenomena: (1) bias amplification and (2) removing the benefit of offsetting biases. Bias amplification occurs when conditioning on the new variable amplifies the bias caused by the omitted, unobserved confounder by increasing the correlation between the unobserved confounder and other included variables of interest. It is also possible that two omitted confounders initially induced bias in opposite directions, and that the benefit of these offsetting biases is lost when one but not both confounders is added to the specification.

5 See, for example, Steiner and Wong (in press).
Step 4: Apply the “best” model from Step 3 to estimate the contribution of other program characteristics of interest \( p_{mj} \) (when these other components or features do not vary randomly among individuals within sites or across sites) to assess their contribution to overall impacts with potentially reduced bias.

CAMIC Method Performance

Bell et al. (2017) report on simulations that investigate whether the CAMIC method might help accomplish its goal of reducing bias in non-experimental estimates of the influence of program characteristics on a program’s total impact. Their simplified framework includes three program components: (1) an “enhancement component” for which we can obtain an unbiased measure of its influence on impact (as depicted by \( P_{ej} \) in Equation (2)); (2) a non-randomized “focal component” for which we are interested in minimizing bias (as depicted by \( P_{ej} \) in Equation (2)); and (3) a “bias-reduction component,” the effect of which is not directly of interest, but can be included in the model to potentially reduce bias in estimates of the influence of the non-randomized focal component (as depicted by \( X_{nj} \) in Equation (2)). All three of these program components may be correlated with an unobserved site-level factor.

To understand the CAMIC method’s potential to reduce bias in the estimate of the focal component, Bell et al. (2017) examined whether the model specification that produced the least biased estimate of the enhancement component, is also the least biased specification for the non-randomized focal component while varying key parameter values.\(^6\) Bell et al. (2017) first

---

\(^6\) The key parameters that determine the bias are the correlations among the three observed program components (the enhancement component, focal component, and the bias-reduction component), the correlation between of each observed program component with the unobserved factor, and the true influence of the bias-reduction component on impact. To calculate bias, one must select a value for each of these seven parameter values. Given the large number of possible combinations of parameter values, Bell et al. (2017) are strategic in selecting a relatively limited number of simulations that help understand a range of possible biases.
identified possible values for the correlation between observed program components (the enhancement component, focal component, and the bias-reduction component) and referred to each combination of correlations in these components as a scenario. Then, for each scenario, they ran 500 simulations to capture a broad range of values of the unobserved parameters. They calculated the proportion of these simulations that are favorable to the CAMIC method for a particular scenario. A simulation is considered favorable to the CAMIC method if the model specification that produced the least biased estimate of the enhancement component is also the model that produced the least biased estimate for the non-randomized focal component of interest and unfavorable otherwise. This structure allows researchers to consider which of the scenarios is most similar to the correlations in observed program components they might observe in their data.

Bell et al. (2017) reach the conclusion that CAMIC is no better than standard, multi-level procedures in producing unbiased estimates of the contribution of certain program characteristics to overall impacts. However, in a key subset of simulated scenarios CAMIC offers a clear advantage: it is the circumstance where the correlation between the enhancement component and the focal component of interest is high ($\rho=0.70$) and the correlation between each of these two components and the bias-reduction component is low ($\rho=0.25$) to moderate ($\rho=0.50$).

Exhibit 3 presents an excerpt from Bell et al.’s (2017) Exhibit 9, and it reveals the following: when the correlation between the enhancement component and the focal component is high ($\rho=0.70$) and the correlation between each of these two components and the bias-reduction component is low ($\rho=0.25$) to moderate ($\rho=0.50$), the CAMIC method performs favorably in more than 80 percent of the simulations (Scenarios A and B in Exhibit 3). In contrast, when the correlations are high ($\rho=0.70$) across the board (as is the case for Scenario C) CAMIC performs
favorably in only 29 percent of the simulations. When the correlation between the enhancement component and the focal component is low ($\rho=0.25$) to moderate ($\rho=0.50$), the findings are inconclusive in that the CAMIC method sometimes performs favorably and sometimes performs unfavorably (Scenarios D through O).

**Exhibit 3. Results for Selected Scenarios, Focused on the Relative Magnitude of the Correlations**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Enhancement component &amp; focal component ($\rho_{12}$)</th>
<th>Enhancement component &amp; bias-reduction component ($\rho_{13}$)</th>
<th>Focal component &amp; bias-reduction component ($\rho_{23}$)</th>
<th>N</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.70</td>
<td>0.50</td>
<td>0.25</td>
<td>416</td>
<td>83</td>
</tr>
<tr>
<td>B</td>
<td>0.70</td>
<td>0.25</td>
<td>0.50</td>
<td>416</td>
<td>83</td>
</tr>
<tr>
<td>C</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
<td>144</td>
<td>29</td>
</tr>
<tr>
<td>D</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>200</td>
<td>40</td>
</tr>
<tr>
<td>E</td>
<td>0.25</td>
<td>0.50</td>
<td>0.25</td>
<td>200</td>
<td>40</td>
</tr>
<tr>
<td>F</td>
<td>0.25</td>
<td>0.25</td>
<td>0.50</td>
<td>204</td>
<td>41</td>
</tr>
<tr>
<td>G</td>
<td>0.25</td>
<td>0.50</td>
<td>0.50</td>
<td>238</td>
<td>48</td>
</tr>
<tr>
<td>H</td>
<td>0.25</td>
<td>0.70</td>
<td>0.50</td>
<td>272</td>
<td>54</td>
</tr>
<tr>
<td>I</td>
<td>0.25</td>
<td>0.50</td>
<td>0.70</td>
<td>272</td>
<td>54</td>
</tr>
<tr>
<td>J</td>
<td>0.50</td>
<td>0.25</td>
<td>0.25</td>
<td>156</td>
<td>31</td>
</tr>
<tr>
<td>K</td>
<td>0.50</td>
<td>0.50</td>
<td>0.25</td>
<td>214</td>
<td>43</td>
</tr>
<tr>
<td>L</td>
<td>0.50</td>
<td>0.25</td>
<td>0.50</td>
<td>274</td>
<td>55</td>
</tr>
<tr>
<td>M</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>188</td>
<td>38</td>
</tr>
</tbody>
</table>

These findings imply that the CAMIC approach—calibrating the model specification by comparing experimental and non-experimental estimates of the enhancement component can reduce bias (and improve upon standard practice) in some other, non-randomized focal component only when there is a meaningful correlation between the enhancement component
and the focal component. Without that correlation, CAMIC cannot be expected to do any better than standard practice. Further, good performance of the CAMIC method requires that the correlation between each of these two components and a third bias-reduction component is low ($\rho=0.25$) to moderate ($\rho=0.50$). If the intercorrelations among the enhancement component, the focal component, and the additional bias-reduction component are all high ($\rho=0.70$), the CAMIC method performs poorly as it may suffer from multicollinearity.

In advance of implementing the CAMIC method, these observations suggest that researchers compute the correlation between the enhancement component and the focal component of interest to ensure that these two program characteristics are highly correlated. If this criteria is satisfied, then the researcher could conduct the CAMIC method, calibrating the model specification using only those additional bias-reduction components that have a low ($\rho=0.25$) to moderate ($\rho=0.50$) correlation with the enhancement component and the focal component of interest. Bell et al. (2017) caution that these simulation findings ignore the role that variance might play in the CAMIC method. Although the CAMIC method seeks to select the model that yields the least biased estimate of the non-experimental estimate of the enhancement component, noise in the estimates may result in selecting the wrong model. This limitation of the CAMIC method is not reflected in the simulations, where perhaps further analysis or application are needed.

**Discussion & Conclusion**

This paper has discussed the problem of bias that arises when estimating the influence of program characteristics that have not been randomized, but instead are selected and implemented in ways and in contexts that correlate with their relative contributions to overall impacts. Both
multi-level analytic methods and instrumental variable estimation are approaches that analysts use to pry the lid from the black box—to use Greenberg, Meyer & Wiseman’s (1993) phrase—and attempt to ascertain what it is about a program that is responsible for its impacts.

Beyond these analytic approaches, recent years have seen an increase in the numbers of experimental evaluations that include both multiple sites and multiple treatment arms. Through design, if individuals are randomly assigned to either a standard treatment group or an enhanced treatment group that receives the standard treatment plus a program enhancement, then the bias in the estimate of the randomized program enhancement can be eliminated. Advancing further still, this kind of design offers additional analytic opportunities for bias reduction in the non-experimental estimation of the contribution of other non-randomized program characteristics to overall impacts.

This paper contributes to the literature in two main ways. First, it advocates for the use of a multi-site, multi-armed experimental design as a way to produce unbiased estimates of an enhancement component. In addition, it suggests a method that could reduce bias in the cross-site non-experimental estimates of non-randomized program characteristics, in the case where the study design includes three-armed randomization that allows for experimental analysis of the contribution of at least one specific program characteristic. Over time, CAMIC method-based tests of cross-site impact attribution specifications may yield consistent results with enough replications. Given the large investment needed to carry out this design, we recognize that it will take time to amass replicate evidence of the influence of program components on impacts. Certainly as the state of the science of within-study comparisons improves as well, we should gain knowledge regarding the calibration of non-experimental to experimental results.
The evolution of CAMIC as a method among those in the evaluation toolkit has implications both for evaluation research and for program practice. Future evaluation research should consider opportunities to test CAMIC in settings where the design is fitting: multi-site, multi-armed experiments, where at least a subset of the sites randomize individuals to either a standard treatment group, an enhanced treatment group (that receives the standard treatment plus a program “enhancement”), or a control group. This design allows the researcher to compute both experimental and non-experimental estimates of the influence of the program enhancement, and CAMIC may be able to leverage that evidence in order to improve the estimation of the contribution of other non-randomized program characteristics of interest. The CAMIC method is likely to achieve its promise of reducing bias in estimates of the relationship between a focal/non-randomized program component and program impact only when the correlation between the randomized program enhancement and the focal/non-randomized program component of interest is high and the correlation between each of these two components and alternative bias-reducing covariates are low to moderate.

One important consideration for researchers interested in using data from multi-site randomized experiments to estimate the influence of site and individual characteristics on variation in program impacts is that of limited sample sizes (Moulton, Bell, & Peck, 2014). Because it can be expensive to conduct randomized experiments in many varied locations, most past experiments have been confined to a small number of sites. Having a small number of sites severely limits researchers’ ability to conduct cross-site comparisons due to lack of statistical power. Additionally, having a small number of sites increases the potential for omitted variable bias of non-experimental estimates due to the need to leave out key site-level variables in order to conserve degrees of freedom.
To the extent that the CAMIC method helps reduce bias in estimates of the influence of non-randomized program characteristics on impact magnitude in real-world applications, CAMIC also has implications for program practice: program managers can expect more reliable estimates of the relative contributions of program characteristics to overall impacts. For example, the proliferation of this line of analysis in practice means that the field should be able to learn how programs operate differently in strong versus weak economic conditions, which aspects of multi-faceted programs are important drivers (or suppressors) of impact, which implementation features are essential to a program’s success, and how participant compositional factors influence (or not) programs’ effectiveness. These are practical lessons that inform administrative decisions on the ground as well as higher-level policy decisions about what to fund.
Works Cited


Steiner, Peter M. & Vivian C. Wong. (in press). Assessing Correspondence between Experimental and Non-Experimental Results in Within-Study Comparisons. *Evaluation Review.*