Using the Results from Rigorous National Evaluations to Inform Local Education Policy Decisions[1]

Larry L. Orr, Johns Hopkins Bloomberg School of Public Health
Robert B. Olsen, George Washington Institute of Public Policy
Stephen H. Bell, Abt Associates
Ian Schmid, Johns Hopkins Bloomberg School of Public Health
Azim Shivji, Abt Associates
Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health

DRAFT
July 25, 2017

Abstract: Increasingly, local education policymakers have access to evidence from published reports based on randomized control trials (RCTs) to inform decisions about whether to adopt an educational intervention. A key question is how well the published results from a multi-site RCT predict the potential consequences of adopting an intervention for each of the many local schools or districts that may consider adopting it. This paper offers a set of methods for quantifying the accuracy of the local predictions that can be obtained from multi-site RCTs, using available data from those RCTs, and for assessing the likelihood that prediction errors will lead to errors in local policy decisions. It also provides the first empirical evidence on the accuracy with which local impacts can be predicted from the evidence taken from published reports on RCTs in education.

Many evaluations of education interventions are primarily intended to inform local education decision makers. Adoption of specific curricula or teaching techniques, teacher evaluation and compensation systems, student mentoring programs, and many other educational interventions are all the province of local school administrators or state officials. Indeed, the Every Student Succeeds Act of 2016 shifts even more authority for education policy from the federal government to states (Aragon et al., 2016), and state education authorities are increasingly giving local school boards more authority over education policy decisions in some instances (Burnette, 2017).

Increasingly, local education policymakers have access to evidence from published reports based on randomized control trials (RCTs) to inform decisions about whether to adopt an intervention. These evaluations, which often encompass multiple schools or districts, use random assignment to determine the schools, classrooms, and/or students that receive an intervention and which do not to ensure that impact estimates—derived as differences in outcomes between the two groups—do not suffer from treatment selection bias (Orr 1999). Many RCTs have been sponsored by the U.S. Department of Education. And many of these

[1] Funding for this research was provided by the Institute for Education Sciences, under Grant # R305D100041.

evaluations have been reviewed and disseminated by the What Works Clearinghouse (WWC) "…to provide educators with the information they need to make evidence-based decisions" (front page of the WWC website; https://ies.ed.gov/ncee/WWC/).

But how well do the published results from a given multi-site RCT predict the potential consequences of adopting the tested intervention for each of the many local schools or districts that may consider adopting it? Published results may produce unbiased estimates of the average impact of an intervention in the study sample and still produce inaccurate predictions of the impact for individual schools and districts if the impact of the intervention varies across localities. Unfortunately, there is relatively little evidence on how much the impacts of educational interventions vary and no evidence (of which we are aware) on the implications of this variation for the accuracy with which the local impact of adopting an intervention can be predicted using findings from a national evaluation.

This paper makes three contributions. First, it highlights a potential challenge in making local education policy decisions that has been underappreciated in the literature: Reported evidence from RCTs may not accurately predict the impacts of adopting an intervention in individual localities. Second, it offers a set of methods for quantifying the accuracy of the local predictions that can be obtained from multi-site RCTs and for assessing the likelihood that prediction errors will lead to errors in local policy decisions. Third, it demonstrates these methods, providing the first empirical evidence on the accuracy with which local impacts can be predicted from the evidence taken from published reports on RCTs in education.

To measure the accuracy with which published evidence from national evaluations can predict local impacts, we develop and apply an analytic strategy that involves (1) pretending that one of the localities (henceforth, "sites") that participated in a national, multi-site RCT had been excluded from the study, (2) using statistical methods to predict the impact of the intervention in the excluded site using the data from the other sites, (3) repeating this process for each site in the RCT, and (4) summarizing the resulting prediction errors across the sites. In addition, we extend and apply methods from Bell and Orr (1995) to calculate the probability that these prediction errors would lead localities to make the wrong policy decisions about whether to adopt the intervention.

Applying these methods to data from three multi-site national RCTs in education,[2] we assess the accuracy with which policymakers can predict the local impacts of three potential policy decisions that they may face: (1) whether to allow charter schools to open in a particular school district or community, (2) whether to adopt technology-based classroom interventions in a particular school, grade level, and subject area, and (3) whether to fund a Head Start program in a particular locality. We make no claims that the results presented in this paper are broadly

[2] We term these studies "national evaluations" because they were funded with the intent of informing policymakers throughout the nation, not just those in the study areas. Only one, the Head Start Impact Evaluation, selected its sample to be formally representative of a national population.

generalizable to other education policy decisions that could be informed by RCTs. However, these results provide initial evidence on how accurately local policymakers can predict the consequences of at least some of the policy decisions they face based on the results from multisite RCTs, and, as noted, the analysis demonstrates methods that can be used to investigate that question in other contexts.

The next section describes the problem policy makers face when trying to use evidence from rigorous national evaluations to predict the impact of adopting an intervention locally. We then present the data and methods we used to assess the magnitude of prediction errors that can result from using this evidence and the likelihood that these errors will lead to incorrect policy decisions. We conclude with our empirical results and our interpretation of those results.

STATEMENT OF THE PROBLEM

Evidence-based policy is all about prediction—for example, predicting the impact of an intervention to inform whether it should be adopted. In a perfect world, local policymakers would be able to predict accurately the impact of adopting an intervention locally. Then policymakers could weigh the predicted impact and the costs of adopting the intervention against the status quo and against the predicted impacts and costs of alternative interventions.

In the real world, local policymakers can attempt to predict an intervention's impact using the evidence available, but that prediction will inevitably contain some error. This section discusses both the sources of error and the prediction options available to local education authorities.

Errors in Predicting the Impact of Adopting an Intervention Locally

Localities can conduct local pilot tests to estimate the impact of adopting an intervention locally. But most often, the evidence available to local policy makers comes from an evaluation conducted in other localities. Therefore, evidence-based policy decisions usually involve out-of-sample predictions using data from localities that participated in an evaluation to predict impacts in other localities.

In predicting local impacts from the data or findings from national evaluations, there are two sources of prediction error: bias and variance. The bias component is defined relative to the parameter of policy interest for the local decision-maker (e.g., the average impact that the intervention would have if it were adopted in the decision-maker's school district). If impacts vary across localities, the average impact estimates reported by the evaluation – though unbiased for the evaluation sample – may be biased for the impact in any given locality. For a particular locality, it can be shown that the bias is a function of two factors: (1) the difference between the evaluation sample and the locality on factors that affect the magnitude of—i.e.,

moderate—the intervention's impact,[3] and (2) the strength of the influence of those moderators on impact magnitude (e.g., see Tipton 2013, p. 116). In general, the amount of bias is unknown and difficult to estimate because the factors that moderate the impacts of any intervention are typically unknown or difficult to measure in evaluations. This bias generates errors in predicting the local impact of adopting an intervention.

The second source of prediction error—the variance (or standard error) of the published impact estimates—results from conducting evaluations in finite samples. Even if the bias of prediction to the local level were zero (i.e., if the true impact were the same in the evaluation sites and the decision-maker's school or district), we would still expect the variance of the impact estimate to produce error in the predicted impact of adopting the intervention locally.

A common metric for quantifying the magnitude of prediction errors is the Mean Squared Error (MSE), which in this context we call the Mean Squared Prediction Error (MSPE). The MSPE captures both sources of prediction error: It equals the bias squared plus the sampling variance of the prediction. This metric is indifferent to whether prediction errors result from bias or variance, much as policymakers should be indifferent between the two sources of the prediction error. It is also indifferent to whether the errors are positive or negative. This is the primary metric we will use to quantify the amount of error in predicting local impacts.

Choosing Among Different Impact Estimates for Making Local Predictions

RCTs typically produce multiple impact estimates that can be used to predict the impact of adopting an intervention locally. For example, they often present an overall average effect as well as the effects for particular subgroups of students or sites, such as minority students or schools in urban settings. But it is not clear which estimate or estimates the policymaker should use because it is not clear which estimates yield the smallest prediction errors.

One option is to use the average impact reported for the entire sample. The main advantage of using this estimate is that it minimizes the variance component of the prediction error by using the largest possible sample. However, if the study sample differs in important ways from the students who would receive the intervention if the intervention were adopted locally—or the environment in which the intervention would be implemented differs substantially from the environment in which the intervention was evaluated—this estimate may be biased for the parameter of interest: the average impact in the locality that may adopt the intervention.

Alternatively, policymakers can use subgroup impact estimates—when reported—to predict the impact of adopting an intervention locally. It is very common for RCTs in education to estimate and report the effect in one or more sets of mutually exclusive subgroups of students (e.g., minority students and white students), teachers (e.g., new teachers and experienced teachers), or schools (e.g., urban schools and rural schools). Using subgroup estimates may

[3] For a conceptual description of the types of factors that may moderate the effects of educational interventions, see Weiss, Bloom, and Brock (2014).

reduce the bias if the subgroup sample mirrors the local student population more closely than the overall sample.

However, relying on subgroup estimates will typically increase the variance component of the prediction error since the subgroup estimates are based on smaller samples and thus contain additional sampling error. Therefore, using subgroup estimates will reduce the MSPE if the reduction in bias outweighs the increase in variance, but it will increase the MSPE if the reverse is true.

Finally, some evaluations model the impact of an intervention as a function of *multiple* moderator variables simultaneously. Mechanically, these models are estimated by interacting multiple variables that may moderate the impact of the treatment with the treatment indicator in a regression model of the outcome. Models of this type potentially allow policymakers to use more information about their students and/or local environment to refine their predictions of the impact of adopting the intervention locally. But to use these models, policymakers would need to do some calculations themselves, combining the estimated coefficients from the model—if published (a rarity)—with local information about students and the environment in which the intervention would be implemented. There is also a tradeoff in these models between bias and variance, which we will discuss later, in the section "Selection of Subgroup and Moderator Variables for the Regression Models."

In summary, when local policymakers have access to published evidence from an RCT, they can typically obtain a pooled impact estimate and several subgroup estimates that could help them to predict the impact of adopting the intervention locally. Furthermore, they can produce additional impact estimates that may be relevant for predicting local impacts if the study reported regression models with multiple moderators—or they obtain the micro data necessary to estimate such models themselves. This paper compares the errors that result from using each of these types of estimates to predict the local impacts of the intervention.

RELATED LITERATURE

To our knowledge, little or no attention has been paid to the problem of translating the findings of large-scale, multi-site evaluations for use in local decision making. In contrast to the enormous amount of attention that has been devoted to issues of internal validity, researchers are just starting to consider external validity, also known as "generalizability" or "transportability"—that is, whether the causal effects found in one context or for one population hold in another context or population. Bareinboim and Pearl (2013) provide a theoretical basis for assessing whether findings from a study are "transportable" to another population or context.

Some recent research has focused on transportability from the sample used in an impact evaluation to the population from which it was selected. Shadish, Cook, and Campbell (2002) refer to this as generalizing "from narrow to broad" (p.22). The need for sophisticated methods in this regard arises from evidence of effect heterogeneity in multi-site randomized trials.

Substantial variation in impacts across sites has been found for a variety of educational interventions, including charter middle schools, small high schools of choice in New York City, Job Corps, and—at least for some of the outcomes examined—Head Start.[4] Similarly, Konstantopoulos (2011) found substantial variation in class size effects across schools.

Important theoretical and empirical work has addressed the challenges in making "narrow to broad" generalizations—i.e., accurately predicting the average impact in a population when impacts vary. Tipton (2013) provides a statistical basis for making generalizations across contexts. Stuart et al. (2011) and Tipton (2014) provide methods for assessing the likely transportability of study findings. Olsen et al. (2013) formalizes the external validity bias arising from estimating the population average treatment effect from a sample of sites that were selected or self-selected non-randomly from the population. Bell et al. (2016) presents empirical evidence on the magnitude of this bias in an education evaluation. Kern et al. (2016) test different analysis methods for reducing this bias and for more generally extrapolating from the study to a target population, while Tipton (2013) and Olsen and Orr (2016) offer different design solutions to the problem: Tipton offers methods for selecting sites systematically to match the population on observed characteristics, while Olsen and Orr demonstrate how sites can be selected randomly.

To our knowledge, the current work is the first research focusing explicitly on what Shadish, Cook, and Campbell refer to as generalizing "from the broad to the narrow" (p. 22) – i.e., from a collection of sites to individual sites outside the evaluation sample. This paper may be the first to test empirically how accurately local education policymakers can make that inductive leap, from the "broad" evidence of a multi-site RCT to the "narrow" impact predicted for the policymaker's own locality.

DATA AND METHODS

This section describes and justifies the data and methods used in the analysis to predict site-level impacts for educational interventions and assess the accuracy of those predictions.

Data

The data used in our analysis come from three different multisite RCTs in education/child development: (1) the Evaluation of Charter School Impacts (Gleason et al., 2010), (2) the Evaluation of the Effectiveness of Educational Technology Interventions (Dynarski et al., 2007; Campuzano et al., 2009), and (3) the Head Start Impact Study (Puma et al., 2010, 2012). The first two datasets were obtained via a restricted access license from the National Center for Education Statistics (NCES). The third dataset was obtained from the Interuniversity Consortium for Political and Social Research (ICPSR) at the University of Michigan. Below we briefly describe each of these studies:

[4] See Weiss et al. (forthcoming).

- *The Evaluation of Charter School Impacts* exploited charter school admission lotteries in 2005-06 and 2006-07 at 36 charter middle schools to estimate the impacts of attending a charter school on student achievement. To be eligible for the study, a charter middle school had to be oversubscribed—that is, it had to have more applicants than it could serve at the school's entry grade level—and use a lottery to admit students to the school. Lottery winners were included in the treatment group; lottery losers were included in the control group. The sample included almost 3,000 students who applied to one of the participating schools. The evaluation reported no significant average impact on student achievement, student behavior or progress in schools. However, it found that impacts vary substantially across schools, and in particular, that impacts were more favorable in schools that serve more low-income and low-achieving students.

- *The Evaluation of the Effectiveness of Educational Technology Interventions* randomized teachers to receive training and resources to implement a technology-related intervention in their classrooms in the 2004-05 school year. The study was conducted in grades 1, 4, and 6, as well as in algebra classes; the technology intervention tested varied across grade levels and whether they were focused on reading instruction or math instruction. The total sample included 132 schools, 439 teachers, and 9,424 students. The study reported no significant average impacts on student achievement in any of the grade levels or classes. Also, while the study displayed estimated impacts separately by school, no test of variation across schools was conducted. Finally, in most grade levels, the study found no significant relationship between the impact of the intervention and variables that might moderate the impact of the intervention.

- *The Head Start Impact Study* randomized almost 5,000 eligible 3- and 4-year-olds who had applied for the program in 2003 at one of 84 grantees that were randomly selected for inclusion in the study. Grantees had to be oversubscribed to be eligible for selection; the great majority in the nation were. Children in the sample were followed through the spring of third grade, and outcome data were collected in the areas of cognitive development, social-emotional development, health status and services, and parenting practices. The study found positive average impacts on exposure to high-quality early care and education environments, positive impacts on language and literacy development while enrolled in the program, and generally insignificant impacts on language, literacy, and math achievement in first grade and beyond. Subsequent research has identified substantial heterogeneity in impacts across centers (Bloom and Weiland 2015; Walters 2015) and further established that centers offering full-day service and frequent home visits delivered larger impacts (Walters 2015).

These studies were selected for three reasons. First, they evaluated the impacts of educational interventions that local policymakers could adopt—or could apply for funding to implement. Therefore, these studies are relevant for assessing our ability to inform local policy decisions using evidence from national studies. Second, they are based on randomized trials. We focus on randomized trials because random assignment maximizes the study's internal validity: This

allows us to focus on the external validity of study results when impacts may vary. Third, they are large studies spanning many states, and publicized on the websites of the U.S. Department of Education and the What Works Clearinghouse. Therefore, if randomized trials are going to be visible enough to influence local education policy, it would almost surely be through studies like these.

Because this paper tests different methods for predicting impacts in a single site, we first had to define what constitutes a site for each of the three studies:

- **Charter schools.** Conceptually, we defined the site as the local area from which a prospective charter school would draw its students. Operationally, each site was defined around a charter lottery.[5] The site was composed of the schools that students who entered the lottery would ultimately attend (typically the charter school for students who won admission in the lottery and typically regular public or private schools for students who did not win admission).

- **Education technology.** Because technology interventions can be implemented in individual schools, and principals face decisions about whether to adopt particular interventions in their schools, we defined the site as a single school.

- **Head Start.** Because Head Start funding is awarded through grants to local organizations, and localities must decide whether to apply for Head Start funding, we defined a site as the geographic area covered by a single Head Start grantee.[6]

Empirical Strategy

We simulate the use of results from multi-site randomized trials to predict impacts in a single school district outside the evaluation sample. [7] The simulations involve taking the actual data from a multisite evaluation that randomized students or classrooms within sites, pretending that one of the participating sites did not actually participate in the evaluation, and testing how well the impact in that site can be predicted using the characteristics of that site and evaluation data from the other sites.

Specifically, we apply the following procedure separately for each of the three multisite RCTs described above:

[5] Generally, each lottery was associated with a single charter school, but there were exceptions where multiple charter schools shared a single lottery (and, thus, a single site).

[6] An alternative would be to define the site as a single Head Start center, where each grant supports multiple centers. However, defining sites as grantee instead of centers allows us to focus on local policy decisions about whether to apply for Head Start funding.

[7] We focus on the results of randomized trials, which eliminate internal validity bias, to focus attention on the external validity of the impact estimates.

1. Begin with data from a multisite RCT that allows unbiased site-specific impact estimation (i.e., a study with within-site random assignment).
2. Select a statistical method for predicting the intervention's impact in individual sites. (Details of those methods below).
3. Pretend that one of the n sites in the evaluation was excluded from the sample.
4. Calculate the *predicted impact for the excluded site* by applying the statistical method from step 2 to the data from the other n-1 sites. This prediction may contain both bias and sampling error, as described earlier.
5. Calculate the *estimated impact for the excluded site* by exploiting the experiment conducted in that site. This estimate, derived from data for just the subject site, will be unbiased due to random assignment. It serves as our benchmark for estimating the amount of prediction error in the predicted impact estimate calculated at step 4.
6. Estimate the prediction error by taking the difference between the predicted impact for the excluded site (from Step 4) and the estimated impact for the excluded site (from Step 5).
7. Repeat steps 3-6 for each of the remaining n-1 sites to calculate an estimate of the prediction error for each site.
8. Calculate the Root Mean Squared Prediction Error (RMSPE) for the chosen statistical method across all sites in the RCT. As will be explained later, our approach to calculating the RMSPE accounts for the sampling error in the estimated impacts for the excluded sites.
9. Estimate the share of sites that would make the wrong policy decision due to the prediction error—that is, adopt the intervention when it should not be adopted or vice versa.
10. Repeat steps 2-9 for different statistical methods of predicting the impact in excluded sites and assess the relative performance of the different methods.

Prediction Methods Tested

To predict the impact in the excluded site (Steps 2 and 4), we apply three different methods:

- **Estimate the average, pooled impact for sites in the study sample.** This impact estimate is usually the main finding from an impact analysis; it can be used to predict the impact in the excluded site.
- **Estimate the impact for a subgroup (defined by a single variable) in which the excluded site falls.** Many RCTs produce impact estimates for selected subgroups of sites, such as separate estimates for urban and rural sites. If the excluded site is in an urban area, the estimated impact for urban sites could be used to predict the impact of the intervention for this site.

- **Estimate an equation that models the variation in impacts across sites as a function of multiple site-level variables.** Some impact analyses use "response surface modelling" (Box and Draper, 1987; Rubin, 1992) to model the impact of an intervention as a function of multiple site level moderator variables (e.g., urban/rural location, % low-income students, and baseline performance levels). The estimated regression model is then used to predict the impact in the excluded site, using that site's characteristics.

Regression Models

This section describes the regression models that we estimated to implement each of the three statistical methods described in the previous section. As a benchmark for calculating prediction errors, we estimated the impact in site *s*, which we pretend was excluded from the evaluation (Step 5, as described earlier). This estimate is calculated using only the data from site *s*. Because of random assignment, this estimate is unbiased for the impact in that site. [8]

To estimate the impact in the excluded site—site s —we used the following regression model using data from that site:

(1) $$y_{is} = \alpha_s + X'_{is}\beta_s + \delta_s T_{is} + e_{is},$$
$$e_{is} \sim N(0, \sigma_e^2)$$

where:

- y_{is} is the outcome for student *i* in site *s*.
- X'_{is} is a vector of student-level covariates included to improve the precision of the estimates.
- T_{is} is the treatment indicator, which equals 1 if student *i* in site *s* was assigned to the treatment group and 0 if this student was assigned to the control group.
- e_{is} is a random error term that varies across the students in site *s*.

The first prediction method examined involves estimating the average, pooled impact for sites—denoted by the subscript *j*—that we treat as being part of the study sample ($j \neq s$) and using this estimate as the predicted impact in the excluded site ($j = s$). To estimate this impact, we estimated the following regression model using data from all sites but the excluded site:

(2) $$y_{ij} = \alpha_j + X'_{ij}\beta + \delta_j T_{ij} + e_{ij},$$
$$\alpha_j = \alpha + u_j$$
$$\delta_j = \delta + v_j,$$

[8] For Charter Schools and Educational Technology, we used PROC REG in SAS to estimate the regression models--Ordinary Least Squares for equation (1) and restricted Maximum Likelihood (ML) for equations (2)-(4). For Head Start, we estimated all regression models in R using the nlme package, and results were compared to SAS results to verify correspondence.

$$\begin{pmatrix} e_{ij} \\ u_j \\ v_j \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_e^2 & 0 & 0 \\ 0 & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_v^2 \end{bmatrix} \right),$$

where most of these terms were defined for equation (1), but in addition:

- j indexes each variable for site *j,* where *j* ranges from 1 to n-1 omitting site *s*, and n is the total number of sites.
- u_j is a random component of the intercept that varies across sites.
- v_j is a random component of the impact that captures the difference between the impact in site *j* and the average impact across all sites.

The estimate of δ was used to predict the impact in site *s*.

The second prediction method examined involves estimating impacts for different subgroups of sites defined by a single variable. This approach involves estimating an enhanced version of equation (2) that adds a binary variable that classifies sites into different subgroups (S_j) and an interaction term between the subgroup variable and the treatment indicator:

(3) $$y_{ij} = \alpha_j + X'_{ij}\beta + \delta_j T_{ij} + e_{ij},$$

$$\alpha_j = \alpha + \gamma S_j + u_j$$

$$\delta_j = \delta + \theta S_j + v_j,$$

$$\begin{pmatrix} e_{ij} \\ u_j \\ v_j \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_e^2 & 0 & 0 \\ 0 & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_v^2 \end{bmatrix} \right),$$

where $S_j = 1$ for sites in one subgroup and $S_j = 0$ for sites in the other subgroup. The estimated impact for site *j* is $\hat{\delta} + \hat{\theta} S_j$, where $\hat{\delta}$ is the estimate of δ and $\hat{\theta}$ is the estimate of θ.

The third method examined involves estimating an equation that models impact as a function of one or more site-level variables – a "response surface model." This approach involves augmenting the regression model to include interaction terms between each of the included moderator variables and treatment, as well as estimating main effects for each moderator. The distinctions between this and the previous approach are that 1) moderator variables are included in continuous, rather than binary, form, and 2) multiple moderator variables are potentially included. This approach uses a model of the following form:

(4) $$y_{ij} = \alpha_j + X'_{ij}\beta + \delta_j T_{ij} + e_{ij},$$

$$\alpha_j = \alpha + Z'_j\gamma + u_j$$

$$\delta_j = \delta + Z'_j\theta + v_j,$$

$$\text{Var}\begin{pmatrix} e_{ij} \\ u_j \\ v_j \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_e^2 & 0 & 0 \\ 0 & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_v^2 \end{bmatrix}\right),$$

where Z_j' is a vector of site-level variables that moderate the effect of the intervention. The estimated impact for site *j* is $\hat{\delta} + Z_j'\hat{\theta}$, where $\hat{\delta}$ is the estimate of δ and $\hat{\theta}$ is the estimate of the coefficient vector θ.

Potential Moderator Variables

Two of the prediction methods—the subgroup approach reflected in equation (3) and the response surface modelling approach reflected in equation (4)— require site-level subgroup or moderator variables. Therefore, we used the available data to select or construct site-level variables that could be used to predict the impact of adopting the intervention in a single site. To identify candidate variables, we relied on published reports from the studies involved, each of which tended to identify variables that were likely to moderate the impacts of the intervention.

However, some of the subgroup and moderator variables identified from these studies are not helpful in predicting local impacts in advance because they are not knowable before the intervention is implemented. In selecting or constructing site-level variables for the analysis, we focused on variables that the local policymaker would know or could learn before deciding whether to implement the intervention locally. These include characteristics of the local students (e.g., share of students who are black, share of students who are in poverty) and schools (e.g., urbanicity). But they exclude characteristics of the intervention (e.g., how it was implemented) or characteristics that would solely describe intervention participants (which may not be known until after the intervention is adopted).)

For the education technology study, the subgroup and moderator variables came from data on individual schools. These school-level data were used to directly construct site-level variables for the education technology study because each site consisted of a single school. However, for the charter school study, each site consisted of multiple schools. To construct site-level variables from school-level data, we took the average of the school-level values from schools attended by students who did not win admission in the charter lottery,[9] weighted according to the relative size (by total enrollment) of these schools. For the Head Start study, we constructed grantee-level variables by averaging variables for individual Head Start centers funded by a given grant—including variables that were aggregated from the child level—weighting the center-level variables by the number of children in the center. In this way, the Head Start site-level variables reflect the population of students that would be served by the grantee.

[9] We did not include the characteristics of schools attended by students who *won* admission in the charter lottery (typically the charter school itself) in this aggregated measure because the characteristics of the charter schools would not be known before the policy decision is made about whether to allow charter schools to operate locally.

While most of the moderator variables were continuous, the subgroup analysis from equation (3) requires categorical or binary variables to divide the sample into subgroups that are mutually exclusive and exhaustive. To construct binary variables from the continuous ones, we calculated the median value of the variable across the sites and set the subgroup variable to one for sites that were above the median and zero for sites that were at or below the median.

Selection of Subgroup and Moderator Variables for the Regression Models

For the analysis, we constructed 7 moderator/subgroup variables for the education technology study, 10 moderator/subgroup variables for the Head Start study, and 11 moderator/subgroup variables for the charter school study (see Exhibit 1).

The approach in equation (3) uses a single binary subgroup variable. The approach in equation (4) uses any number of continuous moderator variables. Both approaches require a strategy for selecting the variable or variables that will be included in the regression model, and the response surface modelling approach also requires a decision on how many variables to include.

The optimal number of moderators to include is not clear. Including too many moderators risks overfitting the regression model, which could inflate the variance of the predictions; including too few moderators could yield site-level predictions with a large amount of bias. Since both bias and variance contribute to RMSPE, there is a bias-variance tradeoff in choosing the number of moderators, and it is not clear a priori how to determine the optimal number of moderators.

To address this problem, we estimated equation (4) with one, two, and five moderators. If we found evidence that models with five moderators consistently outperformed models with two moderators, we were prepared to test models with more than five moderators. The differences between the RMSPE across these three models helps to illuminate the value of additional moderator variables in the response surface function.

The optimal strategy for selecting moderator variables for equation (4) is also not clear. Our preferred strategy was to select the moderators that minimized the unexplained variance of impacts across sites—since this variance leads to biased local predictions, as explained earlier. However, when the number of candidate moderators was large, this would require testing a very large number of combinations of moderators to identify the combination for equation (4) with the smallest residual impact variance. Therefore, for tractability, for each study we first selected the five moderators that, when each is interacted individually with the treatment, yielded the smallest *p*-values (the ones most strongly associated with impact magnitude). These five variables became the pool of candidate moderators eligible for inclusion in that model.

From the moderators available in each study, we followed the following protocol to select specific moderators for each analytic approach:

- **Binary subgroup approach (equation 3)**. We created a binary subgroup variable from each of the five candidate moderators, tested all five possible binary subgroup models, and selected the single subgroup variable that minimized the unexplained variance of impacts across sites.

- **One-moderator model (equation 4).** We tested all five possible one-moderator models from the eligible pool of continuous variables and selected the single moderator that minimized the unexplained variance of impacts across sites.

- **Two-moderator model (equation 4).** We tested all ten possible two-moderator models from the eligible pool and selected the two moderators that together minimized the unexplained variance of impacts across sites.

- **Five-moderator model (equation 4)**. All five candidate moderators from the eligible pool were included in the five-moderator model.

Exhibit 1 shows the site-level variables that were candidate moderators for inclusion in the four types of models listed above; it also identifies in its footnotes the site-level variables that were most commonly selected for each model. A complication reflected in this exhibit is that the model selection strategy was implemented separately for each excluded site, or put differently, for each sample of n-1 sites that we treat as having been included in the evaluation. Therefore, a different set of selected moderators could be selected for each of these samples. For each of the four models with site-level moderators Exhibit 1 lists the moderators that were most frequently selected via the protocol described above.

Exhibit 1: Site-Level Moderators for the Analysis

Moderator	Charter Schools [a]	Education Technology [b]	Head Start [c]
Income	% of students eligible for free or reduced-price lunch [2, 3, 4]	% of students eligible for free or reduced-price lunch [4]	% of children in households with income below the median for the study sample [1, 3, 4]
Race and ethnicity	% of students who are white and not Hispanic [4]	% of students who are black [1, 2, 3, 4] % of students who are Hispanic [3, 4]	% of children who are black [2,4] % of children who are Hispanic [4]
Language			% of children with Spanish as home language [4]
Sex			% of children who are female [3,4]
Disability		% of students who have an IEP or Service Agreement	
Student-teacher ratio	# students / # teachers	# students / # teachers [4]	
Urbanicity	% of students enrolled in schools in large cities [4]	% students enrolled in schools in urban areas [4]	% of children at centers in urban areas
School size	Total number of students Total enrollment divided by grades served		
Teacher experience	% students in schools with more than two-thirds of the teachers having at least five years of experience		
Achievement in math and reading [d]	Difference between the school proficiency rate and the state proficiency rate in those grade levels in: • Math [4] • Math and reading [1, 4]		
Instructional approach	Proportion of all students attending control schools in the site who are in schools that use "ability grouping" [2, 3]		
Staffing		Whether the school has a technology specialist on staff	
Availability of similar services in the community			% of children at centers with a lot, some, or little competition from other providers in the area
Affiliations			% of children at centers affiliated with a: • Community-based organization • Government entity • Another type of organization

Notes:

[1] This variable was the most common moderator selected for the subgroup approach.

[2] This variable was the most common moderator selected for the single-moderator response surface modeling approach.

[3] This variable was in the most common set of moderators selected for the two-moderator response surface modeling approach.

[4] This variable was in the most common set of moderators selected for the five-moderator response surface modeling approach.

Measuring the Magnitude of the Errors in Predicting Local Impacts

To assess the accuracy of the predicted impacts under different prediction methods, we estimated an adjusted version of the root mean squared prediction error (RMSPE). The adjustment accounts for the sampling error in our unbiased estimate of the "true" impact in an excluded site—error that without some adjustment, inflates the magnitude of both the MAPE and the RMSPE.

Our adjusted estimate of the RMSPE is:

$$(5) \qquad \widetilde{RMSPE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\tilde{\Delta}_j - \hat{\Delta}_j\right)^2 - \frac{1}{n}\sum_{j=1}^{n}\hat{\sigma}_{\epsilon j}^2},$$

where n is the number of sites, $\tilde{\Delta}_j$ is the predicted impact for site j (using the data from the other sites), $\hat{\Delta}_j$ is the unbiased within-site estimate from the excluded site (using the data from site j), and $\hat{\sigma}_{\epsilon j}^2$ is the variance of the sampling error in $\hat{\Delta}_j$. Subtracting the mean value of $\hat{\sigma}_{\epsilon j}^2$ across the sites adjusts for the sampling error that would not exist in a true simulation framework, where the true impact in each site was known. If we knew the true impact in each site (with no sampling variability), equation (5) would simplify to the more familiar expression for the RMSE: $\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\tilde{\Delta}_j - \Delta_j\right)^2}$. A formal derivation for equation (5) is provided in Appendix A.

Hypothesis Tests

The RMSPE estimates allow us to compare the performance of different prediction methods. However, a different sample drawn from the same population would yield a different estimate of the RMSPE. Therefore, it is useful to test whether the differences in the magnitude of the (squared) prediction errors are statistically significant. To test for significant differences between methods, we conducted a binomial or sign test. We tested the null hypothesis that the true prediction error is the same for both methods in every site. If we can reject this hypothesis, we may conclude that the two methods perform differently in one or more sites (the alternative hypothesis). If the null hypothesis were true, the two methods would have the same true RMSPE across sites but different estimated RMSPEs due to sampling error.

Key implications of the null hypothesis are that:

- For each site, the probability that the estimated squared prediction error is lower for one method than the other method is exactly 50 percent since this difference is purely random.
- In expectation, each method will yield a smaller estimated squared prediction error than the other method in exactly half of the sites.
- In practice, one method may yield a smaller estimated squared prediction error than the other method in more than half of the sites. Call this percentage P1.

- Assuming independence across sites,[10] the binomial distribution can be used to calculate the probability that one of the two methods outperforms the other in more than P1 percent of the sites. Call this probability P2, where P2 is the p-value of the sign test.

We rejected the null hypothesis if the *p*-value of the test was less than .05 (5 percent) using a two-tailed test.

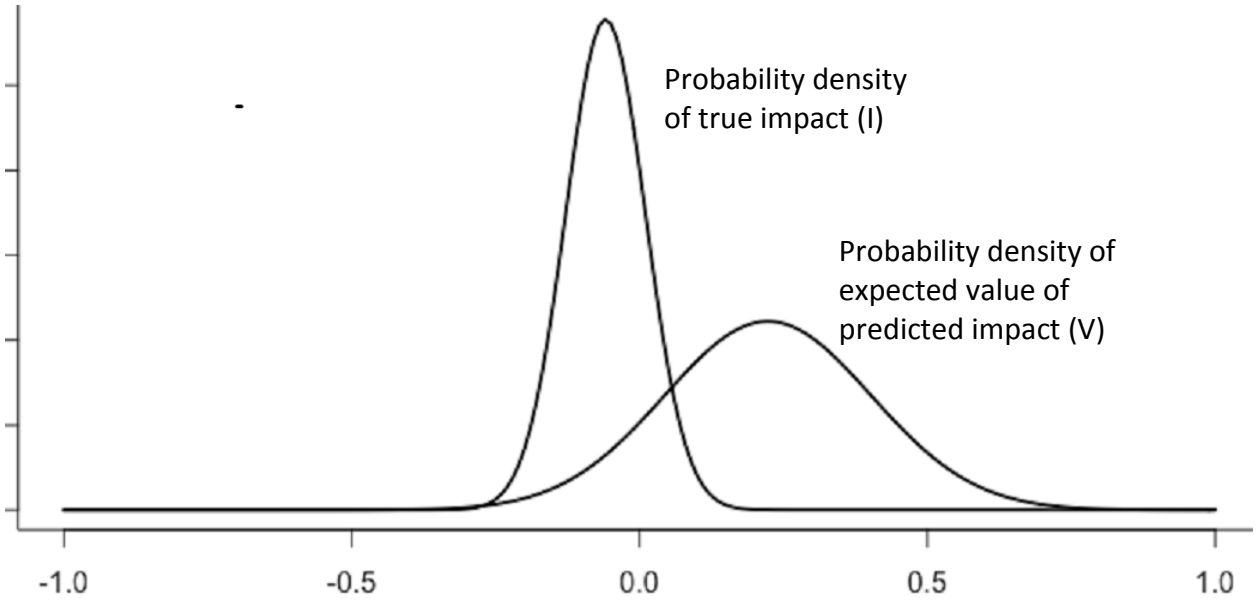
Policy Consequences of the Errors in Predicting Local Impacts

Bell and Orr (1995) developed a Bayesian method to assess the risk of making an incorrect policy decision on the basis of a nonexperimental estimator. We apply that method here to assess the predicted impact estimates obtained from the methods described above. Bayesian methods posit an *a priori* distribution of possible values for a population parameter, such as true impact, by attaching a subjective probability to every possible value of that parameter. A fundamental theorem of Bayesian statistics states that, when one begins with an agnostic view of the size of a parameter, the probability distribution that the analyst should construct for that parameter based on data from a sample should be centered on the parameter estimate produced by the sample (DeGroot, 1970, pp. 190-191). In addition, if the sample estimate has a normal distribution, the probability distribution of possible parameter values also follows a normal distribution, with standard deviation equal to the standard error of the parameter estimate (Bell and Orr, 1995).

Starting with an agnostic view of the true impact, I, in the excluded site, and observing the value and standard error of a single experimental impact estimate based on data from that site (referred to as the "estimated impact" above), it is possible to formulate a "posterior distribution" for the site's true impact. This also applies to the expected value of the estimate, V, for that site taken from data on other sites in the evaluation (referred to as the "predicted impact" above), which we treat as another unknown parameter. Together, these two distributions, one for the true impact in the excluded site and one for the expected value of the prediction for that site, provide a basis for assessing the policy reliability of the prediction method (see Exhibit 2).

[10] The estimated squared prediction errors are not strictly independent across sites because the samples used to predict the impact for each site overlap considerably, given the design of our leave-one-out analysis. However, most of the sampling variation in the estimated prediction error comes from the estimated impact for the excluded site, and these estimated impacts are independent from one another because the samples are non-overlapping. Therefore, the correlation between the squared estimated prediction errors for any two sites is sure to be small.

Exhibit 2: Bayesian Posterior Distributions of True Impact in Excluded Site and Expected Value of Predicted Impact, Under Agnostic Prior



We characterize the policy decision rule in the excluded site as a simple yes/no decision that depends on policymakers' beliefs about the impact of the intervention, which we denote C. If policymakers believe that the impact of the intervention exceeds some cut-off value C* that makes the intervention appealing to adopt, they will adopt the program; if they believe it does not, they will not adopt the program. C* can represent any binary decision rule; it could, for example, be the value that makes the program cost-effective in a benefit-cost analysis, or it could simply be the minimum value that policymakers would judge "practically" significant for policy. Since we do not know the value of C* in any particular application—indeed, the same program may have a different value of C* in different settings—we will consider a range of values of C.

The risk, R, that the predicted impact V will lead to the wrong decision when the true impact is I and the cut-off is C*, is:

(6) $R(C^*) = Pr(V < C^* \text{ and } I > C^*) + Pr(V > C^* \text{ and } I < C^*)$

where:

$Pr(V < C^* \text{ and } I > C^*)$ is the probability that the predicted impact will show the program was ineffective (i.e., $V < C^*$) when the program is effective (i.e., $I > C^*$);

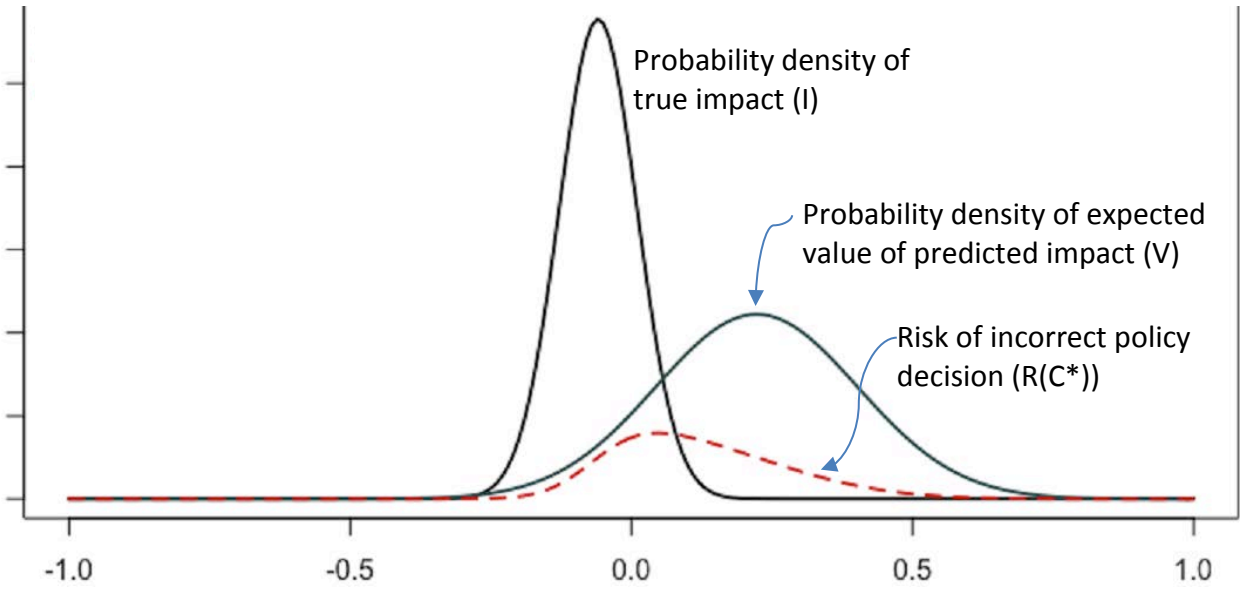
Pr (V > C* and I < C*) is the probability that the predicted impact will show that the program is effective (i.e., V > C*); when the program is ineffective (i.e., I < C*).

Bell and Orr call formula (6), traced out over a range of values for C*, the "risk function." In the special case of zero correlation between V and I, these two random variables (normally distributed) are independent and the risk formula reduces to equation (7):

(7) $R(C^*) = Pr(V < C^*) \cdot Pr(I > C^*) + Pr(V > C^*) \cdot Pr(I < C^*)$

Unfortunately, there is no exact analytic expression for R(C*) when V and I are correlated. However, we were able to develop a very accurate approximation to R(C*) for correlated V and I, and found that the results of our analysis were very insensitive to correlations of +.5 and -.5 (see Appendix B). Therefore, in the absence of any evidence or theory to indicate whether this correlation would be positive or negative, we use the probability given by equation 7, which assumes that V and I are uncorrelated, as our estimate of R(C*). This function, and the two Bayesian distributions from which it is derived, are shown in Exhibit 3 for an illustrative outcome.

Exhibit 3: Risk Function for Making an Incorrect Policy Decision, R(C*)



As with our measure of the accuracy of the predictions, the RMSPE, we compute the risk function for each site in each of three national evaluations, based on the predicted impact from the other sites in that study and the within-site experimental estimate (our estimate of the true

impact). For each possible value of C*, we then sum the calculated risks across all sites from a given evaluation and divide by the number of sites; this yields the expected proportion P(C*) of sites that would make the wrong decision on the basis of the extrapolated estimates if C* is the desired policy support cut-point:

$$(8) \qquad P(C^*) = \frac{\sum_{i=1}^{n} R_j(C^*)}{n},$$

where $R_j(C^*)$ is the probability of an incorrect policy decision in site j for a given value of C*. Thus, P(C*) is the average probability, across sites, of an incorrect decision. P(C*) is our summary measure of the risk that sites will make the wrong policy decision if they rely on predicted impacts derived from a national evaluation. In the empirical results, we show P(C*) as a percentage. If P(C*) for a given C* is quite low, policymakers can be confident that using the predictions will usually lead to the right decision for that cut-point; if it is quite high, the predictions will have a high risk of leading to the wrong decision.

EMPIRICAL RESULTS

In this section, we present our estimates of the accuracy of different prediction methods, followed by the estimated risk of making a wrong policy decision based on each of those predictions.

Accuracy of the Predictions

Exhibit 4 shows the root mean square prediction errors (RMSPEs) for five different prediction methods from the sites that participated in a given multi-site RCT: the pooled impact estimate across sites; subgroup estimates; and estimates from regression models that interacted one, two, or five treatment effect moderators with the treatment indicator. This table presents results separately for the evaluations of charter schools, educational technology, and Head Start. For each study, several different outcomes, denoted in the first two columns of the table, were analyzed. Superscripts to the estimates indicate whether one method produced smaller (squared) prediction errors than another method, using the binomial test described earlier.

The RMSPEs in the table are measured in standard deviations of the outcome variable; they vary from about 0.05 to nearly 0.40, with most clustered in the 0.10 - 0.30 range. Because the RMSPEs are measured in effect size units, comparing them to the average effect sizes shown in the last column of Exhibit 4 provides a sense of the magnitude of the prediction errors.[11] In general, the RMSPE estimates tend to be larger than the average effect size across the sites that participate in the RCT. This suggests that for the typical site, the prediction error may be larger than the impact that local policymakers are trying to predict.

[11] However, the calculation of the Root Mean Squared Prediction Error places a heavier weight on outliers than the average effect size, which places equal weight on larger and smaller values.

Exhibit 4: RMSPE of Different Methods for Predicting Site-Specific Impacts, by Study and Outcome Domain

Outcome Domain	Grade Level	Pooled Analysis	Subgroup Analysis	1-Moderator Model	2-Moderator Model	5-Moderator Model	*Published Impact Estimate**
Charter Schools							
Math	6	0.175	0.176	0.170	0.171	0.190	*-0.06*
Math	7	0.348	0.371[1,2,5]	0.214[s]	0.181[s]	0.156[s]	*-0.06*
Reading	6	0.216[1,2,5]	0.246[2]	0.264[p]	0.283[p,s]	0.284[p]	*-0.07*
Reading	7	0.189[1]	0.164	0.244[p]	0.248	0.259	*-0.08*
Educational Technology							
Math	6	0.272	0.305	0.296	0.297	0.314	*-0.15*
Math	Algebra	0.119	0.146	0.140	0.131	0.203	*0.15*
Reading	1	0.305	0.311	0.304	0.293	0.329	*-0.06*
Reading	4	0.169[s]	0.205[p]	0.171[2]	0.154[1]	0.163	*0.22*
Head Start							
Receptive vocabulary	Pre-K	0.056	0.068	0.083	0.103	0.084	*0.15*
Early numeracy	Pre-K	0.073[2]	0.095	0.089	0.113[p,5]	0.043[2]	*0.12*
Oral comprehension	Pre-K	0.116	0.129	0.141	0.129	0.150	*0.01*
Early reading	Pre-K	0.206	0.209[5]	0.213[2]	0.231[1]	0.234[s]	*0.17*
Self-regulation	Pre-K	0.078[1,5]	0.097	0.137[p]	0.108	0.137[p]	*0.02*
Externalizing	Pre-K	0.201[5]	0.211	0.212[5]	0.230	0.256[p,1]	*-0.05*

[p]Significantly different from the pooled analysis.
[s]Significantly different from the subgroup analysis.
[1]Significantly different from the 1-moderator model.
[2]Significantly different from the 2-moderator model.
[5]Significantly different from the 5-moderator model.
* Published impact estimates for charter schools and education technology come from the final reports from those studies, Gleason et al. (2010) and Campuzano (2009), respectively. Published impact estimates for the Head Start impact study come from a reanalysis of those data in Bloom and Weiland (2015) because the reanalysis pools the 3-year-old and 4-year-old cohorts—as we did in our analysis—while the final evaluation reports presents separate estimates by cohort.

Exhibit 4 shows that more complex models, with treatment effect interactions, did not generally produce smaller prediction errors in these three studies. In fact, the pooled model yields a smaller estimated RMSPE than each of the more complex models for over half of the outcomes examined. For most outcomes, we were unable to reject the null hypothesis that the pooled method yields the same prediction error as each of the more complex methods. However, *all nine of the significant differences between the pooled model and more complex models favored the pooled model*.

Risk of Making the Wrong Policy Decision

In Exhibit 5, we show, for each prediction method and each study, across all outcomes in that study, the average probability of making the wrong policy decision when the policy cut-off C* is set to 0 standard deviations, .25 standard deviations, and .50 standard deviations. The exhibit also indicates the maximum probability of an incorrect policy decision for any outcome across all possible values of the cut-off, and the average RMSPE from Exhibit 4 for each study and for all three studies combined. (For outcome-specific risk estimates and the computer code used to generate these estimates, see Appendix C.)

The cut-off of 0 is relevant in cases where the school can, without additional cost, substitute one policy for another – e.g., by changing a regulation – or when choosing between two equally costly interventions, such as two curricula. The .25 cut-off would require that the proposed intervention be as effective as the typical high school intervention (see Hill et al., 2008 for evidence on the impacts of such interventions). The .50 cut-off corresponds to the effectiveness of the typical middle school intervention or to an atypically expensive elementary or high school intervention that must be correspondingly more effective to be appealing to local policymakers (Hill et al., 2008).

Exhibit 5: Average and Maximum Risk of Incorrect Policy Decision Across All Outcomes, by Study and Method of Predicting Site-Specific Impacts, Alternative Values of C*

Policy Cut-off	Pooled Analysis	Subgroup Analysis	1-Moderator Model	2-Moderator Model	5-Moderator Model
Charter Schools					
Avg. risk at C*=0	45%	47%	43%	42%	43%
Avg. risk at C*=.25	15%	16%	15%	15%	16%
Avg. risk at C*=.50	4%	4%	5%	5%	5%
Maximum risk	56%	61%	64%	66%	58%
Avg. RMSPE	*0.24*	*0.20*	*0.17*	*0.21*	*0.22*
Educational Technology					
Avg. risk at C*=0	49%	54%	49%	45%	48%
Avg. risk at C*=.25	22%	22%	22%	23%	26%
Avg. risk at C*=.50	7%	7%	7%	7%	8%
Maximum risk	55%	64%	60%	48%	52%
Avg. RMSPE	*0.23*	*0.25*	*0.23*	*0.22*	*0.25*
Head Start					
Avg. risk at C*=0	45%	46%	47%	46%	47%
Avg. risk at C*=.25	30%	31%	30%	31%	32%
Avg. risk at C*=.50	13%	13%	13%	13%	13%
Maximum risk	53%	53%	59%	55%	57%
Avg. RMSPE	*0.15*	*0.15*	*0.16*	*0.17*	*0.10*
Combined Studies					
Avg. risk at C*=0	47%	49%	47%	45%	46%
Avg. risk at C*=.25	29%	31%	31%	30%	32%
Avg. risk at C*=.50	12%	12%	12%	12%	13%
Maximum risk	56%	64%	64%	66%	58%
Avg. RMSPE	*0.21*	*0.20*	*0.19*	*0.20*	*0.19*

The risk estimates in Exhibit 5 follow several systematic patterns. First, for a given study and policy cut-off, there is little variation in the probability of an incorrect policy decision across prediction methods. Second, that probability tends to fall as the policy cut-off increases: Policy errors become less likely as the impact required for policy approval increases. Finally, the maximum risk of an incorrect policy decision tends to be in the range between 45% and 60%. (Note that the maximum risk shown for the pooled studies is the maximum of the individual study maxima.) As shown in the first three panels of the exhibit, across studies and prediction methods the average risk ranges from 42% - 54% when the policy cut-off is zero. When the cut-off is .25, this range falls to 15% - 32%. And when the cut-off is .50, the average risk across studies and prediction models is 10% - 21%.

These estimates suggest that in the case of C* = 0 – i.e., when the policymaker would be satisfied with any positive impact – the national evaluation is of little help. The chance of making the right policy decision at the local level based on evidence from such an evaluation is at best only slightly better than 50%—the rate we would expect if policymakers just flipped a coin.[12] At the other extreme, when only an impact of .50 or greater would induce the policy maker to adopt the intervention, on average, s/he will be led astray by the national evaluation only about 1 time in 8.

In the intermediate case of C* = .25, averaging across the three studies (the last panel of Exhibit 5), the probability of making the wrong policy decision by relying on the national evaluation is about 30%. This risk may seem high relative to our usual tolerance for the risk of Type I or II errors in hypothesis testing. But it must be viewed relative to the risk of making the wrong decision in the absence of a national evaluation. In most cases, local policymakers have virtually no information about the likely impact of a new intervention in the absence of a national evaluation, and without that evidence, they are unlikely to have less than a 50% probability of making the wrong decision. If that is true, in this intermediate case where policy makers require an impact of .25, the national evaluation reduces the risk of the wrong policy decision by about two-fifths, from 50 percent to 30 percent.

DISCUSSION AND IMPLICATIONS

The empirical results presented in this paper suggest—at least for the three interventions examined—that most localities will *not* be able to predict accurately in advance the likely consequences of adopting an educational intervention or policy with any of the five approaches tested here. The prediction errors tend to be very large for all five methods, in all three studies, and in fact, there is no evidence that more complex models improve prediction accuracy.

[12] Note that this does *not* mean that the intervention is as likely to be effective as ineffective. Regardless of the probability of the intervention being effective, flipping a coin involves a 50% risk of making the wrong policy decision.

Furthermore, our estimates almost surely understate the size of the typical prediction error because they fail to adjust for the homogenizing effect of conducting the RCT in an unrepresentative sample of sites. Out-of-sample prediction errors for schools that fall outside of the distribution observed in the original study sample cannot be calculated because there is no way to obtain an unbiased estimate of impact for schools that were not part of that sample. However, if there are systematic differences in the types of schools that participated in the RCT and schools that did not, out-of-sample prediction errors for sites not included in the original sample are likely to be larger than for the sites examined here that were included in the original studies.[13]

A unique contribution of this analysis is the ability to estimate the probability of making the wrong local policy decision on the basis of the national evaluation. That probability depends strongly on the size of impact the policy maker requires to justify adopting the intervention. For costless interventions, where any positive impact would justify adoption, a national evaluation is of little help. On the other hand, for interventions where a large (.50 or larger) effect size is required for adoption, the national evaluation substantially reduces the risk of making the wrong policy decision, to a level of 12 - 13%. For intermediate policy cut-offs, the risk is still substantial – in the range of 30% -- but roughly two-fifths less than it would have been in the absence of the evaluation.

Furthermore, more complex models generally did not reduce the probability of making the wrong policy decisions, which is not surprising given that they generally did not reduce the magnitude of the prediction errors. This may be due to the fact that these more complex methods estimated impacts as a function of site characteristics, yet national evaluations are almost never powered for such analyses. That may explain why including site characteristics that moderate impact magnitude does not necessarily help make the predictions more accurate. Evaluations that are powered to estimate site-level subgroup specific effects might show a better ability to predict site-specific impacts as more site characteristics are brought into the analysis as moderators.

These results are, course, based on a small sample of three studies that may or may not be typical in factors that influence findings from this type of analysis. Whether we would reach similar conclusions on the basis of national studies of other educational interventions is an open question. For example, we would expect more accurate predictions for interventions with less cross-site variation in impacts than those examined here.[14] More research is needed on both the magnitude of cross-site impact variation for educational interventions and the accuracy with which we can predict site-level impacts from multi-site impact evaluations in education.

[13] See Stuart et al. (2017) for an analysis of the characteristics of schools that participate in education RCTs, relative to the national population of schools.

[14] See Weiss et al. (forthcoming) for new evidence on the cross-site variance of impacts for 13 educational interventions.

It is important to recognize that national studies may and often do serve purposes other than informing local policy decisions. Many national studies funded by the federal government are designed to inform federal policy decisions, for which estimates of the overall effectiveness of an intervention across a diversity of settings/sites may be most informative. For that purpose, we would expect national studies to produce more accurate evidence to guide policy. However, even when used for that purpose, national studies may provide misleading evidence if impacts vary across sites and sites are selected non-randomly (see Allcott, 2015 and Bell et al., 2015).

While these results suggest caution in extrapolating the results of national evaluations to local schools or school districts, especially for low-cost interventions, our objective here is not to reach definitive conclusions about the usefulness of national evaluations for informing local decisions. Rather, our primary objective is to draw the attention of evaluators and policymakers to the challenges of making local predictions from national studies and to develop and demonstrate a method for analyzing the problem. We hope that this will motivate other researchers to pursue similar analyses and, ultimately, to the development of a literature on external validity similar to the design replication literature that has been built over the last 30 years to assess the internal validity of nonexperimental methods for impact analysis.

REFERENCES

Aragon, S., Griffith, M., Wixom, M. A., Woods, J., & Workman, E. (2016). ESSA: Quick Guides on Top Issues. Denver, CO: Education Commission of the States.

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of External Validity Bias When Impact Evaluations Select Sites Nonrandomly. Educational Evaluation and Policy Analysis, 0162373715617549.

Bloom, H. S., & Weiland, C. (2015). Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the national Head Start Impact Study. MDRC Working Paper.

Box, G. E., & Draper, N.R. (1987). *Empirical model-building and response surfaces*. Hoboken, NJ: Wiley.

Burnette, D. (2017). In Some States, ESSA Means More Power for Local School Boards. State EdWatch blog, Education Week, May 19, 2017.

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts. NCEE 2009-4041. National Center for Education Evaluation and Regional Assistance.

DeGroot, M. H. (1970). Optimal Statistical Decisions. New York, NY: McGraw-Hill.

Gleason, P., Clark, M., Tuttle, C. C., & Dwoyer, E. (2010). The Evaluation of Charter School Impacts: Final Report. NCEE 2010-4029. National Center for Education Evaluation and Regional Assistance.

Hill, C. J., Bloom, H. S., Black, A. R. and Lipsey, M. W. (2008), Empirical Benchmarks for Interpreting Effect Sizes in Research. Child Development Perspectives, 2: 172–177. doi:10.1111/j.1750-8606.2008.00061.x

Hotz, V. J., Imbens, G. W., & Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. Journal of Econometrics, 125(1), 241-270.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. Journal of Research on Educational Effectiveness, 9(1), 103-127.

Konstantopoulos, S. (2011). How consistent are class size effects? Evaluation Review, 35(1), 71-92.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. Journal of Policy Analysis and Management, 32(1), 107-121.

Olsen, R. B. and Orr, L. L. (2016). On the "Where" of Social Experiments: Selecting More Representative Samples to Inform Policy, in *Social Experiments in Practice: The What, Why, When, Where, and How of Experimental Design & Analysis*, special issue of *New Directions for Evaluation*, No. 162 (Winter).

Puma, M., Bell, S., Cook, R., and Heid, C. (2010). Head Start Impact Study. Final Report. U.S. Department of Health and Human Services, Administration for Children & Families.

Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational and Behavioral Statistics*, 17, 363–374.

Stuart, E.A., Bell, S.H., Ebnesajjad, C., Olsen, R.B., and Orr, L.L. (2017) "Characteristics of school districts that participate in rigorous national educational evaluations", *Journal of Research on Educational Effectiveness*, Vol. 10, No. 1.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society: Series A (Statistics in Society), 174(2), 369-386.

Tipton, E. (2013). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. Evaluation review, 37(2), 109-139.

Tipton, E. (2014). How generalizable is your experiment? Comparing a sample and population through a generalizability index. Journal of Educational and Behavioral Statistics, 39(6), 478-501.

Walters, C. R. (2015). Inputs in the production of early childhood human capital: Evidence from Head Start. American Economic Journal: Applied Economics, 7(4), 76-102.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. Journal of Policy Analysis and Management, 33(3), 778-808.

Weiss, M.J., Bloom, H.S., Verbitsky-Savitz, N., Gupta, H., Vigil, A., Cullinan, D. (forthcoming). How much do the effects of education and training programs vary across sites? Evidence from existing multisite randomized trials. Journal of Research on Educational Effectiveness.

Appendix A: Derivation of Adjusted Estimator for Root Mean Squared Prediction Error

As indicated in equation (5), we used an adjusted estimator for the root mean squared prediction error (RMSPE) to summarize the magnitude of the prediction errors across sites. This appendix formally demonstrates why an adjustment is necessary and derives the adjusted RMSPE estimator in equation (5).

To begin, suppose that:

- Δ_j is the true impact in site *j* (unknown for all sites)
- Δ_j can be estimated *without bias* for each site j, thanks to random assignment, with data from site *j* (the "unbiased impact estimate for site *j*")
- Δ_j can be predicted *with bias* for each site *j*, using a statistical model, with data from the other sites (the "predicted impact for site *j*")

With this foundation, we can define the prediction error for a single site, summarize the prediction errors across sites, and define an estimator that adjusts for or "nets out" the sampling error that would not exist if the true impact in each site were known.

1. PREDICTION ERROR

Equation (1) provides an expression for the *unbiased impact estimate for site j* from the randomized trial, using only the data from site *j*:

(1) $\hat{\Delta}_j = \Delta_j + \epsilon_j,$

where Δ_j is the true impact for site *j* and ϵ_j is sampling error in the estimate due to the finite sample in site *j*. This sampling error is assumed to be normally distributed: $\epsilon_j \sim N\left(0, \sigma^2_{\epsilon j}\right)$. The expected value of ϵ_j is zero because the estimator is unbiased.

Equation (2) provides an expression for the *predicted impact for site j* from the randomized trial, using the data for all study sites other than site *j*:

(2) $\tilde{\Delta}_j = \Delta_j + b_j + \omega_j,$

where b_j is the bias in the predicted impact and ω_j is sampling error in the prediction due to the finite sample in the sites used in making the prediction. This sampling error is assumed to be normally distributed—$\omega_j \sim N\left(0, \sigma^2_{\omega j}\right)$—and independent of ϵ_j. Note that b_j is a fixed parameter that is a function of the methodology used to predict the impact in site *j* using the data from other sites.

The prediction error for site *j* is just the difference between the predicted impact and the true impact:

$$
\begin{aligned}
(3) \qquad PE_j &= \tilde{\Delta}_j - \Delta_j \\
&= \left(\Delta_j + b_j + \omega_j\right) - \Delta_j \\
&= b_j + \omega_j
\end{aligned}
$$

Our best estimate of the prediction error for site *j* is the difference between the predicted impact and the unbiased (but noisy) impact estimate:

$$
\begin{aligned}
(4) \qquad \widehat{PE}_j &= \tilde{\Delta}_j - \hat{\Delta}_j \\
&= \left(\Delta_j + b_j + \omega_j\right) - \left(\Delta_j + \epsilon_j\right) \\
&= \left(b_j + \omega_j\right) - \epsilon_j
\end{aligned}
$$

Comparing equation (3) and (4), we can see that the estimated prediction error in equation (4) equals the true prediction error minus the sampling error in the unbiased (but noisy) impact estimate for site *j* (ϵ_j). Since this sampling error has an expected value of zero, the estimated prediction error for site *j* is unbiased for the true prediction error for site *j*. However, the variance of the estimated prediction error ($\sigma^2_{\omega j} + \sigma^2_{\epsilon j}$) exceeds the variance of the true prediction error ($\sigma^2_{\omega j}$).

2. MEAN SQUARED PREDICTION ERROR (MSPE)

In this section, we define the MSPE for an individual site, identify the most obvious estimate for this parameter, note that this estimate is biased upward, provide an alternative estimate that corrects for the bias, and average the corrected MSPE estimates across the sites.

2.1. MSPE for a Single Site

The mean squared error of the predicted impact for site *j*—which we call the mean squared prediction error (MSPE) for that site—is defined as the expected squared prediction error in site *j*. This expectation is defined across repeated samples selected to predict the impact in site *j*:

$$
\begin{aligned}
(5) \qquad MSPE_j &= E\left(PE_j^2\right) \\
&= E\left(b_j + \omega_j\right)^2 \\
&= E\left[b_j^2 + 2b_j\omega_j + \omega_j^2\right] \\
&= b_j^2 + 2b_j E\left(\omega_j\right) + E\left(\omega_j^2\right) && \text{because } b_j \text{ is a fixed parameter} \\
&= b_j^2 + \sigma^2_{\omega j} && \text{because } \omega_j \sim N\left(0, \sigma^2_{\omega j}\right)
\end{aligned}
$$

The equation above shows the familiar result that the Mean Squared Error is the sum of the squared bias and the variance.

For site *j*, the most obvious way to estimate the MSPE is to square the estimated prediction error:

(6) $\widehat{MSPE}_j = \widehat{PE}_j^2$

Unfortunately, this estimator is biased upward: the expected value of this estimator exceeds the true MSPE for site *j* by an amount that equals the variance of the unbiased estimate for site *j*:

(7)
$$\begin{aligned}
&E\left(\widehat{MSPE}_j\right) \\
&= E\left(\widehat{PE}_j^2\right) \\
&= E\left[\left(b_j + \omega_j\right) - \epsilon_j\right]^2 \\
&= E\left(b_j + \omega_j\right)^2 - 2E\left[\left(b_j + \omega_j\right)\epsilon_j\right] + E\left(\epsilon_j\right)^2 \\
&= b_j^2 + E\left(\omega_j^2\right) - 2\left[b_j E\left(\epsilon_j\right) + E\left(\omega_j \epsilon_j\right)\right] + E\left(\epsilon_j\right)^2 \\
&\qquad = b_j^2 + \sigma_{\omega j}^2 - 2\left[E\left(\omega_j \epsilon_j\right)\right] + \sigma_{\epsilon j}^2 \text{ because } \omega_j \sim N\left(0, \sigma_{\omega j}^2\right) \text{ and } \epsilon_j \sim N\left(0, \sigma_{\epsilon j}^2\right) \\
&= b_j^2 + \sigma_{\omega j}^2 + \sigma_{\epsilon j}^2 \qquad \text{because } \omega_j \text{ is independent of } \epsilon_j \\
&= MSPE_j + \sigma_{\epsilon j}^2
\end{aligned}$$

Fortunately, the bias in the estimated MSPE for site *j* can be estimated (without bias) and removed. Equation 7 shows that the bias equals the variance of the unbiased estimate for site *j* ($\sigma_{\epsilon j}^2$). Let $\hat{\sigma}_{\epsilon j}^2$ be the ordinary least squares estimate of the variance of the unbiased impact estimate for site *j* ($\hat{\Delta}_j$). Assuming that this variance estimate is unbiased (or at least consistent), we can construct an unbiased (or at least consistent) estimate of the MSPE for site *j*.

Let us define a new, corrected estimator for the MSPE in site *j*:

(8) $\widehat{MSPE}_j = \widehat{PE}_j^2 - \hat{\sigma}_{\epsilon j}^2$

The expected value of this estimator equals the true MSPE for site *j*:

(9)
$$\begin{aligned}
&E\left(\widehat{MSPE}_j\right) \\
&= E\left(\widehat{PE}_j^2 - \hat{\sigma}_{\epsilon j}^2\right) \\
&= E\left(\widehat{PE}_j^2\right) - E\left(\hat{\sigma}_{\epsilon j}^2\right) \\
&= \left(MSPE_j + \sigma_{\epsilon j}^2\right) - E\left(\hat{\sigma}_{\epsilon j}^2\right) \qquad \text{see equation (7)} \\
&= \left(MSPE_j + \sigma_{\epsilon j}^2\right) - \sigma_{\epsilon j}^2 \qquad \text{since } \hat{\sigma}_{\epsilon}^2 \text{ is unbiased} \\
&= MSPE_j
\end{aligned}$$

2.2. Average MSPE Across Sites

The previous section provides an unbiased estimate for the MSPE for a single site. However, for our leave-one-out exercise, we want to summarize the MSPEs across sites by taking the average. Let us define the parameter that we want to estimate as the average MSPE across the N sites in this sample:

(10) $MSPE = \frac{1}{N}\sum_{j=1}^{N} MSPE_j$

One estimator for this parameter is the simple average of the corrected, unbiased estimates for the MSPEs across the collection of sites:

(11) $\widetilde{MSPE} = \frac{1}{N}\sum_{j=1}^{N} \widetilde{MSPE}_j$

Since the corrected estimator for the MSPE in site *j* is unbiased for the true MSPE in that site, as shown in equation (9), the simple average of those estimators is unbiased for the simple average of the true MSPEs across all sites:

(12) $E\left(\widetilde{MSPE}\right) = E\left(\frac{1}{N}\sum_{j=1}^{N} \widetilde{MSPE}_j\right)$
$= \frac{1}{N}\sum_{j=1}^{N} E\left(\widetilde{MSPE}_j\right)$
$= \frac{1}{N}\sum_{j=1}^{N} MSPE_j$
$= MSPE$

Therefore, using Equations (11), (8), and (4), our measure of the average MSPE across sites is:

(13) $\widetilde{MSPE} = \frac{1}{N}\sum_{j=1}^{N}\left[\left(\tilde{\Delta}_j - \hat{\Delta}_j\right)^2 - \hat{\sigma}_{\epsilon j}^2\right] = \frac{1}{N}\sum_{j=1}^{N}\left(\tilde{\Delta}_j - \hat{\Delta}_j\right)^2 - \frac{1}{N}\sum_{j=1}^{N}\hat{\sigma}_{\epsilon j}^2,$

where the $\frac{1}{N}\sum_{j=1}^{N}\left(\tilde{\Delta}_j - \hat{\Delta}_j\right)^2$ is the average of the squared prediction error estimates and $\frac{1}{N}\sum_{j=1}^{N}\hat{\sigma}_{\epsilon j}^2$ is the average of the variance estimates for the unbiased, site-level impact estimates.

Appendix B: Estimating the Risk Function for Correlated V and I

1. Computation of R(C*) when Estimates are Uncorrelated

The general formula for the risk function is:

B.1) $R(C^*) = Pr(V < C^* \text{ and } I > C^*) + Pr(V > C^* \text{ and } I < C^*)$

where:

$Pr(V < C^* \text{ and } I > C^*)$ is the probability that the predicted estimate will show the program was not effective (i.e., $V < C^*$) when the program is effective (i.e., $I > C^*$);

$Pr(V > C^* \text{ and } I < C^*)$ is the probability that the predicted impact estimate will show that the program is effective (i.e., $V > C^*$); when the program is ineffective (i.e., $I < C^*$).

In the special case of zero correlation between V and I, these two random variables (normally distributed) are independent and the risk formula reduces to:

B.2) $R(C^*) = Pr(V < C^*) \cdot Pr(I > C^*) + Pr(V > C^*) \cdot Pr(I < C^*)$

2. Approximating R(C*) When I and V are Correlated

Unfortunately, there is no closed-form solution for the probabilities in the risk function formula B.1 when the within-site impact (I) and the expected value of the predicted impact for the site (V) are correlated. However, we can approximate the probabilities in equation B.1 as follows.

We express the probability space over which R(C*) is to be calculated as a grid of small squares of width w, each centered on a point (x_{iv}, y_{iv}). Within each of these squares that satisfy either of the conditions in equation B.1 (either ($x_{iv} < C^*$ and $y_{iv} > C^*$) or ($x_{iv} > C^*$ and $y_{iv} < C^*$)), we calculate the probability density of the bivariate normal distribution, for which there *is* a closed-form expression that depends on the correlation between x and y. For each value of C*, we then sum the product of these probabilities times the area of each square (w^2), over the entire space satisfying the conditions in equation B.1. This sum equals R(C*). In principle, the bivariate normal distribution extends from minus infinity to plus infinity; to render the problem computationally tractable, we truncated the space to $\pm$ 4 standard deviations.

B.3) $R(C^*) = \sum_{v=-4/w}^{+4/w} \sum_{i=-4/w}^{+4/w} P_{iv} w^2 f(x_{iv}, y_{iv})$, where:

$P_{iv} = 1$ if either ($x_{iv} < C^*$ and $y_{iv} > C^*$) or ($x_{iv} > C^*$ and $y_{iv} < C^*$)
$= 0$ otherwise

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

= bivariate normal probability density function

$\rho = \text{Correl}(x, y)$

Setting the correlation between V and I to zero, we tested two different grid widths, w=.01 and w=.005 standard deviations, against the values of R computed by formula in the case where the correlation between V and I is zero. The results are shown in Exhibit B.1, which shows our central risk measure, P(C), the average value of R(C*) across all sites, for each of the outcomes in the analysis. The computer code used to generate these estimates is presented in the final section of this appendix.

We found that the estimates given by w=.005 were quite close to the values yielded by the formula (see average differences in last three rows of Exhibit B.1). For that reason, we use areas of width .005 in the approximations shown in the following section of this appendix.

Exhibit B.1 P(C) Computed by Formula and by Two Approximations, ρ = 0

	By formula (A)	.01 Approx (B)	.005 Approx (C)	Difference (B) - (A) (D)	Difference (C) - (A) (E)
Charter year 1 math					
C* = 0	0.412	0.389	0.400	-0.023	-0.012
C* = .25	0.112	0.108	0.110	-0.004	-0.002
C* = .50	0.019	0.018	0.018	-0.001	-0.001
Charter year 2 math					
C* = 0	0.357	0.344	0.351	-0.013	-0.006
C* = .25	0.217	0.211	0.214	-0.006	-0.003
C* = .50	0.111	0.108	0.110	-0.003	-0.001
Charter year 1 reading					
C* = 0	0.572	0.545	0.559	-0.027	-0.013
C* = .25	0.144	0.139	0.141	-0.005	-0.003
C* = .50	0.034	0.034	0.034	0.000	0.000
Charter year 2 reading					
C* = 0	0.421	0.407	0.414	-0.014	-0.007
C* = .25	0.148	0.144	0.146	-0.004	-0.002
C* = .50	0.030	0.029	0.030	-0.001	0.000
Ed Tech math 6					
C* = 0	0.481	0.463	0.472	-0.018	-0.009
C* = .25	0.327	0.317	0.322	-0.010	-0.005
C* = .50	0.099	0.096	0.097	-0.003	-0.002
Ed Tech algebra					
C* = 0	0.465	0.445	0.455	-0.020	-0.010
C* = .25	0.178	0.174	0.176	-0.004	-0.002
C* = .50	0.051	0.050	0.050	-0.001	-0.001
Ed Tech TOWRE					
C* = 0	0.529	0.507	0.518	-0.022	-0.011
C* = .25	0.263	0.256	0.259	-0.007	-0.004
C* = .50	0.083	0.081	0.082	-0.002	-0.001
Ed Tech reading 1					
C* = 0	0.464	0.442	0.453	-0.022	-0.011
C* = .25	0.256	0.250	0.253	-0.006	-0.003
C* = .50	0.109	0.108	0.108	-0.001	-0.001
Ed Tech reading 4					
C* = 0	0.502	0.464	0.483	-0.038	-0.019
C* = .25	0.224	0.218	0.221	-0.006	-0.003
C* = .50	0.060	0.058	0.059	-0.002	-0.001

Head Start PPVT					
C* = 0	0.296	0.288	0.292	-0.008	-0.004
C* = .25	0.408	0.395	0.402	-0.013	-0.006
C* = .50	0.155	0.152	0.154	-0.003	-0.001
Head Start WJ AP					
C* = 0	0.398	0.386	0.392	-0.012	-0.006
C* = .25	0.334	0.324	0.329	-0.010	-0.005
C* = .50	0.126	0.124	0.125	-0.002	-0.001
Head Start WJ LW					
C* = 0	0.347	0.341	0.344	-0.006	-0.003
C* = .25	0.450	0.433	0.441	-0.017	-0.009
C* = .50	0.190	0.187	0.188	-0.003	-0.002
Head Start WJ OC					
C* = 0	0.525	0.476	0.500	-0.049	-0.025
C* = .25	0.226	0.221	0.223	-0.005	-0.003
C* = .50	0.079	0.078	0.078	-0.001	-0.001
Head Start self regulation					
C* = 0	0.537	0.503	0.520	-0.034	-0.017
C* = .25	0.276	0.272	0.274	-0.004	-0.002
C* = .50	0.122	0.120	0.121	-0.002	-0.001
Head Start externalizing					
C* = 0	0.501	0.470	0.486	-0.031	-0.015
C* = .25	0.250	0.246	0.248	-0.004	-0.002
C* = .50	0.122	0.120	0.121	-0.002	-0.001
Average across all outcomes					
C* = 0	0.454	0.431	0.443	-0.022	-0.011
C* = .25	0.254	0.247	0.251	-0.007	-0.004
C* = .50	0.093	0.091	0.092	-0.002	-0.001

3. Estimating P(C) for Alternative values of ρ

To test the sensitivity of the risk estimates to the correlation between the estimates of V and I, we used the approximation with a .005 grid to estimate P(C) for correlations of -.5, 0, and +.5, for C* = 0, C=.25, and C* = .5, for all outcomes in the analysis. The results are shown in Exhibit B.2.

As can be seen in the last two columns of the last three rows of the exhibit, the average estimated values of P(C) for correlated I and V are virtually identical to the values for

uncorrelated I and V when C* = 0 or C* = .25, and differ by only about .04 standard deviations when C* = .5. We conclude that the estimates are quite insensitive to this correlation, and therefore, in the absence of any evidence or theory to suggest whether the correlation should be positive or negative, we use the more exact (because it allows calculation of P(C) by formula) and computationally efficient ρ = 0 to produce the results shown in the text and in Appendix C.

Exhibit B.2 P(C) for Correlated V and I

	Correlation				
	ρ = 0	**ρ = -.5**	**ρ = +.5**	**Diff (B) - (A)**	**Diff (C) - (A)**
	(A)	(B)	(C)	(D)	(E)
Charter year 1 math					
C* = 0	0.400	0.428	0.370	0.028	-0.030
C* = .25	0.110	0.111	0.108	0.001	-0.002
C* = .50	0.018	0.018	0.018	0.000	0.000
Charter year 2 math					
C* = 0	0.351	0.373	0.327	0.022	-0.023
C* = .25	0.214	0.223	0.204	0.009	-0.010
C* = .50	0.110	0.114	0.106	0.004	-0.004
Charter year 1 reading					
C* = 0	0.559	0.617	0.502	0.058	-0.057
C* = .25	0.141	0.143	0.139	0.002	-0.003
C* = .50	0.034	0.034	0.034	0.000	0.000
Charter year 2 reading					
C* = 0	0.414	0.437	0.389	0.024	-0.025
C* = .25	0.146	0.146	0.145	0.000	-0.001
C* = .50	0.030	0.030	0.030	0.000	0.000
Ed Tech math 6					
C* = 0	0.472	0.505	0.438	0.033	-0.035
C* = .25	0.322	0.333	0.309	0.011	-0.013
C* = .50	0.097	0.097	0.097	0.000	0.000
Ed Tech algebra					
C* = 0	0.455	0.504	0.400	0.049	-0.055
C* = .25	0.176	0.179	0.172	0.003	-0.004
C* = .50	0.050	0.050	0.050	0.000	0.000
Ed Tech towre					
C* = 0	0.518	0.555	0.484	0.037	-0.034
C* = .25	0.259	0.264	0.254	0.004	-0.005
C* = .50	0.082	0.082	0.082	0.000	0.000
Ed Tech reading 1					

C* = 0	0.453	0.491	0.414	0.038	-0.039
C* = .25	0.253	0.258	0.247	0.005	-0.006
C* = .50	0.108	0.108	0.108	0.000	0.000
Ed Tech reading 4					
C* = 0	0.483	0.552	0.415	0.069	-0.068
C* = .25	0.221	0.227	0.216	0.005	-0.005
C* = .50	0.059	0.059	0.059	0.000	0.000
Head Start PPVT					
C* = 0	0.292	0.298	0.285	0.006	-0.007
C* = .25	0.402	0.414	0.387	0.013	-0.014
C* = .50	0.154	0.154	0.153	0.000	0.000
Head Start WJ AP					
C* = 0	0.392	0.410	0.373	0.018	-0.019
C* = .25	0.329	0.344	0.313	0.015	-0.017
C* = .50	0.125	0.126	0.124	0.001	-0.001
Head Start WJ LW					
C* = 0	0.344	0.349	0.339	0.004	-0.006
C* = .25	0.441	0.473	0.409	0.032	-0.033
C* = .50	0.188	0.188	0.188	0.000	-0.001
Head Start WJ OC					
C* = 0	0.500	0.577	0.425	0.076	-0.075
C* = .25	0.223	0.224	0.222	0.001	-0.001
C* = .50	0.078	0.078	0.078	0.000	0.000
Head Start self regulation					
C* = 0	0.520	0.586	0.456	0.066	-0.064
C* = .25	0.274	0.277	0.270	0.003	-0.004
C* = .50	0.121	0.121	0.120	0.000	-0.001
Head Start externalizing					
C* = 0	0.486	0.542	0.427	0.057	-0.059
C* = .25	0.248	0.248	0.247	0.000	0.000
C* = .50	0.121	0.121	0.121	0.000	0.000
Average across all outcomes					
C* = 0	0.443	0.482	0.403	0.039	-0.040
C* = .25	0.251	0.258	0.243	0.007	-0.008
C* = .50	0.092	0.092	0.091	0.000	-0.001

4. Computer Code (in R) Used to Generate Risk Estimates When V and I Are Correlated

```
#this script adapts the bell-orr formula to settings with correlated
#impact estimates using a grid approximation of the bivariate
#normal density. it requires a data set with columns for outcome,
#model, site, within-site estimated impact, within-site estimated
#standard error, nonexperimental estimated impact, and
#nonexperimental estimated standard error.
#
#the main function 'bell.orr.corr' takes in 3 arguments:
#1) dat.subset subsets the data frame according to outcome and
#model
#2) rho is the correlation between estimates
#3) approx.crit is the number indicating the width of the grid
#intervals over which the bivariate normal density is evaluated.
#
#the main function also calls an outer function bv.norm that
#approximates the bivariate normal density
#
#the loop at the end of the script allows one to loop over
#all combinations of outcomes, models, correlations, and
#grid approximation interval widths of interest. if there are many
#combinations of interest, this task is best parallelized for
#efficiency.

#DEFINE INNER FUNCTIONS AND OBJECTS FOR LOOP

# function to evaluate bivariate normal pdf on grid
bv.norm <- function(x, y, cutoff, mu, sigma) {
  z <- cbind(x,y)
  cond.bv.norm <- ifelse((x < cutoff & y > cutoff) | (x > cutoff & y < cutoff),
                         dmvnorm(z,mean=mu,sigma=sigma),
                         0)
  return(cond.bv.norm)
}

# specify vector holding cutoff values
c <- seq(-4,4,by=.01)

#DEFINE FACTORS FOR LOOP
# outcome
outcomes <- unique(dat$outcome)
# model
models <- unique(dat$model)
# correlation
rhos <- 0
```

```
# approximation grid
approx.crits <- c(.01,.005)

#DEFINE MAIN FUNCTION
bell.orr.corr <- function(dat.subset,rho,approx.crit) {

  require(mvtnorm)
  require(plyr)
  require(ggplot2)

  x <- seq(-4, 4, by=approx.crit)
  y <- x

  rjc <- adply(dat.subset,1,function(df) {

    sapply(c, function(cutoff) {

      mu <- c(df$unbiased.impact,df$modeled.impact)

      sigma <- matrix(c(df$unbiased.se^2,df$unbiased.se*df$modeled.se*rho,
                        df$unbiased.se*df$modeled.se*rho,df$modeled.se^2),
                      nrow=2)


      # use outer function to evaluate pdf on 2D grid of x-y values
      fxy <- outer(x, y, bv.norm, cutoff, mu, sigma)

      return(sum(approx.crit^2*fxy))

    })

  })

rjc[c("X","outcome","site","unbiased.impact","unbiased.se","model","modeled.impact"
,"modeled.se")] <- NULL

  rc <- data.frame(c,apply(rjc,2,mean))

  colnames(rc) <- c("C","rc")

  rc$rc <- sapply(rc$rc,function(x) ifelse(x>1,1,x))

  c0 <- format(round(rc$rc[rc$C==0],3),nsmall=3)
  c25 <- format(round(rc$rc[rc$C==0.25],3),nsmall=3)
  c50 <- format(round(rc$rc[rc$C==0.5],3),nsmall=3)
  max.rc <- format(round(max(rc$rc),3),nsmall=3)
```

```
  rc.plot <- ggplot(rc, aes(x=C, y=rc)) +
    geom_line() +
    ylab("R(C*)") +
    xlab("C")

  return.list <- list(print(rho),

dat.subset$unbiased.impact,dat.subset$unbiased.se,dat.subset$modeled.impact,dat
.subset$modeled.se,
                      c0,c25,c50,max.rc,
                      rc.plot)
  names(return.list) <- c("rho",
                          "ij.impact","ij.se","ijx.impact","ijx.se",
                          "c0","c25","c50","max.rc",
                          "plot")
  return(return.list)

}

results <- list()
#LOOP
for (outcome in outcomes) {
  for (model in models) {
    for (rho in rhos) {
      for (criterion in approx.crits) {
        results[[paste(outcome,model,rho,criterion,sep="_")]] <-
          bell.orr.corr(dat.subset=subset(dat,outcome==outcome & model==model),
rho=rho,approx.crit=criterion)
      }
    }
  }
}
```

Appendix C. Detailed Estimates of P(C)

1. Estimates by Outcome and Method of Predicting Site-Specific Impacts, Alternative Values of C

To give the reader an overall sense of how P(C), the average risk of an incorrect policy decision across sites, varies with C and the prediction method, the results presented in text Exhibit 5 were averaged across all outcomes in each study. In this appendix, we present detailed estimates of P(C), by outcome, for three different values of C and five prediction methods. Results for each of the three studies are shown in a separate exhibit. The computer code used to generate these estimates is presented in the final section of this appendix.

Exhibit 1: Charter Schools: Risk of Wrong Policy Decision, for Alternative Outcomes, Prediction Methods, and Values of C

Policy Cut-off	Pooled Analysis	Subgroup Analysis	1-Moderator Model	2-Moderator Model	5-Moderator Model
Math, 6th Grade					
C* = 0	0.443	0.457	0.386	0.363	0.409
C* = .25	0.105	0.105	0.111	0.115	0.123
C* = .50	0.019	0.019	0.019	0.019	0.019
Maximum Risk	0.556	0.458	0.447	0.414	0.460
Math, 7th Grade					
C* = 0	0.507	0.469	0.295	0.249	0.265
C* = .25	0.225	0.271	0.225	0.180	0.184
C* = .50	0.097	0.097	0.119	0.128	0.115
Maximum Risk	0.558	0.479	0.307	0.278	0.280
Reading, 6th Grade					
C* = 0	0.493	0.570	0.606	0.637	0.556
C* = .25	0.129	0.129	0.142	0.149	0.172
C* = .50	0.033	0.033	0.034	0.034	0.037
Maximum Risk	0.552	0.605	0.644	0.660	0.578
Reading, 7th Grade					
C* = 0	0.370	0.389	0.430	0.447	0.470
C* = .25	0.134	0.134	0.139	0.154	0.179
C* = .50	0.028	0.028	0.028	0.030	0.036
Maximum Risk	0.551	0.391	0.636	0.559	0.530

Exhibit 2: Educational Technology: Risk of Wrong Policy Decision, for Alternative Outcomes, Prediction Methods, and Values of C

Policy Cut-off	Pooled Analysis	Subgroup Analysis	1-Moderator Model	2-Moderator Model	5-Moderator Model
Math, 6th Grade					
C* = 0	0.468	0.513	0.475	0.469	0.482
C* = .25	0.277	0.317	0.329	0.347	0.365
C* = .50	0.089	0.089	0.096	0.097	0.123
Maximum Risk	0.561	0.524	0.509	0.474	0.499
Algebra					
C* = 0	0.444	0.463	0.472	0.444	0.502
C* = .25	0.164	0.166	0.168	0.180	0.213
C* = .50	0.050	0.050	0.050	0.050	0.055
Maximum Risk	0.549	0.524	0.522	0.478	0.513
TOWRE					
C* = 0	0.497	0.602	0.569	0.461	0.514
C* = .25	0.247	0.250	0.249	0.259	0.308
C* = .50	0.082	0.082	0.082	0.082	0.089
Maximum Risk	0.545	0.635	0.601	0.470	0.524
Reading, 1st Grade					
C* = 0	0.515	0.494	0.434	0.424	0.455
C* = .25	0.237	0.250	0.241	0.255	0.296
C* = .50	0.105	0.105	0.105	0.106	0.126
Maximum Risk	0.547	0.499	0.471	0.432	0.463

Reading, 4th Grade					
C* = 0	0.493	0.603	0.488	0.459	0.467
C* = .25	0.219	0.219	0.220	0.231	0.233
C* = .50	0.060	0.060	0.060	0.060	0.060
Maximum Risk	0.535	0.641	0.496	0.459	0.468

Exhibit 3: Head Start: Risk of Wrong Policy Decision, for Alternative Outcomes, Prediction Methods, and Values of C

Policy Cut-off	Pooled Analysis	Subgroup Analysis	1-Moderator Model	2-Moderator Model	5-Moderator Model
Receptive vocabulary					
C* = 0	0.283	0.283	0.287	0.309	0.318
C* = .25	0.392	0.402	0.408	0.421	0.415
C* = .50	0.154	0.154	0.154	0.156	0.155
Maximum Risk	0.528	0.535	0.541	0.528	0.483
Early numeracy					
C* = 0	0.383	0.400	0.393	0.407	0.406
C* = .25	0.318	0.344	0.321	0.347	0.342
C* = .50	0.125	0.125	0.125	0.126	0.130
Maximum Risk	0.532	0.497	0.537	0.529	0.455
Early reading					
C* = 0	0.339	0.345	0.339	0.345	0.367
C* = .25	0.413	0.450	0.445	0.459	0.482
C* = .50	0.189	0.189	0.189	0.189	0.195
Maximum Risk	0.530	0.507	0.513	0.552	0.535
Oral comprehension					
C* = 0	0.508	0.528	0.548	0.527	0.514
C* = .25	0.223	0.223	0.223	0.223	0.236
C* = .50	0.079	0.079	0.079	0.079	0.081
Maximum Risk	0.527	0.531	0.566	0.543	0.515

Self-regulation					
C* = 0	0.527	0.520	0.588	0.517	0.532
C* = .25	0.273	0.273	0.277	0.276	0.282
C* = .50	0.122	0.122	0.122	0.122	0.122
Maximum Risk	0.533	0.521	0.590	0.519	0.533
Externalizing					
C* = 0	0.484	0.501	0.471	0.506	0.543
C* = .25	0.248	0.248	0.248	0.249	0.256
C* = .50	0.122	0.122	0.122	0.122	0.122
Maximum Risk	0.532	0.518	0.502	0.527	0.565

2. Computer Code (in R) Used to Generate Estimates of P(C)

```
#the function takes in 3 arguments:
#1) site.walk, which is a vector of the site ids;
#2) ij, which is a matrix of 2 columns (impact and standard error) and
#n rows corresponding to n sites, holds the within-site estimates
#3) ijx, of same size as ij, holds the nonexperimental estimates
#
#it returns a list of 9 objects; 4 vectors that are the impact estimates and
#standard errors for both methods; 4 numbers corresponding to the 4
#relevant values of R(C*); and a plot of R(C*).

c <- seq(-4,4,by=.01)
#this is a vector holding the 801 values of C

bell.orr <- function(site.walk,ij,ijx) {

  require(plyr)
  require(ggplot2)

  all.sites <- sapply(site.walk, function(site) {
    sapply(c, function(cutoff) {
      fj <- pnorm(cutoff,mean=ij[paste(site),1],sd=ij[paste(site),2])
      fjx <- pnorm(cutoff,mean=ijx[paste(site),1],sd=ijx[paste(site),2])
      rj <- (1-fj)*fjx + (1-fjx)*fj
      return(rj)
    })
  })
  rownames(all.sites) <- c

  #above, inside the two sapply statements:
  #1) the fj line evaluates the normal cdf for a given value of C and
  #for a given site from the matrix of within-site estimates of impacts
  #and standard errors;
  #2) the fjx line evaluates the normal cdf for a given value of C and
  #for a given site from the matrix of nonexperimental estimates of
  #impacts and standard errors
  #3) the rj line then evaluates the risk function at that given value
  #of C and for that given site
  #
  #the inner sapply statement is then applying this to each value of C.
  #it returns a column vector of length 801 which is R(C*) evaluated at
  #each value of C for a given site j.
  #
  #the outer sapply statement then follows by creating one of these
  #column vectors for each site. in the case of the PPVT outcome, we
```

```
  #have 73 sites, so this leaves a 801 x 73 matrix.

  rc <- adply(all.sites,1,mean)

  #this single command above then (following the PPVT example)
  #takes the 801 x 73 matrix of Rj(C) values and finds the mean for
  #each value of C across all sites. the way it's coded here, it results
  #in an 801 x 2 matrix, in which the first column is the value of C and
  #the second column is the corresponding value of R(C*).

  colnames(rc) <- c("C","rc")
  rc$C <- c

  c0 <- format(round(rc$rc[rc$C==0],3),nsmall=3)
  c25 <- format(round(rc$rc[rc$C==0.25],3),nsmall=3)
  c50 <- format(round(rc$rc[rc$C==0.5],3),nsmall=3)
  max.rc <- format(round(max(rc$rc),3),nsmall=3)
  #the first three of the above four commands evaluate the risk function
  #at C=0, 0.25, and 0.5, respectively. the fourth command finds the
  #maximum value of the risk function.

  rc.plot <- ggplot(rc, aes(x=C, y=rc)) +
    geom_line() +
    ylab("R(C*)") +
    xlab("C")

  #this plots the 801 values of C on the X axis and the corresponding 801
  #values of R(C*) that are the mean values of Rj(C) across all sites on the Y axis.

  return.list <- list(ij[,1],ij[,2],ijx[,1],ijx[,2],c0,c25,c50,max.rc,rc.plot)
  names(return.list) <- c("ij.impact","ij.se","ijx.impact","ijx.se","c0",
                          "c25","c50","max.rc","plot")
  return(return.list)
}

#to run, substitute vector of site ids, matrix containing unbiased impact
#estimates and standard errors for all sites, and matrix containing
#nonexperimental impact estimates and standard errors for all sites into
#site.walk, ij, and ijx arguments, respectively, of below function.

bell.orr(site.walk=,ij=,ijx=)
```