Using the Results from Rigorous Multisite Evaluations to Inform Local Policy Decisions

Larry L. Orr
Robert B. Olsen
Stephen H. Bell
Ian Schmid
Azim Shivji
Elizabeth A. Stuart

Abstract

Evidence-based policy at the local level requires predicting the impact of an intervention to inform whether it should be adopted. Increasingly, local policymakers have access to published research evaluating the effectiveness of policy interventions from national research clearinghouses that review and disseminate evidence from program evaluations. Through these evaluations, local policymakers have a wealth of evidence describing what works, but not necessarily where. Multisite evaluations may produce unbiased estimates of the average impact of an intervention in the study sample and still produce inaccurate predictions of the impact for localities outside the sample for two reasons: (1) the impact of the intervention may vary across localities, and (2) the evaluation estimate is subject to sampling error. Unfortunately, there is relatively little evidence on how much the impacts of policy interventions vary from one locality to another and almost no evidence on the implications of this variation for the accuracy with which the local impact of adopting an intervention can be predicted using findings from an evaluation in other localities. In this paper, we present a set of methods for quantifying the accuracy of the local predictions that can be obtained using the results of multisite randomized trials and for assessing the likelihood that prediction errors will lead to errors in local policy decisions. We demonstrate these methods using three evaluations of educational interventions, providing the first empirical evidence of the ability to use multisite evaluations to predict impacts in individual localities—i.e., the ability of "evidence-based policy" to improve local policy. © *2019 by the Association for Public Policy Analysis and Management.*

INTRODUCTION

Increasingly, local policymakers have access to published research evaluating the effectiveness of policy interventions. National research clearinghouses, such as the U.S. Department of Education's What Works Clearinghouse (WWC) and the U.S. Department of Justice's CrimeSolutions.gov, review and disseminate evidence from program evaluations that can help local policymakers decide whether to adopt an intervention. Some of these clearinghouses have especially promoted rigorous evidence obtained through randomized control trials (RCTs). These evaluations, which often encompass multiple localities, use random assignment to determine who receives an intervention, in order to ensure that impact estimates—derived as differences in outcomes between those who received the intervention and those who did not—do not suffer from treatment selection bias (Orr, 1999).

Journal of Policy Analysis and Management, Vol. 0, No. 0, 1–26 (2019)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.22154

Through these evaluations, local policymakers have a wealth of evidence describing *what* works, but not necessarily *where*. A local policymaker may have access to a report describing the results of a multisite RCT on an intervention the policymaker is considering, but how well would that report predict the potential consequences of adopting the intervention in the policymaker's own local jurisdiction? 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 localities for two reasons: (1) the impact of the intervention may vary across localities, or (2) the evaluation estimate, and therefore the predicted impact in the local site, is subject to sampling error. Unfortunately, there is relatively little evidence on how much the impacts of policy interventions vary and almost 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. And while the sampling variance of the prediction can be estimated, there is no agreed-upon method for taking it into account in local policy decisions. Finally, there is no reason to assume that multisite RCTs produce accurate predictions of the impact locally, in a single site. These studies are typically designed to produce accurate (i.e., with low bias and variance) estimates of the average impact across participating sites. Whether they also produce accurate impact estimates for individual sites is an open question.

This paper makes three main contributions. First, it highlights a potential challenge in making local 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, using several different prediction methods, from published reports of multisite RCTs and for assessing the likelihood that prediction errors will lead to errors in local policy decisions. We focus on multisite RCTs because of their high internal validity and the high visibility of their findings, making them more accessible to policymakers. The same methods could be applied to any multisite evaluation that includes both treated and control units within each site and are potentially useful for any intervention that can be adopted at the local level. Third, it demonstrates these methods using three evaluations of educational interventions, providing the first empirical evidence of the ability to use multisite RCTs to predict impacts in individual localities.

To measure the accuracy with which published evidence from multisite evaluations can predict local impacts, we develop and apply a "leave-one-out" analytic strategy that involves (1) assuming that one of the localities (henceforth, "sites") that participated in a multisite 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 decision about whether to adopt the intervention.

Applying these methods to data from three multisite RCTs in education, 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 operate a Head Start program in a particular locality. We make no claims that the results presented in this paper are broadly generalizable to other 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.

Our methods are in some ways similar to those employed in the "within-study comparison" literature on the internal validity of nonexperimental impact estimation methods (e.g., Cook, Shadish, & Wong, 2008; Fraker & Maynard, 1987; Glazerman, Levy, & Myers, 2003; LaLonde, 1986). In both cases, data from a randomized trial are used to create a benchmark against which to compare a nonexperimental estimate of the same parameter—in our case, prediction of within-site impacts on the basis of an RCT in other sites.

It must be noted that our methods produce lower-bound estimates of prediction error and the probability of an incorrect policy decision because we are unable to account for three possible sources of prediction error when using the results from a multisite impact study to inform policy decisions in a single site. First, our analysis does not account for nonrandom selection of sites into the study sample (Olsen et al., 2013). Nonrandom site selection can lead to systematic differences between the study sample and the population of sites that might benefit from the evidence (e.g., Stuart et al., 2017). However, our analysis strategy ignores this problem because we use sites that were included in the study to represent sites outside that study. Second, our analysis does not account for the fact that the impacts of interventions may change over time. Our analytic strategy is unable to capture these changes because it exploits data that were collected from multiple sites at roughly the same time. Third, it does not account for the fact that local policymakers may be focused on related but different outcome measures than the ones used in our analysis. However, we believe the analysis presented in this paper is useful because it accounts for (1) variation in impacts across sites that results from variation in program implementation, since sites in the three multisite RCTs that we reanalyzed had wide latitude to implement the interventions differently,[1] and (2) variation in impacts across sites due to differences in the characteristics of students and schools in those sites. As we shall see, even these lower-bound estimates are cause for concern in basing local policy on the results of multisite trials.

The next section describes the problem policymakers face when trying to use evidence from rigorous multisite 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 at the local level requires 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 predicted impacts and costs of alternative interventions, including the status quo.

[1] Two of the studies—studies of charter schools and Head Start—simply evaluated interventions that had already been put in place and were not modified in any way for purposes of the evaluation. One of the studies, of education technology, evaluated programs that were implemented specifically for the evaluation. But the evaluation reports (e.g., Campuzano et al., 2009; Dynarski et al., 2007) suggest that implementation was left to the developers and the schools in which the interventions were implemented, with no assistance from the study team. This suggests that the variation in implementation in the study sample should be similar to the variation in implementation that occurs when the interventions are implemented outside of the study.

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 policymakers.

Errors in Predicting the Impact of Adopting an Intervention Locally

Localities can conduct pilot tests to estimate the impact of adopting an intervention locally. But, most often, the evidence available to local policymakers comes from one or more evaluations conducted in other localities, and these policymakers need to extrapolate from this evidence to predict the impact in their 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 decisionmaker—the average impact that the intervention would have if it were adopted in the decisionmaker's locality. 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,[2] 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 decisionmaker's site), we would still expect the variance of the impact estimate to produce random 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 (since positive and negative errors are both squared in the calculation). This is the primary metric we 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 individuals or sites, such as, in the case of educational interventions, minority students or schools in urban settings. But it is not clear which estimate or estimates the policymaker

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

should use because it is not clear which estimates yield the smallest prediction errors for the policymaker's locality.

One option is to use the average impact reported for the entire evaluation 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 individuals who would receive the intervention if it 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. For example, 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 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 local population and environment to refine their predictions of the impact of adopting the intervention locally. Including multiple moderators in the model may reduce the bias for making local predictions but may also increase the variance through "overfitting" if the number of moderators included is too large. To use these models, policymakers would need to do some calculations themselves, combining the estimated coefficients from the model—if published—with local information about students and the environment in which the intervention would be implemented. This could be facilitated if evaluators developed interactive online impact models, but that would only be worthwhile if research like ours showed that such models substantially improved the accuracy of predictions for sites outside the evaluation.

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 developed interactive online impact models. This paper compares the errors that result from using each of these types of estimates to predict the local impacts of the intervention.

It should be noted that we test only prediction methods that could reasonably be expected to be available to local policymakers from published evaluation reports. We do not, for example, test methods that would require the locality considering adoption of the intervention to access the evaluation microdata or to apply advanced statistical modeling techniques. And we assume that local policymakers have no evidence from within their own sites as to the impact of the intervention—e.g., they have not conducted their own evaluation.

RELATED LITERATURE

To our knowledge, little attention has been paid to the problem of translating the findings of large-scale, multisite evaluations for use in local decisionmaking. 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. DiNardo and Lee (2011) discuss how evaluations can make out-of-sample predictions to address "ex ante evaluation questions" about the future effects of the program under different policy scenarios; they also note that any claim about a particular study's external validity is undefined without a precise definition of the target population.

Shadish, Cook, and Campbell (2002) distinguish between generalizing "from narrow to broad" and generalizing "from broad to narrow." A prototypical example of generalizing from narrow to broad might involve selecting a modest number of sites, conducting an experiment or quasi-experiment in those sites, and using the study estimates to make inferences about impacts in a larger population from which the sample was selected. In contrast, generalizing from broad to narrow could involve using impact evidence produced from a study conducted in multiple sites to predict the impact of adopting the intervention in a single site that was not part of the study sample.

The challenge of generalizing from multiple sites to a single site is heavily determined by the extent to which impacts vary across sites. 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 outcomes—Head Start (see Weiss et al., 2017). Similarly, Konstantopoulos (2011) found substantial variation in class size effects across schools. Significant variation in impacts has been found for a range of other programs outside of education, including welfare-to-work programs (Bloom, Hill, & Riccio, 2003) and energy conservation programs (Allcott, 2015).

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) formalize 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) present empirical evidence on the magnitude of this bias for one educational intervention. 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 obtain a representative sample.

This paper is one of the first to present empirical evidence on the challenge in going from broad to narrow. Hotz, Imbens, and Mortimer (2005) may have provided the first paper of this type: They tested the ability to predict the impacts of job training programs in a single locality using data from three other sites. However, the present paper may be the first to conduct a similar test for educational interventions and is almost surely the first to estimate the risk of making the wrong policy decision from the errors that can result when generalizing from broad to narrow.

Despite the paucity of rigorous research on the external validity of evaluation findings, as the idea of evidence-based policy has become more prominent a number of tools have emerged to help local policymakers identify interventions that evidence suggests will succeed in their communities. Clearinghouse websites inventory and rate the strength of evidence for interventions that have been evaluated (e.g., the What Works Clearinghouse [WWC] in education, the Arnold Foundation's Social Programs that Work, the Corporation for National and Community Service's Evidence Exchange, the National Institute of Justice's CrimeSolutions.gov site, the Department of Labor's Clearinghouse for Labor Evaluation and Research [CLEAR], the interagency Youth.gov site, and others). Almost all strength-of-evidence ratings are based on the internal validity of studies examining the subject interventions (e.g., evidence from well-conducted randomized trials receive a top rating, while findings from nonexperimental evaluations must emanate from well-matched treatment and comparison groups to receive a second-tier rating). Most of these sites pay very little attention to external validity. A notable exception is the WWC's "Find Research with Students Like Yours" feature, which allows a user interested in education interventions to specify (fairly general) characteristics of her or his own student body and then identifies studies with "similar" or "very similar" sample characteristics. Most clearinghouses simply offer very general guidance, such as, "... it is important to compare the populations studied with your target population and to look at whether the [referenced evaluation] provides information to address variations in the impact of the intervention across different individuals" (Selecting Evidence-Based Programs, n.d.). Most do describe the samples used in included studies and give subgroup results, to the extent they are available, and some require that the intervention be tested in multiple studies in different settings to receive a top rating (https://evidencebasedprograms.org/).

There is also a growing literature on implementation of evidence-based programs (e.g., Gorman-Smith, 2006; Horner, Blitz, & Ross, 2014; Janta, 2018; Lee et al., 2016a, 2016b; Metz, Bartley, & Maltry, 2017; Pew-MacArthur Results First Initiative, 2014; and Wiseman et al., 2007). But, again, these sources offer only very general guidance for translating results from broad-scope studies to local implementation. For example, one of the best (Gorman-Smith, 2006) notes, "[a] key thing to look for is...[whether] the rigorous evaluation tested the intervention in a population and setting similar to the one you wish to serve. The effectiveness of an intervention may vary greatly depending on the characteristics of the population (e.g., age, average income, educational attainment) and setting (e.g., neighborhood crime and unemployment rates) in which it is implemented."

It is our hope that our current and planned research, based on empirical analysis of the transportability of findings from rigorous impact studies to individual localities, can provide much more specific guidance to local policymakers. In this paper, we conduct an initial test of the proposition underlying much of the evidence-based policy literature—that adjusting for differences between the evaluation sample and the local policymaker's population of interest leads to acceptably accurate predictions of the impact of the tested intervention locally. In future work, we plan to investigate other methods of predicting local impacts.

DATA AND METHODS

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

Data

The data used in our analysis come from three multisite RCTs in education/child development: (1) the Impact Evaluation of Charter School Strategies (Gleason et al., 2010), (2) the Evaluation of the Effectiveness of Educational Technology Interventions (Campuzano et al., 2009; Dynarski et al., 2007), and (3) the Head Start Impact Study (Puma et al., 2010). 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 Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan. Below we briefly describe each of these studies:

- *The Impact Evaluation of Charter School Strategies* exploited charter school admission lotteries in 2005/2006 and 2006/2007 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 varied substantially across schools, and, in particular, that impacts were more favorable in schools that served more low-income and low-achieving students (Gleason et al., 2010).
- *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 2004/2005 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 (Campuzano et al., 2009; Dynarski et al., 2007).
- *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. 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 (Puma et al., 2010). Subsequent research has identified substantial heterogeneity in impacts across centers (Bloom & 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 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 ensures the study's internal validity. This allows us to focus on external validity. Third, the results from these studies are highly visible. Reports from these studies have been published on federal websites,[3] and the WWC has reported on their findings with the goal of informing policy decisions. Therefore, these studies may be prominent enough to be used by local policymakers.

Because this paper tests different methods for predicting impacts in a single site, we first 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.[4] 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.[5]

Empirical Strategy

We simulate the use of results from multisite randomized trials to predict impacts in a single site outside the evaluation sample. To assess the accuracy of local predictions based on multisite evaluation evidence, we apply the "leave-one-out" simulation methods used in machine learning (see Efron & Tibshirani, 1997; and Seni & Elder, 2010). In our case, these methods involve taking the actual data from a multisite evaluation that randomized students or classrooms within sites, assuming that one of the J 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 (and then repeating the last two steps for each site in the evaluation).

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

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.
3. Assume that one of the J sites in the evaluation was excluded from the sample.

[3] For reports from the charter school and educational technology studies, see https://ies.ed.gov/ncee/pubs/. For reports from the Head Start study, see https://www.acf.hhs.gov/opre/research/project/head-start-impact-study-and-follow-up.

[4] 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).

[5] 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.

4. Calculate the *estimated impact for the excluded site* by exploiting the experiment conducted within 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 5.
5. Calculate the *predicted impact for the excluded site* by applying the statistical method from step 2 to the data from the other J-1 sites. This prediction may contain both bias and sampling error, as described earlier.
6. Estimate the prediction error by taking the difference between the predicted impact for the excluded site (from step 5) and the estimated impact for the excluded site (from step 4).
7. Repeat steps 3 through 6 for each of the remaining J-1 sites to estimate 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 for 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 fail to adopt the intervention when it should be adopted—using the method described below.
10. Repeat steps 2 through 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:

- **Use the average, pooled impact estimate for sites in the study sample.** This impact estimate is usually the main finding from an impact analysis.
- **Use the impact estimate 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 can be used to predict the impact of the intervention for this site.
- **Use a predicted estimate from an equation that models the variation in impacts across sites as a function of multiple site-level variables.** Some impact analyses use "response surface modeling" (Box & 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, percent low-income families, and baseline levels of the outcome). The estimated regression model is then used to predict the impact in the excluded site, based on that site's characteristics.

The first two of these prediction methods are nearly always available to local policymakers from published evaluation reports. The third is generally not but could be if research showed it to be a superior prediction method. More sophisticated prediction methods exist but are generally beyond the capability of local policymakers to apply using published study results. In this analysis, we focus our attention on methods that could reasonably be expected to be available to local policymakers. We now present more details on these methods.

Regression Models

This section describes these methods in more detail, especially the regression models used.[6] To estimate the impact δ^w within the excluded site, we estimated the following regression model using data from that site:

$$\begin{aligned} y_i &= \alpha + X_i'\beta + \delta^w T_i + e_i \\ e_i &\sim N\left(0, \sigma_e^2\right) \end{aligned} \tag{1}$$

where:

- y_i is the outcome for student i.
- X_i' is a vector of student-level covariates included to improve the precision of the estimates.
- T_i is the treatment indicator, which equals 1 if student i was assigned to the treatment group and 0 if this student was assigned to the control group.
- e_i is a random error term.

The estimate of δ^w, which we designate $\hat{\delta}^w$, was used as our benchmark for the impact in the excluded site.

The first prediction method that we examined is simply the average, pooled impact across sites included in the evaluation. For this, we estimated the following two-level regression model of students nested within schools[7] using data from all sites except the excluded site (j); the model is indexed by j to reflect the fact that it is run repeatedly, once for each j:

$$\begin{aligned} y_{ij} &= \alpha_j + X_{ij}'\beta + \delta_j^p T_{ij} + e_{ij} \\ \alpha_j &= \alpha + u_j \\ \delta_j &= \delta + v_j, \end{aligned} \tag{2}$$

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

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 J-1 omitting the excluded site, and J 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.

[6] 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) through (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.

[7] For the education technology study, a three-level model of students, teachers, and schools would be ideal since the study randomly assigned teachers, not students. However, we instead used a two-level model of students within schools out of concerns that the three-level model may not converge with only three to four teachers per school, and because the limitation of ignoring the teacher level—downwardly biased standard error estimates—does not affect our analysis because we do not use the estimated standard errors in measuring the accuracy of local predictions made with the regression models.

The estimate of the population parameter δ_j^p is used for the prediction of the impact in the excluded site (j); we denote the estimate as $\hat{\delta}_j^p$.

The other prediction methods involve enhanced versions of equation (2) where one or more site-level variables are interacted with the treatment indicator, and the impact in site j is predicted by inserting the values of these variables for site j into the estimated regression model. In selecting these variables, we used a three-step process. First, we relied on the original study authors' expertise to select the initial pool of potential site-level moderators—presumably based on some combination of theory and evidence. Second, we excluded variables that could not be used to make *ex ante* local predictions because they could not be known prior to implementing the intervention. Third, we analyzed the data from the study to identify the most important moderators from among the remaining variables. For more details on these models and how the moderator variables were selected, see Appendix A.[8]

Table 1 shows the site-level subgroup or moderator variables that were chosen for inclusion for each study; it also identifies in footnotes the site-level variables that were most commonly selected for each model. A complication reflected in this table is that the model selection strategy was implemented separately for each excluded site, or, put differently, for each sample of J-1 sites that we treat as having been included in the evaluation. Therefore, a different subgroup variable or set of potential moderators could be selected for each of these samples. For each of the subgroup or moderator models, Table 1 lists the variables that were most frequently selected via the protocol described above.

Measuring the Magnitude of the Errors in Predicting Local Impacts

We use the Root Mean Squared Prediction Error (RMSPE) to capture the "typical" magnitude of the errors when predicting the impact for individual sites using evidence from a multisite impact evaluation. The prediction error is just the predicted impact in a site minus the true impact in that site: $(\hat{\delta}_j^p - \delta_j^w)$, where $\hat{\delta}_j^p$ is the predicted impact in site j, and δ_j^w is the actual impact within site j. Squaring the prediction error converts all errors to positive values, so that negative errors do not offset positive errors. Taking the simple average of the squared prediction errors across sites yields the Mean Squared Prediction Error. To convert the measure back to more intuitive units, we take the square root to produce the RMSPE.

If the true impact in each site were known (δ_j^w for site j), we could use the standard formula for the RMSPE:$\sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\delta}_j^p - \delta_j^w)^2}$, where J is the number of sites. Since the true impact is not known for any site, we instead use an unbiased estimate of the impact within site j, $\hat{\delta}_j^w$, based on the data from that site: $\sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\delta}_j^p - \hat{\delta}_j^w)^2}$.

[8] For the charter school and education technology studies, we also examined student-level moderators (e.g., student demographics and prior achievement, as well as disability status and English proficiency for the charter school study). These variables could be used by local policymakers to predict average site-level impacts when they have access to data on the characteristics of the students who are likely to receive the intervention if the policymaker adopts it locally (e.g., data on all students in selected schools that would adopt a schoolwide intervention). The results from this investigation (not presented in this paper) suggest that student-level moderators would not improve the accuracy of local predictions—perhaps because they explain relatively little variation in impacts across sites that was not explained by the site-level moderators, many of which were aggregates of student-level moderators (e.g., percent minority). Therefore, for simplicity, we present the results from analyses that include only site-level moderators. All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

Table 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	

Table 1. Continued.

Moderator	Charter Schools[a]	Education Technology[b]	Head Start[c]
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.

Replacing the true impact in site j with the estimated impact inflates the RMSPE because the estimated impact ($\hat{\delta}_j^w$) is based on a finite sample, so the estimate contains sampling error. To correct for this bias, we subtract off the average sampling variance of these estimates across sites ($\frac{1}{J}\sum_{j=1}^{J}\hat{\sigma}_{\epsilon j}^2$):

$$\widetilde{RMSPE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\hat{\delta}_j^p - \hat{\delta}_j^w\right)^2 - \frac{1}{J}\sum_{j=1}^{J}\hat{\sigma}_{\epsilon j}^2}. \tag{3}$$

Appendix B[9] derives this equation and shows that it provides a conservative estimate of the RMSPE.

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; i.e., the RMSPE is subject to sampling error. To test for significant differences between methods, we conducted a binomial or sign test. For each pair of methods, we tested the null hypothesis that the true prediction error is the same for both methods in every site. If this were true, we would expect each method to outperform the other—that is, produce a smaller estimated squared prediction error than the other method—in exactly half of the sites. Assuming independence across sites,[10] we used the binomial distribution to calcu-

[9] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

[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

late the probability that one of the two methods outperforms the other by chance for as many or more of the sites as was actually observed, and we rejected the null hypothesis if this probability was less than 5 percent.

Policy Consequences of the Errors in Predicting Local Impacts

The policy consequences of errors in predicting local impacts depend on how local policymakers use the evidence to inform policy decisions. We consider a stylized approach to making local evidence-based decisions in which the intervention is adopted locally if and only if the policymaker believes the impact exceeds some pre-specified threshold—call it C*. For simplicity, we assume that the policymaker's beliefs about the impact of the intervention are based entirely on the evidence of its impact on a single outcome from a single study. In practice, those beliefs may be influenced by evidence for multiple outcomes or from multiple studies—and by factors that are unrelated to evidence from any study (e.g., educational philosophy, political considerations). The simplifying assumptions made for this analysis allow us to take an initial step in translating prediction errors for individual sites into consequences for policy. Finally, our approach takes C* as given and set by the policymaker. For example, a policymaker may set C* equal to the minimum impact that would ensure the intervention is cost-effective relative to alternative interventions or to equal the minimum impact that policymakers would judge to be practically significant.

The methods used to assess the risk of making an incorrect policy decision are adapted from Bell and Orr (1995), who developed a Bayesian method to assess the risk of making an incorrect policy decision on the basis of a potentially biased, nonexperimental estimator. We apply that method here to assess the predicted impact estimates obtained from the prediction methods described earlier. 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 posterior probability distribution for that parameter based on data from a sample should be centered on the parameter estimate produced by the sample. In addition, if the sample estimate has a normal distribution, the posterior distribution of possible parameter values also follows a normal distribution, with standard deviation equal to the standard error of the parameter estimate (DeGroot, 1970, pp. 190–191).

Thus, starting with an agnostic view of the true impact in site j, δ_j^w, 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. Figure 1 illustrates how such a distribution might look, calibrated in effect size units (i.e., as impact divided by the standard deviation of the outcome measure). Similarly, a posterior distribution of the expected value of the predicted impact, δ_j^p, can be derived from data on the other J-1 sites in the evaluation (referred to as the "predicted impact" above). Figure 1 contains an illustrative posterior distribution for this parameter, also measured in effect size units. Together, these two distributions—the posterior

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.

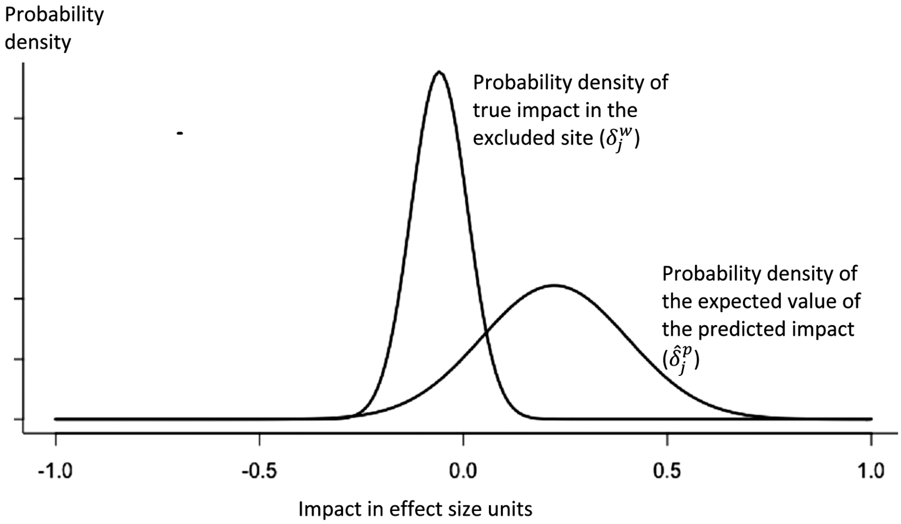

Figure 1. Bayesian Posterior Distributions for the Excluded Site Under Agnostic Prior.

distributions for the true impact and the predicted impact—provide a basis for assessing the policy reliability of the prediction method.

As noted above, we assume that the local policymaker has some policy cut-point C*. If the predicted impact of the policy exceeds C*, he or she will adopt the policy; if the predicted impact is less than C*, he or she will not. Under this decision rule, the risk, R(C*), that the predicted impact will lead to the wrong decision is:

$$\mathrm{R_j}\,(\mathrm{C}^*) = \Pr(\delta_j^p < \mathrm{C}^*\text{and } \delta_j^w > \mathrm{C}^*) + \Pr(\delta_j^p > \mathrm{C}^* \quad \text{and} \quad \delta_j^w < \mathrm{C}^*), \tag{4}$$

where $\Pr(\delta_j^p < \mathrm{C}^*$ and $\delta_j^w > \mathrm{C}^*)$ is the probability that the predicted impact suggests the program would be ineffective in site j ($\delta_j^p < \mathrm{C}^*$) when it would actually be effective in that site ($\delta_j^w > \mathrm{C}^*$), and $\Pr(\delta_j^p > \mathrm{C}^*$ and $\delta_j^w < \mathrm{C}^*)$ is the probability that the predicted impact suggests the program would be effective in site $j(\delta_j^p > \mathrm{C}^*)$ when it would actually be ineffective in that site ($\delta_j^w < \mathrm{C}^*$).

Bell and Orr call equation (4), traced out over a range of values for C*, the "risk function." In the special case of zero correlation between δ_j^p and δ_j^w, these two random variables are independent, and the risk formula reduces to:

$$\mathrm{R_j}(\mathrm{C}) = \Pr\left(\delta_j^p < \mathrm{C}\right) \cdot \Pr\left(\delta_j^w > \mathrm{C}\right) + \Pr\left(\delta_j^p > \mathrm{C}\right) \cdot \Pr\left(\delta_j^w < \mathrm{C}\right). \tag{5}$$

Unfortunately, there is no exact analytic expression for $\mathrm{R_j(C)}$ when δ_j^p and δ_j^w are correlated. However, we were able to develop a very accurate approximation to $\mathrm{R_j(C)}$ for correlated δ_j^p and δ_j^w and found that our estimates of the risks of incorrect policy conclusions were insensitive to correlations as large as +.9 or as small as −.9 (see Appendix C[11]). Therefore, absent any evidence or theory to indicate whether the

[11] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

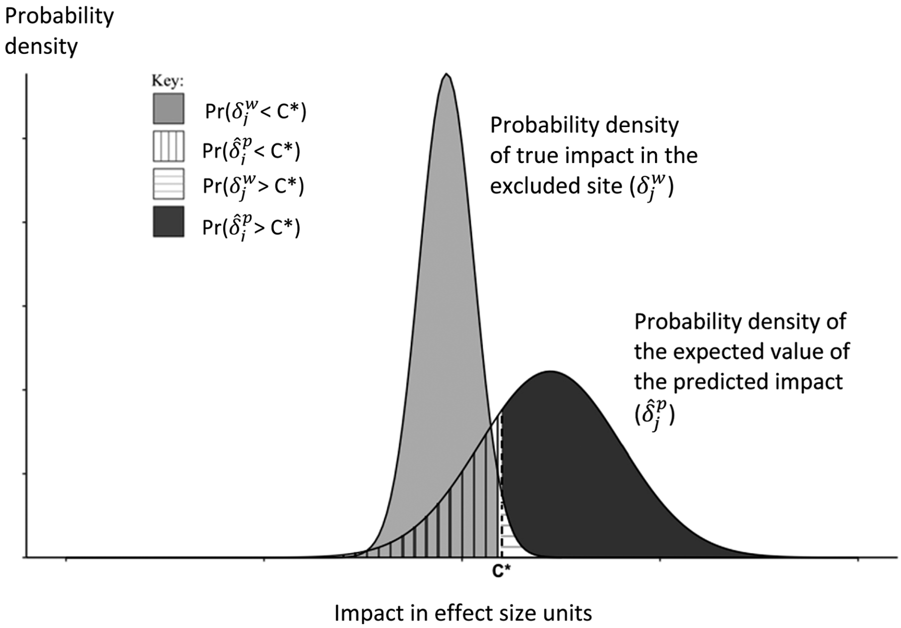

Figure 2. Components of the Risk Function for Illustrative Cutoff C*.

correlation would be positive or negative, we use the probability given by equation (5), which assumes that δ_j^p and δ_j^w are uncorrelated, as our estimate of $\mathrm{R_j(C)}$.

The four components of equation (5), for a given value of C (C*), are depicted in Figure 2, using shadings of different areas underneath the two density curves. The area under the probability density curve of δ_j^w to the left of C* is Pr (δ_j^w < C*); the area under that curve to the right of C* is Pr (δ_j^w > C*). Similarly, the area under the probability density of the expected value of the predicted impact to the left of C* is Pr (δ_j^p < C*); the area under that curve to the right of C* is Pr (δ_j^p > C*). Given the within-site impact estimate and the predicted impact in site *j*, along with their standard errors, these areas can be computed and plugged into equation (5) to compute $\mathrm{R_j(C)}$ for any value of C. Figure 3 shows the probability of an incorrect policy decision calculated from this equation (the dashed line $\mathrm{R_j(C)}$) and the two Bayesian distributions from which it is derived, for all possible cutoff values 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 multisite evaluations, using the predicted impact in that site based on data in the other sites in that study and the within-site experimental estimate (our estimate of the true impact). For each possible policy cut-point C that a policymaker might adopt, we sum the calculated risks across all sites for a given evaluation and divide by the number of sites; this yields the expected share S(C) of sites that would make the wrong policy decision based on the predicted impact in their sites if C is the policy support cut-point:

$$S(C) = \frac{\sum_{j=1}^{J} \boldsymbol{R_j}\,(\boldsymbol{C})}{\boldsymbol{J}}. \tag{6}$$

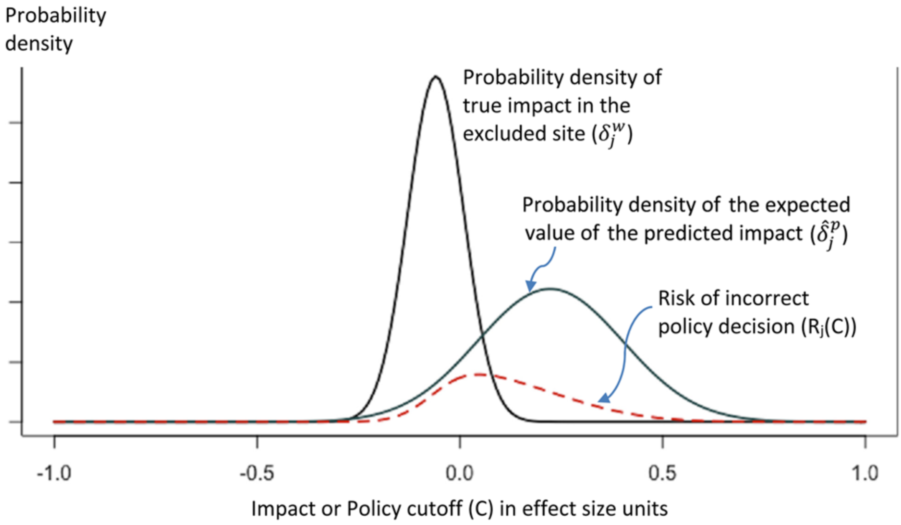

Figure 3. Risk Function for Making an Incorrect Policy Decision, $R_j(C)$. [Color figure can be viewed at wileyonlinelibrary.com]

If S(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 estimates of the accuracy of different prediction methods and the risk of making a wrong policy decision based on those predictions.

Accuracy of the Predictions

Table 2 shows the estimates of the RMSPE in standard deviations of the outcome variable—that is, in effect size units. Estimates are reported separately for each study, outcome variable, and prediction method. Superscripts to the estimates indicate whether one method produced smaller (squared) prediction errors than another method, using the binomial test described earlier.

Focusing first on the pooled analysis, the estimates vary from about 0.06 to 0.35, with somewhat larger estimates for charter schools and education technology than for Head Start. To give the reader a sense of the scale of these estimates, Table 2 also includes the published impact estimate for each of these outcomes (see the final column). A comparison of the estimated RMSPEs with the published impact estimates suggests that the local prediction errors may be large relative to the average impacts of these educational interventions.[12]

The next question is whether the other prediction methods that take advantage of data on treatment effect moderators—subgroup analysis and models with one, two, and five treatment effect moderators—yield more accurate local predictions of impact. The evidence presented in Table 2 suggests that they do not. In fact, the

[12] 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.

Table 2. 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*

Notes:

[p] Significantly different from the pooled analysis at the 5 percent level using a binomial or sign test, as described earlier.

[s] Significantly different from the subgroup analysis at the 5 percent level using a binomial or sign test, as described earlier.

[1] Significantly different from the 1-moderator model at the 5 percent level using a binomial or sign test, as described earlier.

[2] Significantly different from the 2-moderator model at the 5 percent level using a binomial or sign test, as described earlier.

[5] Significantly different from the 5-moderator model at the 5 percent level using a binomial or sign test, as described earlier.

[*] Published impact estimates for charter schools and education technology come from Gleason et al. (2010) and Campuzano et al. (2009), respectively. Published impact estimates for the Head Start Impact Study (HSIS) come from Bloom and Weiland (2015), which pooled the 3-year-old and 4-year-old cohorts from Puma et al. (2010) in their reanalysis of HSIS data.

pooled analysis yields a smaller estimated RMSPE than each of the other methods 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. Moreover, *all nine of the significant differences between the pooled analysis and more complex models favored the pooled analysis*.

Risk of Making the Wrong Policy Decision

In Table 3, 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 cutoff C is set to 0 standard deviations, .25 standard deviations, and .50

Table 3. Average probability of incorrect policy decision across all outcomes, by study and method of predicting site-specific impacts, for alternative values of C.

Policy Cutoff	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%
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%
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%
Avg. RMSPE	*0.15*	*0.15*	*0.16*	*0.17*	*0.10*

standard deviations. The table also indicates the average RMSPE from Table 2 for each study. (For outcome-specific risk estimates and the computer code used to generate these estimates, see Appendix D.[13])

The three different cutoffs—0, 0.25, and 0.50 standard deviations—were selected to represent different places where the local policymaker could "set the bar" when considering whether the intervention is effective enough to adopt. The policymaker may set the bar at zero if the treatment intervention could be implemented at no cost (e.g., by changing a regulation or curriculum), or if the policymaker is simply indifferent about whether to implement the intervention absent evidence on its impacts. Alternatively, the policymaker may set the bar relatively high, like 0.25 standard deviations, if the costs of adopting the intervention are high or the policymaker is just disinclined to adopt it. We think it is unlikely that a local education policymaker would set the threshold as high as 0.50 standard deviations, the third threshold that we test. A threshold of 0.50 standard deviations implies that a policymaker would choose *not* to adopt an intervention that he or she knew would increase student achievement by 0.49 standard deviations—which is large relative to the impacts of most educational interventions and to typical year-to-year achievement gains by students on broad tests of math and reading achievement (Hill et al., 2008). However, we consider a threshold 0.50 as an extreme case to see whether policymakers who set the cutoff extremely high benefit from evidence likely to show that the true effect falls below that cutoff.

The risk estimates in Table 3 follow several systematic patterns. First, consistent with the findings for prediction errors, for a given study and policy cutoff, there is little variation in the average probability of an incorrect policy decision across prediction methods. Second, that probability tends to fall as the policy cutoff increases: Policy errors become less likely as the impact required for policy

[13] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

approval increases. As shown in the table, across studies and prediction methods, the average risk of making the wrong policy decision ranges from 42 to 54 percent when the policy cutoff is zero. When the cutoff is .25, this range falls to 15 to 32 percent. And when the cutoff is .50, the share of sites expected to make an incorrect policy decision ranges from 4 to 13 percent across the studies and prediction models.

DISCUSSION AND IMPLICATIONS

As noted at the outset, current tools and guidance for identifying and implementing evidence-based interventions (i.e., clearinghouses and implementation guides) offer little more guidance (and in a few instances, help) with respect to the external validity of evaluation results than the suggestion that local policymakers find research conducted with samples and settings similar to their own intended clientele. The results presented here call that strategy into question, at least for education interventions. Simply adjusting evaluation results for differences in school and student characteristics between the evaluation sample and the local population of interest did little to improve the predictions of impacts in local school districts.

The empirical results presented in this paper suggest—at least for the three education interventions examined and the five different prediction approaches tested in this paper—that most localities will not be able to use the results of large-scale multisite evaluations to accurately predict the likely consequences of adopting an intervention or policy. It is perhaps not surprising that pooled analysis, which ignores variation in impacts across sites—produces large prediction errors for individual sites. However, other methods performed no better. The prediction errors for methods that modeled cross-site impact variation using site-level data, including site-level aggregates of individual-level data (e.g., percent minority and percent proficient in math and reading), were as large or larger than the prediction errors for the pooled analysis. These findings reveal a serious challenge when using multisite impact evaluations, which are typically designed to estimate the average impact of the program, to predict the impacts in individual sites—a challenge which may extend beyond education evaluations.

Furthermore, our estimates almost surely understate the size of the typical prediction error because extrapolating to sites that actually participated in the evaluation, as we did in the analysis, will surely yield smaller prediction errors than extrapolating to sites that did *not* participate and may differ substantially from those that did. Out-of-sample prediction errors for sites 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 sites 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, reinforcing our finding of relatively large prediction errors. Ongoing work is examining how the results differ when the participating sites systematically differ from the target site.

A unique contribution of this analysis is the ability to estimate the probability of making the wrong local policy decision based on the results from a multisite evaluation. This analysis assumes a decision rule under which the policymaker adopts the intervention locally if he or she believes the impact of the intervention will exceed some threshold, and incorrect policy decisions result from inaccurate local impact predictions—in particular, predictions that fall on the opposite side of this threshold from the true impact. Our findings indicate that the probability of an incorrect local policy decision depends strongly on the size of impact the policymaker requires to justify adopting the intervention. In these three evaluations, if any positive impact would justify adoption (i.e., if the policy cut-point is zero), a national evaluation would be of little help to the local policymaker. In this case, the chance of making the right policy decision at the local level based on evidence from a large, multisite evaluation is at best only slightly better than 50 percent—the rate we would expect if policymakers just flipped a coin.

On the other hand, if a large (.50 or larger) effect size is required for adoption, the national evaluation substantially reduces the risk of making the wrong policy decision from 50 percent, assuming that the policymaker has no prior knowledge of the impact of the intervention, to 4 to 13 percent. For intermediate policy cutoffs, the risk is still substantial—in the range of 15 to 32 percent—but much lower than it would have been in the absence of the evaluation (50 percent). These findings suggest that the cutoff needs to be high for the evidence to substantially improve local policy decisions. If the threshold is high relative to the likely impact of the intervention, the main value of the evidence is to persuade local policymakers to forgo interventions that are probably not effective enough to warrant adopting.

We also found that 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 multisite evaluations are almost never powered for such analyses and thus the subgroup and moderator model effect estimates have large standard errors. 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, of 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 multisite RCTs of other 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] In addition, while a benefit of using actual multisite trials for this leave-one-out exercise is that the data fully reflect real world settings, a drawback is that because we observe only a noisy estimate of each site's true impact, our results cannot separate bias from variance. More research is needed on both the magnitude of cross-site impact variation for policy interventions and the accuracy with which we can predict site-level impacts from multisite impact evaluations. Ongoing work is conducting these leave-one-out exercises with simulated outcomes, which allows us to separate bias from variance, and to specify the cross-site variation in impacts explained by observed and unobserved factors. These simulations will also allow us to consider questions such as the trade-offs in terms of the number of sites and number of individuals per site, and how sampling sites on the basis of impacts may affect the conclusions regarding the ability to predict site-specific impacts.

It is important to recognize that large multisite studies may and often do serve purposes other than informing local policy decisions. Many multisite 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 large multisite studies to produce more accurate evidence to guide policy. However, even when used for that purpose, large multisite studies may provide misleading evidence if

[14] See Weiss et al. (2017) for new evidence on the cross-site variance of impacts for 13 educational interventions.

impacts vary across sites and sites are selected non-randomly (see Allcott, 2015; and Bell et al., 2016).

While these results suggest caution in extrapolating the results of national evaluations to local jurisdictions, our objective here is not to reach definitive conclusions about the usefulness of multisite 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 such 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, lead 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.

LARRY L. ORR is an Associate in the Department of Health Policy and Management at the Johns Hopkins Bloomberg School of Public Health, 4402 Leland Street, Chevy Chase, MD 20815 (e-mail: lorr5@jhu.edu).

ROBERT B. OLSEN is Associate Director and Lead for Impact Evaluation Science at Westat, 1600 Research Boulevard, Rockville, MD 20850 (e-mail: robolsen@westat.com).

STEPHEN H. BELL is Head of Evaluation Research at Westat, 1600 Research Boulevard, Rockville, MD 20850 (e-mail: stephenbell@westat.com).

IAN SCHMID is a Research Associate at the Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, Room 810, Baltimore, MD 21205 (e-mail: ian_schmid@jhu.edu).

AZIM SHIVJI is a Senior Analyst at Abt Associates, Inc., 6130 Executive Boulevard, Rockville, MD 20852 (e-mail: azim_shivji@abtassoc.com).

ELIZABETH A. STUART is a Professor in the Departments of Mental Health, Biostatistics, and Health Policy and Management at the Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, W1033, Baltimore, MD 21205 (e-mail: estuart@jhu.edu).

REFERENCES

Allcott, H. (2015). Site selection bias in program evaluation. The Quarterly Journal of Economics, 130(3), 1117–1165.

Bareinboim, E., & Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. Journal of Causal Inference, 1, 107–134.

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, 38, 318–335.

Bell, S. H., & Orr, L. L. (1995). Are nonexperimental estimates close enough for policy purposes? A test for selection bias using experimental data. Proceedings of the American Statistical Association, Section on Government Statistics, 228–233.

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014). Experimental evidence on distributional effects of Head Start. NBER Working Paper 20434. Cambridge, MA: National Bureau of Economic Research.

Bloom, H. S., Carolyn, J. H., & James, A. R. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of Welfare-to-work experiments. Journal of Policy Analysis and Management, 22, 551–575.

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. New York, NY: MDRC.

Box, G. E., & Draper, N. R. (1987). Empirical model-building and response surfaces. New York, NY: Wiley.

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). Effectiveness of reading and mathematics software products: Findings from two student cohorts. NCEE 2009–4041. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management, 27, 724–750.

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

DiNardo, J., & David, S. L. (2011). Chapter 5: Program Evaluation and Research Designs. Handbook of Labor Economics, 4, Part A, 2011, 463–536.

Ding, P., Feller, A., & Miratrix, L. (2019). Decomposing treatment effect variation. Journal of the American Statistical Association, 114, 304–317.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., . . . Emery, D. (2007). Effectiveness of reading and mathematics software products: Findings from the first student cohort. NCEE 2007–4005. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The .632 + Bootstrap Method. Journal of the American Statistical Association, 92, 548–560.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. Journal of Human Resources, 22, 194–227.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. The Annals of the American Academy of Social and Political Science, 589, 63–93.

Gleason, P., Clark, M., Tuttle, C. C., & Dwoyer, E. (2010). The evaluation of charter school impacts: Final report. NCEE 2010–4029. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gorman-Smith, D. (2006). How to successfully implement evidence-based social programs: A brief overview for policymakers and program providers. Washington, DC: Coalition for Evidence-Based Policy.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. Child Development Perspectives, 2, 172–177.

Horner, R. H., Blitz, C., & Ross, S. W. (2014). The importance of contextual fit when implementing evidence-based interventions. Washington, DC: Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services.

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, 241–270.

Janta, B. (2018). Implementing evidence-based practices effectively: A practical guide. Luxembourg: Publications Office of the European Union.

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, 103–127.

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

LaLonde, R. (1986). Evaluating the econometric evaluations of training with experimental data. American Economic Review, 76, 604–620.

Lee, L., Hughes, J., Smith, K., & Foorman, B. (2016a). An SEA guide for identifying evidence-based interventions for school improvement. Tallahassee, FL: Florida Center for Reading Research.

Lee, L., Hughes, J., Smith, K., & Foorman, B. (2016b). An LEA or school guide for identifying evidence-based interventions for school improvement. Tallahassee, FL: Florida Center for Reading Research.

McCoy, D. C., Morris, P. A., Connors, M. C., Gomez, C. J., & Yoshikawa, H. (2016). Differential effectiveness of Head Start in urban and rural communities. Journal of Applied Developmental Psychology, 43, 29–42.

Metz, A., Bartley, L., & Maltry, M. (2017). An implementation science and service provider-informed blueprint for integration of evidence-based/evidence-informed practices into New Jersey's child welfare system. Chapel Hill, NC: The National Implementation Research Network.

Olsen, R. B., & Orr, L. L. (2016). On the "where" of social experiments: Selecting more representative samples to inform policy. New Directions for Evaluation, 152, 61–71.

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, 107–121.

Orr, L. L. (1999). Social experiments: Evaluating public programs with experimental methods. Thousand Oaks, CA: Sage Publications.

Pew-MacArthur Results First Initiative. (2014). Evidence-based policymaking: A guide for effective government. Washington, DC: The Pew Charitable Trusts.

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

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

Selecting Evidence-Based Programs. (n.d.). Retrieved April 2, 2019, from https://youth.gov/evidence-innovation/selecting-programs.

Seni, G., & Elder, J. (2010). Ensemble methods in data mining: Improving accuracy through combining predictions. San Rafael, CA: Morgan and Claypool.

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of school districts that participate in rigorous national educational evaluations. Journal of Research on Educational Effectiveness, 10, 168–206.

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, 369–386.

Tipton, E. (2013). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. Evaluation Review, 37, 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, 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, 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, 778–808.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multi-site randomized trials. Journal of Research on Educational Effectiveness, 10, 843–876.

Wiseman, S. H., Chinman, M., Ebener, P. A., Hunter, S. B., Imm, P., & Wandersman, A. (2007). Getting To Outcomes^{TM}: 10 steps for achieving results-based accountability. Santa Monica, CA: Rand Corporation.

APPENDIX A: PREDICTION METHODS THAT INVOLVE SITE-LEVEL TREATMENT EFFECT MODERATORS

This appendix describes the methods used to specify regression models and predict impacts using the methods involving site-level variables that potentially moderate the impact of the intervention.

As described in the text, the first predictor is simply the mean effect of the intervention in the evaluation sample, other than the excluded site, based on equation (2). The second prediction method examined involves estimating impacts for different subgroups of sites. 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:

$$\begin{aligned} y_{ij} &= \alpha_j + X'_{ij}\beta + \delta^w_j T_{ij} + e_{ij}, \\ \alpha_j &= \alpha + \gamma S_j + u_j \\ \delta_j &= \delta + \theta S_j + v_j, \\ \begin{pmatrix} e_{ij} \\ u_j \\ v_j \end{pmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_e & 0 & 0 \\ 0 & \sigma^2_u & 0 \\ 0 & 0 & \sigma^2_v \end{bmatrix}\right), \end{aligned} \tag{A.1}$$

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 $\widehat{\delta^w_j} + \hat{\theta} S_j$, where $\widehat{\delta^w_j}$ is the estimate of δ^w_j 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 Model (3) 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:

$$\begin{aligned} y_{ij} &= \alpha_j + X'_{ij}\beta + \delta_j T_{ij} + e_{ij}, \\ \alpha_j &= \alpha + Z'_j\gamma + u_j \\ \delta_j &= \delta + Z'_j\theta + v_j, \\ \text{Var}\begin{pmatrix} e_{ij} \\ u_j \\ v_j \end{pmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_e & 0 & 0 \\ 0 & \sigma^2_u & 0 \\ 0 & 0 & \sigma^2_v \end{bmatrix}\right), \end{aligned} \tag{A.2}$$

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

Potential Moderator Variables

To identify candidate moderator variables for the regression models specified in the previous section, we relied on published reports and papers—including published reports from the studies that collected the data—and assumed that the authors used theory or evidence to identify factors that are likely to influence the magnitude of impact. In selecting or constructing site-level variables for the analysis, we focused on the subset of these 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 or children (e.g., share of students who are black, share of students who are in poverty) and schools (e.g., urbanicity).

In particular, we identified potential moderators for the analysis as follows:

- **Charter schools.** The evaluation of charter schools (Gleason et al., 2010) analyzed whether impacts were associated with six measures of the policy environment, six measures of school operations, and four measures of student and school characteristics (see Gleason et al., 2010, Table V.2). We included the measures of school operations and student and school characteristics;[15] we excluded the measures of the policy environment that were specific to the charter schools in the study because this information would not be known to local policymakers before deciding whether to authorize additional charter schools in their sites.
- **Education technology.** The evaluation of educational technology (Dynarski et al., 2007) analyzed whether impacts were associated with nine measures of classroom instruction and six measures of school characteristics, including five student-level measures aggregated to the school level (see Dynarski et al., 2007, Table II.8). We included all six school characteristics but excluded the measures of classroom instruction since these measures would not be known before deciding whether to adopt the technology in an individual site.
- **Head Start.** The Head Start Impact Study (Puma et al., 2010) did not analyze which grantee-level or center-level factors were associated with the program's impacts. Therefore, to identify candidate moderators, we reviewed several papers that examined such factors (Bitler, Hoynes, & Domina, 2014; Bloom & Weiland, 2015; Ding, Feller, & Miratrix, 2019; McCoy et al., 2016; Walters, 2015). Based on this review, we selected variables that could plausibly be known or at least estimated by a local policymaker who was trying to decide whether to apply for a Head Start grant[16] and excluded variables that could not plausibly be estimated in advance for a site.

This selection process yielded 11 potential moderators for the charter school study, seven potential moderators for the education technology study,[17] and 10 potential moderators for the Head Start study (see Table 1). While most of the potential 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

[15] To construct site-level variables from school-level measures, we identified all the schools that were part of the same site and took a weighted average of the school-level values, weighting by schools' total enrollment.

[16] To construct site-level variables from center-level measures, we identified all of the Head Start centers that were supported by the same grant and took a weighted average of the center-level values, weighting by the total number of children enrolled in the center.

[17] This includes the six analyzed by the study authors plus a classroom-level variable that we aggregated to the school-level.

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

To estimate different regression models, we selected one or more variables from the pool of potential moderators shown in Table 1. Equation (3) requires a single binary subgroup variable. We specified different versions of equation (4) with one, two, or five moderators.[18]

In selecting moderators for these models, we first chose 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). From these five variables, we selected moderators for each prediction approach as described below:

- **Binary subgroup approach (equation 3**). We tested all five possible models with a binary subgroup variable 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 continuous one-moderator models and selected the single moderator that minimized the unexplained variance of impacts across sites.
- **Two-moderator model (equation 4).** We tested all 10 possible two-moderator models and selected the two moderators that together minimized the unexplained variance of impacts across sites.
- **Five-moderator model (equation 4)**. All five candidate moderators were included in the five-moderator model.

[18] We considered implementing models with more than five moderators. However, the five-moderator model did not perform better than the two-moderator model, so it seemed implausible that additional moderators would improve the predictive accuracy of the regression models.

APPENDIX B: DERIVATION OF ADJUSTED ESTIMATOR FOR ROOT MEAN SQUARED PREDICTION ERROR

As indicated in equation (3), 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 (3).

To begin, suppose that:

- $\delta_j^w$is the true impact in site *j* (unknown for all sites)
- δ_j^w 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^w 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.

PREDICTION ERROR

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

$$\widehat{\delta_j^w} = \delta_j^w + \epsilon_j, \tag{B.1}$$

where δ_j^w 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: $\epsilon_j \sim N(0, \sigma_{\epsilon j}^2)$. The expected value of ϵ_j is zero because the estimator is unbiased.

Equation (B.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*:

$$\widehat{\delta_j^p} = \delta_j^w + b_j + \omega_j, \tag{B.2}$$

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—$\omega_j \sim N(0, \sigma_{\omega j}^2)$—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:

$$\begin{aligned} PE_j &= \widehat{\delta_j^p} - \delta_j^w \\ &= \left(\delta_j^w + b_j + \omega_j\right) - \delta_j \\ &= \delta_j^w + \omega_j. \end{aligned} \tag{B.3}$$

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

$$\begin{aligned}\widehat{PE}_j &= \widehat{\delta}_j^p - \widehat{\delta}_j^w \\ &= \left(\delta_j^w + b_j + \omega_j\right) - \left(\delta_j^w + \epsilon_j\right) \\ &= \left(b_j + \omega_j\right) - \epsilon_j.\end{aligned} \tag{B.4}$$

Comparing equation (B.3) and (B.4), we can see that the estimated prediction error in equation (B.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 ($\sigma_{\omega j}^2 + \sigma_{\epsilon j}^2$) exceeds the variance of the true prediction error ($\sigma_{\omega j}^2$).

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, and provide an alternative estimate that corrects for the bias and averages the corrected MSPE estimates across the sites.

MSPE for a Single Site

The mean squared prediction error (MSPE) for site j is defined as the expected squared prediction error in site j, where this expectation is defined across repeated samples selected to predict the impact in site j:

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

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:

$$\widehat{MSPE}_j = \widehat{PE}_j^2. \tag{B.6}$$

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:

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

Fortunately, the bias in the estimated MSPE for site *j* can be estimated (without bias) and removed. Equation (B.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*:

$$\widetilde{MSPE}_j = \widehat{PE}_j^2 - \hat{\sigma}_{\epsilon j}^2. \tag{B.8}$$

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

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

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 *J* sites in this sample:

$$MSPE = \frac{1}{J}\sum_{j=1}^{J} MSPE_j. \tag{B.10}$$

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

$$\widetilde{MSPE} = \frac{1}{J}\sum_{j=1}^{J} \widetilde{MSPE}_j. \tag{B.11}$$

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

$$\begin{aligned} E\left(\widetilde{MSPE}\right) &= E\left(\frac{1}{J}\sum_{j=1}^{J}\widetilde{MSPE}_j\right) \\ &= \frac{1}{J}\sum_{j=1}^{J}E\left(\widetilde{MSPE}_j\right) \\ &= \frac{1}{J}\sum_{j=1}^{J}MSPE_j \\ &= MSPE. \end{aligned} \tag{B.12}$$

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

$$\widetilde{MSPE} = \frac{1}{J}\sum_{j=1}^{J}\left[\left(\widehat{\delta_j^p} - \delta_j^w\right)^2 - \hat{\sigma}_{\epsilon j}^2\right] = \frac{1}{J}\sum_{j=1}^{J}\left(\widehat{\delta_j^p} - \delta_j^w\right)^2 - \frac{1}{J}\sum_{j=1}^{J}\hat{\sigma}_{\epsilon j}^2, \tag{B.13}$$

where $\frac{1}{J}\sum_{j=1}^{J}(\widehat{\delta_j^p} - \delta_j^w)^2$ is the average of the squared prediction error estimates and $\frac{1}{J}\sum_{j=1}^{J}\hat{\sigma}_{\epsilon j}^2$ is the average of the variance estimates for the unbiased, site-level impact estimates.

For interpretability of the size of the MSPE, we present the Root Mean Square Prediction Error (RMSPE) by taking the square root of the $\widetilde{MSPE}$. Since the square root is a nonlinear function, $\sqrt{\widetilde{MSPE}}$ is not necessarily an unbiased estimator of $\sqrt{MSPE}$, but a Taylor Series approximation shows that our estimate of the RMSPE is an underestimate of the true RMSPE.

APPENDIX C: ESTIMATING THE RISK FUNCTION FOR CORRELATED δ_j^w AND $\widehat{\delta_j^p}$

Computation of R(C*) when Estimates are Uncorrelated

The general formula for the risk function is:

$$\mathrm{R}\,(\mathrm{C}) = \Pr\left(\widehat{\delta_j^p} < \mathrm{C} \text{ and } \delta_j^w > \mathrm{C}\right) + \Pr\left(\widehat{\delta_j^p} > \mathrm{C} \text{ and } \delta_j^w < \mathrm{C}\right) \tag{C.1}$$

where:

Pr ($\widehat{\delta_j^p}$ < C and δ_j^w > C) is the probability that the predicted impact will show that an effective program (i.e., δ_j^w > C) is ineffective (i.e., $\widehat{\delta_j^p}$ < C);

Pr ($\widehat{\delta_j^p}$ > C and δ_j^w < C) is the probability that the predicted impact will show that an ineffective program (i.e., δ_j^w < C) is effective (i.e., $\widehat{\delta_j^p}$ > C).

In the special case of zero correlation between $\widehat{\delta_j^p}$ and δ_j^w, these two random variables are independent, and the risk formula reduces to:

$$\mathrm{R_j}\,(\mathrm{C}) = \Pr\left(\widehat{\delta_j^p} < \mathrm{C}\right) \cdot \Pr\left(\delta_j^w > \mathrm{C}\right) + \Pr\left(\widehat{\delta_j^p} > \mathrm{C}\right) \cdot \Pr\left(\delta_j^w < \mathrm{C}\right). \tag{C.2}$$

Approximating R(C) When δ_j^w and $\widehat{\delta_j^p}$ are Correlated

Unfortunately, there is no closed-form solution for the probabilities in the risk function formula (C.1) when the true impact (δ_j^w) and the predicted impact for the site ($\widehat{\delta_j^p}$) are correlated. However, we can approximate the probabilities in equation (C.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 ($\mathrm{x_{iv}}$,$\mathrm{y_{iv}}$). Within each of these squares that satisfy either of the conditions in equation (C.1) (either ($\mathrm{x_{iv}}$ < C and $\mathrm{y_{iv}}$ > C) or ($\mathrm{x_{iv}}$ > C and $\mathrm{y_{iv}}$ < 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 (C.1). This sum is approximately equal to 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 plus-or-minus four standard deviations.

$$R(C) = \sum_{v=-4/w}^{+4/w} \sum_{i=-4/w}^{+4/w} P_{iv} w^2 f\left(x_{iv}, y_{iv}\right), \tag{C.3}$$

where:

$$\begin{aligned} \mathrm{P_{iv}} &= 1 \text{ if either} \left(\mathrm{x_{iv}} < \mathrm{C} \text{ and } \mathrm{y_{iv}} > \ \mathrm{C}\right) \text{ or } \left(\mathrm{x_{iv}} > \ \mathrm{C} \text{ and } \mathrm{y_{iv}} < \ \mathrm{C}\right) \\ &= 0 \text{ otherwise} \end{aligned}$$

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

= bivariate normal probability density function

ρ = the correlation between x and y.

Setting the correlation between δ_j^w and $\widehat{\delta_j^p}$ 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 δ_j^w and $\widehat{\delta_j^p}$ is zero. The results are shown in Table C1, which shows our central risk measure, S(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 Table C1). For that reason, we use areas of width .005 in the approximations shown in the following section of this appendix.

Estimating S(C) for Alternative Values of ρ

To test the sensitivity of the risk estimates to the correlation ρ between the estimates of δ_j^w and $\widehat{\delta_j^p}$, we used the approximation with a .005 grid to estimate S(C) for correlations at .1 intervals from -.9 to +.9, for C = 0, C = .25, and C = .5, for all outcomes in the analysis. The results are shown in Table C2. S(C) for $\rho = 0$ is shown in the center column of the table, with differences from that value for each correlation tested shown in the remaining columns.

As can be seen in the last three rows of the table, when C = .25 or C = .50, the average estimated values of S(C) for correlated δ_j^w and $\widehat{\delta_j^p}$ differ by at most .015 standard deviations from the values for uncorrelated δ_j^w and $\widehat{\delta_j^p}$, and differ by only about .07 standard deviations when C = 0, a difference of only about 16 percent, even in the extreme cases of −.9 or +.9 correlation. 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 S(C) by formula) and computationally efficient $\rho = 0$ to produce the results shown in the text and in Appendix D.

Computer Code (in R) Used to Generate Risk Estimates When δ_j^w and $\widehat{\delta_j^p}$ 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, predicted impact, and
# standard error of prediction.
#
#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)
      }
    }
  }
}
```

Table C1. S(C) computed by formula and by two approximations, $\rho = 0$.

	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

Table C1. Continued.

	By formula (A)	.01 approx (B)	.005 approx (C)	Difference (B) - (A) (D)	Difference (C) - (A) (E)
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

Table C2. Differences in S(C) from S(C) for $\rho = 0$, alternative values of ρ.

	Diff $\rho = -.9$	Diff $\rho = -.8$	Diff $\rho = -.7$	Diff $\rho = -.6$	Diff $\rho = -.5$	R(C), $\rho = 0$	Diff $\rho = +.5$	Diff $\rho = +.6$	Diff $\rho = +.7$	Diff $\rho = +.8$	Diff $\rho = +.9$
Charter year 1 math											
C = 0	0.047	0.043	0.038	0.034	0.028	0.400	−0.030	−0.036	−0.042	−0.048	−0.053
C = .25	0.002	0.001	0.001	0.001	0.001	0.110	−0.002	−0.002	−0.003	−0.003	−0.004
C = .50	0.000	0.000	0.000	0.000	0.000	0.018	0.000	0.000	0.000	0.000	0.000
Charter year 2 math											
C = 0	0.041	0.036	0.031	0.027	0.022	0.351	−0.023	−0.028	−0.033	−0.038	−0.043
C = .25	0.017	0.016	0.013	0.011	0.009	0.214	−0.010	−0.012	−0.015	−0.017	−0.020
C = .50	0.008	0.007	0.006	0.005	0.004	0.110	−0.004	−0.004	−0.005	−0.005	−0.006
Charter year 1 reading											
C = 0	0.105	0.094	0.081	0.070	0.058	0.559	−0.057	−0.069	−0.080	−0.091	−0.103
C = .25	0.002	0.002	0.002	0.002	0.002	0.141	−0.003	−0.003	−0.004	−0.005	−0.006
C = .50	0.000	0.000	0.000	0.000	0.000	0.034	0.000	0.000	0.000	0.000	0.000
Charter year 2 reading											
C = 0	0.041	0.037	0.033	0.028	0.024	0.414	−0.025	−0.030	−0.036	−0.041	−0.047
C = .25	0.000	0.000	0.000	0.000	0.000	0.146	−0.001	−0.001	−0.001	−0.001	−0.001
C = .50	0.000	0.000	0.000	0.000	0.000	0.030	0.000	0.000	0.000	0.000	0.000
Ed Tech math 6											
C = 0	0.062	0.055	0.048	0.041	0.033	0.472	−0.035	−0.042	−0.049	−0.056	−0.064
C = .25	0.015	0.014	0.012	0.010	0.011	0.322	−0.013	−0.015	−0.018	−0.021	−0.024
C = .50	0.000	0.000	0.000	0.000	0.000	0.097	0.000	0.000	0.000	0.000	−0.001
Ed Tech algebra											
C = 0	0.082	0.074	0.066	0.058	0.049	0.455	−0.055	−0.067	−0.080	−0.094	−0.109
C = .25	0.004	0.004	0.003	0.003	0.003	0.176	−0.004	−0.006	−0.007	−0.009	−0.011
C = .50	0.000	0.000	0.000	0.000	0.000	0.050	0.000	0.000	0.000	0.000	−0.001
Ed Tech towre											
C = 0	0.069	0.060	0.052	0.044	0.037	0.518	−0.034	−0.041	−0.047	−0.053	−0.058
C = .25	0.008	0.007	0.006	0.005	0.004	0.259	−0.005	−0.007	−0.008	−0.009	−0.011
C = .50	0.000	0.000	0.000	0.000	0.000	0.082	0.000	0.000	0.000	0.000	0.000
Ed Tech reading 1											
C = 0	0.065	0.059	0.052	0.044	0.038	0.453	−0.039	−0.047	−0.055	−0.063	−0.072
C = .25	0.008	0.008	0.006	0.005	0.005	0.253	−0.006	−0.007	−0.009	−0.010	−0.012
C = .50	0.000	0.000	0.000	0.000	0.000	0.108	0.000	−0.001	−0.001	−0.001	−0.001

Table C2. Continued.

	Diff $\rho = -.9$	Diff $\rho = -.8$	Diff $\rho = -.7$	Diff $\rho = -.6$	Diff $\rho = -.5$	R(C), $\rho = 0$	Diff $\rho = +.5$	Diff $\rho = +.6$	Diff $\rho = +.7$	Diff $\rho = +.8$	Diff $\rho = +.9$
Ed Tech reading 4											
C = 0	0.129	0.114	0.098	0.084	0.069	0.483	−0.068	−0.082	−0.096	−0.110	−0.124
C = .25	0.011	0.009	0.008	0.007	0.005	0.221	−0.005	−0.006	−0.007	−0.008	−0.009
C = .50	0.000	0.000	0.000	0.000	0.000	0.059	0.000	0.000	0.000	0.000	0.000
Head Start PPVT											
C = 0	0.009	0.008	0.007	0.006	0.006	0.292	−0.007	−0.008	−0.010	−0.011	−0.013
C = .25	0.022	0.019	0.017	0.015	0.013	0.402	−0.014	−0.017	−0.020	−0.023	−0.025
C = .50	0.001	0.001	0.000	0.000	0.000	0.154	0.000	−0.001	−0.001	−0.001	−0.001
Head Start WJ AP											
C = 0	0.030	0.027	0.024	0.021	0.018	0.392	−0.019	−0.022	−0.026	−0.030	−0.034
C = .25	0.024	0.022	0.020	0.017	0.015	0.329	−0.017	−0.020	−0.024	−0.028	−0.032
C = .50	0.001	0.001	0.001	0.001	0.001	0.125	−0.001	−0.001	−0.001	−0.001	−0.002
Head Start WJ LW											
C = 0	0.007	0.007	0.006	0.005	0.004	0.344	−0.006	−0.007	−0.008	−0.009	−0.010
C = .25	0.058	0.052	0.045	0.039	0.032	0.441	−0.033	−0.040	−0.046	−0.053	−0.060
C = .50	0.000	0.000	0.000	0.000	0.000	0.188	−0.001	−0.001	−0.001	−0.001	−0.001
Head Start WJ OC											
C = 0	0.140	0.124	0.108	0.092	0.076	0.500	−0.075	−0.090	−0.106	−0.122	−0.137
C = .25	0.001	0.001	0.001	0.001	0.001	0.223	−0.001	−0.001	−0.001	−0.002	−0.002
C = .50	0.000	0.000	0.000	0.000	0.000	0.078	0.000	0.000	0.000	0.000	0.000
Head Start self-regulation											
C = 0	0.120	0.107	0.093	0.079	0.066	0.520	−0.064	−0.077	−0.089	−0.102	−0.114
C = .25	0.006	0.005	0.005	0.004	0.003	0.274	−0.004	−0.005	−0.006	−0.006	−0.007
C = .50	0.000	0.000	0.000	0.000	0.000	0.121	−0.001	−0.001	−0.001	−0.001	−0.001
Head Start externalizing											
C = 0	0.105	0.093	0.082	0.071	0.057	0.486	−0.059	−0.067	−0.079	−0.092	−0.106
C = .25	0.003	0.003	0.003	0.003	0.000	0.248	0.000	0.002	0.001	0.001	0.001
C = .50	0.000	0.000	0.000	0.000	0.000	0.121	0.000	0.000	0.000	0.000	0.000
Average											
C = 0	0.070	0.063	0.055	0.047	0.039	0.443	−0.040	−0.048	−0.056	−0.064	−0.072
C = .25	0.012	0.011	0.009	0.008	0.007	0.251	−0.008	−0.009	−0.011	−0.013	−0.015
C = .50	0.001	0.001	0.000	0.000	0.000	0.092	−0.001	−0.001	−0.001	−0.001	−0.001

APPENDIX D: DETAILED ESTIMATES OF S(C)

In this appendix, we provide estimates of S(C), the probability of making an incorrect policy decision, for each study outcome and prediction method, for alternative values of C. In the final section of the appendix, we also provide the computer code that was used to compute these estimates.

Estimated S(C) by Outcome and Method of Predicting Site-Specific Impacts, Alternative Values of C

To give the reader an overall sense of how S(C), the average risk of an incorrect policy decision across sites, varies with C and the prediction method, the results presented in Table 3 in the main paper were averaged across all outcomes in each study. In this appendix, we present detailed estimates of S(C) by outcome, for three different values of C and five prediction methods. Results for each of the three studies are shown in separate tables.

Computer Code (in R) Used to Generate Estimates of S(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 predicted impact 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 predicted 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 then applies 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, for example, we
  #have 73 sites, so this leaves a 801 × 73 matrix.
  rc ← adply(all.sites,1,mean)
  #this single command above then (following the PPVT example)
  #takes the 801 × 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 × 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
#predicted mpact estimates and standard errors for all sites into
#site.walk, ij, and ijx arguments, respectively, of below function.
bell.orr(site.walk = ,ij = ,ijx = )
```

Table D1. Charter Schools: Probability of wrong policy decision, for alternative outcomes, prediction methods, and values of C.

Policy Cutoff	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
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
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
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

Table D2. Educational Technology: Probability of wrong policy decision, for alternative outcomes, prediction methods, and values of C.

Policy Cutoff	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
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
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
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
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

Table D3. Head Start: Probability of wrong policy decision, for alternative outcomes, prediction methods, and values of C.

Policy Cutoff	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
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
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
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
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
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