Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix

Lant Pritchett and Justin Sandefur

Abstract

In this paper we examine how policymakers and practitioners should interpret the impact evaluation literature when presented with conflicting experimental and non-experimental estimates of the same intervention across varying contexts.

We show three things. First, as is well known, non-experimental estimates of a treatment effect comprise a causal treatment effect and a bias term due to endogenous selection into treatment. When non-experimental estimates vary across contexts any claim for external validity of an experimental result must make the assumption that (a) treatment effects are constant across contexts, while (b) selection processes vary across contexts. This assumption is rarely stated or defended in systematic reviews of evidence. Second, as an illustration of these issues, we examine two thoroughly researched literatures in the economics of education—class size effects and gains from private schooling—which provide experimental and non-experimental estimates of causal effects from the same context and across multiple contexts. We show that the range of "true" causal effects in these literatures implies OLS estimates from the right context are, at present, a better guide to policy than experimental estimates from a different context. Third, we show that in important cases in economics, parameter heterogeneity is driven by economy- or institution-wide contextual factors, rather than personal characteristics, making it difficult to overcome external validity concerns through estimation of heterogeneous treatment effects within a single localized sample.

We conclude with recommendations for research and policy, including the need to evaluate programs in context, and avoid simple analogies to clinical medicine in which "systematic reviews" attempt to identify best-practices by putting most (or all) weight on the most "rigorous" evidence with no allowance for context.

JEL Codes: D04, I2, O2

Keywords: external validity, treatment effects, policy evaluation, causal inference.

Center for Global Development
www.cgdev.org

Working Paper 336
August 2013

Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix

Lant Pritchett

Justin Sandefur

Lant Pritchett: Center for Global Development, lpritchett@cgdev.org; Justin Sandefur: Center for Global Development, jsandefur@cgdev.org. We are grateful to Paul Glewwe, Eric Hanushek, Sarah Humpage and Renato Ravina for sharing data. Sarah Dykstra provided excellent research assistance. We received helpful comments from Chris Blattman, Angus Deaton, and Andrew Zeitlin. All errors and opinions are ours.

CGD is grateful to its board of directors and funders, including the UK Department for International Development and the Swedish Ministry of Foreign Affairs, for support of this work.

Lant Pritchett and Justin Sandefur. 2013. "Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix." CGD Working Paper 336. Washington, DC: Center for Global Development.

http://www.cgdev.org/content/publications/context-matter-for-size

Center for Global Development
1800 Massachusetts Ave., NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors or funders of the Center for Global Development.

Contents

1 Introduction

There are two fundamentally distinct approaches to development. Easterly (2006) summarizes the dichotomy as "planners" and "searchers" but many other scholars, from different disciplines and with different politics, posit a similar dichotomy.[1] Rodrik (2008) called the "new development economics" an approach of "experimentation" which emphasized the adaptation to local context and a search for "best fit" rather than "best practice" (Crook and Booth, 2011).[2]

The other popular movement in development economics has been the rise of a methodological concern with the identification of causal impacts of development projects, programs and policies, particularly the advocacy of the use of randomization as a technique in program evaluation. As a methodological issue this rise of randomization is, on the face of it, neutral with respect to the development approach – "planners" or "searchers" – to which it is applied.

However, there is an interpretation of the use of RCTs that combines a "planning" approach to development with a "rigorous" approach to evidence that we argue is superficially attractive but, on closer examination, logically incoherent. That is, people speak of generating evidence of "what works" and then using that evidence to eliminate programs or policies that "don't work" (or "lack evidence") and so "scale up" those that are "demonstrated" to work. The paradigmatic example of this "planning with rigorous evidence" approach is vaccinations – once a vaccination has been demonstrated to be medically efficacious and cost-effective – then it is merely top-down logistics to fund and scale the implementation of the vaccine. However, the scope of application of the "planning with rigorous evidence" approach to development is vanishingly small. In nearly all development contexts it cannot be assumed that the rigorous demonstration of "what works" (as both efficacious and cost-effective) in one context has superior evidentiary value for any other context. We show that the claims of "external validity" that are a necessary component of the "planning with rigorous evidence" approach to development are not just unlikely but actually embody logical

[1] This distinction is old and cuts across ideological and disciplinary boundaries. In *Seeing Like a State*, Scott, a political scientist who is a "Marxist" (of the American academic type) contrasts "high modernism" with *metis* (1998). Elinor Ostrom, a political scientist by training who won the Nobel Prize in Economics, contrasted "hierarchical" with "polycentric" systems. This distinction goes back to the very foundations of development with the contrasting approaches of "central planning" to allocate the capital budget and Hirschmann's notions of "unbalanced growth."

[2] Pritchett, Woolcock, and Andrews (2012) label the "planner" approach to development of state capability or "good governance" as "accelerated modernization through the transplantation of best practice."

incoherence when existing non-experimental evidence shows widely varying impacts. New evidence from RCTs and other rigorous approaches to program evaluation must interpreted in a way that encompasses all of the known facts–including the facts embodied in the non-experimental evidence.

We are wary of the criticism that we are assembling a straw man here – that nobody actually believes we should run a randomized trial in one non-random setting, under potentially artificial conditions, extrapolate the results around the world and ignore strong contradictory evidence. So before we dive into the crux of our argument, it's useful to highlight four concrete features of current thinking and practice around impact evaluation in development that we would like to contest. Each of these four tenets contains strong, albeit often tacit claims to the external validity of impact evaluation results.

The first feature is a lexicographic preference for internal over external validity, as evidenced by strict rankings of empirical methodologies. For example, the U.S. Department of Education publishes a handbook outlining the standards of evidence for its well-known catalog of evaluation results, the "What Works Clearinghouse" (Institute of Education Sciences, 2008). The first hurdle in evaluating evidence is randomization; failure to randomize disqualifies any study from meeting the standards of evidence. Thus if the hypothetical principal of a school serving low-income children in Brooklyn, New York, was looking for new ideas, the Department of Education would point her to randomized evidence from Boise, Idaho, and discount careful non-randomized research in much more relevant contexts.

Second, development agencies are increasingly commissioning "systematic reviews" of impact evaluations and encouraging the use of formal meta-analysis methods to aggregate results across studies. For instance, the protocol for a recent DFID systematic review of voucher programs for private schools noted that "each study will be represented by a single effect size for each outcome variable, and we will use CMA [Comprehensive Meta-Analysis software] to statistically combine results from the evaluations" (Fronius, Petrosino, and Morgan, 2012). The stated goal is to produce an average effect size and confidence interval for all studies, with secondary focus on a small set of (four) contextual variables that might explain variation across studies.

Third, funding for experimental evaluation in development economics and political science is highly concentrated on a small set of large studies (often with total budgets in the millions of dollars), clustered in a small set of relatively peaceful, democratic and very poor settings (Blair, Iyengar, and Shapiro, 2013). For funding institutions with a global remit, this allocation of resources appears rational only on the basis of bold claims to external validity.

The fourth and most obvious example of ambitious external validity claims is in the formulation of global policy prescriptions from one or a few localized studies. For instance, Banerjee and He (2008) proposed a list of proven interventions from randomized and quasi-experimental studies which, they argued, the World Bank should scale up globally.

In response to the impact evaluation paradigm described by these four features – (i) evidence rankings that ignore external validity, (ii) meta-analysis of the *average* effect of a vaguely-specified 'intervention' which likely varies enormously across contexts, (iii) clustering evaluation resources in a few expensive studies in locations chosen for researchers' convenience, and (iv) the irresistible urge to formulate global policy recommendations – we argue for far greater attention to context and heterogeneity.

This is not an argument against randomization as a methodological tool for empirical investigation and evaluation. It is actually an argument for orders of magnitude *more* use of randomization, but with far fewer grand claims to external validity. To be part of an effective development practice RCTs have to embed themselves firmly into a "searcher" paradigm of development, in which rather than RCTs being mostly used by "outsiders" for "independent" evaluation, RCTs and other methods are brought into the learning process of development organizations themselves (Pritchett, Samji, and Hammer, 2012).

This paper has three parts. First, we show that claims to external validity of estimated impacts from RCTs *must* be wrong because they are logically incoherent. Second, we demonstrate this claim with two specific examples from the economics of education: the effects of class size, and the return to private schooling. Third, drawing on the parameters from the education literature, we show that a rule of giving preference to RCT estimates of causal impact can lead to less accurate decisions than relying on non-experimental estimates in spite of their potential bias. More broadly, once extrapolated from its exact context (where context includes everything) RCT estimates lose *any* claim to superior "rigor."

2 The logical incoherence of external validity claims in the social sciences

Science advances by encompassing all previous observations into a new conceptual framework or theory that generates superior understanding of the phenomena at hand. That is, general relativity had to explain why many observations were consistent with Newton's formulation of gravitation and more. Quantum mechanics had to explain why observations of particles could generate observations of both particle-like and wave-like behavior. Evolution had to encompass previous factual observations about species and speciation. The emphasis on the structure and key role of DNA in biology had to encompass previous observations about, say, inheritance.

Given the many uses to which it is put, it is easy to forget that statistical procedures like Ordinary Least Squares (OLS) produce empirical facts . The mean height of a sampled population is a statistic that is an empirical fact. The standard deviation of weight in a sampled population is a statistic that is an empirical fact. In exactly that same way, the OLS coefficient of regressing weight on height is a summary statistic and is an empirical fact.

While OLS coefficients can sometimes be "tests" of models or hypotheses they are primarily themselves observations to be explained by any adequate characterization of the world. Any discipline's "best available theory" (Barrett 2002) has to adequately encompass all available observations about the world, including encompassing the empirical facts of existing OLS coefficients.

To illustrate our point that external validity claims from RCT results are logically incoherent we want to decompose economic models into two sets of parameters, each of which represent different aspects of causal structures of the world. One are the "causal impact parameter(s) of T on Y" and the other are the "parameter(s) that cause OLS statistical procedures of T on Y to be inconsistent as estimates of causal impact."

In the notation of Rubin's (1974) potential outcome framework, let $T_{ij} \in [0, 1]$ be the treatment indicator, where $i = 1, \ldots, N$ denotes the units of observation and $j = 1, \ldots, J$ denotes different contexts or samples. $Y_{ij}(T)$ for $T = 0, 1$ denotes the potential outcome for unit i given treatment T. For each unit i we observe the treatment T_{ij}, the outcome conditional on that treatment, $Y_{ij}(T_{ij})$, and a vector of covariates, $\mathbf{X}_{ij}$.

Following Imbens (2003), we make three starting assumptions:

$$\begin{aligned} Y_{ij}(0), Y_{ij}(1) &\not\perp T_{ij} | \mathbf{X}_{ij}, && (1) \\ Y_{ij}(0), Y_{ij}(1) &\perp T_{ij} | \mathbf{X}_{ij}, U_{ij} && (2) \\ \mathbf{X}_{ij} &\perp U_{ij} && (3) \end{aligned}$$

Equation (1) relaxes the assumption of strict exogeneity, allowing, for instance, that OLS estimates of non-experimental data yield biased estimates of true treatment effects. Equations (2) and (3) introduce a hypothetical, unobserved covariate. Without loss of generality, it is possible to define this omitted variable such that unconfoundedness holds after conditioning on U_{ij}, and U_{ij} is independent of $\mathbf{X}_{ij}$.

For the sake of simplicity, we assume a linear functional form in which the true causal model relating the treatment T_{ij} to outcome Y_{ij} is:

$$\begin{aligned} \mathrm{E}(Y_{ij}|T_{ij}, X_{ij}, U_{ij}) &= \mathbf{X}_{ij}\gamma_{2j} + \delta_2 U_{ij} + \epsilon_{2ij} + \beta_j T_{ij} \qquad (4) \\ \mathrm{E}(T_{ij}|X_{ij}, U_{ij}) &= \mathbf{X}_{ij}\gamma_{1j} + \delta_{1j} U_{ij} + \epsilon_{1ij} \qquad (5) \end{aligned}$$

We refer to (4) as the treatment equation, β_j as the treatment effect of T on Y, and 5 as the selection equation. In the absence of a "clean" identification strategy, researchers will produce biased estimates of both the treatment and selection parameters. Equations (4) and (5) imply that the bias in non-experimental estimates, which we denote with a tilde ($\tilde{\beta}_j$) is given by the standard expression for omitted variable bias in OLS:

$$\delta_j \equiv \delta_{2j} \frac{\widehat{\mathrm{cov}(T_{ij}, U_{ij})}}{\widehat{\mathrm{var}(T_{ij})}} = \tilde{\beta}_j - \beta_j. \qquad (6)$$

We refer to $\tilde{\beta}$, interchangeably, as an OLS estimate, a non-experimental estimate, a naive estimate, an estimate using observational data, or a non-identified estimate of the true causal effect.[3] In contrast, we assume that studies using experimental methods, instrumental variables, or regression discontinuity designs are able to produce a consistent estimate of the underlying causal parameter, which we denote with a hat, ($\hat{\beta}$).[4]

Note that we index all parameters by j, to indicate possible heterogeneity across contexts – i.e., a possible lack of external validity even for well-identified, internally valid estimates.

[3] We recognize that referring to $\tilde{\beta}$ as an OLS estimate is technically imprecise, as OLS yields unbiased estimates of true causal effects in the context of a controlled experiment. However, this terminology is commonly used and understood to refer to OLS estimates of causal effects using observational data.

[4] For the sake of argument, we set aside concerns outlined by, e.g., Deaton (2010) about the internal validity of experimental estimates and the practical relevance, even within the same context, of local average treatment effects. Instead, we follow common practice in the recent microeconometric literature by drawing a sharp distinction between methods that rely on an explicit source of exogenous variation (and thus deliver "clean" identification), and those that rely on controlling for potential confounding variables (e.g., multiple regression including differences-in-differences and fixed effects estimators, as well as various fomrs of matching estimators).

Take the example of regressing some metric of "student learning" on some measure of "class size." Context could be country (Austria versus Bolivia), region (Alabama versus Idaho), time period (1976 versus 2012), institutional setting (e.g. public versus private schools, regular versus contract teachers), intervention implementation responsibility (line ministry versus NGO), sampling frame (e.g. only rural schools, only "disadvantaged" children).

This framing is quite general across development projects as it could involve some metric of health status on some measure of clinic availability or some metric of individual/household income on some measure of access to finance or some metric of income volatility on some measure of access to insurance or some metric of malarial incidence on treated bed net price, etc.

Suppose we have OLS estimates of $\tilde{\beta}$ using non-experimental data from a variety of contexts. These OLS results either vary across contexts in practically meaningful ways or they don't. Figure 1 shows possible distributions of OLS results across contexts (assuming they follow a roughly normal distribution) where the metric is scaled between "zero", "threshold" and "large." That is, if one interpreted the OLS coefficients as representing a causal impact the "threshold" magnitude would be that such that, based on some cost-benefit, cost-effectiveness, or return on investment calculus, the decision rule would be to "expand X" as a means of achieving gains in Y. In this context "large" means that the magnitude of the OLS coefficient, if interpreted causally, implies a very large benefit-cost ratio or very large return on investment.

Suppose we do one rigorous experiment that estimates causal impact of T on Y in context j. For purposes of argument let us assume this estimate is consistent and hence estimates the "true" causal impact in the usual LATE sense: i.e., $p\text{lim}\hat{\beta}_j = \beta_j$ for individuals whose treatment status affected by random assignment.

The question of external validity can be posed as "how does a rigorous estimate of causal impact in one context (j) affect our beliefs about of causal impact in other contexts (k)?" We can divide this into cases were either (a) there already are non-experimental estimates of $\tilde{\delta}_k$ in the context of interest or (b) there are no estimates at all.

Imagine the simple case of forming a new estimate of causal impact in context k as a linear weighted average of the OLS estimate from k and the RCT estimate from context j. The OLS estimate in context k depends both on parameters of the causal mechanism of impact (δ_k) and on parameters of the causal mechanism of bias in OLS ω_k and the vagaries of sampling ($\mathrm{E}(\epsilon_{ij})$) – all of which are at least possibly context specific.

$$\beta_k = \alpha\hat{\beta}_j + (1-\alpha) \times \tilde{\beta}_k \tag{7}$$

We divide the discussion by whether the experimental or quasi-experimental estimate from context j ($\hat{\beta}_j$) significantly differs from existing OLS estimates from context k ($\tilde{\beta}_k$) or not, and whether the RCT study produces either just an estimate of causal impact, or also an estimate of the bias which OLS would have yielded in context j (i.e., whether the RCT estimates only $\hat{\beta}_j$, or also $\tilde{\beta}_j$ and $\hat{\delta}_j$). We argue that estimates of the bias in OLS contexts from context j can be especially helpful in judging whether experimental estimates from j posess external validity in context k.

In all cases, we focus on the situation in which there is large variance, centered around the threshold – which is the situation in which there is likely to be the most active controversy about causal impacts as evidence will exist on all sides of the debate because both zero and the "threshold" level of impact are within the range of existing empirical estimates. In fact, some might think that this is the situation in which new and "rigorous" evidence might help the most but where actually it cannot be helpful at all. The location of the existing estimates is irrelevant, what is key, as we see below, is the magnitude of the variance.

2.1 Good estimates from the wrong place versus bad estimates from the right place

What can "good" estimates from the wrong place (i.e., experimental estimates of $\hat{\beta}_j$) tell us about causal effects in context k or l? And what can we can conclude when those good estimates from the wrong place contradict "bad" estimates from the right place (i.e., non-experimental estimates from the context of interest (e.g., $\tilde{\beta}_k$)?

First, consider the case where experimental evidence from context j falls within the range of existing non-experimental estimates from contexts k and l, such that $\tilde{\beta}_k < \hat{\beta}_j < \tilde{\beta}_l$. Suppose, following equation (7), we form our estimate of causal impact in contexts k and l as a linear weighted average of the existing OLS estimates and the experimental evidence $\hat{\beta}_j$ with weight α on the latter. The weight represents the degree of external validity we impose; $\alpha = 1$ ignores context specific evidence altogether, placing all weight on the internally valid estimate; conversely, $\alpha = 0$ would imply giving complete priority to external validity concerns over internal validity concerns.

Clearly in this case, any assumption of external validity ($\alpha > 0$) implies that the preferred estimate of causal impact in k is larger than the OLS estimate from k, while it is smaller than the OLS estimate in context l. But – ignoring for the moment the role of idiosyncratic sampling error – this implies that the structural bias (δ) in context k is negative while the structural bias in l is positive.

$$\begin{aligned}
\beta_k &= \alpha\hat{\beta}_j + (1-\alpha)\tilde{\beta}_k \\
\Rightarrow \delta_k &= \alpha(\hat{\beta}_j - \tilde{\beta}_k) < 0 \\
\beta_l &= \alpha\hat{\beta}_j + (1-\alpha)\tilde{\beta}_l \\
\Rightarrow \delta_l &= \alpha(\hat{\beta}_j - \tilde{\beta}_l) > 0
\end{aligned}$$

Thus, in the context of the widely-used model of treatment effects sketched in equations (4) and (5), claims regarding the external validity of one set of parameters contradict any claim to external validity for others. Yet such claims are frequently made with no attempt to reconcile the assertion of complete homogeneity of the β parameter in equation (4) and the simultaneous, unstated assertion of wide heterogeneity across contexts in the δ_1 parameters in equation (5).

This is not a reasonable assertion. As a general approach to judging the external validity of experimental evidence, it is in fact logically incoherent.

Suppose that a separate research team ran a parallel experiment in the same context j which, instead of measuring the causal impact of T on Y, was designed to estimate the δ_1 parameter determining selection into treatment. In the face of widely variant OLS estimates of $\tilde{\beta}_k$ and $\tilde{\beta}_l$, any claim to the external validity of the RCT estimate of $\hat{\beta}_j$ would directly contradict the identically external valid claim of the RCT estimate of $\hat{\delta}_j$.

Second, consider the case where RCT estimates from context j fall outside the range of non-experimental OLS estimates of $\tilde{\beta}_j$ from contexts k and l. In the previous example, the estimates of structural bias had different signs due to the assumption that the RCT estimate was within the range of the existing estimates (in the assumed case of large variability of those OLS estimates.) Alternatively, the RCT estimate could be outside the existing range and assume (without loss of generality, by symmetry) that the RCT is larger than any OLS estimate. This doesn't change the implication that any positive proportional weight on the

RCT estimate from context j in estimating the causal impact in contexts k and l implies very different estimates of structural bias as the OLS estimates (by assumption) are far apart.

Figure 2 illustrates the point that claims to external validity necessarily reduce the variance of the distribution of estimates of causal impact relative to the naive assumption that OLS estimates represent causal estimates. But this reduction in variance is not itself based on any evidence. That is, something about the world produced the observed variation in OLS coefficients across contexts. That something can be decomposed into (a) true variation in causal impact across contexts (b) variation in the structural bias of the existing OLS coefficients as estimates of causal impact and (c) idiosyncratic error in the existing OLS estimates. Assigning any given weight, α, to an RCT result makes a strong empirical claim about the relative sources of variation in OLS estimates that is itself evidence free and hence not rigorous at all.

* * *

Any useful statement of "the evidence" has to be a statement about the evidence about a complete causal representation which explains both the RCT evidence and the OLS evidence in terms of underlying models. This means that a statement about how the rigorous evidence about causal impact from context j affects one's beliefs about the causal impact in context k is necessarily a statement about how evidence from j should affect priors about both causal impacts (β) and selection parameters determining who gets treated ($\gamma_1, \gamma_2, \delta_1, \delta_2$).

A common slogan is that "one good experiment trumps a thousand bad regressions."[5] This suggests that the weight on the randomized evidence is complete, $\alpha \approx 1$. In this view we collapse our distribution of priors about true causal impact in all contexts ($\beta_k \forall k$) to a single parameter $\hat{\beta}_j$. While this slogan might have some merit if one could be completely confident that the "good experiment" and all thousand "bad regressions" were in exactly the same context and estimating exactly the same parameters, this is rarely (if ever) the case in development economics.[6] One cannot "trump" a thousand OLS coefficients with a randomized experiment any more than one can "trump" estimates of the height of children in Nepal, Kenya, and Indonesia with "better" estimates of the height of children in the USA.

[5] The reference is presumably to Krueger (1999) who said "One well designed experiment should trump a phalanx of poorly controlled, imprecise observational studies."

[6] Das, Shaines, Srinivasan, and Do (2009) show most developing countries have very few (often only one or two) published papers in economics on all topics, so the odds any given country has a published paper addressing empirically any given question is near zero.

The correlations and partial associations (for any given set of conditioning variables) in the data are themselves facts about the world. Of course, a "good experiment" may affect how we interpret OLS regression coefficients, but how exactly they do that is actually a quite difficult question.

2.2 Using measures of selection on observables and unobservables from here and there

So far we have been assuming that the RCT study in context j only produces an estimate of causal impact, but we can also explore the case in which the study is able to produce both an estimate of the causal impact from randomized variation and an estimate of that an OLS regression would have produced. For instance, if a study collects baseline data on test scores, class sizes and characteristics then the baseline data can estimate the OLS estimate of class size while the experiment can then produce an unbiased experimental estimate of causal impact for context j. This therefore produces an estimate of the difference between $\tilde{\beta}_j$ and $\hat{\beta}_j$, and hence by simple decomposition, an estimate of the structural bias (plus idiosyncratic error) $\hat{\delta}_j$.

This doesn't make the logical incoherence of claims to external validity any better (or worse) but does clarify what the options are.

Consider the first case discussed above, where experimental estimates from context j fall within the range of non-experimental estimates from other contexts, i.e., $\tilde{\beta}_k < \hat{\beta}_j < \tilde{\beta}_l$. Again, we face two competing external validity claims. On the one hand, assuming any degree of external validity for $\hat{\beta}_j$ ($\alpha > 0$) implies opposite structural biases in contexts k and l. Specifically, we know that $\delta_k < 0$. If we have a direct estimate of $\hat{\delta}_j > 0$ accompanying the RCT evidence in context j, that's a fairly strong sign that the underlying causal model in contexts j and k simply aren't the same. We can remain agnostic about external validity claims between contexts j and l.

So far we've focused on potential logical inconsistencies, but estimation of structural bias parameters can also help in constructing a positive case for external validity.

Let's return to the second case discussed above, where RCT evidence from j falls outside the range of OLS estimates from k and l. Suppose an RCT of private schooling in j finds zero effect, while OLS estimates from k and l show significant positive returns to private schooling, i.e., $\hat{\beta}_j < \tilde{\beta}_k < \tilde{\beta}_l$. If the RCT also reveals that OLS estimates in j are biased upward ($\delta_j > 0$), one could argue we are a step closer to a coherent explanation of all the available facts. While asserting external validity of the β estimates would still require us to gloss over the heterogeneity in the implied δ parameters, at least all parameters are of a potentially consistent sign. In contrast, if the RCT finds that OLS estimates in j are biased *downward* ($\delta_j < 0$), this would have to be interpreted as positive evidence against the external validity of the RCT estimates of the treatment effect $\hat{\beta}_j$, as it is clear the underlying structural model in j differs substantively from that of k and l.

A similar logic applies to patterns of selection into treatment on observable characteristics. To see how estimation of selection effects due to observable characteristics might inform external validity claims, define $\tilde{\tilde{\beta}}$ as the unconditional difference in the mean of Y between treated and untreated individuals, and $\tilde{\beta}$ as the selection bias in this unconditional mean as revealed by controlling for X, such that:

$$\begin{aligned} \mathrm{E}(Y_{ij}|T_{ij}) &= \tilde{\tilde{\beta}}_j T_{ij}, \text{ and} \\ \tilde{\delta}_j &\equiv \tilde{\tilde{\beta}}_j - \tilde{\beta}_j \end{aligned}$$

Now imagine that a combination of RCT and observational evidence from context j yields a vector of parameters $\{\hat{\beta}_j, \tilde{\beta}_j, \tilde{\tilde{\beta}}_j\}$. If non-experimental evidence from context k yields parameters $\{\tilde{\beta}_k, \tilde{\tilde{\beta}}_k\}$ that are consistent with the evidence from j, this is further evidence for the external validity of estimates from j in k. Not only can the RCT evidence encompass all the known facts from context k, but this includes direct evidence that the selection process into treatment operates in a similar fashion across contexts. A core premise of the recent literature on sensitivity analysis in the estimate of treatment effects is that the parameters guiding this selection on observables (i.e., the gap between $\tilde{\beta}_k$ and $\tilde{\tilde{\beta}}_k$, or γ_2) is a useful guide to the likely effect of the size and sign of structural bias due to unobservable characteristics (cf. Altonji, Elder, and Taber (2005); Imbens (2003); Harada (2013)).

Note the approaches outlined in this section – using patterns in observational data analysis alongside experimental evidence – are not feasible for all experimental studies. An impact evaluation of the introduction of a truly novel technological innovation would find no observational variation in the use of said innovation at baseline. But for many of the questions studied in the applied microeconomics of development – including the examples reviewed in depth below, i.e., class size effects, public-private schooling test-score differentials– OLS estimates are readily producible. When baseline observational variation does not exist, experimental designs (such as encouragement designs, or randomization of the cost or information hurdles associated with take-up) that provide direct evidence on demand for and incidence of the intervention are not only of independent interest, but may also greatly assist in assessing the external validity of study findings.

3 Illustration: Class size effects in Tennessee, Tel Aviv, and Teso

In the early 2000's, the World Bank came under increased pressure to adopt a more "evidence based" approach to lending. As noted in the introduction, Banerjee and He (2008) proposed that the World Bank should immediately stop lending for anything that had not been proven to work by a rigorous evaluation. Responding to the objection that this would grind operations to a halt while people waited for proof, they argued that no, this wasn't so, and provided a list of "proven" interventions which, if scaled-up globally, could easily absorb the entirety of the World Bank's lending portfolio.

One example from that list was a study of class size by Angrist and Lavy (1999) which estimated the effect of class-size on test performance in Israeli primary schools. The study helped popularize regression discontinuity designs in applied economics. In this case, the discontinuity hinged on an administrative rule passed down from the rabbinic scholar Maimonides, stipulating that class sizes should not exceed forty. Exploiting this cutoff in an IV framework, they found a negative, significant effect of class size of -.26 standard deviations on both English and math tests in fifth grade, and negative albeit insignificant effects of -.07 and -.03 for fourth grade English and math, respectively. They found no significant effects for grade three, which they speculated may reflect the cumulative nature of class-size effects over time.[7]

On the basis of these results, rigorous for their context, it was argued that class size reductions should be prioritized globally in World Bank education programs. This example illustrates several of our points about external validity claims in development economics.

[7] The paper, which helped popularize regression discontinuity designs in applied economics, is informative not only about the causal effect of class size, but also about the selection processes that create contradictory class-size "effects" in observational data. Regressing scores scores on class-size alone, Angrist and Lavy find a strong positive correlation, ranging from .141 in fourth-grade reading tests to 0.322 for fifth-grade math, all of which significant at the 1% level. Controlling for percent of disadvantaged students, this positive association is attenuated, turning negative for reading and remaining positive for math but with a much smaller magnitude.

3.1 A selective review of the class-size literature

At least three other published papers attempt to replicate Angrist and Lavy's identification strategy in other settings. The results demonstrate the challenge of cross-national generalizations about the effects of education policies. One of the three studies reached broadly similar findings, the second reached essentially opposite results, and the third found the identifications strategy to be invalid in the setting proposed.

Transplanting Maimonides rule from Israel to Bolivia, Urquiola (2006) exploits a similar administrative rule allowing schools who pass certain class size threshold to apply for an additional instructor. Without controls and without exploiting the discontinuity, OLS regressions of test scores from third-grade exams show positive coefficients of .09 in language and .07 in math, both significant at the 5% level. Adding controls for student, teacher, and school characteristics, these coefficients are reduced to approximately zero in both cases. Notably, the coefficient on class size is negative and significant at the 5% level for both language and mathematics – with coefficients of -0.22 (-.23) and -.19 (-.19) without (with) controls – for a sub-sample of rural schools with enrollments of 30 or fewer pupils. Urquiola argues class size is more plausibly exogenous in these circumstances, as isolated schools with a single class per grade cannot sort higher- or lower-ability students into smaller classes. The findings from small, rural schools are corroborated by IV results from the full sample, which show negative and significant effects of class size for both language and math scores, though the significance of the effect on math scores is not robust to the full set of school-level controls.

So far so good: Bolivian data behaves like Israeli data. Not only do both studies find significant, negative effects of larger class sizes, but in both cases selection on unobservables obscures these effects, and furthermore, selection on observable and unobservable characteristics both point in the same direction – with erstwhile higher scoring students being grouped into larger classes.

But this same pattern does not appear to be true for Bangladeshi data, where both the putative causal effect of class size and the selection process into larger classes appears to operate in the opposite direction.

Asadullah (2005) applies the Angrist and Lavy (1999) identification strategy to secular Bangladeshi secondary schools and finds very different results. He exploits a government policy, similar to those in Israel and Bolivia, that allows registered secondary schools to recruit an additional teacher whenever enrollment in a single grade exceeds 60 or an integer multiple of 60. Regressing school-level scores from the national secondary certificate examination on controls for school type (public vs private, single gender) and geographic fixed effects yields a positive (i.e., 'wrongly signed') coefficient on class size that is significant at the 1% level. IV estimates are also positive, significant at the 1% level, and roughly fourteen-times larger in magnitude. While the OLS coefficient implies a 0.25 standard deviation increase in exam scores for each increase in class size of ten pupils, the IV estimates imply a 3.5 standard deviation increase.

There is reason to be cautious in assuming even the internal validity of these estimates. Urquiola and Verhoogen (2009) construct a model of school and household behavior in a setting where parents are free to choose between schools and schools are free to adjust prices and reject pupils. The model predicts discontinuities in household characteristics near the class-size cutoffs. They show these concerns undermine the validity of the identification strategy in the case of Chile, where parents have considerable scope to choose schools.

Lest we conclude, however, that the variability of results between Israel, Bolivia, and Bangladesh is an artefact of an unreliability regression discontinuity design, it is noteable that the same heterogeneity across contexts turns up in the few existing randomized trials of class size in the economics literature.

Krueger (1999) re-analyzes the data from the Tennessee STAR experiments, in which both teachers and pupils in kindergarten were randomly assigned to small (13-17 students) or regular (22-25 sudents) class sizes starting in 1985-6 and tracked through third grade. After examining various potential threats to the internal validity of the experiment – including non-random attrition, non-compliance, and complex re-randomization protocols – Krueger concludes that the causal effect of small class-size on test performance ranged from 0.28 standard deviations in first grade to 0.19 in third grade – equivalent to about 82% of the black-white score gap. (In Table 1 we scale the grade 3 effects to reflect an increase of 10 pupils per teacher for comparability with other studies, yielding an effect of -0.27.)

Do more experiments in the developing world show similar results? The short answer is no.

Banerjee, Cole, Duflo, and Linden (2007) report results from a remedial education intervention in Indian primary schools that provides an indirect, experimental estimate of class size effects on test-score value added. The remedial instructor worked with the twenty lowest-performing pupils for half of the school data, implying a reduction in class size for the remaining pupils, but no change in their instructor. Results show that the experimental intervention had a statistically insignificant negative effect on pupils not directly participating in the remedial classes, implying a statistically insignificant, positive effect of class size equivalent to 0.064 standard deviations from an increase in class sizes of ten pupils. Unfortunately, Banerjee et al. do not report the relationship between class size and test scores or value added using the observational variation in their data, as this was not the main focus of the paper. However, using the public data release it is possible to estimate a simple regression of value added on class size for grades which were not affected by the intervention. This observational variation also yields a positive coefficient on class size but of somewhat smaller magnitude, equivalent to 0.027 standard deviations from an increase in class size of ten pupils. This coefficient is statistically significant at the 5% level after clustering standard errors at the school level. (The data are available here, and the regression specification in Stata for the result reported here is "reg vad numstud if bal==0".)

Turning from India to Kenya, Duflo, Dupas, and Kremer (2012) report on a very similar experiment in which random assignment of a contract-teacher intervention created experimental variation in the class size for the remaining children working with the normal civil service teacher. Results show an increase in scores from .042 to .064 standard deviations in total scores (math and English) for a 10-pupil reduction in class size, depending on the controls included. While these effects are statistically significant at the 5% level in each case, the authors note that – as in the Indian case – they are much of a significantly smaller magnitude than the successful results of the STAR experiment.

Figure 4 shows a summary of the estimates of class-size effects from a systematic review of the empirical literature on school resources and educational outcomes in developing countries conducted by Glewwe, Hanushek, Humpage, and Ravina (2011). They distinguish between studies with and without a "clean" identification approach. In the top panel we report estimates from studies using OLS or propensity matching techniques based on observable characteristics[8], and in the bottom panel we report estimates from experimental, RDD, and

[8] The sample of studies here consists of: Arif and us Saqib (2003); Aslam (2003); Bacolod and Tobias (2006); Banerjee, Cole, Duflo, and Linden (2007); Bedi and Marshall (2002); Behrman, Khan, Ross, and Sabot (1997); Brown and Park (2002); Cerdan-Infantes and Vermeersch (2007); Chin (2005); Du and Hu (2008); Engin-Demir (2009); Glewwe, Grosh, Jacoby, and Lockheed (1995); Gomes-Neto and Hanushek (1994); Hanushek and Luque (2003); Lee and Lockheed (1990); Marshall (2009); Marshall, Chinna, Nessay, Hok, Savoeun, Tinon, and Veasna (2009); Michaelowa (2001); Nannyonjo (2007); Psacharopoulos, Rojas, and Velez (1993); Urquiola (2006); Warwick and Jatoi (1994); and Yu and Thomas (2008).

IV estimates.[9]

The top panel shows a fairly uniform distribution of results across negative significant, negative insignificant, and positive significant results among non-experimental studies, with a somewhat higher concentration of positive significant findings.[10] All signs are defined so that positive is "good", i.e., a reduction in class size leads to an increase in scores. The bottom panel, focusing on "cleanly identified" results, shows a slightly different pattern, again lopsided in favor of positive findings, but with a stronger tendency toward insignificant effects.

Finally, an independent sample of estimates of class-size effects illustrating the same points is provided by Wöβmann and West (2006). They use comparable test data from the TIMSS project to estimate class-size effects across 18 countries, mostly in the OECD. The top panel of Figure 5 shows the distribution of "naive" OLS class-size effects estimates. It is centered well below zero (implying 'perverse' class size effects), with a wide range from roughly -6 to 2.[11]

In addition to these naive OLS estimates, Wöβmann and West (2006) also report estimates using school fixed effects and instrumenting class size with the average class size for the relevant grade and school. This approach overcomes endogenous selection of stronger or weaker pupils into small classes within the same school and grade, but is of course still vulnerable to endogenous sorting of pupils across schools. The results in the middle panel of Figure 5 show that IV estimates are centered just above zero. Comparing the OLS and IV estimates provides an estimate of the structural bias in OLS; the distribution of these biases across countries is shown in the bottom panel. As anticipated, IV estimates push the class size effects in the "correct" direction, but both the IV effects and the estimates of OLS bias evince large variance across contexts.

[9] The sample of studies here consists of Angrist and Lavy (1999); Asadullah (2005); Bedi and Marshall (1999); Khan and Kiefer (2007); Suryadarma, Suryahadi, Sumarto, and Rogers (2006); and Wöβmann (2005).

[10] When studies report multiple estimates, all are included but weighted so that each study receives equal weight.

[11] For comparability with the large literature on TIMSS, we report raw coefficients, reflecting the effect of a one pupil increase on a TIMSS score, which has a mean of 500 and standard deviation of 100. To compare with other estimates in this section, divide the coefficients by 100 to convert to traditional effect sizes and multiply by -10 to consider the hypothetical experiment of increasing class-size by ten pupils.

3.2 Lessons

This example reveals three important lessons about the use of randomized trials for evidence-based policymaking.

First, there have been literally hundreds of studies of class size from non-experimental data, many of which used quite plausible methods of identification. The cumulated evidence – which mostly shows very small impacts (sufficiently small to be statistically indistinguishable from zero) – is ignored when making bold external validity claims from a single experiment. That is, there is no "encompassing" explanation offered as to why all of these previous results are – as empirical facts – consistent with this one piece of evidence from Israel.

Second, our review of the literature shows that probably the most notable feature of the distribution of class-size effects in the larger literature is not that it is centered around a small, statistically insignificant effect, but that it is widely varying across contexts.

Third, the heterogeneity in class-size effects is real. It affects all the parameters of the underlying model, including but not limited to well-identified causal treatment effects. In principle, as outlined in detail in Section 2, it is possible that widely varying OLS estimates reflect a homogenous treatment effect (β) and heterogeneous parameters of structural bias or selection into treatment (δ) – or vice versa. In practice, we have seen that *both* the causal impact on learning of class size reductions, *and* the selection bias in OLS estimates from observational data are widely variant across contexts. Applying the identical instrumental variables strategy to TIMSS data from multiple countries produces not only a wide-range of $\hat{\beta}$ estimates, but shifts these estimates in opposite directions depending on the country. Similarly, both RDD estimates and RCTs using very similar methodologies across contexts produce different results in Israel and the USA versus Bangladesh, India, and Kenya. Nor, we would stress, do all developing countries produce similar results: Bolivia looks more like Israel or the USA than India or Kenya in this respect.

4 Illustration: The return to private schooling when public schools work, and when they don't

At a recent conference on the economics of education a paper was presented in which student selection accounted for all of the difference in outcomes between private and public schools.

The justification provided by the authors for this extreme assumption was that most rigorous evidence–all from the United States shows near zero causal impact on student learning of private over public schools. The argument, made explicit by the authors at the conference, was that the "best" estimate of the impact of private schools on learning for *all* was to extrapolate the most internally valid estimates, even if all those estimates were from one particular (if not peculiar) context.

We would argue instead that any review of the existing literature – some of it experimental, most of it not – on the return to private schooling in developing countries would lead to very different starting assumptions. Private schools serve a very different function in contexts where public schools function relatively well (such as the U.S.), and in contexts where they don't. The process of endogenous selection into private schools will also vary widely based on the school finance model in a given country or school district.

Perhaps the 'cleanest' test of the causal effect of private schooling in a developing country context is provided by Angrist, Bettinger, Bloom, King, and Kremer's (2002) study of Colombia's voucher program for private secondary schools, in which eligibility was decided by a random lottery. The headline result, expressed in the most comparable terms to the other results here, is that lottery winners scored 0.2 standard deviations higher on combined math, reading, and writing tests – the equivalent of a full additional year of schooling. Furthermore, note that this is an intent-to-treat (ITT) effect, where all lottery winners are coded as 'treated', even though actual take-up of the vouchers was 49% per annum and 88% overall. Rather than an OLS estimate of the ITT effect, instrumental variables estimates of the average treatment on the treated (ATT) would be more comparable to the coefficients described in other studies below. The instrumental variables estimates which Angrist et al report use a slightly different treatment variable, defined as using any scholarship (not restricted to the program voucher, and not limited to private schools). This yields a treatment effect of 0.29.[12]

Unfortunately, for our purposes here, Angrist, Bettinger, Bloom, King, and Kremer (2002) do not provide various pieces of information that would help us adjudicate the external validity of this internally valid causal effect estimate. First, no information is provided on how lottery participants compare to the general population. On the one hand, program eligibility was restricted to neighborhoods from the lowest two of six socio-economic strata,

[12] In the first stage, use of any scholarship was observed for 24% of those who did not get a voucher and 90% of those who did. The OLS estimate of the effect of using a scholarship on test scores was actually higher than the IV estimate mentioned in the main text, 0.38 vs 0.29.

but on the other hand applicants had to take the initiative to respond to radio ads, apply, and provide proof of eligibility. Second, conditional on applying and winning the lottery, we know that only 88% of voucher winners ever redeemed their vouchers and claimed a private school scholarship; we don't know how those 88% compare to the 12% who did not. Third, Angrist et al. do not report the raw difference in means (or non-experimental OLS estimates of the gap) between public and private schools at baseline, either in terms of test scores or socio-economic characteristics. Thus we learn nothing about the selection process into private schools which the voucher program aims to affect, and it is impossible to know from this study whether ex ante non-experimental estimates of the program would have produced significantly biased estimates of its effects – and thus whether non-experimental estimates elsewhere should be treater with greater or lesser confidence on the basis of these findings.

Earlier studies struck a very different balance between internal and external validity concerns. Cox and Jimenez (1991) was one of the first papers examining the returns to private versus public schooling in a developing country context, using college-entrance examination results for secondary school students in Colombia and Tanzania. While they pay careful attention to the selection process into public and schooling, their estimation technique remains vulnerable to selection on unobservable characteristics (i.e., criticisms of its internal validity).

Based on simple averages, Cox and Jimenez (1991) show that scores in Colombia were higher in private schools by 0.22 standard deviations, and higher in Tanzanian public schools by 0.14 standard deviations. But there were strong a priori reasons to anticipate very different selection processes into private secondary schools in these two countries. At the time of the study, Tanzanian public secondary schools screened on the basis of competitive entrance exams, were heavily subsidized, and attracted the best students, while in Colombia affluent households disproportionately sent their children to elite private secondary schools perceived to be of higher quality. Using survey data on students' socio-economic backgrounds, Cox and Jimenez estimated – separately for each country – a two-stage 'Tobit' model to explicitly account for the non-random selection of pupils into private schools. Results confirmed the hypothesis of opposite patterns of selection into private schools on the basis of household economic and demographic characteristics, i.e., 'positive selection on observable characteristics' in Colombia and 'negative selection' in Tanzania. Once controlling for this selection process, Cox and Jimenez find large, positive score differentials in favor of private schooling in both countries, equivalent to 0.42 standard deviations (4.51 points) in Colombia and 0.75 standard deviations (6.34 points) in Tanzania. Interestingly, these estimates for Colombia are roughly double the magnitude of those reported in Angrist, Bettinger, Bloom, King, and Kremer (2002) a decade later, but it is impossible to know whether this reflects the cleaner identification of causal effects in Angrist et al, or idiosyncrasies of their non-representative sample.

Chile is arguably the only developing country to have adopted a voucher model of school finance at the national level, beginning in 1981. In a widely cited study, Hsieh and Urquiola (2006) introduce a novel strategy to tease out the causal effect of private schooling on test performance in Chile in the absence of clean natural experiment. At baseline in 1982, public schools scored just 3% below the average score of private schools, and public school pupils ranked just 4% lower an index of socioeconomic status. To get at causal effects, Hsieh and Urquiola examine the evolution of test scores over time, from 1982 to 1996, regressing changes in aggregate scores – combining both public and private schools – on the share of private schools in the commune, an administrative area encompassing all relevant school choices for most households. Effects measured in this way remove any potential bias due to the endogenous sorting of richer, or more able students into private schools. Contrary to the Angrist, Bettinger, Bloom, King, and Kremer (2002) findings from Colombia, results show a negative, though mostly insignificant effect of private schooling on test performance, and a positive effect on the repetition rate. These (non-) results are robust to alternative identification strategy, using the baseline urbanization rate as an instrumental variable – though effect sizes vary quite widely from roughly zero to a negative effect of more than one standard deviation (Table 4). Interestingly, Hsieh and Urquiola (2006) also examine the effect of vouchers on sorting into private schools on the basis of socioeconomic status and academic performance, finding a robust, statistically significant increase in the differential between public and private schools under the voucher program.

In short, the best available evidence suggests Chile's large-scale voucher program did nothing to improve academic performance. Do these results generalize to other settings? At least two other studies have adapted Hsieh and Urquiola's identification strategy to other contexts and found very different results.

Tabarrok (2013) draws on household survey data from India, where the share of pupils in private schooling is high — 27% among 6 to 14 year-olds nationwide in 2005, and 50% in urban areas – despite the lack of large-scale voucher programs as in Chile. Controlling for demographic and socioeconomic characteristics, pupils in private schools score 0.36 standard deviations higher on reading and 0.23 standard deviations higher on arithmetic tests. When pooling public and private scores at the district level to overcome selection bias a la Hsieh and Urquiola (2006), Tabarrok finds even larger, albeit only marginally significant effects of private schooling on test-score performance, equivalent to a 0.64 standard deviation increase in reading and a 0.4 standard deviation increase in math for a hypothetical move from 0% to 100% private schooling in a district.

The explosion of private schooling in South Asia has been more muted in East Africa, but private enrollment appears to have increased quite quickly in Kenya in the wake of the abolition of user fees in government primary schools in 2003, rising from 4.8% of pupils completing grade eight in 1998 to 9.7% in 2005. Bold, Kimenyi, Mwabu, and Sandefur (2013) adopt a similar strategy to Hsieh and Urquiola (2006) and Tabarrok (2013)) to analyze the effects of this private enrollment on school-leaving exams administered nationwide at the end of grade eight. Comparing school level scores, they find a 51 point (unconditional) difference in scores between public and private schools across all subjects, equivalent to roughly 0.78 standard deviations. When aggregating scores at the district level and including district fixed effects to overcome endogenous sorting, the coefficient rises to 64 points or 0.98 standard deviations, significant at the 5% level. Unfortunately, the national exam data used by Bold et al are not linked to any survey information on household characteristics, so the authors cannot observe the pattern of selection on socioeconomic characteristics into private schools in this sample, but it is striking that the unconditional difference in means and the econometric estimates controlling for selection on all pupil characteristics are very similar in magnitude.

5 Calibration: Too much weight on "rigorous" evidence can be worse than useless

So far we have shown that for two prominent questions in the economics of education, experimental and non-experimental estimates appear to be in tension. Furthermore, experimental results across different contexts are often in tension with each other. The first tension presents policymakers with a trade-off between the internal validity of estimates from the "wrong" context, and the greater external validity of observational data analysis from the "right" context. The second tension, between equally well-identified results across contexts, suggests that the resolution of this trade-off is not trivial. There appears to be genuine heterogeneity in the true causal parameter across contexts.

These findings imply that the common practice of ranking evidence by its level of "rigor", without respect to context, may produce misleading policy recommendations. In principle, this possibility is fairly obvious and well known, yet in practice appears to be heavily discounted in both academic and policy discussions. Here we present a simple calibration of the widely-used treatment effects model outlined in Section 2. Our goal is to calculate the errors implied by reliance on OLS estimates (due to structural bias) versus the errors implied by reliance on a single experimental estimate (due to the limits of external validity). This calibration exercise draws on the parameters from the education literature surveyed above, including not only estimated treatment effects and biases, but the variance across studies. Given the current state of the field, the evidence here suggests that policymakers would do well to prioritize external validity over internal validity concerns when surveying the development economics literature.

Our measure of the accuracy of the experimental and non-experimental estimates is their mean squared error (MSE), i.e., how much the estimates deviate from the true effect. For the non-experimental estimate, the MSE is given by the sum of the sampling error and the omitted variable bias due to the failure to observe and control for U_{ik}.

$$\text{MSE}(\tilde{\beta}_k) = \underbrace{\text{Var}(\tilde{\beta}_k)}_{\substack{\text{Sampling} \\ \text{error}}} + \underbrace{(\tilde{\beta}_k - \beta_k)^2}_{\substack{\text{Omitted} \\ \text{var. bias}}} \tag{8}$$

As shown above, the omitted variable bias depends not only on the size of the selection parameter δ_1, but also the magnitude of the effect of U_{ik} on Y_{ik}, as well as the overall variance of U_{ik}.[13]

On the experimental side, the key parameter of interest in the MSE is the underlying variance of the true β parameter across contexts. When using an experimental estimate from one context ($\hat{\beta}_j$) as an estimate of the causal effect in another (β_k) the mean squared error

[13] Following Imbens (2003), the discrepancies in non-experimental estimates can be summarized as a function of the partial R-squared of the omitted variable. This is helpful in two respects: (i) it reduces the source of bias to a single summary measure, and (ii) although the partial R-squared of the omitted variable is by definition not observable, using this metric allows us to discuss in a meaningful way how 'strong' an omitted variable would need to be – relative to the explanatory power of the observed X_{ik} characteristics – to bias the estimates of β by a given amount. For instance, following Altonji, Elder, and Taber (2005), we could examine the bias in OLS estimates of the return to private schooling implied by the existence of an unobserved pupil characteristic U_{ik} – pushing pupils into private school and raising their test scores – with the same explanatory power as the observed socioeconomic characteristics X_{ik}.

is:

$$\mathrm{MSE}(\hat{\beta}_j) = \underbrace{\mathrm{Var}(\hat{\beta}_j)}_{\substack{\text{Sampling error} \\ \text{in context } j}} + \underbrace{\mathrm{Var}(\beta)}_{\substack{\text{Variance of true effect} \\ \text{across contexts}}} \tag{9}$$

To operationalize these equations, we return to Tables 1 and 2 which provide – as best as we are able to glean from the existing literature – comparable experimental (or quasi-experimental IV or RDD) estimates alongside non-experimental OLS estimates from a variety of contexts for the effect of class size and attending a private school. Reading across a single row provides a measure of structural bias, and hence MSE, in non-experimental estimates, by comparing them to the cleanly-identified experimental or quasi-experimental evidence from the same context. Reading down the penultimate column privates a measure of the MSE in cleanly identified, experimental or quasi-experimental estimates, by comparing these unbiased parameters across contexts.

Figure 6 presents the results of the MSE calculations. The y-axis shows the MSE of non-experimental estimates and the x-axis shows the MSE of experimental, IV, or RDD estimates. For a given body of evidence, if these MSE statistics fall above the forty-five degree line, this implies that the experimental estimates from the "wrong" context are a better guide to the true causal effect (equivalent to a high α in equation 7). On the other hand, if the MSEs fall southeast of the forty-five degree line, this implies that we would do well to rely on OLS estimates of observational data from the context of interest. Note that the number of data points available to compute the relevant MSE statistics is vanishingly small, so this exercise should be seen as illustrative more than definitive.

We graph the points for five distinct literatures. First, the RCT evidence on class size provides no examples we are aware of which estimate both experimental and observational, OLS parameters from the same context. However, as noted above, we have computed an OLS statistic using the public data release from Banerjee, Cole, Duflo, and Linden (2007), which suggests the bias overcome through randomization moves the point estimate from (positive, i.e., wrongly signed) 0.027 to 0.064. In contrast, estimates for the U.S. from Krueger (1999) show an effect of -0.27.[14] Comparing these two discrepancies shows an MSE of just 0.0014 for the OLS estimate in India, versus an MSE of 0.112 if one were to attempt to naively apply the RCT estimate from the U.S. to India.

[14] In computing the MSE, we treat this pair as a random draw from the possible distribution of RCT parameters across contexts – ignoring the presumably non-random choice to focus on the U.S. in the early literature.

Second, we show the results for RDD estimates of the class size effect. Qualitatively the results are similar, though the variance in estimates within countries and across countries is much larger. Angrist and Lavy (1999) find an OLS coefficient of 0.019 compared to an RDD estimate of -0.26. However, this variation is dwarfed by the finding of a positive (i.e. perverse) effect of 3.5 standard deviations by Asadullah (2005). The latter results are clearly suspect; however, this may be considered a fair reflection of the risks associated with widespread reliance on IV estimates which may be extremely fragile to, e.g., weak first-stage regressions as highlighted by Stock and Yogo (2005).

Third, the evidence on class-size effects reported by Wöβmann and West (2006) provides a immediate measure of the MSE for both OLS and IV estimates, as they provide both and their sample spans multiple countries. Once again, we see in Figure 6 that the mean squared error implied by the variance of IV estimates across countries is far, far greater than the structural bias in OLS estimates from a given context (in this case, the IV error is greater by more than two orders of magnitude).

Fourth, the literature on private schooling is an area many might expect huge selection biases in OLS estimates, but which also reveals highly variable estimates of causal effects across contexts after accounting for this selection bias. For Chile, Hsieh and Urquiola (2006) show roughly equal effects of private schooling before and after allowing for non-random sorting of pupils into private schools (-0.12 versus -0.10). Applying a similar estimation strategy in Kenya, however, Bold, Kimenyi, Mwabu, and Sandefur (2013) find effects of 0.98 (and again, little evidence of structural bias due to selection).

Fifth, we discuss the Mincerian labor market returns to education in more detail in Section 6.2 below. Here we preempt one point from that discussion. Some of the best known examinations of how endogenous selection into schooling might undermine OLS estimates of the Mincerian return to education have found little or no evidence of structural bias. We focus here on Duflo (2001), who finds a range of IV estimates from 3.5% to 10.6%, and closely matched OLS estimates of 3.4% to 7.8%, yielding an MSE across estimates of 0.0012. Meanwhile, cross-country data on OLS returns to education from Montenegro and Patrinos (2013) reveals variance of 12.25. Note that, in contrast to our other examples, we must rely on variance across context in OLS rather than IV parameters here; nevertheless, this provides suggestive evidence of the relative magnitude of the potential for error here.

In sum, despite the fact that we have chosen to focus on extremely well-researched literatures, it is plausible that a development practitioner confronting questions related to class size, private schooling, or the labor-market returns to education would confront a dearth of well-identified, experimental or quasi-experimental evidence from the country or context in which they are working. They would instead be forced to choose between less internally valid OLS estimates, and more internally valid experimental estimates produced in a very different setting. For all five of the examples explored here, the literature provides a compelling case that policymakers interested in minimizing the error of their parameter estimates would do well to prioritize careful thinking about local evidence over rigorously-estimated causal effects from the wrong context.

6 Learning from experiments when parameter heterogeneity is assumed

6.1 What external validity looks like (e.g. in the physical sciences)

This is not to say that external validity of all parameters – both causal impact and structural bias – across contexts is impossible; it is just that it requires a set of validated invariance laws that encompass both the "true" parameters and also explain all of the empirical observations.

An analogy is special relativity. We think of objects of having a "length" and we have invariance laws about length such as "translational invariance" so that if we displace an object from one location in $\{x, y, z\}$ coordinate space to another location in $\{x, y, z\}$ coordinate space the length is unchanged. Length is also invariant with respect to non-accelerating reference frames. However, an object may appear shorter if it is accelerating relative to the observers reference frame. So, if we had many experiments measuring the length of the same object then either (a) the variance should be low (if all measurements are made in non-accelerating reference frames) or (b) the apparent differences in length should be explained by the encompassing theory that predicts the experimentally observed length and its variation even when the "true" length is invariant.

Another analogy is the boiling point of water. By definition water boils at 100° Celsius at normal sea level atmospheric pressures. But as the atmospheric pressure decreases (say as altitude increases) the boiling temperature of water decreases. A series of experiments measuring the boiling temperature of water should find either (a) all produce roughly the same temperature (if all done at the same atmospheric pressure) or (b) the differences should follow the invariance law that describes observed boiling point of water with respect to some specified contextual variable – like altitude.

In contrast, there is “intrinsic” heterogeneity in the boiling point of various substances. That is, at “normal” atmosphere different substances boil at very different temperature (see Figure 3 and each substance also has an adjustment for boiling point by atmospheric pressure and hence has a known invariance law that allows us to adjust for known contextual conditions. Experiments which find the boiling point of water and Butane will produce different results at different atmospheres but can be adjusted for context but the differences in boiling points between water and Butane are intrinsic.

But this would mean a case in which an RCT of causal impact in context j would be strong and rigorous evidence for changing one’s beliefs about causal impact in all contexts would look like Figure 3. That is, external validity should have three features. One, the existing non-experimental evidence should be tightly clustered (or be able to be adjusted according to known observables and known transformations into being tightly clustered) because there is no “intrinsic” heterogeneity. Two, the RCT results would cluster around the “true” value (again, adjusted by known invariance laws if need be). Three, this implies estimates of the structural bias from comparing RCT and OLS estimates would also be tightly clustered (again, adjusted by invariance laws) because the parameters would be constant.

6.2 Heterogeneous treatment effects

One commonly advocated approach to addressing external validity concerns in experimental studies is to model heterogeneity in treatment effects. If causal parameters vary along, say, observable demographic characteristics, then the distribution of those same demographic characteristics can be used to formulate better predictions of the average effect of the intervention in a new setting. (For examples of this approach, see *inter alia* Stuart, Cole, Bradshaw, and Leaf (2011), Hartman, Grieve, and Sekhon (2010), and Tamer and Kline (2011).)

While paying attention to heterogeneity is laudable, it is unlikely to provide a route to generalizable findings from microeconomic experiments in development economics. Simply put, the heterogeneity within most experimental samples – drawn from a single country, evaluating a single ‘intervention’ implemented by a single institution – is of a much smaller magnitude in comparison to the huge heterogeneity encountered when extrapolating empirical results to wholly new settings.[15]

Consider the familiar challenge of estimating the Mincerian return to education, i.e., the causal effect of education on earnings measured as the percentage increase in an individual's earnings rate per period resulting from one additional year of schooling. Theory can provide a guide as to when, where, and for whom we should expect to see a large return to schooling and when (where, and for whom) we should not. For instance, theory might suggest that the Mincerian return might be higher for individuals with higher initial cognitive skills. Or, given labor market imperfections, and in an economy where most labor market earnings come through self-employment (either agricultural or non-), we might anticipate higher returns to human capital acquired through schooling for individuals with greater access to finance for complimentary physical capital investments. Building on the model in equations (4) and (5), we can summarize all such speculations by allowing the treatment effect of schooling to vary with whatever observable characteristics are measured in the data, such that the single β_j parameter is replaced by $\beta_{ij} = \beta_{ij}(\mathbf{X}_{ij})$.

To make this more concrete, suppose we estimate β as a linear, additively separable function of individual characteristics ($\mathbf{X}_{ij}$) and context-specific characteristics ($\mathbf{Z}_j$)

$$\beta_{ij} = \mathbf{X}_{ij}\beta_x + \mathbf{Z}_j\beta_z + \nu_{ij}. \tag{10}$$

This expression can be substituted into (4) in the form of a series of interaction terms with the treatment variable, T_{ij}. Estimates of β_x provide a firmer basis on which to predict β_k by applying those same parameters to the distribution of individual characteristics in context k.

However, this process assumes that the sample of individuals within context j captures the relevant sources of heterogeneity. If, instead, the context characteristics $\mathbf{Z}_j$ explain a

[15] Our focus here on extrapolating across contexts should not be confused with a related concern, i.e., that within a given context, observational studies using representative samples will produce more representative estimates of average treatment effects than will experimental studies relying on small, non-random samples. Contrary to this claim, Aronow and Samii (2013) demonstrate that representative sampling does not guarantee that observational studies will produce representative estimates of treatment effects for the sampled population when other factors are controlled for using multiple regression techniques. OLS estimates of $\tilde{\beta}_j$ can be seen as a weighted average of $\tilde{\beta}_{ij}$, where the weights vary with individual i's values for the control variables in the regression. At the risk of oversimplifying, the Aronow and Samii (2013) result can be seen as another point in favor of experimental identification within a given context, but this does not affect our argument about the risks associated with transplanting results – experimental or non-experimental – to a different context altogether.

sizeable share of the variation in β, it becomes less clear what analysis of any single context-specific sample can teach about parameters in other context.

Such is the case for the Mincerian returns to education. The World Bank's International Income Distribution Database (I2D2) presents harmonized micro data for the key variables in a typical Mincerian specification (income, education, experience, sex, occupation, and rural/urban residence) for 750 surveys spanning 158 countries. Clemens, Montenegro, and Pritchett (2008) draw on the I2D2 data to to quantify the enormous differences in average incomes across countries for observationally identical workers along these measured dimensions. Montenegro and Patrinos (2013) turn from measuring differences in the intercept to looking at differences in slope coefficients – estimating separate Mincerian returns to education for each available country-year cell using harmonized variable definitions.[16]

Figure 7 tabulates Mincerian coefficients from 128 surveys, based on Montenegro and Patrinos (2013) as reported in the appendix of King, Montenegro, and Orazem (2010). The blue bars show the return to education for the full sample – on average around 8%, but with a standard deviation across surveys of approximately 3.5%. The 5th percentile of the distribution of coefficients is below 3% (e.g., Egypt and Yemen) while the 95th percentile is nearly 15% (e.g., South Africa and Rwanda).

Can heterogeneous returns within these individual country samples help to explain this pattern across countries? Labor economics and human capital theory provide a long list of factors that might create variation in returns: the sector of employment, the quality of education, complementarity with non-cognitive skills or innate ability, and simple supply and demand of skilled labor in the local labor market, to name a few. Montenegro and Patrinos (2013) explore a variety of these hypotheses; here we focus on just one, differential returns between rural and urban sectors. If returns to schooling are higher in urban areas relative to rural, predominantly agricultural areas, then modeling the heterogeneity and adjusting for the share of the working population in rural areas might help explain heterogeneity in the rate of return to education across countries – analogously to the invariance laws in the physical sciences discussed above.

In practice, Figure 7 suggests the prospects for "explaining away" cross-country variation are dim. As anticipated, returns in urban areas are nearly a full point higher than in rural

[16] Note that Montenegro and Patrinos estimate the Mincerian function solely for the sample of wage earners in each country. Given the high share of workers involved in subsistence farming or informal self-employment, this implies a highly selective sample in many developing country data sets. Modeling this selection process would be an obvious path to pursue in explaining cross-country variation in addition to the hypotheses discussed here.

areas on average (8.3% versus 7.4%). But when restricting ourselves to the urban sub-sample, the variance across countries goes up, not down (and likewise when limiting ourselves to the rural sub-sample). Furthermore, the direction of the gap in returns between urban and rural areas is highly variant. In 25% of cases, the return to education is *higher* in rural areas, frustrating any attempt to use within-country patterns of heterogeneous treatment effects to explain cross-country heterogeneity in returns.

Clearly we could estimate a richer model here, with many other possible determinants of within-sample heterogeneity in the Mincerian return. But this example illustrates that even for an economic model that has been studied *ad nauseum*, we are not currently in a strong position to combine theory and empirics to make externally valid claims about parameters' magnitude. If you want to know the return to schooling in country X, there is no reliable substitute for data from country X.

Finally, we note that all the estimates discussed here are estimated by OLS using observational data and thus subject to structural bias. How concerned should we be about this bias? Interestingly, current academic consensus is that OLS and credible IV estimates of Mincerian returns differ very little. Evidence from the U.S. (Card, 2001) and Indonesia (Duflo, 2001) suggests OLS coefficients in a simple Mincerian specification are surprisingly good predictors of the best estimates produced by clever natural experiments. To echo the conclusions of the previous section, the current state of the literature appears to suggest that external validity is a much greater threat to making accurate estimates of policy-relevant parameters than is structural bias undermining the internal validity of parameters derived from observational data.

So far we have focused on the simple fact that true causal parameters vary across contexts. We have assumed, implicitly, that while experimental estimates of treatment effects may – just like non-experimental estimates – lack external validity, in the aggregate the accumulation of more and more studies should be converging to the average global effect. There are strong reasons to believe this is not in fact the case, to which we now turn.

6.3 Non-random placement of RCTs

The movement towards greater (if not exclusive) use of randomized evidence is the current stop on one track of the debate about "identification" in economics, a debate that goes back to the Cowles Commission in the 1950s. There is a logic to exclusive reliance on randomized empirics, with several premises. First, one's primary concern is about internal validity of the study for identifying causal impact. Second, a view that economic models of agent behavior cannot provide empirical identification (e.g. there are no sufficiently compelling exclusion restrictions from existing theory). Third, any, even potential, flaw in identification and hence internal validity is completely fatal to the contribution of a study.

However, the downplaying of the need for economics in empirics has perhaps led to a blind spot in thinking through the economics of randomized studies themselves. That is, since randomized studies of actual projects and programs (as opposed to "field experiments" implemented by the researchers themselves) require the active cooperation of those in authority over the projects or programs this ought to immediately lead to the question, "What are the incentives that would lead these policymakers or project managers to allow independent rigorous evaluation of their activities?" After all, allowing outside evaluation is risky as it may reveal that something the advocates want to happen is ineffective.

This raises the possibility that not only is there no external validity of the typical randomized study but also that the typical randomized study is *systematically* biased. That is, the estimate of the total causal impact from context j might not come from a random sample of all possible contexts to estimate the total causal impact. If there is any association between places/persons/organizations that choose to do randomized evaluations and RCT estimated causal impact then RCT evidence is biased (a bias which could run either way, towards more positive or more negative findings–depending on the agents responsible and their views).

This means that assessing evidence between non-experimental and experimental requires assessing the trade-off between the potential gains to internal validity from RCTs and the potential losses from a bias in external validity from the small and potentially non-random scope of RCTs.

* * *

There is an obvious intellectual puzzle reconciling the three claims that: (a) RCT techniques for policy evaluation have been well known for decades, (b) RCTs help organizations become more effective by providing rigorous evidence about what works to achieve their objectives, and (c) completed RCTs of policy are (extremely) rare even in domains that appear amenable to the method. The fact that RCTs are rare raises the concern that RCT estimates of causal impact might come from atypical contexts. Without a positive model of organizational engagement in, and completion of , RCTs there is no evidence against the claim that RCT results are atypical because, say, more effective organizations are likely to undertake and complete RCTs or because advocacy movements are more likely to promote and complete RCTs in contexts with larger results.

For instance, Duflo, Dupas, and Kremer (2012) showed in a field experiment in the Busia region of Kenya that the introduction of contract teachers improved test scores, whereas hiring additional civil service teachers did not (as discussed above). As the authors note, this suggests that the causal impact of class size on student learning is contingent on the context in complex ways – not only do Duflo et al.'s estimates differ markedly from similar estimates for the U.S., but reductions in class size due to hiring contract and civil service teacher produce very different effects. Bold, Kimenyi, Mwabu, Ng'ang'a, Sandefur, et al. (2013) replicate the Duflo, Dupas, and Kremer (2012) contract teacher program across the other provinces of Kenya. Schools in the replication study were randomly assigned to participate in a contract teacher managed by an NGO, World Vision Kenya, while others were assigned to an erstwhile identical program managed by the Kenyan Ministry of Education. When implemented by an NGO, the contract teacher produced very similar results to those in Duflo, Dupas, and Kremer (2012), but when the program was implemented by the Ministry of Education it had exactly zero impact on student learning. Hence the "context" of implementing organization determined the rigorous estimate of causal impact. Suppose organizations with stronger drive for performance or generally stronger capability are both more likely to undertake and complete RCTs and causal impact of the same program is larger when implemented by a strong and/or performance driven organization. In that case existing RCTs will be a biased sample of contexts.

The other possibility is that contexts are purposively chosen. A controlled experiment was carried out to investigate the impact of increasing the supply of contraceptives on fertility in the Matlab district of Bangladesh beginning in October 1977 (Phillips, Stinson, Bhatia, Rahman, and Chakraborty, 1982; Schultz, 2009). The experiment found the program reduced the total fertility rate (TFR) by about 25 percent (about 1.25 births per woman). This "rigorous" evidence of "what works" was widely touted – but essentially never replicated in other settings. Cross-national evidence had a very difficult time finding any causal role of family planning effort or contraceptive services (Pritchett, 1994). There are four reasons to believe that the Matlab district of Bangladesh setting would be one in which the impact of a program of increased supply of contraceptive services accomplished though female health workers bringing contraception directly to the household for free would be contextual large: (a) the region was isolated from markets and hence transport costs were high, (b) women in this region typically did not leave their household compound without a male escort, (c) women were very poor so that even small costs of contraception were large relative to budgets (and hence not concealable), and (d) culturally the use of contraception was not the norm so the "demonstration effect" of being encouraged to do so by high status females might create an aura of cultural acceptability in an otherwise hostile context. The point is that, having found an RCT result that was to their liking the family planning advocacy movement had no interesting in funding and fielding another study.[17]

Another example comes from the literature on energy efficiency programs. Allcott and Mullainathan (2012) study fourteen randomized trials of energy conservation programs conducted across the U.S. involving over half a million households. They examine the characteristics of power companies who did – and who did not – endogenously select into participation in the RCTs, showing that participating implementing partners had characteristics that are significantly correlated with larger treatment effects. This implies RCT estimates in this literature systematically overstate the population parameter of interest.

Allcott and Mullainathan (2012) also find suggestive evidence of the same phenomenon – which they term partner selection bias – in the microcredit industry. Brigham, Findley, Matthias, Petrey, and Nelson (2013) study a similar problem in the microfinance industry, by mailing out invitations to participate in an evaluation to 1,419 microfinance institutions, but randomizing the invitation to include reference to either positive or negative past findings about microfinance's effectiveness. Brigham, Findley, Matthias, Petrey, and Nelson (2013) find that the optimistic pitch yields a response rate that is twice as high.

To end on a positive note, there is evidence that RCTs may help to ameliorate another factor which drives a wedge between research findings and the true distribution of underlying

[17] The alternative explanation for variation in observed fertility rates is variation in the desired number of children, driven by changes in economic opportunity and social norms. These explanations are not mutually exclusive, and obviously need not apply in fixed proportion across diverse contexts. It is noteworthy that some more recent studies of large-scale natural experiments in contraceptive access have found considerable effects in other contexts. Salas (2013) finds a statistically significant relationship between short-term fluctuations in contraceptive supply and fertility in the Philippines, particularly among rural, less-educated women, though it is difficult to quantify the magnitude of these facts relative to other studies. Pop-Eleches (2010) finds a large, 25 to 30 percent, increase in fertility associated with restrictions on access to contraceptives under pro-natalist policies in Romania from 1966 to 1989 – but this natural experiment measures the effect of a legal prohibition on both family planning and abortion under a totalitarian state. It is no criticism of the Pop-Eleches (2010) study to question whether this result has relevance to marginal increases in contraceptive supply in low-income countries where they are already legally provided and promoted.

causal parameters: publication bias. For instance Brodeur, Le, Sangnier, and Zylberberg (2012) examine 50,000 published hypothesis tests in top economics journals from 2005 to 2011. Looking at the distribution of p-values, they find a large number of 'missing' test statistics, just above the traditional 0.10 threshold for statistical significance, and excess mass just below the lower threshold of .05. Interestingly, this evidence of publication bias (or data mining) is considerably weaker in RCT studies. One explanation is that editors and referees may be more willing to publish non-results from experimental work.

7 Conclusion

Our point here is not to argue against any well-founded generalization of research findings, nor against the use of experimental methods. Both are central pillars of scientific research. As a means of quantifying the impact of a given development project, or measuring the underlying causal parameter of a clearly-specified economic model, field experiments provide unquestioned advantages over observational studies.

But the popularity of RCTs in development economics stems largely from the claim that they provide a guide to making "evidence-based" policy decisions. In the vast majority of cases, policy recommendations based on experimental results hinge not only on the interior validity of the treatment effect estimates, but also on their external validity across contexts.

Inasmuch as development economics is a worthwhile, independent field of study – rather than a purely parasitic form of regional studies, applying the lessons of rich-country economies to poorer settings – its central conceit is that development is different. The economic, social, and institutional systems of poor countries operate differently than in rich countries in ways that are sufficiently fundamental to require different models and different data.

It is difficult if not impossible to adjudicate the external validity of an individual experimental result in isolation. But experimental results do not exist in a vacuum. On many development policy questions, the literature as a whole – i.e., the combination of experimental and non-experimental results across multiple contexts – collectively invalidate any claim of external validity for any individual experimental result.

We end with a set of suggestions for researchers conducting experimental studies in development economics. These suggestions are aspirational, and we would not claim that all good studies should achieve all of them – or that we have always done so in our own work.

1. Specify the context. This should include details on the institutional context of the intervention, and how it might or might not differ from real-world policy implementation.

2. Focus on attempting to explain, rather than "trump" the existing literature. Knowledge is advanced by explaining all known facts, and any study claiming to external validity in another context should attempt to explain the known facts about that context. Well-identified experimental results can help adjudicate between competing theories that might encompass existing evidence. They cannot override existing evidence.

3. Report OLS estimates of $\tilde{\beta}$, as well as the δ parameters related to selection into treatment.

To a great extent, concerns about context and external validity are not the concern of researchers engaged in an individual study, but rather should be given greater emphasis by consumers of experimental research. We advocate the following:

1. Avoid strict rankings of evidence. These can be highly misleading. At a minimum, evidence rankings must acknowledge a steep trade-off between internal and external validity. We are wary of the trend toward meta-analyses or "systematic reviews" in development, as currently sponsored by organizations like DFID and 3ie. In many cases, the transplantation of meta-analysis techniques from medicine and the natural sciences presupposes the existence of a single set of universal underlying parameters, subject to the same type of conditional invariance laws discussed in Section 6.1.

2. Learn to live without external validity. Experimentation is indispensable to finding new solutions to development challenges Rodrik (2008). But the "context" which defines the scope of the internal validity of an experimental study is complex and (as yet) unknown in its effects: geographic setting, time period, as well as the organizational or institutional setting of the treatment, to name a few. Moreover, the "design space" of any class of interventions (e.g. "microcredit" or "remedial instruction" or "pre-natal preventive care") of which any given intervention or treatment is an element is also hyper-dimensional Pritchett, Samji, and Hammer (2012)). Given the evidence of significant heterogeneity in underlying causal parameters presented here, experimental methods in development economics appear better suited to contribute to

a process of project management, evaluation, and refinement – rather than a elusive quest to zero in on a single set of universally true parameters or universally effective programs. But integrating rigorous evidence into the ongoing organizational implementation and learning processes requires an approach that is lower cost per finding, has faster feedback loops, is more integrated in decision making cycles than the way in which "independent impact evaluation" has traditionally been conceived.

References

Allcott, H., and S. Mullainathan (2012): "External validity and partner selection bias," National Bureau of Economic Research.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," Journal of Political Economy, 113(1), 151–184.

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer (2002): "Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment," American Economic Review, pp. 1535–1558.

Angrist, J. D., and V. Lavy (1999): "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," The Quarterly Journal of Economics, 114(2), 533–575.

Arif, G. M., and N. us Saqib (2003): "Production of cognitive and life skills in public, private, and NGO schools in Pakistan," The Pakistan Development Review, 42(1), 1–28.

Aronow, P. M., and C. Samii (2013): "Does Regression Produce Representative Estimates of Causal Effects?," EPSA 2013 Annual General Conference Paper, 585.

Asadullah, M. N. (2005): "The effect of class size on student achievement: evidence from Bangladesh," Applied Economics Letters, 12(4), 217–221.

Aslam, M. (2003): "The Determinants of Student Achievement in Government and Private Schools in Pakistan," The Pakistan Development Review, 42(4), pp–841.

Bacolod, M. P., and J. L. Tobias (2006): "Schools, school quality and achievement growth: Evidence from the Philippines," Economics of Education Review, 25(6), 619–632.

Banerjee, A., and R. He (2008): "Making Aid Work," in Reinventing Foreign Aid, ed. by W. Easterly. M.I.T.

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden (2007): "Remedying education: Evidence from two randomized experiments in India," The Quarterly Journal of Economics, 122(3), 1235–1264.

Bedi, A. S., and J. H. Marshall (1999): "School attendance and student achievement: Evidence from rural Honduras," *Economic Development and Cultural Change*, 47(3), 657–682.

——— (2002): "Primary school attendance in Honduras," *Journal of Development Economics*, 69(1), 129–153.

Behrman, J. R., S. Khan, D. Ross, and R. Sabot (1997): "School quality and cognitive achievement production: A case study for rural Pakistan," *Economics of Education Review*, 16(2), 127–142.

Blair, G., R. Iyengar, and J. Shapiro (2013): "Where Policy Experiments are Conducted in Economics and Political Science: The Missing Autocracies," mimeo.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, J. Sandefur, et al. (2013): "Scaling-up what works: experimental evidence on external validity in Kenyan education," Center for Global Development Working Paper.

Bold, T., M. Kimenyi, G. Mwabu, and J. Sandefur (2013): "The High Return to Private Schooling in a Low-Income Country. Africa Growth Initiative.," Brookings Institution Working PaperArif2003.

Brigham, M., M. Findley, W. Matthias, C. Petrey, and D. Nelson (2013): "Aversion to Learning in Development? A Global Field Experiment on Microfinance Institutions," Technical report, Brigham Young University.

Brodeur, A., M. Le, M. Sangnier, and Y. Zylberberg (2012): "Star Wars: The Empirics Strike Back," Paris School of Economics Working Paper No. 2012-29.

Brown, P. H., and A. Park (2002): "Education and poverty in rural China," *Economics of Education Review*, 21(6), 523–541.

Card, D. (2001): "Estimate the return to schooling: Progress on some persistent econometric problems," *Econometrica*, 69(5), 1127–1160.

Cerdan-Infantes, P., and C. Vermeersch (2007): "More time is better: An evaluation of the full time school program in Uruguay," World Bank Policy Research Working Paper No. 4167.

Chin, A. (2005): "Can redistributing teachers across schools raise educational attainment? Evidence from Operation Blackboard in India," *Journal of Development Economics*, 78(2), 384–405.

Clemens, M., C. Montenegro, and L. Pritchett (2008): "The place premium: wage differences for identical workers across the US border," World Bank Policy Research Working Paper No. 4671.

Cox, D., and E. Jimenez (1991): "The relative effectiveness of private and public schools: Evidence from two developing countries," *Journal of Development Economics*, 34(1), 99–121.

Crook, R. C., and D. Booth (2011): "Working with the Grain? Rethinking African Governance," *IDS Bulletin*, 42(2).

Das, J., K. Shaines, S. Srinivasan, and Q.-T. Do (2009): "U.S. and Them: The Geography of Academic Research," World Bank Policy Research Working Paper No. 5152.

Deaton, A. (2010): "Instruments, randomization, and learning about development," *Journal of Economic Literature*, pp. 424–455.

Du, Y., and Y. Hu (2008): "Student academic performance and the allocation of school resources: Results from a survey of junior secondary schools," *Chinese Education & Society*, 41(5), 8–20.

Duflo, E. (2001): "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review*, 91(4), 795–813.

Duflo, E., P. Dupas, and M. Kremer (2012): "School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools," National Bureau of Economic Research.

Easterly, W. (2006): *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. Oxford University Press.

Engin-Demir, C. (2009): "Factors influencing the academic achievement of the Turkish urban poor," *International Journal of Educational Development*, 29(1), 17–29.

Fronius, T., A. Petrosino, and C. Morgan (2012): "What is the evidence of the impact of vouchers (or other similar subsidies for private education) on access to education for poor people?," London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Glewwe, P., M. Grosh, H. Jacoby, and M. Lockheed (1995): "An eclectic approach to estimating the determinants of achievement in Jamaican primary education," The World Bank Economic Review, 9(2), 231–258.

Glewwe, P. W., E. A. Hanushek, S. D. Humpage, and R. Ravina (2011): "School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010," National Bureau of Economic Research.

Gomes-Neto, J. B., and E. A. Hanushek (1994): "Causes and consequences of grade repetition: Evidence from Brazil," Economic Development and Cultural Change, 43(1), 117–148.

Hanushek, E. A., and J. A. Luque (2003): "Efficiency and equity in schools around the world," Economics of Education Review, 22(5), 481–502.

Harada, M. (2013): "Generalized Sensitivity Analysis and Its Application to Quasi-Experiments," NYU Working Paper.

Hartman, E., R. Grieve, and J. S. Sekhon (2010): "From SATE to PATT: The essential role of placebo test combining experimental and observational studies," in The Annual Meeting of the American Political Science Association, Washington DC.

Hsieh, C.-T., and M. Urquiola (2006): "The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program," Journal of Public Economics, 90(8), 1477–1503.

Imbens, G. W. (2003): "Sensitivity to Exogeneity Assumptions in Program Evaluation," American Economic Review, 93(2), 126–132.

Institute of Education Sciences (2008): What Works Clearinghouse: Procedures and Standards Handbook (Version 2.0). U.S. Department of Education.

Khan, S. R., and D. Kiefer (2007): "Educational production functions for rural Pakistan: A comparative institutional analysis," Education Economics, 15(3), 327–342.

King, E., C. Montenegro, and P. Orazem (2010): "Economic freedom, human rights, and the returns to human capital: an evaluation of the Schultz hypothesis," World Bank Policy Research Working Paper No. 5405.

Krueger, A. B. (1999): "Experimental estimates of education production functions," Quarterly Journal of Economics, 114(2), 497–532.

Lee, V. E., and M. E. Lockheed (1990): "The effects of single-sex schooling on achievement and attitudes in Nigeria," Comparative Education Review, 34(2), 209–231.

Marshall, J. H. (2009): "School quality and learning gains in rural Guatemala," Economics of Education Review, 28(2), 207–216.

Marshall, J. H., U. Chinna, P. Nessay, U. N. Hok, V. Savoeun, S. Tinon, and M. Veasna (2009): "Student achievement and education policy in a period of rapid expansion: Assessment data evidence from Cambodia," International Review of Education, 55(4), 393–413.

Michaelowa, K. (2001): "Primary education quality in francophone Sub-Saharan Africa: Determinants of learning achievement and efficiency considerations," World Development, 29(10), 1699–1716.

Montenegro, C., and H. Patrinos (2013): "Returns to Schooling Around the World," Background paper for the World Bank World Development Report 2013.

Nannyonjo, H. (2007): Education inputs in Uganda: An analysis of factors influencing learning achievement in grade six. World Bank Publications.

Phillips, J. F., W. Stinson, S. Bhatia, M. Rahman, and J. Chakraborty (1982): "Estimate the return to schooling: Progress on some persistent econometric problems," Studies in Family Planning, 13(5), 131–140.

Pop-Eleches, C. (2010): "The Supply of Birth Control Methods, Education, and Fertility Evidence from Romania," Journal of Human Resources, 45(4), 971–997.

Pritchett, L., S. Samji, and J. Hammer (2012): "It's All About MeE: Using Structured Experiential Learning (e) to Crawl the Design Space," Center for Global Development Working Paper No. 322.

Pritchett, L., M. Woolcock, and M. Andrews (2012): "Looking Like a State: Techniques of Persistent Failure in State Capability for Implementation," Center for International Development at Harvard University, Working Paper No. 239.

Pritchett, L. H. (1994): "Desired fertility and the impact of population policies," Population and Development Review, pp. 1–55.

Psacharopoulos, G., C. Rojas, and E. Velez (1993): "Achievement Evaluation of Colombia's" Escuela Nueva": Is Multigrade the Answer?," Comparative Education Review, 37(3), 263–276.

Rodrik, D. (2008): "The New Development Economics: We Shall Experiment, but How Shall We Learn?," HKS Working Paper Series No.RWP08-055.

Rubin, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology, 66(5), 688–701.

Salas, J. I. (2013): "Consequences of withdrawal: Free condoms and birth rates in the Philippines," .

Schultz, T. P. (2009): "How Does Family Planning Promote Development? : Evidence from a Social Experiment in Matlab, Bangladesh, 1977 1996," Presented at Population Association of America meetings Detroit MI, April 30, 2009.

Scott, J. C. (1998): Seeing like a state: How certain schemes to improve the human condition have failed. Yale University Press.

Stock, J. H., and M. Yogo (2005): "Testing for Weak Instruments in Linear IV Regression," National Bureau of Economic Research.

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

Suryadarma, D., A. Suryahadi, S. Sumarto, and F. H. Rogers (2006): "Improving student performance in public primary schools in developing countries: Evidence from Indonesia," Education Economics, 14(4), 401–429.

Tabarrok, A. (2013): "Private Education in India: A Novel Test of Cream Skimming," Contemporary Economic Policy, 31(1), 1–12.

Tamer, E., and B. Kline (2011): “Using Observational vs. Randomized Controlled Trial Data to Learn About Treatment Effects,” Northwestern University - Department of Economics http://bit.ly/15KRVT1.

Urquiola, M. (2006): “Identifying class size effects in developing countries: Evidence from rural Bolivia,” Review of Economics and Statistics, 88(1), 171–177.

Urquiola, M., and E. Verhoogen (2009): “Class-size caps, sorting, and the regression-discontinuity design,” The American Economic Review, 99(1), 179–215.

Warwick, D. P., and H. Jatoi (1994): “Teacher gender and student achievement in Pakistan,” Comparative Education Review, 38(3), 377–399.

Wöβmann, L. (2005): “Educational production in East Asia: The impact of family background and schooling policies on student performance,” German Economic Review, 6(3), 331–353.

Wöβmann, L., and M. R. West (2006): “Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS,” European Economic Review, 50(3), 695–736.

Yu, G., and S. M. Thomas (2008): “Exploring school effects across southern and eastern African school systems and in Tanzania,” Assessment in Education: Principles, Policy & Practice, 15(3), 283–305.

Table 1: Class Size Effects

		Point estimate & std. error			
			Controlling	Controlling	
		No	for observable	for unobservable	Identification
	Country	Controls	characteristics	characteristics	strategy
Angrist & Lavy (1999)	Israel	0.322	0.019	-0.261	RDD
		(.039)	(.044)	(.113)***	
Urquiola (2006)	Bolivia	0.07	0.01	-0.21	RDD
		(0.03)**	-0.03	(0.07)**	
Asadulla (2005)	Bangladesh		0.25	3.5	RDD
			(0.115)***	(1.03)***	
Krueger (1999)	USA			-0.271	RCT
				(.072)***	
Banerjee et al (2007)	India	0.027[A]		0.064	RCT
		(.0125)**		(.118)	
Duflo et al (2012)	Kenya			-0.064	RCT
				(.024)**	

Note that point estimates and standard errors may differ from those reported in the original studies. For comparability, all results are expressed as the effect of a 10-pupil increase in class size measured in standard deviations of the test score. Where estimates are provided for multiple subjects or grade levels from the same study, we have opted for effects on math at the lowest grade level available.

[A] Authors' calculation based on public data release.

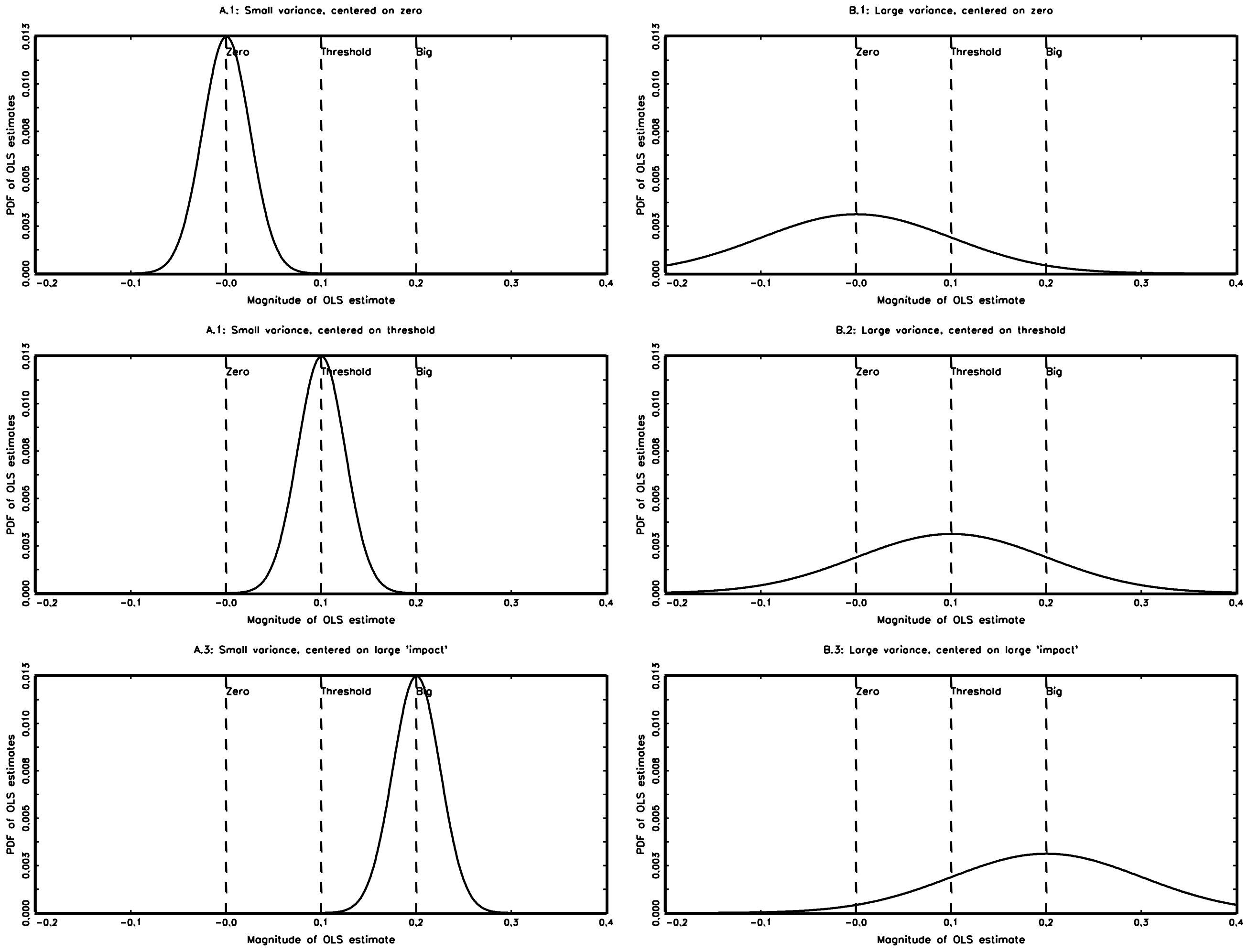
Table 2: The Return to Private Schooling

			Point estimate & std. error			
	Country	Note	No Controls	Controlling for observable characteristics	Controlling for unobservable characteristics	Identification strategy
Angrist et al (2002)	Colombia	(Won lottery)			0.205 (0.108)***	RCT (ITT)
Angrist et al (2002)	Colombia	(Ever used a scholarship)			0.291 (0.153)**	RCT (IV)
Altonji et al (2005)	USA	(Reading)	0.529^{A} (0.225)	0.091 (0.171)		Bounds
Altonji et al (2005)	USA	(Math)	.448 (.123)	0.183 (.073)		Bounds
Cox & Jimenez (1991)	Colombia		0.22	0.55		Selection correction
Cox & Jimenez (1991)	Tanzania		-0.14	0.97		Selection correction
Hsieh & Urquiola (2006)	Chile	(Language)	-0.571 (1.190)		-0.357 (1.381)	Aggregation
Hsieh & Urquiola (2006)	Chile	(Math)	-0.714 (1.188)		-0.51 (1.390)	Aggregation
Tabarrok (2013)	India	(Reading)			0.269 (0.059)***	Aggregation
Tabarrok (2013)	India	(Math)			0.224 (0.036)***	Aggregation
Bold et al (2012)	Kenya	(Reading & Math)	0.79 (0.046)***		0.98 (0.41)**	Aggregation

Note that point estimates and standard errors may differ from those reported in the original studies. For comparability, all results are expressed as the effect of binary change from public to private schooling, measured in standard deviations of the test score. For studies that use aggregate data, we consider a shift from a 0% to 100% private enrollment share. Where estimates are provided for multiple subjects or grade levels from the same study, we have opted for effects on math at the lowest grade level available.

[A] Altonji et al. (2005) report that test scores among Catholic high school students were roughly 0.4 standard deviations higher than those of public school students. This figure was used in combination with the score differences reported in Table 1 to calculate standardized coefficients.

Figure 1: Hypothetical distributions of OLS, non-experimental estimates prior to any RCT evidence



A.1: Small variance, centered on zero

B.1: Large variance, centered on zero

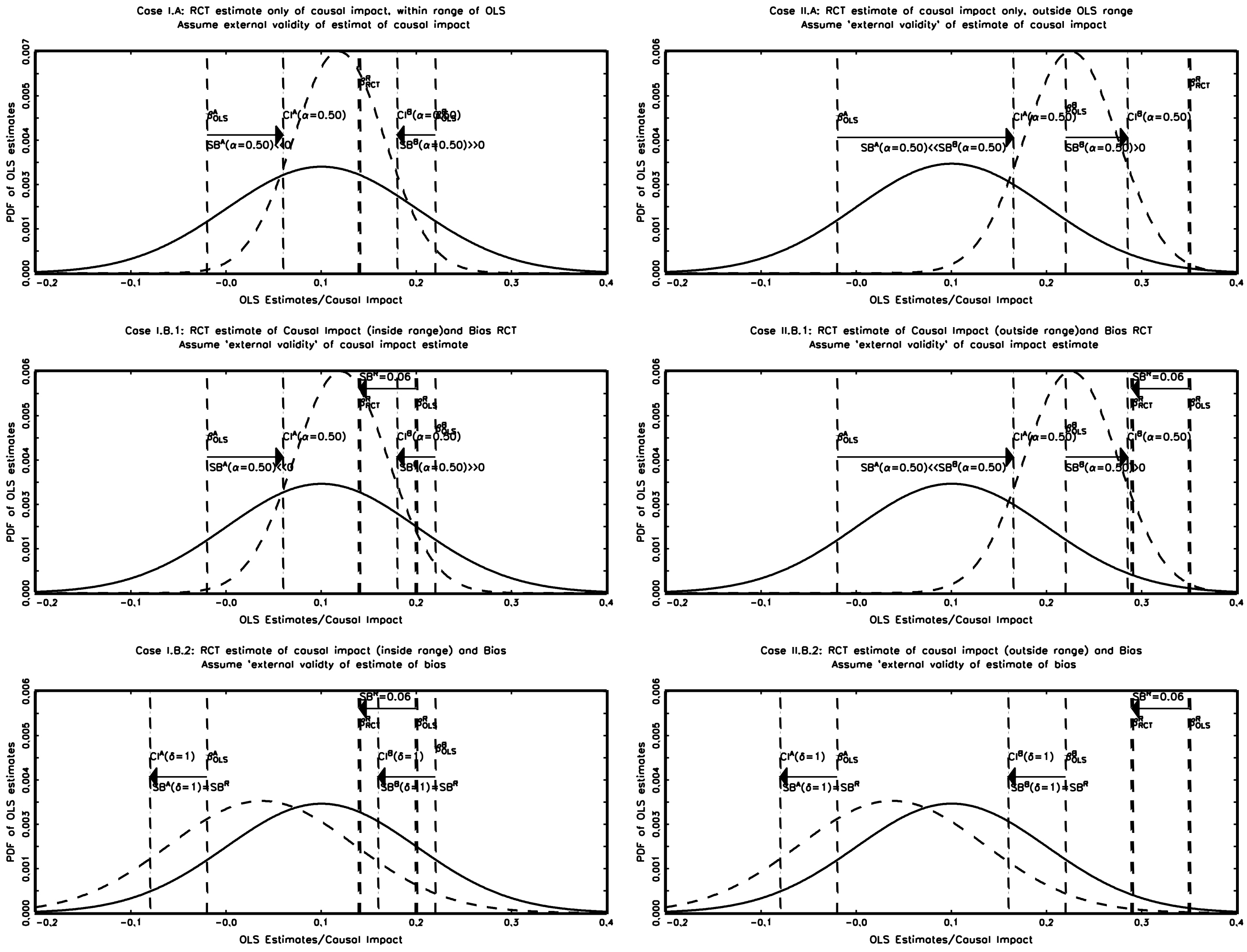
A.1: Small variance, centered on threshold

B.2: Large variance, centered on threshold

A.3: Small variance, centered on large 'impact'

B.3: Large variance, centered on large 'impact'

Figure 2: Any use of RCT evidence from one context to infer causal impact in another context when contextually varying non-experimental estimates already exist suffers from the underlying logical incoherence of selectively asserting the external validity of some, but not all, parameters



Case I.A: RCT estimate only of causal impact, within range of OLS
Assume external validity of estimat of causal impact

Case II.A: RCT estimate of causal impact only, outside OLS range
Assume 'external validity' of estimate of causal impact

Case I.B.1: RCT estimate of Causal Impact (inside range)and Bias RCT
Assume 'external validity' of causal impact estimate

Case II.B.1: RCT estimate of Causal Impact (outside range)and Bias RCT
Assume 'external validity' of causal impact estimate

Case I.B.2: RCT estimate of causal impact (inside range) and Bias
Assume 'external validty of estimate of bias

Case II.B.2: RCT estimate of causal impact (outside range) and Bias
Assume 'external validty of estimate of bias

Figure 3: External Validity with Parameter Heterogeneity

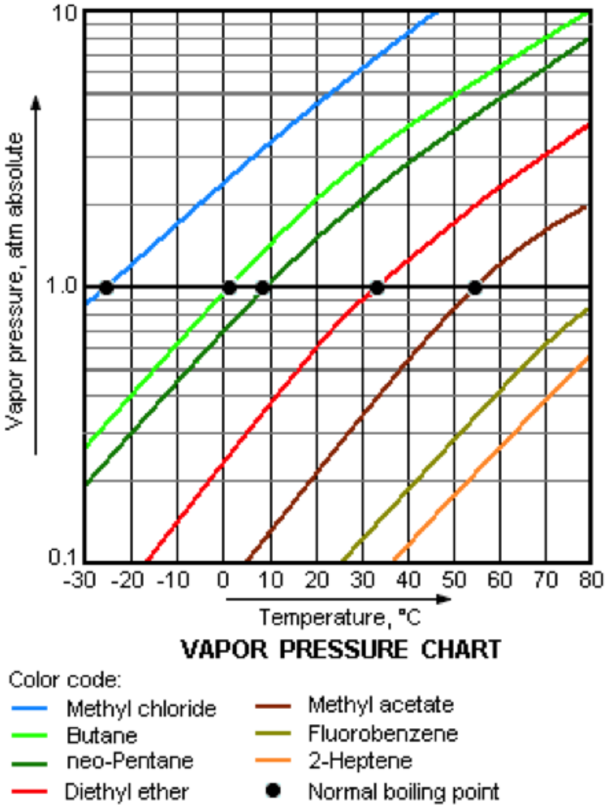
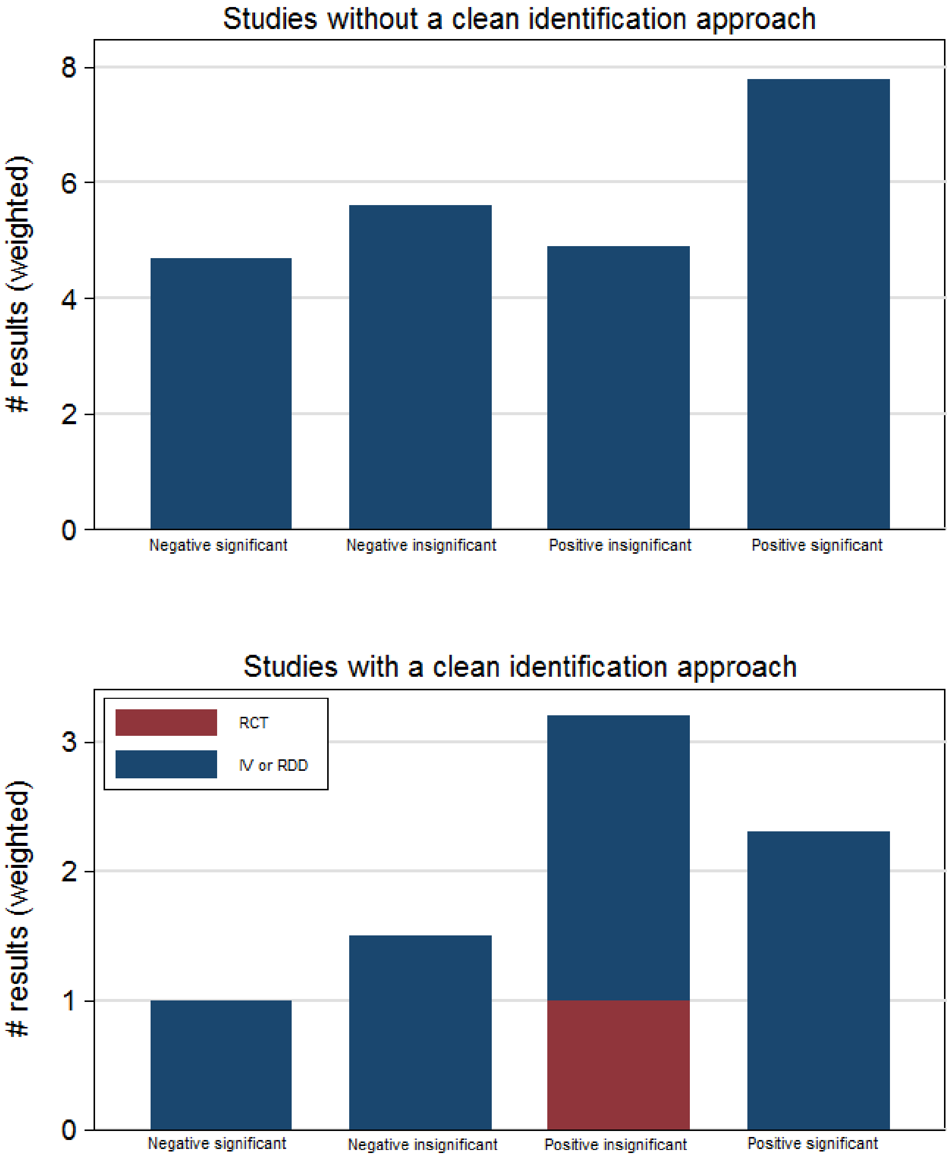


Figure 4: Class Size Effects, Multiple Studies



Studies without a clean identification approach

Studies with a clean identification approach

Figure 5: Class Size Effects, Woessman & West (2006) TIMSS data

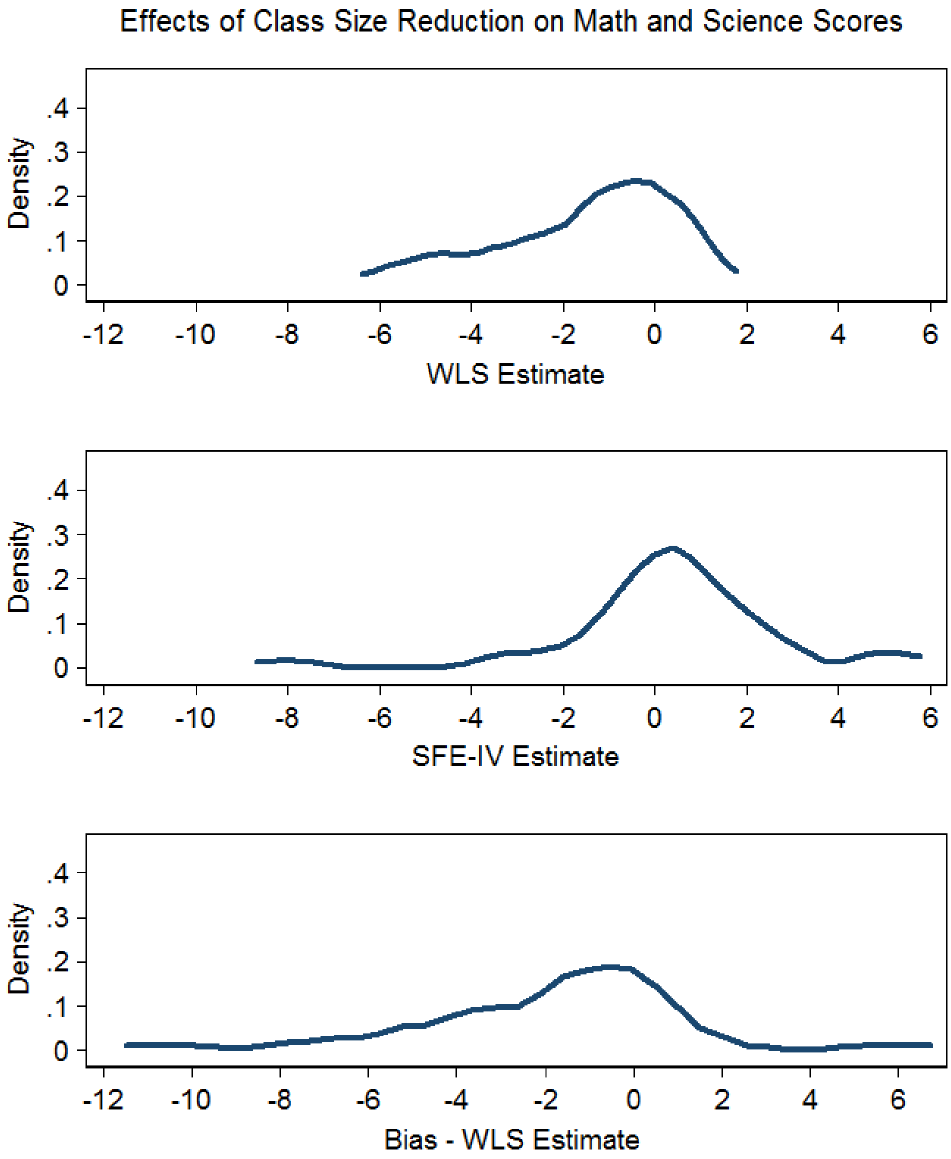
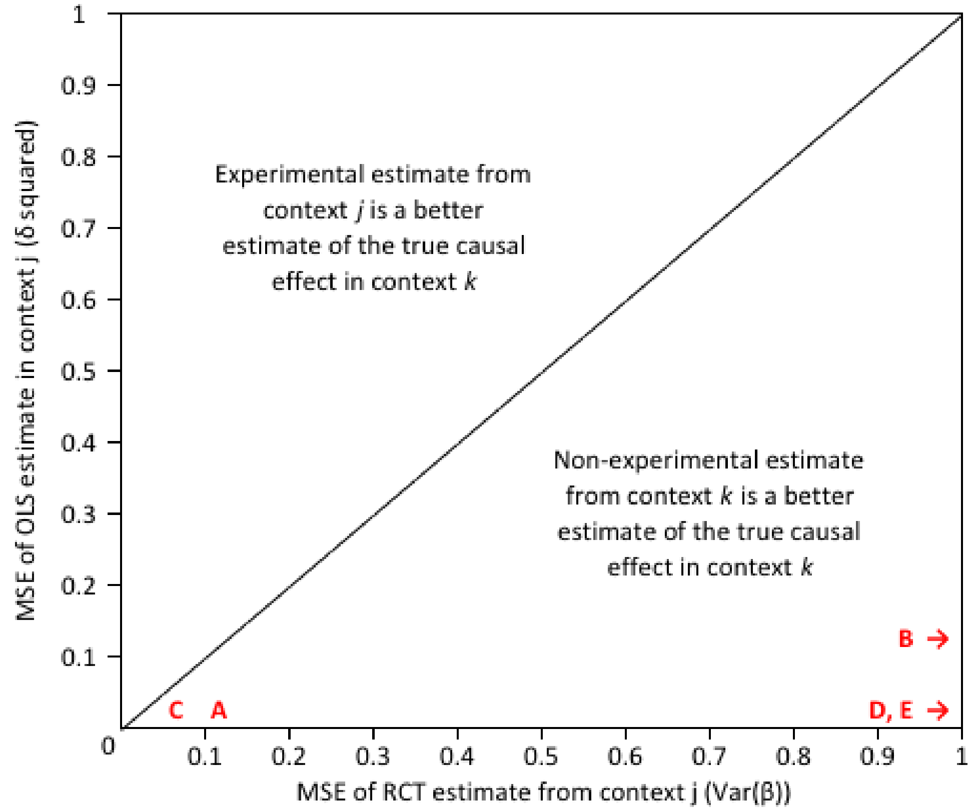
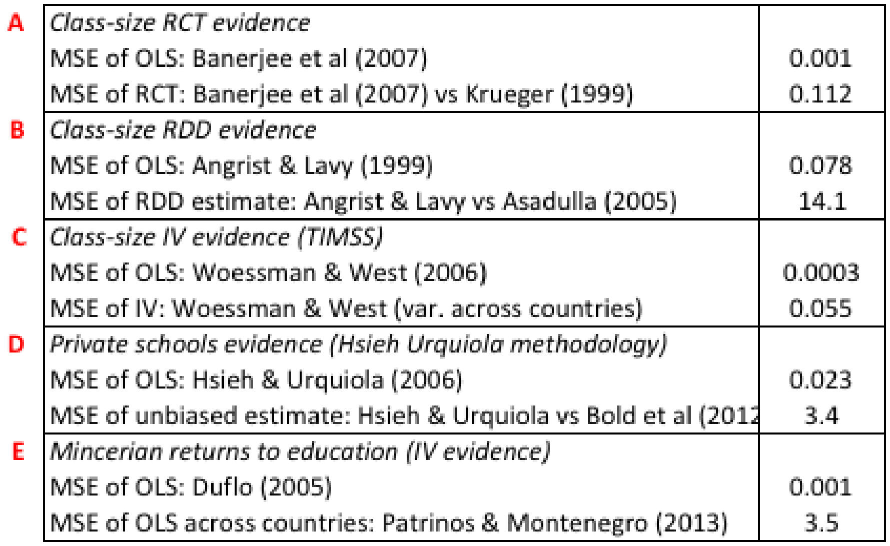


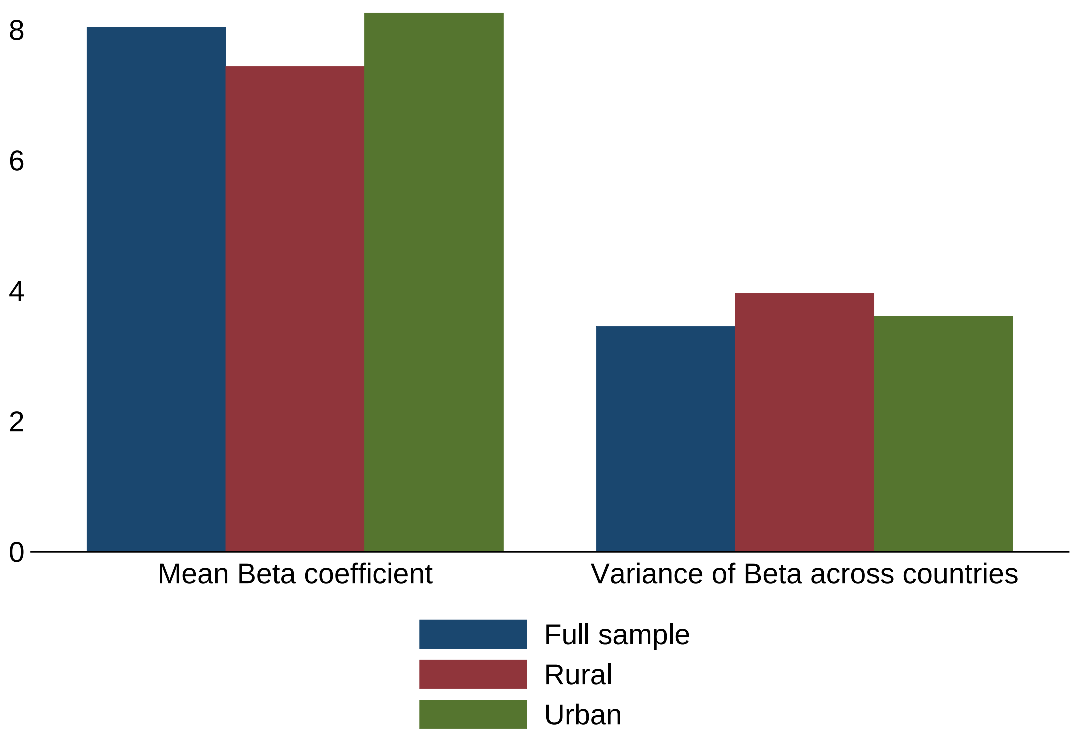
Figure 6: Quantifying the risks associated with limited internal and external validity of treatment effect estimates: Mean-squared error (MSE) of experimental, IV, or RDD estimates from a different country vs observational OLS estimates from the right country.





A	*Class-size RCT evidence*	
	MSE of OLS: Banerjee et al (2007)	0.001
	MSE of RCT: Banerjee et al (2007) vs Krueger (1999)	0.112
B	*Class-size RDD evidence*	
	MSE of OLS: Angrist & Lavy (1999)	0.078
	MSE of RDD estimate: Angrist & Lavy vs Asadulla (2005)	14.1
C	*Class-size IV evidence (TIMSS)*	
	MSE of OLS: Woessman & West (2006)	0.0003
	MSE of IV: Woessman & West (var. across countries)	0.055
D	*Private schools evidence (Hsieh Urquiola methodology)*	
	MSE of OLS: Hsieh & Urquiola (2006)	0.023
	MSE of unbiased estimate: Hsieh & Urquiola vs Bold et al (2012	3.4
E	*Mincerian returns to education (IV evidence)*	
	MSE of OLS: Duflo (2005)	0.001
	MSE of OLS across countries: Patrinos & Montenegro (2013)	3.5

Figure 7: Mincerian returns to schooling across 128 surveys


Mean Beta coefficient
Variance of Beta across countries
Full sample
Rural
Urban