Let's Take the Con Out of Randomized Control Trials in Development: The Puzzles and Paradoxes of External Validity, Empirically Illustrated

Lant Pritchett

Oxford Blavatnik School of Government

May 1, 2021

Abstract. The enthusiasm for the potential of RCTs in development rests in part on the assumption that the use of the rigorous evidence that emerges from an RCT (or from a small set of studies identified as rigorous in a "systematic" review) leads to the adoption of more effective policies, programs or projects. However, the supposed benefits of using rigorous evidence for "evidence based" policy making depend critically on the extent to which there is external validity. If estimates of causal impact or treatment effects that have internal validity (are unbiased) in one context (where the relevant "context" could be country, region, implementing organization, complementary policies, initial conditions, etc.) cannot be applied to another context then applying evidence that is rigorous in one context may actually *reduce* predictive accuracy in other contexts relative to simple evidence from that context—even if that evidence is biased (Pritchett and Sandefur 2015). Using empirical estimates from a large number of developing countries of the difference in student learning in public and private schools (just as one potential policy application) I show that commonly made assumptions about external validity are, in the face of the actual observed heterogeneity across contexts, both logically incoherent and empirically unhelpful. Logically incoherent, in that it is impossible to reconcile general claims about external validity of rigorous estimates of causal impact and the heterogeneity of the raw facts about differentials. Empirically unhelpful in that using a single (or small set) of rigorous estimates to apply to all other actually leads to a *larger* root mean square error of prediction of the "true" causal impact across contexts than just using the estimates from non-experimental data from each country. In the data about private and public schools, under plausible assumptions, an exclusive reliance on the rigorous evidence has RMSE *three times* worse than using the biased OLS result from each context. In making policy decisions one needs to rely on an *understanding* of the relevant phenomena that encompasses *all* of the available evidence.

Let’s Take the Con out of Randomized Control Trials in Development:

The Puzzles and Paradoxes of External Validity, Empirically Illustrated

Introduction

Ed Leamer’s (1983) nearly 40 year old and justly famous paper “Let’s take the con out of econometrics” is worth (re)reading for four reasons. One, it will remind all who need it that economists have understood the pros and cons of inferences from experimental versus non-experimental data for a very long time. Two, he emphasizes that “all knowledge is human belief; more accurately human opinion” hence the purpose of theory and models and empirical evidence is to form more reliable opinions. Third, he proposes adding two words to the applied econometrics lexicon: “whimsy” and “fragility.” He argues the conventions at the time for reporting of non-experimental results without extensive sensitivity analysis of the fragility of the results with respect to variation in the assumptions made about model specification led much of existing published econometrics to be just whimsy. Fourth, he acknowledges that what is and isn’t accepted as evidence (and of what degree of persuasiveness) is a human convention that is specific to a community, which could be a discipline (or sub-discipline) or field or people engaged in a particular practice.

This paper updates Leamer’s classic to address a new issue[1]. In the sub-discipline of development economics there is a proposed new set of conventions about evidence. This convention privileges certain types of evidence as “rigorous” (with the paradigm, though not exclusive, example of “rigorous” being evidence from a Randomized Control Trial (RCT)). This convention also (though mostly implicitly) proposes that “rigorous” evidence from one context (or set of contexts) can, and should be, be applied to other contexts, perhaps though via a “systematic review.” This new convention disparages all other types and forms of evidence, often “*the* evidence” means exclusively the “rigorous” evidence (sparse as that may be) and proposes “systematic” reviews that exclude all but the “rigorous” evidence[2].

[1] As such, and since Leamer’s paper was originally a speech, I am going to retain something of the informal and at times sardonic tone of the original.

[2] A corollary to the desire to rely on rigorous evidence is to bias the research agenda in development economics towards those questions amenable to RCT techniques. This biases research away from questions of “national development” and towards questions amenable to individualizable treatments, independently of their likely empirical *importance* (something explicitly embraced in a strong ideological commitment to the “small” in Banerjee and Duflo 2011). For instance, this biases studies towards anti-poverty programs and their impacts (e.g. Banerjee et al 2015 about “graduation” type programs, Banerjee et al 2015 about micro-credit, Bastagli et al 2107 about cash transfers) in spite of the fact that (i) it is well known that very nearly all variation across countries in headcount poverty outcomes is associated with the level and growth of typical incomes in a country (Dollar et al 2016 , Pritchett 2020), (ii) there was never any argument made that these programs were major determinants of poverty status either across countries or across households within countries. However, these programs were amenable to RCTs as they affected individuals/households and hence allowed treatment effects to be identified and studies to reach sufficient statistical power (even though, as David McKenzie (2020) suggests, if you need a power calculation is probably doesn’t matter (that much) for poverty). I think the single most important critique of the RCT movement is that it does not particularly help with the questions most important for improving human

In previous papers (Pritchett and Sandefur 2014, 2015) we have made the point that, when non-experimental estimates show large variability across countries, external validity is impossible[3]. Moreover, I think that after Vivalt (2020) it is widely accepted that rigorous estimates of the impact of the same class of programs have large variability and hence most people have abandoned the possibility of external validity of causal impact. This paper's claim is *not*: "rigorous estimates of the LATE of classes or types of interventions lack external validity" as, in 2021, that might characterized, fairly, as attacking a straw man.

The point of this paper is that *given* the lack of external validity across contexts of LATE estimates of interventions (policies/programs/projects) the case for the use of "rigorous" evidence outside of its context in forming predictions about the causal impact of an intervention in a specific context must be defended *empirically* (not rhetorically, not ideologically, not theoretically) based on an assessment of *facts* about the relative magnitudes of the various sources of prediction error. I precent an empirical example in which the proposed standard: "Use the (optimally) weighted average of the rigorous LATE estimates from a systematic review as the point estimate for causal impact in your context" does empirically *worse* in RMSE (root mean square error) than just ignoring the rigorous evidence entirely and using the biased non-experimental evidence from each context. Using data on private and public school learning outcomes I show that, under plausible assumptions, the RMSE of the "rely on the rigorous evidence" approach is *three times* worse than the naïve approach of using context specific OLS.

I present just one example but I argue this example is important for five reasons. First, the slogan "rely on the rigorous evidence" is often treated as a truism or a theorem. A single counter-example disproves a theorem. Any *general* assertion that one should "rely on the rigorous evidence" is just pure con. Second, this example illustrates (literally, I illustrate this with graphs) the relevant empirical considerations, which is the magnitude of the bias in non-experimental estimates versus the variability across contexts of the "true" LATE. If the true LATE variability across contexts is large then even using an estimate of the LATE that is correct on average across all contexts can lead to higher prediction error than using estimates that vary across contexts, even if every one of those estimates is biased. Third, the score is now 0-3 *against* rigorous evidence. The present calculations repeat calculations done in Pritchett and

well being in developing countries (Pritchett 2020) and is just a symptom of the larger shift within the field of development away from national development towards "kinky" development (Pritchett and Kenny 2013, Pritchett 2015, Pritchett 2021). This paper however brackets this important issue of the bias in the questions being studied and just examines whether, *even for those limited questions for which RCTs are viable*, the approach produces reliable evidence for decision making.

[3] Impossible for the simple reason that non-experimental estimates are the result of model structures and parameters that determine the true causal impact and that model structures and parameters that determine the bias of any given non-experimental estimate. Therefore if there is large variability in the non-experimental estimates across contexts then either the model structure/parameters of causal impact or of the processes that generate bias must vary across contexts and hence there cannot be external validity of both. But claims of external validity of estimates of causal impact both (a) has no logical or empirical basis and (b) leads to absurd implications about the model structure/parameters that generate bias in non-experimental estimates.

Sandefur (2015) that show OLS was better than "rigorous" evidence for the impact of micro-credit and this is also true of estimates of the impact of class size. The present example is another counter-example against a claim or theorem for which there are, as yet, no empirical examples. Any correct statement about the conditions about when and how one should "rely on the empirical evidence" is going to be conditional on various empirical magnitudes and at this stage we have no idea if "rely on the rigorous evidence" in the strong sense isn't only rarely (if ever) an empirically justified recommendation. Four, while I may be accused of attacking an unrealistic and strong interpretation of "rely on rigorous evidence," once one acknowledges that the strong version is indefensible and one moves to a reasonable "balance the rigorous evidence from other contexts with all other evidence" view much of the rhetorical allure of the of the superiority of "rigorous" evidence and "doing science" and "labs" versus the practitioners more pragmatic forming of judgments under uncertainty ("craft" and "metis") is gone. Fifth, without defensible claims about external validity there is no way to defend the value-for-money or cost-effectiveness of doing RCTs, the overwhelming majority of which are not evaluations "at scale" (Muralidharan and Niehaus 2017) and hence have to justify their worth by application of their findings beyond the context in which the RCT was conducted.

Decision makers should rely on their *understanding* and their *understanding* should *encompass* all of the evidence. I am not proposing that the "rigorous" evidence should be ignored but only that its relevance to the actual decisions in the specific context has to be assessed. To propose that people making decisions should 'rely on evidence" in a way not mediated by their own understanding of the reasons, explanations, and causes is not what "evidence based" decision making implies.

I) The proposed new conventions for "evidence": "Clean sweep" plus external validity of only causal impacts

Speak to us only with the killer's tongue,
The animal madness of the fierce and young
Conrad Aiken, Sursum Corda

A large set of development policy relevant questions can be framed as "If policy/program/project A(Ψ) (where A() is a mapping from 'states of the world' (Ψ) to specific actions a) were implemented what would be the impact on (a set of) outcome measures Y?" This is a necessary input into any *ex ante* evaluation of a policy/program/project as it describes the feasible vectors of netputs used in, say, a cost benefit or rate of return analysis based on the costs of the actions A and the benefits to an objective function of gains in outcome measures Y[4].

[4] The classical theory of cost-benefit analysis in the context of evaluation public sector actions (e.g. Dreze and Stern 1987) was primarily concerned with valuation issues (e.g. discount rates, shadow prices, distributional concerns) as public sector projects frequently produce outputs without market prices and/or are undertaken in the face of market prices that do not reflect social marginal costs or benefits. In their classic treatment Dreze and Sen (1987) state: "We shall not be concerned with assessing the feasibility of projects, but rather with appraising the desirability of a priori specified, and presumably feasible,

Choices may hinge on the distribution of beliefs/opinions about the Local Average Treatment Effect (LATE), the *average*[5] gain in outcomes Y from the policy/program/project (state contingent actions $A(\ \psi)$) in a given context z.

$$1)\ f(LATE(A(\ \psi)^{z})$$

f is a *distribution* of beliefs about the LATE and hence, as a distribution, has (at least) a central tendency and a variability. Decisions may depend not just on some simple rule like: "do A if the mean of *f* is above some threshold value" but may take into account the confidence in positive returns like: "do A if the 20th percentile of *f* is above some critical value."

I.A) The powerful case for RCTs

Leamer (1983) characterizes our uncertainty about estimates of regression coefficients (α,β) as consisting of two matrices, the sampling covariance matrix, S, and the misspecification matrix, M, which is the covariance matrix of the bias parameters of the regression coefficients for the "true" parameters.

$2\ (eqn\ 4\ from\ Leamer)\ var(\hat{\alpha},\hat{\beta})$ = S+M

The sampling variance can be reduced with larger samples but M, the misspecification covariance matrix, can be independent of sample size. If we suspect there *might* be bias but have no prior knowledge or beliefs about the sign of the bias then the expected value of the regression parameters may not be changed by the suspicion of bias but the *variance* is going to be larger than the reported estimated precisions of estimation, S. The main argument of Leamer (1983) is that the then accepted convention in the economics literature of reporting only S (the regression standard errors and t-tests and what not) while ignoring the uncertainty about the reliability of estimates from the misspecification matrix M (or dealing with this misspecification uncertainty in *ad hoc* ways) was the "con" built into the then standard conventions about econometrics. Leamer makes the case for experimental approaches. With a correctly designed experiment the misspecification covariance matrix M can be driven to zero.

There are three very different ways of dealing with the uncertainty about estimating parameters or causal effects from potential misspecification errors.

One is an empirically driven *robustness* approach. This limits the fragility of estimates by reporting estimates over a wide range of specifications of the estimating equation using non-experimental data (e.g. adding various other co-variates to the regression, exploring other

projects." The assumption was that developing country governments had many more projects known to be feasible (and hence effective) than resources/capability to carry them out and hence the challenge was allocating resources to the best projects on some consistent valuation.

[5] I focus on the average because it is the set-up in which the case for RCTs is strongest. It is well known that RCTs do not generate unbiased estimates of anything about the treatment impact but the average (Heckman 2020, Cartwright and Deaton 2018). But this is a severe limitation as in many instances there is interest in the distribution of the treatment effect, for instance, whether a program that has a mean gain of 100 dollars across 100 HHs is the result of each HH gaining 1 dollar or 99 households losing a dollar and one household gaining 199 dollars is obviously important.

estimation techniques like instrumental variables (IV), using different dimensions of covariation such as using fixed effects to estimate coefficients with only the "within" variation, using synthetic controls, regression discontinuity techniques, etc.).

Two, (and somewhat related) is what I call the *understanding* approach. This seeks to improve our theories and models and understanding of the underlying determinants of the outcome Y. This improved understanding of the phenomena with can be used either to (i) "sign and bound" the magnitude of the bias from estimates of the impact of X on Y with non-experimental data or (ii) add knowledge generated from RCTs. In this way M, the magnitude of the uncertainty about misspecification bias, is limited because we believe we have a correct specification that adequately controls for the known determinants of Y other than X and/or a model of the conditions for the non-experimental variation in X (of the type that would provide adequate instruments or estimates) and hence can recover consistent estimates of the causal impact/treatment effect of X on Y.

I argue the "hard" sciences and their applications generally take the "understanding" approach. They work with a fundamentally correct and empirically validated theory/model of the phenomena at hand, a model within which observational and experimental evidence from all sources is encompassed. Astronomy is a non-experimental hard science. In measuring Hubble's constant astronomers use methods that depend on a commonly agreed theory and that theory is used to interpret non-experimental observations to create estimates of model dependent quantities. These in turn are fitted into a larger model that attempts to encompass all of the known facts, such as the ΛCDM (Lambda Cold Dark Matter) approach.   This approach does not necessarily lead to expected results (such as the recent discovery the pace of expansion of the universe was accelerating[6]) or a complete or immediate consensus, as is illustrated by what is currently called the Hubble Tension, that different approaches to measuring the Hubble Constant are producing mutually incompatible estimates (the standard errors bounds produced by each method do not include the estimates from other methods).

RCTs can be used within the *understanding* approach, but they are neither the whole of, nor particularly central, to the approach. For instance, many of the techniques of RCTs were derived from field trials in agriculture (e.g. the work of R.A. Fisher (1926, 1935), Neyman (1923)) and hence a commonly used example in discussions of RCTs is the application of fertilizer (Leamer 1983). But the idea of applying fertilizer (and even what is considered a "fertilizer") comes from an understanding of the processes of plant growth at the chemical, cellular and plant level and a model of the interaction of plants with the soil conditions within the science of agronomy. This combination of correct theory combined with both clinical and field experimentation allows the creation of a reliable "handbook" type knowledge of the optimal fertilizer treatment to apply given the exact circumstances of soil and the existing nutrients in it, moisture, stage of plant growth, etc.[7] While field trials contributed to the science of agronomy,

[6] Kirshner's (2002) book *The Extravagant Universe* is an excellent and readable account of how empirical, non-experimental, science can produce surprising new results.

[7] A (distant) relative of mine made his living in Idaho for a period by providing other farmers in the area a service that during the growing season he (his firm) would provide *daily* soil testing on their fields at

the science of agronomy builds on an understanding of plant growth and is not a collection of reports of "what works" in agriculture.

The example of drug trials is frequently cited as an example of the successful application of RCTs. But drugs to be tested are typically derived from a correct and validated model of the chemistry and biology of the cell and how those operate functionally within specific organisms. The economists discussion of RCTs in drug testing often ignore that in the FDA process any of the four phases of clinical trials are part of Step 3 of the drug approval process, following Step 1 Drug Discovery and Development and Step 2 Preclinical Trials, which depends on a vast body of knowledge of chemistry and biology and an understanding of why a given substance is likely to have the predicted effects in a human being.

The use of RCTs in economics to examine social policy is at least 50 years old. This early use, like the negative income tax experiments, was designed to produce better estimates of underlying model dependent parameters, like the labor supply elasticity, with the idea these parameters could then be used to better *understand* the impact of non-labor sources of income on labor supply and that understanding (and empirical estimates) could be use in the evaluation of a wide range of possible policy designs (Heckman 2020).

The third way of reducing both bias and uncertainty in estimation from misspecification (M) is the *balance* approach. Experiments can be designed to achieve unbiased estimates of the LATE of X on Y by "balancing" all other factors that affect Y between treatment arms and control groups. An unbiased LATE can be estimated without any theory or model or understanding of the phenomena Y. The case for the "balance" model depends on two, related, skepticisms.

The first is a skepticism that "sign and bound" approach to bias can produce useful ranges on the uncertainty of estimates from non-experimental data. Estimating causal impacts using non-experimental data in the social sciences is especially hard for two reasons. One, non-experimental data are the result of purposive choices by agents who may act on information not available to the econometrician, which is not true of corn, or a cancer cell, or Jupiter. Two, the potential magnitude of the bias is related to the magnitudes of the correlations with the error term and many estimation models produce very low explanatory power and hence the magnitude of the phenomena that is "unexplained" by the estimation is very large and hence the potential magnitude of the bias is very large. Even if standard methods, like OLS, produce very precise estimates of the coefficient of X (a small S matrix) the interpretation of that coefficient as a parameter or causal effect depends on assumptions about the behavior of "unobserved" variables and hence if the explanatory power is low this can produce very large misspecification uncertainty M and hence make inferences very fragile with respect to assumptions.

multiple sites and provide *daily* fertilizer and watering plans tailored to their crop and those soil conditions. The farmers were willing to pay because of the savings from applying less than the "recommended" amounts of fertilizer from the suppliers of fertilizer, which, not surprisingly, tend to err on the side of over-generous use of fertilizer and, when amortized over the thousands of acres of the farms the cost savings from optimization to conditions exceeded my relative's fees.

The second skepticism is about the usefulness of theory or models (or at least existing theory and models) and hence skepticism about the very notion of the "structural parameters" suggested by models.

The case for randomization in development is often a case for a "balance" approach which is a skeptical stance that claims that M is (i) empirically very large and unlikely to be reduced by a robustness approach of collecting more and more variables to add to estimating equations and (ii) progress in *understanding* by refining theory and model to limit the magnitude of M is also unlikely to be successful.

I.B) The question of the application of evidence across contexts

A key question about the use of RCT evidence in development is how much the distribution of beliefs about the LATE of doing project/program/policy A(Ψ) in context z should change in response to evidence from RCTs (or other rigorous[8] methods) from other context(s) *c* (equation 3).

$$3)\ f\left(LATE\left(A(\Psi)\right)^{Z}\middle|E^{z}, LATE\left(A(\Psi)\right)^{c}\right) - f(LATE(A(\Psi))^{Z}|E^{z}, E^{-z})$$

People have, at least implicitly, some prior distribution of beliefs/opinions about the LATE of A in z. This distribution is based on the existing body of evidence from context z, E^z, and evidence from other contexts, E^{-z}. In this descriptive sense "evidence" includes everything on which beliefs or opinions are actually be formed, including non-experimental estimates from context z, non-experimental estimates from other contexts, theories and models, analogies and comparisons to other experiences from other sectors or domains. I am making no assumptions that any actual person's "prior" distribution of beliefs is fully rational or fully Bayesian or anything else.

Suppose in context, z, I am deciding on a program of (i) building schools and want to know the elasticity of enrollment with respect to distance (perhaps by groups, like for girls or children from poorer households), or (ii) reducing class size in middle schools and want to know the likely impact on learning or (iii) allowing "money follows the student" that would defray the costs of children attending (some set of) private schools and want to know the impact on learning or (iv) creating "school improvement plans" and want to know how schools will respond and how those responses with change student learning.

And suppose in each case I have an OLS estimate of the relevant "causal impact" parameter, using control set W from context z (e.g. an OLS regression of student enrollment on distance, or an OLS regression of student learning on class size, etc.). In addition, suppose there

[8] I am torn about whether or not to continue to use rigorous in quotes. On the one hand I would prefer to use the scare quotes to make it clear my use is reference, not use, of the word as I do not think there is any defensible clear line between evidence that is or is not rigorous and moreover, nearly all uses of rigorous evidence is not rigorous at all. Even less so a conflation of "rigorous" with "randomized" (Leamer (1983) points to this usage 40 years ago). On the other hand, it is just too pedantic so I will stop using scare quotes but with the shared understanding with the reader the word "rigorous" is reference not use.

are one or more RCTs done in other contexts c and a systematic review of that rigorous evidence produces an (optimally) weighted average of those causal impact estimates:

$$4)\ \beta_{CI}^{SR} = \sum_{c=1}^{C} w_c * \beta_{CI}^{c}$$

Suppose my belief is a weighted average of OLS and systematic review (SR) with weight (α_{SR}):

$$5)\ \beta_{CI}^{z} = (1-\alpha_{SR}) * \beta_{OLS|W}^{z} + \alpha_{SR} * \beta_{CI}^{SR}$$

A fair interpretation of “rely on the rigorous evidence” (RORE(1)) is that α_{SR} should be one:

$$6)\ \widehat{\beta_{CI}^{z}} = \beta_{CI}^{SR}$$

The standard framing the problem of combining different sources of evidence is to choose α_{SR} in order to minimize the RMSE (root mean square error) of prediction over all contexts z϶Z:

$$7)\ \begin{matrix} min \\ \alpha_{SR} \end{matrix} RMSE(\alpha_{SR}\ ) = sqrt\left(\frac{\sum_{z=1}^{N_z}\left(\beta_{CI}^{z,True} - \beta_{CI}^{z}\widehat{(\alpha_{SR})}\right)^2}{N_z}\right)$$

One might think that the “rely on the rigorous evidence” recommendation is based on some evidence that α_{RE}=1 is the optimal weight. But it isn’t.

The variance of the RORE(1) estimate in (somewhat abused) Leamer-like notation is:

$$8)\ var\left(\widehat{\beta_{CI}^{z}} = \beta_{CI}^{SR}\right) = S + M^{c} + M^{z,c}$$

If there is variability in the “true” causal impact across contexts then the variance of RORE(1) has to take into account the variance of applying evidence across contexts, $M^{z,RE(c)}$. The *variability* of the true causal impacts across contexts ($M^{z,RE}$) might be sufficiently large that, even if there is bias and OLS in z therefore lacks internal validity, the RMSE of RORE(1) is larger than ignoring the rigorous evidence completely (α_{SR}=0).

A simple analogy is helpful. Suppose men routinely lie about their height and these lies are nearly always upward. Hence, we know self-reported height lacks internal validity. But suppose (with IRB approval, of course) we sample men and discover both their self-reported and their true height. We would find that the true height of men in the USA is about 5’9’’ with a standard deviation of 3 inches. Suppose each man’s self-report adds 1 inch to his height then RMSE of reliance on the biased self-report is 1 inch. If rely on the mean of the rigorous evidence about men’s height to predict each man’s height, RORE(1),the RMSE is three times as big, 3 inches. If a man says he is 6’4’’ tall rejecting this self-report because self-report is internally biased and instead predicting his height is really rigorously measured true average value of 5’9’’ from a “systematic review” of true heights is just silly. Moreover, if the study had measured the average selectivity bias and predicted each man’s height as the self-report less the selectivity bias (and ignored the estimated average height altogether) the RMSE would be lower than either approach.

I.C) The proposed new convention about "evidence": Rely on rigorous evidence

The proposed new convention for evidence is that the distribution of beliefs about the LATE of A in context z conditional on all previous evidence from z and elsewhere plus the rigorous evidence from other contexts should be roughly the same as that based just on the rigorous evidence from context(s) c.

$$9)\ f(LATE(A)^z|E^z, E^{-z}, LATE(A)^c) \approx f(LATE(A)^z|LATE(A)^c)$$

When Leamer (1983) was arguing against the professional acceptance of the convention of ad hoc, whimsical, specification searches that produced fragile results it wasn't that people were *explicitly* making the case against the need for robustness analysis. Rather the practice of publishing papers with only S (the standard t-statistics, etc.) and ignoring M revealed the actual professional convention. Similarly the problem with RCT is that everything about the practices, publications, rhetoric, and slogans is consistent with a practice of just ignoring $M^{z,c}$ (the misspecification variance due from applying estimates across contexts). The proposed convention has two parts: (i) a "clean sweep" of previous evidence and (ii) reliance only on the rigorous estimates of casual impact, and (iii) assuming external validity of causal impact, not of selectivity bias.

I.C.1) The "clean sweep" approach to previous evidence

Bedecarrats, Guerin and Roubaud (2020) call the new approach to evidence a "clean sweep" approach: pretending previous evidence doesn't exist. This is revealed in three practices: (i) systematic reviews, (ii) using the word "the", and (iii) citations.

Systematic reviews. Systematic reviews consist of a systematic way of dredging up all of the relevant literature and then a filter applied to that body of work that systematically excludes any paper that doesn't meet some criteria for method. The rest of the review then completely, totally, ignores the rest of the evidence.

"The" evidence. Conventions are revealed in the way language is used and a common current usage is to use the definite article, "the" in talking about "the evidence" when what is being referred to is in reality the very narrow slice of the relevant evidence, which meets the speaker's criteria for rigorous.

Citation practices. Two examples. One, in a paper published in a top journal in economics Burde and Linden (2013) examined the impact of distance to a "community based" school using 13 treatment villages in one rural region of Afghanistan. Their review of the voluminous literature estimating the impact of proximity on school enrollment was a single footnote citing just two (!) papers[9].

[9] This example is particularly striking as their findings show that proximity to a school *does* matter for enrollment, which what literally *everyone* who works in or around education already believed and has

The second example is from the Bedecarrats et al (2020) paper which reviews a journal special issue touted as a review of “the evidence” about the impact of microcredit.

> *Randomistas’ results are often presented as unprecedented “discoveries,” whereas they are often only the replication of conclusions obtained from previous studies, primarily those obtained from non-experimental methods that are almost never cited (Labrousse 2010). The General Introduction is a good illustration of this. The results are presented as the first scientific evidence of the impacts of microcredit. “The evidentiary base for anointing microcredit was quite thin” (Banerjee, Karlan, and Zinman 2015: 1). Up to this point, available empirical evidence had been based on “anecdotes, descriptive statistics or impact studies that are unable to distinguish causality from correlation” (pp. 1–2). The authors claim to be part of “the debates that took place in the 2000s and continue today” (p. 2) but these debates are actually taking place in a surprisingly cloistered world. Of the 18 references in the General Introduction, 12 (two-thirds) come from the authors themselves and 17 (94.4 percent) from J-PAL members. Only one article escapes this endogamic principle.*
>
> *No non-randomized studies are cited. Looking at the six articles in the Special Issue, the article on Morocco is equally exclusive (only RCTs are mentioned).*

I.C.2) External validity of only causal impacts, ignoring heterogeneity of impacts

At a conference reviewing papers for a Handbook of Education Economics one of the authors made the case that since essentially all of the methodologically “best” evidence about the causal learning impact of private schools was from the United States and since (in his view) that evidence suggested all of the observed raw differences in student learning between public and private were due to selectivity and the causal impact was zero therefore “we” (development economists) should hold the belief that the causal impact is zero in all countries.

The case for exclusive reliance on the “best” (read: RCT) evidence from *any* context for inference about causal impacts for *all* contexts is almost never made in print so explicitly but is nevertheless the current belief and practice among many academics. Without an explicit answer to the question of “how should I weigh various sources of evidence in forming my distribution of beliefs in my context?” (equation 3) the default (if implicit) answer is “rely (exclusively) on rigorous evidence” (equation 9).

The clamor for “evidence-based” policy making is vacuous (if not fatuous) without clarity on what constitutes “evidence” and how to reconcile the various strands of “evidence.” For instance, there have been a number of systematic reviews of the evidence about the impacts of various actions in education[10]. All of these are intended to provide guidance to policy makers in their choices. The general practice is to show the averages of estimates across “types” or “classes” of interventions and perhaps some indication of the range. While there might be some

been a working premise of education policy in every country in the world for decades. The only way one could pretend this finding was “new” was to use “clean sweep” and feign ignorance.

[10] There have been so many reviews of “the evidence” of “what works” in education there is even a meta-review of the reviews—which shows, not surprisingly the “systematic reviews” don’t come to the same conclusions (Evans and Popova 2015).

lip service about applying these results to context, there is no explicit guidance as to how to do that. If the recommendation is that your beliefs should be 99 percent OLS from your context and 1 percent the evidence from systematic reviews of the rigorous evidence (and there is no reason *a priori* this could not be the optimal weight) then the value of the whole "do RCTs and then a systematic review" is *de minimis*. The implicit answer is that one should rely on "the" evidence in the way systematic reviews define it such that only "rigorous" evidence counts at all. But what is "rigorous" depends on views about external validity and evidence that is rigorous in one context is not rigorous when applied to another (which is, after all, the predominant use of "rigorous" evidence in development)[11].

Even if the rigorous evidence correctly estimates the average of the causal impacts across all contexts z, without a consideration of the variability of causal impacts across contexts there is no way of making any claims about how much "reliance" should be placed on the (average of the) rigorous estimates. The current practice is an asymmetric skepticism in which, even without any evidence presented to "sign and bound" the bias from lack of internal validity and its variance (M^z) it is assumed that the variability to the prediction error in context z from relying on evidence from other contexts ($M^{z,c}$ or $M^{z,RE}$) is low. This is a proposed language game of: "I get to doubt everything you say based on evidence that might lack internal validity that introduces some magnitude of bias and I get to choose to believe it might be really big, but you should believe what I say based on evidence that has internal validity in some other context and ignore the possibility that my evidence may be a first order correct but very high variability estimate."

I.C.3) …and ignoring estimates of selectivity bias

The third element of the proposed convention about evidence is that essentially all of the attention is given to the estimates of LATE whereas an RCT can generate evidence about both the causal impact and the selectivity bias and both of these are rigorous evidence. Since both the LATE and the selectivity bias emerge from some underlying model of (constrained) choices in a given context there is no possible way to choose *a priori* which of these two has greater "external validity" across contexts or, alternatively, which would lead to the lower prediction

[11] For instance, **Angrist, Joshua D. and Victor Lavy.** 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics*, 114(2), 533-75.Angrist and Lavy 1999 have a paper estimating the impact of class size on learning using variation induced by the Maimonides Rule that requires that classes not exceed a certain size which I am happy for argument's sake to call a rigorous estimate of class size in Israel. At a conference I heard a prominent randomista make the case the World Bank could fund class size reductions around the world because this paper was rigorous evidence of the impact of class size reductions on learning (I am not attributing this view about policy applicability to either Angrist or Lavy themselves). A few years back I found that of the first 150 (non-self) citations to this paper (Google Scholar sorts by the number of times the citing paper has itself been cited) *not one* mentioned Israel, the *only* context to which this evidence is arguably rigorous. In the title or abstract of the top 150 citations China, India, Bangladesh, Cambodia, Bolivia, UK, Wales, USA (various states and cities), Kenya and South Africa all figured. Angrist and Lavy 1999 is *not* rigorous evidence about class size impacts in any of these places (an assertion I am confident both its authors would agree with).

error. Suppose we modify equation 5 to allow the OLS estimate in context z to be adjusted for an estimate of the selectivity bias (γ) specific to the OLS estimate with conditioning variables W:

$$5. a)\ \beta^{z}_{CI} = (1 - \alpha_{SR}) * (\beta^{z}_{OLS|W} - \gamma^{z}_{OLS|W}) + \alpha_{SR} * \beta^{SR}_{CI}$$

Now suppose we assume that estimates of selectivity bias have external validity and hence we take the rigorous estimate of the selectivity bias from a systematic review of the estimates of selectivity bias to be our estimate of selectivity bias in context z:

$$5. a)\ \widehat{\gamma^{z}_{OLS|W}} = \gamma^{SR}$$

Which leads to the question of what is the optimal (say, RMSE minimizing) value of α_{SR} in equation 5.b:

$$5. b)\ \beta^{z}_{CI} = (1 - \alpha_{SR}) * (\beta^{z}_{OLS|W} - \gamma^{SR}) + \alpha_{SR} * \beta^{SR}_{CI}$$

Now RORE(α_{SR}=1) says: "Rely exclusively on rigorous evidence about causal impacts but ignore completely the equally rigorous evidence about selectivity bias."

It is clear that applying the seemingly simple slogan "rely on the rigorous evidence" implies one must consider: (i) the mean and variance (M^z) of the bias in the particular non-experimental evidence in context z, (ii) the mean (which might be zero) and the variance ($M^{z,c}_{CI}$) of the application of the rigorous evidence about causal impacts from context(s) c to z, and (iii) the mean (which might be zero) and the variance ($M^{z,c}_{SB}$) of the application of the rigorous evidence about selectivity bias from context(s) c to z.

In the simple analogy above of the bias in self-reported height if the mean is 1 inch and the standard deviation across men in self-reported bias is 1 inch then the RMSE of α_{SR}=1 is 3.27, the RMSE of α_{SR}=0 is less than half that, 1.41, the RMSE minimizing choice is (roughly) without adjustment for selectivity bias α_{SR}=.18 producing a RMSE of 1.28 and the RMSE with α_{SR}=0 (ignoring rigorously estimated average height altogether) but adjusting for the average bias in self-report is 1.01.

I.D) Conclusion of first section

The conclusion of this section is that, even in situations in which there is reason to believe non-experimental estimates have bias (lack internal validity) and hence there is a powerful case for RCTs, the case for "rely on the rigorous evidence" (RORE) interpreted as (i) ignore all evidence from context z that isn't rigorous and (ii) put all the weight on the rigorous estimates of causal impact and, hence, (iii) ignore the evidence about selectivity bias evidence, is both the currently conventionally accepted practice of RCTs (and other) plus systematic reviews that estimate average impacts of classes of interventions and is just completely unsupported by theory, empirics, or logic as a proposal for use in predicting causal impacts in development[12].

[12] It is perhaps worth pointing out, if only in a (long) footnote, that, while not tenable as a way of forming better predictions of causal impacts in development in a practical or pragmatic sense, one can see the immense attraction of the "clean sweep" and "balance" approach to RCTs plus the pretense of external

II) Facts (and the variance of facts across contexts) are evidence too: Estimates about the private sector learning premium

The empirical example I will use to illustrate the issues around external validity is the magnitude of the *causal* private school premium in learning[13]. Everyone agrees that the raw differences in assessed learning between students in private and public schools reflect both selection effects and (possibly) causal effects and hence that standard non-experimental estimates are biased and lack internal validity.

New estimates, from two completely different sources, of the private-public learning difference across a fairly large number of both 'developed' and 'developing' countries—or more neutrally, (old) OECD and non-OECD countries, make the present empirical illustration possible. Since I am a focused on the role of RCTs in generating evidence relevant for development I will focus (almost) exclusively on developing countries.

With the recent participation of the PISA-D (D for Development) countries, there are now 35 non-OECD countries with estimates of the average scores for public and private school students on Math, Reading, and Science. I combine these into a single estimate for each country of the raw private sector premium across the three subjects by dividing the score gap for each subject by its country/subject specific standard deviation to put the learning premium estimates into standard deviation units or "effect sizes" that are standard in the education literature[14].

A recent paper Patel and Sandefur (2019) (henceforth P-S) use gives a sample of children in Bihar India an assessment with and instrument that has questions from different global and regional assessments. This "Rosetta Stone" allows them to translate scores from each assessment into a common metric. They then use the data from the assessments to estimate the private premium/deficit for these countries (and more, see below). I average their math and reading estimates and divide by an assumed country standard deviation of the assessment of 90 points for all countries and subjects[15]. This produces another 30 non-OECD estimates of the raw

validity for young academics. First, "clean sweep" means all that tedious "literature review" and understanding of the current state of understanding in a field can be avoided, if "the" evidence means "evidence about causal impacts from RCTs" then almost whatever one does you can say "this is the first rigorous evidence on this subject" and act as if this is, in and of itself, a contribution to the field. Second, the "balance" approach means that the effort to understand theory can be avoided (which is good, because it is hard and complicated) because one doesn't need to actually understand the broader causes to just do X and see if Y changes. Third, the pretense that a key constraint to improving human well-being is lack of the kind of knowledge your RCT will generate and that your research will have practical application gives you some "warm glow" about "changing the world" in the dark cold night of being a graduate student. This is a partial answer to the question "If RCTs have so little value why has there been such a large expansion in their number?" They are demonstrably of substantial value in terms of career progression to the academics doing them.

[13] This adds to previous illustrations using class size impacts on learning (Pritchett and Sandefur 2014) and the impact of microcredit (Pritchett and Sandefur 2015).

[14] The scores and the observed premium are very highly correlated across the three subjects across countries so little is lost by aggregating the scores.

[15] Three countries (Bahrain, Indonesia, and Chile) have estimates only for math or reading and I use the one that is available.

private sector premium. Their estimates are generally for grade 4 versus the PISA which tests all children aged 15 in grade 7 or above.

Together, these two papers give a fairly large (but by no means either comprehensive or representative sample) coverage of developing countries.

II.A) Raw public-private sector differences

Figure 1 shows the box-plot of each of the two estimates (PISA and P-S) with their three letter country codes[16]. The overall median estimate of the raw private premium is about .6 standard deviations and is remarkably similar for both the PISA (.63) and Patel-Sandefur (.60) estimates. An impact of any intervention on learning of .1 to .2 sd is considered an impact worth reporting and a common estimate of the gain from a year of schooling is .3 to .4 sd units.

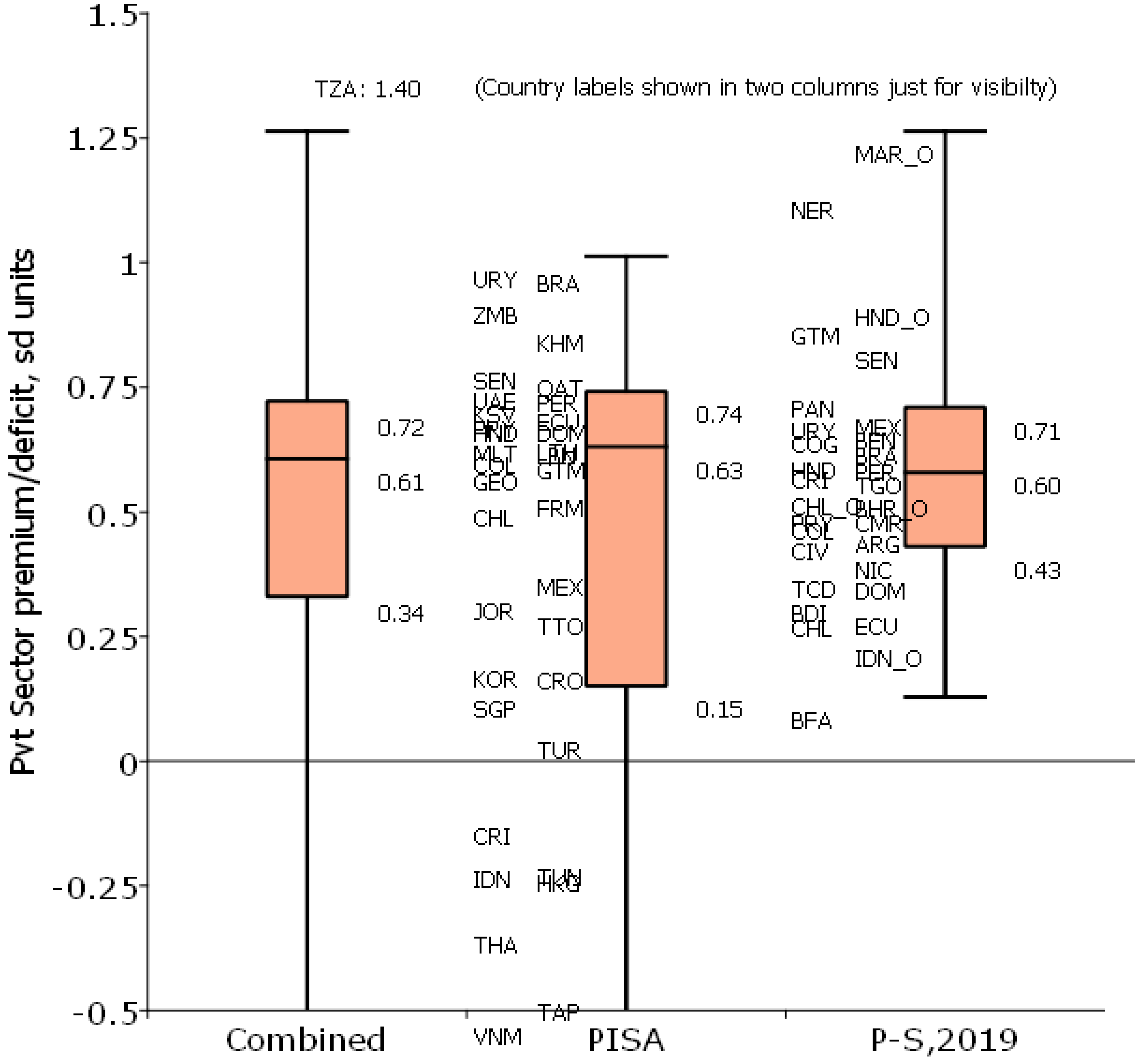
Figure 1 illustrates the striking *differences* across countries in the raw private sector premium. The raw private-public difference ranges from a one standard deviation or more positive difference (e.g. Brazil, Uruguay, Morocco, Niger) to some pretty substantial negative values, with private school students scoring *less* well than public school students, in seven countries (e.g. Indonesia, Tunisia, Thailand, Tapei, Vietnam). Across all countries the inter-quartile range (25th to 75th percentile) in the raw private sector premium is roughly .4 units (.72 to .34) (larger in PISA, .6, and smaller in P-S, .3 as the P-S data has no negative values).

These (estimates of) the raw differences in scores between private and public and the variation in these differences across countries are just facts. These facts are part of "the evidence" to be encompassed by any adequate understanding of the phenomena. One hard question is the causal interpretation of these facts: "*Why* is it that, in the typical non-OECD country, the operations of the education system(s) is such that the observed raw private sector learning advantage is around .6 sd?" The answer might be entirely, or in any part, selection effects.

Another hard question in understanding of private school impacts in developing countries is: "Why is it that the operation of the education systems across countries produces a large raw private sector premium in some countries, of modest size in others, and zero or negative in still others?" Again, the answer to the differences across countries could be variation across countries in either causal impacts or selection effects.

[16] In the P-S estimates some of the country estimates are using the TIMSS or PIRLS data directly whereas others use the estimates for the students from the Rosetta Stone adjustment of the original data set. Those countries with "_O" are those with "original" data and the rest (the majority of the developing country cases) are the result of the Rosetta Stone adjustment.

Figure 1: Heterogeneity in the raw private sector premium/deficit



Source: Author's calculations.

II.B) Private sector learning premium, adjusted for household SES

Everyone has always understood that it was common that children from higher socio-economic status (or just higher income/wealth) households were more likely to do well in school and that these children were also more likely to enroll in private school. Because the (budget) constrained choice of enrollment into private or public schools is correlated with determinants of learning outcomes, the raw difference in scores obviously lacks internal validity as an estimate of the LATE/causal impact on learning of enrolling a given child in a private versus private school.

Both the PISA and the P-S include an estimate of a private sector premium conditional on a measure of the student's household socio-economic status (SES)—and some other demographic characteristics. The standard PISA reports show the raw difference and the difference in each country "adjusted for" the PISA constructed index of Economic, Social and Cultural Status (ESCS) index, which is a combination of variables about household assets that affect child learning (e.g. access to books, computers), parental education levels, and parental occupation. P-S construct an asset index and adjust that for the income distribution of each country by a percentile method and estimate the private sector premium conditional on this HH asset/income index. These two SES indices are *not* comparable.

Figure 2 shows for each country *c* the (i) raw private sector premium ($\beta(\emptyset)_{CI}^{c}$), (ii) the premium adjusted for SES ($\beta(SES,Z)_{CI}^{c}$) and (iii) the magnitude of the adjustment of the premium based on selectivity on the observed SES indicators ($\beta(\emptyset)_{CI}^{c} - \beta(SES,Z)_{CI}^{c}$). Figure 2 illustrates four facts.

First, the adjustment for student HH SES reduces the mean/median estimate of the private sector premium substantially. The median raw private sector premium is .62 and the SES adjusted premium is .34, about .28 points lower, so the SES adjusted premium is roughly in half that of the raw.

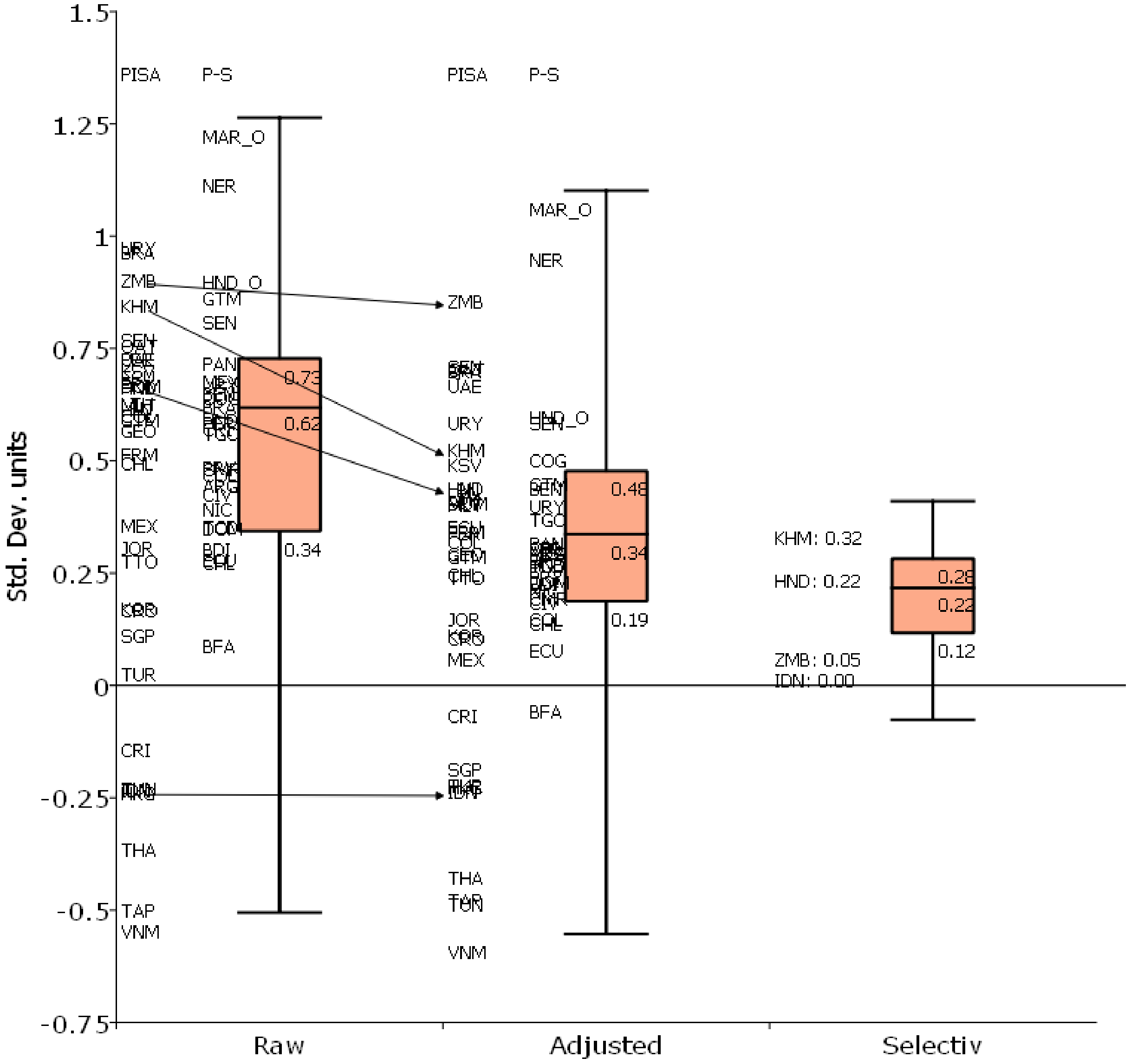
Second, the typical (median) SES adjusted private sector premium is still positive and large. The median is .34 standard deviations, which is a large "effect size" and roughly the equivalent of a year of schooling.

Third, the dispersion of the SES adjusted premium is large. The 25th-75th spread across countries from .19 to .48 is about .3 standard deviations, which implies the dispersion across countries is about the same size as the median.

Fourth, the magnitude of the adjustment for selectivity across countries on the two different SES variables is itself typically substantial (a median of .22).

Figure 2 uses four selected countries (all from PISA) as illustrations of the adjustment for SES. Honduras happens to have about the typical adjustment and so the ESCS adjustment moves a raw estimate of .70 to an ESCS adjusted private premium of .48, a raw to SES adjusted move of .22. The figure also shows Zambia (ZMB) and Cambodia (KHM) as their estimate of the raw premium is similar (.94 and .88) but the magnitude of the SES adjustment is very different, a much smaller adjustment than the typical country for Zambia (.05), making the ESCS adjusted estimate .89, whereas the adjustment is much larger than the typical country for Cambodia (.32) making the ESCS adjusted estimate only .56. For Indonesia, with a negative raw estimate (-.19), the adjustment for ESCS is basically zero (.002) hence the SES adjusted estimate is the same as the raw.

Figure 2: Private sector premium, Raw, Adjusted, and Adjustment



Source: Author's calculations with data from PISA and Patel and Sandefur (2019).

II.C) Correcting the estimate of causal impact for variables that are not observed and which create bias

The most sophisticated current approach to the long tradition of "sign and bound" is the Oster (2016) adjustment to estimate a lower bound on the estimate of the impact X[17]. The Oster adjustment is an estimate of what the unbiased estimate of the LATE would be, using

[17] The Oster adjustment extends the Altonji, Elder, and Taber 2005 adjustment, which was aimed at assessing the degree of selection bias in non-experimental estimates of the impact of Catholic schools in the USA.

assumptions about the unobserved variables. These assumptions are often made to produce a lower bound of what the coefficient on X would be under specified assumptions about the correlation of the unobserved variables with X and their importance in explaining Y.

Patel and Sandefur (2019) use the student level data from the assessments to estimate the Oster adjustment[18]. Figure 3 shows the raw, adjusted for SES (assets), and the Oster lower bound for the 32 countries in the P-S sample (this adds in Portugal and Denmark to the 30 non-OECD countries). The Oster bound estimates show three facts.

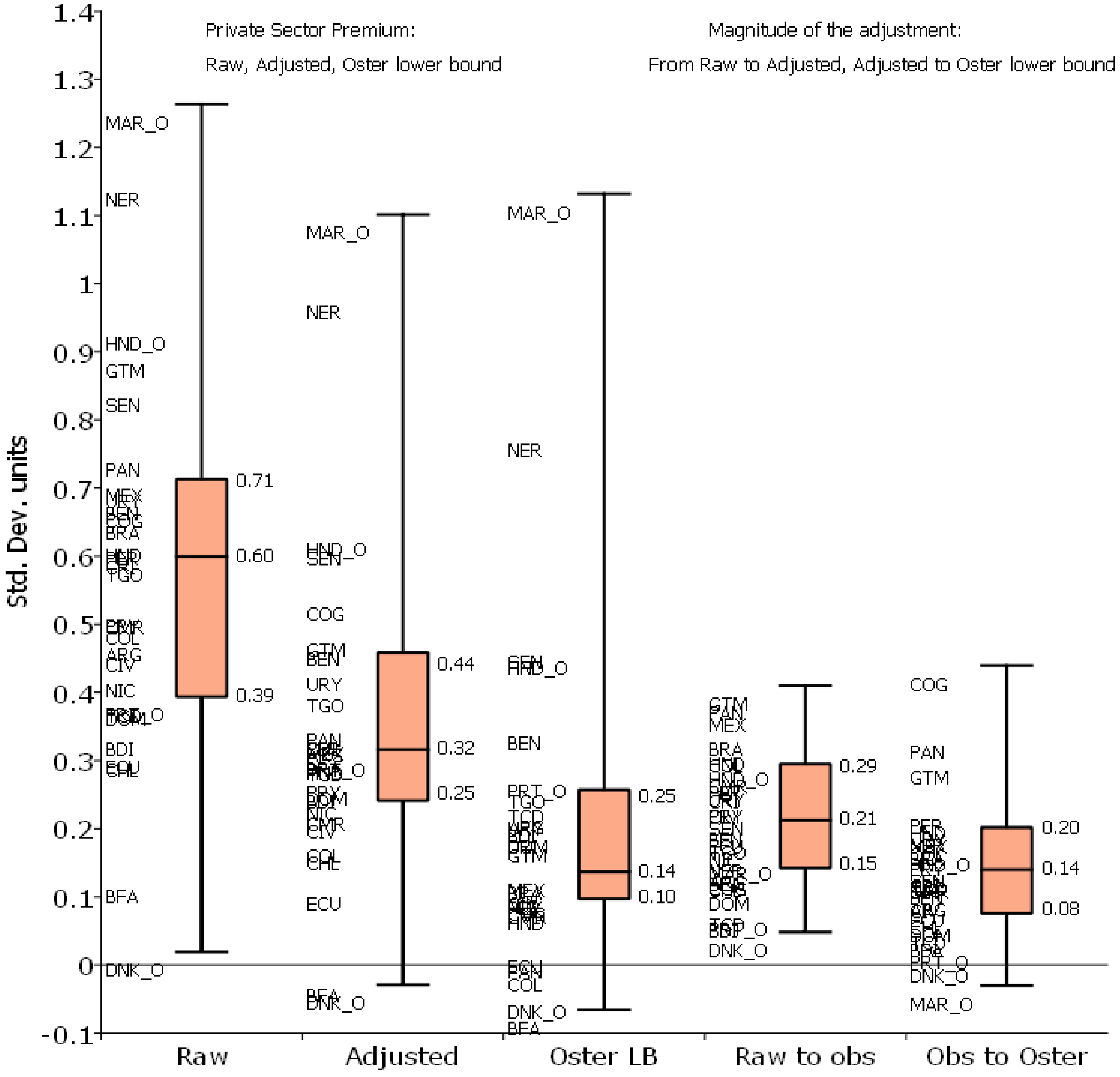
First, the private sector premium is still positive, but smaller than the "SES adjusted" estimates. Each adjustment reduced the median by about half (from .60 to .32 and from .32 to .14).

Second, there remains in the Oster adjusted estimates, massive heterogeneity as the range still goes from zero (or negative) to a full standard deviation (Morocco). The 25th-75th range is as big as the median, from .25 to .10 (a spread of .15)

Third, there is variation across countries in the extent to which the Oster adjustment affects the point estimates (as there was for the magnitude of adjustment to estimates of the private school premium from the observed variables—and this is not happenstance). The median was to reduce the estimated premium by .14 sd, but for some countries it *increased* the estimate of the premium (e.g. Denmark, Morocco) and for other countries the adjustment was much larger, for instance, lowering the premium from the SES adjusted by .30 in Guatemala.

[18] In principle one could do Oster bounds for the PISA data but it would require working with the student level data country by country.

Figure 3: Patel-Sandefur private sector premium and adjustments



Source: Author’s calculations with P-S data.

II.D) Could one reasonable expect external validity across contexts of either estimates of causal impact or selectivity bias?

The observed private sector premium in any given context is the cumulative result of potentially millions and millions of individuals making choices, over extended periods of time, subject to constraints (time, material resources, individual capacity, available information, etc.). Every day students decide how hard to work in school (versus goofing off, daydreaming), how much to study after school (versus play), and what effort to devote which subjects. Parents have decided which school their child should attend (given their financial and other constraints). how much attention and support they give to their child’s studies (again, given their capacity and constraints), and what level (if any) of tutoring or outside help to seek (and in what subjects) and use information known only to the parents/students (e.g. estimates of student capability, ambition, preferences).

Teachers are deciding every day what to teach, how to teach it, how much effort to put into preparation, how much feedback of which type to give each on work of each student. Those who manage public and own and manage private schools are making choices about whether to stay in business or close, who to hire, how to structure compensation, what textbooks and instructional materials to provide. These choices subject to the constraints and conditions[19] aggregate up to the observed outcomes of enrollments of students between public and private schools (school selection) and to observed distributions of learning across students in the two types of schools.

In our best available theories neither "causal impact" or "selectivity bias" between private and public schools are constants, like electron mass or the gravitational constant. Nor are they parameters of a physical or biological nature that might vary in myriad ways with respect to conditions but vary in understood ways that can be put into Handbooks: like the boiling point of water (which depends on atmospheric pressure), or the compressive strength of concrete (which depends on the water content when poured, the degree of compaction, the temperature at which the concrete dries, etc), or, the relationship between plant growth and fertilizer application (which depends on soil conditions, etc.). And causal impact or selectivity bias are not even "structural parameters" of a given model (like a price elasticity of demand, or an elasticity of labor supply). Both "causal impact" and "selectivity bias" are outcomes of constrained choices by individuals and households and influenced by some complex mix of model structures, their parameters, and relevant variables.

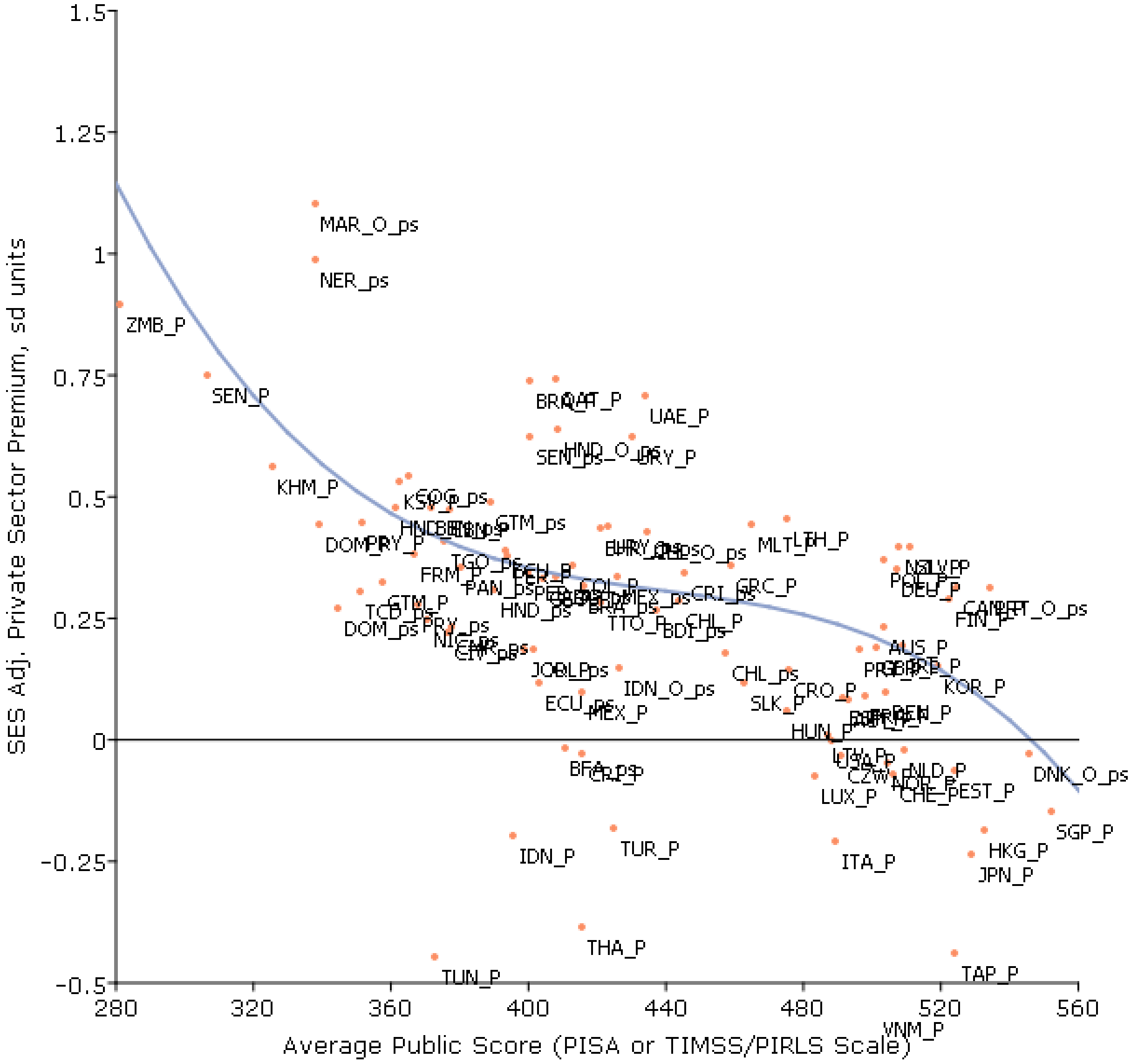
One possible way of understanding the observed differences across countries in the private school learning premium is that the "true" causal impact is the same across countries and the observed differences in estimates are defects of method in adjusting for selection. However, there is no reason to expect "external validity" of estimates of causal impact. The estimate of a causal impact in context c is going to depend on the relative efficacy of public and private schools in that context, each of which is the outcome of a "theory" or model structure, parameters, and variables. For instance, teachers in public schools in some contexts face an accountability system that is not "coherent for learning" (Pritchett 2015) and hence face little or no motivation to achieve learning results whereas in other countries the public sector does promote effective accountability for learning results and hence the public sector teachers are motivated for results (with some combination of intrinsic and extrinsic motivations).

Figure 4 shows a simple scatter plot of the private sector learning premium adjusted for observables including household SES ($\beta(SES.Z)^{c}_{CI}$) against the average score of students in the public sector (with a fitted cubic relationship). In very poorly performing public systems (bottom tercile, public sector score less than 400) the median adjusted premium is high (.39), in

[19] I am not implying the credibility of any particular model of those constrained choices, like that all agents are perfectly intertemporally maximizing some well-defined and stable utility function, just that constrained choices are being made.

moderate performance systems (public sector score between 400 and 475) the estimated SES adjusted premiums is moderate and has high variance, with some large positive and large negative estimates, and is small (median .08) for high performing public systems (above 475), many positive and many negative estimates.

Figure 4: Average public scores and private premium adjusted for SES



Source: Author's calculations. _P are those from PISA, _ps from P-S.

Variance in public sector performance across countries in learning is to be expected as there is variance across countries in many dimensions that are potentially relevant. Developing country governments vary along many dimensions of political structure, administrative capability, and public sector outcomes. They have very different omnibus measures state capability (Andrews, Pritchett and Woolcock 2016). Countries have very different outcomes on other public sector tasks, even when they have exactly the same policies, like returning misaddressed mail (Chong et al 2014) or controlling corruption (e.g. Kaufman, Kraay and Mastruzzi 2010). Direct evidence of the management practices of schools suggest large differences in the implementation of effective practices across countries and that these are related to learning outcomes (Leaver, Lemos, and Scur (2019)). No reasonable model of the causal impact of private versus schools across the developing world could start from the premise that the public sectors are equally efficacious in all countries in producing learning.

Note that the PISA ESCS adjusted estimate for the USA is -.003—almost exactly zero. This makes it plausible to believe that well-identified estimates in the USA could produce low or zero values. But, at the same time, it is very hard to look at Figure 4 take at all seriously the argument made above that the rigorous estimates from the USA should be applied to all other countries. Even if (i) I believe the best and most rigorous estimates of causal impact of private schools (correcting completely for both types of selection) in the USA are zero and (ii) I believe that these estimates from the USA are the best and most rigorous estimates available in the world, it is implausible to argue that therefore I should also believe that (iii) that the causal impact in Morocco or Zambia or Brazil is, like the USA, zero (or, for that matter, the causal impact of private schools in Indonesia is not negative).

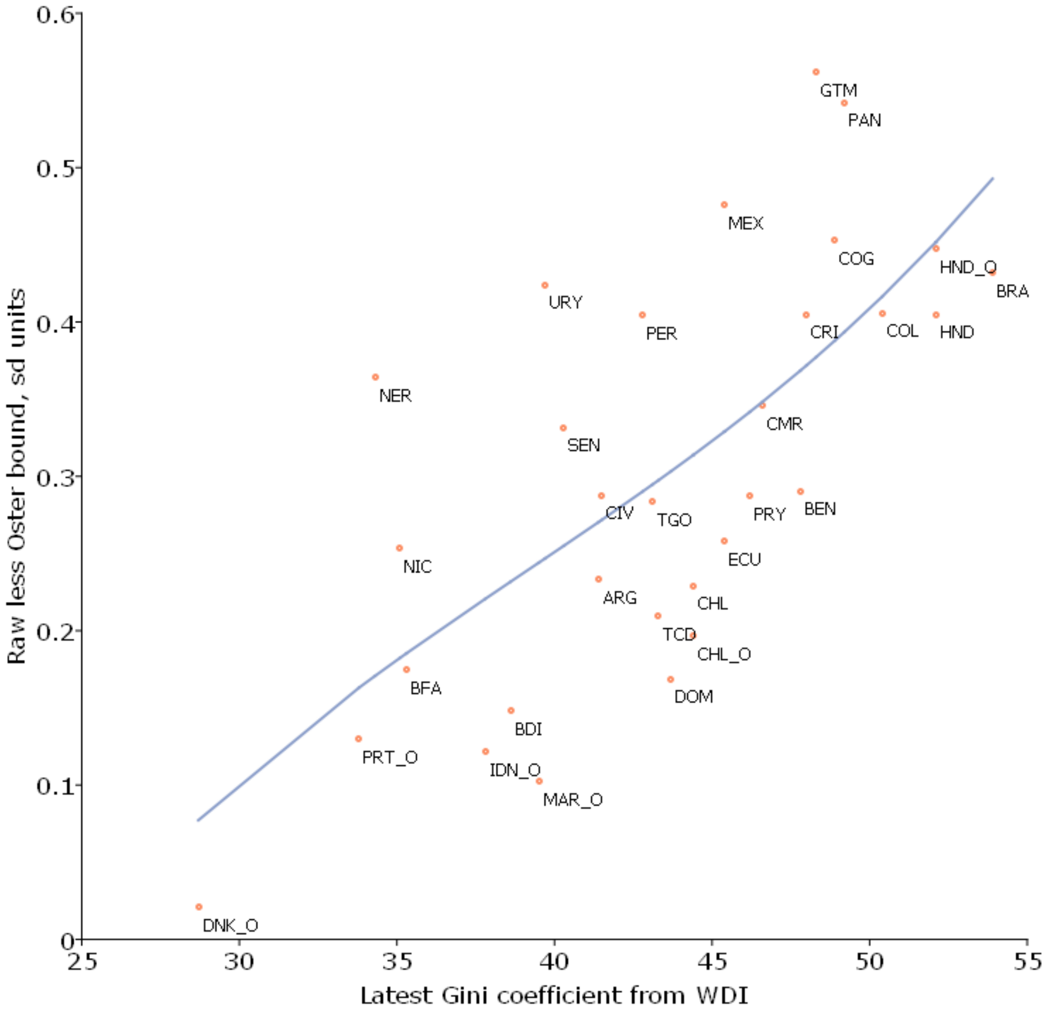
We should also expect large differences in the efficacy of the private sector in producing learning across countries. For one thing, we observe very different levels of average productivity across private firms generally (Klenow and Hseih 2009) and there is no reason to believe firms producing education would be any different. Moreover, we observe firms in different countries engaged in the production of private goods to have very difficult management practices (Bloom, Genakos, Sadun and van Reene 2012) and these extend to school management in the private sector (Lemos, Muralidharan and Scur 2021).

In addition, there is evidence that the structure of the markets for private schools can affect the average efficacy in producing learning. For instance, a team of researchers have done a series of experiments on the operation of the market for low cost schools in Pakistan and find that “interventions” like providing information (Andrabi, Das, Khwaja 2015) or giving capital grants to school operators (Andrabi et al 2020) affect the learning efficacy (and pricing) of the typical private firm (in part through exit of low quality schools). The obvious (and common sense) implication is that the efficacy of the private sector itself is a function of market structure, parameters, and variables.

I am not promoting any particular model of the determination of public or private efficacy as being “correct” but just pointing out that expecting “external validity” of LATE/causal impact of PSLP across contexts in spite of large observed differences in SES adjusted estimates is unreasonable.

The same is true of selection bias. Countries around the world differ in so many ways: the resources parents have available on average, the inequality in income/consumption/asset across households, how much labor markets reward academic credentials (and with what structure, like how convex is the relationship between academic decrees and incomes, for instance). One of the many intriguing facts thrown up by the Patel-Sandefur (2019) estimates is that the degree of selection bias, as measured by the difference between the raw private sector premium and the Oster bound, varies with measures of a country's income inequality. Figure 5 (which is adapted from Figure 17 in the Patel Sandefur paper) shows that countries with a high degree of income inequality have a large raw premium but about the same Oster bound as countries with low income inequality. And this variation is selectivity bias is not completely determined by the estimates of causal impact. For instance, both Morocco (MAR_O) and Indonesia (IDN_O) have relatively small selectivity bias (around .1) but the Oster estimate is ten times larger for Morocco (.738) than Indonesia (.075).

Figure 5: Patel-Sandefur estimates of selectivity bias, by inequality (Gini)



Source: Adapted from Figure 17 in Patel and Sandefur (2019)

III) *Relying exclusively on rigorous evidence of causal impact makes for worse predictions*

This section is simple[20]. I use the data on estimates of the PSLP to compute the RMSE (Root Mean Square Error) of predicting the "true" causal impact across various weights on a "rigorous" estimate (equation 5) and various assumptions about what the "true" estimate for each country is[21]. The results show that "relying on the rigorous evidence" (RORE(α_{SR}=1)) produces worse prediction errors, sometimes much worse, than ignoring the rigorous evidence altogether and just using the OLS estimates with SES (α_{SR}=0). The lower RMSE of predictions of causal impact is the result even though we know that the SES adjusted OLS estimates are biased by selection on unobserved variables. The worse prediction error is because the variation in the "true" estimates is large and hence collapsing predictions onto a single value (even if that value is, on average, correct) increases prediction error.

III.A) OLS alone produces better RMSE of prediction

Obviously the "true" estimate is unknown, but we do have from P-S an estimate of the raw, HH SES adjusted, and Oster lower bound PSLP for 32 countries. I assume the "true" causal impact is a weighted average of the OLS estimate and the Oster bound for each country:

$$10)\ \beta^{z}_{CI,True} = (1-\alpha_{Ost}) * \beta^{z}_{CI,OLS|Z} + \alpha_{Ost} * \beta^{z}_{CI,Oster}$$

With this assumption we can take any set of estimates of the causal impact and compute two estimates of the RMSE (root mean square error) relative to this assumption. The most naïve possible prediction is to just use the OLS estimate (α_{SR}=0). Alternatively, one could take any estimate regarded as "rigorous" and compute the RMSE of RORE(α_{SR}=1).

Figure 6 shows the ratio of the RMSE of RORE (α_{SR}=1) to OLS (α_{SE}=0), where values above 1 imply exclusive reliance on the rigorous evidence does worse than ignoring it altogether:

$$11)\ Ratio_{RMSE} = \frac{RMSE(\alpha_{EV}=1, \alpha_{Ost}=\{.5,1\})}{RMSE(\alpha_{EV}=0, \alpha_{Ost}=\{.5,1\})}.$$

Since the assumptions about "true" values matter, Figure 6 shows the ration of RMSE for two values of assumptions about the "true" value: (i) α_{Ost}=.5 splits the difference of OLS and the Oster bound) and (ii) $\alpha_{Ost} = 1$ assumes the "true" value for each country is the estimated Oster bound.

Figure 6 shows different ways of deciding what the "rigorous" evidence is for this calculation:

- For each of the 32 countries from the P-S data assume that the "rigorous' evidence is either the Oster bound or the weighted average of OLS and Oster for a single country and

[20] Compared to the next section, In previous papers the complex section came first and I now think that was a mistake as it made the point seem more complex and difficult than it really is.

[21] This was Justin Sandefur's idea and he did this calculation for estimates of the impact of micro-credit in a previous paper (Pritchett and Sandefur 2015).

use that one country's estimate as the available rigorous estimate (this produces the lines showing the ratios of RMSE across all possible countries)

- Assume the average of the "true" value estimates for all S countries (with $\alpha_{Ost} = .5\ or\ 1$)
- I do a "causal un-systematic review" taking the simple average of the estimated private sector premium from three RCT (or well identified) estimates: Colombia (Angrist et al 2002), Andhra Pradesh India (Muralidharan and Sundararaman 2015) and Pakistan (forthcoming).
- I mimic a "systematic review" from the existing P-S estimates by an random sampling of 3 countries from the P-S countries for 1000 choices and taking the average of those 3 as a possible outcome of a "systematic review" so I can generate not the mean of any particular set but also the distribution.

I discuss the results of each scenario.

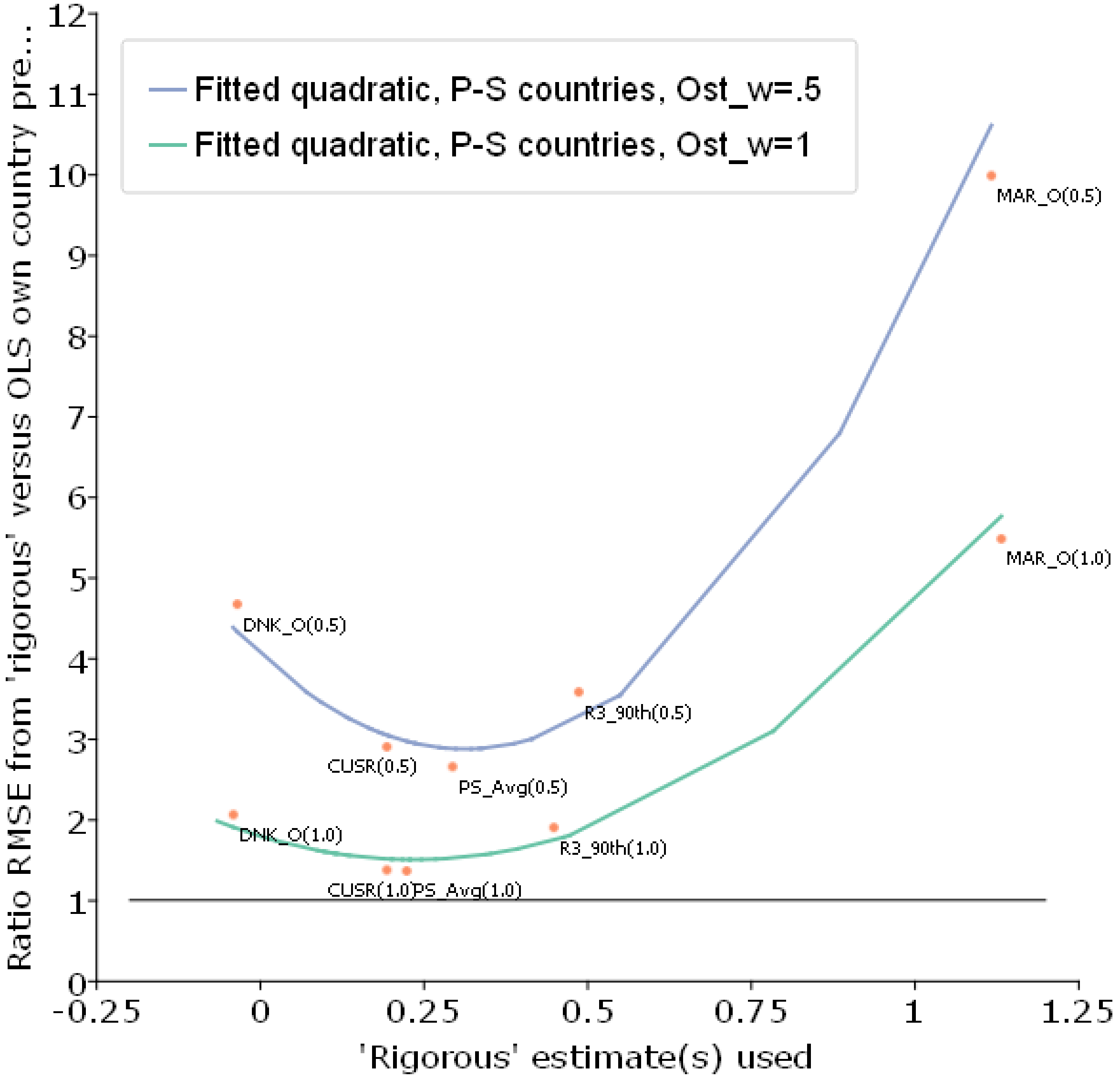
Variation across all countries. The lower line in Figure 6 under the assumption the "true" LATE for each country is the Oster bound ($\alpha_{Ost} = 1$). This is a strong case for using rigorous estimates as it assumes the selection bias from unobserved variables is large. The idea of taking each country one by one is to simulate "what if a researcher did a RCT for one particular country, recovered its "true" LATE, and then one used that rigorous estimate to predict the LATE for all countries?"

First, the ratio of RORE(1) to OLS RMSE has a quadratic shape across countries because the closer the country's estimate is to the average across all countries the lower the RMSE. If the rigorous estimate country happens to be far from the mean then RMSE disadvantage of RORE(1) is larger. Morocco's (MAR_O) Oster bound estimate is over 1 sd hence, doing an RCT in Morocco and then using that rigorous evidence to predict for all other countries would produce a RMSE five times worse than just ignoring the Moroccan estimate and using OLS for each country. Similarly, extrapolating a rigorous estimate from Denmark (DNK_O) to other countries produces a RMSE of prediction twice as high as just ignoring it altogether.

Second, in these data RORE for evidence from any country is *always* worse than naïve OLS. As we saw in the simply analogy of predicting men's height, if the variability is large relatative to the internal bias justing relying on the country specific estimates known to have bias can still be a better RMSE error prediction than relying on any singly estimate that lacks any bias. While the private school premium is an example where the case for using rigorous evidence is strong because the selectivity bias on both observed SES (Figure 2) and on unobserveds (Figure 3) is strong, it is still the case the cross-country variablity is sufficiently large that the RMSE of using OLS (that allows country specific variation) is better than the best single value from rigorous evidence.

Three, the lower line with the assumption the 'true' estimate is $\alpha_{Ost} = 1$ is the "worst case" scenario for using OLS and there is no reason to assume the Oster bound is correct, as it is a bound. The higher line in Figure 6 assumes that "true" coefficient for each country is a weighted average with $\alpha_{Ost} = .5$. The changes are intuitive. As this assumes there is less selection bias on unobserveds, the gains of shifting the distribution to be centered on an estimate with internal validity are smaller and the losses from the collapse onto a single value are still large, so the RMSE advantage of ignoring the rigorous estimate from a single country are larger. Even in the *best* case the RMSE prediction error is 3 times larger for RORE than OLS.

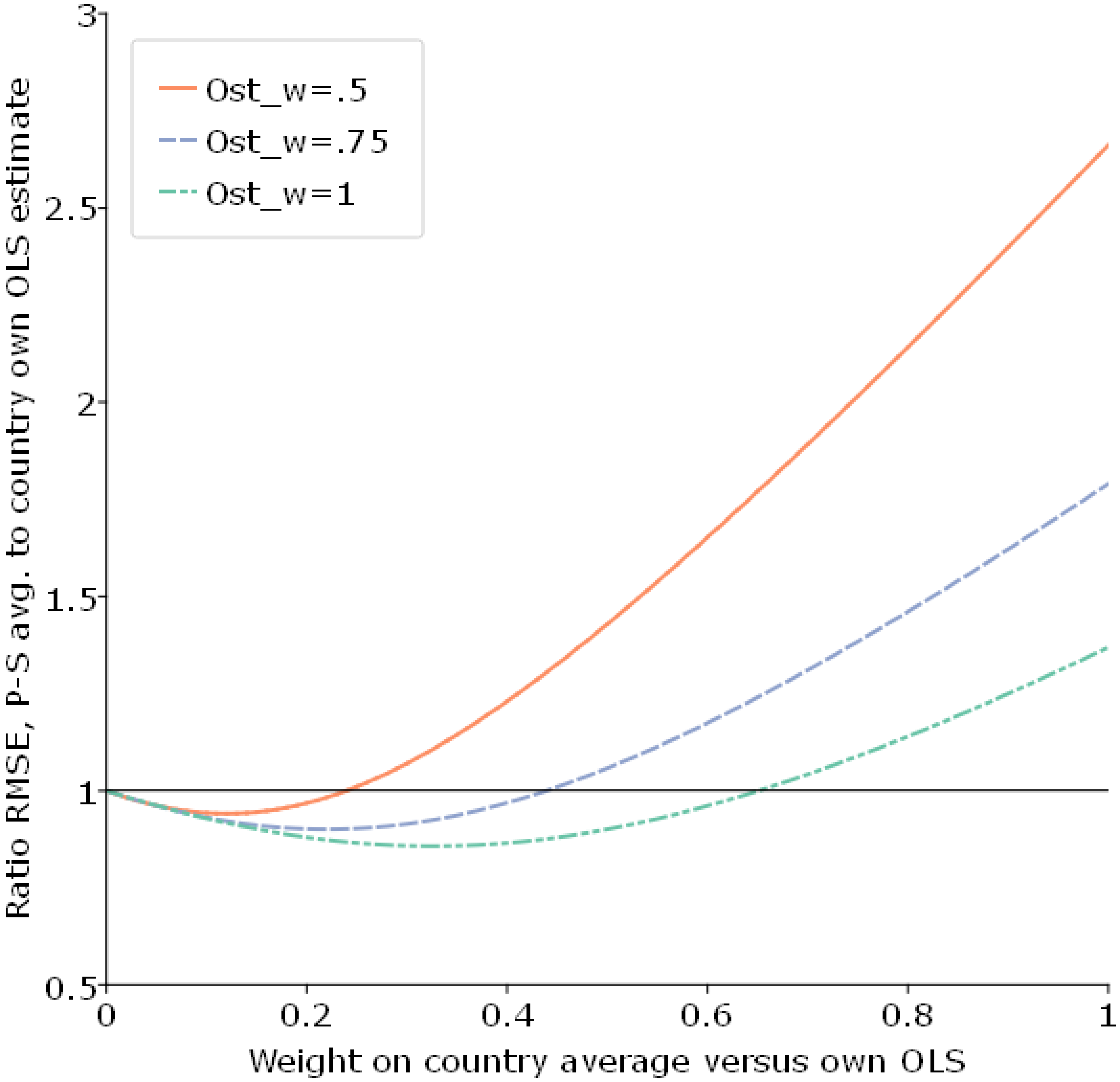
Figure 6: Ratio RMSE of prediction across Patel-Sandefur estimates



Using the average of all the countries. Also shown in Figure 6 is the ratio of RMSE using the average of the assumed 'true' value (depending on the Oster weight, either 1 or .5) for all 32 P-S countries, which is.223, as the rigorous estimate. This is as low as the ratio of RMSE can get and still it is better to just use OLS, a little bit better if $\alpha_{Ost} = 1$ and a lot better if $\alpha_{Ost} =$ .5 (that is, selectivity bias from unobserveds is less severe).

That the RMSE of prediction is better for OLS than for RORE does not implying using OLS is optimal. The RMSE error minimizing weight is, in general, non-zero and the precise value depends on the relative magnitudes of the bias from OLS and the underlying heterogeneity in the 'true' LATE across countries. Figure 7 iterates across possible weights on the rigorous evidence (α_{RE}), which in this case is assumed to be the actual cross-country average of the assumed 'true' values, for different assumptions about the weighted average of the OLS and Oster in estimating the 'true' coefficient. Intuitively, when the selectivity bias less severe (e.g. assumed lower values of α_{Ost}) then the optimal weight on the rigorous evidence is lower (less than .2). Even when it is that $\alpha_{Ost} = 1$ the optimal weight is less than .4 and simple OLS outperforms any values above about .65.

Figure 7: RMSE of prediction using country average versus own OLS



Casual Un-Systematic Review. The obvious weakness of the calculations above is that I am not in fact relying on actual rigorous estimates but rather a set of assumptions about what they would be, if they were done. But for my point this matters much less than you think. I do a casual and unsystematic review of the actual rigorous estimates of the causal impact of private schools on learning in development countries. I take an estimate of .2 from Colombia (Angrist et al 2002), an estimate of .23 from Andhra Pradesh (Muralidharan and Sundararaman 2015), and an estimate of .15 for Pakistan (forthcoming). The average is .19. Using this as the rigorous estimate produces ratios of RMSE of prediction slightly worse than the P-S average and for both $\alpha_{Ost} = 1$ and $\alpha_{Ost} = 1$ (shown as CUSR) and one still gets better predictions ignoring these three actual studies entirely and using own context OLS.

Simulated Systematic Reviews. One of the risks of trying to rely on the rigorous evidence is that RCTs are uncommon (because they are difficult to do well, time consuming, and they are very expensive relative to studies done on non-experimental data)[22]. This creates two risks. One is that the places and organizations that do RCTs are not representative of the typical causal impact (Allcott 2012, Brigham et al 2013)[23]. The other risk is that, just by chance, since the sample is small one might be low or high values. I simulate a distribution "systematic reviews" on evidence from just three countries produces by taking a 1000 samples each of size 3 from the assumed 'true' values for the 32 countries. Obviously the average of the RMSE of these simulated systematic reviews is same using the country average. The label R3_90th shows the RMSE from using the 90th percentile of the distribution of estimates from systematic reviews of size 3, which is much larger than the ratio RMSE using the country average. This is just to make the obvious point that if there is large heterogeneity in the true values across contexts the results of RCTs from just a few can lead to very large mispredictions.

I am not making the case for "ignore rigorous evidence not from your context." Your estimate of the causal impact in your context should be based on your understanding of all the evidence, from your own context and from other contexts. But the powerful rhetoric and slogan like "rely on rigorous evidence" that suggests an *exclusive* reliance on the rigorous evidence, or anything like equation 9 of forming your beliefs about the distribution of LATE in your context exclusively on the rigorous evidence from other contexts, has no support. Once one acknowledges that the evidence from other contexts has to be weighed together with other

[22] I would wager than the *incremental* cost of all of the PISA and P-S estimates *combined* is less than just the cost of the Muralidharan and Sundararaman (2015) estimate alone. That is, once one had decided to incur the cost of creating a nationally representative estimate of the learning proficiency of 15 year old students then the incremental cost of generating an OLS estimate of the adjusted private sector learning premium is just the cost of (i) making sure the questionnaire identifies the type of school (as sampling private schools is not an incremental cost as it is necessary to achieve representative estimates), (ii) making sure data on SES is generated (although this is likely to be done for other reasons and so isn't necessarily an incremental cost of the estimates) and (iii) doing the programming to run the regressions. A rough guess is that the incremental cost of both the PISA and P-S estimates is $250,000 whereas the AP study was, my guess, twice that.

[23] Another of the many asymmetries the randomistas try to maintain: "Because individuals act in self-interested ways that can induce selection bias no non-experimental result can be trusted at all, but the obvious self-interested selection bias of who does an RCT, on what, and why (Pritchett 2002) should be ignored and the resulting RCT should be considered a rigorous estimate."

evidence the whole rhetorical power of the phrase “rigorous” and “the evidence” is lost as even if the estimate was rigorous in context c (or the set of them was each ‘true’ for c) the weight is not also “rigorous.”

III.B) The many absurd implications of assuming external validity of causal impact

A fundamental issue with external validity of causal impact estimates is addition. One can decompose any non-experimental estimate into the ‘true’ causal impact ($\beta_{CI}^{z,True}$), the bias in estimation of the particular non-experimental estimate (γ_{SB}^{z}) and sampling error.

$$13)\ \beta_{NE}^{z} \equiv \beta_{CI}^{z,True} + \gamma_{SB}^{z} + sampling\ error$$

Given addition I cannot rationally change my beliefs/opinions about the causal impact in context z in response to evidence from context(s) c without also changing my beliefs/opinions about the bias. The proposed convention of RORE(1) by ignoring evidence about selection bias altogether (Section II.C) ignores addition, and ignoring addition, not surprisingly, leads to goofy implications.

Figure 8: Illustrating the absurd implications of assuming external validity

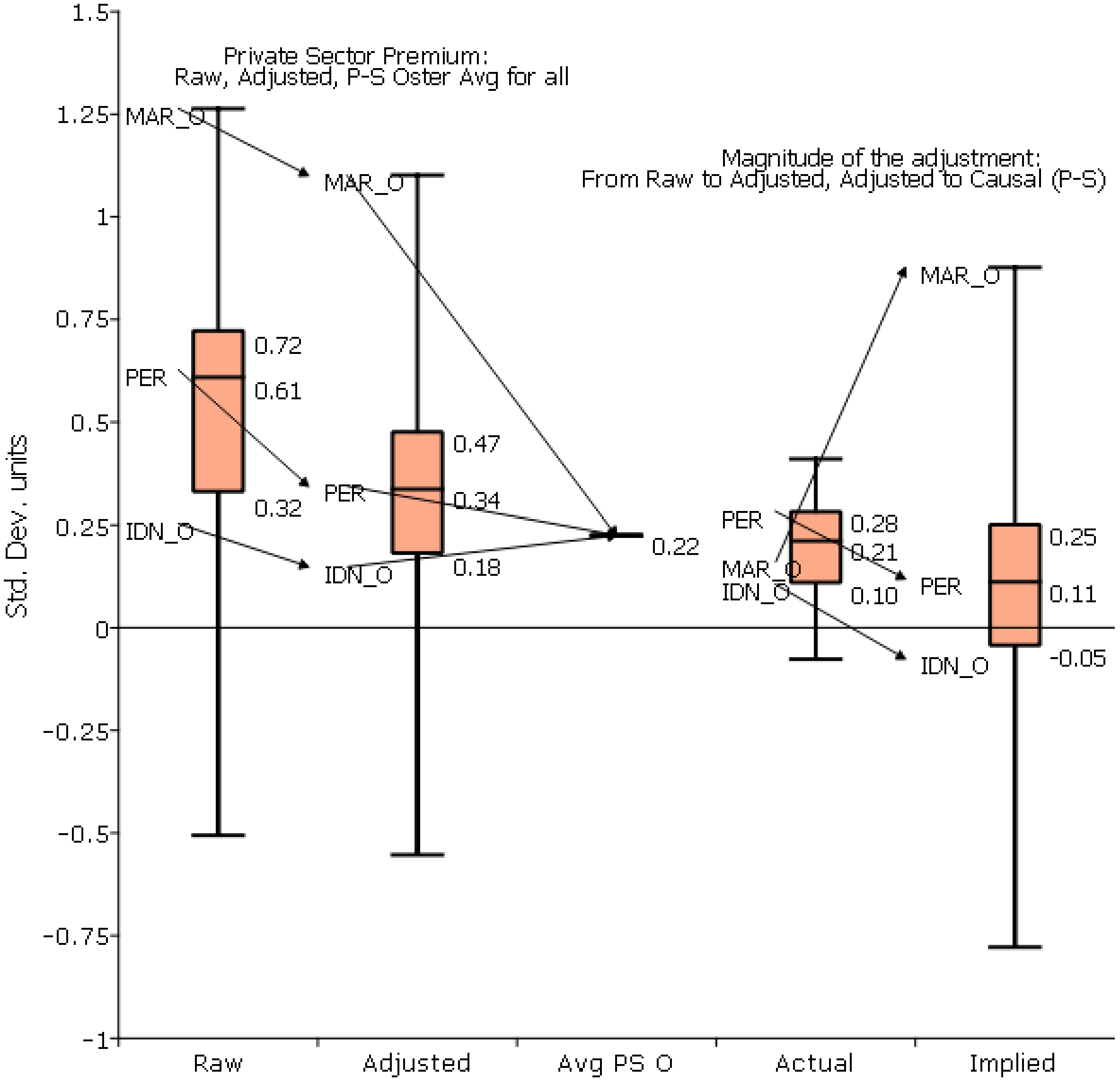


Figure 8 shows the six dubious empirical implications of assuming that a single estimate (in this case the average across countries of the P-S Oster bounds, .22, which is that a "systematic review" would produce, but it doesn't really matter which particular estimate I use) was the "best" and "rigorous" estimate and collapsing one's beliefs for all countries onto that value.

First, since the estimates of the raw and SES adjusted estimates of private sector premium have large heterogeneity, RORE-CI forces one's beliefs about causal impact in different countries to move in different directions in different countries, by a lot. The raw estimate for Morocco is a premium of 1.23 and its HH SES adjusted estimate is 1.1, much higher than .22. The raw P_S estimate for Indonesia is .25 and the HH SES adjusted estimate is .15. You have to believe that the non-experimental estimates for Morocco were much too high and those for

Indonesia too low. For no good reason. That is, just reliance on rigorous estimates of causal impacts from some set of contexts isn't a theory or model or rationale of why we should change our beliefs in this way.

Second, RORE-CI implies that a belief that the variability/heterogeneity of the causal estimates was much, much less than the observed variance of either the raw, adjusted for HH SES, or Oster bound estimates (that we say in Figure 3). There is no reason to believe this. It is equally plausible that all of the observed heterogeneity across countries is not actual, real, heterogeneity in causal impacts that exists in the world because contexts really differ.

Three, RORE-CI implies that the selectivity bias from the non-HH SES variables (the "unobserved") was massively positive in Morocco and negative in Indonesia (just as two examples). This implies that selectivity bias cannot have external validity.

Fourth, RORE-CI implies the direction/sign of the selectivity bias from unobservables is the same as that for SES in some countries and different in other countries. So, for instance, in Peru the adjustment for HH SES moves from a raw premium of .62 to an adjusted premium of .34 (reducing the estimate by .28) and hence if I think the causal impact in Peru is .22 I have to believe the selection bias on non-SES variables was .12, which is arbitrary but perhaps not so hard to believe. But for Indonesia (using the TIMSS data from P-S) the raw estimate is .25 and the SES adjusted is .15. This means that, if I am to believe that the point estimate of causal impact for Indonesia is .22 I have to believe that the selection effect of the non-SES variables has *negative* bias of -.07—of the opposite sign from the adjustment for selectivity bias from HH SES. This is not impossible, but there is no reason to believe it is true.

Fifth, the collapse onto a single value for the causal impact implies that the variance across countries of the selectivity bias from unobservables has to increase massively. Since the causal impact and the selectivity bias have to add up to the raw for each country and the variance in the raw premium across countries is fixed fact any reduction in the variance of causal impacts leads to an increase in the variance of selectivity bias. Again, one has to believe this for no good reason.

Sixth, since we can observe (for these countries) the selectivity bias from the observables for HH SES we can compare the variability the selectivity bias from adjusting for observed variables to the adjustment for the unobserved. The fourth box in Figure 6 shows that the spread in the adjustment from HH SES adjusted premium to causal impact has to be much larger than the spread in the adjustment from raw to HH SES selectivity. Again, for no good reason.

Believing in the external validity of causal impact produces implied estimates of selection bias across countries that are not just a set of measure a weird set of measure zero: not only have to take particular values to reconcile the addition of non-experimental estimate and the causal impact estimate and the selection effect, but those values have no model or theoretic reason or rationale (and seem pretty implausible)[24].

[24] While these are all illustrated in Figure 8 with α_{RE}=1 Pritchett and Sandefur (2014, 2015) show these problems emerge in general with positive weight on estimates of causal impact. The problem is that the

III.C) Prediction error is much better using estimates of selection bias rather than causal impact

Given that the large errors of prediction from trying to use one value for casual impact (RORE-CI) an obvious alternative idea is to “correct” the OLS for each country for selection bias. Even crude corrections for the selection bias of OLS estimate for each country bias produce much lower prediction errors than RORE of causal impacts.

Equation 12 allows the prediction for each country to be a weighted average of the “own OLS adjusted for selection bias” and some version of the rigorous evidence.

$$12)\ \beta^{z}_{CI} = (1-\alpha_{RE}) * (\beta^{z}_{OLS|W} - \gamma^{z}_{OLS|W}) + \alpha_{RE} * \beta^{RE}_{CI}$$

Figure 9 shows the prediction RMSE assuming the ‘true’ is the Oster bound for each country ($\alpha_{Ost} = 1$). Using the P-S average and $\alpha_{RE} = 1$ the RMSE=.234 and with $\alpha_{RE} = 0$ the RMSE=.171 (these are exactly the values of RMSE that produce the ratios shown in Figure 6). In Figure 9 these are compared these two RMSE of prediction to three different *ad hoc* ways of estimating the selection bias of OLS with SES for each country.

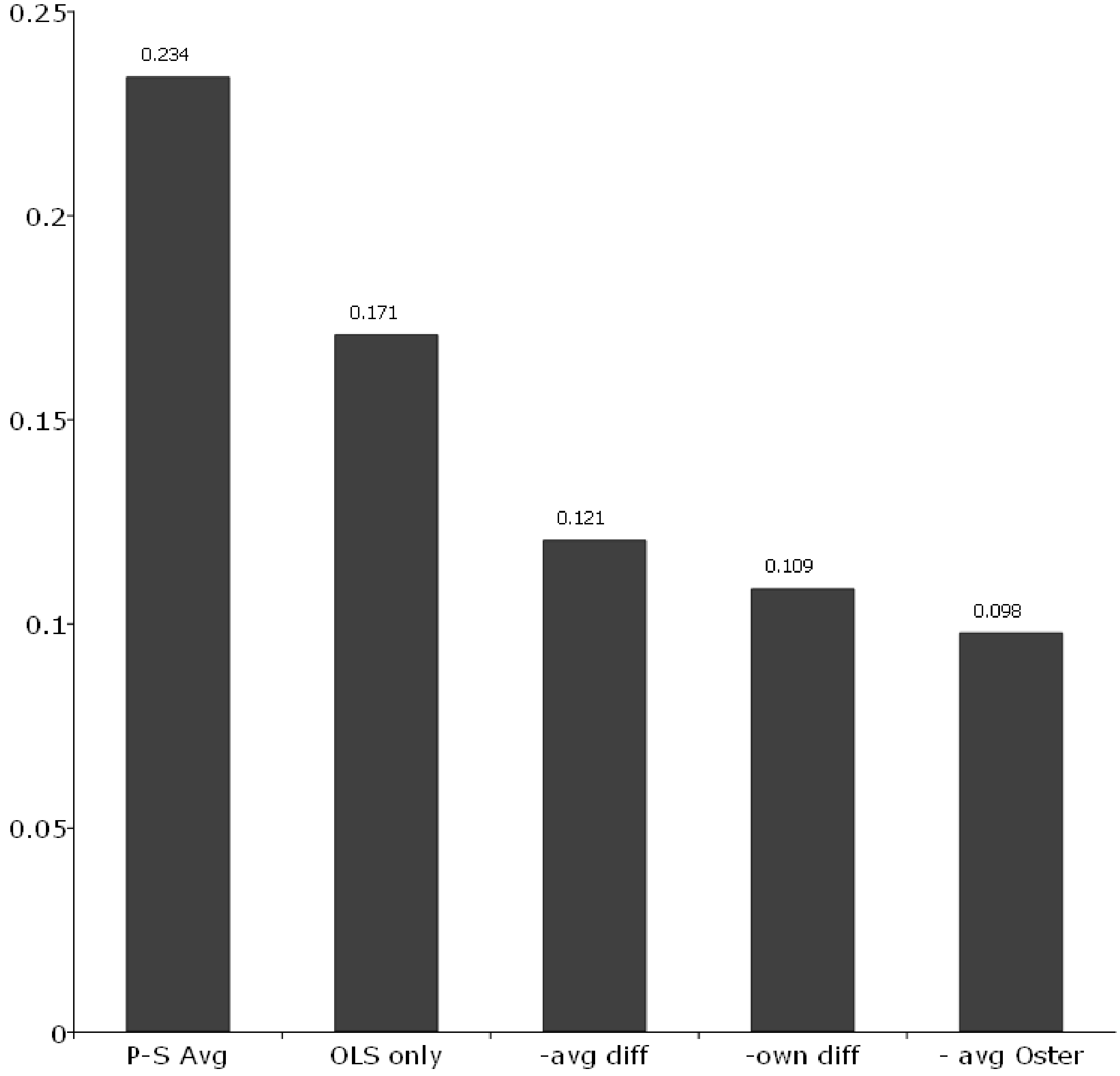
Average difference between the raw and OLS. Since we have the raw and OLS private sector premium for each country we can just assume that the selection bias from the unobserved variables in each country are the same magnitude as the average of the selection bias for SES for all countries. This average is: $\overline{\beta(\emptyset)^{z}_{CI} - \beta(SES)^{z}_{CI}}$=.210. Hence an ad hoc guess of the causal impact for each country is their OLS estimate less .21. This simple adjustment and ignoring the evidence about causal impact ($\alpha_{SR} = 0$) produces a RMSE about half that of RORE($\alpha_{SR} = 1$): .234 versus .121.

Country’s difference of raw and OLS. Alternatively, one could that the selection bias from OLS with SES to the ‘true’ value for each country was just the same as the raw to OLS with SES (rather than using the cross-country average). In this case the RMSE is lower still, .109.

Average of the OLS and Oster bound. Finally, one could assume that the selection bias for each country was the average of the difference between the OLS with SES and the Oster bound for each country. The $\alpha_{SR} = 0$ prediction for each country is the OLS estimate less .140 (Figure 3). This produces a RMSE of .098. This counter-factual is just like, in our simple analogy of guess men’s height, adjusting each man’s self-report for the average difference of self-report and true. This is assuming that the part of the “rigorous” evidence that has the most useful external validity is the estimate of selections, not the estimate of height.

fact that there is an adding up problem plus heterogeneity in the distribution of the non-experimental estimates implies that any given weight given to the causal impact implies paradoxical implications for the weight given to the estimate for selectivity bias.

Figure 9: RSME adjusting for selectivity bias vs causal impact



III.D) "Balance" approaches to RCTs are self-limiting in answering key questions

The case *for* the use of the "balance" approach to RCTs in development is also the case *against* the usefulness of RCTs in development. That is, by eschewing the "understanding" approach the embeds results in a paradigm/theory/model that encompasses all of the evidence in favor of a "let's just try this and see what happens" one is left with RCT results that cannot make any legitimate claim to application beyond their narrow context. Without a theory or model one cannot even say what "context" is and hence cannot talk meaningfully about the scope of the applicability of the results as without a model one cannot say what is a "similar" context or what the "local" in LATE means except in the most whimsical (in Leamer's sense) ways.

What is "context"? Suppose I did an experiment of dropping a marble onto a bucket of H_2O and traced the time it took to sink to the bottom of the bucket. What would be a "similar" experiment? If a dropped a round object of exactly the same weight as the marble made of salt it would not just sink, it would also dissolve and the notion "time to sink to the bottom" would not even be well-defined. If I dropped a marble of exactly the same size made of cork it would float and time to sink is (essentially) infinite. If the water was 36 degrees or 42 degrees I would likely get the (nearly) same results but the H_2O was at 30 degrees changed to ice the same marble would not sink at all. Certainly the range of interacting and background conditions for the efficacy of development projects is at least at complicated and complex as an object sinking in water. In the balance approach results cannot have any "rigorous" interpretation or lessons beyond: "Q: What do I learn from your study?" A: "This was done and that happened. That is all that can be said with rigor."

While here I have often used "context" to mean "country" but without a theory it is impossible to say what the relevant "context" is and hence evidence is only rigorous about the past. Context can be the presence or absence of complementary conditions. Mbiti et al (2019) for instance, show strong complementarities between teacher incentives based on student performance and school grants, hence in the "context" of "no school grants" the impact of teacher incentives in Tanzania will be lower than in the "context" of "with school grants." Context could be implementing organization. Vivalt (2020) shows different impacts on average between evaluations of programs implemented by government versus non-government organizations. Bold et al (2018) did an experimental evaluation of the scaling in Kenya of a previous experimental evaluation shown to "work" in Kenya and found that in the scaling of the original intervention when the implementation was done by an NGO (as it was in the original experiment) the scale up had roughly the same impact but when implemented by the government it had no impact at all. "Context" can even lead to different results for the same treatment by the same individuals but in a different setting. In Bihar India a "teaching at the right level" style pedagogical approach was implemented in a summer camp by public school teachers and had the large impacts demonstrated when implemented by an NGO. But when the same public school teachers implemented the same program in their schools in the regular school year it failed (Banerji and Walton 2011). The list of elements that are potentially the "context" that affects the efficacy of a policy/program/project is long (and difficult to know in advance (Pritchett and Nadel 2016) and theory and models are the only way to begin to delineate and document those.

Conclusion[25]

So here I am, in the middle way, having had twenty years-
Twenty years largely wasted, the years of l'entre deux guerres-
Trying to use words, and every attempt
Is a wholy new start, and a different kind of failure
Because one has only learnt to get the better of words
For the thing one no longer has to say, or the way in which
One is no longer disposed to say it. And so each venture
Is a new beginning, a raid on the inarticulate,
With shabby equipment always deteriorating
In the general mess of imprecision of feeling,
Undisciplined squads of emotion. And what there is to conquer
By strength and submission, has already been discovered
Once or twice, or several times

T.S. Eliot, East Coker, Four Quartets.

Development actions and interventions (policies/programs/projects/practices) should be based on the evidence. This truism now comes with a radical proposal about the meaning of "the evidence." In development practice, where there are hundreds of complex, sometimes rapidly changing, contexts seemingly innocuous phrases like "rely on the rigorous evidence" are taken to mean: "Ignore evidence from your context and rely in *your* context on evidence that was 'rigorous' for another place, another time, another implementing organization, another set of interacting policies, another set of local social norms, another program design and do this without any underlying model or theory that guides your understanding of the relevant phenomena."

This paper is (another) extended empirical counter-example to the claims that, even in a case where the bias from non-experimental estimates is demonstrably large (estimating the private sector learning premium), that the proposed convention for "evidence" of (i) "clean

[25] As a last, long footnote, this paper is about what is, to my mind, only the fifth biggest problem with the use of RCTs in development. First is that very few of the big and important questions that face developing countries or development agencies are amenable to these methods (see Pritchett (2020) and the "Interviews" chapter containing interviews with four different development policy makers in Bedecarrats, Guerin, and Roubaud 2020). Second, the estimates of treatment effects of specific interventions lack "construct validity" (Pritchett 2017, Vivalt 2020) in that they are estimating empirical features of a single instance of a class even though the within class variation due to the particulars of the design is massive. Estimating the horsepower of a single car, or a small set of cars, does not produce reliable answers to the question of 'what is the horsepower of cars?" because that question is itself is a mistake (as our questions like "what is the impact of class size on student learning?"). Third, RCTs ignore organizational capability and hence the likelihood the adoption of "what worked" in one organization will work in another. Fourth, the "theory of change" embedded in the arguments that RCT evidence will be useful in changing policy/program/project design is wrong—if not self-refuting (Pritchett 2009).

sweep" of the previous evidence, based on reviews that are systematic mainly in systematically (by design) ignoring most of the evidence, (ii) reliance only on the estimates of casual impact and (iii) ignoring evidence about selectivity bias from rigorous studies--actually leads to worse predictions of causal impacts. Alternative, *ad hoc*, but empirically feasible estimates of the context specific LATE—like OLS--produce prediction errors (RMSE) *half* that of the *best* exclusive "rely on the rigorous evidence."

What I advocate is "rely on an *understanding* of *all* of the evidence." This has to involve a working theory/model/conceptual frame for the phenomena at hand that can address both the question of the typical impacts, the large heterogeneity in impacts across contexts (countries, over time, settings, complementary policies, organizations) and which encompasses all of the facts (e.g. raw outcomes), and the evidence about associations and impacts (both experimental and non-experimental).   There are alternative approaches that organizations can (and do) take to discovering what is effective in their contexts and there is, as yet, no evidence these are not, in fact, far superior to the use of RCTs for organizational learning (Pritchett and Nadel 2016).

A final, hopefully deeply annoying, point. The advocates of RCTs and the use and importance of rigorous evidence, who are mostly full-time academics based in universities, have often taken a condescending (e.g. books with titles like "More than good intentions"), if not outright *ad hominem*, stance towards development practitioners. They have often treated arguments against exclusive reliance on RCT evidence, like that the world is complex, getting things done in the real world is a difficult craft, that RCTs don't address key issues, that results cannot be transplanted across contexts, not as legitimate arguments but as the self-interested pleadings of "bureaucrats" who don't care about "the evidence" or development outcomes. Therefore, it is striking that it is the practitioner objections about external validity that are actually *technically* right about the unreliability of RCTs for making context-specific predictions and it is the academics that are wrong, and this in the technical domain that supposedly is the acamedicians comparative advantage.

Andrabi, Tahir; Jishnu Das; Asim Khwaja; Selcuk Ozyurt and Niharika Singh. 2020. "Upping the Ante: The Equilibrium Effects of Unconditional Grants to Private Schools." *American Economic Review*, 110(10), 3315-49.

Andrabi, Tahir; Das, Jishnu; Khwajam, Asim Ijaz. 2009. "Report Cards: The Impact of Providing School and Child Test-Scores on Educational Markets." *BREAD Working Paper No. 226*.

Andrews, Matthew; Lant Pritchett and Michael Woolcock. 2016. *Building State Capability: Evidence, Analysis, Action*. Oxford, UK: Oxford Univerity Press.

Angrist, Joshua; Eric Bettinger; Erik Bloom; Elizabeth King and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review*, 92(5), 1535-58.

Angrist, Joshua D. and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics*, 114(2), 533-75.

Banerjee, Abhijit; Esther Duflo; N. Goldberg; Dean Karlan; Robert Osei; William Pariente; Jeremy Shapiro; Bram Thuysbaert and Christopher Udry. 2015a. "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries." *Science*, 348(6236).

Banerjee, Abhijit; Dean Karlan and Jonathan Zinman. 2015b. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics*, 7(1), 1-21.

Banerji, Rumini and Michael Walton. 2011. "What Helps Children to Learn? Evaluation of Pratham's Read India Program in Bihar and Uttarakhand." *Research Note*.

BASTAGLI, FRANCESCA; JESSICA HAGEN-ZANKER; LUKE HARMAN; VALENTINA BARCA; GEORGINA STURGE and TANJA SCHMIDT. 2018. "The Impact of Cash Transfers: A Review of the Evidence from Low- and Middle-Income Countries." *Journal of Social Policy*, 48(3), 569-94.

Bloom, Nicholas; Christos Genakos; Raffaella Sadun and John Van Reenen. 2012. "Management Practices across Firms and Countries." *Academy of Management Perspectives*, 26(1).

Bold, Tessa; Mwangi Kimenyi; Germano Mwabu and Justin Sandefur. 2018. "Experimental Evidence on Scaling up Education Reforms in Kenya." *Journal of Public Economics*, 168, 1-20.

Brigham, Matthew R.; Michael G. Findley; William T. Matthias; Chase M. Petrey and Daniel L. Nielson. 2013. "Aversion to Learning in Development? A Global Field Experiment on Microfinance Institutions."

Burde, Dana and Leigh L Linden. 2013. "Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools." *American Economic Journal: Applied Economics*, 5(3), 27-40.

Chong, Alberto; La Porta, Rafael ; Lopez-de-Silanes, Florencio; Shleifer, Andrei. 2012. "Letter Grading Government Efficiency " *NBER Working Papers* 18268.

Deaton, Angus and Nancy Cartwright. 2018. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine*, 210, 2-21.

Evans, David and Anna Popova. 2015. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *World Bank Policy Research Working Paper*, 7203.

Heckman, James J. 2020. "Epilogue: Randomization and Social Policy Revisisted," F. Bedecarrats, I. Guerin and F. Roubaud, *Randomized Control Trials in the Field of Development: A Critical Perspective.* Oxford, UK: Oxford University Press, 304-30.

Hsieh, Chang-Tai and Peter J Klenow. 2009. "Misallocation and Manufacturing Tfp in China and India." *The Quarterly Journal of Economics*, 124(4), 1403-48.

Kaufmann, Daniel; Kraay, Aart; Mastruzzi, M;. "Governance Matters V: Governance Indicators for 1996-2005,"

Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." *The American Economic Review*, 73(1), 31-43.

Leaver, Clare; Renata Lemos and Daniela Scur. 2019. "Measuring and Explaining Management in Schools: New Approaches Using Public Data." *RISE Working Paper 19/033*.

Lemos, Renata; Karthik Muralidharan and Daniela Scur. 2021. " Personnel Management and School Productivity: Evidence from India." *NBER Working Paper*, 28336.

Mbiti, Isaac; Karthik Muralidharan; Mauricio Romero; Youdi Schipper; Constantine Manda and Rakesh Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania." *The Quarterly Journal of Economics*, 134(3), 1627-73.

McKenzie, David. 2020. "If It Needs a Power Calculation, Does It Matter for Poverty Reduction?" *World Development*, 127.

Muralidharan, Karthik and Paul Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives*, 31(4), 103-24.

Muralidharan, Karthik and Venkatesh Sundararaman. 2015. "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India." *The Quarterly Journal of Economics*, 130(3), 1011-66.

Nadel, Sara and Lant Pritchett. 2016. "Searching for the Devil in the Details: Learning About Development Program Design." *Center for Global Development Working Paper*, 434.

Pritchett, Lant. 2002. "It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." *Journal of Policy Reform*, 5(4), 251-69.

____. 2009. "The Policy Irrelevance of the Economics of Education: Is 'Normative as Positive' Useless, or Worse?," W. Easterly and J. Cohen, *What Works in Development: Thinking Big and Thinking Small.* Washington DC: Brookings Institution Press,

Pritchett, Lant and Justin Sandefur. 2014. "Context Matters for Size: Why External Validity Claims and Development Practice Do Not Mix." *Journal of Globalization and Development*, 42, 161-97.

____. 2015. "Learning from Experiments When Context Matters." *American Economic Review*, 105(5), 471-75.