“Rely (only) on the Rigorous Evidence” is Bad Advice

Lant Pritchett

LSE School of Public Policy

May 20, 2023

The slogan that development policy/program/project decisions should be “evidence based” is vacuous without specificity as to what counts, with what reliability, as “evidence” for which decisions. One popular interpretation of “evidence based” is to “rely (only) on the rigorous evidence” (RORE), in which “systematic review” filters the literature, retaining only estimates causal impact or a “treatment effect” (TE) which pass a threshold as “rigorous.” The systematic review summary of average (and distribution) of these treatment effects is intended to guide decisions across the variety of developing countries. I use two sets of cross-country empirical estimates of outcomes (wage gains for migrants, private school learning gains) which provide a “raw” difference (with/without), an OLS estimate (with/without controlling for observables), and an (arguably) rigorous estimate of the TE (with/without controlling for observables and unobservables). In both empirical examples relying on the “systematic review average treatment effect estimates” leads to *worse* decisions in RMSE (root mean squared error) than the completely naïve procedure of just relying on country specific OLS estimates or than adjusting country specific OLS estimates for its selectivity bias from unobserved variables.

“Rely (only) on Rigorous Evidence” is Bad Advice[1]

Introduction

A simple example helps set the stage. Suppose men generally self-report they are taller than they really are and hence their self-report of height is a biased estimate. An objective measurement of the true height of a relevant sample of men would create rigorous evidence, a distribution of unbiased estimates. In the practical task of assigning pants sizes to a group of men there are three options. One is to rely only on the rigorous evidence, ignore the biased evidence altogether, not even ask men about their height, and just assign every man the pants that fit the average man. Two, one could ignore the bias altogether and just give each man the pants corresponding to their self-report. Three, one could take self-reported height and scale it back based on a rigorous estimate of the average self-report bias. Which of these produces the smallest mismatch in pant sizes depends on three empirically contingent facts: (i) the true variability across men in height, (ii) the average magnitude of the bias from self-report, and (iii) the true variability across men in the self-report bias. As the standard deviation of men’s height in the USA is about 3 inches, if the average self-report bias is just one or two inches (a man who is 5’9’’ saying he is six feet tall is immediately implausible) then just relying on biased self-reported height will produce better results than using the mean of the rigorous evidence and if the self-report bias variability is small, just knocking off each man’s self-reported height the average self-report bias might be the best.

In development circles the urge to make “evidence based” decisions and in particular the slogan to “rely (only) on the rigorous evidence” (RORE) has been rhetorically powerful and seemingly persuasive. Resources have flowed into (i) dramatically expanding the number of randomized control trial (RCT)--and other “rigorous”--estimates of the causal impact or treatment effect (TE) of potential “interventions” (policies, programs, projects) and (ii) carrying out “systematic reviews” that privilege studies with “rigorous” studies that provide unbiased estimates of treatment effects (TE).

For instance, a recent Global Education Evidence Advisory Panel report (World Bank 2020) gave what they regarded as globally relevant advice about “best buys” in education. The report’s main figure provides average estimates of the cost-effectiveness of various (classes of) potential interventions based on “learning adjusted years of schooling” (LAYS) gain per $100. The figure shows that the intervention with the highest average cost effectiveness is: “Giving information on education quality, costs, and benefits.” That average estimate is based on *two* studies, one which showed nearly zero impact and one which showed astronomically high cost-effectiveness, as the improvements were obtained at very low cost. The report also reports estimates of the average cost effectiveness of “Teacher accountability and incentive reforms” and “Giving merit scholarships to disadvantaged children and youth,” each based on just three studies. In the domain of basic education, this is a “best practice” example (because, unlike previous reviews, it integrates “school completion” and “learning augmenting” interventions on a

[1] I would like to thank the editors of the special issue, Anke Hoeffler and Reetika Khera, the anonymous referees, Justin Sandefur, and Gulzar Natarajan for helping me produce this work.

common metric) of what “evidence based” recommendations from “systematic reviews” produce.

But, as the simple example of men’s height illustrates Whether or not the average of rigorous estimates of treatment effects produces better estimates of country specific treatment effects than the naïve approach of using each country’s biased OLS estimate is an empirically contingent question. Vivalt (2020) shows there is massive variability in the rigorous treatment effect estimates, both across and within studies. Angrist and Meager (2022) show that the effect size estimates of the impact of a single class of pedagogical intervention, “teaching at the right level,” differ across available studies by an order of magnitude. In previous work (Pritchett and Sandefur 2014, 2015) we provided an empirical example in which the RMSE (Root Mean Squared Error) of predicting treatment effects was smaller using the OLS estimates than using the average of the rigorous estimates.

In this paper I extend this work with two additional empirical examples with much larger cross-national samples: 42 countries for estimating the treatment effect of migration on wages and 29 countries for estimating the treatment effect on measures of learning from private sector schooling. Using cross-national data from these two empirical examples I illustrate four points.

One, there is no single obvious and empirically plausible interpretation of “rely on the rigorous evidence.” A rigorous study (potentially) produces both a consistent estimate of the treatment effect (TE) and, because the TE estimate, the OLS estimate, and the estimate of the selectivity bias on unobserved variables (SBU) of the OLS estimate are linked by an identity, a rigorous study could produce an estimate of the OLS SBU bias. The most common practice in systematic reviews of focusing exclusively on average of the TE estimates--SR($\overline{TE}$)--has no conceptual justification over a systematic review focus on the average SBU--SR($\overline{SBU}$). Given the large cross-national variance in the OLS estimates, it is arithmetically impossible that *both* TE *and* SBU estimates have external validity. Given the variability in both OLS and rigorous estimates the assumption that either TE or SBU estimates have is clearly wrong.

Two, in both empirical examples predicting impact in each country with the average of the treatment effects--SR($\overline{TE}$) produces *worse* decisions across countries in RMSE than either (i) just using each country’s OLS estimate or (ii) predicting the TE in each country by adjusting the OLS estimate for average estimated bias, β^c_{OLS}- SR($\overline{SBU}$).

Three, in both examples standard economic models predict heterogeneity in the true TE across countries in ways strongly confirmed by the data.

Four, in both examples there is also evidence for systematic heterogeneity across contexts in the magnitude of OLS SBU.

The widely practiced approach of doing systematic reviews that filter out nearly all of the relevant evidence and then acting as the average of the resulting treatment effect (or causal impact) estimates is “the” evidence for “evidence based” decisions is both empirically wrong and conceptually “not even wrong” (in the sense of Wolfgang Pauli) as this approach relies on conceptually indefensible assumptions about external validity.

I) The Horse Race: Methods

Here is a pretty simple empirical question: "would using the average of rigorously estimated treatment effects improve the predictive accuracy of estimating treatment effects across countries compared to the alternative of just using each country's own OLS estimate?" Pritchett and Sandefur (2015) use the RCT results of estimating the impact of microcredit across six countries from Banerjee, Karlan, and Zinman 2015 and use the raw data from these same studies to estimate OLS estimates and show the answer is "only sometimes." For the "reported profits" variable OLS always outperforms the (non-context specific) average RCT predictions. For the "consumption" variable OLS outperforms when the sample of RCTs is small but when all RCTs are used the results are slightly better for RCTs. I feel these results have been underappreciated as this was for a small sample (six countries) for a single intervention (micro-credit) and so are ignored or treated as possibly just an anomaly. Replication is difficult because the suitable data for these calculations are scarce as it requires both an OLS estimate and a consistent estimate of the treatment effect across a number of countries.

This paper uses cross-national empirical results that have raw, OLS and a consistent estimate of the (lower bound of the) treatment effect for two phenomena. One, estimates of the wage gains from migration for a specific worker moving from one of 42 countries to the USA, using the ratio of PPP wages. Two, for the learning increment of math from enrollment in private school we have OLS and TE estimates for 29 countries.

I.a) The problem of selection effects and the array of estimates: Raw, OLS, TE(Oster) Selectivity and bias in estimating treatment effects

In a standard set up (e.g. Altonji, Elder and Taber 2005) suppose that an outcome Y for individual *i* in context *C* depends on whether or not individual *i* gets "treatment" X (in the examples X is discrete, either migrant or non-migrant or enrolled in private school or not enrolled in private school) and also on other determinants of the outcome, divided into those determinants observed by the econometrician and those unobserved (equation 1):

$$1)\ Y_i^C = \beta^C X_i^C + W_{i,observed}^C + W_{i,unobserved}^C$$

The difficulties of recovering a consistent estimate of the treatment effect in this situation have long been well-known (e.g. the classic treatment in Leamer 1983 drawing on earlier literature). Just comparing the raw average scores of, say, students in public versus private schools is likely to overstate the treatment effect on learning of private schools as students select into private schools on characteristics of their household that also have a direct causal impact on learning (such as parental education, household wealth/income, socio-economic statue). This selectivity bias can be reduced with estimation methods, say simple multivariate OLS, that include a range of observed characteristics (W) of the student and their HH and hence $\hat{\beta}_{OLS(W_{observed})}^C$ is an estimate of the outcome difference for "observationally equivalent on $W_{observed}$" individuals.

However, selection into treatment status is plausibly based on characteristics unobserved by the econometrician. This implies that any given $\hat{\beta}_{OLS(W_{observed})}^C$ suffers from omitted variables bias to the extent there are W$_{\text{unobserved}}$ which are correlated with selection into treatment. Even conditioning on all observed variables, students with, say, more unobserved grit or ambitious parents, are likely to both have higher measured learning outcomes and to be enrolled in a private school.

One can *define* for any given country and any set of observables the OLS selectivity bias on unobservables (SBU) as the gap between the OLS estimate and the true treatment effect (or, by extension, a consistent estimate of SBU is the gap between OLS and a consistent estimate of the TE).

$$2)\ \widehat{SBU}^c \equiv \hat{\beta}^c_{OLS(W_{observeds})} - \beta^c_{TE}$$

Oster (2019) shows that a consistent estimate of the treatment effect of X in equation 1, $\tilde{\beta},$ can be recovered from observational data and some assumptions via equation 3.

$$3)\ \tilde{\beta} = \hat{\beta} - \delta(\dot{\beta} - \hat{\beta})\frac{\bar{R} - \hat{R}}{\hat{R} - \dot{R}}$$

Oster estimates require two empirical estimates and two assumptions. The two empirical estimates are: (i) the difference in the estimated β with and without W$_{\text{observed}}$, $\dot{\beta} - \hat{\beta},$ which, for a discrete variable X is just the raw difference in averages less the OLS estimate on X, and (ii) the difference in the regression R-squared without and with W$_{\text{observed}}$, $\hat{R} - \dot{R},$ how much higher the OLS R-squared is when the co-founders W$_{\text{observed}}$ are included.

In addition to the estimated quantities equation 3 requires two assumptions: (i) an assumption about a proportionality parameter, δ, between selectivity on the observables and unobservables and (ii) an assumption about the R-squared of equation 1 with the unobservables included. That is, $\bar{R}$ is the R-squared if both W$_{\text{observed}}$ and W$_{\text{unobserved}}$ were included in the regression, and this is usually parameterized as $\bar{R} = \Pi\hat{R}$.

Obviously neither the proportionality parameter δ or Π can be estimated from the data as they depend on “unobserved” variables. Oster (2019) does a review of the literature, comparing estimates of $\tilde{\beta}(\delta, \Pi)$ to estimates of treatment effects from other methods, like RCTs. Based on comparisons from the existing literature, she shows the assumptions of δ=1 and Π=1.3 are quite conservative, in that these assumptions would produce treatment effect estimates lower, not higher, than would result from consistent estimation methods. The proportionality assumption of δ=1implies that there is as much selectivity into treatment from the unobserved variables as from the unobserved. The assumption that Π=1.3 implies the inclusion of W$_{\text{unobserved}}$ would raise the R-squared by 30 percent. These values have become quite widely adopted.

A key innovation of this paper is to use estimates of treatment effects from the Oster (2019) method as consistent estimates (of lower bounds) of treatment effects as this allows, for the first time, the comparisons of large numbers of country estimates. Our two empirical examples, of wage gains from migration to the USA and of learning gains from private school both report Oster estimates with these values. I am going to treat the Oster (2019) estimates with those values as consistent estimates of treatment effects, which is likely to be conservative in that the “true” treatment effect is larger (in absolute value) as the absence of $W_{unobserved}$ from the estimation likely creates less SBU than the assumptions of δ=1, Π=1.3 imply.

I.B) Horse race for predictive accuracy

If we accept the $\tilde{\beta}^{c}_{Oster}(\delta = 1, \Pi = 1.3)$, or TE(O)^c, estimates as consistent estimates of the true TE for each country the Root Mean Square Error (RMSE) or Average Absolute Deviation (AAD) of prediction errors can be calculated for three “horse race” possibilities.

One, use the average of the TE(O)^c estimates across all countries as the prediction for each country. This prediction using SR($\overline{TE}$) mimics the “systematic review report of the average of the treatment effects” approach.

$$4)\ RMSE(\overline{TE(O)^c}) = sqrt(\frac{\sum_{c=1}^{c=N}\left(TE(O)^c - \overline{TE(O)}\right)^2}{N})$$

Two, naïve OLS predicts each country’s TE is its OLS estimate.

$$5)\ RMSE(\beta^c_{OLS}) = sqrt(\frac{\sum_{c=1}^{c=N}\left(TE(O)^c - \beta^c_{OLS}\right)^2}{N})$$

Three, the estimate for selectivity bias from unobservables (SBU) is, for each country, defined to be equal to the gap between the estimate of the treatment effect, TE(O) and the OLS estimate, which conditions on observeds.

$$6)\ RMSE(\overline{SBU(O)^c}) = sqrt(\frac{\sum_{c=1}^{c=N}\left(TE(O)^c - (\beta^c_{OLS} - \overline{SBU}(O))\right)^2}{N})$$

The formulae for the average absolute deviation, which is a more robust measure of predictive accuracy as it does not heavily penalize large prediction errors, are self-explanatory.

II) Estimates of gains from labor mobility

Suppose you were a developing country government (say, Guatemala) who was initiating a bilateral agreement to increase labor mobility with another country (say, the USA) and you wanted an estimate of the earnings gain to an incremental mover with specific characteristics (e.g., a given level of schooling). You would understand both that observational methods comparing the wages of Guatemalans in the USA to Guatemalans in Guatemala would fail to account for selectivity on unobserved covariates. At the same time, you would understand that rigorous, RCT, estimates of treatment effects from other pairs of countries might not have external validity for your country. Which would be better, just to rely on estimates using OLS on observational data about Guatemalans or to rely on rigorous evidence from other contexts?

II.A) Estimates of gains to migrants: Raw, OLS, TE(Oster)

Clemens, Montenegro, and Pritchett (2019) use US Census data and labor market surveys from 42 other countries, which jointly allow the comparison of the earnings differences (in PPP dollars, so “real” consumption units) between, say, people born in Guatemala and educated in

Guatemala (inferred from the USA census questions about a person's age at migration) working in the USA (migrants) versus those born and educated in Guatemala and working in Guatemala (non-migrants). For 42 countries CMP (2019) provide estimates of (i) the raw wage ratio of migrants and non-migrants, (ii) a standard OLS wage regression in the USA and in the sending country with observables (e.g., age, sex, education, sector, and urban residence to estimate the earnings wage ratio for "observationally equivalent" migrants and non-migrants (at specific values of the covariates) and (iii) an Oster (2019) lower bound, $\tilde{\beta}_{Oster}(1{,}1.3)$ (TE(O)^c).

Figure 1 shows four results.

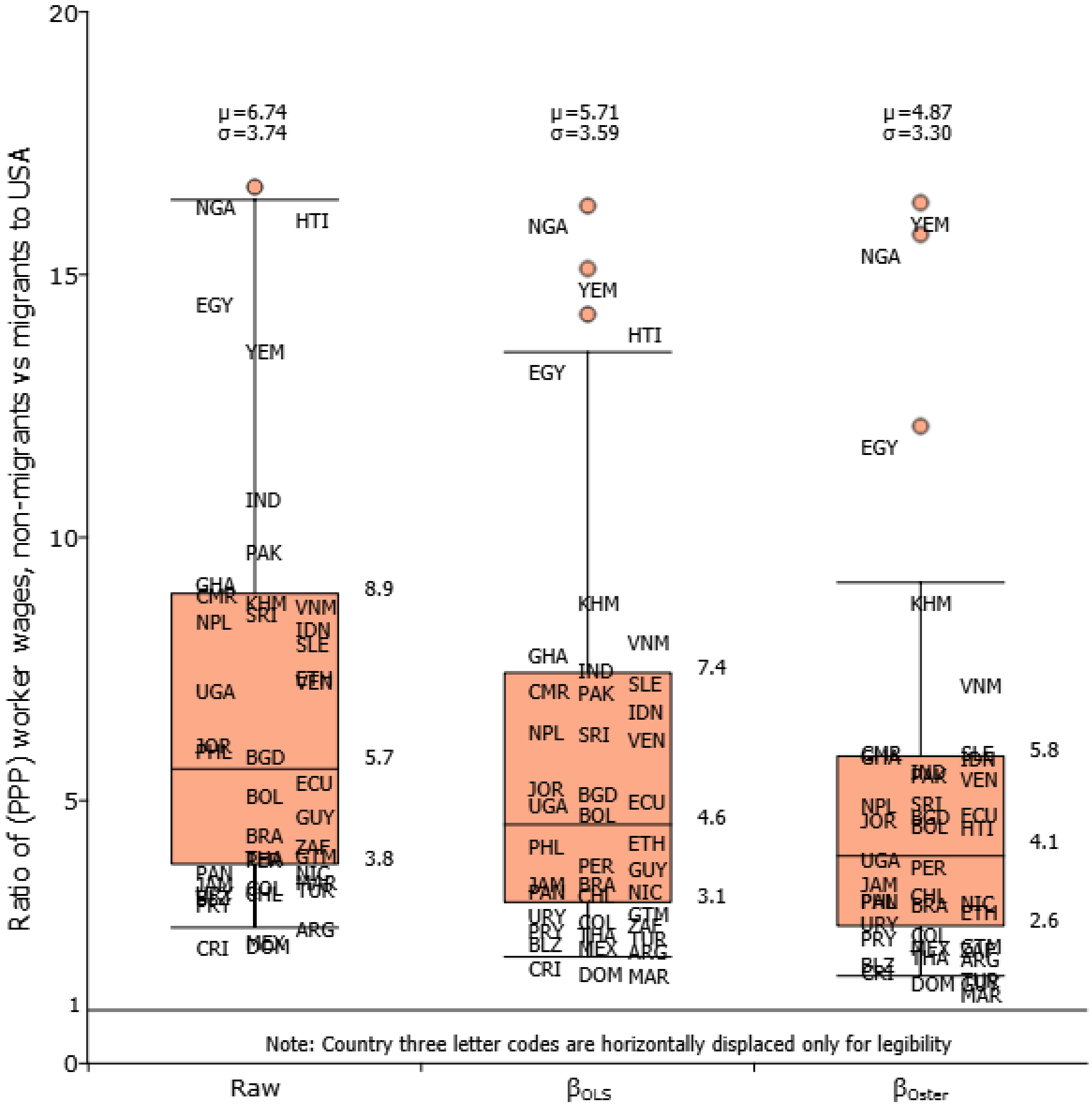
One, the TE(O) estimates of the wage gains are large. Averaged across the 42 countries, a randomly selected low-school worker would make 4.87 times (median 4.1) times higher earnings (in PPP) in the USA than in their home country. This is an average (labor force aged population weighted) wage gain of P$13,715 (in 2001 dollars). These estimates are consistent with a variety of other approaches to estimating the causal wage gains to a low skill worker moving from poor to rich country (Pritchett and Hani 2020).

Two, the variation in the TE(O) estimates across countries is substantial: some countries have very high estimates (Egypt at 12.1) while other countries have low estimates (Dominican Republic at 1.9). The 25th percentile (about Uruguay) is 2.6 and the 75th percentile (about Indonesia) is 5.8, more than twice as high. The standard deviation is 3.3

Three, there typically is quite strong positive migrant selectivity on observables. Positive migrant selectivity on the observed variables leads to OLS estimated wage ratios substantially lower than the raw ratio (5.71 versus 6.74). This in turn (via equation 3) produces $\tilde{\beta}_{Oster}(1{,}1.3)$ estimates lower than OLS, an average of 4.87 versus 5.71. The average OLS SBU of .84 (=5.71-4.87) is substantially less that the cross-national standard deviation of TE(O) of 3.3.

Fourth, (and this has to be inferred from Figure 1 by comparing the results for specific countries across the box-plots) the extent of selectivity on observables varies substantially across countries. Some countries have very small differences between the raw and OLS wage ratio estimates: the difference in Jamaica is 0, in Mexico is .09 and Peru is .10. In other countries there are very large differences: the difference for India is 3.26, for Ethiopia it is 3.14, and for Morocco is 1.81). Migrant selectivity bias on unobserved productivity in estimating wage gains is often raised in the context of highly skilled professions (e.g., doctors, engineers, academics) and economic "superstars" (e.g., CEOs, entrepreneurs) but the CMP (2019) estimates are for workers with less than high school completed (9-12 years of schooling) for which massive "long-tailed" selectivity wage gains are likely less common.

Figure 1: Comparison of Raw, OLS and Oster estimates of ratios of PPP wages for migrants to home country workers from 42 different countries to the USA



Source: Author's graph based on Clemens, Montenegro and Pritchett 2019 estimates.

II.B) There is no single interpretation of "rely on the rigorous evidence"

There is a well identified estimate of the wage gains to migrants from a program in New Zealand that allowed Tongan workers to migrate on a temporary basis for agricultural work (McKenzie, Gibson and Stillman 2010). Given an oversubscription of visa applicants, recipients were chosen randomly from the applicants, which allow researchers to correct for the potential bias from self-selection of applicants on unobservable characteristics. Their estimated TE was a wage gain ratio of 3.63. This same study also reported an OLS estimate of the migrant/home wage ratio of 4.83 so that the estimated magnitude of the OLS selectivity bias on unobservables (SBU) was 1.2 (=4.83-3.63).

This study (as pretty much any RCT study could) produced not one, but two, pieces of rigorous evidence: an estimate of the treatment effect of 3.63 (ratio of wage gain) and an estimate of the SBU of 1.2. This implies there are two equally valid interpretations of "rely on the rigorous evidence." One is to use SR($\overline{TE}$) the average estimate of the rigorous estimates of treatment effects. The other possibility is SR($\overline{SBU}$), use the average estimates of the SBU. The conceptual problem is that, given the large heterogeneity across countries in the OLS evidence about wage gain ratio in Figure 1, these two, equally plausible, approaches to the rigorous evidence: (i) will give will give contradictory advice about how to adjust the OLS evidence to produce an TE estimate and (ii) either interpretation will generate completely implausible implications.

As a simple example, suppose that at first the only evidence we had about wage gains were the OLS regression results for Guatemala (GTM in Figure 1, $\hat{\beta}_{OLS} = 3.2$) and for Bangladesh (BGD in Figure 1, $\hat{\beta}_{OLS} = 5.5$) and these informed our priors to those countries. Then the McKenzie et al (2010) study was done for Tonga-NZ, which on the assumption this was, at the time, the only "rigorous" study by the filter of a systematic review, implies SR($\overline{TE}$)=3.63 and SR($\overline{SBU}$)=1.2. What would "rely on the rigorous evidence" mean?

The SR($\overline{TE}$) interpretation would imply that we should revise our estimate of the wage gains in Guatemala *upward* from 3.2 to 3.7. The SR($\overline{SBU}$) interpretation would imply we should revise our estimate of the wage gains in Guatemala *downward* from 3.2 to 2.0. Moreover, the SR($\overline{TE}$) interpretation suggests we should revise our estimate of the wages gains in Bangladesh *downward* from 5.5 to 3.7. This *necessarily* means the SR($\overline{TE}$) interpretation of "rely on the rigorous evidence" requires us to believe, given the identity in equation 2, that the SBU for Tonga-NZ is 1.2, the SBU for Guatemala is *negative* .5 (implying migrants are negatively selected and hence that OLS is too low relative to the true treatment effect) and that the SBU for Bangladesh is 1.8.

Things don't get any better from this simple example if there are multiple rigorous studies, as the two potential interpretations of "rely on the rigorous evidence" still contradict each other and either interpretation has obviously counter-factual implications. Suppose we treat the TE(O)^c as the rigorous evidence a systematic review would be based on. If we assume treatment effects have external validity then all countries should believe their country's wage gain ratio should be SR($\overline{TE}$)=4.87 (equation 7a). However, if we assume estimates of the OLS SBU have external validity then each country should believe that the wage gain ratio for their country should be the OLS estimate less the average estimated SBU (equation 7b).

7a) $\beta_{c,USA}^{True} = SR(\overline{TE}) = 4.87$

7b) $\beta_{c,USA}^{True} = \hat{\beta}_{c,USA}^{OLS(Wobs)} - SR(\overline{SBU}),\ where\ SR(\overline{SBU}) = mean(\beta_{c,USA}^{OLS} - \tilde{\beta}_{c,USA}^{Oster})$=.84

The identity in equation 2 linking OLS, SBU and the true treatment effect implies equation 7c

$$7c)\ \beta^{OLS}_{c,USA} \equiv \beta^{True}_{c,USA} + SBU_{c,USA}$$

The OLS estimate for each country c, $\beta^{OLS(Wobs)}_{c,USA}$is just an empirical fact (the determinate outcome of applying a given statistical procedure to a given data set) and cannot be freely chosen.

Table 1 illustrates five serious logical and empirical problems with assuming external validity.

First, the gaps between the TE(O)^{c} estimates and SR($\overline{TE}$)=4.87 are large and arbitrary and there is no theoretical or empirical justification for believing there is external validity and the country specific TE(O)^{c} estimates are just mistaken.

Second, assuming external validity of TE and adopting that each country's estimate should SR($\overline{TE}$) implies the cross-national standard deviation of the "true" TE is zero (column V). But the cross-national standard deviation of the TE(O)^{c} estimates is 3.30. Hence assuming external validity, SR($\overline{TE}$), implies strongly counter-factual beliefs about the cross-national variation in the true TEs. The same problem arises with assuming external validity about the bias, taking SR($\overline{SBU}$) as the rigorous estimate of the SBU for all countries, as the estimated standard deviation of SBU is 1.5.

Third, column VII shows the estimate of the OLS SBU for each country is implied by the identity in equation 2 and SR($\overline{TE}$). Given India's OLS estimate of 7.86 the implied SBU is 2.99 (7.86-4.87) which implies that Indian (low schooled) migrants to the USA are strongly positively selected on unobservables, more strongly than India's Oster estimated SBU estimate of 1.93 (column IV). Conversely, the OLS estimate for the Dominican Republic is 2.08, which produces an OLS SBU estimate implied by SR($\overline{TE}$) of -2.79 (=2.08-4.87) which implies (low skill) workers from the Dominican Republic are massively *negatively* selected on unobservables—even though the Oster estimate for the Dominican Republic suggests migrants are modestly *positively* selected on unobservables, at .18 (=2.08-1.90).

Moreover, the combination of assuming external validity of treatment effects, which implies zero variation in the true TE across countries and the actual variation of the OLS estimates (which are an empirical fact) implies that all of the variation in OLS versus SR($\overline{TE}$) must be due to variation in the country SBU, which implies a variation in the SBU of 3.59, much higher than its estimated value of 1.5.

Fourth, the estimates of the OLS SBU implied by the assumption of external validity of SR($\overline{TE}$) bear no relationship to the actual country specific empirical estimates of the SBU, $\beta^{OLS(Wobs)}_{c,USA}$-$\tilde{\beta}^{c,USA}_{Oster}$, or common sense, or the existing literature. Excluding one outlier country, Haiti, the correlation between the Oster SBU estimates in column IV and the SR($\overline{TE}$) estimates in column VII is modestly *negative*, at -.11. One implication of Column VII estimates of the OLS SBU is that 22 of 42 countries have *negative* selectivity on unobservables in spite of *positive* selection on observables, which would be extremely odd. Moreover, empirical estimates of migrant selectivity across a large number of countries suggest positive selectivity on both observables and unobservables (Clemens and Mendola 2020).

Fifth, the convention that “rely on the rigorous evidence” implies some assumptions about external validity (as otherwise it is obvious an RCT evidence from one country isn’t rigorous evidence at all for any other country), but there is no logical or theoretical reason to believe that treatment effects have external validity versus that selection bias has external validity—and both cannot have external validity. For instance, the OLS estimate for Thailand is 2.83 (column II). With SR($\overline{TE}$) (external validity of treatment effects) the estimate would be 4.87 (column V), much higher, with SR($\overline{SBU}$) the estimate would be 1.99 (column VIII), which is much lower. Does “rely on the rigorous evidence” mean SR($\overline{TE}$) or SR($\overline{SBU}$)? It cannot mean both and there is no rational reason to prefer one over the other.

Table 1: Thought experiment comparing SR($\overline{TE}$) and SR($\overline{SBU}$) with actual country specific OLS and TE estimates

Country	Actual country estimates				SR($\overline{TE}$)			SR($\overline{SBU}$)		
	Raw	OLS	Oster	Oster estimated OLS SBU	Average of Oster Estimates	Gap with SR($\overline{TE}$)	Implied OLS SBU	Estimate	Gap with Oster estimate	Gap SR($\overline{SBU}$) and SR($\overline{TE}$)
Column number:	I	II	III	IV (=II-II)	V	VI (=V-III)	VII (=II-V)	VIII (=II-.84)	IX (=VIII-III)	X (=.84-IV)
Highest ten wage ratio countries by Oster estimate										
Yemen	13.92	15.11	16.37	-1.25	4.87	11.50	10.24	14.28	2.09	9.41
Nigeria	16.67	16.31	15.76	0.54	4.87	10.89	11.44	15.47	0.29	10.60
Egypt	14.82	13.53	12.12	1.41	4.87	7.25	8.66	12.69	-0.57	7.82
Cambodia	9.13	9.14	9.15	-0.01	4.87	4.28	4.27	8.30	0.85	3.43
Vietnam	9.06	8.40	7.55	0.84	4.87	2.68	3.52	7.56	0.00	2.69
Cameroon	9.27	7.48	6.29	1.19	4.87	1.42	2.61	6.64	-0.35	1.77
Sierra Leone	8.35	7.61	6.27	1.34	4.87	1.40	2.74	6.77	-0.50	1.90
Ghana	9.51	8.16	6.23	1.93	4.87	1.36	3.29	7.32	-1.09	2.45
Indonesia	8.64	7.07	6.19	0.88	4.87	1.32	2.20	6.23	-0.04	1.36
India	11.12	7.86	5.93	1.93	4.87	1.06	2.99	7.02	-1.09	2.15
Average	6.74	5.71	4.87	0.84	4.87	0.00	0.84	4.87	0.00	0.00
Std. Dev.	3.74	3.59	3.30	1.50	4.87	3.30	3.59	3.59	1.50	3.59
Smallest ten wage ratio countries by Oster estimate										
Mexico	2.68	2.59	2.56	0.03	4.87	-2.31	-2.28	1.75	0.81	-3.12
South Africa	4.49	2.99	2.52	0.46	4.87	-2.35	-1.89	2.15	0.38	-2.72
Thailand	4.30	2.83	2.40	0.43	4.87	-2.47	-2.04	1.99	0.41	-2.88
Argentina	2.93	2.49	2.36	0.12	4.87	-2.51	-2.38	1.65	0.72	-3.22
Belize	3.52	2.63	2.25	0.39	4.87	-2.62	-2.24	1.80	0.45	-3.07
Costa Rica	2.58	2.19	2.10	0.10	4.87	-2.77	-2.68	1.36	0.74	-3.51
Turkey	3.68	2.74	1.95	0.79	4.87	-2.92	-2.14	1.90	0.05	-2.97
Guyana	5.08	4.07	1.90	2.17	4.87	-2.97	-0.80	3.23	-1.33	-1.64
Dom. Rep.	2.62	2.08	1.90	0.19	4.87	-2.97	-2.79	1.25	0.65	-3.62
Morocco	3.84	2.03	1.67	0.36	4.87	-3.21	-2.84	1.19	0.48	-3.68
Root Mean Square Error			1.70			3.26			1.48	
Average Absolute Deviation			.90			2.21			0.77	

Source: Author’s calculations with estimates from CMP(2019), table 2.

These five problems are quite general, in three senses.

One, one could use any single values of TE or SBU derived in whatever way from any set of rigorous estimates, that is, the "systematic review" could be filtered in any way, and still have exactly the same five issues.

Two, one could modify equation 7a to 7d so the country specific prediction was a weighted average of the OLS and SR($\overline{TE}$) with any α on SR($\overline{TE}$) (7a with α=1 is a special case). One still has all the same five problems, just moderated somewhat (Pritchett and Sandefur 2014)—and one loses the rhetorical appeal to "rigorous" as neither the OLS estimates nor the weight α are "rigorous."

$$7d)\ \beta_{c,USA}^{True} = (1-\alpha) * \beta_{c,USA}^{OLS(W)} + \ \alpha * \beta_{c,USA}^{SR(\overline{TE})}$$

Three, one could combine equations 7a and 7b so that the estimate of the "true" TE in country c was a weighted combination of the average treatment effect and the OLS adjusted for the average selection bias, equation 7e. This however implies that "rely on the rigorous evidence" can mean pretty much anything, depending on the choice of the weight θ. For instance, the treatment effect for Belize, estimated ratio of wages in the USA to wages in Belize for an equal productivity mover, could be anywhere from 4.87 (θ=1, column 5) to 1.80 (θ=0, column VIII), including producing as the "rigorous" estimate the OLS estimate of 2.63 using θ=.27 or the Oster estimate of 2.25 with θ=.15, or with the arbitrary but focal point of equal weights, θ=.5, one could estimate the true gain as 3.33.

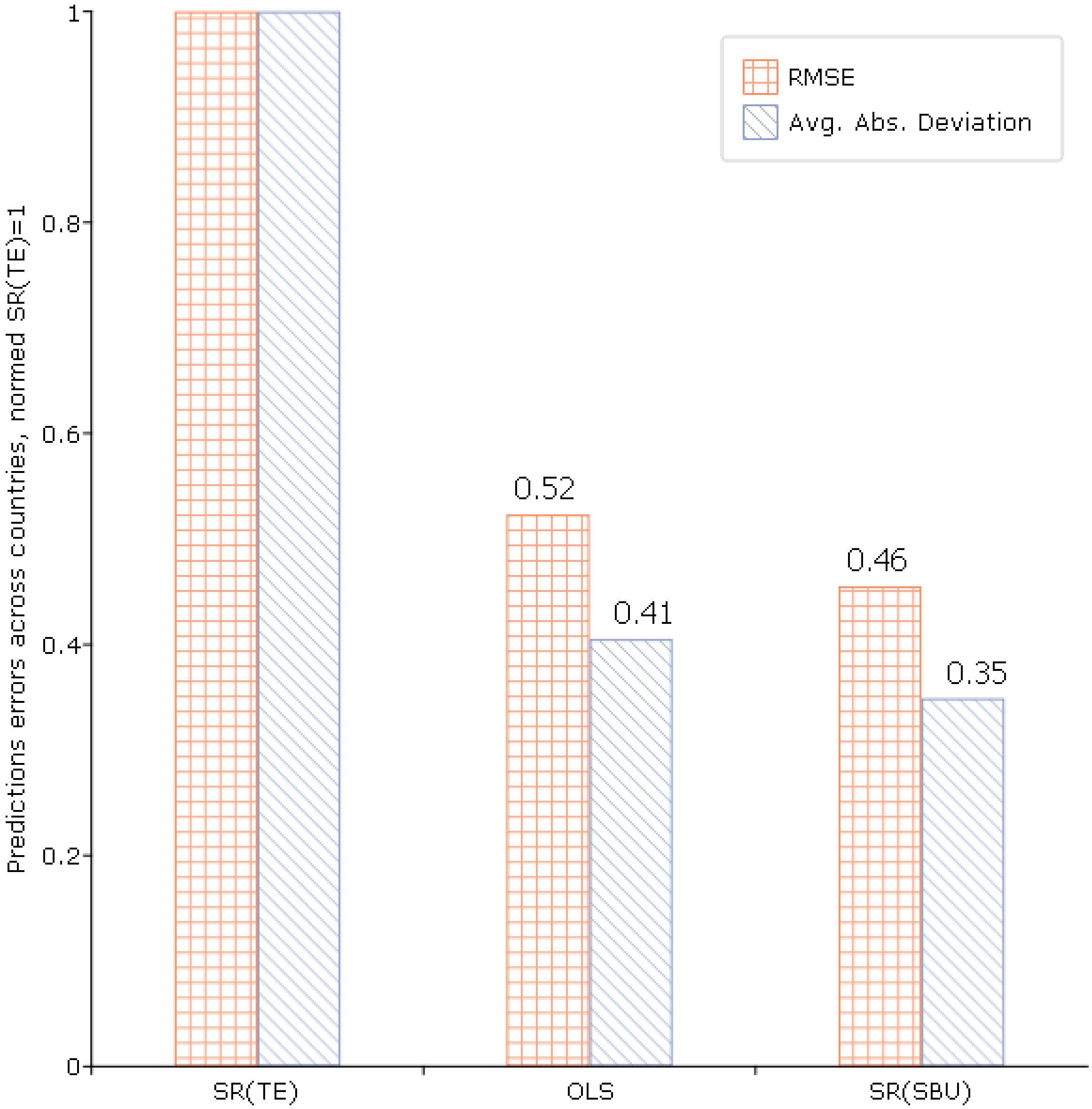
$$7e)\ \beta_{c,USA}^{True} = (1-\theta) * (\hat{\beta}_{c,USA}^{OLS(Wobs)} - SR(\overline{SBU})) + \ \theta * SR(\overline{TE})$$

The slogan "rely on the rigorous evidence as summarized by systematic reviews" is vacuous, but any specific interpretation of that faces enormous challenges in even achieving logical coherence and even minimal empirical plausibility. A standard approach is for systematic reviews to completely ignore the OLS evidence, so that α=1, and completely ignore the estimates of the SBU, so that θ=1, and hence the slogan is the special case, RORE(1,1) which implies the best prediction for each country is SR($\overline{TE}$). This however implies the assertion that TEs have external validity but estimates of SBU have no external validity, and worse, the estimates of SBU implied by SR($\overline{TE}$) have to take on country by country values that are an arbitrary set of measure zero. The existing systematic review never acknowledges these conceptual challenges by "feigned ignorance" (Pritchett 2020), which ignores that the OLS estimates exist and ignores that studies which can produce a rigorous estimates of TE also (can) produce OLS and hence via an identity, rigorous estimates of the SBU of OLS.

II.C) SR($\overline{TE}$) *produces worse cross-national predictions than OLS or RORE(bias)*

SR($\overline{TE}$) produces worse predictions of the "true" wage gains. Figure 2 (and the bottom two rows of Table 1) show the results of the horse race, where the RMSE and AAD of SR($\overline{TE}$) is normed to 1. OLS country by country produces a RMSE about half that of SR($\overline{TE}$)), and performs even worse for AAD. SR($\overline{SBU}$) outperforms either SR($\overline{TE}$) (by a wide margin) or OLS (by a modest margin).

Figure 2: Prediction errors from SR($\overline{TE}$), using the average estimated treatment effect from an systematic review, the standard interpretation of "rely on the rigorous evidence," are much worse than OLS or SR($\overline{SBU}$)



Source: Author's calculations with data in Figure 1 and Table 1.

The intuition of this result is clear. In this data standard deviation of the TE(O)^c of 3.3 (Table 1) is much larger than the typical OLS SBU of .84. Hence, ignoring the cross-national variance in the true TE in order to use an average of the "rigorous" estimates leads to worse predictions on average.

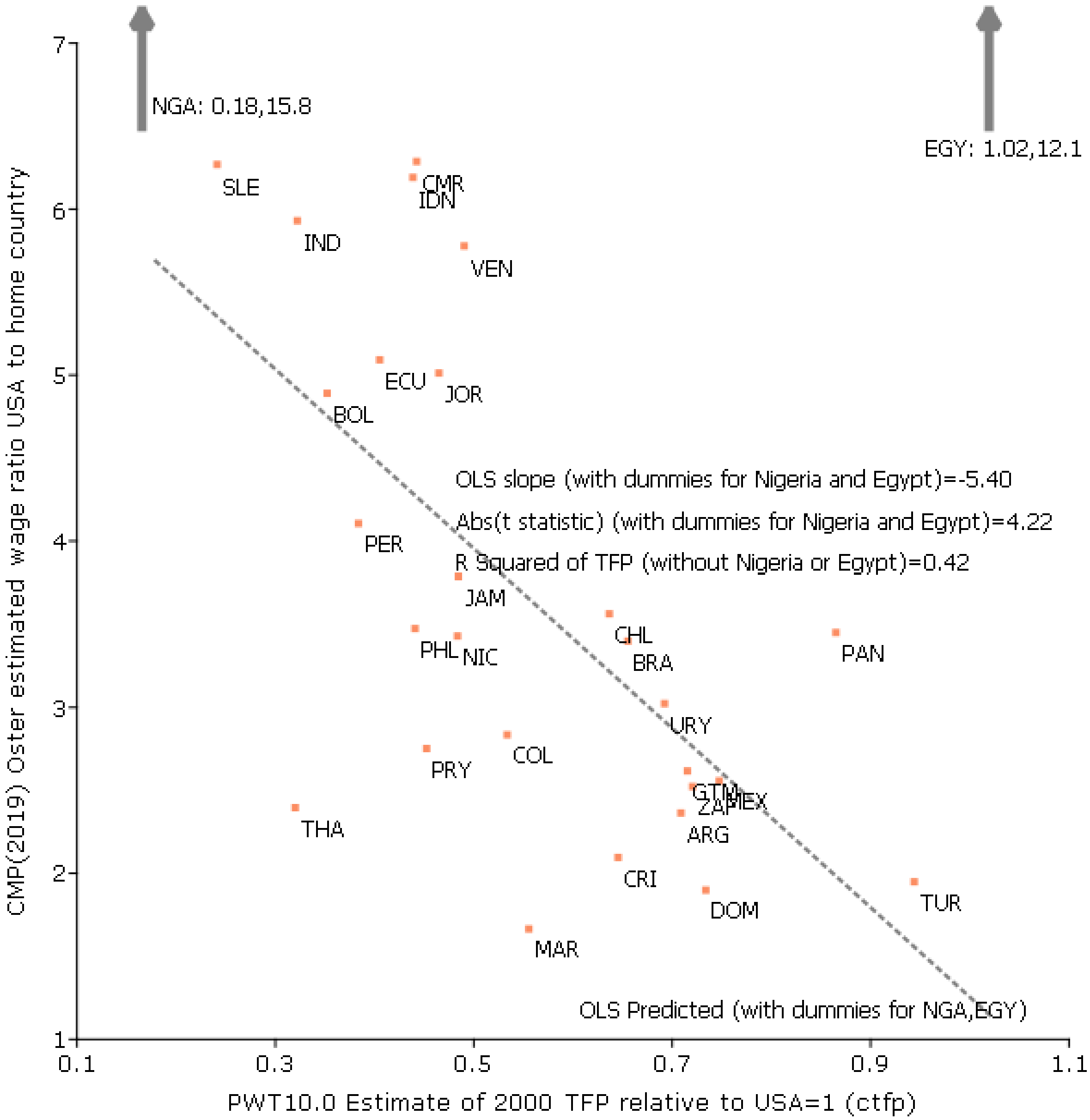
II.D) SR($\overline{TE}$) *is not based on any plausible theory*

A prediction of the wage gain from actions that allow a worker *i* to move from country *h* to country *d* should be based on our best available model predicting the wages of the mover in those two places. Implicitly, assumptions of external validity of estimated policy relevant quantities and the proposed method of SR($\overline{TE}$) assume that most (or all) of the variation in the biased observational estimates of the treatment effect are due to flawed methods and not due to true cross-contextual variation in the treatment effect. Therefore if substantial variation in the TE(O)^c estimates are associated with *any* model of cross-national wage differences for workers with equal intrinsic (personal) productivity this assumption is false and hence SR($\overline{TE}$) is scientifically dubious.

In Solow-Swan models there are cross-national differences in A or TFP and these differences imply different marginal products of factors, capital or human capital. If factors are paid their marginal product then wages in countries *h* and *d* for a worker with same human capital will be higher where A is higher. Figure 3 shows the scatter-plot of the 29 countries that have both a Penn World Tables 10.0 estimate of TFP relative to the USA at current PPPs and also a CMP(2019) estimate of wage differences. The association between country level TFP relative to the USA and the estimated TE(O)^c wage ratios is strong and negative (there are two large, Nigeria (NGA) and Egypt (EGY), and the regression includes a dummy for each of those countries). The regression is strongly consistent with the idea that equal intrinsic productivity workers gain more by moving to the USA from moving from countries with lower TFP relative to the USA.

This is not so say this simply cross-national association based on a simple model of aggregate output is the “best” model of gains from migration, the point that even this “quick and dirty” economics suggests that “external validity” cannot be assumed as it is not based on *any* economics at all and contradicts even simple, but empirically validated models.

Figure 3: Wage gains of movers to the USA and TFP relative to the USA are strongly (negatively) correlated



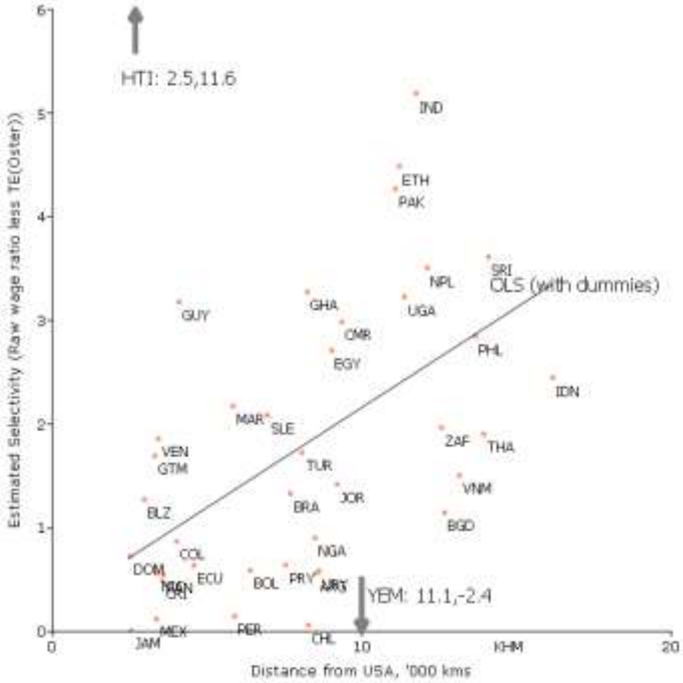
Source: Author's calculations with CMP (2019) estimates. OLS regression includes dummy variables for the extreme observations for Nigeria (NGA), Egypt (EGY) (whose data are indicated in the graph).

II.E) SR($\overline{SBU}$) is also not based on any valid theory

The primary reason why it is so difficult to recover reliable causal estimates about economic phenomena from observational data is that the data reflect the results of agents making purposive decisions. But this implies that the *magnitude* of selectivity bias depends on the underlying economics of the choices of agents make, subject to the constraints they face. This then implies that the selectivity bias, on both observables and unobservables, may vary from context to context. Assuming external validity for $SR(\overline{SBU})$ faces the same problems as assuming external validity for treatment effects, the assumption is not based on the any (much less the best available) understanding/model/theory.

For instance, it is plausible that the higher the fixed costs of a given move the larger the selectivity bias, as only those who anticipate larger gains are willing to make the move. Figure 4 shows a simple scatter plot between an estimate of selectivity of migrants on both observables and unobservables (the gap between the country raw wage ratio and $TE(O)^c$) and the distance from the country to the USA (Meyer and Zignago 2011). As can be seen there are some massive outliers (Haiti, Yemen, Cambodia (KHM)) but if one allows for dummy variables there is a strong positive association between distance and estimates of selectivity bias consistent with a simple economic model. Again, the point is not that Figure 4 illustrates a complete and correct model of cross-national differences in the selectivity bias of wage gains for (low schooling level) migrants, but rather only that assuming "external validity" of bias estimates has no justification as the best available understanding of the migrant selection process as "external validity" of selectivity would imply that the variance in selectivity across countries should not be predictable on the basis of *any* economic model.

Figure 4: Estimated association of impact of selectivity in migration to the USA on estimated wage gains and distance to the USA



Source: Author's calculations from CMP (2019) estimates and data on distance from CPEII. *OLS regression includes a binary variable for Haiti (HTI), Yemen (YEM), and Cambodia (KHM).*

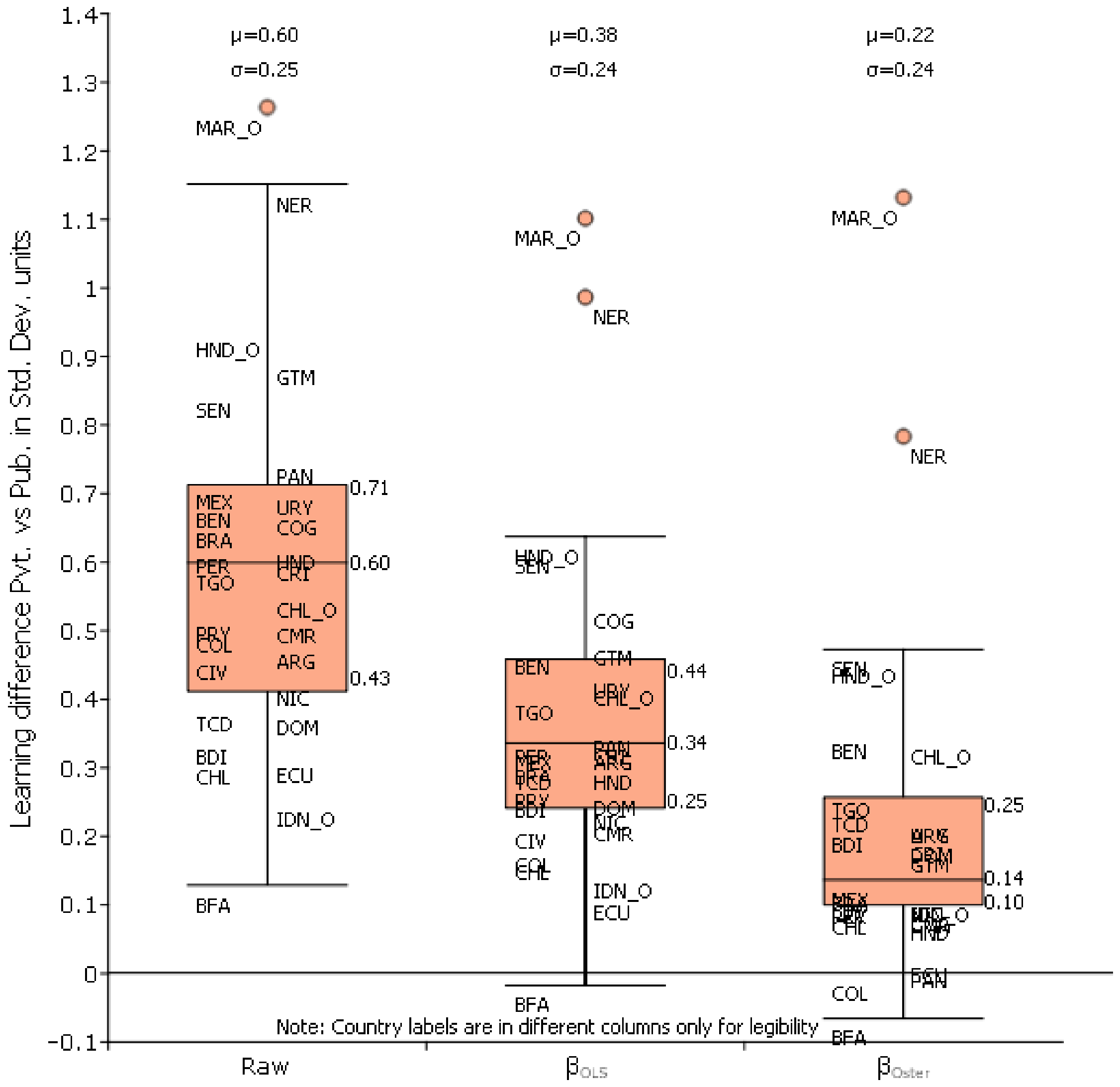
III) *Second empirical example: Private School Learning Premium*

The section above is a complete paper and lays out the insuperable conceptual and empirical objections to adopting the implicit assumptions embedded in the apparently straightforward and seemingly good advice to "rely on the rigorous evidence." This section add to that by showing that all of the conceptual and empirical points made above using the cross-national data on wage ratios hold for a completely different phenomena, the private sector learning premia. This is important in reassuring the reader the wage ratio example was not a "special case" that is uniquely unfavorable for the case for "rely on the rigorous evidence" but that the problems raised by large variability across countries in the true TE are generic. Given the space constraints of this paper, the hopefully explication of the points above, and the fuller discussion in Pritchett (2021), this section will be telegraphic.

A recent study by Patel and Sandefur (2020) uses a "Rosetta Stone" approach of giving students in a single setting an assessment with items from different assessments to create comparable estimates of mathematics capabilities for large samples of individual students in 29 developing countries. They estimated: (i) the "raw" private sector learning premium (PSLP), (ii) an OLS estimate of the PSLP controlling for a set of student and household covariates, and (iii) TE(O)^c estimates of the PSLP, using δ=1 and Π=1.3.

Figure 5 shows the same four key points about the distribution of the empirical estimates of the PSLP as shown for the wage ratios. One, on average the TE(O) is substantial, an average of .22 standard deviations (a median of .14, as the estimates are skewed). Two, the heterogeneity of the TE(O) is substantial, the 25th-75th spread is .15 and standard deviation is .24. Three, there is quite strong selectivity on observables as the average OLS is .38 versus the average raw PSLP is .60. Four, (and again this has to be inferred from compared across the box-plots) there are substantial differences in the degree of selectivity on observables and unobservables, from quite small for Morocco (MAR_O) to very large for Mexico (MEX), which falls from a raw of .72 to a TE(O) of only .14.

Figure 5 Box-plots of raw, OLS, and TE(O) estimates of the Private Sector Learning Premium (PSLP)



Source: Patel and Sandefur (2020). The notation ABC_O in the country labels (e.g. CHL_O) for Chile) indicates estimates "original" data, not based on the "Rosetta Stone" estimates.

Table 2—same calculations as Table 1--illustrates with PSLP data that the assumption of external validity of the TE estimates and the identity linking OLS, TE and bias necessarily leads to unreasonable implications. For instance, in Ecuador (ECU) the estimates imply strong positive selectivity bias on observables, the gap between the raw and the OLS is .20 (.32-.12), but the SR($\overline{TE}$) estimate of .22 implies selection into private school on unobservables must be *negative*, as .22 is *higher* than Ecuador's OLS estimate of .12. SR($\overline{TE}$) implies there are five countries with *positive* selectivity on observables but *negative* selectivity on unobservables. Conversely, in Niger (NER) the selection on observables is relatively strong and the TE(O) is .20

units lower than the OLS. But the SR($\overline{TE}$) of .22 implies the OLS SBU in Niger was not .20, but three times larger, .76. Again, these dubious empirical implications are *necessarily* implied by the SR($\overline{TE}$) assumption of the external validity of TE estimates.

Table 2: Estimates of the Private Sector Learning Premium, SR($\overline{TE}$) and SR($\overline{SBU}$)

Country	Patel and Sandefur (2020) estimates				SR($\overline{TE}$)			SR($\overline{SBU}$)		
	Raw	OLS	TE(O)	Implied OLS SBU	Avg. TE(O)	Gap with TE(O)^c	Implied OLS SBU	TE estimate	Gap with TE(O)^c	SBU Gap
Columns:	I	II	III	IV	V	VI (=V-III)	VII (=II-V)	VIII (=II-V(avg)	IX (VIII-III)	X (=IV-IV(avg)
BFA	0.13	-0.02	-0.07	0.05	0.22	0.29	-0.24	-0.17	-0.10	0.10
COL	0.51	0.19	0.00	0.19	0.22	0.23	-0.04	0.04	0.04	-0.04
PAN	0.75	0.36	0.02	0.34	0.22	0.21	0.13	0.21	0.19	-0.19
ECU	0.32	0.12	0.02	0.09	0.22	0.20	-0.11	-0.04	-0.06	0.06
HND	0.63	0.31	0.09	0.22	0.22	0.14	0.08	0.16	0.07	-0.07
CHL	0.31	0.18	0.10	0.08	0.22	0.13	-0.05	0.03	-0.07	0.07
CMR	0.52	0.23	0.10	0.14	0.22	0.13	0.01	0.08	-0.02	0.02
COG	0.68	0.54	0.10	0.44	0.22	0.12	0.32	0.39	0.29	-0.29
PRY	0.52	0.28	0.11	0.17	0.22	0.11	0.05	0.13	0.02	-0.02
PER	0.62	0.34	0.11	0.23	0.22	0.11	0.12	0.19	0.08	-0.08
NIC	0.43	0.25	0.11	0.14	0.22	0.11	0.02	0.10	-0.02	0.02
IDN_O	0.25	0.15	0.12	0.03	0.22	0.11	-0.08	0.00	-0.12	0.12
CIV	0.47	0.22	0.12	0.11	0.22	0.11	0.00	0.07	-0.05	0.05
BRA	0.66	0.32	0.13	0.18	0.22	0.09	0.09	0.16	0.03	-0.03
MEX	0.72	0.34	0.14	0.20	0.22	0.09	0.11	0.19	0.05	-0.05
GTM	0.90	0.49	0.19	0.30	0.22	0.04	0.27	0.34	0.15	-0.15
DOM	0.39	0.27	0.20	0.07	0.22	0.02	0.05	0.12	-0.08	0.08
CRI	0.61	0.35	0.20	0.14	0.22	0.02	0.12	0.19	-0.01	0.01
BDI	0.34	0.27	0.22	0.05	0.22	0.01	0.04	0.11	-0.10	0.10
URY	0.71	0.44	0.23	0.21	0.22	0.00	0.22	0.29	0.06	-0.06
ARG	0.48	0.34	0.23	0.11	0.22	0.00	0.11	0.18	-0.04	0.04
TCD	0.39	0.31	0.25	0.06	0.22	-0.02	0.08	0.15	-0.09	0.09
TGO	0.60	0.41	0.27	0.14	0.22	-0.04	0.18	0.26	-0.01	0.01
CHL_O	0.56	0.43	0.34	0.09	0.22	-0.12	0.20	0.28	-0.07	0.07
BEN	0.69	0.48	0.35	0.13	0.22	-0.13	0.25	0.33	-0.03	0.03
HND_O	0.94	0.64	0.46	0.17	0.22	-0.24	0.41	0.49	0.02	-0.02
SEN	0.85	0.62	0.47	0.15	0.22	-0.25	0.40	0.47	0.00	0.00
NER	1.15	0.99	0.78	0.20	0.22	-0.56	0.76	0.83	0.05	-0.05
MAR_O	1.26	1.10	1.13	-0.03	0.22	-0.91	0.88	0.95	-0.18	0.18
Mean	0.60	0.38	0.22	0.15	0.22	0.00	0.15	0.22	0.00	0.00
Median	0.60	0.34	0.14	0.14	0.22	0.09	0.11	0.18	-0.01	0.01
RMSE		0.18				0.24			0.10	
Avg. Abs. Dev.		0.15				0.16			0.07	

Source: Author's calculations with estimates from Patel and Sandefur (2020).

Again the SBU estimates implied by assuming external validity of treatment effects must fall onto exactly (the arbitrary set of measure zero) results in column VII. As argued above, there is no reason to prefer SR($\overline{TE}$) over SR($\overline{SBU}$) as it is *a priori* at least as plausible there is cross-country external validity in the estimates of the OLS SBU and hence that the country specific estimates of the TE should be the result of adjusting the country specific OLS estimates for the average cross-national estimated bias (as in Column VIII).

Table 2 shows the RMSE and AAD from using either SR($\overline{TE}$), country specific OLS, or SR($\overline{SBU}$), assuming TE(O)^c estimate is the "true" TE for each country. As with wage ratios, the RMSE error for SR($\overline{TE}$) is more than twice as large as that for SR($\overline{SBU}$) (.24 vs .10) and larger than for OLS (.24 vs .18). In this case there is a modest caveat as the AAD is only slightly lower for OLS than SR($\overline{TE}$) and the RMSE without Morocco is slightly larger for OLS than for SR($\overline{TE}$).

The PSLP results reinforce the intuition that SR($\overline{TE}$) will produce worse prediction errors than OLS when the cross-national variation in the true TE is large relative to the selectivity bias (average OLS SBU). With wage ratios the variation in the true TE was large and OLS SBU modest, whereas with the PSLP these are of roughly similar magnitude and hence SR($\overline{TE}$) and OLS RMSE prediction errors (excluding Morocco) are quite similar.

A model that the PSLP is constant across all countries is easily rejected. In the Patel and Sandefur (2020) data there is a strong negative, non-linear association (R2 of .305) between the TE(O)^c PSLP estimates and the math assessment results in the public sector. One need not have a complete and fully articulated model of the cross-national variation in the PSLP to think it is plausible that some governments are reasonably effective and can produce learning outcomes are near the efficiency frontier and hence in those cases the PSLP will be low. But when governments are not effective (and economists have no validated positive model suggesting all governments are equally effective) they may be very bad at producing learning. This low government efficacy creates the possibility of a private sector outperforming the public sector by a wide margin and the PSLP is high.

Similarly, there is no plausible case that there is external validity in the estimates of selectivity bias. Patel and Sandefur (2020) show (Figure 17 in their paper) that the measured total selectivity bias—the gap between the raw PSLP and the TE(O) estimate—is associated with the country's income inequality. Countries with larger inequality (e.g. Guatemala and Honduras) tend to have a higher selectivity bias than do lower income inequality countries (e.g. Indonesia or Morocco). Assuming equal selectivity bias across countries is not consistent with the data.

Conclusion

This paper, together with Pritchett and Sandefur (2015), makes the horse race score in predicting the actual country specific treatment effects or causal impacts 3-0, with "rigorous evidence" at zero. Across three very different subjects--micro-credit, wage gains from migration, private school learning premium—using naïve OLS from the specific country gives better RMSE than "rely on the rigorous evidence" interpreted as SR($\overline{TE}$). And worse, as emphasized in Pritchett and Sandefur (2014), the standard approach to "systematic reviews" which focuses on summarizing treatment effects on the premise these summaries have evidentiary value relies on assumptions about external validity that are indefensible; conceptually (why is it that TE estimates have external validity and not estimates of bias?), theoretically (economic models predict heterogeneity of TE and are inconsistent with external validity across countries), or empirically (external validity of TE implies impossible beliefs about the cross-national variance in observational estimates).

To conclude on a more positive note, the alternative to the "evidence based" approach to decision making is "understanding based" decision-making. This paper had its origins as an encomium to Edward Leamer, who, for me, emphasized that the goal of empirical economics was a correct understanding rather than methodological purity. A correct understanding of the relevant phenomena needed to be capable of *encompassing* all of the available evidence into an overall theory or model or narrative, which in turn means that our understanding needs to be dynamic, as even past reliable empirical associations can break down in new circumstances (Leamer 2010). In a social science like economics this is the sense understanding of the German word *verstehen* (a concept whose consequences for method were elaborated (for me) by Gadamer (1975)), an interpretive understanding, while, in a social science like economics, appreciating that this interpretive understanding needs to encompass and embed empirical findings. And the application of a correct understanding to concrete decisions needs to reflect something like the Greek word *phronesis*, or practical wisdom. And the implementation that leads to improvements often relies on knowledge as both *metis* and *techne*, as articulated by Scott (1998). In this sense, many of papers that result from RCTs are important, not because of their specific numerical findings, or that their results have immediate applicability to "policy," or that the results have external validity (or even because I agree with the authors' own views of their findings), but because, as well-documented and empirical anecdotes about directed perturbations to the existing equilibria of complex systems, they force us to adapt our interpretive understanding and expand our collective practical wisdom and hence point to potential new pathways of betterment. I don't "reject" RCTs as a method, there are dozens of RCT studies that have changed my understanding and which I routinely use in my writing and teaching (e.g., Olken (2007), Banerjee, Duflo and Glennerster (2008), Bertrand, Djankov, Hanna and Mullainathan (2007), Glewwe, Kremer, and Moulin (2009), Andrabi et. al. (2020), Andrabi et. al. (2015), Dhaliwal and Hanna (2017), Muralidharan and Sundararaman (2015), Muralidharan and Singh (2020), Kerwin and Thornton 2020). However, none of these excellent papers provide "rigorous" evidence of anything beyond what happened when they did exactly what they did, where they did it, when they did it, and how they did it.

Approaches like Problem Driven Iterative Adaptation (PDIA) (Andrews, Pritchett, Woolcock 2017, Andrews and Samji 2020) that build capability of organizations to solve problems *by* the *practice* of solving problems do not so much "reject" methods that generate academic papers with rigorous estimates of intervention impact as encompass them into practical pathways to doing things better.

References

Altonji, Joseph; Todd Elder and Christopher Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, 113(1), 151-84.

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, Jishnu Das, Asim Ijaz Khwaja, Tara Vishwanath, Tristan Zajonc, The LEAPS Team 2007. *Learning and Educational Achievements in Punjab Schools (Leaps): Insights to Inform the Education Policy Debate*. Washington DC: World Bank.

Andrews, Matt and Salimah Samji. 2020. "How to Implement Policies with Impact: A Policy-Maker's Toolkit." *Dubai Policy Review*, (February).

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

Angrist, Noam and Rachael Meager. 2022. "The Role of Implementation in Generalizability: A Synthesis of Evidence on Targeted Educational Instruction and a New Randomized Trial." *CEDIL Syntheses Working Paper*, 4.

Banerjee, Abhijit; Esther Duflo and Rachel Glennerster. 2008. "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System." *Journal of European Economic Association*, 6(2-3), 487-500.

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

Bertrand, Marianne; Simeon Djankov; Rema Hanna and Sendhil Mullainathan. 2007. "Obtaining a Driver's License in India: An Experimental Approach to Studying Corruption." *The Quarterly Journal of Economics*, 122(4), 1639-76.

Clemens, Michael A. and Mariapia Mendola. 2020. "Migration from Developing Countries: Selection, Income Elasticity, and Simpson's Paradox." *IZA Discussion Papers, Institute of Labor Economics (IZA).* 13612.

Clemens, Michael A.; Claudio E. Montenegro and Lant Pritchett. 2019. "The Place Premium: Bounding the Price Equivalent of Migration Barriers." *The Review of Economics and Statistics*, 101(2), 201-13.

Dhaliwal, Iqbal and Rema Hanna. 2017. "The Devil Is in the Details: The Successes and Limitations of Bureaucratic Reform in India." *Journal of Development Economics*, 124, 1-21.

Gadamer, Hans-Georg. 1975. *Truth and Method*. New York: Continuum.

Glewwe, Paul; Michael Kremer and Sylvie Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics*, 1(1), 112-35.

Leamer, Edward E. 2010. "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives*, 24(2), 31-46.

McKenzie, David; Steven Stillman and John Gibson. 2010. "How Important Is Selection? Experimental Vs. Non-Experimental Measures of the Income Gains from Migration." *Journal of the European Economic Association*, 8(4), 913-45.

Muralidharan, Karthik and Abhijeet Singh. 2020. "Improving Public Sector Management at Scale? Experimental Evidence on School Governance in India." *RISE Working Paper*, 20/056.

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.

Olken, Benjamin A. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy*, 115(2), 200-49.

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business and Economic Statistics*, 37(2).

Pritchett, Lant. 2021. "Let's Take the Con out of Randomized Control Trials in Development: The Puzzles and Paradoxes of External Validity, Empirically Illustrated." *CID Faculty Working Paper Series*, 399.

____. 2020. "Why "Feigned Ignorance" Is Not Good Economics (or Science Generally)," *LantRant Blog.*

Pritchett, Lant and Farah Hani. 2020. "The Economics of International Wage Differentials and Migration," *Oxford Research Encyclopedias* Oxford UK: Oxford University 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.

Scott, James. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven CT: Yale University Press.

Vivalt, Eva. 2020. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association*, 18(6), 3045-89.

World, Bank. 2020. "Cost-Effective Approaches to Improve Global Learning : What Does Recent Evidence Tell Us Are "Smart Buys" for Improving Learning in Low and Middle Income Countries (English). Washington, D.C. : World Bank Group. ."