Learning from Experiments when Context Matters

By Lant Pritchett and Justin Sandefur*

Suppose a policymaker in context j is contemplating expanding some social program in hopes of raising welfare as measured by Y. She confronts two contradictory pieces of evidence. The first is analysis of a representative sample of eligible individuals, i, some of whom self-selected into the program. Researchers have used this observational data to estimate the "effect" of program participation, T, using the following regression:

$$Y_{ij} = \mu_j + \beta_j T_{ij} + \eta_{ij}. \tag{1}$$

Clearly, the OLS estimates of β_j suffer potential bias. Second, the policymaker reviews results from multiple experimental studies of similar programs in other contexts (β_k for $k \neq j$) that use random assignment to overcome the risk of selection bias, some of whose findings contradict the observational data analysis.

The premise of this short paper is that this is a more or less accurate description of many real-world policy decisions, in which policymakers must weigh observational data from the right context against experimental evidence from a different program and context.

Building on related work in Pritchett and Sandefur (2014), we explore this trade-off between internal and external validity in development economics empirically, drawing on recent experimental work in development economics – i.e., the impact of microcredit, as well as education and health interventions, and the impact of migration – where we can compare (a) the discrepancies between experimental and non-experimental estimates within the same study, to (b) variation in experimental estimates between studies.

Root mean squared error (RMSE) provides a measure of reliability that encompasses these threats to both internal and external validity. For non-experimental estimates of the program in the relevant context, we calculate the RMSE by comparing it to the experimental estimate from within the same study:

$$\text{RMSE}(\hat{\beta}_j^o) = \sqrt{\text{Var}(\hat{\beta}_j^o) + (\hat{\beta}_j^o - \hat{\beta}_j^e)^2} \tag{2}$$

Superscripts o and e denote parameters estimated with observational and experimental variation, respectively. The first term of the RMSE reflects sampling error in the observational estimate, while the second term reflects selection bias. The bias estimate takes the experimental estimate in context j as the truth.

To compute the RMSE for the experimental estimate we contemplate the errors a policymaker in context j would incur if she relied on experimental evidence from other contexts $k \neq j$.

$$\text{RMSE}(\hat{\beta}_{\neq j}^e) = \frac{1}{K} \sum_{k \neq j} \sqrt{\text{Var}(\hat{\beta}_k^e) + (\hat{\beta}_k^e - \hat{\beta}_j^e)^2} \tag{3}$$

Once again, the experimental estimate of the program in context j is our benchmark, and the RMSE consists of both sampling variance and cross-context parameter heterogeneity in treatment effects measured across K other experiments.

The following section illustrates these calculations for the case of microcredit, re-analyzing publicly available data from recently published experimental studies. Based on their mean-square error, we find that non-experimental evidence within context empirically outperforms a single experimental estimate from another context. This advantage disappears for one outcome variable (consumption), but not for another (profit), as more experimental evidence accumulates from diverse contexts. Section

* Pritchett: Harvard Kennedy School and Center for Global Development, lpritchett@cgdev.org. Sandefur: Center for Global Development, jsandefur@cgdev.org.

II briefly illustrates these points with examples from other literatures, and discusses their broader relevance to development policymaking.

I. Case study: microcredit

A recent crop of experimental studies on the impact of microcredit provides a rare opportunity within development economics to compare cleanly identified treatment effects from broadly similar interventions across very disparate contexts. Banerjee, Karlan and Zinman (Forthcoming) summarize the six experiments in a forthcoming volume of the *American Economic Journal: Applied Economics*. While the studies were not coordinated *ex ante*, the authors have conformed *ex post* to parallel reporting formats to facilitate direct comparison of the results. None of the six studies reports treatment effects based on observational data analysis outside of the experiment, but all six study designs make this possible, and all six have transparently released their survey data and programs to enable replication.

We compare impacts on two outcome variables that feature prominently in all or most of the studies, and which are measured fairly consistently: business profits and household consumption.[1] Profits are measured in all six studies,[2] and household consumption in five.[3] To aid comparability, we standardize both dependent variables using the mean and standard deviation of the control group.[4]

[1]All of the regression specifications described below are close variants of results reported in the original six studies. We began by replicating the intention-to-treat (ITT) estimates fo profit (Table 3) and consumption (Table 6) from each study. See the notes on Table 1 for further details.

[2]Where multiple measures are available, we use the most comprehensive measure. Unlike the other studies, profits in the Bosnia and Herzegovina study (Augsburg et al., Forthcoming) refer to the client's main busines only.

[3]The exception here is Ethiopia (Tarozzi, Desai and Johnson, Forthcoming) which finds significant negative impacts on food security, but does not have a monetary measure of household consumption.

[4]Of the six studies, only one (Attanasio et al., Forthcoming) uses a logarithmic transformation of the consumption variable in the original specification. For comparability, we exponentiate this variable before standardizing.

Our non-experimental measure of the treatment is taken from an OLS regression as shown in equation (1), with one significant modification. We control for the random assignment variable, Z_i, so that our estimate of β^o 'naively' compares people who did or did not self-select into microcredit within the treatment group (i.e., within the group invited or encouraged to participate for experimental purposes), and likewise within the control group.[5]

Results of estimating this modified version of equation (1) are modest but varied (Table 1 columns 1 and 3). For business profits, effect sizes range from a positive and statistically significant 0.2 standard deviations in Bosnia and Morocco, to a very small, but negative and statistically significant effect in Mongolia. For consumption, the range of effects is even wider, from positive 0.2 to negative 0.4 standard deviations, though the latter is not significantly different from zero.

As our experimental benchmark, we estimate the effect of treatment on the treated (TOT) by instrumenting T_i in equation (1) with the random assignment variable, Z_i.[6] We focus on the TOT, rather than the effect of the intention-to-treat (ITT) as reported in the original studies, because it is conceptually closest to the treatment effect produced using observational data under the assumption of ignorability of treatment. In both cases, we attempt to measure the effect of treatment for people who self-select into treatment. In addition, estimates of β^{itt} will vary dramatically across studies due to differences in take-up rates, potentially exaggerating heterogeneity in treatment effects for a given beneficiary.

Results of estimating equation (**??**) for each study are reported in columns 2 and 4 of Table 1. Due to modest take-up rates in many studies, confidence intervals on the experimental IV estimates are larger, but the range of effects is broadly similar. For business profits,

[5]For each study, T is defined as a binary indicator of participation in the specific microcredit program under experimental evaluation. In each case, we also include the same set of exogenous covariates employed in the original studies.

[6]For simplicity, our notation ignores the clustered nature of most of the experiments, though we adjust standard errors accordingly where appropriate.

these range from positive and statistically significant in Morocco (0.3 sd) to a small, negative, and insignificant effect in Mongolia (-0.01 sd). For consumption, the only statistically significant effect is negative (-0.2 sd in Bosnia). The highest estimate is huge (0.65 sd in Mongolia) but very imprecisely estimated.

Calculations based on (2) show that on average, observational analysis produces estimates with a RMSE of 0.18 standard deviations for profit and 0.36 for consumption (Table 1, bottom panel). This bias is not trivial, since most of the estimated experimental effects are within 0.1 standard deviation of zero.

However, the RMSE based on () from using an experimental estimate from another context is, in most cases, even greater than the bias afflicting observational data analysis. The reliability of the experimental evidence improves as more experiments accumulate, both because of a decline in sampling error and because meta-analysis averages out the contextual variation in parameters. With only one experiment from another context, the RMSE for profit is 0.33 standard deviations and for consumption, 0.49. Contextual variation is, in this sense, much larger than selection bias. For the profit variable, the observational data analysis continues to outperform experimental estimates even when averaging over five experimental estimates. For consumption, however, once three other experiments are available, policymakers would incur a smaller RMSE by relying on the experimental literature than observational data within context.

II. General discussion

We argue that this inability of internally valid evidence on program impact in one context to improve on predictive accuracy over simple non-experimental estimates in a different context is not specific to this example of microcredit, but is likely a generic (if not universal) feature of development policies and programs.

First, the empirical heterogeneity across contexts in non-experimental estimates of treatment effects for other broad classes of interventions contemplated in development economics is large. For instance, suppose one wanted to estimate the wage gain to movers from a marginal relaxation of barriers to labor mobility in rich countries. Clemens, Montenegro and Pritchett (2008) compare wages of observationally equivalent low-skill workers from 41 countries working in the USA. These non-experimental estimates imply a wage gain of 1.48 log points (287%), encompassing an enormously broad range from 0.7 to 2.7, depending on country of origin. Similarly, consider the Mincerian return to schooling. Montenegro and Patrinos (2014) estimate the wage gain from an additional year of secondary schooling for males for 103 different countries. The average of these estimates is 6.8 percent and the heterogeneity is again massive; the standard deviation is 5.1 and inter-quartile range is 5.0.

Obviously, these non-experimental estimates of the returns to migration or schooling could overstate or understate the true wage gain to the marginal mover or student. For migration, the only experimental evidence of the gains from a program promoting low-skill migration comes from McKenzie, Stillman and Gibson (2010), who exploit a lottery rationing access to New Zealand for Tongan workers. Using observational variation, they estimate the log wage differential is 1.64, while the lottery produces an unbiased estiamte of the wage gain of 1.36 log points – hence $\hat{\beta}^e = 1.36$ and the selection bias which we define as $\omega \equiv \hat{\beta}^o - \hat{\beta}^e$ is equal to $1.64 - 1.36 = 0.28$. Any assertion of the generalizability of β estimates implies the opposite for estimates of ω. This bias parameter is also of interest: policymakers may be genuinely interested in how policy effects the self-selection of workers into migration (or students in schooling, entrepreneurs into microcedit, etc.). This highlights our second point: in the presence of large variability in non-experimental estimates of impact, external validity of treatment effects is improbable, and external validity of all the relevant behavioral parameters is impossible.

Third, we actually dont know what context means. Parameters that are known with engineering precision, like the boiling point of water or tensile strength of steel, are subject to context, in the sense of a denumerable and hope-

Table 1—Treatment effects from microcredit: within- and between-study comparisons

	Profit		Consumption	
	(1) Observational	(2) Experimental	(3) Observational	(4) Experimental
Bosnia, Augsburg et al. (2014) data	0.193**	0.179	-0.0368	-0.233**
	(0.0793)	(0.144)	(0.0690)	(0.119)
	994	994	994	994
Ethiopia, Tarozzi et al (2014) data	-0.0280	0.510		
	(0.0502)	(0.429)		
	12675	12675		
India, Banerjee et al (2014) data	0.0415	0.262	-0.0174	0.0515
	(0.0295)	(0.235)	(0.0367)	(0.301)
	6190	6190	6775	6775
Mexico, Angelucci et al (2014) data	0.0845*	0.000583	0.0113	-0.0376
	(0.0429)	(0.198)	(0.0172)	(0.210)
	16005	16005	16496	16496
Mongolia, Attanasio et al (2014) data	-0.0317***	-0.0105	-0.372	0.654
	(0.00764)	(0.0119)	(0.381)	(0.513)
	608	608	611	611
Morocco, Crepon et al (2014) data	0.194**	0.281*	0.199**	-0.140
	(0.0895)	(0.169)	(0.0932)	(0.142)
	4934	4934	4924	4924
Within-study RMSE based on:				
Non-experimental data	0.18		0.36	
Between-study RMSE based on:				
1 other experiment		0.33		0.49
2 other experiments		0.28		0.40
3 other experiments		0.24		0.34
4 other experiments		0.22		0.32
5 other experiments		0.21		

Note: Each coefficient reports a separate treatment effect from a separate regression, with the dependent variable listed in the first row and the data source listed in the first column. All results are based on re-analysis of micro data from the original studies listed, using publicly available replication files provided by the authors. The TOT effects reported here correspond most closely to the ITT estimates from the following tables and columns in each of the original studies: Bosnia and Herzegovina, Table 3, column 5 for profit and Table 6, column 1 for consumption (Augsburg et al., Forthcoming); Ethiopia, Table 3, column 7, top panel for profit (Tarozzi, Desai and Johnson, Forthcoming); India, Table 3, top panel, column 4 for profit and Table 6, column 1, top panel for consumption (Banerjee et al., Forthcoming); Mexico, Table 3, column 3 for profit, and the sum of the variables in Table 6, columns 3 and 4 for consumption (Angelucci, Karlan and Zinman, Forthcoming); Mongolia, Table 3, column 5 for profit and Table 6, column 1 for consumption (Attanasio et al., Forthcoming); Morocco, Table 3, column 5 and Table 6, column 1 for consumption (Crépon et al., Forthcoming).

fully short list of interacting variables: air pressure, heat. Social programs, in contrast, are embedded in contexts which encompass a long list of unknown factors which interact in often unknown ways. Take the current evidence about the relationship between learning and various education interventions. As noted in Pritchett and Sandefur (2014), experimental evidence shows that lower class sizes produce greater student learning (at least at some ages, for some students, in one state, in some subjects) in the United States (Krueger, 1999). There is also good experimental evidence that class size has next to no impact in Kenya and India (Banerjee et al., 2007; Duflo, Dupas and Kremer, 2012). This lack of external validity could be accounted for by context, but what about the context? The education level of the teachers? The accountability of the education system? The subject matter?

Fourth and finally, we doubt the construct validity (Shadish, Cook and Campbell, 2002) of classes like "microcredit", "pay for performance", "information campaigns", "conditional cash transfers", or "expanding contraceptive access"' to compare program or policy interventions. In the course of implementation, any specific intervention has to make choices within a high-dimensional design space of attributes. A microcredit program for instance, must choose intended beneficiary borrowers, interest rates, repayment frequencies, group or individual liability, how to hire loan officers.[7] Any given program is an instance of a class. What one can infer about a class from an instance depends on dimensionality of the design space and shape of the impact function over the design space.

For instance, there is evidence that NGO versus government implementation is a key element of the design space for social programs in the developing world – as demonstrated by the contrasting performance of parallel experiments NGO- and government-led experiments in Bold et al. (2013) for Kenyan education, the contrast in results between Duflo, Hanna and Ryan (2012), Banerjee, Duflo and Glennerster (2008) and Dhaliwal and Hanna (2014) from attendance monitoring experiments in Indian NGO and civil service contexts, and confirmed by meta-analysis of a broader range of studies in Vivalt (2014). But even smaller, more *ad hominem* factors may also be crucial. Denizer, Kaufmann and Kraay (2013) review the success and failure of World Bank projects and find that the quality of the task manager assigned to the project has as much impact on project success as many country or project characteristics. If a sufficient description of an intervention includes the name of the implementing organization and the project manager, then we are a very far distance from generating evidence from impact evaluations that has external validity.

[7] Banerjee, Karlan and Zinman (Forthcoming) implicitly make this point by avoiding any calculation of an average effect across the six microcredit studies reviewed above, while emphasizing the wide variations in program design across the studies.

III. Conclusion

We analyze the trade-off between internal and external validity faced by a hypothetical policymaker weighing experimental and non-experimental evidence. Empirically, we find that for several prominent questions in development economics, relying on observational data analysis from within context produces treatment effect estimates with lower mean-square error than relying on experimental estimates from another context. Our results suggest that as policymakers draw lessons from experimental impact evaluations, they would do well to focus attention on heterogeneity in program design, context, and impacts, and may learn little from meta-analyses or 'systematic reviews' that focus on average effects for broad classes of interventions.

REFERENCES

Angelucci, Manuela, Dean Karlan, and Jonathan Zinman. Forthcoming. "Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco." American Economic Journal: Applied Economics.

Attanasio, Orazio, Britta Augsburg, Ralph De Haas, Emla Fitzsimons, and Heike Harmgart. Forthcoming. "The Impacts of Microfinance: Evidence from Joint-Liability Lending in Mongolia." American Economic Journal: Applied Economics.

Augsburg, Britta, Ralph De Haas, Heike Harmgart, and Costas Meghir. Forthcoming. "Microfinance at the margin: Experimental evidence from Bosnia and Herzegovina." American Economic Journal: Applied Economics.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman. Forthcoming. "Six randomized evaluations of microcredit: introduction and further steps." American Economic Journal: Applied Economics.

Banerjee, Abhijit V, 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 the European Economic Association, 6(2-3): 487–500.

Banerjee, Abhijit V, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. Forthcoming. "The miracle of microfinance? Evidence from a randomized evaluation." American Economic Journal: Applied Economics.

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

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2013. "Scaling-up what works: experimental evidence on external validity in Kenyan education." Center for Global Development Working Paper.

Clemens, Michael A, Claudio E Montenegro, and Lant Pritchett. 2008. "The place premium: wage differences for identical workers across the US border."

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William Pariente. Forthcoming. "Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco." American Economic Journal: Applied Economics.

Denizer, Cevdet, Daniel Kaufmann, and Aart Kraay. 2013. "Good countries or good projects? Macro and micro correlates of World Bank project performance." Journal of Development Economics, 105: 288–302.

Dhaliwal, Iqbal, and Rema Hanna. 2014. "Deal with the Devil: The Successes and Limitations of Bureaucratic Reform in India." National Bureau of Economic Research.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2012. "School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools." National Bureau of Economic Research.

Duflo, Esther, Rema Hanna, and Stephen P Ryan. 2012. "Incentives work: Getting teachers to come to school." The American Economic Review, 102(4): 1241–1278.

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

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): 913–945.

Montenegro, Claudio E, and Harry A Patrinos. 2014. "Comparable estimates of returns to schooling around the world." World Bank Policy Research Working Paper, , (7020).

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, 4(2): 161–197.

Shadish, William R., Thomas D. Cook, and Donald Thomas Campbell. 2002. Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage Learning.

Tarozzi, Alessandro, Jaikishan Desai, and Kristin Johnson. Forthcoming. "The Impacts of Microcredit: Evidence from Ethiopia." American Economic Journal: Applied Economics.

Vivalt, Eva. 2014. "How Much Can We Generalize from Impact Evaluations?"