American Economic Journal: Applied Economics 2019, 11(1): 57–91
https://doi.org/10.1257/app.20170299

Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments[†]

By Rachael Meager*

Despite evidence from multiple randomized evaluations of microcredit, questions about external validity have impeded consensus on the results. I jointly estimate the average effect and the heterogeneity in effects across seven studies using Bayesian hierarchical models. I find the impact on household business and consumption variables is unlikely to be transformative and may be negligible. I find reasonable external validity: true heterogeneity in effects is moderate, and approximately 60 percent of observed heterogeneity is sampling variation. Households with previous business experience have larger but more heterogeneous effects. Economic features of microcredit interventions predict variation in effects better than studies' evaluation protocols. (*JEL* D14, G21, I38, O12, O16, P34, P36)

Questions surrounding the effectiveness of microcredit as a tool to alleviate poverty have motivated researchers to implement several randomized evaluations of microfinance institutions (Banerjee, Karlan, and Zinman 2015). These studies were designed to test whether microcredit might help poor households by fostering entrepreneurship or potentially harm them by creating credit bubbles (Ahmad 2003; Yunus 2006; and Roodman 2012). Yet consensus on the overall result of these studies has been impeded by concerns about external validity; that is, concerns that the studies may be too different from each other and from future policy settings to permit general conclusions (Pritchett and Sandefur 2015). On this question, the results of multiple studies in heterogeneous contexts offer more than the sum of their parts: they collectively provide the opportunity to estimate not only the average impact but also the heterogeneity in effects across contexts. I perform

* The London School of Economics, STICERD, LSE STICERD, Houghton Street, London, WC2A 2AE, United Kingdom (email: r.meager@lse.ac.uk); I thank Esther Duflo, Abhijit Banerjee, Anna Mikusheva, Rob Townsend, Jeff Harris, Victor Chernozhukov, Andrew Gelman, Jerry Hausman, Oriana Bandiera, Gharad Bryan, Greg Fischer, Aluma Dembo, Xavier Jaravel, Jonathan Huggins, Ryan Giordano, Tamara Broderick, Lars Hansen, Shira Mitchell, Kirill Borusyak, Cory Smith, Arianna Ornaghi, Greg Howard, Nick Hagerty, John Firth, Jack Liebersohn, Peter Hull, Matt Lowe, Donghee Jo, Yaroslav Mukhin, Tetsuya Kaji, Xiao Yu Wang, Aaron Pancost, five anonymous referees, and the participants of NEUDC 2015, The Chicago-MIT student conference 2016, the MIT Economic Development Lunch Seminar, MIT Econometrics Lunch Seminar, and Yale PF/Labor Lunch Seminar for their suggestions, critiques, and advice. I also thank the authors of the seven studies that I use in my analysis, and the journals in which they were published, for making their data and code public.

[†] Go to https://doi.org/10.1257/app.20170299 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

this joint estimation using Bayesian hierarchical models to aggregate the evidence from seven randomized trials of microcredit.

The concerns raised about the external validity of the microcredit literature apply to impact evaluations and social science more broadly. It is common for studies in a literature to vary in their economic and social environments, as well as in the specific implementations of the policy interventions and the evaluation protocols chosen by the researchers. These factors make it unlikely that impact evaluations of social and economic interventions measure exactly the same effect. Yet, understanding the general impact of an intervention is important when deciding whether to implement or subsidize this intervention in settings not yet studied. Evidence aggregated across multiple contexts can provide a reasonable basis for general policy recommendations, but generalizable conclusions may or may not be within reach for any given policy. If the average effect estimated using all the data is in fact composed of substantially heterogeneous effects, predicting the impact of the intervention in a new context can be uncertain or even infeasible. Thus, heterogeneity in observed effects is often interpreted as a measure of the literature's external validity (Vivalt 2016, Allcott 2015, Pritchett and Sandefur 2015).

Aggregating the evidence on microcredit in the presence of concerns about generalizability requires joint estimation of the average effect and the heterogeneity in effects across studies; this motivates the use of the Bayesian hierarchical framework. The core challenge is to use the observed heterogeneity in estimated effects as a signal of the heterogeneity in effects across some broader class of sites, while correcting for the fact that some of the observed heterogeneity is sampling variation (Rubin 1981). The hierarchical framework can separate genuine heterogeneity from sampling variation and simultaneously use this variation to inform the uncertainty on the general treatment effect. However, this correction implies an adjustment on the estimated effects from each study, and thus an adjustment of their corresponding estimated average effect (Gelman et al. 2009). This interdependent uncertainty creates a potentially challenging joint inference problem, particularly with a small number of studies. In this setting, Bayesian methods may offer improved tractability and estimation relative to popular frequentist counterparts such as random effects or Empirical Bayes (Chung et al. 2013, Chung et al. 2015, and Gelman 2017).

My analysis complements previous efforts to aggregate the evidence in the microcredit literature. Review articles such as Banerjee, Karlan, and Zinman (2015) and Banerjee (2013) have assessed the overall literature incorporating expert judgment; I build on their qualitative insights to estimate an average treatment effect and quantitatively assess heterogeneity across studies.[1] Previous formal analyses of

[1] I do not always confirm the results of the review articles. They often employ simple but misleading aggregation techniques such as "vote counting" the number of results in a literature that are statistically significant versus not (see Hedges and Olkin 1980 or Section 9.4.11 of the Cochrane Handbook (Higgins and Green 2011)). The results of such heuristics often differ from formal aggregation techniques; the specific predictions of Banerjee, Karlan, and Zinman (2015), which do not bear out, are on page 12 and 13: "Our eyeballing suggests that pooling across would yield significant increases in business size and profits." and "One robust finding on consumption is a decrease in discretionary spending (temptation goods, recreation/entertainment/celebrations)." Clearly, expert judgment is an important

multiple studies of microcredit and other interventions have separately estimated the average effect and the heterogeneity, rather than jointly addressing these questions (e.g., Allcott 2015; Pritchett and Sandefur 2015). Most of the work on external validity in economics has considered the problem within experiments or quasi-experiments and leveraged partial compliance structures without necessarily estimating a single general effect (Angrist and Fernández-Val 2013, Bertanha and Imbens 2014, and Kowalski 2016).[2] The Bayesian hierarchical framework offers a joint approach to the problems of evidence aggregation and external validity, and is now being adopted into economics (Dehejia 2003; Burke, Hsiang, and Miguel 2014; and Vivalt 2016).

I focus on seven studies that meet the following inclusion criteria: the main intervention studied must be an expansion of access to microcredit either at the community or individual level, the assignment of access must be randomized, and the study must be published before February 2015 (the period of my literature search). The selected studies are Angelucci, Karlan, and Zinman (2015); Attanasio et al. (2015); Augsburg et al. (2015); Banerjee, Karlan, and Zinman (2015); Crépon et al. (2015); Karlan and Zinman (2011); and Tarozzi, Desai, and Johnson (2015). Due to the policies of the two journals that published these papers—the *American Economic Journal: Applied Economics* and *Science*—the full datasets are accessible online. This data makes it possible to fit a variety of models, incorporating information beyond the scope of traditional meta-analysis and checking sensitivity to modeling choices.

I study the impact of access to microcredit on household business profit, expenditures, and revenues in order to evaluate the initial claim of the Grameen Bank that microloans allow poor entrepreneurs to grow their businesses and make more profit (Yunus 2006). Yunus has also suggested that some households might be able to open businesses for the first time, and perhaps even "beggars can turn to business" (Yunus 2006); I pursue a strategy that may detect this effect. Yet households may benefit from microcredit in other ways, such as increased total consumption, shifting to spending on consumer durables, or decreasing spending on "temptation" goods due to greater hope for the future (Banerjee 2013). These outcomes were not collected by all the studies, but I aggregate the available evidence on consumption, consumer durables spending, and temptation goods spending. In all cases, I examine the effect of increased access to microcredit, which is often considered the intention to treat (ITT) effect, due to the likely failure of the stable unit treatment value assumption (e.g., see Banerjee et al. 2015 or Kaboski and Townsend 2011).

I find that the average treatment effects on these outcomes are typically small and uncertain, around 5 percent of the average control group's mean outcome. The sign of the estimated average impact suggests beneficial effects on all outcomes, but there is moderate to high posterior probability of a zero impact due to uncertainty both within and across studies. I find that classical meta-analytic

component of evidence aggregation, but it is not a substitute for statistical analysis. Banerjee, Karlan, and Zinman (2015) fully acknowledged that formal meta-analysis was a complement to their paper, not a substitute.

[2] These methods are not ideal for literatures such as microcredit in which there are potential spillovers across households, and when the expectation of future borrowing opportunities may affect behavior today even in non-borrowers (Banerjee et al. 2015).

techniques, which do not account for variation across studies, often underestimate the uncertainty in the predicted impact in future contexts, and can misleadingly declare "statistical significance" for the impact on some outcomes in the microcredit data. However, the results of both the hierarchical and classical aggregation suggest that the effects on the six household outcomes are uncertain and may be close to zero.

To assess external validity, I estimate the variation in effects across studies for each outcome; the heterogeneity is moderate compared to the average impact on a given outcome. I compute several metrics of heterogeneity in effects, including a metric that scores genuine heterogeneity in effects against the average sampling variation in the studies, to quantify the percentage of variation that is "local" versus "general" (Gelman and Pardoe 2006). I find that, on average, across metrics and outcomes, approximately 60 percent of the initially observed variation in microcredit treatment effects is due to sampling variation. The genuine heterogeneity in effects is thus smaller than previously thought (Pritchett and Sandefur 2015; Vivalt 2016). The Bayesian hierarchical models detect slightly more general information than local information, on average, such that effects in different sites can indeed predict one another, and thus have some predictive content for the future impact in comparable settings. Overall, while these is some variation in treatment effects, these RCTs appear to be reasonably externally valid.

Finally, I explore the role of covariates at both the household and study level to further understand the observed variation in effects across sites. Conditioning on households' previous business experience, I find that microcredit typically has a precise zero impact on household profits for those with no previous business experience. By contrast, the treatment effect on households with business experience is large, on average, yet more uncertain and more heterogeneous across sites. Conditioning on study-level variables, such as the average loan size, interest rate, unit of randomization, and pre-existing levels of microcredit access is challenging with seven studies: I fit a ridge regression to assess their relative predictive power. I find that economic features of microcredit interventions, such as interest rates, are more predictive of cross-site variation in treatment effects than differences in study protocols, such as the randomization unit; this suggests several avenues for further research.

I. Data

The seven studies I consider in this analysis are summarized in Table 1. Many of the studies aimed to test the initial claim of the Grameen Bank that microcredit fosters entrepreneurship among poor households (Roodman 2012). Access to microcredit should allow poor entrepreneurs to avoid reliance on moneylenders and grow their businesses, thus increasing their business expenditures, revenues and ultimately profits (Yunus 2006). All seven microcredit experiments collected information on these three business outcomes. Advocates of microcredit have also emphasized that loans could allow severely credit-constrained households to open businesses they could not otherwise have opened (Yunus 2006). To ensure that new businesses opening is counted as an average increase in profit, households with no

Table 1—Lender and Study Attributes by Country

Country	Bosnia and Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco	The Philippines
Study citation	Augsburg et al. (2015)	Tarozzi, Desai, and Johnson (2015)	Banerjee, Duflo, Glenneister, and Kinnan (2015)	Angelucci, Karlan, and Zinman (2015)	Attanasio et al. (2015)	Crépon et al. (2015)	Karlan and Zinman (2011)
Treatment	Lend to marginally rejected borrowers	Open branches	Open branches	Open branches, promote loans	Open branches, target likely borrowers	Open branches	Lend to marginal applicants
Randomization level	Individual	Community	Community	Community	Community	Community	Individual
Urban or rural?	Both	Rural	Urban	Both	Rural	Rural	Urban
Target women?	No	No	Yes	Yes	Yes	No	No
MFI already operates locally?	Yes	No	No	No	No	No	Yes
Microloan liability type	Individual	Group	Group	Group	Both	Group	Individual
Collateralized?	Yes	Yes	No	No	Yes	No	No
Any other MFIs competing?	Yes	No	Yes	Yes	Yes	No	Yes
Household panel?	Yes	No	No	Partial	Yes	Yes	No
Interest rate (intended on average)	22% APR	12% APR	24% APR	100% APR	24% APR	13.5% APR	63% APR
Sampling frame	Marginal applicants	Random sample	Households with at least 1 woman age 18–55 of stable residence	Women ages 18–60 who own businesses or wish to start them	Women who registered interest in loans and met eligibility criteria	Random sample plus likely borrowers	Marginal applicants
Study duration	14 months	36 months	40 months	16 months	19 months	24 months	36 months

Notes: The construction of the interest rates here is different to the construction of Banerjee et al. (2015); they have taken the maximal interest rate, whereas I have taken the average of the intended range specified by the MFI. In practice, the differences in these constructions are numerically small.

business or missing business data have their revenues, expenditures, and profits coded as zero. This was the decision made by the researchers in many, though not all, of the seven studies. To give my analysis a chance to detect this potential benefit of microcredit, I have employed this strategy throughout.

However, households may benefit from microcredit in other ways, particularly by altering their consumption choices. This could imply that loan access promotes increased consumption in general, but it may also change the composition of household consumption spending. In an environment with limited savings products, credit and savings may function as substitutes, and microcredit may be used to purchase bulky consumer durables items such as vehicles or school tuition. This would imply greater spending on consumer durables in particular, and perhaps decreased spending on "temptation" goods as a result of this substitution. Reduced spending on temptation goods might also arise if access to microcredit increases a household's expectation of escaping poverty in the future (Banerjee 2013). Thus, although data on consumption, consumer durables, and temptation spending were not collected by all the studies, I aggregate the evidence on these outcomes where available.

For all outcomes, I analyze the effect of a randomly assigned increase in access to credit from microfinance institutions (MFIs). This may occur due to branches opening at the community level, perhaps combined with outreach and targeting, or it may occur due to random offers made at the individual level. In the framework of some of the original studies, this increase in access is the treatment assignment rather than the treatment itself: the access effect could be thought of as capturing the "Intention To Treat" (ITT) effect of taking up a loan. However, as pointed out in Banerjee, Duflo, Glennerster, and Kinnan (2015), spillovers and the expectation of future credit access mean that an analysis defining those who take up microcredit as "compliers" may not capture the causal effect of interest. Credit market interventions are often unlikely to satisfy the stable unit treatment value assumption at the household level due to informal financial links between households and general equilibrium effects (Kaboski and Townsend 2011; Banerjee, Duflo, Glennerster, and Kinnan 2015). Hence, the effect measured here may be best understood as the average impact on an individual in a community with increased access to microcredit. Potential concerns that this understates the impact of microcredit on those who eventually take it up are addressed in Section VB and in other work (see Meager 2016).

Where possible, I conform to the decisions made by the original authors regarding the construction and analysis of the outcome variables, but at times this concern was superseded by the need to construct each variable in a uniform way across the studies. Variables for which this proved to be infeasible in practice have been omitted from the analysis. However in some other cases I was able to construct some covariates not analyzed by the original studies using their datasets (see Section V for further discussion).[3] I do not winsorize outliers because most of the studies did not do so, and moreover Augsburg et al. (2015) found that winsorizing outliers sometimes made results statistically significant when they were not significant in the full sample. If the extreme values do not change the point estimate but increase the uncertainty, then winsorising them may lead to analysis that underestimates the true uncertainty about the impact of microcredit.

II. Methodology

A. *Hierarchical Models*

Consider K study sites in which researchers perform similar interventions and measure similar outcomes. Each study, indexed by k, estimates a treatment effect τ_k averaged across individuals in the study. Suppose a researcher is concerned with estimating the average of these treatment effects across these contexts, an object often defined as $\tau = E[\tau_k]$. The studies don't report $\{\tau_k\}_{k=1}^K$: instead, they report estimates $\{\hat{\tau}_k\}_{k=1}^K$. Some of the observed variation in $\{\hat{\tau}_k\}_{k=1}^K$ is sampling variation,

[3] As few of the microcredit studies collected individual-level baseline surveys, and many household covariates are plausibly affected by microfinance, the covariate analysis is limited. Other potential data issues with the original studies, such as attrition or sample selection, were left as they were in the online datasets and not further addressed here.

yet there is likely to be some genuine variation in effects across settings, often defined as $\sigma_\tau^2 = \text{var}(\tau_k)$. This σ_τ^2 influences the uncertainty researchers should have about the value of τ, and also captures a notion of external validity, as it measures the extent to which any τ_k predicts any other $\tau_{k'}$. The core challenge of evidence aggregation is to separate σ_τ^2 from sampling variation, thus characterizing uncertainty on τ and assessing generalizability.

Hierarchical models address this challenge by jointly modeling sampling variation and true heterogeneity across studies, an approach that owes much of its popularity to the work of Rubin (1981) on parallel randomized experiments. Rubin considers a case in which the analyst has access to a set of estimated effects $\{\hat{\tau}_k\}_{k=1}^K$ and estimates of the associated sampling errors $\{\hat{se}_{\tau_k}\}_{k=1}^K$. Rubin specifies a relationship between the observed estimates and the unobserved $\{\tau_k\}_{k=1}^K$, and in addition specifies a relationship between $\{\tau_k\}_{k=1}^K$ and the aggregate parameters of interest (τ, σ_τ^2). The Rubin (1981) model has a hierarchical likelihood in which each site has its own treatment effect parameter, τ_k, but these effects are all drawn from a common distribution governed by (τ, σ_τ^2) as follows:

$$\hat{\tau}_k \sim N\left(\tau_k, \hat{se}_k^2\right) \quad \forall k, \tag{1}$$

$$\tau_k \sim N\left(\tau, \sigma_\tau^2\right) \quad \forall k.$$

The Rubin (1981) model is fully parametric, yet it is more general than it appears: it nests both the individual analyses in the literature and the results of classical meta-analysis. The choice of the Gaussian link between $\{\hat{\tau}_k, \hat{se}_k\}_{k=1}^K$ and $\{\tau_k\}_{k=1}^K$ is motivated by each study's use of unbiased and asymptotically normal estimators. As RCTs in economics typically carry out inference under these assumptions on their estimators, the Gaussian choice here often imposes no more structure than the original studies. The Gaussian distributional link between $\{\tau_k\}_{k=1}^K$ and $\left(\tau, \sigma_\tau^2\right)$ ensures that the model nests the analytic framework of classical meta-analysis, the results of which are recovered by setting $\sigma_\tau^2 = 0$ (Gelman et al. 2009). If σ_τ^2 is set to be infinite, the Rubin (1981) model returns the original estimates. The Gaussian structure is also tractable and offers lower mean squared error relative to other options in many cases (Efron and Morris 1977). Recent work has shown that Gaussian hierarchical models generally deliver reliable inference on the mean τ and variance σ_τ^2 even when the underlying true distribution is not Gaussian (McCulloch and Neuhaus 2011).

The Rubin (1981) likelihood in equation (1) is a particular case of a more general class of models, which can accommodate a variety of input data and distributional structures both parametric and nonparametric. An important generalization for the purposes of the microcredit literature is to consider the full data from each study rather than just the reported estimates. In particular, information on the control group means $\{\mu_k\}_{k=1}^K$ may be useful to the extent that these means

could be correlated with the treatment effects. Incorporating this information, or any other relevant information, has the potential to improve the inference on $\{\tau_k\}_{k=1}^{K}$ and $\left(\tau, \sigma_\tau^2\right)$.

Consider some outcome of interest, such as profits for a household i in study site k, denoted y_{ik}. Denote the binary indicator of treatment status by T_{ik}, and allow y_{ik} to vary randomly around its conditional mean $\mu_k + \tau_k T_i$. The random variation in y_{ik} may be the result of sampling variation or measurement error, as in the Rubin (1981) model, or it may be the result of unmodeled heterogeneity or uncertainty in outcomes for individuals. Allow the variance of the outcome variable y_{ik} to vary across sites, so $\sigma_{y_k}^2$ may differ across k. Specifying a Gaussian likelihood for y_{ik} provides a close analogue to the Ordinary Least Squares (OLS) regressions performed in the original RCTs.[4] The following joint model, incorporating control means and treatment effects and permitting a correlation between them, can aggregate evidence from the full data from all K studies:

$$y_{ik} \sim N\left(\mu_k + \tau_k T_{ik}, \sigma_{yk}^2\right) \quad \forall i, k, \tag{2}$$

$$\begin{pmatrix}\mu_k \\ \tau_k\end{pmatrix} \sim N\left(\begin{pmatrix}\mu \\ \tau\end{pmatrix}, V\right) \quad \text{where} \quad V = \begin{bmatrix}\sigma_\mu^2 & \sigma_{\tau\mu} \\ \sigma_{\tau\mu} & \sigma_\tau^2\end{bmatrix} \quad \forall k.$$

Including additional covariates into the aggregation process could further improve inference on the treatment effects. Some microcredit studies identified households' previous business experience as a pretreatment predictor of heterogeneous impacts from access to loans (Banerjee, Duflo, Glennerster, and Kinnan 2015; Crépon et al. 2015). If treatment effects vary by subgroup then heterogeneity across sites could be partially explained by differing subgroup prevalence in the samples. This would make the treatment effect in future locations easier to predict and potentially allow for subgroup targeting. On the other hand, it may be that treatment effect heterogeneity across studies is located within a particular subgroup of households. With access to the full data, conditioning on household-level variables is possible as long as they are recorded; they do not need to have been reported by the original studies. For example, most microcredit studies did not report results splitting on previous business experience, yet all studies collected this information.

Extending one study's subgroup analysis to all the studies permits an investigation of how general or replicable the detected subgroup effect really is. This can be done within the hierarchical aggregation framework. Consider for generality L household-level covariates, and denote these covariates X_{ik} for household i in site k. To specify a full interactions model—that is, to examine the power set

[4] The kernel of the Gaussian likelihood is the least squares objective function, so MLE on a Gaussian regression mean delivers analytically identical point estimates to OLS regression. The standard errors may be different unless homoskedasticity is assumed in the regression, as robust or clustered errors do not have simple likelihood counterparts. Of course when considering a data-generating model, the Gaussian assumption on the outcome data here is unrealistic. I discuss this and fit alternative models in my follow-up paper, Meager (2016).

of subgroups—creates 2^L intercept terms and 2^L slope terms, henceforth, indexed by p. Here, X_{ik} are all binary, so let $\pi(p) : \{1, 2, \ldots, 2^L\} \rightarrow \{0, 1\}^L$ be the bijection that defines the full set of interactions of these variables. For $p \in \{0, 1\}^L$, denote $X_{ik}^p = \prod_{p=1}^{L} \left[X_{ik}^p\right]^{\mathbf{1}\{I_p=1\}}$. The model below incorporates these effects and remains tractable by enforcing independence across the treatment effects in the 2^L subgroup blocks:

$$y_{ik} \sim N\left(\sum_{p=1}^{2^L} \left[\mu_k^p + \tau_k^p T_{ik}\right] X_{ik}^{\pi(p)}, \sigma_{yk}^2\right) \quad \forall i, k, \tag{3}$$

$$\begin{pmatrix} \mu_k^p \\ \tau_k^p \end{pmatrix} \sim N\left(\begin{pmatrix} \mu^p \\ \tau^p \end{pmatrix}, V_p\right) \quad \text{where} \quad V_p = \begin{bmatrix} \sigma_{\mu^p}^2 & \sigma_{\tau^p \mu^p} \\ \sigma_{\tau^p \mu^p} & \sigma_{\tau^p}^2 \end{bmatrix} \quad \forall p, k.$$

A different model must be built to incorporate site-level covariates which capture differences in the economic environments, features of the interventions, or study protocols.[5] For example, in the microcredit literature the loans offered were of different sizes with different average introductory interest rates, in environments with different pre-existing levels of financial inclusion or microcredit market saturation, and the studies were randomized in different ways. A similar set of differences exists between studies in most literatures in applied economics. To build a model that considers the role of M such covariates in predicting the variation in the treatment effects, denote the set of relevant site-level covariates W_k. Define β_τ to be the M-dimensional vector of coefficients that capture how W_k predicts variation in $\{\tau_k\}_{k=1}^K$, and β_μ to be the analogous coefficients for $\{\mu_k\}_{k=1}^K$. The following Bayesian hierarchical model estimates these parameters:

$$y_{ik} \sim N\left(\mu_k + \tau_k T_{ik}, \sigma_{yk}^2\right) \quad \forall i, k, \tag{4}$$

$$\begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu + W_k \beta_\mu \\ \tau + W_k \beta_\tau \end{pmatrix}, V\right) \quad \text{where} \quad V = \begin{bmatrix} \sigma_\mu^2 & \sigma_{\tau\mu} \\ \sigma_{\tau\mu} & \sigma_\tau^2 \end{bmatrix} \quad \forall k.$$

Although these models use parametric likelihoods, this structure is less restrictive than it appears. As in the case of the Rubin (1981) model, the full-data models shown above nest many popular aggregation methods, because these approaches often impose restrictive assumptions on σ_τ^2 (Gelman 2006). Pooling all the data

[5] In many cases, however, there will be as many or even more site-level covariates than sites; overfitting and even analytical intractability may follow. This situation arises in the microcredit data and I address the issue in Section VB.

together and running a single OLS regression with site-level fixed effects and a common slope τ, as done in Banerjee, Duflo, Goldberg, Karlan, Osei, Parienté, Shapiro, et al. (2015), is a special case of model II.2.[6] This approach does not estimate heterogeneity in the set $\{\tau_k\}_{k=1}^{K}$, this heterogeneity is not used to inform uncertainty about τ or τ_{K+1}. The same is true for classical meta-analysis that estimates τ by taking an average of $\{\hat{\tau}_k\}_{k=1}^{K}$ weighted by their inverse sampling variances. These aggregation methods offer no channel through which the data can signal that the studied effects are heterogeneous and may not contain information about each other.

Because parametric hierarchical structure permits estimation of σ_τ^2 even with relatively small K, it brings data to the question of heterogeneity in the treatment effects across sites.[7] This corresponds to a common applied definition of external validity as the ability of any τ_k to predict any other $\tau_{k'}$ (Allcott 2015, Pritchett and Sandefur 2015). If the models above estimate σ_τ^2 to be approximately zero, there is no unexplained heterogeneity across sites and thus perfect external validity. However, the models may return an estimate of σ_τ^2 that is so large it signals negligible prediction ability of any treatment effect in the set $\{\tau_k\}_{k=1}^{K}$ for any other effect in any other contexts. It may also be the case that a moderate value of σ_τ^2 is recovered, signaling limited but not zero external validity. The estimation of σ_τ^2 is an advantage of hierarchical models, as this parameter is itself of interest and the resulting flexibility may improve the inference on τ and τ_{K+1}.

The variation captured by σ_τ^2 is also related to the extent of "information pooling" across sites that may occur when the hierarchical model is fit to the data. If the model detects $\sigma_\tau^2 = 0$, and thus perfect external validity, it pools all the data and estimates a single homogeneous effect weighting all the data points equally. In this case, the average τ is a better estimate of τ_k in every site than any site's own $\hat{\tau}_k$, so the hierarchical model "shrinks" the $\{\hat{\tau}_k\}_{k=1}^{K}$ toward each other by using the data from site k' to adjust the estimate the impact in site k and vice versa. But if σ_τ^2 is large, the hierarchical model will not pool information across sites, and thus will not shrink the original point estimates $\{\hat{\tau}_k\}_{k=1}^{K}$ together. The hierarchical models can thus maintain the original partition of the data, which classical meta-analysis cannot do (Gelman et al. 2009, Gelman and Pardoe 2006). In this "no pooling" case, τ is an uninformative object, so the uncertainty intervals on it will be wide. Hierarchical models can also estimate intermediate values of σ_τ^2, and thus implement "partial pooling," shrinking the $\{\hat{\tau}_k\}_{k=1}^{K}$ together to an extent inversely proportional to the size of σ_τ^2 (Rubin 1981).

Hierarchical models do require that the treatment effects be "exchangeable" in order to perform well, which formally means that their joint distribution must be invariant to permutation of the K indices (Diaconis 1977). This means, for example,

[6] Specifically, it would be the same as setting $\sigma_\tau^2 = 0$, $\sigma_{\mu\tau} = 0$ and removing the hierarchy on μ. Results of the two procedures will be identical if the structure on the unexplained residual variance is identical, but may not be identical if standard errors are computed using a standard error correction that does not have a simple likelihood counterpart.

[7] Although the upper level distributions in hierarchical models are non-parametrically identified (Andrews and Kasy 2017), in practice K is often small, and functional form assumptions are required for tractability.

that researchers do not have any knowledge of the relative ordering of the treatment effects before they see the data.[8] Thus, researchers will only be able to assess external validity for the set of sites that are in fact exchangeable, and the predicted effect τ_{K+1} only applies to sites exchangeable with the set of sites already studied. As I aggregate only RCT data, the inferences here may not be generalizable to contexts that could not plausibly have been studied with RCTs. Individual policymakers must use their judgment in determining whether their local context is comparable to those aggregated here.

B. *Bayesian Estimation and Inference*

Although hierarchical models can be estimated using frequentist methods, as in "random effects" meta-analytic models, Bayesian methods have several advantages (Rubin 1981, Gelman 2006, Betancourt and Girolami 2013). Bayesian estimates often have lower mean squared error for the parameters at the upper level of the model $\left(\tau, \sigma_\tau^2\right)$ relative to maximum likelihood (Chung et al. 2013, Chung et al. 2015, Gelman 2017). This can occur because the priors constrain the model to avoid fitting the noise in the sample: the priors "regularize" the estimates (Hastie, Tibshirani, and Friedman 2009; and Gelman 2017). Regularization is the introduction of additional information to constrain an estimation procedure and prevent overfitting, typically reducing the variance of the procedure at the cost of introducing bias (Hastie, Tibshirani, and Friedman 2009). In the microcredit literature, σ_τ^2 must be estimated from seven studies: the main challenge for estimation at this scale is variance, not bias.[9] Priors trade an introduction of bias for a reduction in variance, which tends to be substantial enough in practice as to reduce the mean squared error overall (Chung et al. 2013).[10] Frequentist alternatives produce unbiased yet higher variance estimates of σ_τ^2 in particular, and thus tend to overestimate the magnitudes of τ and $\{\tau_k\}_{k=1}^K$ (Gelman 2017).

In addition, Bayesian methods may better quantify the uncertainty on $\left(\tau, \sigma_\tau^2\right)$ relative to frequentist methods. Joint inference on parameters in hierarchical models can be challenging because of the correlation between the uncertainties at each level of the model: inference on τ and σ_τ^2 depends on inference on all $\{\tau_k\}_{k=1}^K$, and vice versa (Betancourt and Girolami 2013). Markov Chain Monte Carlo simulation methods, particularly Hamiltonian Monte Carlo, can effectively surmount these challenges, but greatly benefit from informative priors that direct the algorithm toward a sensible part of the parameter space. For the microcredit interventions, even before randomized trials are done, researchers can be confident τ is not on the order of US$1 trillion PPP, and that τ is closer to US$100 PPP

[8] If researchers know that a set of M covariates should be correlated with the treatment effects, they can use the model in equation (4) that only requires *conditional* exchangeability. However, problems will arise if $M \geq K$, as in the microcredit data. This issue is discussed and addressed in Section VB.

[9] For those concerned about estimating any parameter from seven data points, consider that the alternative is to use zero data points and assume the parameter takes a single known value with certainty.

[10] Penalized frequentist methods also provide improvement over classical frequentist methods, such as random effects. However, in practice, frequentist penalties need to be tuned, typically via cross-validation. With only seven studies, the cross validation error will be substantial, so Bayesian methods are likely to perform better.

than 100,000 USD PPP. Yet prior-free methods, such as simulated maximum likelihood, do not focus the simulation on these reasonable areas of the parameter space, which makes them slower and less likely to converge in practice. As a result, many frequentist methods, such as Empirical Bayes or random effects, do not perform joint inference at all, and instead condition on the point estimates of $\left(\tau, \sigma_{\tau}^{2}\right)$, resulting in confidence intervals that can be too narrow.

Thus, Bayesian methods allow researchers to fit more complex and realistic models to their data. This can be important when K is small such that even moderately flexible likelihoods may overfit the data in the absence of constraints. This problem arises in the microcredit data when attempting to fit the joint model described by equation (2). Estimating the correlation between the mean value of the outcome in the control groups, the $\{\mu_k\}_{k=1}^{K}$, to the treatment effects $\{\tau_k\}_{k=1}^{K}$ when $K = 7$, is prone to overfitting. Yet assuming this correlation is zero, which is implicitly done in simpler models that ignore the control means or enforce independence, seems too restrictive. A compromise can be reached by fitting the model in equation (2) with a reasonably strong prior on the correlation. Inference may be sensitive to the prior in these cases, and indeed, inference on the joint model (equation 2) is somewhat sensitive to the prior on the upper level covariance matrix, although the final results are similar across specifications (see online Appendix B). I therefore fit the Rubin (1981) model and an "independent" version of the joint model that imposes zero correlation between the mean and the treatment effect, as robustness checks.

Priors are also useful in allowing economic theory and contextual knowledge to enter the inference process formally.[11] Consider estimating the correlation between the average household business profits in the control group $\{\mu_k\}_{k=1}^{K}$, and the treatment effect on business profit $\{\tau_k\}_{k=1}^{K}$. If everyone has similar latent productivity, then diminishing marginal returns would suggest the correlation should be negative, as smaller businesses should be able to produce more and grow faster with the same input. Yet, if individuals' productivity is heterogeneous, high business profits in the control group could signal more latent entrepreneurial talent, so that loans to this talented population should have a larger impact, which suggests the correlation should be positive. Since the net observed correlation is likely composed of these countervailing effects, it is unlikely to take an extreme value of either sign. This information is encoded into the prior as a penalty for a large coefficient in either direction; this choice imposes classical regularization (Hastie, Tibshirani, and Friedman 2009). In cases where the research community disagrees about the implications of economic theory or contextual features, priors will need to be more diffuse to reflect the diversity of beliefs.

In my main specification (equation 2), the covariance matrix of the parent distribution, V, is the parameter that captures this correlation. Covariance matrices

[11] Information about the parameters enters via the likelihood in both cases, but the kind of information that can enter is different. Likelihoods can communicate hard constraints on parameters, such as support constraints or type constraints, but typically cannot communicate fuzzy constraints, such as "it is more likely to be positive than negative," or similar.

encode both information about correlation and the scale of variation, and I seek a strong prior on the former and a weak prior on the latter. Hence, I follow the advice of Gelman and Hill (2007) and decompose V into a correlation matrix Ω and scaling factor θ. A prior of half-Cauchy$(0, 10)$ on θ permits the scaling to vary widely. I use an LKJ-correlation matrix distribution prior with a concentration parameter value of 3, denoted *LKJcorr*(3), which favors correlations close to 0 (an LKJ-corr parameter of 1 produces a uniform prior over the space of correlation matrices). The remaining priors on all other parameters are diffuse, with variance at least five times as large as the variation in the data, to reflect the diversity of beliefs in the research community before these studies were conducted. Thus, I use the following set of priors for the joint model in equation (2):

$$
\begin{aligned}
(5) \qquad \begin{pmatrix} \mu \\ \tau \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1{,}000^2 & 0 \\ 0 & 1{,}000^2 \end{bmatrix} \right), \\
\sigma_{yk} &\sim U[0, 100000] \quad \forall k, \\
V &= \text{diag}(\theta)\, \Omega\, \text{diag}(\theta), \\
\theta &\sim \textit{Cauchy}(0, 10), \\
\Omega &\sim \textit{LKJcorr}(3).
\end{aligned}
$$

Less flexible models, such as the Rubin (1981) model and the independent version of this full data model, are generally fit with weakly informative priors throughout the paper, similar in strength to the priors on (μ, τ) above.

C. *Measuring External Validity*

External validity is often characterized by the extent to which the average impact τ predicts treatment effects across different contexts (Pritchett and Sandefur 2015, Allcott 2015). This can be measured by the heterogeneity in τ_k across studies, which is captured by σ_τ^2 in all the models above. If $\sigma_\tau^2 = 0$, a policymaker can learn as much about the impact of microcredit in Ethiopia from a study in Mexico as from a study in Ethiopia itself. If σ_τ^2 is large, then a policymaker in Ethiopia learns little about the likely impact of microcredit in her country from a study of microcredit in Mexico, no matter how excellent the study. In the extreme case, as σ_τ^2 becomes arbitrarily large, a policymaker learns essentially nothing from studies that were conducted outside her setting. Thus, a larger σ_τ^2 leads to a wider "predictive distribution" of the impact in the next comparable study τ_{K+1}. Because Bayesian inference produces a "posterior" distribution by combining the likelihood and prior, this prediction is a full distribution, often called the "posterior predictive distribution" of τ_{K+1}.

The smaller the posterior estimate $\tilde{\sigma}_\tau^2$, the more data pooling across sites occurs in the estimation, and the more a policymaker should update her beliefs about the

impact of microcredit in her context given the results of the RCTs in other countries. The models will produce correspondingly narrow posterior uncertainty on τ and on the posterior predictive distribution on τ_{K+1}. Therefore, for each outcome of interest I report $\tilde{\sigma}_\tau^2$ as well as the entire posterior distributions of τ and τ_{K+1}. The latter quantity is of particular importance as, conditional on the model, this provides the best guess of the impact of expanding access to microcredit in a new location.

A drawback of using $\tilde{\sigma}_\tau^2$ or the width of the uncertainty interval on τ_{K+1} as a metric of external validity is that it is unclear what exactly constitutes a large or small value of this parameter in any given context. Thus, it is useful to examine the "pooling metrics" associated with the Bayesian hierarchical framework, whose magnitude is easily interpretable. The most prominent metric is the conventional "pooling factor" metric, defined as follows (Gelman and Hill 2007, 477):

$$\omega(\tau_k) = \frac{\hat{se}_k^2}{\tilde{\sigma}_\tau^2 + \hat{se}_k^2}. \tag{6}$$

For each site k, this metric decomposes the potential variation in the estimate in site k into genuine underlying uncertainty $\tilde{\sigma}_\tau^2$ and sampling error $\hat{se}_k^2$. Here, $\omega(\tau_k) > 0.5$ means that $\tilde{\sigma}_\tau^2$ is smaller than the sampling variation, indicating substantial pooling of information and a "small" $\tilde{\sigma}_\tau^2$. In that case, τ_k is a better signal of τ than $\hat{\tau}_k$ is of τ_k, and if policymakers and researchers update their beliefs about the impact of microcredit based on the results of an RCT in a given context, they should also update their beliefs about the impact of microcredit in the general case.

This $\omega(\tau_k)$ conditions on sampling variation in order to score how much researchers can learn about site k' by analyzing the current data from the rest of the sites (captured by σ_τ^2) against what they can learn about site k by analyzing the current data from site k (captured by $\hat{se}_k$). Yet from a frequentist perspective, conditioning on sampling variation may cause some discomfort, and the fact that the sample size can influence this metric may be undesirable. Thus, I also compute two additional metrics as robustness checks. The first such metric is a "brute force" version of the conventional pooling metric, which scores how closely aligned the posterior mean of the treatment effect in site k, denoted $\tilde{\tau}_k$, is to the posterior mean of the general effect $\tilde{\tau}$ versus the separated no-pooling estimate $\hat{\tau}_k$. I define this as follows:

$$\omega_b(\tau_k) \equiv \{\omega : \tilde{\tau}_k = \omega\tilde{\tau} + (1-\omega)\hat{\tau}_k\}. \tag{7}$$

The motivation for $\omega_b(\tau_k)$ is that in the Rubin (1981) model it is identical to the conventional pooling metric, but it is not identical in more complex models that pool across multiple parameters (such as model II.2).[12] As a robustness check I also compute the "generalized pooling factor" defined in Gelman and Pardoe (2006), which takes a different approach using posterior variation in the deviations of each

[12] I manually constrain it to take values between $[0, 1]$ as the rare occasions on which it falls outside this range are due to shrinkage on other parameters rather than due to any feature of the parameters in question.

τ_k from τ.[13] Gelman and Pardoe (2006) suggest interpreting $\lambda_\tau > 0.5$ as indicating a higher degree of general or "population-level" information relative to the degree of site-specific information.

III. The General Impact of Microcredit Expansions on Household Outcomes

The general impact of expanding access to microcredit on key household outcomes is an important parameter for a policymaker deciding whether to recommend microcredit programs, and perhaps even subsidize them, in countries not studied with randomized trials. Aggregating evidence across multiple contexts can provide a reasonable basis for such recommendations.[14] In all the Bayesian hierarchical models from Section II, this general effect is captured by τ, the mean of the parent distribution from which all site-specific treatment effects are drawn. This is the expected value of the treatment effect in all sites that are broadly comparable to the current set of sites. In order to test the claim that microcredit helps households by fostering entrepreneurship, I perform inference on τ for household business expenditures, revenues, and profits. To examine other potential welfare benefits through changes in consumption behavior, I also analyze consumption, consumer durables spending, and temptation goods spending for those sites that recorded them.

To estimate τ for each outcome variable, I fit the model described by equation (2) after standardizing all units to USD PPP over a two week period (indexed to 2010 dollars). Table 2 reports the posterior means, which are the most likely value of the treatment effects, and posterior quantiles, which describe the uncertainty about these parameters, for the full joint model. For comparison, the table also shows the results of a simple OLS full-pooling regression with fixed effects for country (as in Banerjee, Duflo, Goldberg, Karlan, Osei, Parienté, Shapiro, et al. 2015). The graph in Figure 1 shows the posterior distributions of τ for each of the six outcomes, and for comparison, the sampling distribution of the OLS estimator for the full-pooling model's estimate of τ. The independent model specification, which does not exploit the correlation between control means and treatment effects and is fit as a robustness check due to the sensitivity of the joint model, is also shown in the table and in Figure 2. These results are broadly robust

[13] Let $E_{post}[\cdot]$ denote the expectation taken with respect to the full posterior distribution, and define $\epsilon_k = \tau_k - \tau$. Then the generalized pooling factor for τ is defined:

$$\lambda_\tau \equiv 1 - \frac{\frac{1}{K-1}\sum_{k=1}^{K}\left(E_{post}[\epsilon_k] - \overline{E_{post}[\epsilon_k]}\right)^2}{E_{post}\left[\frac{1}{K-1}\sum_{k=1}^{K}(\epsilon_k - \overline{\epsilon}_k)^2\right]}. \tag{8}$$

The denominator is the posterior average variance of the errors, and the numerator is the variance of the posterior average error across sites. If the numerator is relatively large, then there is little pooling in the sense that the variance in the errors is largely determined by variance across the blocks of site-specific errors; if the numerator is relatively small, then there is substantial pooling.

[14] The aggregation exercise conducted in this paper is a necessary input into making decisions about policy itself and about the trade-offs involved in multisite program evaluation, but it is not sufficient for such decision making. This is both because it is left to each policymaker's judgment to determine whether their settings of interest are exchangeable with the settings studied, and because there are unknown risks involved in the decisions made by policymakers, the cost of gathering more information, and other variables. One would need a decision theoretic framework to address this properly; this is left for future research.

Table 2—Average Treatment Effect of Microcredit Intervention (τ)

		Estimate	Posterior distribution quantiles			
Outcome	Model	$\tilde{\tau}$	2.5th	25th	75th	97.5th
Profit	BHM (joint)	6.8	−3.0	1.8	10.4	24.5
	BHM (independent)	7.3	−4.7	1.9	11.2	27.5
	Full pooling	7.3	−1.8	4.1	10.4	16.3
Expenditures	BHM (joint)	6.7	−2.3	2.6	9.7	22.1
	BHM (independent)	8.4	−3.9	3.44	12.0	27.6
	Full pooling	13.0	−2.6	7.7	18.4	28.6
Revenues	BHM (joint)	14.5	−1.4	6.6	19.9	43.5
	BHM (independent)	19.9	−6.2	9.0	28.1	60.1
	Full pooling	22.5	4.6	16.3	28.6	40.4
Consumption	BHM (joint)	3.4	−6.3	0.8	5.9	13.2
	BHM (independent)	3.8	−11.3	0.4	7.1	22.2
	Full pooling	4.6	−1.1	2.6	6.6	10.4
Consumer durables	BHM (joint)	1.8	−3.9	0.7	2.9	8.3
	BHM (independent)	2.1	−11.3	0.5	3.4	16.2
	Full pooling	2.3	−23.9	−6.7	11.3	28.5
Temptation goods	BHM (joint)	−0.8	−3.3	−1.3	−0.2	1.3
	BHM (independent)	−0.8	−3.6	−1.3	−0.2	1.4
	Full pooling	−0.6	−1.1	−0.8	−0.5	−0.2

Notes: All effects are in USD PPP per fortnight. The BHM (joint) refers to the model that estimates effects on both the mean (location) and dispersion of the outcome distribution; in this case the dispersion is measured by the mean absolute deviations. The BHM (independent) does not exploit correlation between the control means and treatment effects, serving as a robustness and sensitivity check. The full pooling model is simply a linear regression of outcome on treatment status with country fixed effects, with clustered standard errors at the country level. The *p*-values and Hochberg-corrected *p*-values for the hypothesis that $\tau = 0$ for each outcome are:

	Profit	Expenditures	Revenues	Consumption	Durables	Temptation
p-values	0.1	0.0	0.1	0.1	0.9	0.0
Adjusted *p*-values	0.2	0.1	0.2	0.2	0.9	0.0

to the omission of any single study, and to alternative prior specifications (see online Appendix B).

The results suggest that the effect of microcredit is likely to be positive but small in magnitude relative to control group average outcomes, and there is a substantial probability of essentially zero impact. For example, the posterior mean $\tilde{\tau}$ for profit is about US$7 PPP per two weeks, while the control group mean is about US$95 PPP per two weeks, and the control group standard deviation is US$160 PPP per two weeks. An average increase of less than 8 percent of the current average profit, and less than 5 percent of the standard deviation, is not likely to be a transformative change for a household. In addition, while there is only a 15 percent approximate chance that the impact is negative, on average, the probability of an impact of US$13 PPP per two weeks is also no more than 15 percent. Although there is little evidence that microcredit generally harms borrowers, as was feared by many of its critics, there is also little evidence of an effect that could transform poor households into prosperous entrepreneurs, as was initially claimed by its advocates.

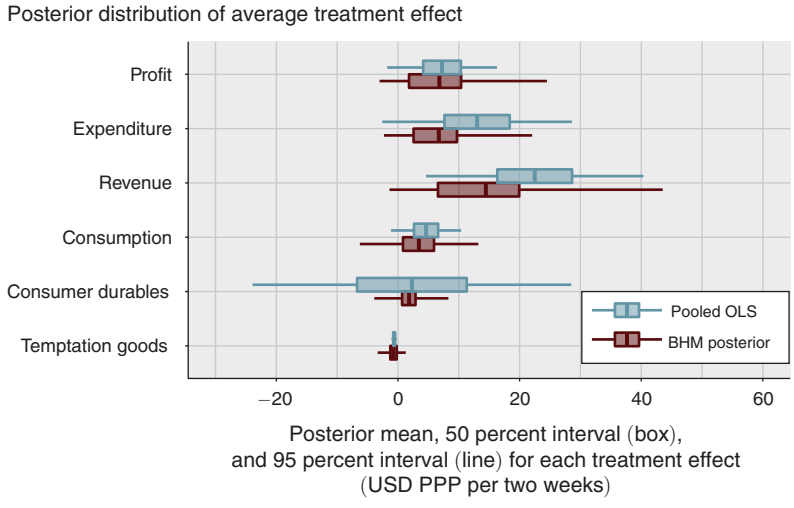

FIGURE 1. GRAPH OF POSTERIORS FOR EACH τ FROM THE MAIN SPECIFICATION OF THE JOINT BAYESIAN HIERARCHICAL MODEL (BHM), WITH THE FULL POOLING OLS INTERVALS FOR COMPARISON

Notes: For the BHM, the thin line covers the central 95 percent posterior interval, the box covers the central 50 percent posterior interval, and the vertical bar within the box marks the posterior mean. For the OLS, the thin line covers the standard 95 percent confidence interval, the box covers a 50 percent confidence interval computed in the same way, and the vertical bar within the box marks the estimate.

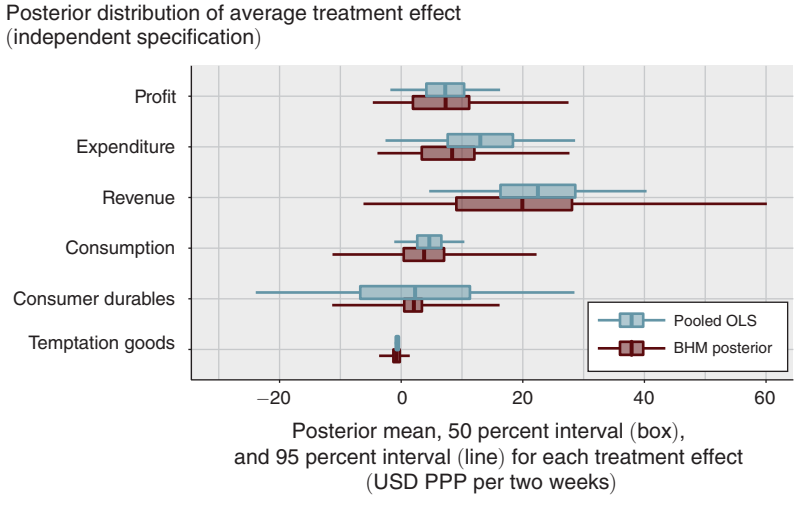

FIGURE 2. GRAPH OF POSTERIORS FOR EACH τ FROM THE INDEPENDENT MODEL, WITH THE FULL POOLING OLS INTERVALS FOR COMPARISON

Note: Interpretation as in Figure 1.

Policymakers are often encouraged to use null hypothesis significance testing or compute *p*-values to make decisions about which interventions have "real" impacts. Within that framework, for revenues and temptation goods, the full pooling model fit with OLS gives a substantially different result to the Bayesian hierarchical models. Testing at the 5 percent level in the full pooling model would declare these two variables "statistically significant," but the hierarchical model finds that their central 95 percent posterior intervals include 0 comfortably. This remains true for temptation goods even when a Hochberg correction for multiple testing is applied (see Table 2). This difference arises because the full pooling model can neither detect heterogeneity nor incorporate this heterogeneity across sites into its estimate of the uncertainty about τ. Bayesian methods—or at least, avoiding a "statistical significance filter"—may give policymakers a more accurate understanding of the impact of microcredit.

To understand why the Bayesian hierarchical model consistently places more probability mass near zero than the full pooling model does for the microcredit data, it is useful to examine the study-specific treatment effects $\{\tau_k\}_{k=1}^{K}$ and their no-pooling estimates $\{\hat{\tau}_k\}_{k=1}^{K}$, as shown in Figure 3.[15] In almost all cases, the more precisely estimated effects are closest to zero. Figure 3 suggests that there is substantial pooling for all outcomes, and the cluster of precise studies near zero pulls the less precise studies dispersed widely around them in toward zero.

These findings differ from the prediction of Banerjee, Karlan, and Zinman (2015) that combining the six 2015 studies and running pooled regressions might find a beneficial and significant impact on profit, business expenditures, and temptation spending. This was a reasonable conjecture: for example, four of the isolated $\hat{\tau}_k$ estimates on profit were positive and reasonably large but statistically not significant, and pooling does tend to increase power and precision. For business revenues and temptation spending, the full-pooling OLS model does indeed exclude 0 in the 95 percent interval, but the hierarchical model overturns this result. My findings also differ from those of Vivalt (2016), which reports a small negative impact of microcredit on profit. However, Vivalt's analysis aggregates a different set of studies, including several observational studies; the quantity being aggregated is potentially different to the treatment effect estimated here.

IV. Heterogeneity in the Impact of Microcredit Expansions

Although the average impact of microcredit is important for policy purposes, this alone does not provide a policymaker with compelling reason to recommend or subsidize microcredit in settings not yet studied. If this average is composed of many heterogeneous effects, it will be at best uncertain and at worst infeasible to predict the impact of microcredit in a new context. Quantifying the heterogeneity of the effects and thus the external validity of this average effect is a concern for both research

[15] Due to the occasionally varying scales of the sampling error, not all intervals have been fully displayed graphically from end to end, but this information can be found in the tables in Appendix A. The independent model results for the same variables are shown in Figure 4 and are similar.

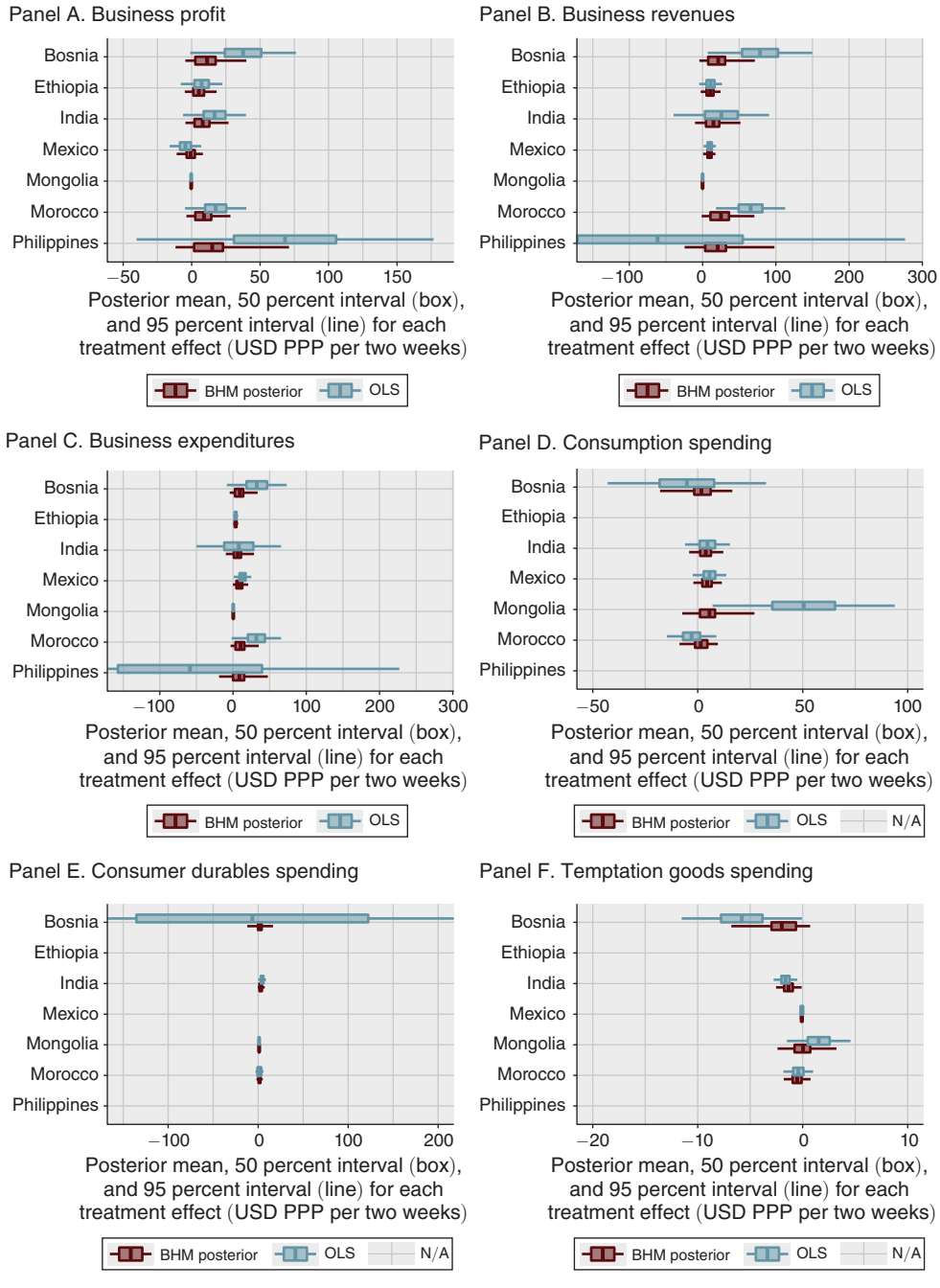

FIGURE 3. GRAPH OF POSTERIORS FOR EACH τ_k FROM THE MAIN SPECIFICATION OF THE JOINT MODEL, WITH THE NO-POOLING OLS INTERVALS FOR COMPARISON

Notes: As the scales of the sampling error differ across sites, some intervals have not been fully shown here. For the BHM, the thin line covers the central 95 percent posterior interval, the box covers the central 50 percent posterior interval, and the vertical bar within the box marks the posterior mean. For the OLS, the thin line covers the standard 95 percent confidence interval, the box covers a 50 percent confidence interval computed in the same way, and the vertical bar within the box marks the estimate. Display is truncated in some cases. The tables in online Appendix A provide the values of these four quantiles and the mean for all marginal posteriors for the main specification model, without truncation.

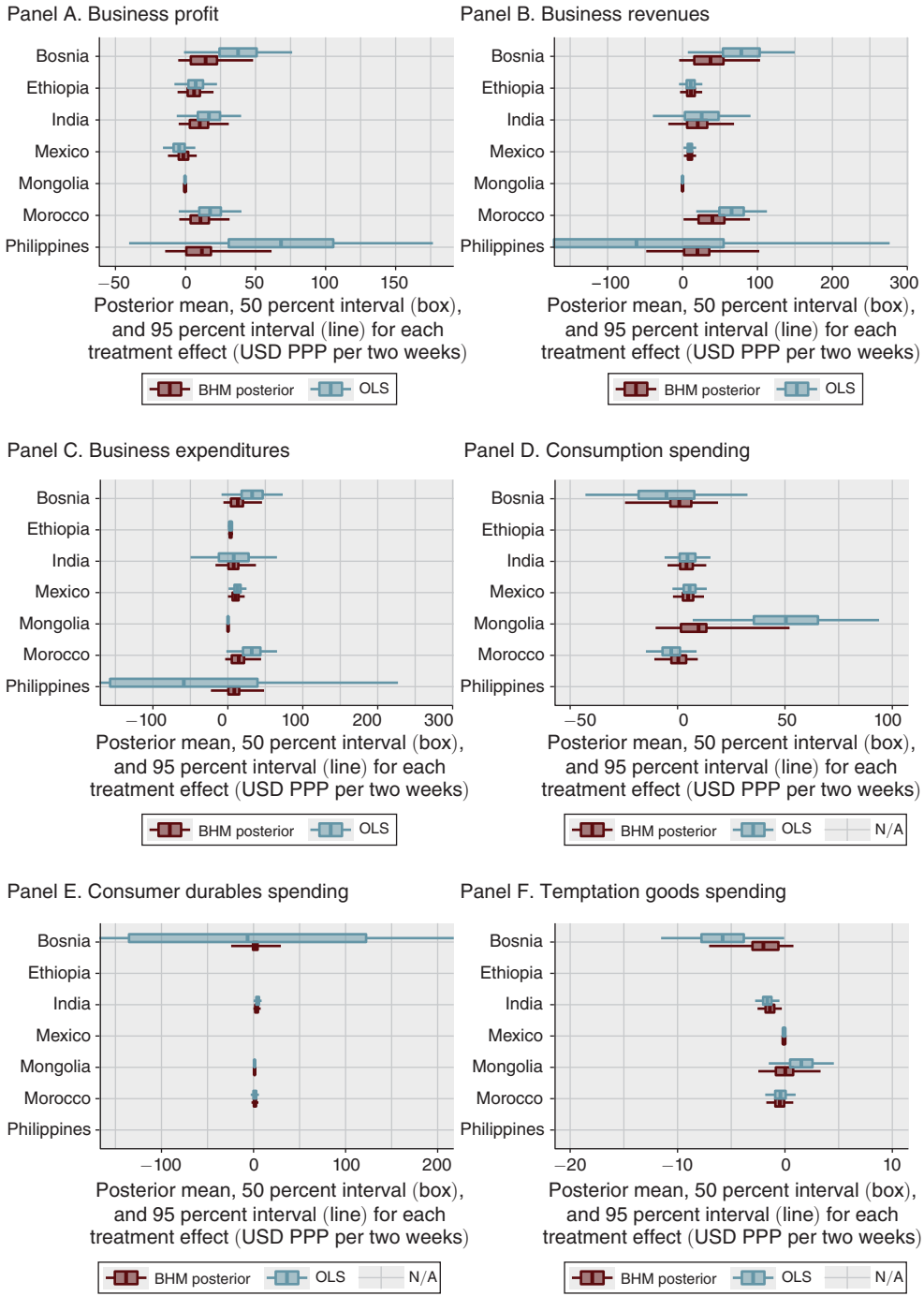

FIGURE 4. GRAPH OF POSTERIORS FOR EACH τ_k FROM THE INDEPENDENT MODEL, WITH THE NO-POOLING OLS INTERVALS FOR COMPARISON

Notes: For the BHM, the thin line covers the central 95 percent posterior interval, the box covers the central 50 percent posterior interval, and the vertical bar within the box marks the posterior mean. For the OLS, the thin line covers the standard 95 percent confidence interval, the box covers a 50 percent confidence interval computed in the same way, and the vertical bar within the box marks the estimate. Display is truncated in some cases.

TABLE 3—POOLING FACTORS FROM THE JOINT MODEL

	Treatment effects			Control group means		
Outcome	$\omega(\tau)$	$\breve{\omega}(\tau)$	$\lambda(\tau)$	$\omega(\mu)$	$\breve{\omega}(\mu)$	$\lambda(\mu)$
Profit	0.5	0.7	0.7	0.0	0.1	0.0
Expenditures	0.5	0.6	0.8	0.0	0.1	0.0
Revenues	0.5	0.5	0.8	0.0	0.1	0.0
Consumption	0.5	0.7	0.9	0.0	0.2	0.0
Consumer durables	0.3	0.5	1.0	0.0	0.0	0.0
Temptation goods	0.2	0.4	0.6	0.0	0.1	0.0

Notes: All pooling factors have support on [0, 1], with 0 indicating no pooling and 1 indicating full pooling. These are simple averages computed across all sites in the data. The $\omega(\cdot)$ refers to the conventional pooling metric that scores signal strength at the general level against average signal strength at the local level. The $\breve{\omega}(\cdot)$ refers to the proximity-based "brute force" pooling metric that measures the geometric proximity of the partial pooling estimate to the no-pooling and full pooling estimates. The $\lambda(\cdot)$ refers to the Gelman and Pardoe (2006) pooling metric that scores the posterior variation at the general level against the average posterior variation at the local level.

and policy. The existing microcredit studies differed in their economic contexts, study protocols, population compositions, and along variety of other dimensions (Table 1). Given these differences, the heterogeneity in the existing studies—and thus their ability to predict each others' treatment effects—provides a signal of the predictive power of this set of studies for a similar yet unstudied context.

To quantify the heterogeneity in the site-specific treatment effects of microcredit expansions, I now report the metrics discussed in Section II. Table 3 displays the conventional pooling metrics, $\omega(\tau)$ for each outcome, which measures the percentage of total variation attributable to sampling variation. I compute the brute force pooling metric $\tilde{\omega}$ and the Gelman and Pardoe (2006) metric λ for a more comprehensive assessment of general versus local information.[16] I also compute the pooling metrics for the control group means $\{\mu_k\}_{k=1}^{K}$, because if the control means are similar, then finding similar treatment effects may only reflect similarities in chosen study locations. If, on the other hand, similar treatment effects accompany dissimilar control group means, then policy recommendations can be more confidently extrapolated to somewhat heterogeneous contexts. I find reasonable similarity among treatment effects, with an average of 60 percent of observed variation attributed to sampling error across all metrics and all outcomes. By contrast, there is no pooling of information on the control group means: whatever similarities are evident in the treatment effects do not reflect preexisting similarities in the study populations. This suggests that microcredit access produces similar, although not identical, treatment effects even in heterogeneous populations.

To get closer to the ideal of predicting treatment effects in a new context, the posterior predictive distributions of τ_{K+1} for all outcomes of interest are shown

[16] As a robustness check, I also compute these metrics for the model that enforces independence between μ and τ, and the results are shown in Table 4.

TABLE 4—POOLING FACTORS FROM THE INDEPENDENT MODEL

	Treatment effects			Control group means		
Outcome	$\omega(\tau)$	$\breve{\omega}(\tau)$	$\lambda(\tau)$	$\omega(\mu)$	$\breve{\omega}(\mu)$	$\lambda(\mu)$
Profit	0.4	0.5	0.7	0.0	0.0	0.0
Expenditures	0.5	0.4	0.8	0.0	0.1	0.0
Revenues	0.4	0.5	0.7	0.0	0.0	0.0
Consumption	0.4	0.6	0.8	0.0	0.2	0.0
Consumer durables	0.3	0.4	1.0	0.0	0.0	0.0
Temptation goods	0.3	0.4	0.7	0.0	0.1	0.0

Notes: All pooling factors have support on [0, 1], with 0 indicating no pooling and 1 indicating full pooling. These are simple averages computed across all sites in the data. The $\omega(\cdot)$ refers to the conventional pooling metric that scores signal strength at the general level against average signal strength at the local level. The $\breve{\omega}(\cdot)$ refers to the proximity-based "brute force" pooling metric that measures the geometric proximity of the partial pooling estimate to the no-pooling and full pooling estimates. The $\lambda(\cdot)$ refers to the Gelman and Pardoe (2006) pooling metric that scores the posterior variation at the general level against the average posterior variation at the local level.

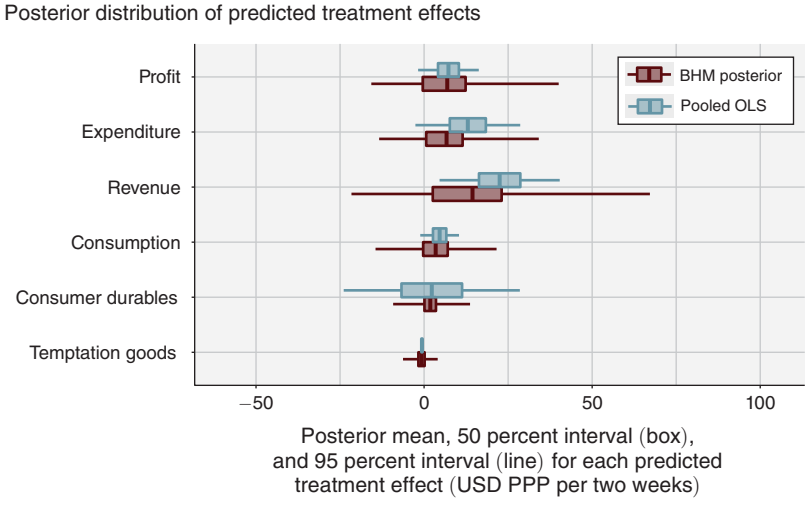

FIGURE 5. POSTERIOR PREDICTIVE DISTRIBUTIONS FOR THE NEXT SITE, τ_{K+1}, COMPARED WITH OLS

Notes: For the BHM, the thin line covers the central 95 percent posterior predictive interval, the box covers the central 50 percent posterior predictive interval, and the vertical bar within the box marks the posterior mean. For the OLS, the thin line covers the standard 95 percent confidence interval, the box covers a 50 percent confidence interval computed in the same way, and the vertical bar within the box marks the estimate.

in Figure 5, with the full-pooling OLS distributions shown for comparison.[17] The predictive intervals are substantially wider than the OLS estimate's intervals, reflecting that there is some heterogeneity in the effects. This heterogeneity increases

[17] The results of the independent model are shown in Figure 6, and are somewhat wider than the joint model because using the observed (typically positive) correlation between the control mean and treatment effect improves fit.

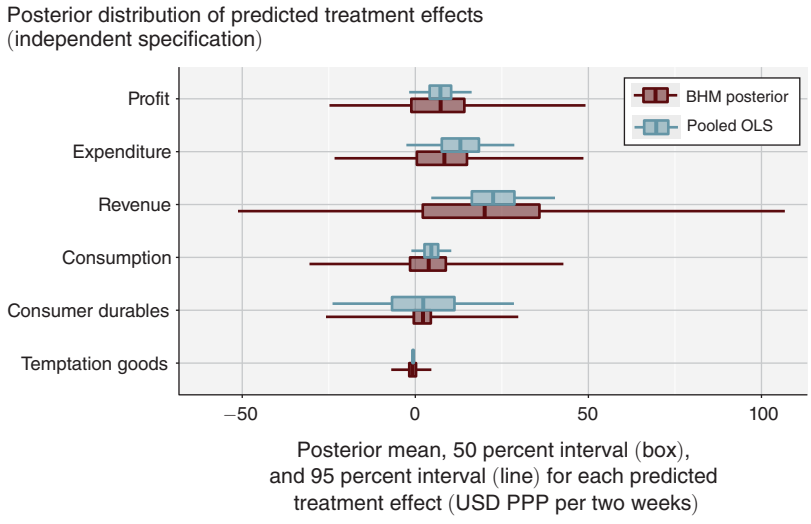

FIGURE 6. POSTERIOR PREDICTIVE DISTRIBUTIONS FOR THE NEXT SITE, τ_{K+1} FROM THE INDEPENDENT MODEL, COMPARED WITH OLS

Note: Interpretation as in Figure 5.

the chances that a new context may not exhibit a positive impact form microcredit: treatment effect for almost all the outcomes has a 25 percent chance of realizing in a socially undesirable direction. For example, the next site's treatment effect on profit has a 50 percent chance of being between US$0 and US$11 PPP, a 25 percent chance of being negative, and a 25 percent chance of being higher than US$11 PPP. The 95 percent prediction interval is almost three times wider than the OLS estimator's 95 percent interval. The full pooling model will tend to underestimate this uncertainty when effects are heterogeneous, as appears to be the case here.[18] Even when there is some similarity in the impact across contexts, the uncertainty generated by a lack of perfect homogeneity across studies remains important for quantifying uncertainty around future impacts.

The finding that the treatment effects of microcredit are reasonably informative for one another differs from the conclusions of Pritchett and Sandefur (2015). This is because they analyzed each study separately before comparing the results: this procedure does not permit pooling of information across studies and thus retains that heterogeneity, which is due to sampling variation. Because they restrict their analysis to the no pooling model, it is not surprising that they find more dispersion in the estimated treatment effects. I also find less heterogeneity in effects than is suggested by Vivalt (2016), perhaps because that analysis includes observational studies. Overall, my results suggest that much of the apparent dispersion in the reported treatment effects is due to sampling variation, and that the genuine underlying heterogeneity is smaller than previously thought.

[18] The one exception here is for consumer durables, where the Bayesian hierarchical results are more precise. This is because the sampling variation is particularly large for the durables variable.

V. Understanding Heterogeneity in Treatment Effects

A. *Household-Level Covariates*

Understanding the genuine heterogeneity in treatment effects across sites remains an important task even when that heterogeneity is smaller than initially thought. Ideally, economists would like to understand how covariates that capture contextual variation between the studies predict the heterogeneity between the observed treatment effects. Household-level covariates may be able to explain the heterogeneity, or they may simply identify the subgroups in which the heterogeneity across sites is located. In the microcredit context, researchers identified several potentially important covariates such as a household's previous business experience, urban versus rural household location, and group versus individual loans (Banerjee, Karlan, and Zinman 2015). Unfortunately, loan type and urban versus rural locations did not vary within site at all for five of the seven studies. By contrast, prior business ownership varied within site for all studies except Karlan and Zinman (2011), so this is the natural variable to examine in detail.[19]

I fit the interactions model from equation (3), incorporating a binary indicator on whether the household already operated a business before any microcredit expansion and enforcing independence in the parent distribution for tractability. Denote this variable PB_{ik}, where $PB_{ik} = 1$ if the household operated a business prior to the microcredit intervention. The results from fitting the fully interacted model with this covariate show that the households where $PB_{ik} = 1$ exhibit muc h more heterogeneity in treatment effects across sites. Figure 7 shows the posterior distributions of the general impacts for the two groups, and Figure 9 shows the posterior distributions of the impacts for each group in each site. While revenues and expenditures seem to rise for both groups—albeit less for the group with new businesses—only the households with prior business experience are likely to be making profits. In fact, for those without prior businesses the treatment effect on profit is almost exactly zero in every site (panel A of Figure 9). Perhaps these new business owners are less productive types, or perhaps it requires learning, experimentation, or time with their business before they can make profit.

The increased heterogeneity across sites in the group with prior business experience is also evident in much wider posterior predictive distributions, as shown in Figure 8. This suggests that when microcredit causes change to occur, its impact is in fact quite heterogeneous across contexts. Overall, these results solidify the conclusion that for households without business experience, microcredit typically has negligible impact: if, as Yunus hoped, beggars can leave the streets because of microcredit, they do not do it often enough or fast enough for it to show up in these seven RCTs.[20] This could be due to the microfinance contract discouraging the kind of risky, illiquid investment often required to begin new businesses.

[19] This analysis is hampered by lack of individual-level baseline surveys, as many household covariates recorded at endline are plausibly affected by microfinance.

[20] It could be that RCT endline surveys are done too close to the intervention time to capture the eventual success of these endeavors, in which case long-term follow-ups should be done to shed light on this question.

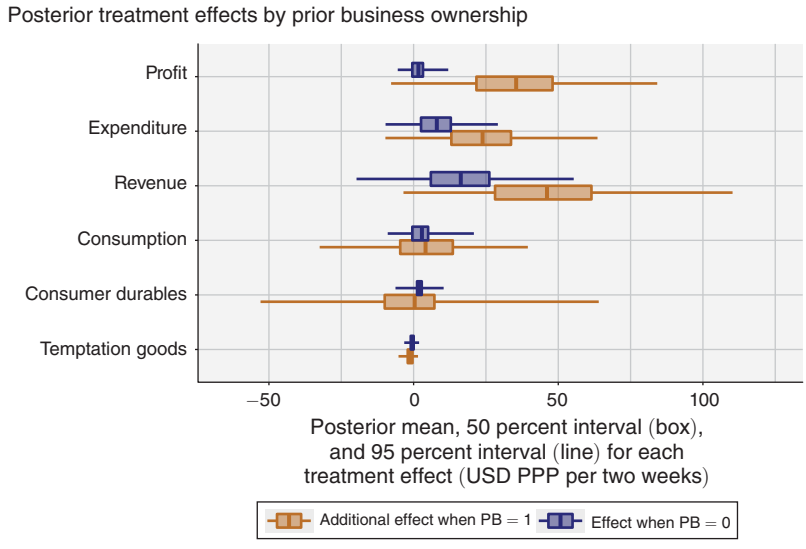

FIGURE 7. POSTERIOR DISTRIBUTIONS OF τ FOR ALL OUTCOMES SPLIT BY PRIOR BUSINESS OWNERSHIP

Notes: For the BHM, the thin line covers the central 95 percent posterior interval, the box covers the central 50 percent posterior interval, and the vertical bar within the box marks the posterior mean. For the OLS, the thin line covers the standard 95 percent confidence interval, the box covers a 50 percent confidence interval computed in the same way, and the vertical bar within the box marks the estimate.

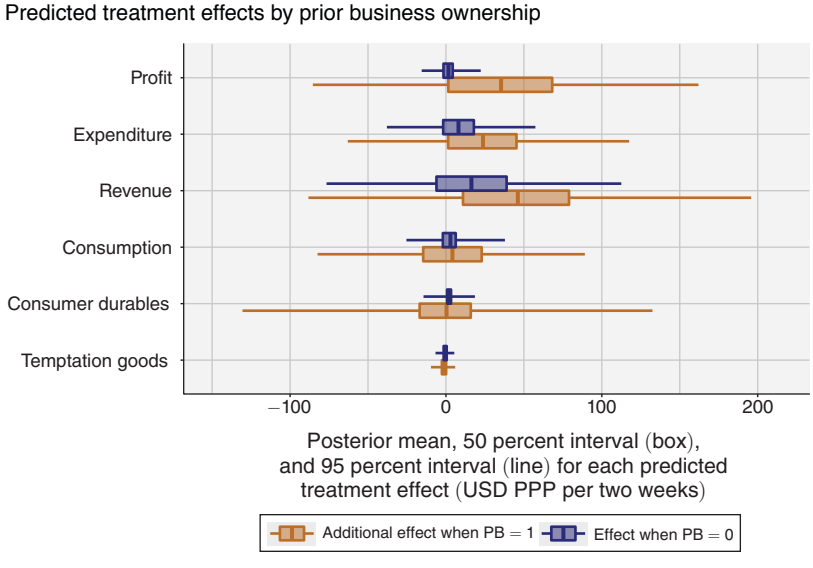

FIGURE 8. POSTERIOR PREDICTIVE DISTRIBUTIONS OF τ_{K+1} SPLIT BY PRIOR BUSINESS OWNERSHIP

Notes: For the BHM, the thin line covers the central 95 percent posterior predictive interval, the box covers the central 50 percent posterior predictive interval, and the vertical bar within the box marks the posterior mean. For the OLS, the thin line covers the standard 95 percent confidence interval, the box covers a 50 percent confidence interval computed in the same way, and the vertical bar within the box marks the estimate.

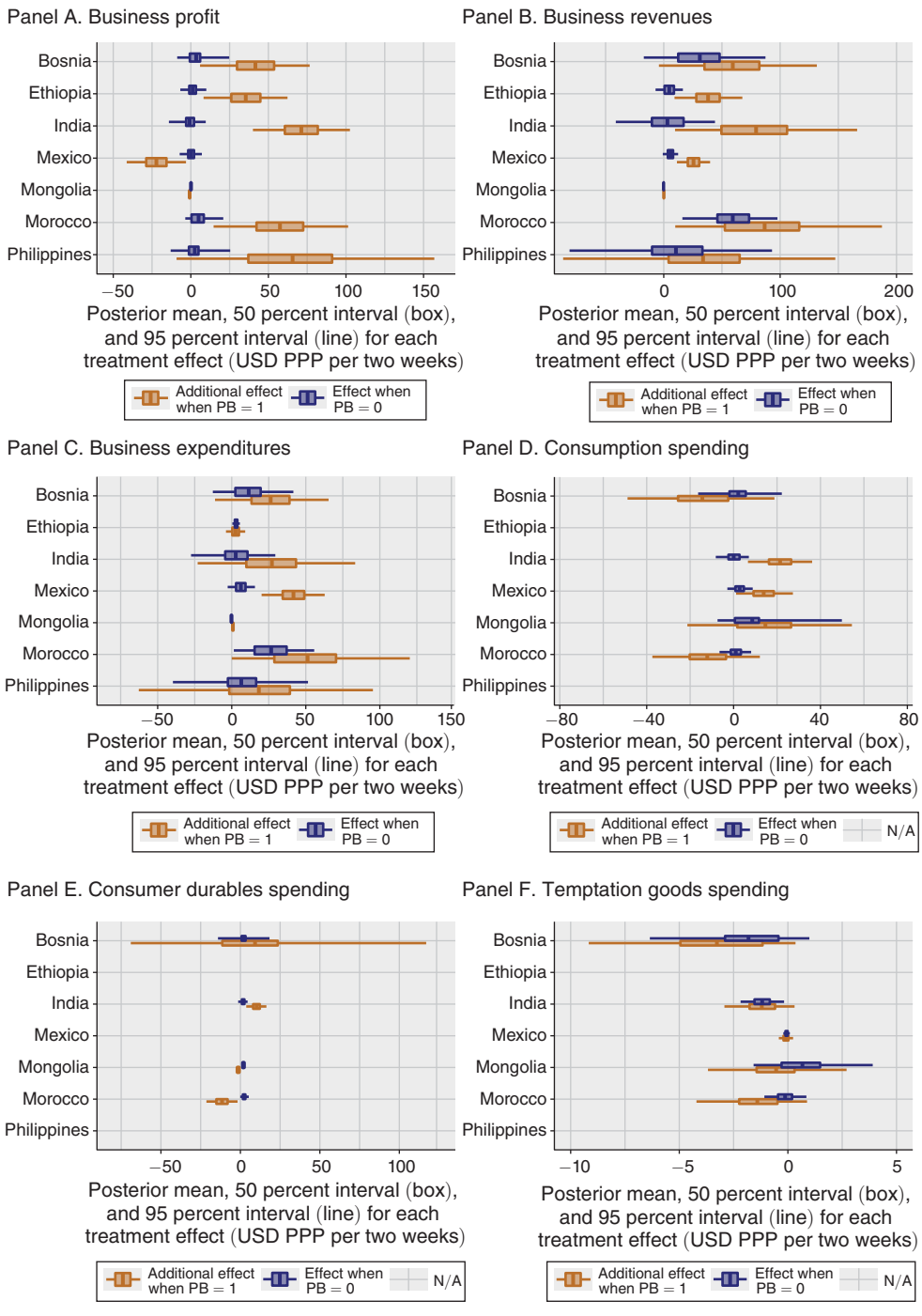

FIGURE 9. POSTERIOR DISTRIBUTIONS OF τ_k FOR ALL SITES AND OUTCOMES SPLIT BY PRIOR BUSINESS OWNERSHIP

Notes: For the BHM, the thin line covers the central 95 percent posterior interval, the box covers the central 50 percent posterior interval, and the vertical bar within the box marks the posterior mean. For the OLS, the thin line covers the standard 95 percent confidence interval, the box covers a 50 percent confidence interval computed in the same way, and the vertical bar within the box marks the estimate. Display is truncated in some cases.

Perhaps households with existing businesses have alternative uses for credit, such as the financing of running costs, that households without businesses do not have. In any case, the results here lend credence to the results of Field et al. (2013) in suggesting that microloans do not reliably stimulate successful new entrepreneurial endeavors.

These results also illustrate how multi-study analysis can help to combat the problems of searching over subgroups for statistically significant effects. When the researchers who ran the RCT in India (Banerjee, Duflo, Glennerster, and Kinnan 2015) checked for this same subgroup, they found a large "statistically significant" effect here. But the Bayesian hierarchical analysis of all the sites shows that this is not always the case. As shown in Figure 9, in some cases this subgroup displays a negligible effect, and in others the effect appears large and negative. While the general treatment effect for the subgroup of households who had a previous business is indeed much higher than for those without, the predictive distribution of this additional effect is diffuse and includes zero comfortably. While there is much more evidence of the potential for large effects in this subgroup, there is also more substantial heterogeneity in the effects across sites in this group.

B. *Site-Level Covariates*

Heterogeneity in treatment effects across studies may be predicted or explained by variables defined at the site level. The MFIs offered different interest rates and loan sizes in each site, and the studies had different protocols and designs which might lead them to estimate different average treatment effects. Five studies randomized at the community level by opening branches in randomly chosen villages or neighborhoods, but two randomized at the individual level by offering loans to randomly chosen members of an existing applicant pool. Group-level randomization may detect spillovers or other General Equilibrium (GE) effects that individually randomized studies cannot. In addition, the applicant pool in the individually randomized studies has signaled strong interest in microcredit, and the take-up is much higher in these studies, leading to concerns that the measured effects may be different. But other economic factors such as large differences in the interest rates offered on the loans, or the loan size, or the MFI's outreach and targeting policies (summarized in Table 1) might also lead to differences in the effects being estimated. Each of these differences in the interventions plausibly creates different impacts at the group and individual level. The randomization unit may evoke more concern, but as this section shows, other variables are stronger predictors of differences in treatment effects.

To understand the influence of all relevant covariates, the ideal procedure would be to condition on these site-level variables and estimate a model that quantifies their role in predicting treatment effects. Unfortunately, the microcredit literature contains only seven experimental studies and at least seven contextual variables of interest. Economists are already analyzing the role of certain contextual variables they deem important, such as credit market saturation, in isolation from the other covariates (see for example Wydick 2015). To estimate these correlations with only seven studies, it is necessary to turn to regression methods that prevent overfitting by

using a penalty function, a procedure referred to as "regularization" in the statistical learning literature (Hastie, Tinshirani, and Friedman 2009). Therefore, I regress the estimated treatment effects from the Bayesian Hierarchical analysis on the relevant contextual variables, regularizing the coefficients to force them to be close to zero unless there is strong evidence of their explanatory power.

Given the limitations of having only seven studies, I perform a Ridge regression with a fixed penalty from which I interpret only the rank ordering of the coefficient magnitudes for each of the covariates.[21] Ridge regression can also be performed within the Bayesian hierarchical models from Section II by modifying the second level of the likelihood to specify the mean as a linear function of covariates. Consider a set of S contextual variables, stored in a vector denoted W_k for site k. Now specify that $\tau_k \sim N(\tau + W_k\beta, \sigma_\tau^2)$ for all sites, and re-estimate the model. The penalty function here consists of a strong prior that each element of the slope vector β is tightly normally distributed around zero. In both the Bayesian and Frequentist Ridge procedures, the variables with the strongest predictive power for the pattern in $\{\tau_k\}_{k=1}^K$ end up with large coefficients despite the penalty (Griffin and Brown 2013). This motivates the interpretation of coefficient magnitude as a ranking of the covariates' relative predictive power, in the absence of sufficient data to assess their absolute predictive power.

I fit a Ridge model with many site-level contextual variables: the site's average value of the outcome in the control group, a binary indicator on whether the unit of study randomization was individuals or communities, a binary indicator on whether the MFI targeted female borrowers, the interest rate (APR) at which the MFI in the study usually lends, a microcredit market saturation metric taking integer values from 0–3, a binary indicator on whether the MFI promoted the loans to the public in the treatment areas, a binary indicator on whether the loans were supposed to be collateralized, and the loan size as a percentage of the country's average income per capita.[22] Table 5 displays the values taken by each of these variables in each site, although they must be standardized to have zero mean and unit variance before regularization.

The results of the Ridge regression at the study level are shown in Figure 10, which displays the absolute magnitude of the coefficients on the various contextual variables for each of the six outcomes. The figure suggests that economic factors, such as whether an MFI targets women, offers a high interest rate or a large loan are more predictive of differences in effects than the randomization unit.[23] All three have negative correlations, suggesting that microloans with lower interest rates, smaller

[21] Statistical learning methods such as Ridge or Lasso procedures are often "tuned" via cross-validation, but this procedure can perform poorly with so few data points. Online Appendix C contains further explanation of this problem, and an illustration of it produced by cross-validation at the study-level in the microcredit literature.

[22] To address concerns that the national average incomes are not capturing the relative income of the actual samples in the studies, I also consider loan size as a percentage of the control group's average income, and the results are shown in the online Appendix C.

[23] This could be because of the selection effect of the randomization level, which potentially raises the average productivity of the whole sample including the control group. In that case, randomization unit will be highly correlated with the control group mean, potentially causing a near-multicollinearity problem and preventing any one of these variables from appearing important. This argument can be made for many variables, and it is still correct to include both the control mean and the unit of randomization in the Ridge regression. But to make the point that the unit of randomization is not a strong predictor, I redo the analysis omitting the control mean as an explanatory

TABLE 5—CONTEXTUAL VARIABLES USED IN RIDGE ANALYSIS (PRE-STANDARDIZATION)

Country	Randomization	Women	APR	Saturation	Promotion	Collateral	Loan size
Bosnia	1	0	22.0	2	0	1	9.0
Ethiopia	0	0	12.0	1	0	0	118.0
India	0	1	24.0	3	0	0	22.0
Mexico	0	1	100.0	2	1	0	6.0
Mongolia	0	1	24.0	1	0	1	36.0
Morocco	0	0	13.5	0	1	0	21.0
Philippines	1	0	63.0	1	0	0	24.1

Note: Contextual variables: Unit of randomization (1 = individual, 0 = community), Women (1 = MFI targets women, 0 = otherwise), APR (annual interest rate), Saturation metric (3 = highly saturated, 0 = no other microlenders operate), Promotion (1 = MFI advertised itself in area, 0 = no advertising), Collateral (1 = MFI required collateral, 0 = no collateral required), Loan size (percentage of mean national income).

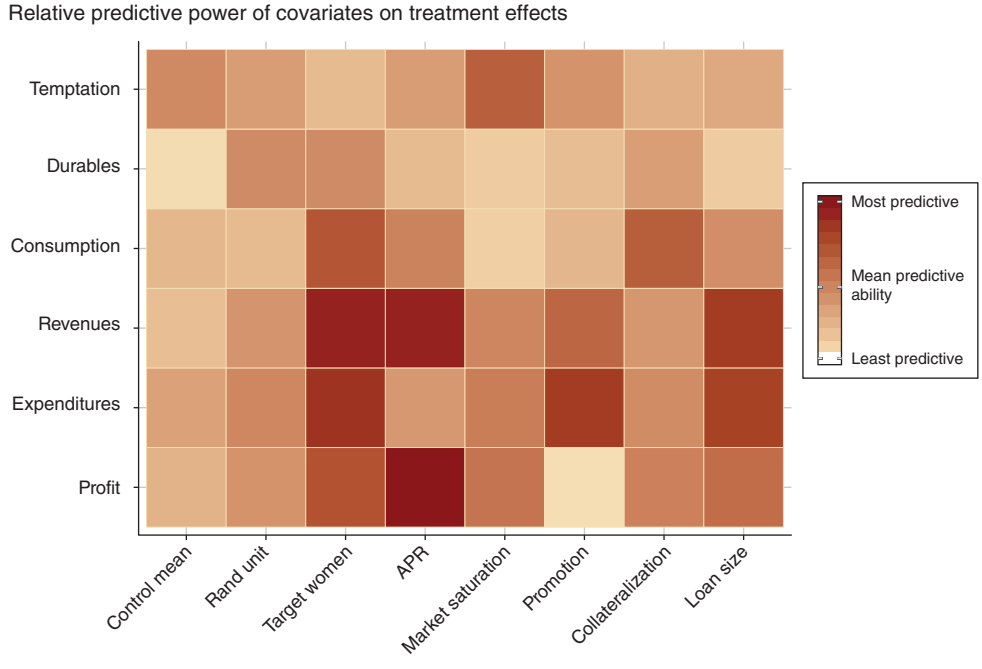

FIGURE 10. ABSOLUTE MAGNITUDE OF THE RIDGE REGRESSION COEFFICIENTS FOR ALL OUTCOMES AND COVARIATES

Note: Results shown for ridge penalty of size 0.1, but the relative ordering of coefficients is largely invariant to penalty size in the regions tested.

loan sizes, and no gender targeting may be associated with better outcomes, though no definitive conclusion is possible on the basis of this exercise. Nonetheless, the

variable, as shown in online Appendix C. I find that the coefficients on economic variables are still larger than that unit of randomization (see Figure B.2 in online Appendix B).

evidence here suggests that the observed heterogeneity in effects across contexts seems more likely to be due to economic differences than to study protocols. This finding is robust to alternative definitions of the variables and alternative specifications (see online Appendix C).

These results may alleviate concerns that the analysis here understates the true impact of microcredit services because of existing market saturation in the studied contexts, or because the take-up of these particular services was low. The studies varied greatly in their pre-existing microcredit market saturation, but this variable is not among the most predictive of the variation in effects; low saturation is not reliably associated with larger effects, as this concern would suggest. The studies also varied greatly in their take-up, which is causally downstream of the unit of randomization; the individually randomized studies both had take-up rates of more than 80 percent, but they do not reliably exhibit larger effects. Instead, microfinance institutions which offer lower interest rates and do not target their services by gender are more reliably associated with somewhat larger treatment effects.

VI. Conclusion

A joint assessment of the average impact and heterogeneity across seven randomized evaluations suggests that microcredit access in general does not transform the lives of poor households in measurable ways, as was initially hoped. Yet there is little evidence that microcredit causes over-indebtedness or destroys livelihoods due to credit bubbles. The moderate heterogeneity in effects across studies may alleviate concerns about the external validity of these RCTs: approximately 60 percent of the initially observed variation in estimated effects is sampling variation. The studies' treatment effects provide reasonable signals of the effects in other study sites, and are thus likely to contain predictive information about a broader class of sites. However, even the moderate heterogeneity detected is enough to generate different results for the Bayesian predictive effect τ_{K+1} relative to the results of classical meta-analysis.

Several potential directions for future work arise from the analysis of covariates. Households with previous business experience often see larger effects from microcredit access, but these effects vary so widely across studies as to prevent general conclusions of positive impact for this group. Investigating the correlation between the treatment effects and study protocols, intervention characteristics and economic contexts, I find that interest rates, loan sizes, and targeting gender are more predictive than the unit of randomization or other evaluation protocols. Thus, further work to assess the causal impact of altering certain features of loan contracts, such as in Field et al. (2013), may be warranted.

Finally, it remains possible that microcredit could affect household or village welfare without affecting average outcomes: perhaps households use these loans to manage risk, or the effects are heterogeneous across quantiles within studies (e.g., Crépon et al. 2015). However, quantile treatment effects have different technical properties to average treatment effects and require substantially different models for aggregation which I address in a separate paper (Meager 2016). Another potential concern with the analysis here is the strict inclusion criterion allowing only RCTs. There are limits to the contexts in which one can randomize treatment; observational

studies of microfinance could provide additional insights. Developing new aggregation methods to accommodate different types of studies may be necessary to improve our understanding of microcredit and many other interventions.

Appendix A: Technical Details of Estimation

Estimating the unknown parameters specified in the hierarchical likelihoods of models such as the one described in equations (1) and (2) is challenging because the likely values of the parameters on the lower level are influenced by the values of the parameters at the upper level, which introduces ripples in the likelihood (Betancourt and Girolami 2013). In theory, either Maximum Likelihood methods or Bayesian methods can be used, but in practice there are strong reasons to prefer Bayesian inference for this problem. The primary issue with Maximum Likelihood is that to get tractability the estimation is done via "Empirical Bayes," which first estimates the upper level parameters and then plugs these point estimates into the lower level to estimate the lower level parameters. By conditioning on a single value of the hyperparameters (τ, σ_τ^2), this procedure systematically underestimates the uncertainty at the lower level of the model. By contrast, Bayesian inference proceeds via estimation of the full joint posterior distribution of all unknown parameters simultaneously, from which the marginal distributions provide accurate uncertainty intervals.

The Bayesian approach does not require the compromises typically made by the MLE method for tractability because it performs estimation using a powerful simulation technique called Markov Chain Monte Carlo methods. These methods require a proper posterior distribution as the target distribution, and this property can be guaranteed by the use of proper prior distributions on the unknown parameters. These priors also allow the researcher to improve the estimation by targeting regions of the parameter space that are more likely to contain relevant values; if only vague knowledge of this is obtainable, then the priors can be made quite diffuse (sometimes called "weakly informative)." If substantial expert knowledge of the likely values is available before seeing the data, this can of course be incorporated via stronger priors. Even if the prior distributions are incorrectly centered, sufficiently diffuse priors can still improve the mean squared error of the estimation by reducing the variance at the cost of some increase in bias—that is, the prior constrains the fit of the model and thus regularizes the estimates (Hastie, Tibshirani, and Friedman 2009).

The posterior distribution for the basic full-data model is proportional to the product of the likelihood in equation (2) and the prior in equation (5):

$$
\begin{aligned}
\text{(A1)}\quad p(\tau, \mu, \tau_1, \tau_2, \ldots | Y) \propto & \prod_{i=1}^{N} \prod_{k=1}^{K} \left(N\left(y_{ik} | \mu_k + \tau_k T_{ik}, \sigma_{yk}^2\right)\right) \\
& \times \prod_{k=1}^{K} \left(N\big((\mu_k, \tau_k) | (\mu, \tau), V\big)\right) \times N\big((\mu, \tau) | (0,0), I_2\big) \\
& \times \mathit{Cauchy}(0, 10) \times \mathit{LKJcorr}(3).
\end{aligned}
$$

This is not a known distribution, but it can be fully characterized via simulation using Markov Chain Monte Carlo methods (MCMC). The basic intuition behind MCMC methods is the construction of a Markov chain, which has the posterior distribution as its invariant distribution, so that in the limit, the draws from the chain are ergodic draws from the posterior. This chain is constructed by drawing from known distributions at each "step" and using a probabilistic accept/reject rule for the draw based on how likely the draw was to have been generated by the posterior. This can be calculated on a draw-by-draw basis without having to evaluate the entire function, and because more likely draws are proportionally more likely to be accepted.

I use a particular subset of MCMC methods called Hamiltonian Monte Carlo (HMC) methods throughout this paper, which are particularly suited to estimating hierarchical models (Betancourt and Girolami 2013). HMC uses discretized Hamiltonian dynamics to sample from the posterior, and has shown good performance especially combined with the No-U-Turn sampling method (NUTS) to auto-tune the step sizes in the chain (Hoffman and Gelman and 2014). HMC with NUTS is easy to implement because it can be done automatically in Stan, which is a free software module that calls C++ to fit Bayesian models from R or Python (Stan Development Team 2014). Stan often requires no more input from the user than typing the equations for the likelihood and priors, although more complex models benefit from code written more efficiently than that. Stan automatically reports the posterior means (e.g., $\tilde{\tau}$ for τ) and their marginalized posterior variances (e.g., $\tilde{se}^2_\tau$), supplying both the parameter values most likely to be true given the data and the degree of certainty we should have about their value. Stan also automatically reports the marginal 95 percent credible intervals and 50 percent credible intervals. Credible intervals are the Bayesian counterpart of confidence intervals, but they admit a direct probability interpretation: the probability that an unknown parameter lies in the α percent credible interval is α percent.

Stan also computes and reports several performance metrics and convergence diagnostics for the HMC in every model it fits. First, it reports the Monte Carlo error of the posterior mean, which should be small relative to the magnitude of the mean if the sampler has converged. Second, it computes the $\hat{R}$ metric of Gelman and Rubin (1992) by randomly perturbing the starting points for the HMC chains and then checking the between variance of the chains relative to the within-chain variance. If all the chains have converged to the posterior, their within variance should be the same as their between variance: the $\hat{R}$ is the ratio of these variances and should be close to 1. For each model, I run 4 chains and accept $\hat{R} < 1.1$.

REFERENCES

Ahmad, M. M. 2003. "Distant Voices: The Views of the Field Workers of NGOs in Bangladesh on Microcredit." *Geographical Journal* 169 (1): 65–74.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130 (3): 1117–65.

Amemiya, Takeshi. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Andrews, Isaiah, and Maximilian Kasy. 2017. "Identification of and Correction for Publication Bias." National Bureau of Economic Research (NBER) Working Paper 23298.

Angelucci, Manuela, Dean Karlan, and Jonathan Zinman. 2015. "Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco." *American Economic Journal: Applied Economics* 7 (1): 151–82.

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66 (2): 249–88.

Angrist, Joshua D., and Iván Fernández-Val. 2013. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." In *Advances in Economics and Econometrics*, Vol. 3, edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 401–34. Cambridge, UK: Cambridge University Press.

Attanasio, Orazio, Britta Augsburg, Ralph De Haas, Emla Fitzsimons, and Heike Harmgart. 2015. "The Impacts of Microfinance: Evidence from Joint-Liability Lending in Mongolia." *American Economic Journal: Applied Economics* 7 (1): 90–122.

Augsburg, Britta, Ralph De Haas, Heike Harmgart, and Costas Meghir. 2015. "The Impacts of Microcredit: Evidence from Bosnia and Herzegovina." *American Economic Journal: Applied Economics* 7 (1): 183–203.

Banerjee, Abhijit Vinayak. 2013. "Microcredit Under the Microscope: What Have We Learned in the Past Two Decades, and What Do We Need to Know?" *Annual Review of Economics* 5 (1): 487–519.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2015. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." *American Economic Journal: Applied Economics* 7 (1): 22–53.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, et al. 2015. "A multifaceted program causes lasting progress for the very poor: Evidence from six countries." *Science* 348 (6236): 1260799.

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.

Bertanha, Marinho, and Guido W. Imbens. 2014. "External Validity in Fuzzy Regression Discontinuity Designs." National Bureau of Economic Research (NBER) Working Paper 20773.

Betancourt, Michael, and Mark Girolami. 2013. "Hamiltonian Monte Carlo for Hierarchical Models." https://arxiv.org/pdf/1312.0906.pdf.

Burke, Marshall, Solomon M. Hsiang, and Edward Miguel. 2014. "Climate and Conflict." National Bureau for Economic Research (NBER) Working Paper 20598.

Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott. 2010. "The horseshoe estimator for sparse signals." *Biometrika* 97 (2): 465–80.

Chung, Yeojin, Andrew Gelman, Sophia Rabe-Hesketh, Jingchen Liu, and Vincent Dorie. 2015. "Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models." *Journal of Educational and Behavioral Statistics* 40 (2): 136–57.

Chung, Yeojin, Sophia Rabe-Hesketh, Vincent Dorie, Andrew Gelman, and Jingchen Liu. 2013. "A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models." *Psychometrika* 78 (4): 685–709.

Crépon, Bruno, Florencia Devoto, Esther Duflo, and William Parienté. 2015. "Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco." *American Economic Journal: Applied Economics* 7 (1): 123–50.

Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii. 2015. "From Local to Global: External Validity in a Fertility Natural Experiment." National Bureau of Economic Research (NBER) Working Paper 21459.

Dehejia, R. 2003. "Was There a Riverside Miracle? A Hierarchical Framework for Evaluation Programs with Grouped Data." *American Statistical Association Journal of Business and Economic Statistics* 21(1): 1–11.

Diaconis, Persi. 1977. "Finite Forms of De Finetti's Theorem on Exchangeability." *Synthese* 36 (2): 271–81

Duvendack, Maren, Richard Palmer-Jones, and Jos Vaessen. 2014. "Meta-analysis of the impact of microcredit on women's control over household decisions: Methodological issues and substantive findings." *Journal of Development Effectiveness* 6 (2): 73–96

Efron, B., and C. Morris. 1975. "Data Analysis Using Stein's Estimator and It's Generalizations." *Journals of the American Statistical Society* 70 (350).

Eysenck, H. J. 1994. "Systematic reviews: Meta-analysis and its problems." *BMJ* 309 (6957): 789–92.

Fahrmeir, Ludwig, Thomas Kneib, and Susanne Konrath. 2010. "Bayesian regularisation in structured additive regression: A unifying perspective on shrinkage, smoothing and predictor selection." *Statistics and Computing* 20 (2): 203–19.

Field, Erica, Rohini Pande, John Papp, and Natalia Rigol. 2013. "Does the Classic Microfinance Model Discourage Entrepreneurship among the Poor? Experimental Evidence from India." *American Economic Review* 103 (6): 2196–2226.

Gelman, A. 2017. "The Future of Null Hypothesis Significance Testing When Studying Incremental Changes and What to Do about It." *Personality and Social Psychology Bulletin* 44 (1): 16–23.

Gelman, Andrew. 2006. "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis* 1 (3): 515–33.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2009. *Bayesian Data Analysis*. 2nd ed. Abingdon: Taylor and Francis.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.

Gelman, Andrew, and Iain Pardoe. 2006. "Bayesian measures of explained variance and pooling in multilevel (hierarchical) models." *Technometrics* 48 (2): 241–51.

Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7 (4): 457–72.

Griffin, Jim E., and Philip J. Brown. 2013. "Some Priors for Sparse Regression Modelling." *Bayesian Analysis* 8 (3): 691–702.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. *Springer Series in Statistics*. Berlin: Springer.

Hedges, Larry V., and Ingram Olkin. 1980. "Vote-Counting Methods in Research Synthesis." *Psychological Bulletin* 88 (2): 359–69.

Higgins, Julian P., and Sally Green. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0. Baltimore: Cochrane Collaboration.

Hoffman, Matthew D., and Andrew Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15 (1): 1593–1623.

Kaboski, Joseph P., and Robert M. Townsend. 2011. "A Structural Evaluation of a Large-Scale Quasi-Experimental Microfinance Initiative." *Econometrica* 79 (5): 1357–1406.

Karlan, Dean, and Jonathan Zinman. 2011. "Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation." *Science* 332 (6035): 1278–84.

Kowalski, Amanda E. 2016. "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments." National Bureau of Economic Research (NBER) Working Paper 22363.

Kroese, Dirk P., Thomas Taimre, and Zdravko I. Botev. 2011. *Handbook of Monte Carlo Methods*. Wiley Series in Probability and Statistics. New York: Wiley.

MacKinnon, James G., and Matthew D. Webb. 2017. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Journal of Applied Econometrics* 32 (2): 233–54.

McCulloch, Charles E., and John M. Neuhaus. 2011. "Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter." *Statistical Science* 26 (3): 388–402.

Meager, Rachael. 2016. "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature." https://economics.mit.edu/files/12292.

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments: Dataset." *American Economic Journal: Applied Economics*. https://doi.org/10.1257/app.20170299.

Park, Trevor, and George Casella. 2008. "The Bayesian Lasso." *Journal of the American Statistical Association* 103 (482): 681–86.

Pritchett, Lant, and Justin Sandefur. 2015. "Learning from Experiments when Context Matters." *American Economic Review* 105 (5): 471–75.

Roodman, David. 2012. *Due Diligence: An Impertinent Inquiry into Microfinance.* Center for Global Development. Baltimore, January.

Rubin, Donald B. 1981. "Estimation in Parallel Randomized Experiments." *Journal of Educational and Behavioral Statistics* 6 (4): 377–401.

Sandefur, Justin. 2015. "The Final Word on Microcredit." *Center for Global Development Commentary and Analysis*, January 22. https://www.cgdev.org/blog/final-word-microcredit.

Stargazer: Well-Formatted Regression and Summary Statistics Tables. 2014. Central European Labour Studies Institute (CELSI). http://CRAN.R-project.org/package=stargazer.

Tarozzi, Alessandro, Jaikishan Desai, and Kristin Johnson. 2015. "The Impacts of Microcredit: Evidence from Ethiopia." *American Economic Journal: Applied Economics* 7 (1): 54–89.

van der Vaart, A. W. 1998. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge, UK: Cambridge University Press

Vivalt, Eva. 2016. "How Much Can We Generalize from Impact Evaluations?" http://evavivalt.com/wp-content/uploads/2014/12/Vivalt_JMP_latest.pdf.

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Berlin: Springer.

Wydick, Bruce. 2015. "Microfinance on the Margin: Why Recent Impact Studies May Understate Average Treatment Effects." https://sites.google.com/a/usfca.edu/wydick/home/research/mfcomment.pdf?attredirects=0.

Yunus, Muhammad. 2006. "Muhammad Yunus—Nobel Lecture." Speech, Nobel Peace Prize, Oslo, December 10, 2006. https://www.nobelprize.org/nobel_prizes/peace/laureates/2006/yunus-lecture-en.html.