Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature

Rachael Meager*†

November 27, 2018

Abstract

Studies of microcredit show positive and negative treatment effects at certain quantiles of household outcome distributions. I develop new Bayesian hierarchical models to aggregate the evidence on these effects and assess their generalizability. I provide a broadly-applicable limited-information model enforcing quantile monotonicity via variable transformation. Partially discrete outcomes such as profit are aggregated using full-data mixture models. Across all outcomes I find a precise, generalizable zero effect from the 5th to 75th quantiles, and large yet heterogeneous and uncertain effects on the right tails. Households who had previously operated businesses account for the majority of the impact and the uncertainty.

*The London School of Economics. Contact: rmeager@lse.ac.uk

†Funding for this research was generously provided by the Berkeley Initiative for Transparency in the Social Sciences (BITSS), a program of the Center for Effective Global Action (CEGA), with support from the Laura and John Arnold Foundation. I am immensely grateful to Ted Miguel and the team at BITSS. I also thank Esther Duflo, Abhijit Banerjee, Anna Mikusheva, Rob Townsend, Victor Chernozhukov, Isaiah Andrews, Tamara Broderick, Ryan Giordano, Jonathan Huggins, Jim Savage, Andrew Gelman, Tetsuya Kaji, Michael Betancourt, Bob Carpenter, Ben Goodrich, Whitney Newey, Jerry Hausman, John Firth, Cory Smith, Arianna Ornaghi, Greg Howard, Nick Hagerty, Jack Liebersohn, Peter Hull, Ernest Liu, Donghee Jo, Matt Lowe, Yaroslav Mukhin, Ben Marx, Reshma Hussam, Yu Shi, Frank Schilbach, David Atkin, Gharad Bryan, Oriana Bandiera, Dean Karlan, Greg Fischer, and the audiences of many seminars for their feedback and advice. I thank the authors of the 7 microcredit studies, and the journals in which they published, for making their data and code public. All my code and data is accessible online at https://bitbucket.org/rmeager/aggregating-distributional-treatment-effects. The paper was pre-registered on the OSF at https://osf.io/tdvc8/ recognising that pre-registration is different for a meta-analysis.

1 Introduction

Financial market expansions in the developing world have the potential to create winners and losers. Increasing access to credit in particular may have heterogeneous effects both because borrowers differ in their investment opportunities and because of general equilibrium dynamics (Banerjee 2013, Kaboski and Townsend 2011). Proponents of financial interventions such as microcredit claim that the positive impact on high-productivity borrowers justifies continued market expansion; detractors claim that the resulting "saturation" of credit markets leads to exploitative lending practices which systematically harm the most vulnerable borrowers (Ahmad 2003, Roodman 2012). The debate continues despite decades of research because the microcredit literature has focused on estimating average treatment effects which cannot evince this heterogeneity. Although recent studies have estimated sets of quantile treatment effects to address this concern, the findings differ across contexts, impeding the formulation of any general consensus and leaving open the possibility of cherrypicking results (Banerjee et al 2015a, Crepon et al 2015, Angelucci et al 2015). Existing meta-studies of microcredit ignored these sets of quantile effects due to a lack of methodology to aggregate them (Meager 2018, Vivalt 2017, Banerjee et al 2015b). In this paper I develop such methodology and aggregate the evidence on the distributional treatment effects of microcredit.

Microcredit institutions reached 132 million low-income clients with a global loan portfolio worth 102 billion dollars in 2016, and the figure is growing yearly (Microfinance Barometer, 2017). At this scale, even small negative impacts for a subset of borrowers would be a concern, and several governments have curtailed microfinance operations ostensibly for this reason (Microfinance Focus 2011, Banerjee 2013, Breza and Kinnan 2018). Yet even in cases where microcredit benefits all households, an unequal distribution of benefits has the potential to affect social and political institutions (Acemoglu and Robinson 2008, Acemoglu et al 2015). Several randomized trials find evidence of negative effects at lower quantiles of household business profits, but others find zero impact on most of the distribution; some found positive effects at the upper quantiles but these were typically imprecise (Augsburg et al 2015, Attanasio et al 2015, Banerjee et al 2015a, Crepon et al 2015, Angelucci et al 2015, Tarozzi et al 2015, Karlan and Zinman 2011). The lack of a general consensus is due to both a lack of power to estimate distributional effects relative to average effects (Leon and Heo 2009) and the possibility of genuine differences in the impact of microcredit across settings.

Aggregating the evidence on these distributional effects across contexts permits

estimation of the typical impact that microcredit is likely to have in future policy contexts. Combining information from multiple studies improves power and prevents cherrypicking, that is, undue focus on the most extreme effects in the literature, which may be the least likely to replicate (Rubin 1981). However, contextual heterogeneity across sites makes it difficult to combine the microcredit studies: pooling the data and performing one analysis is likely to underestimate the true uncertainty (Gelman et al 2004). Here, as in most of social science, the extent of true heterogeneity in the impact of microcredit across settings is unknown. Aggregation raises the question of external validity: the extent to which the recorded impact of a policy in one setting predicts its impact in another, and the extent to which the average of all effects predicts the next effect (Allcott 2015, Bisbee et al 2016). If microcredit has different distributional impacts in different contexts, it may be imprudent to use results from one context to guide policy another context (Pritchett and Sandefur 2015). It is therefore important to use an aggregation method that incorporates uncertainty about the heterogeneity in effects across studies.

Bayesian hierarchical models provide one such framework for evidence aggregation. These models estimate the heterogeneity across studies and use this information to adjust the uncertainty about the typical impact and likely future impact in new settings. By assessing both within-study and across-study variation within a single model, the hierarchical approach is able to separate genuine heterogeneity in effects from the influence of sampling variation (Wald 1947, Rubin 1950, Efron and Morris 1975, Rubin 1981, Gelman et al 2004). This structure nests both the full-pooling case in which the treatment effects are homogeneous, and the no-pooling case in which the effects are so heterogenous as to contain no information about each other. Hierarchical models can also implement a range of intermediate "partial pooling" solutions, borrowing some power across studies to improve inference on all unknown parameters only to the extent suggested to be appropriate by the data (Gelman et al, 2004).[1]

The hierarchical approach is well established in statistics and is increasingly used for evidence aggregation in economics (Chib and Greenberg 1995, Dehejia 2003, Hsiang, Burke & Miguel 2013, Vivalt 2016, Bandiera et al 2017, Meager 2018). The implementation is typically Bayesian due to the potential for improved performance in practice, especially when there are few studies (Rubin 1981, Chung et al 2013). Partial pooling models have also been used to borrow power across regions or sub-

[1] Classical meta-analysis methods typically select the full-pooling solution ex-ante, and modern applied analysis of multiple experiments in economics has tended to compute both full pooling and no pooling models without attempting to access the range of potential solutions in between (Banerjee et al 2015c).

units when studying large geographic areas, or to combine multiple estimates within a study (Hull 2018, Chetty and Hendren 2018, Chetty, Friedman and Rockoff 2014). Yet within the hierarchical framework there are no tools to aggregate distributional effects such as sets of quantile treatment effects. Even outside of the hierarchical framework, the economics literature on external validity and generalizability has focused on different kinds of average effects (Heckman, Tobias, & Vytlacil 2001, Angrist 2004, Angrist & Fernandez-Val 2010, Bertanha & Imbens 2014, Allcott 2015, Dehejia, Pop-Eleches and Samii 2015, Gechter 2015, Athey & Imbens 2016, Andrews and Oster 2018). This necessitates the development of new methods for aggregating evidence on distributional treatment effects in the presence of treatment effect heterogeneity.

Estimating sets of quantile treatment effects provides a general approach to the analysis of treatment effect heterogeneity. With a binary treatment, the quantile treatment effects measure the difference between the unconditional quantiles of the treatment and control groups: this captures *any* difference in the two cumulative distribution functions. Although the set of quantile treatment effects does not estimate the quantiles of the distribution of individual treatment effects, heterogeneity in the effects on different quantiles is evidence of heterogeneity in these individual effects. Other methods such as interacting treatment with covariates in a linear mean or median regression constrains the differences to operate through specific channels. If the researcher does not have access to the right covariates, or does not enter them in the correct way - perhaps themselves transformed or interacted with each other - the heterogeneity will not be detected.[2] If an intervention generally alters household risk exposure then there are no covariates which can predict the resulting heterogeneity in outcomes; sets of quantiles can detect the resulting change in dispersion of outcomes. Sample quantiles also have desirable robustness properties relative to sample means, particularly in the presence of heavy-tailed underlying data (Koenker and Basset 1978).

Aggregating sets of quantiles presents new challenges relative to average effects. Existing Bayesian hierarchical models for average effects such as Rubin (1981) use the knowledge that the within-study sampling variation of both means and quantiles is often asymptotically Gaussian (Mosteller 1946). However, an aggregation

[2]In most cases it is not possible to avoid this problem because the total number of potential combinations of interacted covariates tends to be larger than the data set, necessitating machine learning methods which do not readily provide information on which covariates or which combination best predict the heterogeneity. One way this has been avoided in practice is by researchers selecting the few variables they believe are important ex-ante, but this opens up the possibility of specification searching and cherrypicking results.

model for sets of quantiles must also pass information across the quantiles, because quantiles must be monotonically increasing; failure to incorporate this structure may result in the quantile crossing problem (Koenker 2005).[3] To solve this problem I exploit the fact that Bayesian inference treats unknown objects as random variables: the quantile treatment effects can be constrained to imply monotonic outcome quantiles by transforming the implied unconditional quantiles using functions that only have support on monotonic vectors. In contrast to ex-post rearrangements or smoothing strategies which impose one particular monotonizing procedure (He 1997, Chernozhukov et al 2010), my strategy selects the monotonic vector most likely to be the true parameter from a range of possible solutions. I provide monte carlo simulations showing that the model performs well for inference on the quantile effects in general, as well as showing certain cases where the inference breaks down on the covariation parameters.

A second problem arises in aggregating quantile effects on the microcredit data because business outcomes contain point masses at zero. These are due to households who either do not operate a business or only operate seasonally, and this information must remain in the sample to capture any business creation effects of microcredit. The discrete spikes mean that the asymptotic sampling distribution of the quantiles is no longer Gaussian (Mosteller 1946). To aggregate evidence on quantile effects in this setting, I build richly-parameterised mixture models that capture the economic structure of the variables. The model allows microcredit to affect all aspects of the distribution, and I aggregate by placing a hierarchy on these effects which permits partial pooling across sites. The implied quantile effects can be recovered using the method of Castellaci (2012). This approach automatically satisfies the monotonicity constraints and passes information across quantiles via the functional form assumptions. Model fit assessment and model selection will be necessary to ensure reliable inference; I fit models using both Pareto distributions and Lognormal distributions based on the existing literature on the tail shape of profits and earnings (Piketty 2015, Gabaix 2008, Roy 1950) and find that the Lognormal fits the data better.

Applying these models to seven randomized trials of expanding access to microcredit, I find a precise zero effect on household outcomes from the 5th to 75th percentiles. Above the 75th percentile, there is substantial probability of a large positive impact on most outcomes, but there is greater uncertainty around this effect due to greater heterogeneity within and across studies. Thus, I find some evidence

[3]This problem can occur in the aggregation process even if it is not present in the original studies because weighted averages of monotonic objects need not be monotonic.

of the potential for positive effects and no evidence of systematic harm to any group of borrowers, as there are no generalizable negative quantile effects at any part of the distribution. Part of the greater uncertainty in the tails is due to the tail shape of the business variables, which are so heavy that average treatment effect analysis and Gaussian asymptotics are likely to be unreliable on this data (Koenker and Basset 1978, Mosteller 1946). The likelihood of large, economically important increase in household profits and consumption is far greater than the chance of a zero or negative impact, but there is substantial uncertainty about the specific effect size in any context. By contrast, classical full pooling methods applied to the same data declare "statistically significant" effects in the right tail of business variables as a result of ignoring the heterogeneity across sites.

Further analysis reveals that both the majority of the right tail impact of microcredit and the heterogeneity across studies occurs within the group of households who had previous business experience. This group of households sees increases in consumption above the 75th percentile, although there is still a precise zero effect below that. To better understand the overall pattern of results, in an appendix I pursue a bounding exercise to show that the precise and generalizable impact at zero is unlikely to be due to low take-up of loans. This implies an increase in economic inequality even within the group of households who previously operated businesses, such that the social welfare effects of microcredit are likely to be complex. [4]

2 Data and Context

Microcredit has the potential for heterogeneous effects both because households differentially select into take-up, and because those who do take up have different outcomes depending on how they use the loan or whether they experience shocks (Banerjee 2013). Even certain households who do take up may experience no benefit, perhaps because the terms of the loan may be restrictive and undesirable for investment purposes, or the term to maturity may be too short (Banerjee 2013). But even when borrowers do benefit on average, those who don't take up may end

[4]While the set of studies here may not be a representative sample of all possible microcredit interventions, there is little reason to suspect publication bias in this literature. The papers here published a variety of results most of which were null results in the framework of null-hypothesis significance testing. This leads to less risk of classical publication bias in which only "significant" results appear in the literature. However, it is still possible that these studies are not representative of the world. To address this issue requires substantially more structure and has thus far been ignored in the meta-analysis literature. This may be because such an exercise requires the development of aggregation techniques that can account for differential types of studies in a more complex way than a simple meta-regression, which fails to share information across the study types.

up worse off: there is the potential for winners and losers in general equilibrium due to effects on wages or displacement of informal lending (Kaboski and Townsend 2011, Morduch 1999). Another concern is that multiple microlenders into a community can lead to predatory lending practices and "overlending" to households who cannot feasibly repay the loan (Shicks 2013, Ahmad 2003). This motivated the "No Pago" movement against microcredit in Nicaragua and was ostensibly part of why the government of Andhra Pradesh shut microcredit down during the crisis of 2010 (Microfinance Focus 2011, Banerjee 2013). Even if the groups of households who experience such effects are small, the social welfare consequences could be substantial, particularly if economic inequality across households is affected. Average treatment effects will not reveal these consequences: microcredit access could have zero impact on average, yet large positive or large negative impacts for different types of households.

Motivated by these concerns, several studies of microcredit reported sets of quantile treatment effects for the main outcomes. Some studies did find evidence that microcredit interventions help some households and harm others: many studies found large, positive yet imprecisely estimated impacts on the upper tail, and a few also recorded imprecise negative effects at the lower tail (Angelucci et al. 2015, Augsburg et al. 2015, Banerjee et al. 2015b, Crepon et al. 2015). Yet certain studies recorded noisy positive effects at the lower tail of some outcomes (such as profit in Banerjee et al 2015b) or negative effects at the upper tails (household business income in Angelucci et al 2015). In many of these same cases the quantile treatment effects recorded exact zeroes, estimated with relatively high precision, for the central quantiles. While certain studies, such as Crepon et al 2015, recorded "statistically significant" tail effects on both ends, almost all studies recorded imprecise estimates at the upper tails. In this setting, the gains from aggregating evidence across these studies may be considerable in both precision and an improved understanding of the general pattern of quantile treatment effects.

Estimating sets of quantile treatment effects has many advantages over other approaches to heterogeneous effects in this setting (Banerjee et al 2015b). Microcredit plausibly impacts households' risk exposure - both because they could use credit as consumption insurance, and because they could change their business strategy as a result of the loan - so researchers need to be able to detect changes in the spread of the distribution which by definition won't be predicted by observables. Even when changes in the distribution result from heterogeneous effects rather than changes to risk exposure, microcredit is an intervention for which even households with similar covariates could have considerably different treatment effects due to unobserved het-

erogeneity in ability and access to productive opportunities (Kaboski and Townsend, 2011). The possibility of general equilibrium effects within the villages themselves make it important to assess the impact of treatment on the overall shape of the distribution at the village level, and to characterise heterogeneity based on relative position (ranks or quantiles) rather than absolute values of covariates. In any case, baseline covariates are not available for many of the microcredit studies. Finally, as some of the outcome variables of interest have heavy tails, the robustness properties of quantiles are beneficial in the microcredit setting.

To aggregate the evidence on the quantile treatment effects of microcredit, I use data from seven studies which meet the following inclusion criteria: the main intervention 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 et al. 2015, Attanasio et al. 2015, Augsburg et al. 2015, Banerjee et al. 2015b, Crepon et al. 2015, Karlan and Zinman 2011, and Tarozzi et al. 2015, six of which were published in a special issue of the *American Economics Journal: Applied Economics*.[5] [6] I restrict the sample to randomized controlled trials (RCTs) because they typically have high internal validity for estimating causal effects, and because as yet there is no established methodology designed to aggregate both RCTs and observational evidence in a single framework.[7]

I analyse six outcomes linked to the claim that offering households more credit on more favourable terms should stimulate entrepreneurship (Morduch 1999, Yunus 2006, Roodman 2012). These include: household business expenditures, business revenues, and business profits, household consumption, consumer durables spending and temptation goods spending. Because microfinance insitutions (MFIs) offer lower interest rates relative to informal moneylenders, poor entrepreneurs may be able to start new businesses or grow their existing businesses, increasing their business expenditures, revenues and ultimately profits (Yunus 2006). Greater economic

[5]Other RCTs of microfinance tend to randomly vary certain characteristics of the loans themselves, which allows researchers to understand the impact of these features of the loans but complicates the inference on the general impact of the standard microcredit model (Field et al 2013). Karlan and Zinman 2009 expands access to consumer credit, but microcredit is often considered categorically different to consumer credit; see Banerjee 2013 for a deeper discussion of this.

[6]I focus on expanding access to microcredit because this is the intervention closest to the policy of subsidizing microfinance institutions (MFIs) or promoting interventions under the general umbrella of "microcredit".

[7]Existing meta-analyses and other evidence aggregation exercises have either cherrypicked certain observational studies deemed "good enough", or thrown all types of studies into a single analysis, a strategy which is likely to violate exchangeability assumptions on the treatment effects discussed in section 3.

prosperity should enable households to increase their consumption in the medium and long run. Yet even households without business investment opportunities may use microloans to shift spending away from "temptation goods" (pleasurable but nonproductive or even harmful expenditures) and towards durable goods (Roodman 2012, Banerjee 2013). This situation might arise if access to microcredit increases a household's expectation of escaping poverty in the future, or if microcredit solves a self-control problem (Banerjee and Mullainathan 2010, Banerjee 2013). In general, I analyse the effect of expanding access itself, often called the Intention to Treat Effect in the original studies (Banerjee et al 2015b). The network links between households, the potential for general equilibrium effects, and the impact of the mere expectation of taking up credit in the future even if one does not take it up today means that the Stable Unit Treatment Value Assumption (SUTVA) is likely to be violated within a community that experiences any increase in access (Banerjee 2013, Kinnan and Townsend 2012, Breza 2012).[8]

Despite the restrictive inclusion criteria, the selected studies still differ substantially in their implementations and local contexts (see table 7 in Appendix E). They cover seven different countries, they have different partner NGOs, offering similar but not identical loan contract structures with different interest rates and loan sizes, and they differ in terms of their randomization units - five randomized at the community level and two at the individual level - with various encouragement and sampling designs. Given this heterogeneity across studies, it seems likely that there might be important heterogeneity in the resulting effects. However, the 95% confidence intervals of the quantile effects do overlap across most of the studies, suggesting there may be meaningful similarities across settings.[9] In this context, where the generalizability of the evidence across settings is unclear ex-ante, the Bayesian hierarchical framework is an appropriately cautious way to proceed with evidence aggregation.

The open data policies of the *American Economics Journal: Applied Economics* and *Science* allow me access to the microdata from all of these experiments, such that I can standardize which quantiles I compute across studies and can construct each underlying variable in a uniform manner across studies. The variables were measured in different currencies, in different years, and over different time periods (this matters because these are all flow variables). I standardize all measurements

[8]To investigate the role of take-up in this context, I pursue a bounds analysis explained further in Appendix D.

[9]This pattern was also observed in the average treatment effects in these studies, which turned out to have only moderate underlying heterogeneity despite these contextual differences (Meager 2018). However, similarities in the average treatment effects may be uninformative about the true generalizability of the effects if indeed these averages are composed of heterogeneous quantile effects.

to be USD PPP in 2009 dollars over a two-week period. Business variables require further standardization: to capture the potential for microcredit to allow individuals to open new businesses or to switch to operating any existing seasonal businesses throughout the year, households with no business or missing business data have profits imputed as zero. This was the decision made by the original authors of many of the seven studies, although not all of them. Because the business creation channel is closely tied to the central claims of Yunus (2006), I employ this strategy throughout to business expenditures and revenues as well.[10] Other than standardizing the definition and construction of variables as much as possible, in most other respects I have attempted to conform to the decisions made by the original authors themselves. [11][12]

Household and study-level covariates may play some role in determining heterogeneity in the quantile treatment effects, but there are limitations to pursuing a covariates analysis in this literature. Only three of the microcredit RCTs collected comprehensive individual-level baseline surveys. One pre-treatment variable was recorded at endline in all studies due to its theoretical importance: a binary indicator that a household had previous experience operating a business (Banerjee et al 2015b). Although covariates at the study level may also predict variation in effects across context, there at least seven such covariates and only seven studies, so conventional regression analysis will be overfitted and misleading. It is still useful to aggregate the evidence without conditioning on covariates, as this permits an understanding how much unconditional heterogeneity there is; if there is little or no variation across settings, further analysis is a less pressing concern for future work.

[10] While it would be ideal to examine effects on other variables such as income and assets, the measurement and definition of those variables differed across the studies to such an extent that it is unclear how to aggregate them. This issue was noted in Meager (2018) and in my pre-registration: https://osf.io/tdvc8/ .

[11] Certain potential issues such as attrition or sample selection were left as they were in the studies themselves and in most cases I have used the entire sample available in the online data sets. Ethiopia is the only exception: this study contained a cross-randomized family planning treatment. I use only the pure control and the pure microcredit samples, which is the conservative choice given that we do not know how microcredit interacts with family planning (the study estimates a very imprecise interaction).

[12] I do not winsorize any of the variables because most of the studies did not do so, and 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, winsorising them may underestimate the true uncertainty. As my analysis shows, the behaviour of the upper tails turn out to play an important role in determining the impact of microcredit.

3 Methodology

3.1 Bayesian Hierarchical Models

3.1.1 Hierarchical Models

Consider a body of evidence consisting of K studies indexed by k, each of which provides some k-specific data $\mathcal{Y}_k$ about a given policy intervention. The K data sets taken together form one large data set, denoted $\mathcal{Y} = \{\mathcal{Y}_k\}_{k=1}^K$. Each study has a site-specific parameter of interest $\theta_k \in \Theta_k$, which could be the average treatment effect of microloan access on household business expenditures, or the entire set of quantile treatment effects. The full data in each site k consists of N_k households, summing to N households in the total combined sample of all studies. In some cases, analysts will not have access to the full underlying data, only to the estimated effects and their standard errors from each of the K papers, denoted $\{\hat{\theta}_k, \hat{se}_k\}_{k=1}^K$. The general structure and intuition in the aggregation problem is the same in both cases and I consider models applicable to both situations.

Suppose that the analyst wishes to learn about the expected value of these $\{\theta_k\}_{k=1}^K$ parameters, incorporating uncertainty due to unobserved differences across study sites or settings. This is often of interest because such an object corresponds to some population average value of this parameter, say $\theta = E[\theta_k]$ with the expectation taken across the sites rather than across households. Incorporating site-level uncertainty permits inference not just to unobserved households in the existing sites but to unobserved sites themselves, and thus, such a procedure may permit extrapolation beyond the current set of studies.[13] One can learn about this θ using the evidence on $\{\theta_k\}_{k=1}^K$, but the optimal learning procedure depends on the heterogeneity or dispersion of $\{\theta_k\}_{k=1}^K$ around θ, denoted Σ_θ (Rubin 1981, Gelman et al 2004). This Σ_θ describes the signal strength of any θ_k for inference about the general effect θ, and thus the signal strength of θ as a predictor of θ_{K+1} if the sites are sufficiently comparable.[14] Here Σ_θ parameterizes a notion of generalizability of the evidence contained in $\mathcal{Y}$ to external settings, which captures the definition of external validity in Allcott (2015) and Dehejia et al. (2015). If $\Sigma_\theta = 0$, then θ is a perfect predictor of θ_{K+1}; if not, there will be some extrapolation error which grows large as the parameter Σ_θ grows large. Hence, this Σ_θ determines the optimal aggregation method and the relevance of θ for policy purposes.

[13] If this population parameter effectively does not exist, and it is impossible to update beliefs about economic mechanisms across settings, then much of economics is called into question.

[14] Technically the sites must be "exchangeable", this condition is discussed later in this section.

Joint estimation of θ and Σ_θ is the core challenge of aggregation across studies. Before aggregation occurs, the data has been analyzed separately in each study: this constitutes a "no pooling" model, where each effect θ_k is estimated using only the data from its own site, $\mathcal{Y}_k$. The resulting estimates, denoted $\{\hat{\theta}_k\}_{k=1}^K$, are only optimal for the set $\{\theta_k\}_{k=1}^K$ if $K < 3$ and if indeed no general common parameter θ exists.[15] The heterogeneity of $\{\hat{\theta}_k\}_{k=1}^K$ is generally biased upwards for Σ_θ because it includes the sampling variation of each $\hat{\theta}_k$ around its θ_k (Stein 1951, James and Stein 1961). These estimates or the underling data must be combined in some way to estimate θ, Σ_θ and θ_{K+1}. A "full pooling" aggregation method is an estimation procedure for θ which uses all the data $\mathcal{Y}$ and assumes that $\theta_k = \theta_{k'} \; \forall \; k, k'$. This assumption may be made explicitly or implicitly: any estimator that does not leverage the K-site structure nor estimate Σ_θ is a full pooling estimator. A "partial pooling" estimator uses the full data $\mathcal{Y}$ to estimate θ but does not assume $\theta_k = \theta_{k'} \; \forall \; k, k'$. A partial pooling aggregation procedure provides estimates of θ, Σ_θ as well as new estimates of $\{\theta_k\}_{k=1}^K$ produced by transferring some information across sites, denoted $(\tilde{\theta}, \tilde{\Sigma}_\theta, \{\tilde{\theta}_k\}_{k=1}^K)$.

Hierarchical modeling is a general framework for implementing partial pooling to aggregate evidence across studies which jointly estimates θ and Σ_θ. The defining characteristic of these models is a multi-level structure, which defines a set of parameters at the site level, $\{\theta_k\}_{k=1}^K$, a set of parameters at the population level, θ, and a relationship between them. One way to realize this structure is to use a multi-level likelihood which expresses the dependence of the data on the entire set of parameters (Efron & Morris 1975, Rubin 1981, Gelman et al. 2004). The "lower level" of the model describes the dependence between the data and local parameters in site k:

$$\mathcal{Y}_k \sim f(\cdot|\theta_k) \; \forall \; k. \tag{3.1}$$

The "upper level" of the model describes the potential for statistical dependence between local parameters and general parameters via some likelihood function $\psi(\cdot)$, which contains the parameter Σ_θ either implicitly or explicitly depending on the specific model. Hence, while in general $\psi(\cdot|\theta, \Sigma_\theta)$, this second argument is often implicit and thus, for simplicity, notationally suppressed. This upper level "general" or "parent" distribution is then denoted:

$$\theta_k \sim \psi(\cdot|\theta) \; \forall \; k. \tag{3.2}$$

[15] If K $\geq$ 3 all no-pooling estimators are risk-dominated in terms of MSE by partial pooling estimators. The formal proof of this statement is in Stein 1956, and further discussion is in Efron & Morris 1975.

A hierarchical likelihood contains both levels:

$$\mathcal{L}(\mathcal{Y}|\theta) = \prod_{k=1}^{K} f(\mathcal{Y}_k|\theta_k)\psi(\theta_k|\theta). \tag{3.3}$$

This likelihood structure nests common approaches to understanding the evidence from multiple studies, including both the no-pooling and full-pooling models. The model can detect these cases because the parameters that govern the $\psi(\cdot)$ function, including its implicit structure on Σ_θ, are estimated rather than imposed ex-ante. For example, the model may estimate that $\theta_k \approx \theta_{k'} \; \forall \; k, k'$, and hence that $\Sigma_\theta = 0$, if that is supported by the data. This result would recover the full-pooling model's solution, up to a degrees of freedom correction. Alternatively, the model can estimate very large dispersion in $\{\theta_k\}_{k=1}^K$ such that in fact $\{\tilde{\theta}_k\}_{k=1}^K = \{\hat{\theta}_k\}_{k=1}^K$, and as such recover the no-pooling model's solution. For applications in economics, where it is reasonable to think that neither extreme is likely to describe the data well, the model's main advantage is that it can recover a solution anywhere on the spectrum between these two extremes if that intermediate solution is most supported by the data. The model's estimation of θ and Σ_θ are appropriately influenced by the extent of this "partial pooling" (also called "shrinkage", because the estimates are pulled closer together, so the variance is "shrunk"). Hence, although some efficiency is lost if in reality $\Sigma_\theta \in \{0, \infty\}$, the hierarchical approach is more robust than the full pooling or no pooling approaches.

While in principle the hierarchical model could be specified with a nonparametric likelihood, a parametric structure is often preferable in low-data environments, such as evidence aggregation with a small or moderate number of studies.[16] Any partial pooling model must impose some structure to determine the extent of the pooling and how the pooling will be informed by the data. If the analyst faces a low-data environment at the cross-study level, this structure must not be too flexible or the model risks overfitting the scarce data that is available. Nonparametric methods often lack the power to deliver reliable inference at the general level, because they lack the structure of parametric models. As a result, hierarchical models used for evidence aggregation of scalar parameters often specify $\psi = N(\theta, \Sigma_\theta^2)$ due to the desirable frequentist properties of the resulting model (Efron and Morris 1975). This functional form appears more restrictive than the no-pooling or full-pooling models implemented using ordinary least squares regression, but in fact the Normal model still nests both of these cases since it can estimate $\Sigma_\theta \to \infty$ or $\Sigma_\theta = 0$

[16] A similar point and a proof of the nonparametric identification is provided in Andrews and Kasy 2017.

respectively. The no-pooling and full-pooling models do not specify parametric upper-level structure only because they impose such strong assumptions about Σ_θ. Parametric hierarchical likelihoods relax the assumptions on Σ_θ without providing too many degrees of freedom relative to the number of studies being aggregated.

When parametric structure is needed, the key insight of Rubin (1981) is that one can use knowledge of the sampling behaviour of certain statistics, in his case sample means and differences in sample means, to inform this choice. Even with limited information about the θ_k parameters, usually in the form of reported estimates and standard errors $\{\hat{\theta}_k, \hat{se}_k\}_K^{k=1}$, one often knows their approximate sampling behaviour under assumptions that seem reasonably mild. For example, sample means and by extension parameter estimates from linear regressions estimated by ordinary least squares often satisfy the Law of Large Numbers and the Central Limit Theorem, such that asymptotically

$$\hat{\theta}_k \sim N(\theta_k, \hat{se}_k^2). \tag{3.4}$$

In the full data case, one can analogously specify the within-sample variation using the structure imposed by the original studies. For example, if each study of a binary treatment indicator ran linear regressions of the form $y_{nk} = \mu_k + \tau_k T_{nk}$ for household n in site k, then the point estimates can be analytically replicated by the model

$$y_{nk} \sim N(\mu_k + \tau_k T_{nk}, \sigma_k^2). \tag{3.5}$$

Since these functional forms reflect underlying knowledge of the data or statistics being studied, hierarchical models based on these structures can more effectively separate the sampling variation from the between-study variation in effects (Rubin 1981). With the local variation specified in a particular way, the choice of the parent distribution that governs the between-study heterogeneity in effects can now be made in view of tractability and performance properties measured by the Mean Squared Error. These considerations typically motivate the use of Gaussian structure at the upper level of the model because they implement beneficial forms of shrinkage across the studies and have been shown to perform well for a variety of problems (McCullough and Neuhaus 2011, Efron and Morris 1975, Gelman et al 2004). In particular, if one is concerned with inference on only location and scale parameters, McCullough and Neuhaus (2011) shows that the Gaussian performs well even if the true underlying distribution is not Gaussian.[17]

[17] There are still important limitations to this approach, such as the restriction to single-peaked distributions which prevents for example detection of subclusters in the data, but with only seven studies the microcredit literature is unlikely to provide a fruitful setting for reliable cluster detection.

Hierarchical models do require that $\{\theta_k\}_{k=1}^K$ be "exchangeable", such that their joint distribution is invariant to permutation of the indices (Diaconis, 1977). This means the analyst must have no knowledge of the ordering or any sub-clustering of the treatment effects *a priori* that is not specified in the model (Rubin 1981). If economic theory demands that a particular covariate should be correlated in a certain way with the treatment effects, that can be translated into *conditional* exchangeability by introducing this covariate into the model. Yet theory and prior knowledge rarely provide certainty about these relationships, and building sufficiently weak structure that still permits inference on the role of covariates is typically challenging in a low-data environment. In the absence of strong prior knowledge about the treatment effects, exchangeability is a reasonable structure to impose (Gelman et al 2004).[18] Any future site for which θ_{K+1} is used to predict the effect must be exchangeable with the sites in the sample for this prediction to be valid, which is generally a requirement for predicting out-of-sample effects (see for example Allcott 2015).

3.1.2 Bayesian Implementation

While hierarchical models can be estimated using frequentist methods, in practice Bayesian inference offers several advantages. The first is accurate characterization of the uncertainty on all parameters produced by jointly estimating all unknowns. Commonly used maximum likelihood techniques estimate the upper level first and then condition on the point estimates using the "empirical Bayesian" approach from Efron & Morris (1975). This ignores the uncertainty about the upper level parameters, θ and Σ_θ, when computing uncertainty intervals on the lower level parameters, and thereby systematically underestimates the uncertainty at the lower level (Rubin 1981). This conditioning is required for tractability in the maximum likelihood estimation (MLE) framework as it is commonly implemented, because of the nonlinear interdependencies between $\{\theta_k\}_{k=1}^K$, θ, and Σ_θ.[19] By contrast, Bayesian inference jointly and simultaneously estimates all unknowns, accurately characterizing the uncertainty at every level of the model and producing coherent inference across levels.

Bayesian inference proceeds by specifying a prior on all unknowns, $\mathcal{P}(\theta)$, and combining it with the likelihood via Bayes' rule to generate the the joint posterior

[18] It is possible to build a more complex structure that allows "partial exchangeability" if this is desired. See Albert and Chib 1997.

[19] While MLE methods that do not inappropriately condition on unknowns are theoretically available, they seem to be largely unused in practice.

distribution $f(\theta|\mathcal{Y})$. The specification of a proper prior distribution ensures that $f(\theta|\mathcal{Y})$ is a proper probability distribution with desirable decision-theoretic properties such as admissibility (Efron 1982, Berger 2013). All proper Bayesian posteriors are consistent in the frequentist sense under similar conditions that make MLE consistent, as long as the prior has support over the true parameters, so aggregation performed in this framework will asymptotically deliver the correct answer as more studies are done (for the details see Van der Vaart 1998).

In a low-data environment, specifying informative priors can substantially improve the performance of the hierarchical model. If the analyst only has vague knowledge of the location of this likely region, then the priors can be made quite diffuse or "weakly informative" (Gelman et al 2008). If there is substantial expert knowledge of the likely values before seeing the data, perhaps from economic theory or previous studies, this can be incorporated using stronger priors. Even if the prior distributions introduce some bias due to incorrect centering, they may still improve the mean squared error of the estimation by reducing the variance (Chung et al. 2013, 2015). In low-data environments such as the cross-study level of the hierarchical model, overfitting and high variance can be the major obstacle to making reasonable inferences or predictions. Thus, the priors perform regularization in the classical sense: they introduce new information that constrains the fit of the model in ways beneficial to inference (Hastie, Tibshirani and Friedman 2009, section 10.2). Priors also increase the tractability and speed of the estimation by targeting regions of the parameter space that are more likely to contain relevant values.

Bayesian inference also provides a framework for thinking about the distribution of the treatment effect in a hypothetical future site θ_{K+1}. This is often the object of most interest for policymakers, but the distribution of this object must be computed accounting for the full joint posterior uncertainty rather than conditioning on a particular point estimate.[20]The Bayesian approach delivers the correct uncertainty interval in the form of posterior predictive inference (Gelman et al., 2004), which averages over the posterior uncertainty on the unknowns (θ, Σ_θ). Formally, the

[20]The full joint posterior uncertainty accounts perfectly for the uncertainty about how well the new location matches the old location, if the new site is exchangeable with the old sites, and the model structures are correct. If these conditions do not hold, we have modelling uncertainty, which is not accounted for in any meta-analytic methods at present (nor in any of the popular analytic methods used in empirical economics). However, attempts to account for this uncertainty within the status quo have all been informal, subjective and opaque; for example, experts picking and choosing which existing sites they think are most relevant for some particular new setting is not a transparent or reproducible aggregation method.

posterior predictive distribution is:

$$f(\theta_{K+1}|\mathcal{Y}) = \int \psi(\theta_{K+1}|\theta) f(\theta|\mathcal{Y}) \mathrm{d}\theta \tag{3.6}$$

Specifically for aggregating distributional effects, the Bayesian approach has another advantage in incorporating knowledge about the properties of θ, because it offers a natural mechanism for implementing constraints on parameters. If the parameter θ can only belong to some subset of the parameter space, $\mathcal{A}_\Theta \subset \Theta$, this produces the following restricted likelihood:

$$\mathcal{L}_{\mathcal{A}_\Theta}(\mathcal{Y}|\theta) = \mathcal{L}(\mathcal{Y}|\theta) \cdot \mathbb{1}\{\theta \in \mathcal{A}_\Theta\}. \tag{3.7}$$

While this is conceptually simple, implementing the restriction is not straightforward in some cases, such as the one considered here. However, because Bayesian inference treats unknown parameters as random variables, a statistical transformation of variables can impose constraints throughout the entire estimation without any distortion of the probability space. If θ is a multivariate random variable with PDF $p_\theta(\theta)$ then a new random variable $\theta^* = \mathrm{f}(\theta)$ for a differentiable one-to-one invertible function $\mathrm{f}(\cdot)$ with domain $\mathcal{A}_\theta$ has density

$$p(\theta^*) = p_\theta(\mathrm{f}^{-1}(\theta))|det(J_{\mathrm{f}^{-1}}(\theta))|. \tag{3.8}$$

Therefore to implement inference using $\mathcal{L}_{\mathcal{A}_\Theta}(\mathcal{Y}|\theta)$, leading to the correctly constrained posterior $f_{\mathcal{A}_\Theta}(\theta|\mathcal{Y})$, we specify the model as usual and then implement a transformation of variables from θ to θ^*. We then perform Bayesian inference using $\mathcal{L}(\mathcal{Y}|\theta^*)$ and $\mathcal{P}(\theta^*)$, derive $f(\theta^*|\mathcal{Y})$, and then reverse the transformation of variables to deliver $f(\theta|\mathcal{Y}) \cdot \mathbb{1}\{\theta \in \mathcal{A}_\Theta\}$. Frequentist implementation of constraints typically must reckon with the constraints twice, first in point estimation and second in interval estimation, and it can be costly to ensure coherence between the two or to extent the consequences to other parameters; the Bayesian implementation ensures coherence because the constraint is imposed on the parameter itself and is thus accounted for in both estimation and inference using the resulting joint posterior of all unknowns.

Tractability issues can arise in Bayesian inference on hierarchical models due to the same issues that lead frequentists to adopt Empirical Bayes, these can often be surmounted by the use of Markov Chain Monte Carlo (MCMC) methods. These methods construct 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 the posterior distribution's value at the draw. While these chains always converge to the correct distribution in the limit, popular algorithms such as the Metropolis-Hastings or Gibbs samplers can be prone to inefficient random walk behavior when the unknowns are correlated, as with hierarchical models. Instead, I use Hamiltonian Monte Carlo (HMC) methods, which can better estimate hierarchical models (Betancourt and Girolami, 2013).[21]

3.2 Limited Information Asymptotic Quantile Models

Consider the task of aggregating sets of quantile treatment effects and assessing their generalizability. First recall that the uth quantile of some outcome is the value of the inverse CDF at u:

$$Q_Y(u) = F_Y^{-1}(u). \tag{3.9}$$

Performing quantile regression for some quantile u in site k when the only regressor is the binary treatment indicator T_{nk} requires estimating:

$$Q_{y_{nk}|T}(u) = \beta_{0k}(u) + \beta_{1k}(u)T_{nk} \tag{3.10}$$

For a single quantile u, the treatment effect is the univariate parameter $\beta_{1k}(u)$. If there is only one quantile of interest, a univariate Bayesian hierarchical model can be applied, as in Reich et al (2011). But in the microcredit data, researchers estimated a set of 10 quantiles $\mathcal{U} = \{0.05, 0.15, ..., 0.95\}$ and interpolated the results to form a "quantile difference curve". This curve is constructed by computing the quantile regression at all points of interest:

$$Q_{y_{ik}|T} = \{Q_{y_{ik}|T}(u) = \beta_{0k}(u) + \beta_{1k}(u)T_{ik} \ \ \forall \ u \in \mathcal{U}\} \tag{3.11}$$

The results of this estimation are two $|\mathcal{U}|$-dimensional vectors containing intercept and slope parameters. For the microcredit data, I work with the following vector of

[21] HMC uses discretized Hamiltonian dynamics to sample from the posterior, which can be combined with the No-U-Turn sampling method (NUTS) to auto-tune the step sizes in the chain (Hoffman and Gelman, 2011). This algorithm is automated in the software package Stan, a free statistical library which calls C++ to fit Bayesian models from R or Python (Stan Development Team, 2017).

10 quantile effects:

$$\begin{aligned}\beta_{0k} &= (\beta_{0k}(0.05), \beta_{0k}(0.15), ...\beta_{0k}(0.95)) \\ \beta_{1k} &= (\beta_{1k}(0.05), \beta_{1k}(0.15), ...\beta_{1k}(0.95))\end{aligned} \tag{3.12}$$

The quantile difference curve is the vector β_{1k}, often linearly interpolated. With a binary treatment variable, the parameters in a quantile regression are simple functions of unconditional outcome quantiles. Let $Q_{0k}(u)$ be the value of the control group's quantile u in site k, and let $Q_{1k}(u)$ be the value of the treatment group's quantile u in site k. Then:

$$\begin{aligned}Q_{0k} &= \{Q_{0k}(u) \ \ \forall \ u \in \mathcal{U}\} \\ Q_{1k} &= \{Q_{1k}(u) \ \ \forall \ u \in \mathcal{U}\}.\end{aligned} \tag{3.13}$$

Then the vectors of intercepts and slopes for the quantile regression curves can be reformulated as

$$\begin{aligned}\beta_{0k} &= Q_{0k} \\ \beta_{1k} &= Q_{1k} - Q_{0k}.\end{aligned} \tag{3.14}$$

Hence, while the quantile difference curve β_{1k} need not be monotonic, it must imply a monotonic Q_{1k} when combined with a monotonic β_{0k}. The fact that any inference done quantile-by-quantile may violate monotonicity of $(Q_1, Q_0, \{Q_{1k}, Q_{0k}\}_{k=1}^K)$ is a well-understood problem (Chernozhukov et al. 2010). Partial pooling for aggregation can exacerbate this problem because even if every lower level Q_{1k} and Q_{0k} satisfies monotonicity, their "average" or general Q_1 and Q_0 may not do so. Thus, unlike quantile crossing within a sample, the crossing in this setting is not necessarily the result of an incorrect asymptotic assumption or an extrapolation to a poorly-covered region of the covariate space. Indeed, for binary treatment variables, the within-sample estimators always satisfy monotonicity, but the averaging and pooling of these estimators may introduce crossing where none existed.[22] Ideally, therefore, an aggregation model should fit all quantiles simultaneously, imposing the monotonicity constraint. Aggregating the quantile difference curves, $\{\beta_{1k}\}_{k=1}^K$, requires more structure than aggregating quantile-by-quantile, but permits the transmission of information across quantiles.

I propose a general methodology to aggregate reported information on quantile

[22] Yet even if quantile crossing does not arise, neighboring quantiles contain information about each other not just because of monotonicity but because smooth distributions have quantiles that tend to lie close to each other; using that information can improve the estimation and reduce posterior uncertainty.

difference functions building on the approach of Rubin (1981) and a classical result from Mosteller (1946) about the joint distribution of sets of empirical quantiles. Mosteller shows that if the underlying random variable is continuously distributed, then the asymptotic sampling distribution of a vector of its empirical quantiles is a multivariate Normal centered at the true quantiles and with a known variance-covariance structure. This implies that the difference of the empirical quantile vectors from two independent samples, $\beta_{1k} = (Q_{1k} - Q_{0k})$, is also asymptotically a multivariate Gaussian. The theorem offers a foundation for a hierarchical quantile treatment effect aggregation model using the knowledge that the sampling variation is approximately a multivariate Gaussian, and that as a result modelling the parent distribution as Gaussian will be both tractable and have attractive performance (Rubin 1981, Efron and Morris 1975). The resulting analysis requires only the limited information reported by each study (although it can be fit to the full data) and is applicable to any continuous distribution as long as there is sufficient data in each of the studies to make the asymptotic approximation reasonable.

For this model, the data are the vectors of sample quantile differences $\{\hat{\beta}_{1k}\}_{k=1}^{K}$ and their sampling variance-covariance matrices $\{\hat{\Xi}_{\beta_{1k}}\}_{k=1}^{K}$. Thus, the lower level $f(\mathcal{Y}_k|\theta_k) = f(\beta_{1k}|\beta_{1k})$ is given by the expression:

$$\hat{\beta}_{1k} \sim N(\beta_{1k}, \hat{\Xi}_{\beta 1k}) \; \forall \; k \tag{3.15}$$

The upper level of the model $\psi(\theta_k|\theta)$ is therefore:

$$\beta_{1k} \sim N(\beta_1, \Sigma_1) \; \forall \; k. \tag{3.16}$$

However, the estimated $(\tilde{\beta}_1, \{\tilde{\beta}_{1k}\}_{k=1}^{K})$ from this likelihood may not respect the implied quantile ordering restriction when combined with the estimated control quantiles, even if $\hat{\beta}_{1k}$s do. We need to add the relevant constraints to this model, but these difference functions are not the primary objects on which the constraints operate. While $(\beta_1, \{\beta_{1k}\}_{k=1}^{K})$ need not be monotonic, they must imply monotonic $(Q_1, \{Q_{1k}\}_{k=1}^{K})$ when combined with $(Q_0, \{Q_{0k}\}_{k=1}^{K})$. Since the objects $(Q_1, Q_0, \{Q_{1k}, Q_{0k}\}_{k=1}^{K})$ define the constraints, they must appear in the model.

Once the quantiles $(Q_1, Q_0, \{Q_{1k}, Q_{0k}\}_{k=1}^{K})$ appear in the model, transforming them into monotonic vectors will fully impose the relevant constraint on $(\beta_1, \{\beta_{1k}\}_{k=1}^{K})$. This strategy exploits the fact that Bayesian inference treats unknown parameters as random variables, so applying the transformation of variables formula and then reversing the transform at the end of the procedure completely preserves the pos-

terior probability mass, and hence correctly translates the uncertainty intervals. I proceed with a transform proposed for use in Stan (2016), but in theory any valid monotonizing transform will do, since it is always perfectly reversed.[23] Consider monotonizing the $|\mathcal{U}|$-dimensional vector β_0, with uth entry denoted $\beta_0[u]$. One can map β_0 to a new vector β_0^* as follows:

$$\beta_0^*[u] = \begin{cases} \beta_0[u], & \text{if } u = 1 \\ log(\beta_0[u] - \beta_0[u-1]) & \text{if } 1 < u < |\mathcal{U}| \end{cases} \tag{3.17}$$

Any vector β_0 to which this transform is applied and for which inference is performed in the transformed space will always be monotonically increasing. For the rest of the paper, I denote parameters for which monotonicity has been enforced by performing inference on the transformed object as in equation 3.17 with a superscript m. Thus, by applying the transform, I work with β_0^m rather than an unconstrained β_0.

Employing a monotonizing transform is an appealing alternative to other methods used in the econometrics literature to ensure monotonicity during quantile regression. Restricting the Bayesian posterior to have support only on parameters which imply monotonic quantiles means that, for example, the posterior means are those values which are most supported by the data and prior information from the set which satisfy the constraint. Frequentist solutions such as rearrangement, smoothing or projection each prevent the violation of the constraint in one specific way chosen *a priori* according to the analyst's own preferences (He 1997, Chernozhukov et al. 2010). While each strategy performs well in terms of bringing the estimates closer to the estimand (as shown in Chernozhukov et al. 2010) the Bayesian transformation strategy can flexibly borrow from each of the strategies as and when the data supports their use. Imposing the constraint throughout the inference avoids the additional complications of choosing *when* during aggregation one should implement the ex-post fixes proposed in the frequentist literature; for example, in the case of rearrangement, it would be hard to interpret the result of partially pooling information on the 25th quantile only to have some other quantile substituted in for certain studies ex-post.

Equipped with this monotonizing transform, it is now possible to build models with restricted multivariate Normal distributions which only produces monotonically increasing vectors. I propose the following model to perform aggregation in a hierarchical framework, taking in the sets of empirical quantiles $\{\hat{Q}_{1k}, \hat{Q}_{0k}\}_{k=1}^K$ and

[23] While some transforms may perform better than others in certain cases, to my knowledge there is little research on this issue that presently permits us to choose between transforms.

their sampling variance-covariance matrices $\{\hat{\Xi}_{1k}, \hat{\Xi}_{0k}\}_{k=1}^{K}$ as data. For this hierarchical quantile set model, the lower level $f(\mathcal{Y}_k|\theta_k)$ is:

$$\begin{aligned} \hat{Q}_{0k} &\sim N(\beta_{0k}^m, \hat{\Xi}_{0k}) \ \ \forall\ k \\ \hat{Q}_{1k} &\sim N(Q_{1k}^m, \hat{\Xi}_{1k}) \ \ \forall\ k \\ \text{where} \ \ Q_{1k} &\equiv \beta_{0k}^m + \beta_{1k} \end{aligned} \tag{3.18}$$

The upper level $\psi(\theta_k|\theta)$ is:

$$\begin{aligned} \beta_{0k}^m &\sim N(\beta_0^m, \Sigma_0) \ \ \forall\ k \\ \beta_{1k} &\sim N(\beta_1, \Sigma_1) \ \ \forall\ k \\ \text{where} \ \ \beta_1 &\equiv Q_1^m - \beta_0^m \end{aligned} \tag{3.19}$$

The priors $\mathcal{P}(\theta)$ are:

$$\begin{aligned} \beta_0^m &\sim N(0, 1000 * I_{10}) \\ \beta_1 &\sim N(0, 1000 * I_{10}) \\ \Sigma_0 &\equiv diag(\nu_0)\Omega_0 diag(\nu_0)' \\ \Sigma_1 &\equiv diag(\nu_1)\Omega_1 diag(\nu_1)' \\ \text{where} \ \ \nu_0, \nu_1 &\sim \text{halfCauchy}(0, 20) \text{ and } \Omega_0, \Omega_1 \sim LKJCorr(1). \end{aligned} \tag{3.20}$$

This formulation is convenient as the form of $\hat{\Xi}_{1k}$ is exactly derived in the Mosteller (1946) theorem, though the individual entries need to be estimated. The structure could be modified to take in the empirical quantile treatment effects $\{\hat{\beta}_{1k}\}_{k=1}^{K}$ and their standard errors instead of $\{\hat{Q}_{1k}\}$ if needed. The model imposes no structure on (Σ, Σ_0), other than the logical requirement of positive semi-definiteness. This complete flexibility is made possible by the discretization of the quantile functions; these matrices could not take unconstrained form if the quantile functions had been modelled as draws from Gaussian Processes.[24] Overall, this structure passes information across the quantiles in two ways: first, by imposing the ordering constraint, and second, via the functional form of $\hat{\Sigma}_k$ from the Mosteller (1946) theorem.

The above model implements partial pooling not only on the $\{\beta_{1k}\}_{k=1}^{K}$ parameters

[24] Gaussian Processes in general are too flexible to fit at the upper level of these models for this application, and popular covariance kernels tend to have identification issues that limit their usefulness in the current setting. In particular, most tractable and popular kernels do not permit the separation of dispersion of points within the functional draws from dispersion of points across the functional draws.

but also on the $\{\beta_{0k}\}_{k=1}^{K}$ parameters, that is, the control group quantiles. The technical reason for this is that one needs to define a notion of a general β_0 in order to define the constraint on the general β_1 and the predicted $\beta_{1,K+1}$. However, this structure also provides us with useful insight that allows us to better interpret the results of the partial pooling on $\{\beta_{1k}\}_{k=1}^{K}$. Suppose for example that we observe substantial pooling on $\{\beta_{1k}\}_{k=1}^{K}$, but we also observe this on $\{\beta_{0k}\}_{k=1}^{K}$; in that case, we observe similarities in the treatment effects perhaps only because we have studied places with similar control groups. In that case it will be hard to justify extrapolation to another setting with a substantially different value of β_{0k}. On the other hand, suppose that we observe substantial pooling on $\{\beta_{1k}\}_{k=1}^{K}$ but no pooling at all on $\{\beta_{0k}\}_{k=1}^{K}$. Then we have learned much more generalisable information, because we now know that the treatment effects can be similar even when the underlying control distributions are different.

3.2.1 Model Performance

To assess the performance of the model, I provide Monte Carlo simulations under a variety of data scenarios and report the coverage of the posterior intervals. Ideally, the 50% posterior interval should contain the true parameter 50% of the time, and the analogous property should hold for the 95% posterior interval. When the data is simulated from the model itself this property is guaranteed in Bayesian inference, however, in practice one typically does not have the luxury of fitting data that one knows originates from a particular model. Therefore, all the monte carlo simulations I provide here fit the model above to data generated from a somewhat different model. In particular, I always simulate data for which my priors are incorrect; as the priors are reasonably diffuse they should not compromise the inference, and nor do they.

The results in table 1 show that the model typically provides approximately nominal coverage on β_1, and often provides greater than nominal coverage, regardless of the generating process. However, inference on the β_0 and the covariation parameters Σ_0, Σ_1 is more sensitive to underlying conditions. When there is large variation in the data within sites, the model can have difficulty with achieving nominal coverage on the 50% interval for these parameters, although the 95% interval usually retains its coverage properties. The fact that the model has trouble when the data exhibits large variation reflects one of the conditions of the Mosteller (1946) theorem, namely that the underlying density that generates the data is not vanishing in the neighbourhood of the quantile. While this condition is formally satisfied in the sim-

ulations, the model seems to be affected regardless: the poor average performance in these cases is generated by difficulty characterising the extremal quantiles where the density is thinnest.

Encouragingly, when the data variation is moderate or small, the model does reasonably well on most parameters even when the full pooling or no pooling cases approximately hold. The no pooling case provides some trouble for inference on Σ_0, Σ_1 due to the extreme cross-site variation. Large within-site data variation seems to cause difficulties for the 50% intervals in the full pooling inference, but the 95% intervals retain their coverage properties even in this case. There are some results in the table that do not fit the broad patterns laid out here, but this may be due to the relatively small number of MC runs (due to the relatively long time it takes to run the model). The results point to overall good performance, although suggesting that caution should be applied when approaching data sets that have high variance or heavy tails even when the theoretical conditions for asymptotic normality are formally satisfied.

3.2.2 Interpretation

The goal of the hierarchical model is to estimate the central location and the dispersion in the distributions from which each of the observed $\{\beta_{0k}, \beta_{1k}\}$ are drawn. This permits analysts and policymakers to formulate expectations about what may be likely in future settings. The expectations are taken over the distribution of the vectors $\{\beta_{0k}, \beta_{1k}\}$, not over households; this point is crucial to avoid confusion about how to interpret quantiles in a hierarchical setting. The estimate of β_0 is an estimate of the expected marginal distributions of the control groups' outcomes in the set of sites exchangeable with the K sites at hand, subject to the constraint that these objects all be monotonic. The estimate of β_1 is an estimate of the expected differences between the marginal outcomes of the treatment and control groups in the set of exchangeable sites, subject to the known properties of each of the groups distributions. While the monotonicity constraint complicates the interpretation relative to unconstrained Gaussian models, the broad intuition is that the parameters β_0, β_1 provide an estimate of the centrality of the individual distributions over the vector spaces in which $\{\beta_{0k}, \beta_{1k}\}$ live.

Quantile effects are often subject to misinterpretation. Quantiles do not satisfy laws of iterated expectations, so the treatment effects at the quantiles of the outcome distribution (which is what is delivered here) are not the quantiles of the distribu-

tion of treatment effects.[25] Another source of confusion is that while unconditional quantiles of a distribution correspond to specific data points in the sample, these data points (and the individuals who produce them) are not meaningful in themselves. It does not make sense to think of quantile estimation as applying to specific households nor "tracking" them across time or place. As stressed in Koenker 2005, the fact that one can locate the specific data point that sits at a particular sample quantile does not mean that this datum is deeply related to the population quantile in some way: it happens to be the best estimate of that population quantile, nothing more. The population quantile is a *parameter*, not an "uber household"; the sample household whose outcome value is "selected" as a quantile estimate is not playing the role of an individual in the world but rather the role of a sample order statistic of similar general stature to a sample mean. If the sample had realised differently, a different household would have been selected as the estimate, but the population quantile remains the same.[26]

Koenker's point extends fully to the case of aggregation of quantiles. The hierarchical nature of the model does not imply that the information or outcome of any specific individual household is being passed (or not passed) up to the "general" level of the model. Rather, the model posits the potential existence of a general or meta-distribution which governs the K observed distributions and observed differences between the treatment and control distributions in each site. We then study the means and covariance matrices of these parent distributions in an attempt to understand how useful that structure is for prediction. In this context, the relative positions of e.g. the 25th quantile in site 1 and the 25th quantile in site 2, and how similar they are to the expected value of the 25th quantile taken across all the sites, is simply a question: how similar are the value of the 25th quantiles are in all the sites we have studied? If the answer is "not very similar" this is not a problem for interpretation of the quantiles, but simply a possible state of reality we have fully anticipated in the model and which will be expressed by a very large covariance matrix (or at least a large entry on the diagonal for that quantile).

In fact, the expected value of the quantile treatment effect in the hierarchical context does not find solutions corresponding to single households, but rather, takes weighted averages of the K solutions. Suppose for example that all consumption

[25] While it would be nice to know the latter object, this is not estimable without considerable additional structure

[26] Another common misunderstanding is that "only" the "chosen" household contributes to the inference at any given quantile. In fact, in quantile estimation, as in mean estimation, all the household data points together determine the best estimate of a point (or set of points) corresponding to a population object of interest.

values in one site lie below all those in another site. This does not mean that the lower quantiles of the general distribution are all taken from the first site, nor that the upper quantiles of the general distribution are all taken from the second site. Instead, the procedure examines each site's data set quantile by quantile and asks "What is the average estimate of this quantile, and how similar are these quantiles across sites?". The expected value of a given quantile taken over all sites for example may not correspond to any household's value or any specific quantile effect in any site, any more than the expectation of a set of values would correspond to any one of the values: it wouldn't, except by chance. The expected median is not necessarily the median of the K medians and nor does it need to be such in order to be interpreted: the expectation is formulated over a posited distribution of medians, which corresponds to the question: "What should I expect the median to be in these kinds of places?"

Hence, β_0 and β_1 should not be interpreted as the quantiles or differences of some aggregated data set; rather, they are the *expected* quantiles and differences in any given site with this expectation taken across sites, subject to the monotonicity constraint in this case. The hierarchical model does not attempt to arrange all the individual data points or quantile difference estimates in some kind of grand order (nor would it be clear how to interpret such an exercise). Quantile regression permits one to infer the shapes of distributions, not to track individuals specifically over time or over ranks of relative groups one could decide to place them in. The goal of the hierarchical quantile model is to infer a set of true differences that correspond to a population distribution's response to a treatment, and to understand how different these responses are across settings.

3.3 Full Information Parametric Quantile Models

The strength of the model based on the Mosteller (1946) theorem is that it works for any continuous outcome variable; its weakness is that it *only* works for continuous variables. In the microcredit data, this approach will work for household consumption, consumer durables spending and temptation goods spending. But household business profit, revenues and expenditures are not continuous because many households either did not own or did not operate their businesses in the month prior to being surveyed and therefore recorded zero for these outcomes. This creates large "spikes" at zero in the distributions, as shown in the histograms of the profit data for the sites (Appendix A, figure 7). This spike undermines the performance of the Mosteller theorem and of the nonparametric bootstrap for standard error calcula-

tion. The Mexico data provides the cleanest example of this, shown in Appendix A figure 8: the first panel is the result of using the Mosteller asymptotic approximation, and the second panel is the result of the nonparametric bootstrap applied to the standard errors on the same data. The former produces the dubious result that the uncertainty on the quantiles in the discrete spike is the same as the uncertainty in the tail; the latter produces the dubious result that the standard errors are exactly zero at most quantiles.

The potential for quantile regression techniques to fail when the underlying data is not continuous is a well-understood problem (Koenker and Hallock 2001; Koenker 2011). In some cases, "dithering" or "jittering" the data by adding a small amount of random noise is sufficient to prevent this failure and reliably recover the underlying parameters (Machado and Santos Silva, 2005). As shown in Appendix A this does not work for microcredit data. An alternative method to aggregate the quantile treatment effects must be developed for these three outcomes, and for any outcome of interest which is not continuously distributed.

However, when the Mosteller (1946) approximation cannot be applied due to the presence of discrete probability masses in the distribution of the outcome variable, the researcher typically has some contextual or prior economic knowledge of why these masses arise. It may be possible to explicitly model the processes that generate the probability density functions (PDFs) of household outcomes. I pursue a flexible and richly-parametised approach at the data level using mixtures of distributions in which treatment can affect all aspects of the shapes of the component distributions as well as the weights on each of the components themselves.[27] While this requires substantial input from the researcher and the aggregation model must be tailored to each specific case, this method will automatically solve the two problems discussed with quantile aggregation. Directly modelling the PDFs as proper densities, which therefore integrate to proper and thus weakly monotonic Cumulative Density Functions, will automatically deliver monotonically increasing vectors of quantiles. The model transfers information across neighbouring quantiles because they are directly linked by the functional form assumptions.

For the household business variables in the microcredit data, there is sufficient

[27]I do not use the popular Bayesian nonparametric technique of using mixtures of Gaussians governed by a Dirichlet Process because applying a hierarchical model to these infinite-dimensional PDFs does not provide the information we want in this application. For example, the Dirichlet Process is governed by a base distribution and a parameter α, but placing a hierarchy on either of these does not readily admit a useful interpretation: we actually do not care how many components are used to model the distributions (governed by α) nor how similar are the component distributions to one another (governed by the parameters on the components), we care about how similar is the ultimate change in the overall distributional shape resulting from treatment.

contextual economic information to build a parametric model. In this setting, economic theory predicts that these variables should be mixtures of spikes at zero and continuous tails because they are the output of a partially discrete decision process. First, a household has an extensive margin decision to make about whether to operate a business this season or not. This decision may be different at different times of the year depending on the outside options, as many households in these contexts engage in seasonal agricultural labour or intermittent construction labour for part of the year, only operating their businesses during the "lean season". Only those households who decide to open and operate their businesses go on to make an intensive margin decision, the result of which manifests some continuous expenditures, revenues and profit. This explains the spike at zero observed in all three business variables, which is a real feature of the generating process and not an artefact of the data collection.

Economic theory and prior research suggest that the continuous portions of business variables such as revenues and profit tend to follow power laws or other fat-tailed laws (Stiglitz 1969, Gabaix 2008, Allen 2014, Bazzi 2016). Hence, the outcome PDF can be modeled as a mixture of three distributions: a lower tail, a spike at zero, and an upper tail. As T_{nk} may affect the mass in the components and the shape of the tail components, I specify treatment effects on all aspects of this mixture PDF. The model can then aggregate the effect of the treatment on each of the parameters that govern the distribution, as well as the implied quantile treatment effects.

The risk in specifying any parametric structure based on contextual and prior information is that our knowledge may be insufficient or incorrect, leading to poor inference. It is advisable therefore to assess the sensitivity to the choice of functional form, as well as to assess model fit and avoid reliance on models that fail to approximate the data well. In the case of household business variables the distribution of the tails could reasonably be modelled by a Pareto distribution, as in Piketty 2015 or Bazzi 2016. However, a Log-Normal distribution would allow for more mass near the lower bound of the distribution per Roy 1950 and is analogous to log transforming the positive values in the sample, a common practice in applied microeconomics (see for example Banerjee et al 2015b). I fit both models to the microcredit data and examine the posterior fit of each model in order to select between them, and I determine what if any inferences are robust to the choice of tail distribution.

Consider the following tailored hierarchical PDF model to aggregate the quantile effects on household business profit. Denote the probability mass in the jth mixture component for a household n with treatment status T_{nk} to be $\Lambda_j(T_{nk})$ for $j = 1, 2, 3$. This dependence can be modeled using a multinomial logit specification, denoting

the intercept in site k for mixture component j as α_{jk} and the treatment effect as π_{jk}. For the spike at zero, the Dirac delta function can be used as a distribution, denoted $\delta(x)$ for a point mass at x. If using the LogNormal distribution for the tails, then each are governed by a location parameter and a scale parameter. The latter can only be positive valued so I employ the exponential transform to ensure the support constraint is satisfied. I model the location parameter using a linear regression format in which the value for the control group in site k is μ_k and the value for the treatment group is $\mu_k + \tau_k$. The scale parameter is modelled with the control group's value being $\exp(\sigma_k^c)$ and the treatment group's value being $\exp(\sigma_k^c + \sigma_k^t)$.

The lower level of the likelihood $f(\mathcal{Y}_k|\theta_k)$ is specified according to this mixture distribution. Let $j = 1$ denote the negative tail of the household profit distribution, let $j = 2$ denote the spike at zero, and let $j = 3$ denote the positive tail. Then the household's business profit is distributed as follows:

$$\begin{aligned} y_{nk}|T_{nk} \sim\ & \Lambda_{1k}(T_{nk}) LogNormal(-y_{nk}|\mu_{1k} + \tau_{1k}T_{nk}, \exp(\sigma_{1k}^c + \sigma_{1k}^t T_{nk})) \\ & + \Lambda_{2k}(T_n)\delta_{(0)} \\ & + \Lambda_{3k}(T_n) LogNormal(y_{nk}|\mu_{3k} + \tau_{3k}T_{nk}, \exp(\sigma_{3k}^c + \sigma_{3k}^t T_{nk}) \ \forall\ k \\ \text{where}\ \ \Lambda_{jk}(T_{nk}) =\ & \frac{\exp(\alpha_{jk} + \pi_{jk}T_{nk})}{\sum_{j=1,2,3}\exp(\alpha_{jk} + \pi_{jk}T_{nk}))} \end{aligned} \tag{3.21}$$

The upper level $\psi(\theta_k|\theta)$ is:

$$(\alpha_{1k}, \alpha_{2k}, \alpha_{3k}, \pi_{1k}, ...)' \equiv \zeta_k \sim N(\zeta, \Upsilon)\ \forall\ k \tag{3.22}$$

For tractability and simplicity I enforce diagonal Υ for the microcredit analysis. This prevents the model from using correlations in the distribution of say $\{\alpha_{1k}\}_{k=1}^K$ and $\{\alpha_{2k}\}_{k=1}^K$, a restriction enforced only because the correlations are hard to estimate with seven sites and this can introduce substantial additional variance into the estimation procedure. The restriction also means that the model needs only weak priors $\mathcal{P}(\theta)$ as follows:

$$\begin{aligned} \zeta &\sim N(0, 10) \\ \Upsilon &\equiv diag(\nu_\Upsilon)\Omega_\Upsilon diag(\nu_\Upsilon)' \\ \nu_\Upsilon &\sim \text{halfCauchy}(0, 5) \\ \Omega_\Upsilon &= I_{|\zeta|} \\ \alpha_{mk} &\sim N(0, 5). \end{aligned} \tag{3.23}$$

The Pareto model is defined analogously in the Online Appendix.

3.3.1 Recovering quantile effects from this model

Quantile recovery is nontrivial in this setting because mixture distributions in general do not have analytical quantile functions. However, because the mixture distribution in this particular model has components with disjoint supports, one can apply the method of Castellacci (2012) to compute the quantiles analytically. Given the profit model above I derive the quantile function using this method for each model. The result for the LogNormal model is:

$$\begin{aligned} Q(u) = -\text{LogNormal}^{-1}\left(1 - \frac{u}{\Lambda_1(T_n)} \mid \mu_{1k} + \tau_{1k}T_{nk}, \exp(\sigma^c_{1k} + \sigma^t_{1k}T_{nk})\right) * \mathbb{1}\{u < \Lambda_1(T_n)\} \\ + 0 * \mathbb{1}\{\Lambda_1(T_n) < u < (\Lambda_1(T_n) + \Lambda_2(T_n))\} \\ + \text{LogNormal}^{-1}\left(\frac{u - (1 - \Lambda_3(T_n))}{\Lambda_3(T_n)} \mid \mu_{3k} + \tau_{3k}T_{nk}, \exp(\sigma^c_{3k} + \sigma^t_{3k}T_{nk})\right) * \mathbb{1}\{u > (1 - \Lambda_3(T_n))\} \end{aligned} \tag{3.24}$$

The Pareto model is defined analogously. The full posterior distribution of the entire set of quantiles and thus the implied quantile treatment effects is easily computed from the posterior distribution of the unknown parameters within the Bayesian framework, by applying the computation to every MCMC draw from the joint posterior distribution. This method ensures that the uncertainty on the quantiles implied by the uncertainty on the parameters that govern the tailored hierarchical PDF model is translated exactly.

To analyse household business expenditures, revenues and profits, I fit both the Pareto and LogNormal models and use posterior predictive checking to select the structure that fits the data best (Gelman et al 2004). This requires simulating data from the posterior distribution of each model, and then comparing the quantiles of the simulated data to the quantiles of the real data. As the Online Appendix shows, the LogNormal model outperforms the Pareto in terms of predicting the actual observed control group quantiles, particularly in the right tail. The Pareto shape has too little mass near zero and too much mass in the tail relative to the LogNormal. However, the broad patterns observed in the results of the LogNormal model are also observed in the Pareto model, and are thus robust to choice of tail distribution (see Online Appendix). In particular, both models show a precise zero impact at most quantiles, and then the potential for large increases in the right tails.

4 Results

4.1 Main Results

4.1.1 Consumption Variables

The limited information hierarchical quantile aggregation model based on the Mosteller (1946) theorem is used to estimate the general quantile treatment effects for household consumption, consumer durables spending and temptation goods. Figure 1 shows the posterior distribution of the generalized quantile treatment effects β_1 for each of the consumption outcomes, with the full-pooling aggregation results shown for comparison. Each graph has a line depicting the posterior mean of the quantiles, a shaded area showing the central 50% posterior interval for the quantiles, and a lighter shaded area showing the central 95% posterior interval. The results shows that microcredit has a precise, generalizable zero effect below the 75th percentile. Beyond this point there is an imprecise positive effect that exhibits high variance across sites. A comparison of the no pooling model (the estimates from the papers), the partial pooling model and the full pooling model, is in table 2. The results show that the full pooling model and the BHM typically produce similar output for the 5th to 75th quantile of household outcomes, but diverge in the upper tail. This difference itself may be a signal the presence of heterogeneous effects across settings, although when the difference is small it is challenging to know whether this is simply a degrees of freedom issue.

A clearer signal of the detected heterogeneity in the tail is that the inference on the predicted effect in the next site is substantially more uncertain in the hierarchical context. Posterior predictive distributions of these consumption variables are shown in figure 2 with the full-pooling model for comparison. The results show considerably more uncertainty about the outcomes, particularly at the right tail, than would be suggested by taking either the full-pooling model or the posterior distribution of β_1 from the partial pooling model (figure 1). Particularly for household consumption, the model declines to make any strong prediction at the tail, with a positive effect being only moderately more likely than a negative effect at the 90th percentile and above. Formal pooling metrics from previous literature computed for the three consumption variables are shown in Appendix B, table 4. The level of pooling on the quantile difference curves is intermediate, but typically much more than the level of pooling on the quantiles themselves.[28] In any case, the full-pooling model seems

[28]This should be interpreted with caution, and I have not emphasized these results because

to underestimate the uncertainty on effect on the upper quantiles that one might expect to see in the next site, and thus delivers inference with higher precision than is warranted by the evidence.

The site-specific results from the Bayesian hierarchical model illuminate how these general results arise at the upper level of the model. Table 2 and figures in Appendix C display these results for each site, with the no-pooling results shown for comparison. There is moderate, although not extensive, pooling of the functions together for these outcomes. However, the curves are typically quite similar to each other even in the no-pooling model, with most of their posterior mass located near zero for the majority of the quantiles. This supports the notion of a generalizable and replicable zero effect on the shape of the distribution, except at the upper tail where there is both more uncertainty within each site and less apparent similarity across sites.

4.1.2 Business Variables

The quantile treatment effect results from the the Lognormal hierarchical PDF models for all business outcomes are shown in figure 3 with the full pooling results shown for comparison. A comparison of the no pooling model (the estimates from the papers), the partial pooling model and the full pooling model, is in table 3. The models find a precise and generalizable zero effect below the 75th percentile, although the lower tail of profit is an imprecise zero. Above the 75th percentile there is a large positive point estimate, but much less precision and more uncertainty, due to heterogeneity both within and across sites. By contrast, the full pooling models find much larger and more precise "statistically significant" effects in the tails.[29] In a frequentist sense, the apparently "statistically significant" results in the upper tails "detected" in the full pooling model are eliminated by the application of a hierarchical model. In a Bayesian sense, the full pooling model is misleadingly precise in the upper tail, and the posterior uncertainty we should have about these tail effects

in this case the different pooling metrics often return substantially different results for the same variables. This suggests that perhaps there is more work to be done on pooling metrics in practice. But overall, there is almost zero pooling of the control group quantiles according to two of the three metrics, and intermediate pooling according to the third metric. All the metrics show much more pooling on β_1 than on β_0: this indicates that the control groups are substantially different in across studies, and suggests that the zero impact along most of the distribution is indeed generalizable across heterogeneous contexts.

[29] The difference is dramatic because when the tails are sparse, a little more pooling goes a long way; yet as with consumption, the presence of different point estimates in the tails is itself a signal of heterogeneity across settings, such that the full pooling assumption is unwarranted in this setting and unlikely to produce reliable inference.

is much larger. However, there is more than a 90% probability of a positive effect on the 85th and 95th quantiles of all the distributions, suggesting that microcredit may indeed be affecting these tails in some positive way.

The posterior predicted quantile results for future effects, again computed using the Castellacci (2012) formula, are shown in figure 4 with the full pooling results for comparison. Any detected heterogeneity in the quantile treatment effects on household business outcomes is typically localized above the 85th percentile. Below this point, the effect is zero and reasonably generalizable, but above this point the high variation and sparsity in the tails means that there is great uncertainty about the exact impact microcredit will have on the right tail of the next distribution to which it is applied. Using pooling metrics to assess the heterogeneity in the effects specified within the tailored hierarchical PDF models across sites shows reasonable generalizability, with approximately 60% pooling on average across all metrics (see Appendix B).[30] Yet as before, the full pooling model displays unwarranted precision and magnitude of impact relative to the more moderate and uncertain prediction made by the hierarchical model.

An important reason for the uncertainty in the right tail of business outcomes is that they exhibit extreme kurtosis, that is, the tails of these variables are very heavy. The positive tail of profit, which is less heavy than that of revenues and expenditures, has an excess kurtosis of 811 in the Lognormal Model (see calculations in Appendix E). For reference, the standard Laplace distribution has an excess kurtosis of 3, yet even in that case the sample median is 2-3 times more efficient than the sample mean as an estimator of the location parameter (Koenker and Bassett 1978). The Pareto models fit to the business data find scale parameters close to zero, indicating that the tails are heavy enough to impede the functioning of the central limit theorem and even the law of large numbers (see Online Appendix). This suggests that the average treatment effects estimated via OLS regression in the original studies and thus the analysis in Meager (2018) may be unreliable for these variables, both because they invoke Gaussian asymptotics which do not hold, and because in this case the mean itself is not reliable as a summary statistic of the underlying distribution.

[30] There is again noticeable dispersion in the pooling metrics results, which suggests that these metrics should be interpreted with caution. Nevertheless there is a reasonable amount of commonality across sites, and again there is much more pooling of information on treatment effects than on control group quantile values, suggesting that these results are somewhat generalizable to other sites with different control group distributions. These results are computed separately for the two sets of treatment effects that parameterize these tailored hierarchical PDF models: the categorical logit switching effects, are shown in table 5 and the tail shape effects are shown in table 6. In each table, the same pooling metrics for the control group values of the relevant parameters are shown for comparison. For both sets of effects, there is moderate or substantial pooling on the treatment effects, but only mild to moderate pooling on the control group means.

4.2 The role of business experience

While the results of the hierarchical aggregation display less heterogeneity across the experiments than the disaggregated results suggested, understanding the remaining heterogeneity is important. There are a number of covariates both within and across sites which could predict these differences in the distributional effects of microcredit in theory. At the household level, the most relevant pre-treatment covariate is the previous business experience of the households in the sample, as measured by their operation of a business prior to the microcredit intervention. As different study populations had differing prevalence of households with these prior businesses, conditioning the analysis on this variable could help to explain the remaining heterogeneity in the causal impact of microcredit.

To assess the importance of previous business experience in modulating the causal impact of microcredit, I split the entire sample by a binary indicator of prior business ownership and separately analyze the two subsamples. Fitting the Bayesian hierarchical quantile aggregation models to each group shows that the impact of microcredit differs across the two types of households. Figures 5 and 6 show the general distributional impact of microcredit on the six household outcomes of interest for each of the household types. For most outcomes, households with no prior business ownership see negligible impact of microcredit across the entire distribution, leading to a generalizable and precise impact of zero across all quantiles, with only a small increase in the variance in the right tail. Households with prior businesses are responsible for the positive and large point estimates in the right tails, but also for the noise in that tail, suggesting that they are also the source of the heterogeneous effects. This confirms the results of Banerjee et. al. (2015b) and Meager (2018), which performed similar analyses within a single site and for the average effects respectively, and found differences in the way households with business experience respond to microcredit relative to those without such experience.

A closer examination of the results yields indirect evidence about the different ways in which these two types of households respond to increased access to microcredit. For households with business experience, there is strong evidence of a positive effect on total consumption at the 95th percentile, whereas households without experience see little impact on total consumption at any quantile (figure 5). These experienced households are largely responsible for the large point estimates and the massive uncertainty in the tails of the profit, revenues and expenditures distributions at the general level. However, examining the point estimates here with caution given the great uncertainty surrounding them, it does seem that inexperienced households

are responsible for the imprecise yet positive point estimate at the 95th percentile of consumer durables spending, while the experienced households generally do not alter their durables consumption at all (figure 5). Taken together, this suggests that some households who don't have prior businesses may generally use microcredit to change the composition of their consumption bundles; but even this smaller effect occurs only in the tail and is imprecisely estimated (figure 5).

5 Discussion

The aggregated distributional effects show no evidence that access to microcredit causes any negative shifts in the distribution of household outcomes. While moderately negative impacts are within the 95% posterior interval of the effects on the upper tails of most of the distributions, the point estimate and vast majority of the posterior mass is positive in those cases. The only variable with larger uncertainty at the lower tail is profit, but the point estimate is zero and the uncertainty is symmetric around that point. This provides reasonably strong evidence against the notion that microcredit causes substantially worse outcomes for some group of households than they would have experienced in its absence. While a lack of negative quantile effects does not imply that no household experiences any harm from microcredit, it does imply that any households who does experience harm is approximately canceled out by others who experience benefits, such that these groups are swapping ranks in the outcome distribution rather than contributing to any change in the shape of that distribution. The community as a whole however does not experience any systematic worsening of its economic outcomes.

The precise zero effect from the 5th to 75th percentile of most of the household outcomes is a true zero and not a mechanical artefact of the spike at zero nor an economic consequence of the low takeup. Consumption, consumer durables and temptation goods do not exhibit a spike of households who record an outcome equal to zero (this is almost true by definition, since it is hard to survive on nothing), yet microcredit still has a precisely estimated zero effect for most of the distribution. Even for profit, revenues and consumption the spike only accounts for at most 50% of the outcome distribution, yet the zero effect applies to 75% of the distribution. Similarly, Bosnia and the Philippines had over 90% takeup and yet still exhibited zero effects from the 5th to 75th percentile (see Appendix C). The bounding exercise in Appendix D aggregates all the data on the question of takeup and shows that even the effects on the outcome distribution for those who take up microloans are likely to be zero along most of the distribution.

I do find evidence of large positive effects of microcredit on the right tail of all outcome distributions, although these effects are imprecisely estimated and heterogeneous across contexts. Thus, the quantile analysis effectively decomposes the small and moderately noisy average treatment effect estimates from all the papers, aggregated in Meager (2018), into an imprecise yet large effect on the tail, and a precise zero everywhere else. These tail effects are large enough to be economically important and are typically concentrated among those households who have previous experience operating businesses, for whom we can rule out a zero effect on consumption at the 95th percentile, though the estimate is still quite imprecise, as shown in figure 5. Thus overall the aggregated distributional analysis provides evidence that microcredit is likely to do some good and no systemic harm. While the models are unable to precisely predict the effect on the right tail, and thus cannot confidently predict the impact in the next location into which microcredit expands, it is more likely to be positive than negative. Of course, for most of the community, it appears that no systematic change is occurring and the majority of the outcome distribution looks the same in both treatment and control.

What are the economic consequences of potentially increasing the right tail of consumption and business outcomes while leaving the rest of the distribution unchanged? This pattern means that expanding access to microcredit is likely to cause an ex-post increase in economic inequality across households, which may be important if inequality leads to capture of local political institutions or other adverse social consequences (Acemoglu and Robinson 2008). However, that increase is entirely generated by the right tail expanding rightwards: a probable improvement of economic circumstance for some, with no corresponding systematic loss for any group of households. A rightward expansion of the upper tail does not means that the richer households are getting richer, because quantile effects cannot be localised to any particular households without invoking a rank invariance assumption or some comparable structure (which is unrealistic for credit market interventions). The interpretation of the quantile effect results presented here must remain at the group level, and thus, we cannot infer which households specifically benefit from the likely expansion of the right tail. More detailed baseline data may have permitted an exploration of this question, although such households may well look identical to others along all the covariates we can measure (as suggested in Kaboski and Townsend 2011).

This pattern of probable yet variable expansion in the right tail, combined with the inability to localise the effects to particular households in these data sets, highlights the value of locating and studying these highly productive individuals. Studies

such as Hussam, Rigol and Roth 2017, which leverages local knowledge to lend to borrowers with high marginal returns to capital, are valuable both because these individuals seem to be the only households positively benefiting and because the benefits are large. My aggregated analyses largely confirms those results, yet adds the nuance that we cannot expect the results observed in such papers to replicate elsewhere, and there may well be contexts in which these positive tail effects will not materialise. However, my analysis also demonstrates the challenges of inference on these highly productive households because, almost by definition, their returns follow heavy-tailed distributions. Under such circumstances, studies that appear to be well-powered may be underpowered to detect these effects, which suggests that there are likely benefits to either powering studies to detect effects on heavy-tailed distributions or emphasising aggregated results rather than individual studies.

The heavy tails (extreme kurtosis) in the household business outcomes has both methodological and economic implications. Ordinary least squares regressions such as those performed in the original randomized controlled trials are likely to perform poorly compared to quantile regression techniques or parametric modelling of the tail (Koenker and Basset 1978, Koenker and Hallock 2011). More substantively, heavy tails suggest that in these populations, certain individual households account for large percentages of the total business activity. It may be challenging to understand the economies of developing countries if we trim or winsorize the most productive households who make up a large percentage of total economic activity. It might be more useful to study mechanisms that can produce fat-tailed outcomes, such as multiplicative production functions, experimentation or investments with a relatively high risk exposure and long maturation horizons. The fact that households with prior businesses increase their consumption (figure 5) may suggest they have some expectation of future increases in profits or earnings. This highlights the potential benefits of studying these households over longer time horizons, or perhaps taking multiple observations of the same households as in the Townsend Thai Data (2018) and as suggested in McKenzie (2012).

My analysis is not exhaustive, and the conclusions I can draw are limited by the constraints of my framework and of the original studies. It may be that if microcredit interventions were studied over a 10 or 20 year horizon, the imprecise tail effects we observe after two years could either become precise or could lead to benefits across the entire distribution. If the studies had a richer set of baseline data, a deeper understanding of the household-level distributional impacts of expanding access to microcredit could be generated by including baseline covariates and perhaps leveraging more economics knowledge of the contextual microstructure to the

analysis. It would be informative to apply an individual-level structural model to this data, such that one could infer the distribution of individual-level treatment effects, but there is currently no established methodology for aggregating structural parameters. Finally, by restricting the selected set of studies to be RCTs, there is a possibility of a sample selection bias due to the conditions required to perform field experiments; as yet, there is no established method for combining experimental and observational studies in a single aggregation framework. Despite these cautions, the conclusion of the current analysis remains salient. In general there is likely to be no difference between the treatment and control groups below the 75th quantile in future sites that receive more access to microcredit, and while we cannot reliably predict the effect above the 75th percentile, the aggregated evidence suggests it is likely to be positive.

6 Conclusion

The microcredit results demonstrate the value of analysing and aggregating evidence using appropriate methodology rather than a "default" approach restricting oneself to conditional means. The models developed in this paper could be used to study the distributional effects of other financial interventions, trade and innovation policies, educational subsidies, and local migration incentives, all of which have social welfare implications (Borusyak and Jaravel 2018, Duflo, Dupas and Kremer 2017, Chetty, Hendren, and Katz 2016, Bryan, Chowdhury and Mobarak 2014). There are many settings in which quantiles are implicated in policy directly, often because taxation and welfare policies are explicitly made with reference to quantiles of the income distribution; in such settings, the conditional mean often contains little information about the policy's impacts (Bitler, Gelbach and Hoynes 2006, Ramnath and Tong 2017). Quantile regression also allows detection of changes in total inequality in the distribution of outcomes even when this is not predicted by any of the covariates we can collect, which is increasingly important given the potentially harmful consequences of social and economic inequality in modern economies. In addition, the approaches I provide here can accommodate the heavy tails found in many economic data sets (Bazzi 2016, Pancost 2016, Gabaix 2008, Fama 1965). Multi-study aggregation of quantile effects within the Bayesian hierarchical framework can thus deliver inference that is both more informative and more reliable than analyses of average treatment effects alone.

Table 1: Simulation Results: Coverage of Limited-Information Quantile Model Posterior Inference under Cross-Site Variation (CSV) and Data Variation (DV)

Features	β_1: 50%	95%	β_0: 50%	95%
Little CSV, Little DV	0.524	0.976	0.560	0.952
Little CSV, Moderate DV	0.644	0.980	0.480	0.936
Large CSV, Moderate DV	0.532	0.956	0.448	0.928
Large CSV, Large DV	0.568	0.992	0.512	0.944
Moderate CSV, Large DV	0.632	0.988	0.480	0.940
Little CSV, Large DV	0.596	0.996	0.548	0.984
Moderate CSV, Little DV	0.512	0.952	0.488	0.912
Very Large CSV, Large DV	0.476	0.928	0.266	0.688
Approx No Pooling, Moderate DV	0.480	0.944	0.434	0.870
Approx Full Pooling, Moderate DV	0.668	0.988	0.668	0.996
Approx Full Pooling, Very Large DV	0.758	0.998	0.576	0.972
	Σ_1 (off diag): 50%	95%	Σ_1 (diag) :50%	95% (diag)
Little CSV, Little DV	0.903	1	0.880	0.996
Little CSV, Moderate DV	0.932	1	0.884	0.996
Large CSV, Moderate DV	0.845	1	0.616	0.996
Large CSV, Large DV	0.933	1	0.868	1
Moderate CSV, Large DV	0.887	1	0.810	1
Little CSV, Large DV	0.927	1	0.820	1
Moderate CSV, Little DV	0.807	1	0.520	0.992
Very Large CSV, Large DV	0.336	1	0.850	0.998
Approx No Pooling, Moderate DV	0.593	0.998	0.454	0.926
Approx Full Pooling, Moderate DV	1	1	0.420	0.996
Approx Full Pooling, Very Large DV	1	1	0.026	0.998
	Σ_0 (off diag): 50%	95%	Σ_0 (diag): 50%	95%
Little CSV, Little DV	0.857	1	0.476	0.996
Little CSV, Moderate DV	0.878	1	0.444	1
Large CSV, Moderate DV	0.255	1	0.304	0.984
Large CSV, Large DV	0.083	0.999	0.196	0.992
Moderate CSV, Large DV	0.101	1	0.198	0.994
Little CSV, Large DV	0.675	1	0.228	1
Moderate CSV, Little DV	0.800	1	0.420	0.980
Very Large CSV, Large DV	0.497	0.994	0.426	0.944
Approx No Pooling, Moderate DV	0.560	0.995	0.362	0.866
Approx Full Pooling, Moderate DV	1	1	0.684	0.996
Approx Full Pooling, Very Large DV	1	1	0.052	1

Notes: CSV is Cross-Site Variation, DV is Data Variation. Simulation runs kept small due to relatively long runtime for model fit, but results are relatively stable within run sets. For this exercise, the CSV in β_{0k} space is typically larger than that in β_{1k} space, to reflect the likely reality of comparing quite different places with plausibly similar effects.
[Back to main]

Table 2: Consumption: Comparison Of No Pooling, Partial Pooling and Full Pooling Results

Quantile:	5th	15th	25th	35th	45th	55th	65th	75th	85th	95th
No Pooling										
Bosnia	-2	11	5.1	-2.5	-4.7	-34.4	-2.3	4.6	4.1	0.4
	(-6,2.1)	(3.2,18.7)	(-10.5,20.7)	(-41.6,36.6)	(-17.1,7.6)	(-76.4,7.6)	(-7.8,3.2)	(-6.5,15.8)	(-2.2,10.4)	(-12.7,13.6)
India	-2	13.2	8	-12.8	-7.7	-64.5	-1.2	8.2	-1.1	-6.4
	(-7,3)	(2.1,24.4)	(-12.1,28.1)	(-69.3,43.7)	(-22.3,6.8)	(-135.5,6.5)	(-7.3,4.8)	(-8.2,24.7)	(-7,4.7)	(-23.5,10.7)
Mexico	-9.3	4.1	12.3	0.8	-5.2	0.9	0.2	-2.6	1.6	3.6
	(-13.9,-4.8)	(-1.3,9.5)	(-7.6,32.1)	(-18.6,20.2)	(-9.8,-0.5)	(-22.5,24.2)	(-6.3,6.7)	(-9.6,4.3)	(-5.4,8.5)	(-5.3,12.5)
Mongolia	-1.5	5.5	6.5	-0.9	-7.1	-16.3	-1	2.2	5.3	3.7
	(-5.8,2.7)	(-0.8,11.7)	(-9.4,22.4)	(-30.8,29.1)	(-17.3,3.1)	(-46.1,13.5)	(-5.9,3.9)	(-6,10.4)	(-0.5,11.2)	(-6.7,14)
Morocco	-0.5	16.6	-8.2	87.4	4.1	104	-1.4	40.1	-2.6	-54
	(-6.1,5.1)	(-5.4,38.7)	(-35.6,19.1)	(-27.2,202)	(-11.6,19.9)	(-30.2,238.3)	(-8,5.3)	(-6.7,86.9)	(-9.8,4.6)	(-96.3,-11.8)
Partial Pooling										
Bosnia	-1	7.7	-0.3	5.7	-0.3	5.3	-1	6.2	0.7	5.2
	(-4,2)	(1.7,13.9)	(-5.8,5.9)	(-6,15.5)	(-5.6,5.1)	(-8.8,15.8)	(-4.7,2.5)	(-0.9,13.1)	(-3.2,5.9)	(-3.2,12.4)
India	-1.5	8.3	1.6	3	-3	-1.3	-0.3	7.4	-2.2	3.1
	(-5.1,2.3)	(-0.7,17.9)	(-4.6,11.4)	(-22.9,22.1)	(-10.4,2.6)	(-37.8,17.7)	(-4.2,3.8)	(-3.4,18.8)	(-6.5,1.9)	(-10.9,14.8)
Mexico	-7.3	3.2	2.4	3	-3.5	4.4	-0.3	0.2	-1.1	3.4
	(-12,-2.5)	(-0.9,7.3)	(-7.4,15.9)	(-4,10.5)	(-9,0.8)	(-2.8,14.4)	(-5.5,5.6)	(-5.2,5.1)	(-6.6,5)	(-1.7,9)
Mongolia	-0.3	3.9	-0.1	4.2	-1.5	3.6	-0.3	4.2	1.7	4.6
	(-3.5,2.9)	(-0.7,8.7)	(-7,6.6)	(-3.9,12.2)	(-7.8,3.2)	(-5.6,11.1)	(-3.7,3.2)	(-0.8,9.4)	(-2.2,6.6)	(-0.8,10.4)
Morocco	-0.7	11.4	-5	76.6	4.7	89.9	0.2	36.7	-2.8	-31.9
	(-4.8,3.5)	(-10.2,33.3)	(-18.9,4.5)	(7.7,152.6)	(-2.8,13.8)	(-6.3,215.9)	(-4.6,4.9)	(-1.4,76.9)	(-7.9,1.9)	(-70.9,7.3)
Average	-2	0	-0.4	-1.1	-0.8	2.8	4.1	6	4.1	19.4
	(-9.7,7.2)	(-4.6,4.5)	(-4.4,4.1)	(-6.6,4.8)	(-9.1,6.8)	(-3,9.1)	(-1.8,9.9)	(-2.3,13.5)	(-14.2,16.8)	(-26.1,62.4)
Full Pooling										
Average	-3.9	0.2	-0.9	-1.8	-1.3	2.5	3.6	6.1	6.4	13.9
	(-7.6,-0.1)	(-2.6,3.1)	(-3.7,1.9)	(-4.8,1.2)	(-5,2.4)	(-1.4,6.3)	(-0.8,7.9)	(0.1,12.1)	(-2,14.7)	(-4.8,32.6)

Notes: All units are USD PPP per two weeks. Estimates are shown with their 95% uncertainty intervals below them in brackets. In this case the full pooling and no pooling models are frequentist. [Back to main]

Table 3: Profit: Comparison Of No Pooling, Partial Pooling and Full Pooling Results

Quantile:	5th	15th	25th	35th	45th	55th	65th	75th	85th	95th
No Pooling										
Bosnia	0	0	0	0	0	0	0	122.7	117.1	148.9
	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(57.5,164.1)	(30.1,189.1)	(-53.7,322.2)
Ethiopia	-5.8	-1.8	-0.5	0	0.2	0.5	0.9	1.7	3.4	10.1
	(-11.7,0.7)	(-3.2,-0.4)	(-1.1,0.1)	(-0.7,0.7)	(-0.5,1)	(-0.4,1.5)	(-0.3,2.4)	(-0.3,3.9)	(-0.5,7.4)	(-2.6,21.5)
India	0	0	0	0	0	0	0	0	-6.7	7.9
	(-5.2,1.7)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(-17.6,6.1)	(-36.5,59.3)
Mexico	0	0	0	0	0	0	0	0	0	10.2
	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(-4.3,26.3)
Mongolia	-1.8	-0.8	0	0	0	0	0	0	0	-0.2
	(-5.1,2)	(-2.2,0.6)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(-1.3,0.6)
Morocco	-30.3	1.5	2.2	0	0	3.8	7.8	14.5	28.6	78.5
	(-70.4,12.2)	(-6.4,8.7)	(-1,4.5)	(0,0)	(0,2.7)	(0.2,8.1)	(1.9,14.7)	(4,26.7)	(4.7,54.8)	(-20,183.5)
Philippines	0	-27.8	-13.6	-10.3	-6.3	0.2	10.8	30.6	72.8	222
	(0,0)	(-49.7,6.2)	(-39.5,6.5)	(-42.1,13.8)	(-46.1,24.3)	(-51.1,40.3)	(-58.8,66.5)	(-72.7,113.3)	(-99.2,216.7)	(-206.1,614.1)
Partial Pooling										
Bosnia	0	0	0	0	0	0	0	89.5	79.5	129.2
	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(15.3,144.7)	(22.7,146.3)	(15.3,249.8)
Ethiopia	-3.7	-1.2	-0.4	0	0.2	0.4	0.8	1.5	2.9	8.4
	(-9.3,1.5)	(-2.6,0)	(-0.9,0.1)	(-0.7,0.7)	(-0.5,0.9)	(-0.4,1.3)	(-0.3,2)	(-0.2,3.3)	(-0.2,6.3)	(-1.5,18.6)
India	0	0	0	0	0	0	0	0	-0.1	25.3
	(-3.9,1.3)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(-11.4,10.4)	(-19,64.5)
Mexico	0	0	0	0	0	0	0	0	0	12.2
	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(-1.8,25.7)
Mongolia	-1	-0.6	0	0	0	0	0	0	0	0
	(-3.6,1.7)	(-1.8,0.6)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(-0.7,0.6)
Morocco	-35.8	-0.6	1.2	0	0	3	5.9	11.3	23.6	74.6
	(-69.4,-3.2)	(-7,6.1)	(-1.6,4.1)	(0,0)	(-0.3,2.6)	(-0.7,6.9)	(0.2,12.2)	(1.7,22.6)	(2.5,48)	(-17.5,168.4)
Philippines	0	-4.9	-0.7	2.6	7.1	13.7	24.1	41.5	77	201
	(0,0)	(-33.6,18.7)	(-19.1,14.6)	(-18.2,20.1)	(-17.8,28.4)	(-18.1,40.9)	(-18.5,61.5)	(-20.9,98.1)	(-31.4,180.5)	(-80.2,491.6)
Average	0	0	0	0	0	0	0	2.8	6.9	20.8
	(-49.6,4.4)	(-9.2,0.2)	(-2.9,0.5)	(-1.2,3.1)	(-1.1,7)	(-2,12.5)	(-3,20)	(-4.5,31.5)	(-4.6,57.2)	(-5.5,165.1)
Full Pooling										
Average	2.4	0	0	0	0	0	0	5.6	21	106
	(-2.6,7.4)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(3.5,7.7)	(15.8,26.3)	(80.1,132.6)

Notes: All units are USD PPP per two weeks. Estimates are shown with their 95% uncertainty intervals below them in brackets. In this case all intervals are Bayesian but the No Pooling and Full Pooling models replicate in a frequentist framework. Because these posteriors are fat tailed, the estimate provided is the posterior median value. [Back to main]

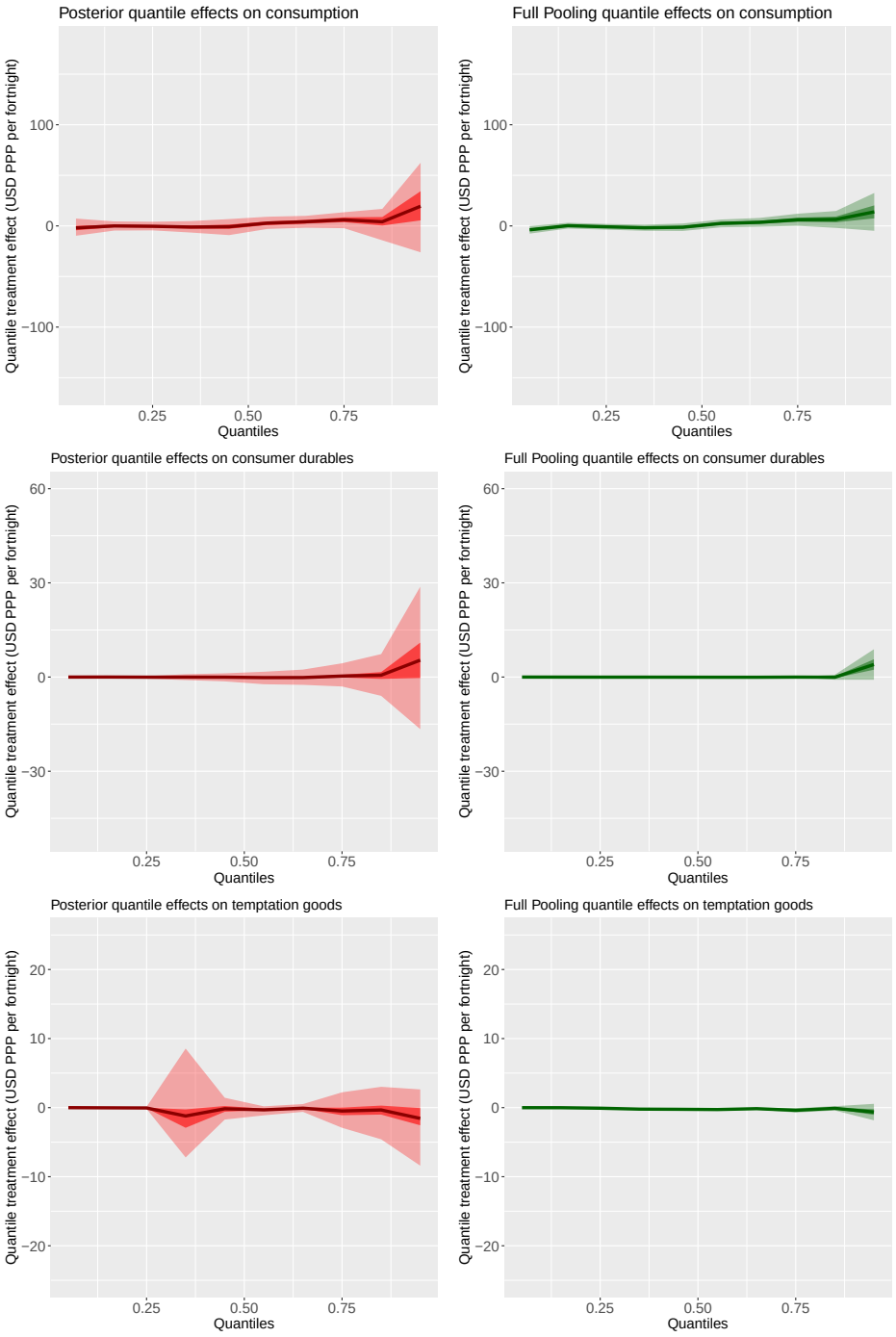

Figure 1: General Quantile Treatment Effect Curves (β_1) for consumption-type variables. The dark line is the posterior mean, the opaque color bands are the central 50% posterior uncertainty interval, the translucent color bands are the central 95% posterior uncertainty interval. [Back to main]

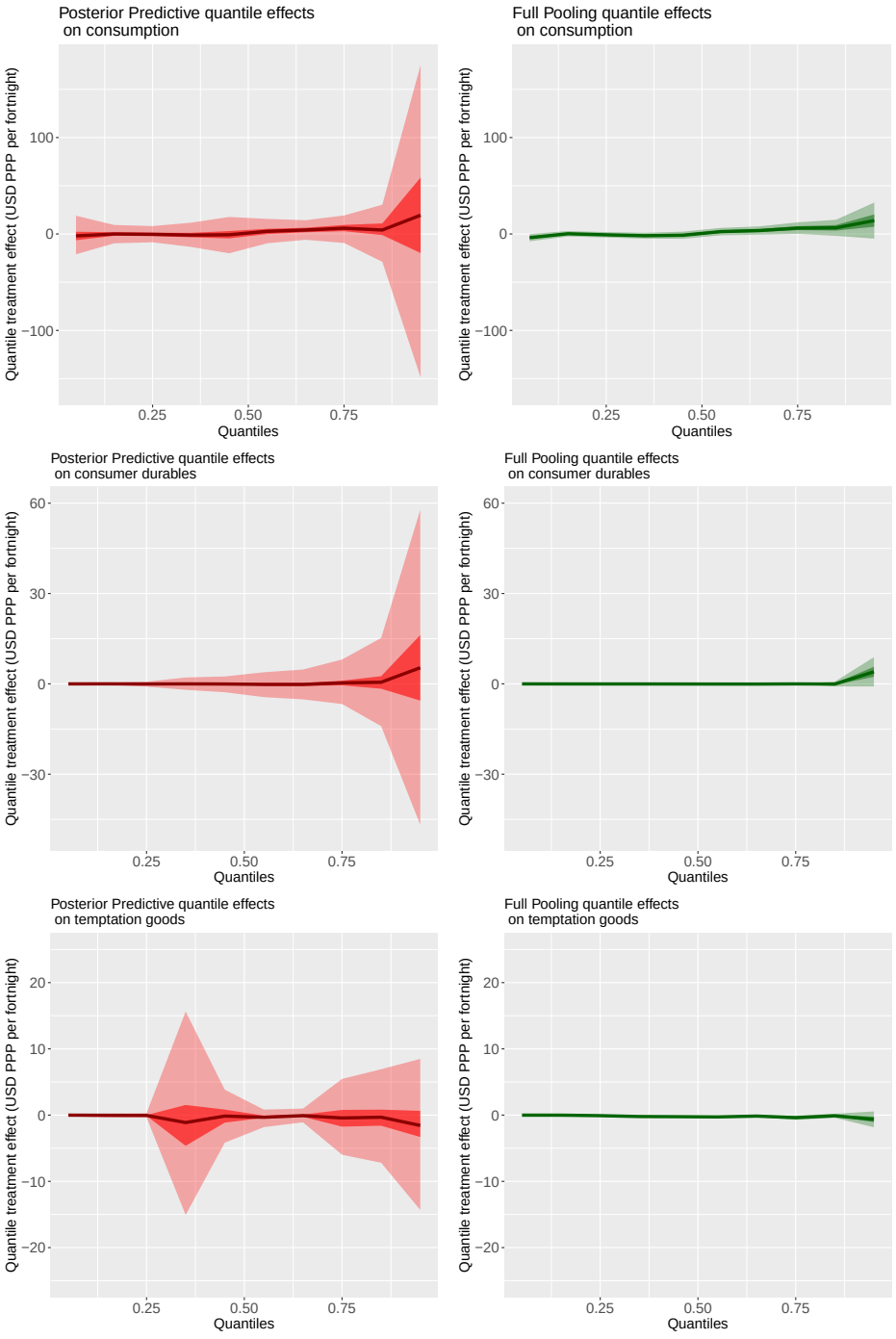

Figure 2: Posterior Predictive Quantile Effect Curves ($\beta_{1,K+1}$) for consumption-type variables.The dark line is the posterior mean, the opaque color bands are the central 50% posterior predictive uncertainty interval, the translucent color bands are the central 95% posterior predictive uncertainty interval. [Back to main]

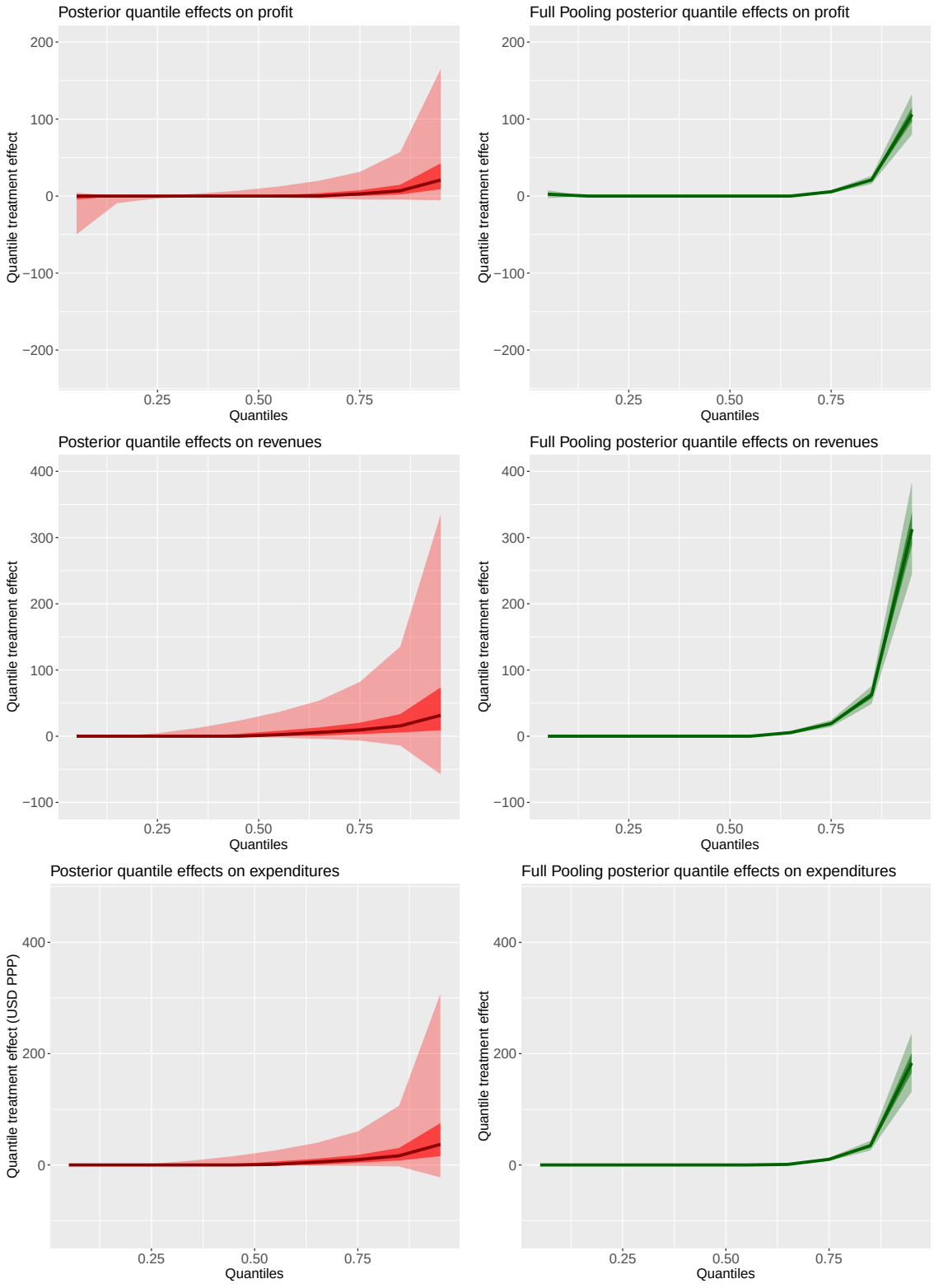

Figure 3: General Quantile Treatment Effect Curves (β_1) for business variables from the LogNormal model. The dark line is the median posterior draw, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. [Back to main]

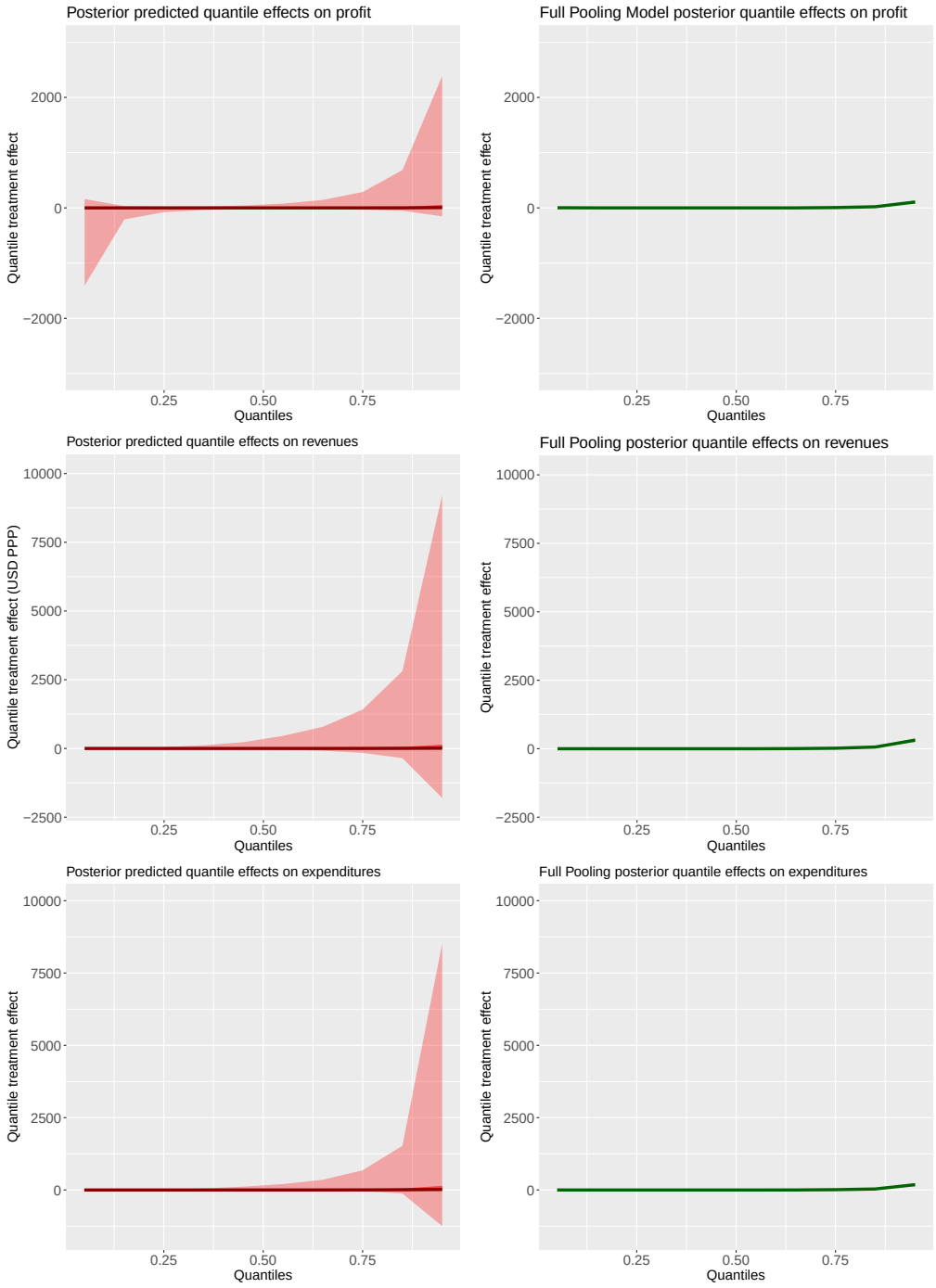

Figure 4: Posterior predicted quantile treatment effect curves for Business Variables from the LogNormal model. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. [Back to main]

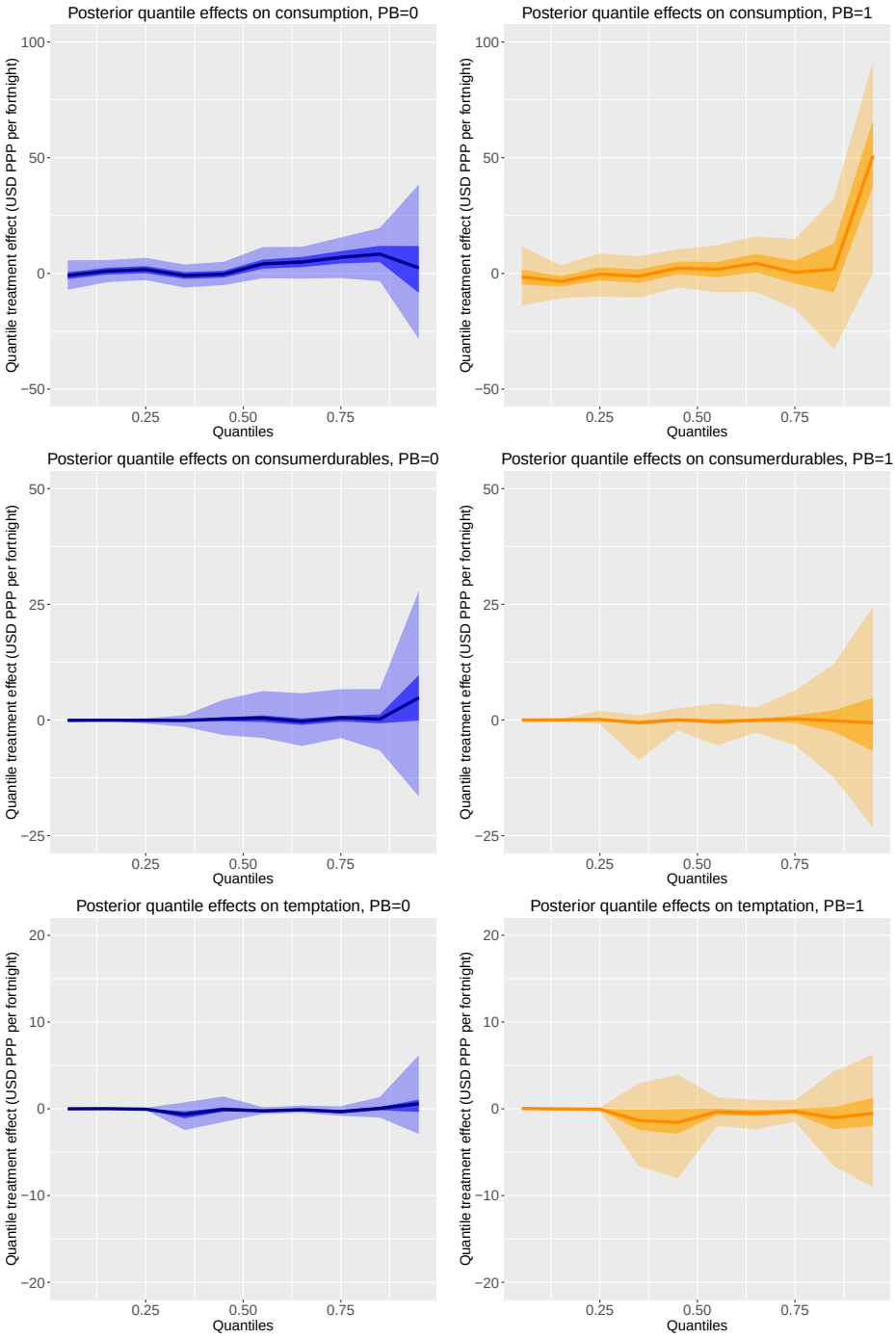

Figure 5: General Quantile Treatment Effect Curves split by prior business ownership (β_1) for consumption-type variables. The dark line is the posterior mean, the opaque color bands are the central 50% posterior uncertainty interval, the translucent color bands are the central 95% posterior uncertainty interval. [Back to main]

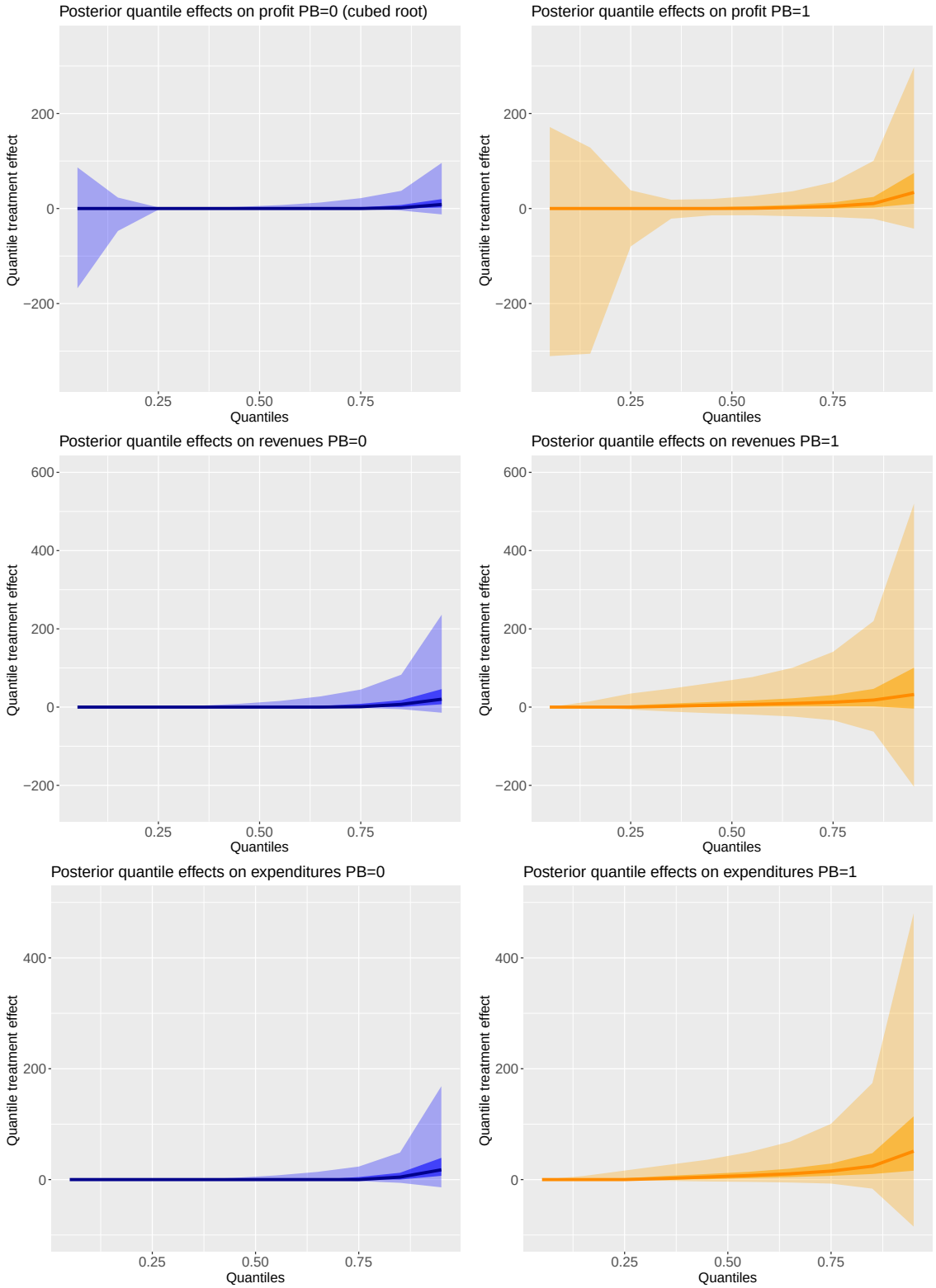

Figure 6: General Quantile Treatment Effect Curves (β_1) for business variables split by prior business ownership. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in cubed root of USD PPP due to the scale differences in the uncertainty at the right tail versus the rest of the distribution. [Back to main]

Appendix A Data and Dithering

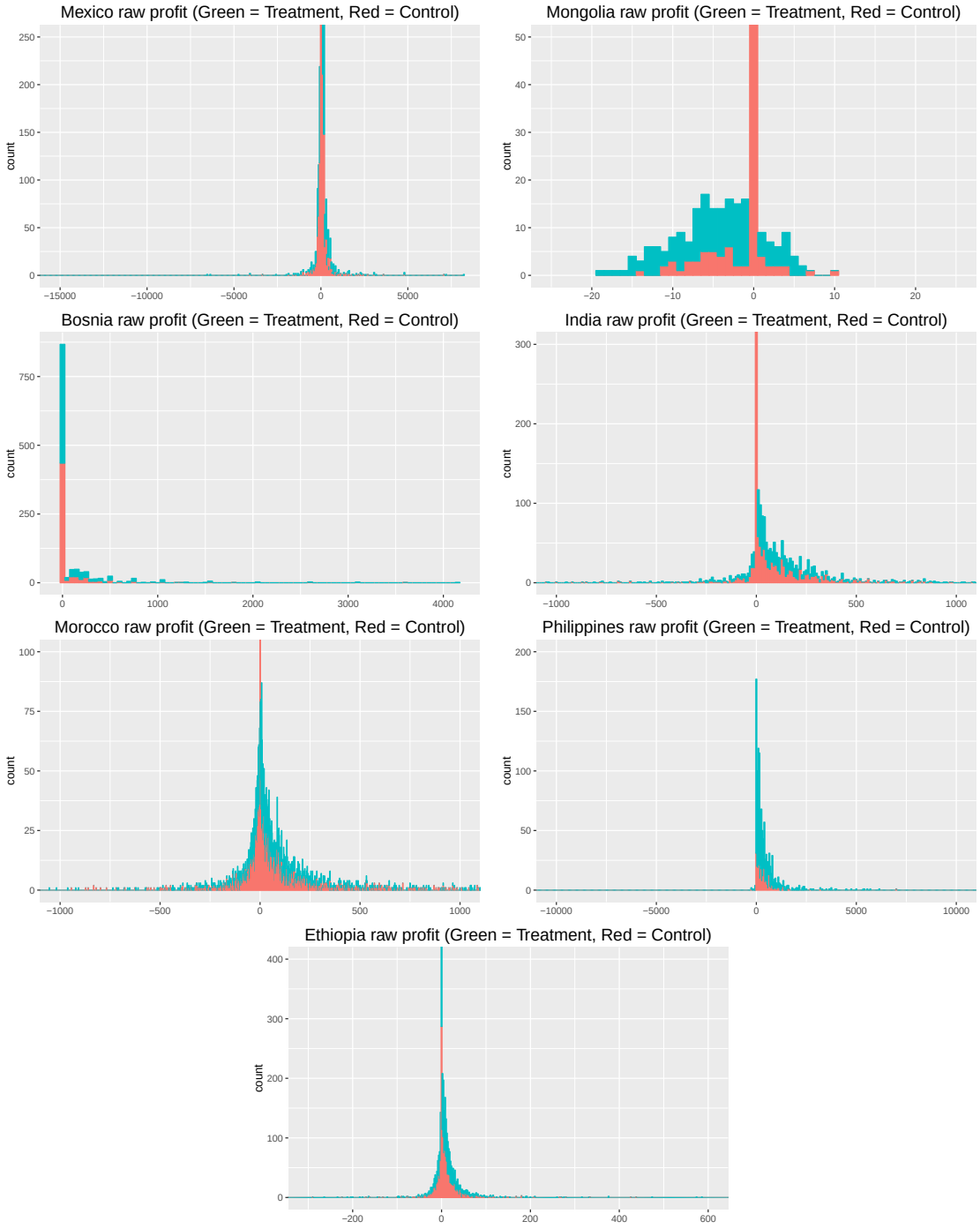

Figure 7: Histograms of the profit data in each site, in USD PPP per 2 weeks. Display truncated both vertically and horizontally in most cases. [Back to main]

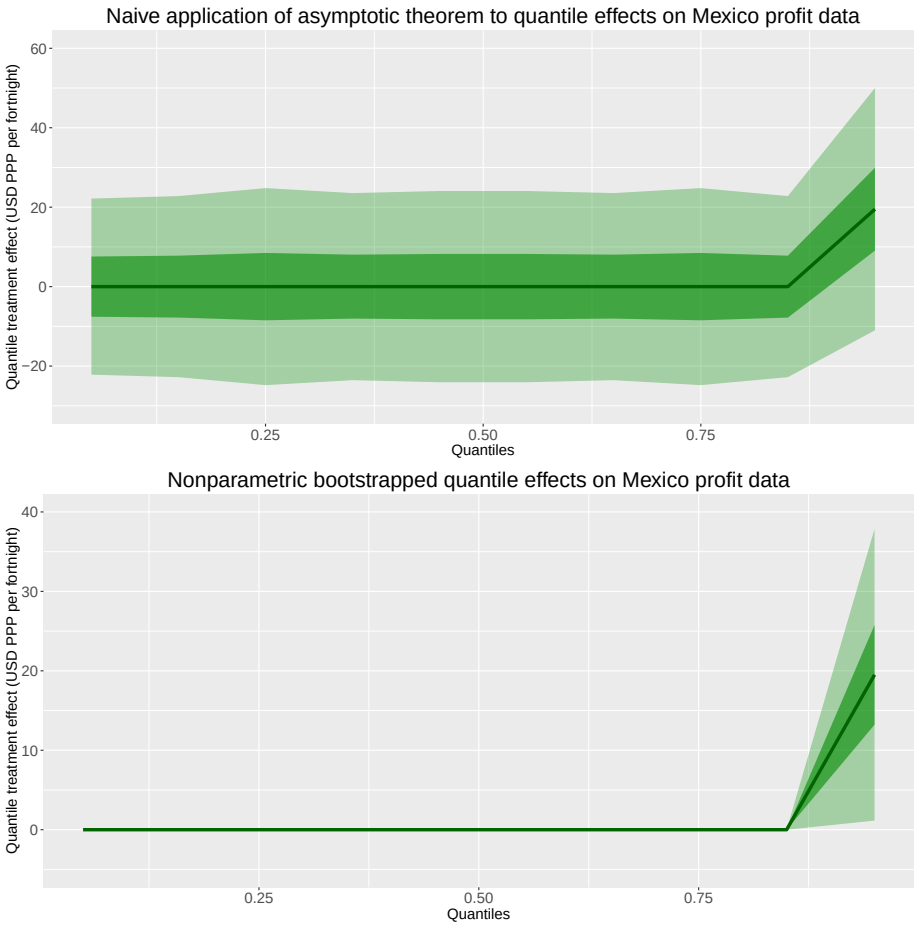

Figure 8: Quantile TEs for the Mexico profit data, Mosteller theorem approximation standard errors (above) and nonparametrically bootstrapped standard errors (below). The green line is the estimated effect, the opaque bands are the central 50% interval, the translucent bands are the central 95% interval. The output of these estimators should be similar if the Mosteller (1946) theorem holds, but it is not similar because profit is not in fact continuously distributed. This is due to a discrete probability mass at zero, reflecting numerous households who do not operate businesses. [Back to main]

Dithering is often an effective strategy for partially discrete data: In fact, a small amount of dithering is necessary for the microcredit data on consumer durables spending and temptation goods spending to conform to the Mosteller approximation, as this data is actually somewhat discrete. However, in the microcredit business data, the complications caused by these spikes at zero are not effectively addressed by dithering. The results in figure 9 show that applying the Mosteller theorem to the dithered profit data leads to inference that is too precise in the tail relative to the results of the bootstrap on the same data.

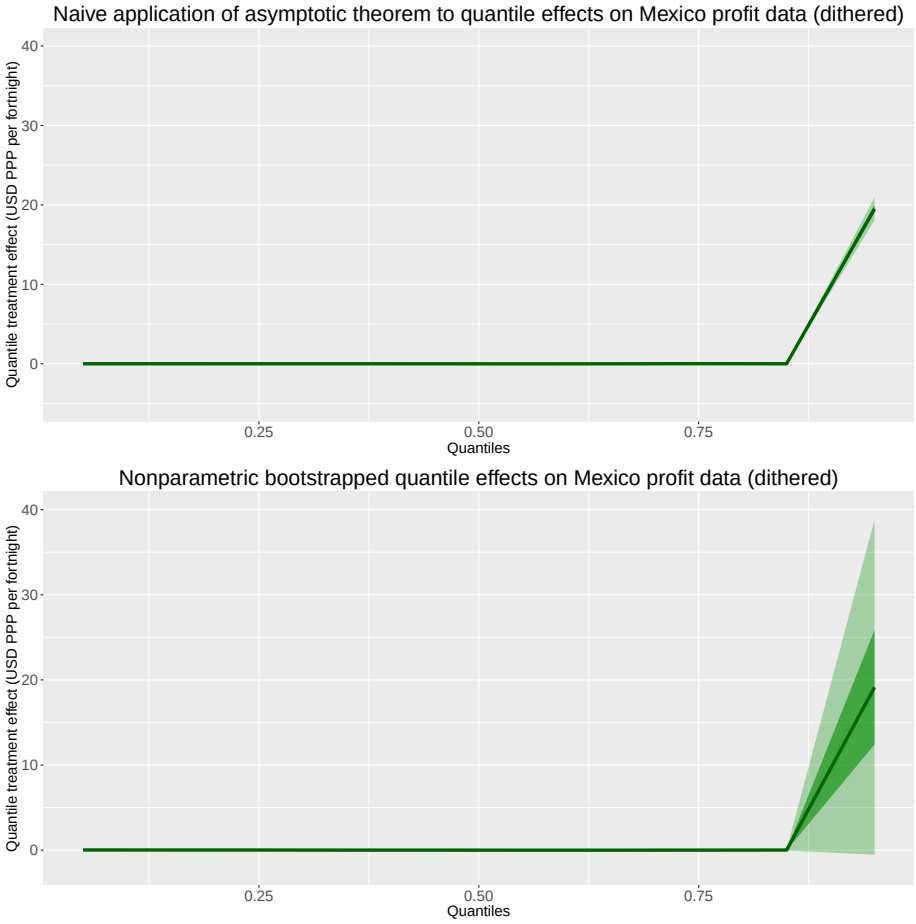

Figure 9: Quantile TEs for the dithered Mexico profit data, Mosteller theorem approximation standard errors (above) and nonparametrically bootstrapped standard errors (below). The green line is the estimated effect, the opaque bands are the central 50% interval, the translucent bands are the central 95% interval. Dithering is a simple strategy which can overcome problems associated with quantile regression on discrete distributions, recommended in Machado & Santos Silva (2005) and Koenker (2011). It has failed in this case. [Back to main]

Appendix B Pooling Metrics

B.1 Pooling Metrics for Hierarchical Models

The hierarchical framework also provides several natural metrics to assess the extent of pooling across sites shown in the posterior distribution (Gelman et al. 2004, Gelman and Pardoe 2006). In the context of multi-study aggregation under the assumption of exchangeability, the extent of pooling across study sites has a natural interpretation as a measure of generalizability. The magnitude of Σ_θ and the magnitude of the uncertainty interval on the predicted effect in the next site θ_{K+1} is of utmost interest. Yet the drawback of using $|\tilde{\Sigma}_\theta|$ as a pooling metric is that it may be unclear what constitutes a large or small magnitude in any given context. Thus, other pooling metrics have been developed for the univariate case, where θ is a scalar and thus Σ_θ is a scalar, denoted σ_θ^2. As I extend these metrics to apply to the multivariate distributional effects typically computed by economists, a general overview of their scalar counterparts is given here.

The most prevalent metric in the literature is the conventional "pooling factor" metric, defined as follows (Gelman and Hill 2007):

$$\omega(\theta_k) \equiv \frac{\hat{se}_k^2}{\tilde{\sigma}_\theta^2 + \hat{se}_k^2}. \tag{B.1}$$

This metric has support on [0,1] because it decomposes the potential variation in the estimate in site k into genuine underlying heterogeneity and sampling error. It compares the magnitude of $\tilde{\sigma}_\theta^2$ to the magnitude of $\hat{se}_k^2$, the sampling variation in the no-pooling estimate of the treatment effect from site k. Here, $\omega(\theta_k)$ close to 1 indicates that $\tilde{\sigma}_\theta^2$ is smaller than the sampling variation, indicating substantial pooling of information and a "small" $\tilde{\sigma}_\theta^2$ (and the reverse if the metric is close to 0). If the average of these K pooling metrics across sites is above 0.5, the genuine underlying heterogeneity is smaller than the average sampling variance.

The fact that the $\omega(\theta_k)$ uses sampling variation as a comparison is inherently Bayesian, and makes a statement about information in the data rather than a statement about the world. Thus it may be beneficial to use an alternative pooling metric. Meager (2018) proposed the use of the following metric based on relative geometric proximity, defined as follows:

$$\breve{\omega}(\theta_k) \equiv \{\omega : \ \tilde{\theta}_k = \omega\tilde{\theta} + (1-\omega)\hat{\theta}_k\}. \tag{B.2}$$

This metric scores how closely aligned the posterior mean of the treatment effect in site k, denoted $\tilde{\theta}_k$, is to the posterior mean of the general effect $\tilde{\theta}$ versus the separated no-pooling estimate $\hat{\theta}_k$. Here, $\breve{\omega}(\theta_k) > 0.5$ indicates that the generalized treatment effect is actually more informative about the effect in site k than the separated estimate from site k is for site k (since $\tilde{\theta}_k$ is our best estimate of θ_k). This $\breve{\omega}(\theta_k)$ is the "brute force" version of the conventional pooling metric because it is identical in models which partially pool on only one parameter, but may differ in models that pool across multiple parameters. I truncate this metric to lie on $[0, 1]$ to

preserve comparable scales across metrics, as the occasions on which it falls outside this range are due to shrinkage on other parameters.

Another pooling metric that can be computed for these models is the "generalized pooling factor" defined in Gelman and Pardoe (2006), which takes a different approach using posterior variation in the deviations of each θ_k from θ. Let $E_{post}[.]$ denote the expectation taken with respect to the full posterior distribution, and define $\epsilon_k = \theta_k - \theta$. Then the generalized pooling factor for θ is defined:

$$\lambda_\theta \equiv 1 - \frac{\frac{1}{K-1}\sum_{k=1}^{K}(E_{post}[\epsilon_k] - \overline{E_{post}[\epsilon_k]})^2}{E_{post}[\frac{1}{K-1}\sum_{k=1}^{K}(\epsilon_k - \bar{\epsilon_k})^2]}. \tag{B.3}$$

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 very little pooling, as 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. Gelman and Pardoe (2006) interpret $\lambda_\theta > 0.5$ as indicating a higher degree of general or "population-level" information relative to the degree of site-specific information.

[Back to main]

B.1.1 Pooling Metrics for Nonparametric Quantile Treatment Effects

Conventional pooling metrics for hierarchical models are designed to be applied to univariate treatment effects. For the multivariate Normal quantile curve aggregation models, the object that governs the dispersion of β_{1k} around β_1 is the parent variance-covariance matrix Σ_1. The raw size of this matrix is the purest metric of that dispersion, but this can only be measured in terms of a certain matrix norm, and different norms will give different answers. I proceed using a statistical argument to determine the appropriate norm.[31] Consider the idiosyncratic k-specific components $\xi_k = \beta_{1k} - \beta_1$, so that $\xi_k \sim \mathcal{N}(0, \Sigma_1)$. The question of how much heterogeneity there is in the set $\{\beta_{1k}\}_{k=1}^K$ is isomorphic to the question of how far away from 0 is the typical draw of ξ_k. The answer turns out to be defined by the trace of Σ_1, or the Frobenius norm of $\Sigma_1^{1/2}$.

To see why the trace of Σ_1 is a sensible metric for the average magnitude of ξ_k, consider the transformed variable $z_k \equiv \Sigma_1^{-1/2}\xi_k \sim \mathcal{N}(0, I)$. Then, considering the variance of ξ_k, we have $\|\xi_k\|^2 = \left\|\Sigma_1^{-1/2} z_k\right\|^2 = z_k'\Sigma_1 z_k$. Thus, we can get the expected squared distance of ξ_k from 0 by computing $E[z_k'\Sigma_1 z_k]$. Since z_k follows a standard multivariate Normal, this expectation is simply the trace of Σ_1. To see this another way, recall that in a finite dimensional Euclidean space, taking *any* orthonormal basis e, we have $\text{tr}(A) = \sum_{i=1}^n \langle Ae_i, e_i\rangle$. Thus, the trace of Σ_1 determines how far away we push any orthonormal basis vector away from itself by premultiplying by Σ_1, and this defines a notion of dispersion in the space spanned by e. In addition,

[31] I thank Tetsuya Kaji for his conceptualization of this approach and his major contribution to this argument.

because $\text{tr}(\Sigma_1)$ is equivalent to the Frobenius norm of $\Sigma_1^{1/2}$, it is submultiplicative and unitarily invariant.

Defining $\text{tr}(\Sigma_1)$ as the preferred metric allows the natural extension of the univariate pooling metrics to the multivariate Normal objects in the hierarchical likelihood. Recalling that the model implies $\hat{\beta}_{1k} \sim \mathcal{N}(\beta_1, \hat{\Xi}_{\beta_1 k} + \Sigma_1)$, we can compute the percentage of total variation of the no-pooling quantile treatment effect curve estimates around their true mean β that is due to sampling variation from $\hat{\Xi}_{\beta_1 k}$. Hence, I construct a matrix-valued version of the conventional pooling metric as follows:

$$\begin{aligned} \boldsymbol{\omega}(\beta) &= \frac{1}{K} \sum_{k=1}^{K} \frac{\text{tr}(\hat{\Xi}_{\beta_1 k})}{\text{tr}(\hat{\Xi}_{\beta_1 k} + \Sigma)} \\ &= \frac{1}{K} \sum_{k=1}^{K} \frac{\text{tr}(\hat{\Xi}_{\beta_1 k})}{\text{tr}(\hat{\Xi}_{\beta_1 k}) + \text{tr}(\Sigma)} \end{aligned} \tag{B.4}$$

The suitability of the trace operator here suggests a general method for constructing pooling factors on multivariate treatment effects. Consider the Gelman and Pardoe (2006) pooling metric which, for univariate treatment effects, compares within-variation in the posterior draws of each β_{1k} to the between variation in the posterior draws of $\{\beta_{1k}\}_{k=1}^{K}$. One can take the sum of this metric evaluated at each quantile treatment effect as the trace did for the conventional pooling metric. Here the sum must be normalized to ensure the result lies on the interval [0,1]. Defining $|\mathcal{U}| = U$ and using $\beta[u]$ to refer to the uth entry in the vector of effects, I define the multivariate analogue of the Gelman & Pardoe (2006) metric for a U-dimensional treatment effect as follows:

$$\boldsymbol{\lambda}_{\boldsymbol{\beta_1}} = \frac{1}{K} \sum_{k=1}^{K} \left(1 - \frac{1}{U} \sum_{u=1}^{U} \frac{var(E[\beta_{1k}[u] - \beta_1[u]])}{E[(var(\beta_{1k}[u] - \beta_1[u]))]} \right). \tag{B.5}$$

I define the multivariate analogue of the "brute force" pooling metric defined in Meager (2015) for a U-dimensional treatment effect as follows, using $\beta[u]$ to refer to the uth entry in the vector of effects:

$$\breve{\boldsymbol{\omega}}(\boldsymbol{\beta_1}) = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{U} \sum_{u=1}^{U} \frac{\beta[u]_{1k} - \hat{\beta}_{1k}[u]}{\beta_1[u] - \beta_{1k}[u]} \right). \tag{B.6}$$

[Back to main]

B.1.2 Pooling Metrics for Parametric Quantile Treatment Effects

In tailored hierarchical PDF models, the upper level variance-covariance matrix V is the object that governs the dispersion of the treatment effects and thus the heterogeneity. The raw size of this matrix is one metric of that dispersion, and as discussed above, the trace of the matrix captures a notion of dispersion on the set of $\{\theta_k\}_{k=1}^{K}$. However, it is unclear in this setting what we should compare against $\|V\|$ because modelling the outcomes explicitly means we do not have recourse to a

sampling variance-covariance matrix within the model itself. In order to construct a sampling variance-covariance matrix, I fit a no-pooling version of the tailored PDF model, omitting the upper level of the hierarchy. I use the set of no pooling model parameters $\{\hat{\zeta}_k\}_{k=1}^K$ and their accompanying posterior variance-covariance matrix $\hat{\Sigma}_\zeta$ to construct the pooling metrics of interest. Hence, the translation of the conventional pooling metric in this case is

$$\begin{aligned} \boldsymbol{\omega}_V(\beta) &= \frac{1}{K}\sum_{k=1}^{K}\frac{\text{tr}(\hat{\Sigma}_{\zeta k})}{\text{tr}(\hat{\Sigma}_{\zeta k}+V)} \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{\text{tr}(\hat{\Sigma}_{\zeta k})}{\text{tr}(\hat{\Sigma}_{\zeta k})+\text{tr}(V)}. \end{aligned} \tag{B.7}$$

In this paper, the matrix V has been constrained to be diagonal for tractability purposes, so I construct a comparably diagonal $\hat{\Sigma}_{\zeta k}$ from each site using the marginal posteriors for each component. The Gelman and Pardoe pooling metric and the brute force pooling metric are extended to the tailored hierarchical PDF as in the multivariate Normal model case.

B.2 Results

Table 4: Pooling Factors for Nonparametric Quantile Models on Consumption

Outcome	Treatment Effects			Control Group Means		
	$\omega(\beta_1)$	$\breve{\omega}(\beta_1)$	$\lambda(\beta_1)$	$\omega(\beta_0)$	$\breve{\omega}(\beta_0)$	$\lambda(\beta_0)$
Consumption	0.252	0.730	0.703	0.004	0.298	0.049
Consumer Durables	0.276	0.658	0.930	0.053	0.532	0.013
Temptation Goods	0.284	0.552	0.589	0.017	0.495	0.004

Notes: All pooling factors have support on [0,1], with 0 indicating no pooling and 1 indicating full pooling. 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. [Back to main]

Table 5: Pooling Factors for Categorical Logit Effects (Reference Category: Positive)

Outcome	Treatment Effects			Control Group Means		
	$\omega(\kappa_j)$	$\breve{\omega}(\kappa_j)$	$\lambda(\kappa_j)$	$\omega(\rho_j)$	$\breve{\omega}(\rho_j)$	$\lambda(\rho_j)$
Profit (Negative vs Positive)	0.378	0.712	0.913	0.146	0.424	0.248
Profit (Zero vs Positive)	0.133	0.496	0.690	0.012	0.381	0.495
Expenditures (Zero vs Positive)	0.085	0.625	0.788	0.010	0.489	0.561
Revenues (Zero vs Positive)	0.137	0.695	0.881	0.010	0.503	0.566

Notes: All pooling factors have support on [0,1], with 0 indicating no pooling and 1 indicating full pooling. 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. [Back to main]

Table 6: Pooling Factors for Lognormal Parameters

Location Parameters						
Outcome	Treatment Effects			Control Group Means		
	$\omega(\tau_j)$	$\breve{\omega}(\tau_j)$	$\lambda(\tau_j)$	$\omega(\mu_j)$	$\breve{\omega}(\mu_j)$	$\lambda(\mu_j)$
Profit (Negative Tail)	0.422	0.786	0.938	0.294	0.252	0.274
Profit (Positive Tail)	0.185	0.711	0.870	0.009	0.019	0.002
Expenditures	0.100	0.592	0.712	0.003	0.017	0.001
Revenues	0.048	0.293	0.393	0.002	0.007	0.001
Scale Parameters						
Profit (Negative Tail)	0.307	0.424	0.681	0.290	0.366	0.465
Profit (Positive Tail)	0.118	0.529	0.739	0.026	0.035	0.064
Expenditures	0.036	0.302	0.392	0.006	0.169	0.017
Revenues	0.051	0.457	0.540	0.007	0.047	0.020

Notes: All pooling factors have support on [0,1], with 0 indicating no pooling and 1 indicating full pooling. 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. [Back to main]

Appendix C Site-Specific Shrinkage Results from All Models

This section provides the results of the site-specific shrinkage from all the models fit in the main body of the paper, in order of appearance in the text.

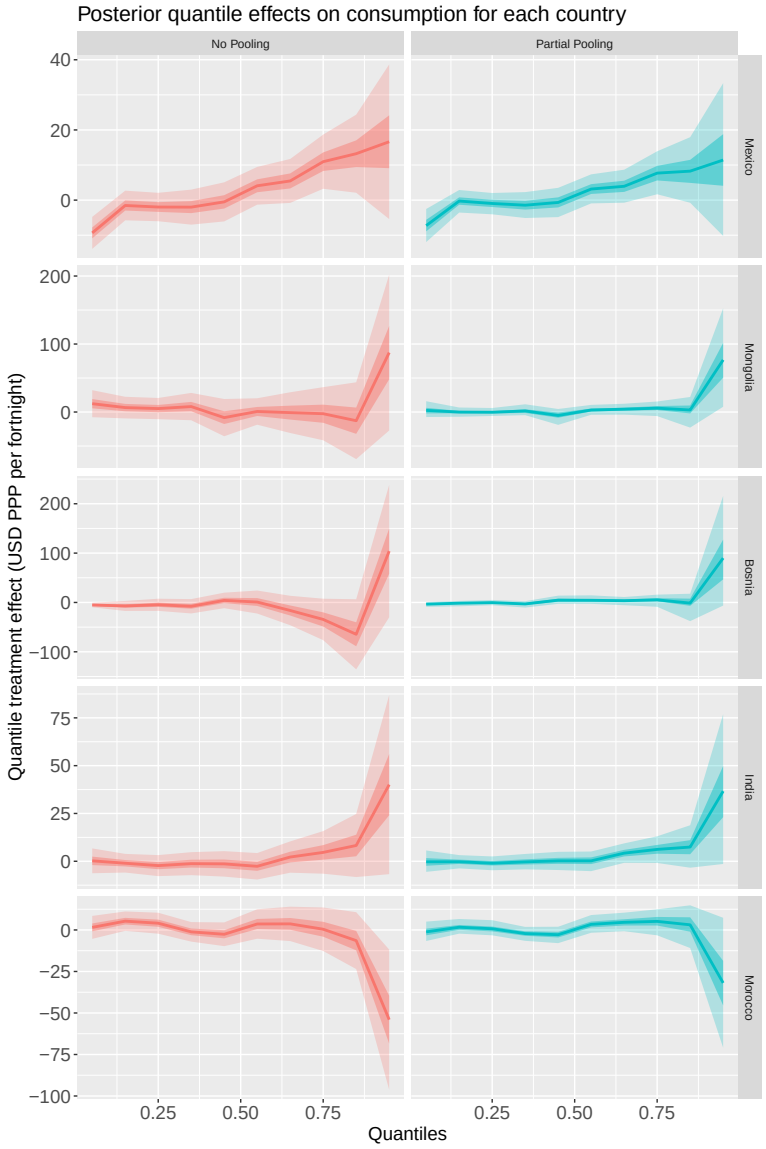

Figure 10: Site by site results for the consumption outcomes. [Back to main]

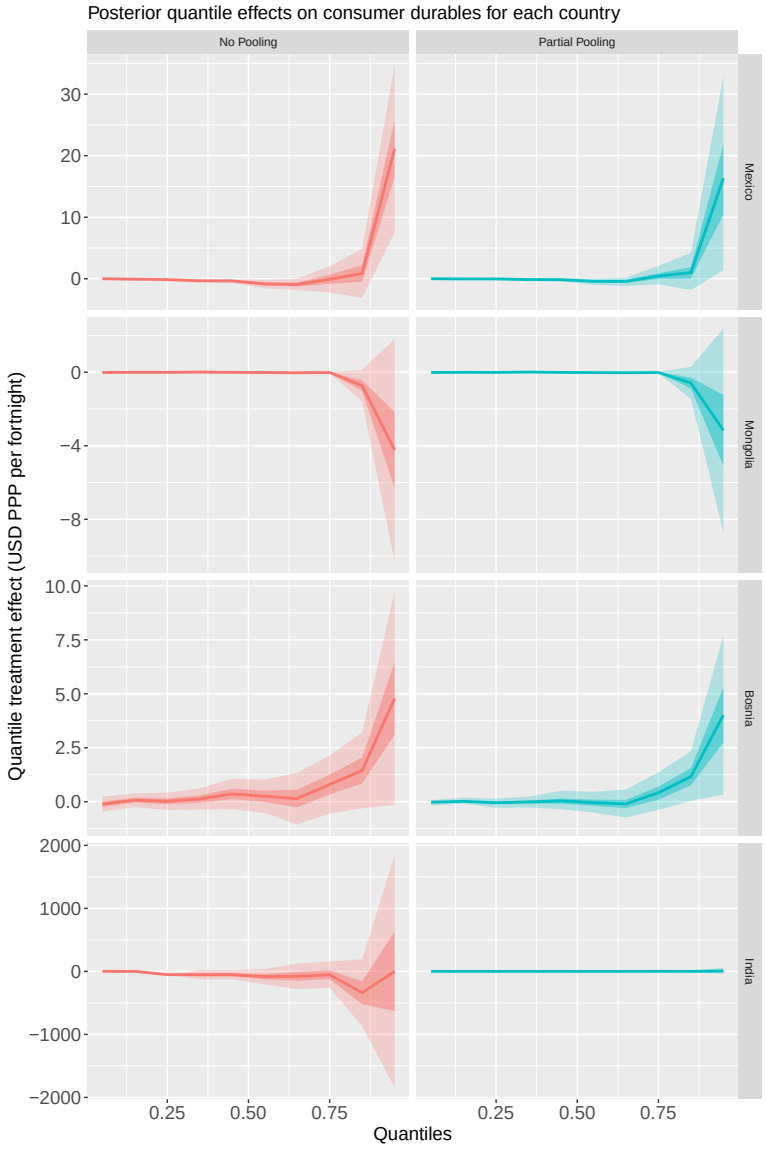

Figure 11: Site by site results for the consumer durables outcomes.[Back to main]

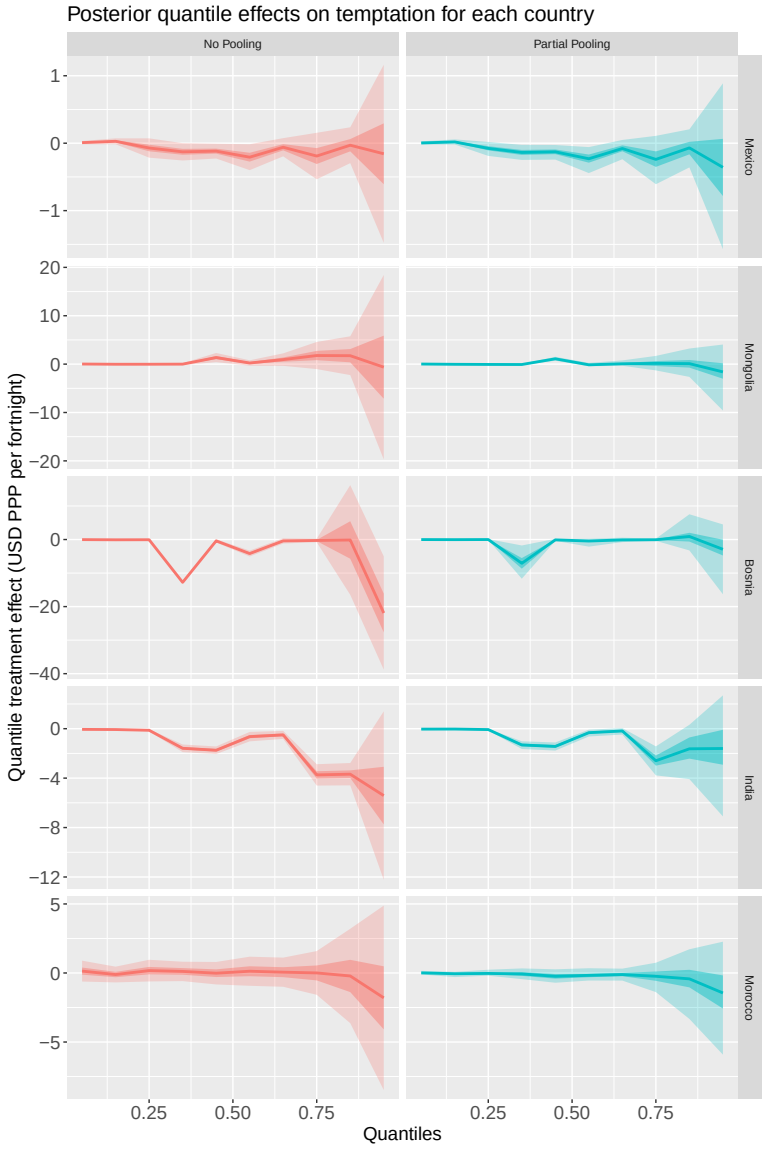

Figure 12: Site by site results for the temptation outcomes. [Back to main]

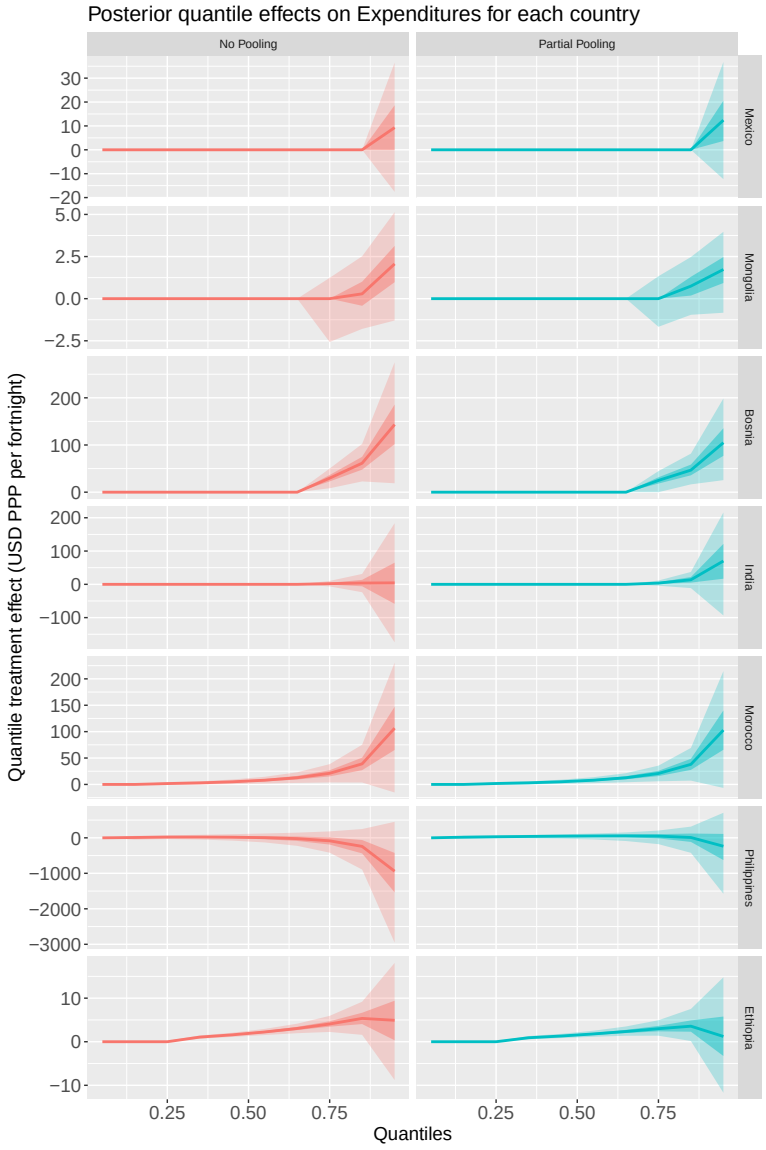

Figure 13: Site by site results for the expenditures outcomes. [Back to main]

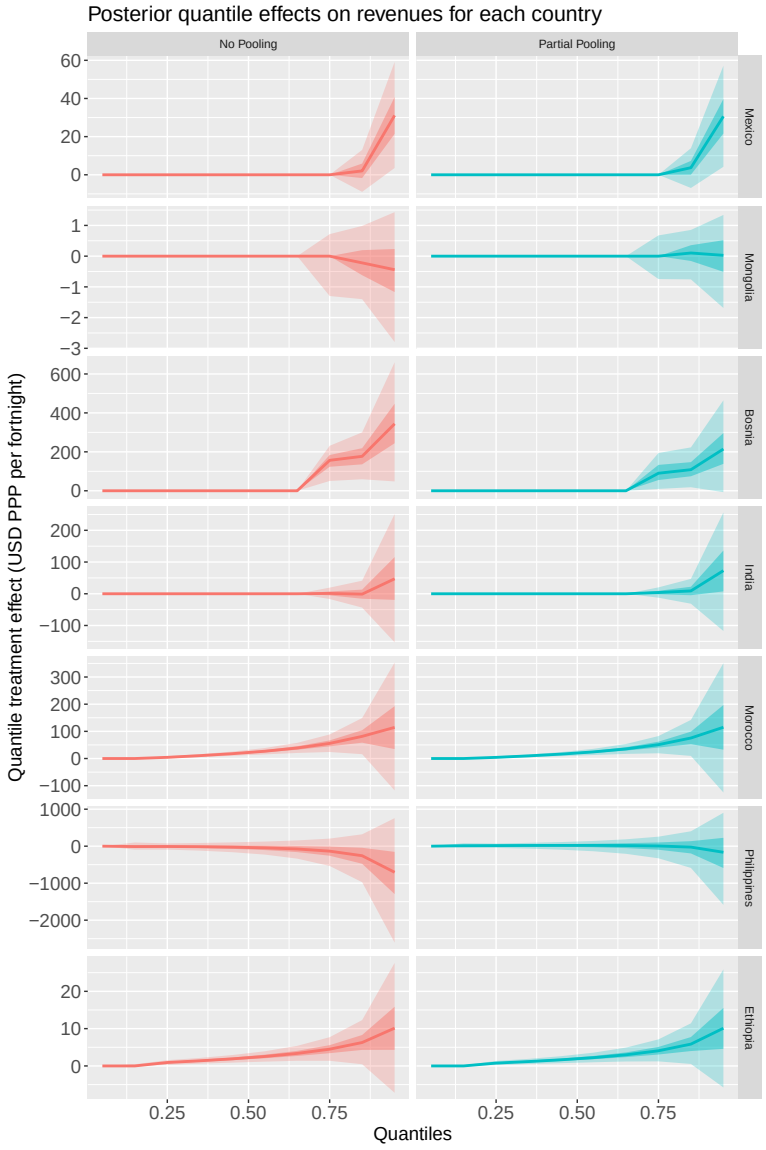

Figure 14: Site by site results for the revenues outcomes. [Back to main]

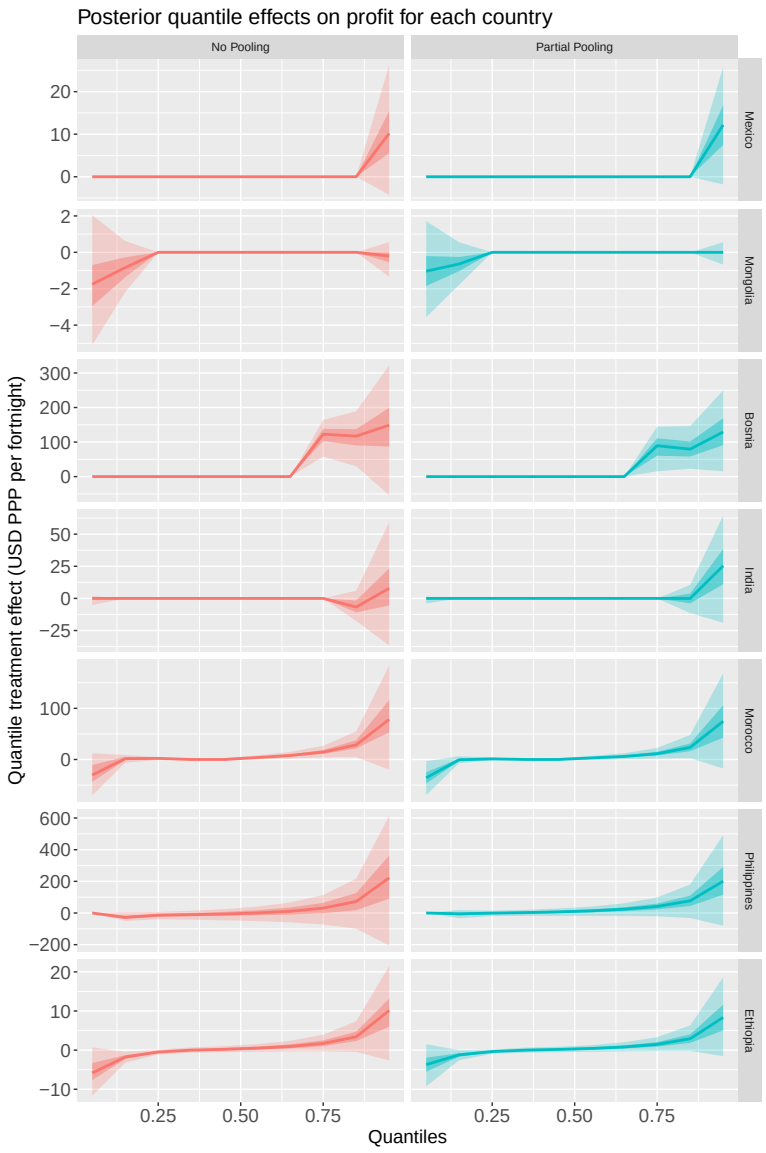

Figure 15: Site by site results for the profit outcomes. [Back to main]

Appendix D The role of take-up

One concern about the models presented in the main analysis is that they ignore the role of differential take-up in explaining the impact of microcredit. While the results of the analysis stand for themselves as group-level causal impacts, the economic interpretation of the results might differ if we knew, for example, that the zero impact along most of the outcome quantiles was entirely due to lack of take-up by most of the households in the sample. The main results contain suggestive evidence that the lack of impact at most quantiles is not solely due to lack of take-up: the 2 sites that randomized loan access rather than branch access and therefore had almost full take-up (Bosnia and the Philippines) displayed the same trend as all the other sites (Appendix C). Yet the observed pattern of zeroes could still be due to low differential in take-up between treatment and control group, which was recorded in most sites. It would be ideal to understand the effect of microcredit on those who are induced to take up loans by this random expansion of access.

There is no satisfactory way to identify the distributional treatment effect only on those households who were randomly induced to take up a loan (the "compliers" in the Neyman-Rubin causal framework), because it is unlikely that the Stable Unit Treatment Value Assumption (SUTVA) holds for individual households within a village.[32]

I pursue a bounding exercise that provides suggestive evidence that take-up patterns are unlikely to be responsible for the precise zero results along most of the distribution. Ideally, the right comparison to make is between the group of households who took up microcredit only due to the random expansion of access, and the same group of households in the control group. This comparison estimates the distributional effect on the compliers. But we cannot identify those households in the control group, because they are indistinguishable from the "never taker" households.[33] Nor can we separate the compliers from the "always takers" in the treatment group. However, under a set of broadly reasonable assumptions for the microcredit setting - i.e. excluding SUTVA - it is possible to develop bounds on the changes in the compliers' distribution.

Consider the conventional Neyman-Rubin causal model with no defiers, and assume that selection into treatment is monotonically increasing in the treatment effect, so that the always-takers have the largest effects, the compliers have moderate effects and the never-takers have the smallest effects. This is reasonable for microcredit because it costs something to access microloans, but that something is usually time and energy rather than physical assets; so poor households with big

[32] If SUTVA did hold, using ITT as an instrument for takeup would identify the average effect on compliers (aka the Local Average Treatment Effect) which in this literature is exactly the object of interest. However, to my knowledge an analogous result is not available for quantiles of the distribution of outcomes for the complier population, and SUTVA most likely does not hold here in any case. Without SUTVA it may still be possible to infer certain average characteristics of the compliers as in Finkelstein and Notowidigdo (2018), but this exercise does not easily extend to quantiles and relies on zero effects for never takers.

[33] The households in the control group who do manage to access microcredit are the always takers and not the compliers against which the appropriate comparison can be made.

effects are more likely to be "always takers" than rich households with small effects. Then, even under moderate violations of SUTVA and moderate rank re-ordering of households (even such that they can cross ranks with households from other groups), it is the case that the following bounds hold.

Denote the three groups of households: always takers, compliers and never takers, $G \in \{AT, C, NT\}$. Suppose that the three groups have ordered distributional treatment effects despite the SUTVA violations, in order to make the notion of selection on treatment effects coherent without SUTVA. Denote the quantiles of a group's outcome distribution $Q_G(TU, T)$ where TU is a binary indicator of taking up the loans, and T is any vector of assigned treatment status for the households in the given village or local area. A sufficient, though strong, condition for ordered distributional effects is that pointwise for any u:

$$Q_{AT}(1,T)(u) - Q_{AT}(0,T')(u) \geq Q_C(1,T)(u) - Q_C(0,T')(u) \geq Q_{NT}(1,T)(u) - Q_{NT}(0,T')(u) \; \forall \, T, T'. \tag{D.1}$$

Thus, regardless of the particular treatment regime, these types are ordered by their quantile treatment effects. This ordering would be implied by a similar ordering on the individual treatment effects, but as this ordering does not imply a full ordering on the individual effects, it is more general. Armed with this condition I can derive bounds without assuming that there is no effect on the never-takers (in contrast to Imbens and Rubin 1997, Abadie Angrist and Imbens 2002, and Finkelstein and Notowidigdo 2018) which is fortunate because this is unlikely when SUTVA is violated in general, and particularly when other households in one's village are taking up new sources of credit.

First, consider comparing the outcome of the households who take up in treatment versus those households who do not take up in control, as a potential upper bound on the distributional effects on compliers. These groups are composed of combinations of the three groups denoted above, so denote the quantiles of a combination of two groups by $\mathcal{Q}(Q_G, Q_{G'})$ such that $\mathcal{Q}(Q_{AT}(1,T), Q_C(1,T))$ applies to households who take up in the treatment group, and $\mathcal{Q}(Q_{NT}(0,T'), Q_C(0,T'))$ applies to those who do not in control. The quantiles of the combined group do not have simple relationships to the quantiles of each underlying group in general, particularly when the groups may be of differing size in the sample (and population). Regardless, there are partial orders on the distributions represented by these quantile functions that allow the derivation of sufficient conditions for the following upper boundary condition to hold pointwise for any quantile u:

$$\mathcal{Q}(Q_{AT}(1,T)(u), Q_C(1,T)(u)) - \mathcal{Q}(Q_{NT}(0,T')(u), Q_C(0,T')(u)) \geq Q_C(1,T)(u) - Q_C(0,T')(u) \; \forall \, T. \tag{D.2}$$

In this case, first order stochastic dominance of certain pairs of group outcomes permits the following sufficient conditions for this bound to hold. First notice that by definition if $F_G \; FOSD \; F_{G'}$, then $Q_G(u) \geq Q_{G'}(u)$ for all u. If these two groups are combined by simple pooling, as in the above exercise, then the quantiles of the pooled data will lie weakly below the more favourable distribution and weakly above the less favourable distribution, such that $Q_G(u) \geq \mathcal{Q}(Q_G, Q_{G'})(u) \geq Q_{G'}(u)$. This is true no matter the relative size of the two groups being pooled. Thus a sufficient condition for the bound to hold is that $Q_{AT}(1,T) \; FOSD \; Q_C(1,T)$ and

$Q_C(0,T')$ $FOSD$ $Q_{NT}(0,T')$. That is, always takers have better outcomes from taking up loans than compliers do, and compliers have better outcomes without taking up loans than never-takers do. The dominance does not have to hold across T allocations in this case, as long as it holds within them. While the sufficient condition is much stronger than the necessary condition, which is simply the bound itself, this exercise provides some understanding of the kinds of situations in which the bound may be reasonable.

Now consider comparing the outcome of the households who take up in treatment versus those households who take up loans in control. For this comparison to form a lower bound on the distributional effects for compliers, it must be that

$$\mathcal{Q}(Q_{AT}(1,T)(u), Q_C(1,T)(u)) - Q_{AT}(1,T')(u) \leq Q_C(1,T)(u) - Q_C(0,T')(u) \; \forall \; T, T' : T' < T. \tag{D.3}$$

This is the place where SUTVA violations make the comparison challenging and require us to compare distributions across values of T. In this case, we are comparing across a case of some T to a case where fewer people are actually taking up the loan because fewer people are treated, hence, the considered alternative T' has fewer non-zero entries, and is thus smaller in magnitude. If SUTVA holds, then the T argument is irrelevant, and the sufficiency condition for D.2 is sufficient for this bound as well. If SUTVA does not hold, then there are two cases to consider. Either the violation pushes up the value of $Q_{AT}(1,\cdot)$, in which case the earlier sufficiency condition still holds. If the violation pushes down the value of the quantiles, however, there is an additional requirement that it must do so in a moderate way: sufficient, but again not necessary for this, would be that the SUTVA violation adjustment not be both negative and larger in magnitude than the treatment effect on compliers at that quantile.

These bounds will make sense when selection probability is monotonically increasing as the treatment effect grows larger, and when the levels of the outcomes are somewhat positively correlated with this treatment effect. They will be unlikely to hold if either of these two patterns does not hold in the data. Thus while these bounds may not be applicable to every situation, they do seem applicable to the microcredit data. As discussed above, since it generally takes a lot of time and effort to access microcredit, households are more likely to do so if their own treatment effects are larger. Further, while we cannot be sure this reflects a full ordering within sites, the cross-site correlation between the average treatment effects and the control groups' levels of consumption, profit, etc is generally positive (Meager, 2018). This at least provides suggestive evidence that a positive correlation between levels and effects may be present within sites as well, and that as a result, these bounds are reasonable.

I compute these bounds and I find that the posited lower bound does lie weakly below the posited upper bound in all cases (and strictly below in the case of consumption). However, the bounds are very close together and overall similar to the main distributional effect estimated by comparing treatment status itself (the "ITT" comparison, or the "access as treatment" comparison). Comparing the households who took up the loans in the treatment group to households in the control group who did not take up loans produces largely similar results - although they are weakly

more positive - as comparing all treated and control households, as shown in figure 16. Consumption is an exception to this trend, and the positive non-zero results for this comparison are interesting, but as an upper bound this does not overshadow the null results on the rest of the variables. The results of comparing the households who took up the loans in the treatment group to households who took up in the control group for all outcomes is shown in figure 17. These effects tend to be broadly similar to the impact of mere access, in that they are zero almost everywhere, although on average the effects are estimated to lie weakly below the ITT effect. While this analysis provides suggestive evidence that microcredit's lack of impact below the 75th quantile is not solely due to lack of take-up, it is not conclusive. A structural analysis of this data or an additional experiment would be required to obtain a more definitive answer to this question.

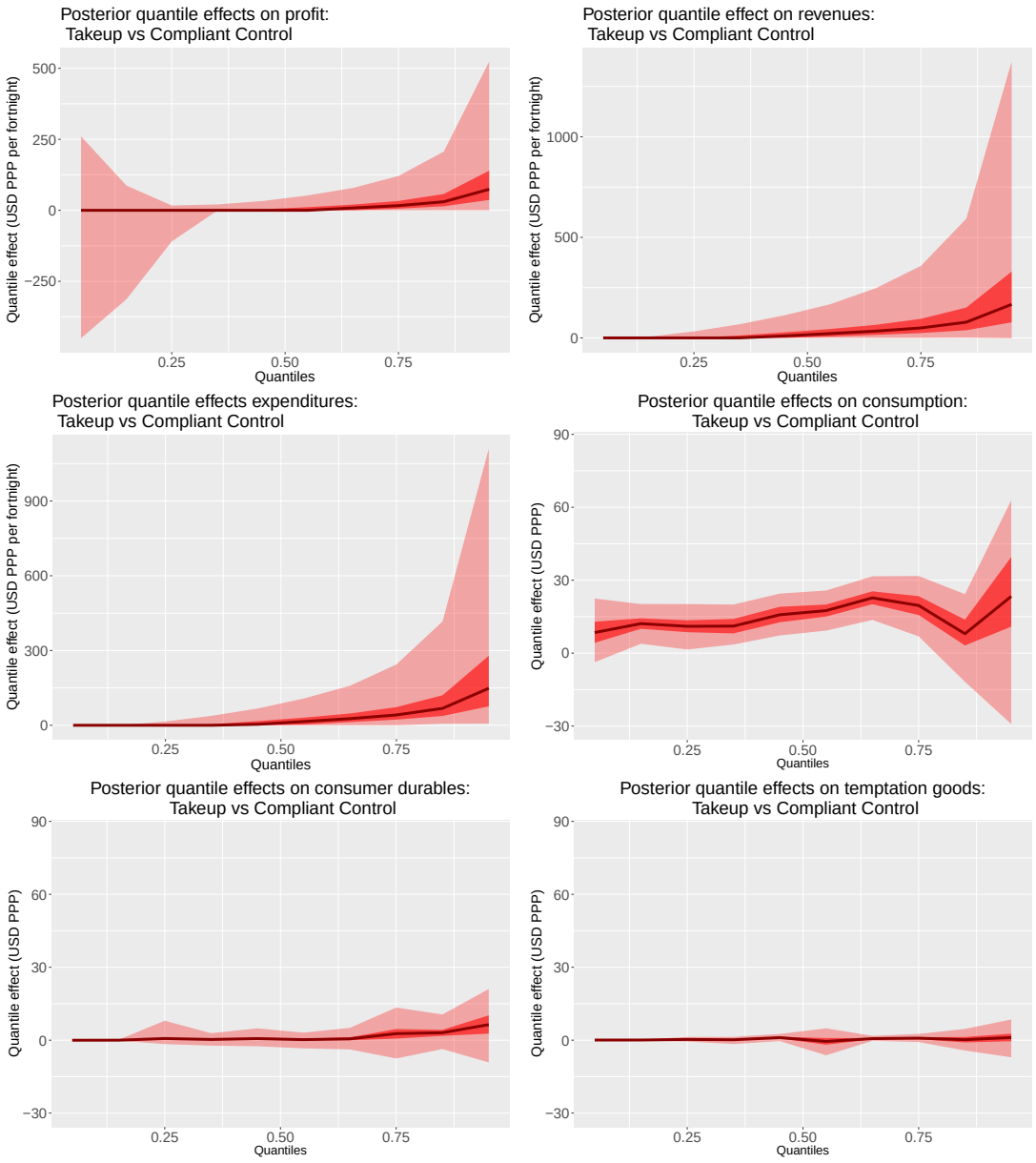

Figure 16: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Compliant control households who did not take up. This effect should overestimate the true impact of microcredit on those who take it up in a simple selection framework. Consumption variables are in USD PPP per two weeks, business variables are in cubed root of USD PPP per two weeks due to the scale differences in their uncertainty intervals. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. [Back to main]

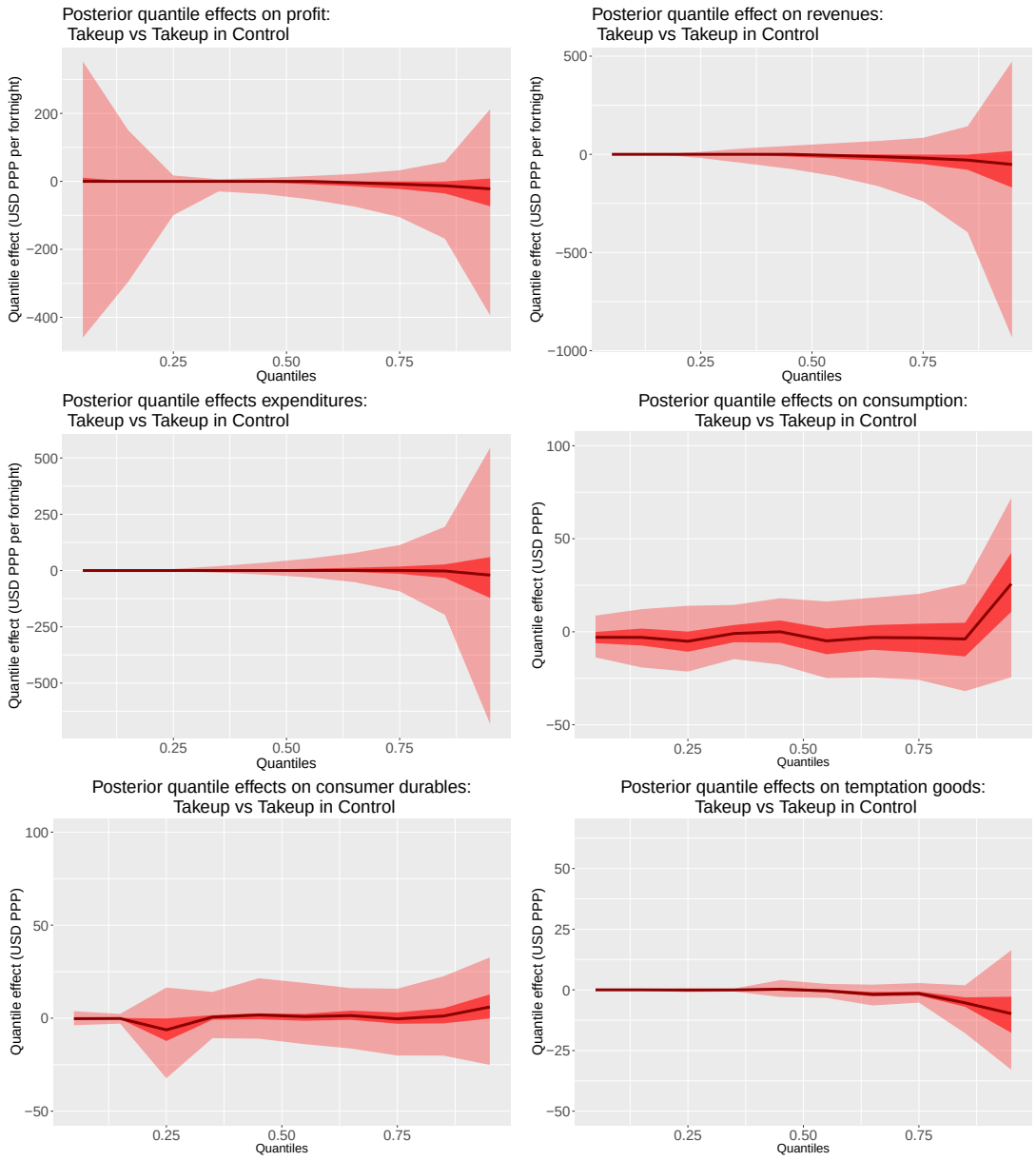

Figure 17: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Control households who took up. This effect should underestimate the true impact of microcredit on those who take it up in a simple selection framework. Consumption variables are in USD PPP per two weeks, business variables are in cubed root of USD PPP per two weeks due to the scale differences in their uncertainty intervals. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. [Back to main]

Appendix E Tabular results

Table 7: Lender and Study Attributes by Country

Country	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco	The Philippines
Study Citation	Augsburg et al (2015)	Tarozzi et al (2015)	Banerjee et al (2015b)	Angelucci et al (2015)	Attanasio et al (2015)	Crepon 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

Note: The construction of the interest rates here is different to the construction of Banerjee et al (2015a); 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. This table was also printed in Meager (2018) which used the same studies.[Back to main]

Excess Kurtosis in LogNormal distributions is the extent to which tail indices are greater, and thus the extent to which the tails are heavier, than those of the Gaussian. For a LogNormal parameterised as

$$LogNormal(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma y}\exp\left(\frac{-(\log(y)-\mu)^2}{2\sigma^2}\right)$$

the excess kurtosis is

$$\exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 6.$$

I compute the kurtosis for the general control group based on the posterior mean values of μ and σ for this group in the tables below. The example in the text is obtained using μ_2 and σ_2^c.

Table 8: All General-Level Posterior Marginals for the LogNormal Profit Model

	mean	MCMC error	sd	2.5%	25%	50%	75%	97.5%	# effective draws	$\hat{R}$
μ_1	3.200	0.008	0.732	1.722	2.784	3.200	3.615	4.698	9,099	1.000
μ_2	3.843	0.007	0.818	2.225	3.356	3.845	4.324	5.496	15,000	1.000
τ_1	0.094	0.001	0.094	-0.099	0.045	0.095	0.143	0.273	6,719.600	1.001
τ_2	0.077	0.0005	0.042	-0.007	0.054	0.078	0.102	0.157	7,566.232	1.000
σ_{μ_1}	1.659	0.008	0.654	0.867	1.227	1.514	1.923	3.302	7,284.792	1.000
σ_{μ_2}	2.033	0.006	0.677	1.153	1.574	1.889	2.332	3.711	15,000	1.000
σ_{τ_1}	0.117	0.004	0.128	0.005	0.035	0.079	0.154	0.459	1,090.338	1.003
σ_{τ_2}	0.055	0.001	0.052	0.002	0.020	0.043	0.075	0.183	1,323.050	1.004
σ_1^c	0.452	0.002	0.145	0.180	0.374	0.447	0.525	0.761	7,205.404	1.000
σ_2^c	0.225	0.001	0.101	0.022	0.167	0.225	0.284	0.428	10,278.910	1.000
σ_1^t	0.022	0.001	0.094	-0.162	-0.024	0.022	0.067	0.206	6,128.028	1.001
σ_2^t	0.017	0.0003	0.029	-0.043	0.001	0.017	0.032	0.072	9,321.264	1.000
$\sigma_{\sigma_1^c}$	0.302	0.002	0.164	0.122	0.196	0.262	0.357	0.724	5,126.273	1.001
$\sigma_{\sigma_2^c}$	0.242	0.001	0.100	0.125	0.176	0.220	0.280	0.499	9,328.806	1.000
$\sigma_{\sigma_1^t}$	0.163	0.002	0.116	0.034	0.089	0.134	0.201	0.467	2,860.338	1.001
$\sigma_{\sigma_2^t}$	0.046	0.001	0.037	0.002	0.020	0.038	0.062	0.140	2,034.778	1.002
β_{11}	-1.965	0.016	1.273	-4.525	-2.715	-1.958	-1.193	0.527	6,334.358	1.000
β_{12}	0.025	0.001	0.114	-0.187	-0.035	0.019	0.080	0.265	6,957.068	1.001
β_{21}	0.390	0.010	0.906	-1.379	-0.168	0.367	0.918	2.255	7,964.995	1.000
β_{22}	-0.067	0.001	0.104	-0.279	-0.124	-0.066	-0.012	0.143	8,309.348	1.001
$\sigma_{\beta_{11}}$	2.767	0.017	1.277	0.770	1.959	2.560	3.346	5.904	5,636.316	1.000
$\sigma_{\beta_{12}}$	0.128	0.002	0.125	0.005	0.047	0.096	0.168	0.446	5,901.720	1.001
$\sigma_{\beta_{21}}$	1.603	0.014	0.902	0.130	0.990	1.532	2.093	3.672	3,987.814	1.002
$\sigma_{\beta_{22}}$	0.146	0.002	0.114	0.007	0.065	0.124	0.197	0.432	5,234.755	1.001
$\sigma_{\beta_{31}}$	1.450	0.014	0.889	0.091	0.815	1.381	1.942	3.493	3,896.658	1.001
$\sigma_{\beta_{32}}$	0.117	0.002	0.109	0.004	0.041	0.089	0.161	0.390	5,085.964	1.002

Note: The β_3 parameters are normalized to be zero at the general level as required for multinomial logit models. The site-specific effects still have variation around this zero anchor as reported.

Table 9: All General-Level Posterior Marginals for the LogNormal Revenues Model

	mean	MCMC error	sd	2.5%	25%	50%	75%	97.5%	# effective draws	$\hat{R}$
μ_1	4.472	0.007	0.873	2.733	3.959	4.479	4.992	6.193	15,000	1.000
τ_1	0.083	0.001	0.068	-0.058	0.045	0.086	0.123	0.211	10,482.840	1.000
σ_{μ_1}	2.181	0.007	0.718	1.258	1.693	2.030	2.496	3.982	10,285.460	1.000
$\sigma_{\tau_1}]$	0.140	0.001	0.080	0.039	0.089	0.124	0.171	0.329	5,189.630	1.001
σ_1^c	0.213	0.001	0.136	-0.063	0.134	0.214	0.292	0.485	11,190.950	1.000
σ_1^t	-0.010	0.0003	0.031	-0.071	-0.028	-0.011	0.008	0.052	9,554.774	1.000
$\sigma_{\sigma_1^c}$	0.331	0.001	0.135	0.171	0.241	0.301	0.383	0.668	8,452.406	1.001
$\sigma_{\sigma_1^t}$	0.062	0.0004	0.033	0.020	0.040	0.055	0.075	0.146	6,447.524	1.000
β_{11}	0.011	0.008	0.734	-1.464	-0.424	-0.004	0.443	1.521	8,107.184	1.001
β_{12}	-0.063	0.001	0.081	-0.235	-0.101	-0.058	-0.020	0.091	6,772.048	1.001
$\sigma_{\beta_{11}}$	1.209	0.010	0.760	0.064	0.637	1.164	1.645	2.912	5,305.339	1.001
$\sigma_{\beta_{12}}$	0.095	0.001	0.091	0.003	0.032	0.071	0.129	0.327	5,418.020	1.001
$\sigma_{\beta_{21}}$	1.192	0.010	0.762	0.062	0.615	1.147	1.631	2.894	5,341.343	1.001
$\sigma_{\beta_{22}}$	0.095	0.001	0.091	0.003	0.033	0.071	0.130	0.328	5,944.329	1.000

Note: The β_3 parameters are normalized to be zero at the general level as required for multinomial logit models. The site-specific effects still have variation around this zero anchor as reported. Note also that σ_1^t can be negative as this is the effect specified on the exponential level.

Table 10: All General-Level Posterior Marginals for the LogNormal Expenditures Model

	mean	MCMC error	sd	2.5%	25%	50%	75%	97.5%	# effective draws	$\hat{R}$
μ_1	4.042	0.006	0.733	2.563	3.593	4.047	4.483	5.528	15,000	1.000
τ_1	0.103	0.001	0.048	0.005	0.076	0.104	0.132	0.198	8,840.624	1.000
σ_{μ_1}	1.867	0.005	0.624	1.061	1.449	1.735	2.135	3.449	15,000	1.001
σ_{τ_1}	0.078	0.001	0.060	0.004	0.035	0.067	0.106	0.226	1,919.668	1.002
σ_1^c	0.303	0.002	0.171	-0.037	0.204	0.304	0.401	0.649	8,974.738	1.001
σ_1^t	-0.008	0.001	0.045	-0.092	-0.033	-0.009	0.016	0.082	5,069.866	1.000
$\sigma_{\sigma_1^c}$	0.421	0.002	0.171	0.218	0.309	0.382	0.489	0.845	8,374.404	1.001
$\sigma_{\sigma_1^t}$	0.094	0.001	0.051	0.035	0.062	0.082	0.111	0.217	3,164.881	1.001
β_{11}	0.234	0.009	0.694	-1.177	-0.180	0.233	0.653	1.645	6,027.909	1.000
β_{12}	-0.116	0.001	0.117	-0.349	-0.177	-0.114	-0.053	0.112	7,262.210	1.000
$\sigma_{\beta_{11}}$	1.148	0.011	0.712	0.062	0.613	1.102	1.565	2.729	4,414.652	1.001
$\sigma_{\beta_{12}}$	0.157	0.002	0.125	0.007	0.071	0.132	0.209	0.465	5,601.528	1.000
$\sigma_{\beta_{21}}$	1.119	0.011	0.707	0.056	0.580	1.075	1.535	2.714	4,076.193	1.001
$\sigma_{\beta_{22}}$	0.159	0.002	0.124	0.007	0.074	0.136	0.212	0.463	5,427.373	1.001

Note: The β_3 parameters are normalized to be zero at the general level as required for multinomial logit models. The site-specific effects still have variation around this zero anchor as reported.Note also that σ_1^t can be negative as this is the effect specified on the exponential level.

For visual ease, the figures below graph the treatment effects and posterior predicted effects for each of the dimensions of change permitted in the model.

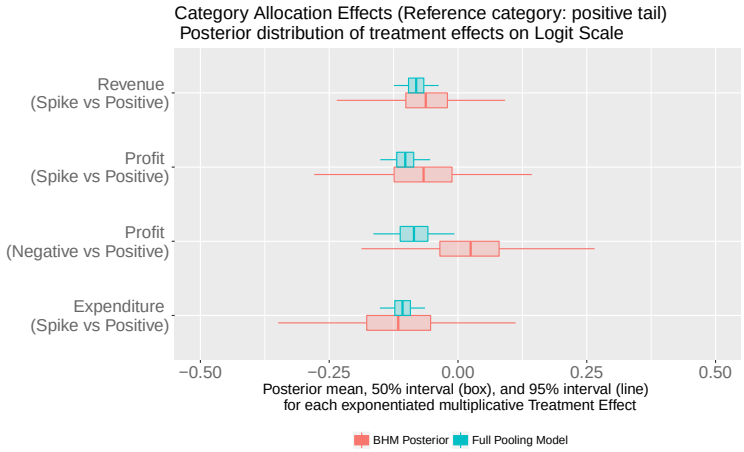

Figure 18: Posterior distributions for the logit treatment effects (π_j) on category assignment. These treatment effects are specified as an exponentiated multiplicative factor on the control group proportion of households in the category: if $\tilde{\pi_j} = 0$ the effect is zero, if $\tilde{\pi_j} < 0$ the treatment increases the proportion of households in the positive tail relative to other categories. [Back to main]

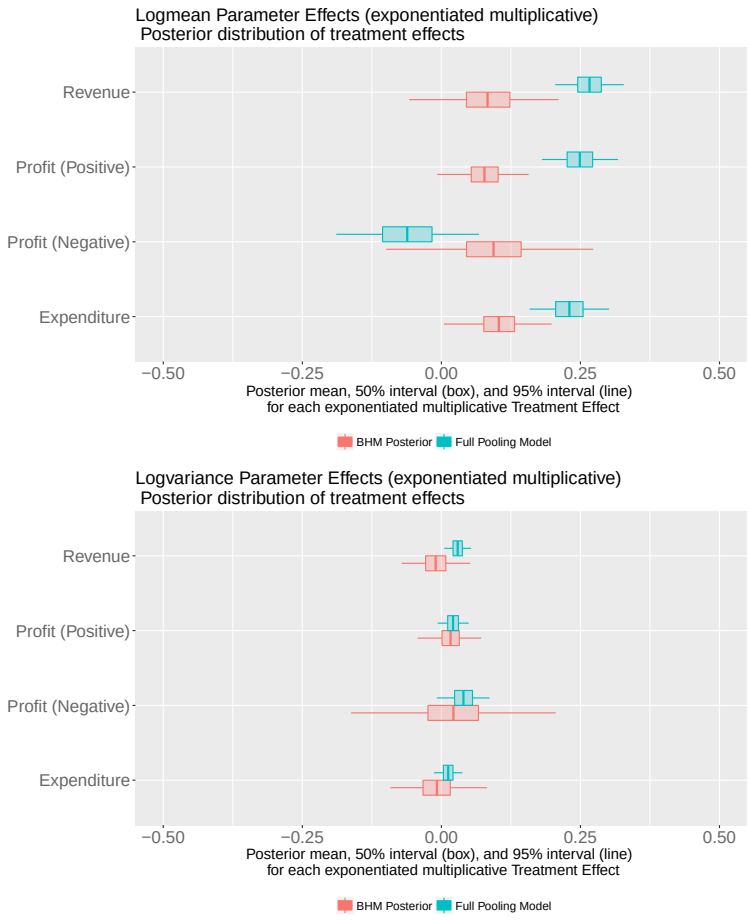

Figure 19: Posterior distributions for the location treatment effects (τ_j) and the scale treatment effects (σ_j^t). [Back to main]

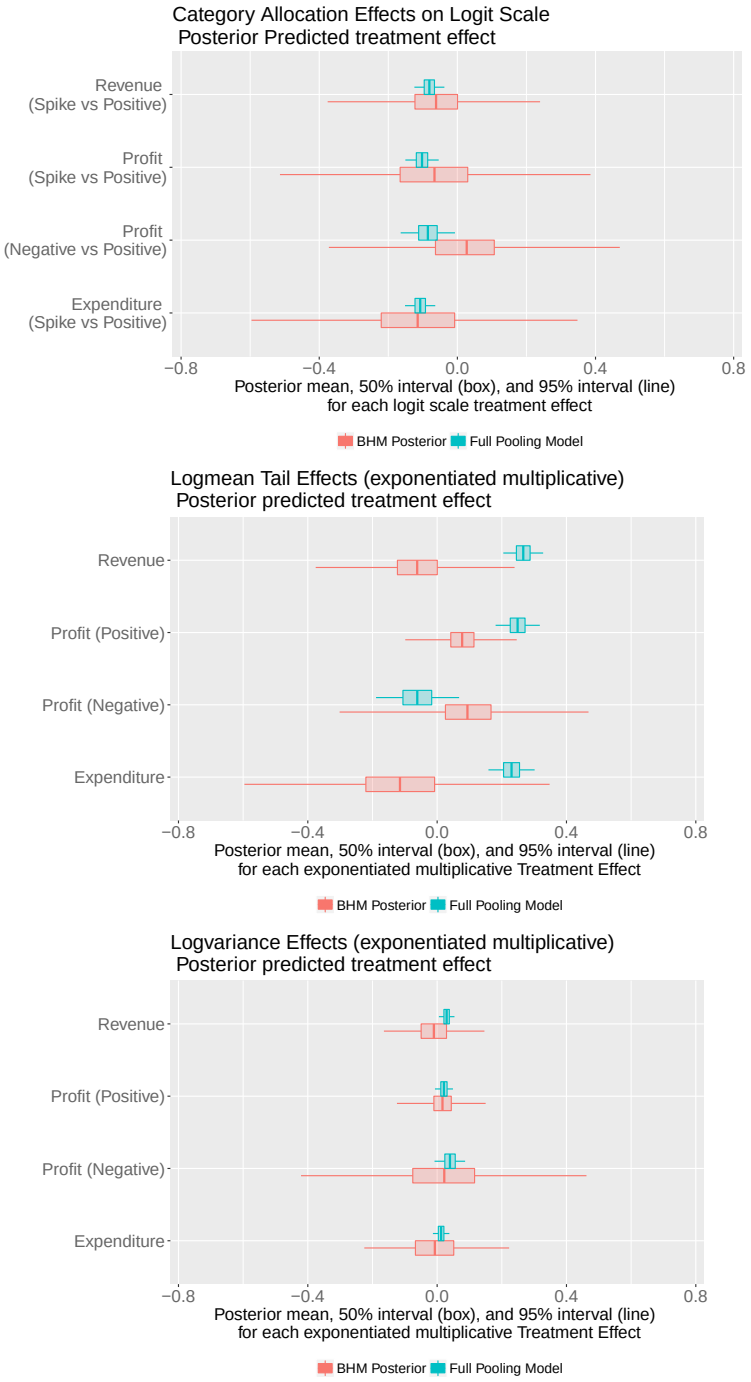

Figure 20: Posterior predicted distributions for the logit treatment effects on category assignment and tail shape effects. [Back to main]

References

[1] Abadie, A., Angrist, J. and Imbens, G. (2002) "Instrumental Variables estimates of the effect of subsidised training on quantiles of trainee earnings." Econometrica, 70(1) 91-117.

[2] Acemoglu, D., and Robinson, J. A. (2008). "Persistence of power, elites, and institutions." American Economic Review, 98(1), 267-93.

[3] Acemoglu, D. Suresh Naidu, Pascual Restrepo, James A. Robinson, (2015) "Chapter 21 - Democracy, Redistribution, and Inequality", The Handbook of Income Distribution, Editor(s): Anthony B. Atkinson, François Bourguignon, Elsevier, Volume 2, 2015, Pages 1885-1966.

[4] Acharya, A., Blackwell, M., & Sen, M. (2016). "Explaining causal findings without bias: Detecting and assessing direct effects." American Political Science Review, 110(3), 512-529.

[5] Ahmad, M. M. (2003). "Distant voices: the views of the field workers of NGOs in Bangladesh on microcredit." The Geographical Journal, 169(1), 65-74.

[6] Albert, J and Siddhartha Chib (1997) "Conditionally Independent Hierarchical Models", Journal of the American Statistical Association, September 1997, Vol. 92, No. 439

[7] Allcott, H. (2015). "Site Selection Bias in Program Evaluation." The Quarterly Journal of Economics, 130(3), 1117-1165.

[8] Allen, T. (2014). "Information frictions in trade". Econometrica, 82(6), 2041-2083.

[9] Andrews, I., and Maximilian Kasy (2017) "Identification of and Correction for Publication Bias", NBER Working Paper No. 23298, March 2017, Revised November 2017

[10] Andrews, I., and Oster, E. (2017). "Weighting for External Validity" (No. w23826). National Bureau of Economic Research.

[11] Angelucci, M., 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.

[12] Angrist, J. D. (2004). "Treatment effect heterogeneity in theory and practice". The Economic Journal, 114(494), C52-C83.

[13] Angrist, J., and Ivan Fernandez-Val . (2010). "Extrapolate-ing: External validity and overidentification in the late framework" (No. w16566). National Bureau of Economic Research.

[14] Athey, S., and Guido Imbens. "Recursive partitioning for heterogeneous causal effects." Proceedings of the National Academy of Sciences 113, no. 27 (2016): 7353-7360.

[15] Attanasio, O., 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.

[16] Augsburg, B., 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.

[17] Autor, D. H., Katz, L. F., and Krueger, A. B. (1998). "Computing inequality: have computers changed the labor market?". The Quarterly Journal of Economics, 113(4), 1169-1213.

[18] Autor, D. H., Dorn, D., Hanson, G. H., and Song, J. (2014). "Trade adjustment: Worker-level evidence". The Quarterly Journal of Economics, 129(4), 1799-1860.

[19] Bandiera, O, G. Fischer, A. Prat and E.Ytsma (2017) "Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments" Working Paper Version October 2017

[20] Banerjee, A. (2013). "Microcredit under the microscope: what have we learned in the past two decades, and what do we need to know?". Annu. Rev. Econ., 5(1), 487-519.

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

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

[23] Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Pariente, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. (2015c) "A multifaceted program causes lasting progress for the very poor: Evidence from six countries." Science 348, no. 6236 (2015): 1260799.

[24] Banerjee, A., & Mullainathan, S. (2010). The shape of temptation: Implications for the economic lives of the poor (No. w15973). National Bureau of Economic Research. NBER Working Paper No. 15973, Issued in May 2010

[25] Bazzi, S. (2016) "Wealth Heterogeneity and the Income Elasticity of Migration" American Economic Journal: Applied, 9(2), 219-55.

[26] Bell, A., Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen (2017) "Who Becomes an Inventor in America? The Importance of Exposure to Innovation" NBER Working Paper No. 24062, December 2017

[27] Bertanha, M., and Guido Imbens (2014). "External validity in fuzzy regression discontinuity designs" (No. w20773). National Bureau of Economic Research.

[28] Betancourt, M. J., and Mark Girolami. (2013). "Hamiltonian Monte Carlo for hierarchical models." arXiv preprint arXiv:1312.0906

[29] Bisbee, J., Rajeev Dehejia, Cristian Pop-Eleches, Cyrus Samii. "Local Instruments, Global Extrapolation: External Validity of the Labor Supply". Journal of Labor Economics, Volume 35, Number S1, pp. S99-S147.

[30] Borusyak, K., & Jaravel, X. (2018). "The Distributional Effects of Trade: Theory and Evidence from the United States." Working Paper Version.

[31] Breza, E., & Kinnan, C. (2018). "Measuring the equilibrium impacts of credit: Evidence from the Indian microfinance crisis" (No. w24329). National Bureau of Economic Research

[32] Breza, E. (2012). "Peer effects and loan repayment: Evidence from the krishna default crisis." Working paper.

[33] Bryan, G., Chowdhury, S., & Mobarak, A. M. (2014). "Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh." Econometrica, 82(5), 1671-1748.

[34] Castellacci, Giuseppe, (2012) "A Formula for the Quantiles of Mixtures of Distributions with Disjoint Supports". Available at SSRN: http://ssrn.com/abstract=2055022 or http://dx.doi.org/10.2139/ssrn.2055022

[35] Chernozhukov, V., Ivan Fernandez-Val, and Alfred Galichon.(2010) "Quantile and probability curves without crossing." Econometrica 78.3 1093-1125.

[36] Chetty, R., Hendren, N., and Katz, L. F. (2016). "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment." American Economic Review, 106(4), 855-902.

[37] Chetty, R., and Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects. The Quarterly Journal of Economics, 133(3), 1107-1162.

[38] Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. American Economic Review, 104(9), 2593-2632

[39] Chib, S., & Greenberg, E. (1995). Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models. Journal of Econometrics, 68(2), 339-360.

[40] Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. Journal of Educational and Behavioral Statistics, 40(2), 136-157.

[41] Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A non-degenerate penalized likelihood estimator for variance parameters in multilevel models. Psychometrika, 78(4), 685-709.

[42] Crepon, Bruno, Florencia Devoto, Esther Duflo, and William Pariente. 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.

[43] Dehejia, R., Pop-Eleches, C., and Samii, C. (2015). "From Local to Global: External Validity in a Fertility Natural Experiment" (No. w21459). National Bureau of Economic Research.

[44] Dehejia, R. H. (2003). "Was there a Riverside miracle? A hierarchical framework for evaluating programs with grouped data." Journal of Business & Economic Statistics, 21(1), 1-11.

[45] Diaconis, Persi (1977) "Finite Forms of De Finetti's Theorem on Exchangeability" Synthese, 36, 1977, 271-281

[46] Duflo, E., Pascaline Dupas and Michael Kremer, (2017) "The Impact of Free Secondary Education: Experimental Evidence from Ghana" Working Paper, 2017

[47] Duvendack, M., Richard Palmer-Jones & 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

[48] Efron, B., and Morris, C. (1975). "Data analysis using Stein's estimator and its generalizations". Journal of the American Statistical Association, 70(350), 311-319.

[49] Fama, Eugene F., (1963), "Mandelbrot and the Stable Paretian Hypothesis", The Journal of Business, 36, http://EconPapers.repec.org/RePEc:ucp:jnlbus:v:36:y:1963:p:420.

[50] Fama, Eugene F. (1965) "Portfolio Analysis In A Stable Paretian Market." Management Science 11.3 : 404-419. Business Source Complete. Web. 10 Aug. 2016.

[51] Fogli, A and Guerrieri, V. (2017). "The End of the American Dream? Inequality and Segregation in US cities." In 2017 Meeting Papers (No. 1309). Society for Economic Dynamics.

[52] Gabaix, X. (2008) "Power Laws in Economics and Finance" NBER Working Paper No. 14299, accessed online August 12th 2016, http://www.nber.org/papers/w14299

[53] Gechter, M. (2015). "Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India". manuscript, Pennsylvania State University.

[54] Gelman, A., and Carlin, J. (2014). "Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors." Perspectives on Psychological Science, 9(6), 641-651.

[55] Gelman, A., John B. Carlin, Hal S. Stern and Donald B. Rubin (2004) "Bayesian Data Analysis: Second Edition", Taylor & Francis

[56] Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). "A weakly informative default prior distribution for logistic and other regression models". The Annals of Applied Statistics, 2(4), 1360-1383.

[57] Gelman, A.,& Jennifer Hill (2007) "Data anlysis using regression and multilevel heirarchical models" Cambridge Academic Press.

[58] Gelman, A., and Pardoe, I. (2006). "Bayesian measures of explained variance and pooling in multilevel (hierarchical) models". Technometrics, 48(2), 241-251.

[59] Gelman, A., and Rubin, D. B. (1992)." Inference from iterative simulation using multiple sequences". Statistical science, 457-472.

[60] Giordano, R., Tamara Broderick, Rachael Meager, Jonathan Huggins, Michael Jordan (2016) "Fast robustness quantification with variational Bayes" ICML Workshop on Data4Good: Machine Learning in Social Good Applications, New York, NY, arXiv:1606.07153

[61] Hartley, H. O., and Rao, J. N. (1967). "Maximum-likelihood estimation for the mixed analysis of variance model". Biometrika, 54(1-2), 93-108.

[62] Hussam, R., Rigol, N., and Roth, B. (2017). "Targeting high ability entrepreneurs using community information: Mechanism design in the field." Working Paper, Nov 2017

[63] Hastie, T., Tibshirani, R., and Friedman, J. (2009). "The elements of statistical learning". Second Edition. Springer Series in Statistics.

[64] He, X. (1997). "Quantile Curves without Crossing". The American Statistician, 51(2), 186-192. doi:10.2307/2685417

[65] Heckman, J., Tobias, J. L., and Vytlacil, E. (2001). "Four parameters of interest in the evaluation of social programs". Southern Economic Journal, 211-223.

[66] Higgins, J. P. and Sally Green (Eds) (2011) "Cochrane handbook for systematic reviews of interventions" (Version 5.1.0). Chichester: Wiley-Blackwell.

[67] Hlavac, Marek (2014). "stargazer: LaTeX/HTML code and ASCII text for well-formatted regression and summary statistics tables". R package version 5.1. http://CRAN.R-project.org/package=stargazer

[68] Hoffman, M. D.,& Gelman, A. (2014)."The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo". The Journal of Machine Learning Research, 15(1), 1593-1623.

[69] Hsiang, S. M., Burke, M., and Miguel, E. (2013). "Quantifying the influence of climate on human conflict." Science, 341(6151), 1235367.

[70] Hull, P. (2018). Estimating hospital quality with quasi-experimental data. Chicago Working Paper.

[71] Imbens, G. W., & Rubin, D. B. (2015). "Causal inference in statistics, social, and biomedical sciences." Cambridge University Press.

[72] Imbens, G.W and Rubin D.B (1997) "Estimating outcome distributions for compliers in instrumental variables models" The Review of Economic Studies, 64(4), 555-574

[73] James, William, and Charles Stein. (1961) "Estimation with Quadratic Loss." In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1:361?379, 1961.

[74] Jones, C. I. (2015). "Pareto and Piketty: The macroeconomics of top income and wealth inequality." The Journal of Economic Perspectives, 29(1), 29-46.

[75] Kaboski, J. P., and Townsend, R. M. (2011). "A structural evaluation of a large?scale quasi?experimental microfinance initiative". Econometrica, 79(5), 1357-1406.

[76] Karlan, D., and Zinman, J. (2009). "Observing unobservables: Identifying information asymmetries with a consumer credit field experiment." Econometrica, 77(6), 1993-2008.

[77] Karlan, Dean and Jonathan Zinman (2011) "Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation", Science 10 June 2011: 1278-1284

[78] Katz, L. F., Kling, J. R., & Liebman, J. B. (2001). "Moving to opportunity in Boston: Early results of a randomized mobility experiment." The Quarterly Journal of Economics, 116(2), 607-654.

[79] Kinnan, C., & Townsend, R. (2012). Kinship and financial networks, formal financial access, and risk reduction. American Economic Review, 102(3), 289-93.

[80] Koenker R and Gilbert Bassett, Jr. (1978) "Regression Quantiles" Econometrica, Vol. 46, No. 1. (Jan., 1978), pp. 33-50.

[81] Koenker, R, and Kevin F. Hallock. (2001). "Quantile Regression." Journal of Economic Perspectives, 15(4): 143-156

[82] Koenker, R (2005) "Quantile Regression", Econometric Society Monographs No. 38, Cambridge University Press 2005

[83] Koenker, R, (2011) "Additive models for quantile regression: Model selection and confidence bandaids" Brazilian Journal of Probability and Statistics, 2011, Vol. 25, No. 3, 239-262

[84] Leon, A. C., and Heo, M. (2009). "Sample Sizes Required to Detect Interactions between Two Binary Fixed-Effects in a Mixed-Effects Linear Regression Model." Computational Statistics & Data Analysis, 53(3), 603?608.

[85] Machado, J.A.F, and J. M. C. Santos Silva (2005) "Quantiles for Counts" Journal Of The American Statistical Association Vol. 100 , Iss. 472

[86] McCulloch, Charles and Neuhaus, John M. "Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter". Statist. Sci. 26 (2011), no. 3, 388–402.

[87] McKenzie, D. (2012). "Beyond baseline and follow-up: The case for more T in experiments." Journal of development Economics, 99(2), 210-221.

[88] Meager, R. (2018). "Understanding the Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of 7 Randomised Experiments." Accepted in the American Economics Journal: Applied, January 2018

[89] Microfinance Barometer (2017) "Microfinance Barometer 8th Edition: Is Microfinance Still Working?", Convergences, France, www.convergences.org, 2017

[90] Microfinance Focus (2011) "Six Microfinance Crises the sector does not want to remember." http://www.microfinancefocus.com/6-microfinance-crises-sector-does-not-want-remember

[91] Morduch, J. (1999). "The microfinance promise." Journal of economic literature, 37(4), 1569-1614.

[92] Mosteller (1946) "On Some Useful "Inefficient" Statistics" The Annals of Mathematical Statistics, Vol. 17, No. 4. (Dec., 1946), pp. 377-408

[93] Pancost, A. (2016) "Do Financial Factors Drive Aggregate Productivity? Evidence from Indian Manufacturing Establishments" Working Paper, accessed online August 2016

[94] Piketty, T. (2015). About capital in the twenty-first century. American Economic Review, 105(5), 48-53.

[95] Pritchett, Lant & J. Sandefur (2015) "Learning from Experiments when Context Matters" American Economic Association 2015 Preview Papers, accessed online February 2015

[96] Reich, B. J., Fuentes, M., & Dunson, D. B. (2011). Bayesian spatial quantile regression. Journal of the American Statistical Association, 106(493), 6-20.

[97] Roodman, D. (2012). "Due diligence: An impertinent inquiry into microfinance". CGD Books.

[98] Roy, A. D. (1950). "The distribution of earnings and of individual output." The Economic Journal, 60(239), 489-505.

[99] Rubin, D. B. (1981). "Estimation in parallel randomized experiments. Journal of Educational and Behavioral Statistics", 6(4), 377-401.

[100] Rubin, H. (1950). "Note on random coefficients". Statistical inference in dynamic economic models, 419-421.

[101] Schicks, J. (2013). "From a supply gap to a demand gap? The risk and consequences of over-indebting the underbanked." In Microfinance in Developing Countries (pp. 152-177). Palgrave Macmillan, London.

[102] Stan Development Team (2017) "Stan Modeling Language: User's Guide and Reference Manual." Version 2.17.0.

[103] Stiglitz, J. E., and Weiss, A. (1981). "Credit rationing in markets with imperfect information". The American economic review, 71(3), 393-410.

[104] Stein, C. (1956) "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Berkeley and Los Angeles, University of California Press, 1956, Vol. 1, pp. 197-206.

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

[106] Townsend,R.M. (2018) "Townsend Thai Project Household Annual Resurvey, 2017 (Rural)", https://doi.org/10.7910/DVN/UW4VKE, Harvard Dataverse, V1

[107] Vivalt, E. (2016) "How much can we generalise from impact evaluations?" Working Paper, NYU

[108] Van der Vaart, A.W. (1998) "Asymptotic Statistics", Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press

[109] Wald, A. (1947). "Foundations of a general theory of sequential decision functions". Econometrica, Journal of the Econometric Society, 279-313.

[110] Wickham, H. (2009) "ggplot2: elegant graphics for data analysis". Springer New York, 2009.

[111] Yunus, M, (2006) "Nobel Lecture", Oslo, December 10, 2006.