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

Rachael Meager*†

November 28, 2017

Abstract

This paper develops methods to aggregate evidence on distributional treatment effects from multiple studies conducted in different settings, and applies them to the microcredit literature. Several randomized trials of expanding access to microcredit found substantial effects on the tails of household outcome distributions, but the extent to which these findings generalize to future settings was not known. Aggregating the evidence on sets of quantile effects poses additional challenges relative to average effects because distributional effects must imply monotonic quantiles and pass information across quantiles. Using a Bayesian hierarchical framework, I develop new models to aggregate distributional effects and assess their generalizability. For continuous outcome variables, the methodological challenges are addressed by applying transforms to the unknown parameters. For partially discrete variables such as business profits, I use contextual economic knowledge to build tailored parametric aggregation models. I find generalizable evidence that microcredit has negligible impact on the distribution of various household outcomes below the 75th percentile, but above this point there is no generalizable prediction. Thus, there is strong evidence that microcredit typically does not lead to worse outcomes at the group level, but no generalizable evidence on whether it improves group outcomes. Households with previous business experience account for the majority of the impact in the tails and see large increases in the upper tail of the consumption distribution in particular.

*Massachusetts Institute of Technology. Contact: rmeager@mit.edu

†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, and the audiences of the MIT Economic Development Lunch Seminar, MIT Econometrics Lunch Seminar, the MIT Summer Applied Micro Lunch Series, and NEUDC 2016 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. I welcome any further feedback via email.

1 Introduction

It is increasingly recognized that translating research into policy requires aggregating evidence from multiple studies of the same economic phenomenon (Allcott 2015, Dehejia et al. 2015, Banerjee et al. 2015). This translation requires not only an estimate of the impact of an intervention across different contexts, but also an assessment of the generalizability of the evidence and hence its applicability to policy decisions in other settings. Moreover, making these policy decisions often requires information about the distributional treatment effects of the intervention, because average treatment effects can hide large negative and positive impacts. Researchers who study interventions such as microcredit often estimate sets of quantile treatment effects to characterize the impact across the entire distribution of household outcomes. Yet there is currently no methodology to formally aggregate the evidence on sets of quantile effects across settings when the generalizability of the results is not known. This paper develops these methods in a Bayesian hierarchical framework and uses them to generate new insights about the causal impact of expanding access to microcredit.

Distributional treatment effects are a concern in the microcredit literature because the causal impact of access to loans is likely to be heterogeneous across households (Banerjee et al., 2015). Despite studies showing that microcredit may not have large effects on average outcomes, microfinance organizations and policy groups have suggested that microloans may still provide major increases in household profit and consumption for certain types of borrowers (CGAP, 2011). Yet it is equally possible that microcredit could actually harm some borrowers due to restrictive loan contracts and the potential for over-indebtedness (Schicks, 2013). The applied theoretical literature suggests there may be winners and losers from credit market interventions in general (Kaboski and Townsend 2011, Buera, Kaboski and Shin 2012, Buera and Shin 2013). Such heterogeneous effects may also influence the resource distribution across households, which has consequences for social welfare. Furthermore, household welfare is not solely determined by the level of consumption or income but also by the volatility or risk attached to these variables, which microcredit could alleviate (Collier et al. 2011, Banerjee 2013). Thus, policymakers who must consider these issues when making decisions about microcredit interventions need reliable evidence about the distributional impact of microcredit and the generalizability of this impact across settings.

The experimental literature on microcredit to date provides some empirical evidence that these interventions might help some households and harm others (Angelucci et al. 2015, Attanasio et al. 2015, Augsburg et al. 2015, Banerjee et al. 2015, Crepon et al. 2015, Karlan and Zinman 2011, and Tarozzi et al. 2015). Some studies found negative impacts on the lower tail of the distribution of household profits, such that some treated households had worse outcomes than ever observed in control households (Crepon et al. 2015). Yet some studies also found positive impacts on the upper tail, such that some treated households saw higher profits than any in the control group (Angelucci et al. 2015, Augsburg et al. 2015, Banerjee et al. 2015, Crepon et al. 2015). But in most studies the impact was imprecisely estimated in the tails, which in some cases prevented any firm conclusions (Tarozzi et al. 2015). The important question for policy purposes is whether these tail effects are a general or robust feature of microcredit interventions,

and, if so, what their expected magnitude may be in another location. This question can only be answered by aggregating all the evidence on the quantile treatment effects and assessing their generalizability.

The task of extracting generalizable information about the causal impact of microcredit or other interventions studied across diverse contexts presents several challenges. Evidence aggregation requires estimating the general impact of an intervention using information about local effect in several different studies or sites. Intuitively, this involves computing some kind of weighted average of the effects, but as the goal is to apply this inference to a broad class of future study sites, these weights should emphasize information that is detected to be common across the sites in the sample. The extent to which the site-specific effects contain common information is intrinsically linked to the generalizability of this body of evidence: the similarity of the local effects to each other and to the average effect is a signal of how close a comparable future site's effect will be to this average effect. This notion of generalizability measures how accurately we can predict the effect in future sites given the evidence from the current sites, which is precisely the definition of external validity in Allcott (2015) and Dehejia et al. (2015). Hence, assessing generalizability is central to evidence aggregation, as it informs both how the general impact is estimated and how the result is interpreted for policy in future locations.

Hierarchical modeling provides the tools to simultaneously aggregate evidence across settings and assess the generalizability of the resulting estimates. The key idea is to specify parameters at multiple levels of the data, with relationships between the levels parameterized and estimated (Wald 1947, Rubin 1950, Hartley & Rao 1967, Efron & Morris 1975, Rubin 1981). In particular, Bayesian estimation of these models accurately characterizes the joint uncertainty over the parameters at all levels, and provides regularization which is beneficial in the low-data environment of meta-analysis (Gelman et al. 2004, and Hastie, Tibshirani & Friedman 2009). Hierarchical models allow heterogeneous effects at the study level, and detect the extent to which the data supports pooling information across studies or passing information up to the general level. If little or no generalizable information is detected across the studies, then the hierarchical model reports wide uncertainty intervals on the general effect and associated pooling metrics will be small. By contrast, popular fixed-effects or "full pooling" techniques neither measure generalizability nor produce estimates that account for it. These simpler methods are only optimal when the studies have homogeneous effects: a strong assumption for the microcredit literature which spans seven different countries. The hierarchical approach is more appropriate for most applications in social science, and is increasingly used by economists (Dehejia 2003, Hsiang, Burke & Miguel 2015, Vivalt 2015, Meager 2015). Yet within this framework there are no tools to aggregate distributional effects such as sets of quantile treatment effects.

This paper fills the methodological gap and performs to my knowledge the first estimation of the general distributional impact of any intervention in the literature. The current literature on external validity and generalizing from experimental data has focused exclusively 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). Thus, despite the growing attention to evidence aggregation and external validity in economics, the task of aggregating distributional treatment effects has been

ignored thus far. Using a Bayesian hierarchical framework, I develop methods to aggregate treatment effects on the dispersion in household outcomes and to aggregate sets of quantile treatment effects that characterize the entire distributional impact. I also develop new metrics of generalizability by extending the existing pooling metrics to accommodate multidimensional treatment effects. The extension of these techniques to sets of distributional treatment effects poses several new methodological challenges, which I address within the Bayesian framework.

The first challenge for aggregating distributional effects is that these estimates must be constrained to satisfy support constraints on the estimands. Otherwise the models may produce inference and policy conclusions that we know to be incorrect ex-post. Sets of quantile treatment effects must imply monotonic sets of quantiles in the treatment and control groups because these quantiles are transposed cumulative density functions. Methods that do not impose these constraints fail to exploit all the information in the sample, since the constraint indicates that neighboring quantiles contain information about each other that can be used to improve the inference for each quantile. Moreover, unconstrained estimation may produce results that imply that a household at the 50th percentile of the consumption distribution is consuming more than a household at the 60th percentile, which is logically impossible. Yet implementing the monotonicity constraint may be challenging, because the support of each effect depends on the true values of the neighboring effects and control quantiles, all of which are estimated rather than known. This interdependence in the support constraints means that the constraint must be jointly imposed on the entire vector of quantile outcomes throughout the estimation process.

For continuous outcome variables, I solve this problem using variable transformation on the unknown parameters to implement the support constraints. The quantile treatment effects vector can be constrained to imply monotonic outcome quantiles by introducing these vectors of quantiles and explicitly transforming them using functions that only have support on monotonic vectors. Because Bayesian inference treats unknowns as random variables, this implements the support constraint without distortion or loss of information about the statistical uncertainty by using the classical transformation of random variables formula. Imposing the constraint via estimation of a transformed variable passes more information across quantiles when implied monotonicity is likely to be violated. The resulting quantile treatment effect estimate is the one most supported by the data among the set of those that satisfy the constraints on the true effects.[1] I use this transformation approach to construct models that aggregate the distributional effects of continuously distributed variables, building on the structure of the classical Mosteller (1946) theorem about the asymptotic distribution of empirical quantiles.

The second challenge for aggregating the distributional effects of microcredit access is that not all household outcomes are continuously distributed: business profit, revenues and expenditures do not satisfy the assumptions of the Mosteller (1946) theorem. The challenges of performing quantile inference for partially discrete outcomes are well understood in the quantile regression literature, but the existing statistical solutions do not work well for the household outcome variables in the microcredit data (e.g. Machado and Santos Silva 2005). The solutions fail because these household outcomes are fundamentally mixture variables, composed of a point mass or

[1] By contrast, existing approaches in the frequentist literature impose a particular monotonizing procedure on the estimated effects ex-post, such as rearranging or smoothing: see for example Chernozhukov et al, 2010.

"spike" at zero for households that did not operate businesses and a continuous distribution of expenditures, revenues and profits for those that did. These point masses cannot be ignored because microcredit interventions might cause net changes in the proportion of households that make zero profits versus positive or negative profits, an effect that can only be detected by including the point mass in the analysis.

The general solution I propose to the problem of aggregating quantile inference on mixture variables is to build parametric models that capture the underlying economic structure that generates each specific outcome variable. The aggregation can then be performed on the rich set of parameters that govern the tailored mixture distribution, and the intervention's effect on those parameters. This permits the analyst to break down the quantile treatment effects into a set of effects on the economic mechanisms that generate the outcome, although the implied quantile effects can be recovered via simulation from the Bayesian posterior. Explicit parameterization of the outcome variable automatically satisfies the support constraints on the implied quantile treatment effects and the outcome quantiles. This procedure also leverages information from neighboring quantiles because they are linked by the functional form assumptions.

More specifically, I propose a set of mixture models for the microcredit data based on economic and contextual knowledge about the activities of the households in the sample. The key economic insight is that the household business variables are produced by a partially discrete underlying decision structure: first, a household chooses (or receives a random shock) to operate a business or not, and second, they choose how much to invest given their local conditions which in turn produces some revenue and profit. Therefore, the impact of microcredit can be broken up into its impact on the discrete decision margin - a "category switching" effect for the households - and its impact on the continuous decision margin of how much to invest or how to turn expenditures into profit. I model the microcredit intervention as having some potential impact on both of these margins, although future work could microfound this impact to produce a fully structural model for aggregation. Both theoretical and applied work on wealth, income and profit in many settings suggests that these variables often follow power laws, perhaps due to the way that latent distributions of productivity interact with local market structures, so the continuous tails can be modeled using Pareto distributions (Gabaix 2008, Allen 2014, Jones 2015, Bazzi 2016). This parameterization solves the problem of passing information across quantiles and imposes monotonicity on the outcome distributions in the microcredit data.

Implementing these approaches, I provide models to aggregate treatment effects both on sets of quantiles for continuous and mixture variables, and on the dispersion or variance of household outcomes. Fitting my models to the experimental microcredit literature yields strong and generalizable evidence that microcredit does not affect outcomes below the 75th quantile, but no generalizable evidence on whether microcredit improves the upper tail beyond this point. This pattern is detected in all six household outcomes studied: consumption, consumer durables spending, temptation goods spending, business profit, business expenditures and business revenues. I also find moderate increases in the dispersion of household business outcomes, signaling some increase in inequality across households ex-post, and this finding is reasonably generalizable. The parametric models that decompose the quantile effects into two channels find only weak support for the category switching effect and no support for any direct expansion of the

continuous tails. The lack of generalizability in the right tail of household outcomes is partially driven by extremely high kurtosis in the business variables, such that inference based on Gaussian asymptotics for the mean and variance effects is unreliable for these outcomes (Fama 1965, Koenker and Basset 1978).

An analysis of the role of household covariates reveals that the majority of the impact of microcredit and the heterogeneity at the study level occurs within the group of households who had previous business experience. There is strong evidence that these experienced households increase their consumption above the 75th percentile in the general case, although there is still little change below this percentile. Households without business experience see negligible impact at every quantile for most outcomes, although they do see a noisy but large impact in the upper tail of consumer durables spending. Hence, while entrepreneurial households use microcredit to expand their businesses and increase total consumption, other households will see at best a change in the composition of their consumption. However, as both types of households have extremely heavy tails in their business outcomes, there may be large individual variation within these groups. The presence of high kurtosis in business outcomes suggests that it is important for economists to study individuals even when the overall goal is to understand broader phenomena or aggregate output, as certain individuals make major contributions to total output even in rural village economies. Attempts to handle heavy-tailed distributions by trimming the top 1% of observations are misguided in this context, as it is critically important to study the far right tail of the distribution despite the econometric challenges involved.

These results demonstrate the value of analysing and aggregating evidence using appropriate methodology, particularly when Gaussian approximations may be unreliable for the variables of interest. Quantile regression provides robust estimates of distributional treatment effects at the group level, and parametric models can accommodate the discrete components and high kurtosis often found in economic variables (e.g. Bazzi 2016, Pancost 2016, Gabaix 2008, Fama 1965). By contrast, common methods used to analyse average treatment effects such as ordinary least squares regression are not robust to high kurtosis, and may provide inaccurate results when applied to heavy tailed distributions (Koenker and Basset 1978). In such cases, meta-analytic methods based on Gaussian asymptotic approximations may compound these errors. Hence, the use of distributional analysis rather than average effects analysis as well as the application of the aggregation models developed in this paper could improve the reliability of meta-analyses in many areas of economics. These methods are particularly relevant to any interventions for which heterogeneous effects present a policy concern, such as education interventions, deworming pills, and labour market interventions, for which these models can deliver results that are both more informative and more reliable than analyses based only on average treatment effects.

2 Data and Economic Context

This paper considers evidence from seven randomized experiments on expanding access to microcredit services in the field. The selected studies are Angelucci et al.(2015), Attanasio et al.(2015), Augsburg et al.(2015), Banerjee et al.(2015), Crepon et al.(2015), Karlan and Zinman (2011), and Tarozzi et al.(2015). The selection was limited to randomized controlled trials

(RCTs) because there is no established methodology for combining experimental and observational studies that addresses selection bias within and across studies. In this case there is little risk of selection bias within studies, because the sample is restricted to RCTs and the treatment I consider is branch or loan access.[2] In addition, issues of publication bias seem unlikely to be a major risk, because they reported mostly null results for most household outcomes. There is some risk of selection bias of study sites due to conditions required to perform an RCT, so I restrict the interpretation of "generalizability" to the set of possible locations which could plausibly have RCTs or are broadly similar to such places.

I examine six household outcomes: business profit, business revenues, business expenditures, consumption, consumer durables spending and temptation goods spending. Each of these is important for testing the theory that microcredit improves household welfare by creating new opportunities for small-scale entrepreneurship, or by allowing people to shift their consumption into durable items and away from temptation goods. The household business outcomes are measured in all sites. The set of household consumption variables was not measured in each site, but as they are centrally important to testing the welfare impact of microcredit, they must be analyzed regardless (Banerjee 2013). Fortunately, they are measured in at least four sites in each case. The six variables selected here are measured in reasonably comparable ways across sites. While it would be ideal to examine effects on income and assets, the measurement and definition of these variables differed considerably across the studies such that it is unclear how to proceed with aggregation. While many NGOs are interested in microcredit as a tool for women's empowerment, this was measured using localized site-specific indices of variables which differed substantially across sites and thus are similarly challenging to aggregate.

All selected studies consider an expansion of access to loans from microfinance institutions, a policy intervention that is fundamentally equivalent to a relaxation of a credit constraint. It is reasonable to expect that this intervention may have a core economic mechanism that operates to some extent regardless of context.[3] Yet the studies have many dimensions of heterogeneity. They cover seven different countries: Mexico, Mongolia, Bosnia and Herzegovina, India, Morocco, the Philippines, and Ethiopia. They had different partner NGOs, offering similar but not identical loan contract structures with different interest rates and loan sizes. The studies also differed in terms of their randomization units - five randomized at the community level and two at the individual level - with different encouragement and sampling designs. Given this contextual heterogeneity, there is little justification for assuming homogeneous treatment effects across studies. Hence, Bayesian hierarchical models are the right aggregation framework for this data.

The policy debate around the economic and social impact of microcredit provides a setting in which distributional effects are highly salient. The main policy concern is that households may have heterogeneous treatment effects from branch access: this is why the original studies computed quantile treatment effects at multiple points in the distribution (e.g. Banerjee et al.

[2]I consider access and not take-up as the treatment because the Stable Unit Treatment Value Assumption is unlikely to hold within villages or groups which is the unit of randomization in 5 of the 7 studies, so that any IV analysis is likely to be invalid.

[3]Indeed, an assumption of this sort is the underlying premise of the discipline of economics. If no such common mechanism exists for economic phenomena across contexts, much of applied economics research is called into question.

2015). Not only do households differentially select into loan take-up, but they are also likely to experience heterogeneous effects of take-up depending on how they use the loan or whether they experience idiosyncratic shocks. It is therefore plausible that microcredit access could produce large positive and negative impacts for different types of households (CGAP 2011). Various theoretical models of microcredit predict winners and losers even in cases where average welfare is increased (Kaboski and Townsend 2011, Buera, Kaboski and Shin 2012). Thus, microcredit could affect cross-sectional inequality ex-post, which may have consequences for local economic and political systems. In addition, households may use microloans to smooth consumption or investment, which affects the dispersion of outcomes but not necessarily their mean level. Yet these changes in volatility affect household welfare and should be studied where possible (Collier et al. 2011, Banerjee 2013). Another concern is that the functionality of credit markets depends on the risk profile of borrower projects, which implicates the entire distribution of business profits, not just the mean (Stiglitz and Weiss, 1981). The average treatment effect is therefore an insufficient measure of the true impact of microcredit.

While distributional treatment effects are important for policy, their causal interpretation can be subtle. Dispersion metrics and outcome quantiles do not satisfy a law of iterated expectations, so group-level differences cannot be interpreted as expected individual differences. Quantile treatment effects do not estimate the expected causal effect for a household located at the quantile in question unless the treatment satisfies rank-preservation, which is a strong assumption for credit market interventions. Instead, quantile effects must be interpreted as estimating the causal effect on the distribution of outcomes at the group level. However, while positive quantile effects at the upper tail of the profit distribution do not necessarily mean that rich households are getting richer, they do mean that some treated households are better off relative to the control group. Similarly, negative effects at the lower tail do not necessarily mean that poor households are affected negatively, but do imply that some households experience significantly worse outcomes if their community is given access to microcredit. Overall, increases in dispersion at the group level can be interpreted as increases in ex-post inequality across households, and quantile effects can be interpreted as changes in the distribution of group outcomes.

Due to the policies of the *American Economic Journal: Applied* and *Science*, the two journals in which these studies were published, all the raw data is available online. Hence, it is possible to fit rich models for aggregation that move beyond the techniques used by the original study authors. This is crucial for aggregating quantile treatment effects in particular, many of which were reported in the papers with standard errors of zero at certain quantiles due to applying the nonparametric bootstrap to the problem (see for example Angelucci et al, 2015). The quantile effects aggregation methods I develop in this paper are explicitly designed to address the root of this problem, which is that some of the outcomes are mixture variables, composed of a continuous distribution with discrete probability masses at zero. This is due to the underlying economic structure of the household decision process that generates the business outcomes, which has a discrete decision component (whether to operate a business or not) and a continuous decision component (how much to invest once operating). Understanding this aspect of the data will be crucial to both the within-site analysis and the aggregation procedures I present in this paper.

3 Methodology

3.1 General Approach

This section describes the general methodological framework of Bayesian hierarchical modeling for aggregation of evidence from multiple study sites. 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 set of K data sets contains all the evidence relevant to evaluating the impact of this intervention, 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.

The premise of aggregation is that there may exist some general parameter $\theta \in \Theta$ which is common across study sites at the population level. We can learn about this θ using the evidence at hand because it is related to the set $\{\theta_k\}_{k=1}^K$ in some way, but the nature of this relationship is typically not known. The key unknown variable in this relationship is the heterogeneity or dispersion of $\{\theta_k\}_{k=1}^K$ around θ, denoted Σ_θ. 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.[4] Hence, 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.

Estimation of θ and Σ_θ is the core challenge of aggregation. 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 indeed no general common parameter θ exists.[5] The heterogeneity of $\{\hat{\theta}_k\}_{k=1}^K$ may be a poor estimator of Σ_θ because it includes the sampling variation of each $\hat{\theta}_k$ around its θ_k. 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, here denoted $\bar{\theta}$. By contrast, 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

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

[5] In fact, they may be suboptimal even in this case, if K > 3. A proof of this is in Stein 1951, and further discussion is in Efron & Morris 1975.

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 is then denoted:

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

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 may appear restrictive, but in fact it nests all common meta-analytic techniques, including 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 endorse the full-pooling model. Alternatively, the model can estimate very large dispersion in $\{\theta_k\}_{k=1}^K$ such that in fact $\{\tilde{\theta}_k\}_{k=1}^K = \{\hat{\theta}_k\}_{k=1}^K$. This result would endorse the no-pooling model. For applications in economics, where it is reasonable to think that neither extreme is true, the model's real strength is that it may choose some interior point on the spectrum between these two extremes if that interior point 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" that occurs. Hence, this method is more robust than the full pooling or no pooling approaches, although it sacrifices some efficiency if in reality $\Sigma_\theta \in \{0, \infty\}$. Employing some flexible structure of this kind is the only way to estimate a general θ while simultaneously estimating Σ_θ.

While in principle the hierarchical model could be specified with a nonparametric likelihood, a parametric structure is typically preferable in low-data environments, such as evidence aggregation with $K < 50$. 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 can lack the power to deliver reliable inference at the general level. 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 $\sigma_\theta = 0$ respectively. The no-pooling and full-pooling models do not specify upper-level structure only because they impose 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.

In order to perform well, hierarchical models require that $\{\theta_k\}_{k=1}^K$ be "exchangeable", so that their joint distribution is invariant to permutation of the indices. This means the analyst must have no knowledge of the ordering of the treatment effects *a priori* that is not present in the model. Hence, if economic theory demands that a particular covariate should be correlated in a certain way with the treatment effects, we can require *conditional* exchangeability and build 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 any case, the hierarchical model can only be used to assess generalizability for the set of sites which are in fact exchangeable given the structure imposed. 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, a point also noted in the framework of Allcott (2015).

3.1.1 Pooling Metrics for Hierarchical Models

Hierarchical models also provide 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, the extent of pooling across study sites has a natural interpretation as a measure of generalizability. The magnitude of Σ_θ, or relatedly, the magnitude of the uncertainty interval on the predicted effect in the next site θ_{K+1}, provides a natural metric. 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, while it is important to report and interpret $\tilde{\Sigma}_\theta$ and the uncertainty on θ_{K+1}, it is also useful to examine pooling metrics whose magnitude is easily interpretable. Pooling metrics have only 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):

$$\omega(\theta_k) \equiv \frac{\hat{se}_k^2}{\tilde{\sigma}_\theta^2 + \hat{se}_k^2}. \tag{3.4}$$

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, $\omega(\theta_k) > 0.5$ indicates that $\tilde{\sigma}_\theta^2$ is smaller than the sampling variation, indicating substantial pooling of information and a "small" $\tilde{\sigma}_\theta^2$. 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. In that case, the extrapolation from θ_k to θ is more reliable than the signal of $\hat{\theta}_k$ for θ_k: a strong indicator of cross-study generalizability.

The fact that the $\omega(\theta_k)$ uses sampling variation as a comparison is both a strength and a weakness of the metric. In one sense this is exactly the right comparison: it scores how much we learned about site $K+1$ by analyzing data from site k against how much we learned about site k by analyzing data from site k, which is captured by the sampling variation in $\hat{\theta}_k$. Yet in another sense, if the sampling variation is very large or small due to an unusually small or large sample size or level of volatility or noise in the data, it may be beneficial to use an alternative pooling metric. Meager (2015) proposed the use of the following metric based on relative geometric proximity, defined as follows:

$$\breve{\omega}(\theta_k) \equiv \{\omega : \ \tilde{\theta}_k = \omega\tilde{\theta} + (1-\omega)\hat{\theta}_k\}. \tag{3.5}$$

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:

$$\lambda_\theta \equiv 1 - \frac{\frac{1}{K-1}\sum_{k=1}^{K}(E_{post}[\epsilon_k] - \overline{E_{post}[\epsilon_k]})^2}{E_{post}[\frac{1}{K-1}\sum_{k=1}^{K}(\epsilon_k - \bar{\epsilon_k})^2]}. \tag{3.6}$$

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.

3.2 Bayesian Implementation

While hierarchical models may be estimated in several ways, there are reasons to prefer Bayesian inference in practice. The major advantage of Bayesian methods is the 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. But 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.[6] 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 posterior:

$$f(\theta|\mathcal{Y}) = \frac{\mathcal{L}(\mathcal{Y}|\theta)\mathcal{P}(\theta)}{\int_\Theta \mathcal{L}(\mathcal{Y}|\theta)\mathcal{P}(\theta)d\theta}. \tag{3.7}$$

The joint posterior distribution $f(\theta|\mathcal{Y})$ characterizes all the information and uncertainty about all the unknown parameters conditional on the data. This is one reason why the tractability problems faced by the MLE method do not arise in Bayesian inference: the same object that generates the point estimate also provides the joint, conditional and marginal uncertainty intervals on all the unknowns. The specification of a proper prior distribution is essential to Bayesian inference, ensuring that $f(\theta|\mathcal{Y})$ is a proper probability distribution with desirable decision-theoretic properties such as admissibility, as described in Efron (1982) and 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 (for more detail of Doob's theorem and other relevant results, see Van der Vaart 1998).

In a low-data environment, specifying informative priors can substantially improve the performance of the hierarchical model. Priors increase the tractability and speed of the estimation by targeting regions of the parameter space that are more likely to contain relevant values. If the analyst only has vague knowledge of the location of this likely region, then the priors can be made quite diffuse (sometimes called "weakly informative"). If the analyst has access to 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: the prior regularizes the estimates. In low-data environments such as the cross-study level of the hierarchical model, overfitting

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

and high variance can be the major obstacle to making reasonable inferences or predictions. Here, as in many other statistical problems, regularization often improves performance (Hastie, Tibshirani and Friedman 2009, section 10.2).

Bayesian inference also provides a framework for decision-making about policy and future research, and a coherent conceptualization of the parameter θ_{K+1}. The distribution of the treatment effect in a hypothetical future site θ_{K+1} is the key object of interest for policymakers, and this distribution must be computed accounting for the full joint posterior uncertainty rather than conditioning on a particular point estimate or even a particular interval estimate. The Bayesian approach delivers the correct uncertainty interval in the form of posterior predictive inference (Gelman et al., 2004), which marginalizes out the posterior uncertainty on the unknowns (θ, Σ_θ). Formally, the posterior predictive distribution is:

$$f(\theta_{K+1}|\theta) = \int \psi(\theta_{K+1}|\theta) f(\theta|\mathcal{Y}) \mathrm{d}\theta \tag{3.8}$$

The Bayesian framework is well-suited to providing these objects because the entire goal of aggregating towards generalizable evidence itself is underpinned by Bayesian thinking: we seek to update our understanding of the unknown parameters in one location using the information about the parameters from other locations. On a conceptual level, if we wish to make policy decisions accounting for our uncertainty about any of these unknown parameters, the correct object to take expectations over is the posterior distribution of the unknown, not the sampling distribution of some estimator.

The Bayesian approach also offers a natural mechanism for implementing constraints on θ. 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:

$$\mathcal{L}_{\mathcal{A}_\Theta}(\mathcal{Y}|\theta) = \mathcal{L}(\mathcal{Y}|\theta) \cdot \mathbb{1}\{\theta \in \mathcal{A}_\Theta\}. \tag{3.9}$$

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. Recall that 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

$$p(\theta^*) = p_\theta(\mathrm{f}^{-1}(\theta))|det(J_{\mathrm{f}^{-1}}(\theta))|. \tag{3.10}$$

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 $f(\theta|\mathcal{Y}) \cdot \mathbb{1}\{\theta \in \mathcal{A}_\Theta\}$.

Where tractability issues arise in Bayesian inference, typically because the posterior distribution does not take a known or well-understood functional form, these have been effectively surmounted by the use of Markov Chain Monte Carlo (MCMC) methods. These methods con-

struct 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 are ideally suited to estimating hierarchical models (Betancourt and Girolami, 2013). HMC uses discretized Hamiltonian dynamics to sample from the posterior, which achieves excellent performance when combined with the No-U-Turn sampling method (NUTS) to auto-tune the step sizes in the chain (Gelman and Hoffman, 2011). This algorithm is straightforward to implement because it has been largely automated in the software package Stan, a free statistical library which calls C++ to fit Bayesian models from R or Python (Stan Development Team, 2014).

3.3 Dispersion Treatment Effects

I now turn to the specific modeling choices involved in my development of a model to aggregate the treatment effects of an intervention on the dispersion of a distribution, within the general framework of section **??**. In the case of microcredit, we have access to data on economic outcomes such as household business profit or consumption measured at the household level. Any particular scalar outcome is denoted y_{nk} for household n in site k. These outcomes may be continuous, discrete or mixture variables. Treatment is a binary indicator $T_{nk} \in \{0, 1\}$ throughout. Consider a decomposition of any household outcome y_{nk} into a control group mean μ_k and an additive treatment effect of microcredit τ_k. Similarly, decompose the standard deviation of y_{nk} into the control group's standard deviation and a treatment effect. To impose the constraint that standard deviation must be non-negative for each group at every level of the model, I specify these effects on the exponentiated scale rather than on the raw scale.[7] Hence, the standard deviation for a household n in site k with treatment status T_{nk} is:

$$\sigma_{y_k} = \exp(\eta_k + \gamma_k T_{nk}). \tag{3.11}$$

In this specification, γ_k captures the treatment effect on the standard deviation. If $\gamma_k = 0$, then there is no treatment effect on the variance. If $\gamma_k < 0$ then the standard deviation in the treatment group is reduced by a factor of $\exp(\gamma_k)$ relative to the control group standard deviation. If $\gamma_k > 0$ then the standard deviation in the treatment group is increased' by a factor of $\exp(\gamma_k)$. For example if $\gamma_k = 1$ then the treatment group standard deviation is 2.7 times the size of the control group standard deviation.

I propose the following hierarchical model to aggregate the effects on the mean and standard deviation of household outcomes. The lower level $f(\mathcal{Y}_k|\theta_k)$ describing the data's dependence on

[7] I thank Anna Mikusheva for her contribution to the development of this idea.

the local parameters, is:

$$y_{nk} \sim N(\mu_k + \tau_k T_{nk}, (\exp(\eta_k + \gamma_k T_{nk}))^2) \; \forall \; k. \tag{3.12}$$

This specifies a linear regression on the outcome's mean and on the log of its standard deviation. Estimating a model with this level alone would provide the same point estimates as a simple ordinary least squares regression, with standard errors adjusted for any difference in the standard deviation between the treatment and control groups.[8] Adding the upper level of the model then shrinks these site-level parameters together jointly towards the upper-level parameters, both allowing and estimating correlations between them. The upper level $\psi(\theta_k|\theta)$ for this model is:

$$\begin{pmatrix} \mu_k \\ \tau_k \\ \eta_k \\ \gamma_k \end{pmatrix} \sim N \left(\begin{pmatrix} \mu \\ \tau \\ \eta \\ \gamma \end{pmatrix}, V \right) \; \forall \; k \tag{3.13}$$

Together, equations **??** and **??** form the hierarchical likelihood. To perform Bayesian inference via the full joint posterior distribution, I use weakly informative priors $\mathcal{P}(\theta)$. I pursue the strategy from Lewandowski et al.(2009) of decomposing the variance-covariance matrix V on the upper level into a scale parameter ν and a variance-covariance matrix Ω. In this case however the ν parameter's prior needs to be split up in order to reflect the differing scales of these parameters: (μ, τ) are in USD PPP per fortnight, while (η, γ) are on the multiplicative exponential scale. These priors are diffuse except for the prior on Ω which pushes the posterior towards detecting independence across parameters. Because economic theory predicts two possible countervailing relationships between baseline wealth and the impact of microcredit - microcredit may have diminishing marginal returns, or perhaps it only works on relatively rich households, or both - with only 7 data points we should temper the conclusion in the data if it suggests an extreme correlation in either direction. The priors are:

$$\begin{aligned} \begin{pmatrix} \mu \\ \tau \\ \eta \\ \gamma \end{pmatrix} &\sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1000^2 & 0 & 0 & 0 \\ 0 & 1000^2 & 0 & 0 \\ 0 & 0 & 100^2 & 0 \\ 0 & 0 & 0 & 100^2 \end{bmatrix} \right) \\ V &\equiv diag(\nu)\Omega diag(\nu) \\ \text{where } \nu[1,2] &\sim Cauchy(0, 50) \\ \nu[3,4] &\sim Cauchy(0, 5) \\ \Omega &\sim \text{LKJcorr}(3) \end{aligned} \tag{3.14}$$

The results of this model are sensitive to the prior on Ω, as pointed out in Giordano et al.(2016), precisely because there is so little cross-sectional data at the upper level. Therefore,

[8] This is not the same as the White or Eicker-Huber-White generalized correction for heteroskedasticity. It has more in common with the Welch adjustment to the t-test under the Behrens-Fisher problem (which is the problem that arises if $\gamma_k \neq 0$).

as a robustness check, I also fit an alternative model with an "independent" specification, which does not display sensitivity to the upper level variance priors. While this restrictive functional form cannot exploit correlations which are very likely to exist, its resulting lack of sensitivity makes this model a useful check against researcher degrees of freedom. The derivation and results from this independent version of the model are provided in Appendix **??**.

Standard deviation is not the only metric of dispersion relevant to household outcomes. In fact, standard deviation can be an unreliable or unstable measure of spread in fat-tailed distributions; in cases with extremely high kurtosis, the standard deviation may not even exist in the underlying population distribution. A more robust metric of dispersion is the mean absolute deviation (MAD) of the outcome values from their mean, or from their median value (Fama 1965, Pham-Gia and Hung 2001). Therefore, I propose a hierarchical model to jointly aggregate the results on the MAD and the mean for a given household outcome. Because it can be challenging or even analytically impossible to specify an outcome distribution entirely as a function of its mean and MAD, I propose a model which takes in as data the no-pooling estimates of these parameters and their standard errors $\{\hat{\theta}, \hat{se}_k\}_{k=1}^K$, in the tradition of Rubin (1981).

The following model works for any metric of dispersion, but for my application I consider the mean absolute deviations from the sample mean, defined

$$MAD(\mathcal{Y}_k) \equiv \frac{1}{N_k} \sum_{n=1}^{N_k} |y_{nk} - \bar{y}_k|. \tag{3.15}$$

I split the MAD for any given outcome in site k into a control group MAD, defined by $\exp(\Delta_k)$, and a treatment group MAD defined by $\exp(\Delta_k + \Gamma_k)$. These may be estimated using any consistent and asymptotically Normal no-pooling estimator of choice. For this application I use frequentist plug-in estimators (i.e. the analogous sample statistics) and nonparametrically bootstrapped standard errors. This generates the objects $\{\hat{\Delta}_k, \hat{\Gamma}_k, \hat{se}_\Delta, \hat{se}_\Gamma\}_{k=1}^K$. Because the model should adjust the uncertainty on the average treatment effects for the detected effects on the MAD, the no-pooling estimates on the mean $\{\hat{\mu}_k, \hat{\tau}_k, \hat{se}_\mu, \hat{se}_\tau\}_{k=1}^K$ should also be computed and incorporated into the model as data. To do this, I propose the following model. The lower level now describes the dependency of $\hat{\theta}_k$ on θ_k, so $f(\mathcal{Y}_k|\theta_k) = f(\hat{\theta}_k|\theta_k)$ for this case as follows:

$$\begin{aligned}
\hat{\tau}_k &\sim N(\tau_k, \hat{se}_\tau^2) \; \forall \; k \\
\hat{\mu}_k &\sim N(\mu_k, \hat{se}_\mu^2) \; \forall \; k \\
\hat{\Delta}_k &\sim N(\Delta_k, \hat{se}_\Delta^2) \; \forall \; k \\
\hat{\Gamma}_k &\sim N(\Gamma_k, \hat{se}_\Gamma^2) \; \forall \; k.
\end{aligned} \tag{3.16}$$

The upper level of the model is conceptually identical to the full data case, and describes the

relationship $\psi(\theta_k|\theta)$ as follows:

$$\begin{pmatrix} \mu_k \\ \tau_k \\ \Gamma_k \\ \Delta_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \tau \\ \Delta \\ \Gamma \end{pmatrix}, V\right) \ \forall\ k \tag{3.17}$$

To complete this model, I use the same priors as specified in equations **??**. In addition, the pooling metrics developed for average treatment effects $\{\tau_k\}_{k=1}^K$ can be directly applied to the dispersion effects $\{\gamma_k\}_{k=1}^K$ or $\{\Gamma_k\}_{k=1}^K$. This is possible because all the models above specify the effect on the dispersion using a single scalar parameter.

3.4 Nonparametric Quantile Treatment Effects

I now discuss the specific modeling choices involved in the construction of a method to aggregate sets of quantile treatment effects and assess their generalizability. The uth quantile of some outcome is the value of the inverse CDF at u:

$$Q_Y(u) = F_Y^{-1}(u). \tag{3.18}$$

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

$$Q_{y_{nk}|T}(u) = \beta_{0k}(u) + \beta_{1k}(u)T_{nk} \tag{3.19}$$

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 Hassan (2014) and Reich et al.(2011). But in the microcredit data, researchers estimated a set of 10 quantiles $\mathcal{U} = \{0.05, 0.1, 0.15, ..., 0.95\}$ and interpolated the results to form a "quantile difference curve". This curve is constructed by computing the quantile regression at all points of interest:

$$Q_{y_{ik}|T} = \{Q_{y_{ik}|T}(u) = \beta_{0k}(u) + \beta_{1k}(u)T_{ik} \ \ \forall\ u \in \mathcal{U}\} \tag{3.20}$$

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 10 quantile effects:

$$\begin{aligned} \beta_{0k} &= (\beta_{0k}(0.05), \beta_{0k}(0.15), ...\beta_{0k}(0.95)) \\ \beta_{1k} &= (\beta_{1k}(0.05), \beta_{1k}(0.15), ...\beta_{1k}(0.95)) \end{aligned} \tag{3.21}$$

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:

$$\begin{aligned} Q_{0k} &= \{Q_{0k}(u) \ \ \forall \ u \in \mathcal{U}\} \\ Q_{1k} &= \{Q_{1k}(u) \ \ \forall \ u \in \mathcal{U}\}. \end{aligned} \tag{3.22}$$

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

$$\begin{aligned} \beta_{0k} &= Q_{0k} \\ \beta_{1k} &= Q_{1k} - Q_{0k}. \end{aligned} \tag{3.23}$$

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. For binary treatment variables, the no-pooling estimators necessarily satisfy monotonicity, but partial pooling may introduce crossing where none existed. Yet even if quantile crossing does not arise, neighboring quantiles contain information about each other, and using that information can improve the estimation and reduce posterior uncertainty. 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 quantile difference curves which builds on a classical result from Mosteller (1946) about the joint distribution of sets of empirical quantiles. The theorem states that if the underlying random variable is continuously distributed, then the asymptotic sampling distribution of a vector of empirical quantiles is 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 multivariate Normal. The theorem therefore offers a foundation for a hierarchical quantile treatment effect aggregation model using multivariate Normals. The resulting analysis is nonparametric at the data level, as it is applicable to any continuous distribution as long as there is sufficient data in each of the studies.

For this model, the data are the vectors of sample quantile differences $\{\hat{\beta}_{1k}\}_{k=1}^K$ and their sampling variance-covariance matrices $\{\hat{\Xi}_{\beta_1 k}\}_{k=1}^K$. Thus, the lower level $f(\mathcal{Y}_k|\theta_k) = f(\beta_{1k}|\beta_1 k)$ is given by the expression:

$$\hat{\beta}_{1k} \sim N(\beta_{1k}, \hat{\Xi}_{\beta_1 k}) \ \forall \ k \tag{3.24}$$

At the upper level of the model, a Normal specification offers tractability and has generally desirable properties (Efron and Morris, 1976). The upper level of the model $\psi(\theta_k|\theta)$ is therefore:

$$\beta_{1k} \sim N(\beta_1, \Sigma_1) \ \forall \ k. \tag{3.25}$$

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

$$\beta_0^*[u] = \begin{cases} \beta_0[u], & \text{if } u = 1 \\ log(\beta_0[u] - \beta_0[u-1]) & \text{if } 1 < u < |\mathcal{U}| \end{cases} \tag{3.26}$$

Any vector β_0 to which this transform is applied and inference 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 above with a superscript m. Thus, by applying the transform above, I work with β_0^m rather than an unconstrained β_0, to ensure monotonicity.

Employing a monotonizing transform is an appealing alternative to other methods used in the econometrics literature to ensure monotonicity during quantile regression. This transformation enforces the constraint in a flexible and adaptive manner, passing more information across quantiles in cases where the draws from the posterior are close to violating the constraint. 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. By contrast, 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. 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.

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 their sampling variance-covariance matrices $\{\hat{\Xi}_{1k}, \hat{\Xi}_{0k}\}_{k=1}^K$

as data. The lower level $f(\mathcal{Y}_k|\theta_k)$ is:

$$\begin{aligned} \hat{Q}_{0k} &\sim N(\beta_{0k}^m, \hat{\Xi}_{0k}) \ \ \forall\, k \\ \hat{Q}_{1k} &\sim N(Q_{1k}^m, \hat{\Xi}_{1k}) \ \ \forall\, k \\ &\text{where} \ \ Q_{1k} \equiv \beta_{0k}^m + \beta_{1k} \end{aligned} \tag{3.27}$$

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

$$\begin{aligned} \beta_{0k}^m &\sim N(\beta_0^m, \Sigma_0) \ \ \forall\, k \\ \beta_{1k} &\sim N(\beta_1, \Sigma_1) \ \ \forall\, k \\ &\text{where} \ \ \beta_1 \equiv Q_1^m - \beta_0^m \end{aligned} \tag{3.28}$$

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

$$\begin{aligned} \beta_0^m &\sim N(0, 1000 * I_{10}) \\ \beta_1 &\sim N(0, 1000 * I_{10}) \\ \Sigma_0 &\equiv diag(\nu_0)\Omega_0 diag(\nu_0)' \\ \Sigma_1 &\equiv diag(\nu_1)\Omega_1 diag(\nu_1)' \\ \text{where} \ \ \nu_0, \nu_1 &\sim \text{halfCauchy}(0, 20) \text{ and } \Omega_0, \Omega_1 \sim LKJCorr(1). \end{aligned} \tag{3.29}$$

This model can be modified to take in the empirical quantile treatment effects $\{\hat{\beta}_{1k}\}_{k=1}^K$ and their standard errors instead of $\{\hat{Q}_{1k}\}$. The current 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 model here permits completely arbitrary functional form on (Σ, Σ_0), although they are logically required to be positive semi-definite. 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 modeled as draws from Gaussian Processes.[9] 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 strength of this approach 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 did not operate businesses 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 (figure **??**). This spike undermines the performance of the Mosteller theorem and of the nonparametric bootstrap for standard error calculation. The Mexico data provides the cleanest example of this, shown in figure **??**: 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;

[9] 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.

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).[10] But in the microcredit data, the complications caused by these spikes at zero are not effectively addressed by dithering. The results in figure **??** 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. This situation arises because business variables are generated by a partially discrete process: first, a household has a binary choice of whether to operate a business or not, and second, if the business exists it has some continuous expenditures, revenues and profit. Hence, 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.

3.4.1 Pooling Metrics for Nonparametric Quantile Treatment Effects

Conventional pooling metrics for hierarchical models are designed to be applied to univariate treatment effects. Hence, while these metrics were appropriate to apply to the variance effects, and could perhaps be applied pointwise to the quantile results, it would be ideal to have pooling metrics on the entire set of quantiles. 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.[11] 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, because $\text{tr}(\Sigma_1)$ is equivalent to the Frobenius norm of $\Sigma_1^{1/2}$, it is submultiplicative and unitarily invariant.

[10] 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.

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

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:

$$\begin{aligned} \boldsymbol{\omega}(\beta) &= \frac{1}{K}\sum_{k=1}^{K} \frac{\text{tr}(\hat{\Xi}_{\beta_1 k})}{\text{tr}(\hat{\Xi}_{\beta_1 k} + \Sigma)} \\ &= \frac{1}{K}\sum_{k=1}^{K} \frac{\text{tr}(\hat{\Xi}_{\beta_1 k})}{\text{tr}(\hat{\Xi}_{\beta_1 k}) + \text{tr}(\Sigma)} \end{aligned} \tag{3.30}$$

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 $\{\beta_{1k}\}_{k=1}^{K}$. The simplest generalization of this to multivariate treatment effects is to simply take the sum of this metric evaluated at each quantile treatment effect; this is exactly what the trace did for the conventional pooling metric. To ensure the metric retains an easily interpretable scale, 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:

$$\boldsymbol{\lambda}_{\boldsymbol{\beta_1}} = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{1}{U}\sum_{u=1}^{U}\frac{var(E[\beta_{1k}[u] - \beta_1[u]])}{E[(var(\beta_{1k}[u] - \beta_1[u])]}\right). \tag{3.31}$$

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:

$$\breve{\boldsymbol{\omega}}(\boldsymbol{\beta_1}) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{U}\sum_{u=1}^{U}\frac{\beta[u]_{1k} - \hat{\beta}_{1k}[u]}{\beta_1[u] - \beta_{1k}[u]}\right). \tag{3.32}$$

3.5 Parametric Quantile Treatment Effects

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. Hence, it may be possible to explicitly model the processes that generate the probability density functions (PDFs) of household outcomes. I propose a flexible parametric approach, which exploits the researcher's knowledge of the shape or sensible parameterization of a potentially complex mixture distribution.[12] While this requires substantial input from the researcher and the specific aggregation model must be tailored to each specific case, this method will automatically satisfy quantile function monotonicity since it directly models the PDFs as proper densities. This approach

[12] I do not use nonparametric mixtures of Gaussians because it is unclear how to apply a hierarchical model to these infinite-dimensional PDFs.

also transfers information across quantiles because they are linked together by the parametric functional form assumptions.

For the specific household business variables in the microcredit data, there is sufficient contextual economic information to build a parametric model. These variables are the output of a partially discrete decision process: first, a household has a binary choice of whether to operate a business or not, and then those households who operate businesses manifest 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 artifact of the data collection. Economic theory and prior research suggest that the continuous portions of business variables such as revenues and profit follow power laws, and can be modeled using Pareto distributions (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 effect of the treatment on each of the parameters that govern the distribution, as well as the implied quantile treatment effects.

I propose 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 $j = 1, 2, 3$. I model this dependence 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. I proceed using a Pareto distribution for the tails, in each case governed by a location parameter which controls the lower bound of the support and a scale parameter which controls the thickness of the tail. The location parameter ι_{jk} is exactly known because I externally defined the domain of each of the components by manually splitting the data. However the shape parameter is unknown and may be affected by treatment, which I model using a multiplicative exponential regression specification to impose a non-negativity constraint on the parameter. The shape parameter in mixture component j for household n in site k is therefore $\exp(\rho_{jk} + \kappa_{jk}T_{nk})$.

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:

$$\begin{aligned} y_{nk}|T_{nk} \sim\ & \Lambda_{1k}(T_{nk})Pareto(-y_{nk}|\iota_{1k}, \exp(\rho_{1k} + \kappa_{1k}T_{nk}) \\ & +\Lambda_{2k}(T_n)\delta_{(0)} \\ & +\Lambda_{3k}(T_n)Pareto(y_{nk}|\iota_{3k}, \exp(\rho_{3k} + \kappa_{3k}T_{nk}) \ \forall\ k \\ \text{where}\ \ \Lambda_{jk}(T_{nk}) =\ & \frac{\exp(\alpha_{jk} + \pi_{jk}T_{nk})}{\sum_{j=1,2,3}\exp(\alpha_{jk} + \pi_{jk}T_{nk})} \end{aligned} \tag{3.33}$$

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

$$(\kappa_{1k}, \kappa_{3k}, \pi_{1k}...)' \equiv \zeta_k \sim N(\zeta, \Upsilon) \; \forall \; k \tag{3.34}$$

For tractability and simplicity I enforce diagonal Υ for this paper. Therefore, the model needs only weak priors $\mathcal{P}(\theta)$ as follows:

$$\begin{aligned} \zeta &\sim N(0, 10) \\ \Upsilon &\equiv diag(\nu_\Upsilon)\Omega_\Upsilon diag(\nu_\Upsilon)' \\ \nu_\Upsilon &\sim \text{halfCauchy}(0, 5) \\ \Omega_\Upsilon &= I_{|\zeta|} \\ \alpha_{mk} &\sim N(0, 5) \end{aligned} \tag{3.35}$$

The tailored hierarchical PDF aggregation models for revenues and expenditures are constructed as above, but with no negative tail and hence only 2 mixture components. This set of treatment effects ζ has the advantage of decomposing the quantile effects into two mechanisms, each with an economic interpretation. The general κ_j captures the effect of microcredit on the extensive margin: the allocation of households to the general category of making no business profit, making positive profit or making negative profit, as indicated by the jth component. The general π_j captures the effect of microcredit on the intensive margin: the general distribution of profits for those households which do operate businesses, and hence populate the jth component's continuous tail. This decomposition is estimated within each site and at the aggregate level, and this model thereby estimates the generalizability of these two different channels.

Despite the economic insight that these specific parameters provide, it is still useful to recover the implied quantile treatment effects from this model. This is a nontrivial challenge because mixture distributions in general do not have analytical forms for their quantile functions. However, because the mixture distribution in this model has components with disjoint supports, I 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, and get the following result:

$$\begin{aligned} Q(u) = &-\text{Pareto}^{-1}\left(\frac{u}{\Lambda_1(T_n)} \mid \iota_{1k}, \rho_{1k}(\exp(\kappa_{1k}T_n))\right) * \mathbb{1}\{u < \Lambda_1(T_n)\} \\ &+ 0 * \mathbb{1}\{\Lambda_1(T_n) < u < (\Lambda_1(T_n) + \Lambda_2(T_n)\} \\ &+ \text{Pareto}^{-1}\left(\frac{u - (1 - \Lambda_3(T_n))}{\Lambda_3(T_n)} \mid \iota_{3k}, \rho_{3k}(\exp(\kappa_{3k}T_n)\right) * \mathbb{1}\{u > (1 - \Lambda_3(T_n))\} \end{aligned} \tag{3.36}$$

As this is a function of the existing unknown parameters, the full posterior distribution of the entire set of quantiles and the implied quantile treatment effects is easily computed within the Bayesian framework. 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.

3.5.1 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 the purest metric of that dispersion, and as discussed above, the trace of the matrix is the norm that captures the 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 modeling 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

$$\begin{aligned} \boldsymbol{\omega}_V(\beta) &= \frac{1}{K}\sum_{k=1}^{K}\frac{\text{tr}(\hat{\Sigma}_{\zeta k})}{\text{tr}(\hat{\Sigma}_{\zeta k}+V)} \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{\text{tr}(\hat{\Sigma}_{\zeta k})}{\text{tr}(\hat{\Sigma}_{\zeta k})+\text{tr}(V)}. \end{aligned} \tag{3.37}$$

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.

4 Results

4.1 Dispersion Treatment Effects Results

The results of fitting the dispersion models to the 6 household outcomes in the microcredit data show some evidence for a generalizable increase in the dispersion, particularly in the household business outcomes. Yet the findings differ substantially across the different dispersion metrics. The more robust metric, the effect on the MAD (Γ), shows on average a 15% increase in the dispersion on the household business outcomes but no conclusive movement on the consumption outcomes (see table **??** for full results). The less robust metric, the effect on the standard deviation (γ), shows much larger point estimates with an average increase of 40%, but the posterior intervals on γ are much wider than the intervals on Γ, and always include zero (see table **??**). The difference is most salient for household business outcomes, which show evidence of a small but generalizable increase in the MAD, and evidence of a potentially large yet non-generalizable increase in the standard deviation. In all cases the full-pooling aggregation is shown to severely underestimate the uncertainty in comparison; imposing the full-pooling assumption can be highly misleading in cases where it is not warranted.

This pattern is confirmed by examining the local effects on each metric for each site: there is essentially zero shrinkage across sites for the standard deviation, but there is moderate shrinkage

on the MAD effects (see the figures in Appendix **??**). Many of the local effects on the standard deviation are large, even more than 100% in some cases, but they do not aggregate to any generalizable information. It may seem incongruous that in the case of profit, 6 out of 7 γ_k effects are large and precisely estimated and yet the aggregate γ for profit is imprecisely estimated. But that is exactly what it means for a result to lack generalizability: the effects are so heterogeneous that the model cannot infer that the effect in the next site will be similar to any one of them nor to their average value. By contrast the site-specific effects on the MAD are smaller and closer together, providing strong evidence of a moderate but generalizable increase in the dispersion of business outcomes and weak evidence of a general increase in dispersion of consumption outcomes.

These models also produce new results on the average treatment effects which adjust the inference for the effects on the dispersion, which in this case substantially revises the treatment effects downwards towards zero. This is shown in table **??**, which compares the results on the location effect from the joint location/MAD model in equations **??** and **??** to the results in Meager (2015) which did not correct for any dispersion effects, and to the full pooling aggregation. The results suggest that some of the upper tails of the posterior distributions of the average effects in Meager (2015) were due to increases in dispersion that were misattributed to changes in the mean. But overall the new results strengthen the conclusion of Meager (2015), suggesting that the average effect of microcredit is smaller than previously estimated, and in general may be zero or close to it. The new results also have tighter posterior intervals, indicating the model performs at least as much pooling as the model in Meager (2015), and thus that the results on the average household outcomes are reasonably generalizable.

4.1.1 Pooling Metrics for Dispersion Treatment Effects

Examining the three pooling metrics for the two metrics of dispersion effects confirms that the MAD effects exhibit substantial generalizability, while the standard deviation effects exhibit virtually zero generalizability. The results for the effect on the MAD (Γ) are shown in table **??** with the pooling results on the average MAD in the control group (Δ) shown for comparison. The model displays substantial pooling on Γ, around 60% averaged across all three metrics, but little pooling on Δ with an average of 10% across all metrics. The detected similarity in the Γ_ks across sites is therefore not due to similar baseline dispersion across sites: it is the mechanism, not the context, which appears to be similar here. As expected, however, the results for the variance tell a different story: all pooling metrics for both the control group's standard deviation (η) and the effect (γ) are less than 5% (see table **??**). The Bayesian hierarchical model effectively selects the no-pooling model on the variances, but chooses substantial pooling on the MAD, confirming the results of section **??**.

This result is reflected in the relatively tight 50% and 95% posterior predictive intervals on the distribution of Γ_{K+1} relative to γ_{K+1}, which are in both cases the forecasted results of the hypothetical next experiment. These intervals are shown in figure **??**. Although the posterior predictive intervals should be larger than the posterior intervals on Γ or γ because $\tilde{\Sigma}_\theta \neq 0$, in this case the intervals on Γ_{K+1} are more than twice as precise as the intervals on γ_{K+1}. For

example, there is a 25% chance that $\gamma_{K+1} < 0$ on profit, and a 25% chance of an effect of 1 or larger, which would create a 300% increase in the dispersion of profit across households relative to the control group. By contrast, the posterior predictive inference on Γ_{K+1} displays more than a 50% chance of seeing a result between 0 and 30% on most outcomes. In all cases, the full-pooling aggregation results underestimate the uncertainty by several orders of magnitude, and are thus inappropriate tools for the prediction of θ_{K+1}.

Interpreting these results together is challenging because the two metrics of dispersion provide different conclusions about the magnitude of the effect and its generalizability, particularly for the business outcomes. While both of the dispersion metrics display more evidence of a real impact on the outcomes than the mean treatment effects did, only the MAD shows similar generalizability to the means. Moreover, while the MAD is more robust in general, it is not immediately clear why the variance metric results should be so different; this may indicate an issue with the modeling assumptions underlying the computation of the variance, or it may be that the two metrics are simply illuminating different aspects of the data. As it turns out, the results of the quantile aggregation will be able to illuminate the origin of these differences.

4.2 Nonparametric Quantile Treatment Effects Results

Aggregating the quantile treatment effects for household consumption, consumer durables spending and temptation goods spending shows that microcredit generally and precisely does not affect outcomes below the 75% percentile, and above this point the prediction exhibits high variance at the general level. Figure **??** shows the posterior distribution of the generalized quantile treatment effects β_1 for each of these 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 show that the full pooling model and the BHM typically produce similar output for the 5th to 75th quantile of household outcomes, but sharply diverge in the upper tail. It is generally true that the effect of expanding access to microcredit is close to zero for quantiles below 0.75, and while larger effects may be possible in some sites at higher quantiles, the results are much noisier in the tail at the general level. The full-pooling model results typically underestimate the uncertainty, particularly on the upper quantiles, and thus indicate more 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. Figures in Appendix **??** 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 pattern suggested by the general results at the upper level, that in general there is a common effect of essentially zero on the shape of the distribution, except at the upper tail where there is both more uncertainty within each site and less apparent commonality across sites.

4.2.1 Pooling Metrics for Nonparametric Quantile Treatment Effects

The results of calculating the pooling metrics for the multivariate quantile models show that the level of pooling on the quantile difference curves is intermediate, around 50% on average. Results for the three consumption variables are shown in table **??**. The level of pooling on the control group quantiles is lower on average: nearly zero according to two of the three metrics, and intermediate according to the third metric. The posterior predictive distributions of these consumption variables are shown in figure **??** with the full-pooling model for comparison. The results show substantially 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. 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.

These results lie between the slightly stronger 60% pooling detected on the mean and MAD treatment effects, and the lack of pooling on the variance effects. This pattern further confirms that the quantile curve effects model is capturing both the effects on the mean and on the variance. The effects on the mean are close to zero but very generalizable, because for most of the consumption outcome quantiles, the quantile treatment effect is precisely zero and very generalizable. The effects on the variance are large and not generalizable because the effect on the quantiles in the right tail is often large but heterogeneous across sites. The shifting of the tail affects the posterior inference on the mean treatment effects to some degree, pulling the effect away from zero, but it affects the variance to a much larger degree because extreme values enter the variance squared. The MAD does not square its components, so extreme values have less influence; this is part of what makes this metric robust to issues such as high kurtosis.

4.3 Parametric Quantile Treatment Effects Results

The results of applying the tailored hierarchical PDF model to the household business profit, business expenditures and business revenues show that microcredit has a generalizable and precise zero impact below the 80th quantile, and above this point no prediction is possible. The economic structure of the model permits the decomposition of any detected impact into an effect on the extensive margin ("category switching") and the intensive margin ("tail expansion") but in this case neither can be conclusively detected. Figure **??** show the effects of expanding access to microcredit on the probability of component assignment on the logit scale, from both the hierarchical model and the full-pooling model. A negative coefficient means that treatment makes households more likely to be assigned to the positive tail than to the spike at zero or the negative tail; a positive coefficient shows the opposite. The model finds only weak support for the idea that microcredit increases the proportion of households in the positive tail relative to the spike at zero, while the full-pooling model is overly confident in the effect precision.

There is even less evidence of any change in the shapes of the Pareto distributions that govern the continuous tails of profit, revenues and expenditures. Figure **??** shows the effect of microcredit on the shape parameter of the Pareto distributions for all relevant tails for the

business outcomes. These results suggest substantial pooling, and indeed there is virtually zero cross-site heterogeneity in the treatment effects (see Appendix **??**). Moreover, the control group's tail shape parameters all indicate extremely fat tails: the tail parameters are smaller than 2, implying that within the model the variance is estimated to be infinite (see Appendix **??**). As a result, there is little room in the parameter space to indicate any expansion of the tails in the treatment group. This is the first indication that these distributions have extremely heavy tails, which suggests that the results on the mean and variance treatment effects are unreliable for the household business variables, both because they relied on Gaussian asymptotics which do not hold for distributions with such heavy tails, and because in this case the mean and variance are not reliable as summary statistics of the underlying distribution.

The result that microcredit has a generalizable and precise zero impact below the 80th quantile with no generalizable prediction beyond this point is derived using the Castellacci (2012) formula. I apply the formula to every draw from the joint posterior, in effect drawing joint posterior quantiles from the model itself. In this instance, the translation is somewhat challenging because the Pareto tails of this model are estimated to have shape parameters that imply infinite variation in the distribution for each outcome in both treatment and control groups (i.e. they lie between 0 and 2). This is not an uncommon feature of profit data, and the study of such heavy tails can be found as early as Fama (1963, 1965) and as recently as Gabaix (2008) and Pancost (2016). As a result, when I compute the posterior quantiles and the posterior predicted quantile differences, the standard errors at the tails are 10-15 orders of magnitude larger than those on the rest of the quantiles (see Appendix **??**). This extreme sparsity may be partially responsible for the model's refusal to provide a prediction in the upper tail at the general level, although it may also be due to heterogeneity across sites in these tails.

The quantile treatment effect results recovered from the tailored hierarchical PDF models for all outcomes are shown in figure **??** with the full pooling results for comparison. The results demonstrate a major difference in the uncertainty at the quantiles below the 80th percentile and those above it for all outcomes. Because the upper tail of all business variables is so sparse and the tails are extremely heavy, the model produces no prediction at the upper quantiles, with 95% posterior intervals many orders of magnitude larger than the uncertainty intervals at the median. The hierarchical model refuses to draw inference on these quantiles from the data, and thus communicates the vast uncertainty we should have about these upper tail effects. By contrast, the full pooling models are misleadingly precise and overconfident in predicting extremely large "statistically significant" general effects in the tails. The difference is dramatic because when the tails are sparse, a little more pooling goes a long way, yet the full pooling assumption is likely to be false and so these conclusions are likely misleading. These apparently "statistically significant" results in the upper tails "detected" in the full pooling model are eliminated by the application of a hierarchical model.

4.3.1 Pooling Metrics for Parametric Quantile Treatment Effects

Assessing 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. 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 **??** and the tail shape effects are shown in table **??**. 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. However, there is noticeable dispersion in the results across each of the metrics, which suggests that the results should be interpreted with caution. Nevertheless there is a reasonable amount of commonality across sites, suggesting that these results are at least partially generalizable to other sites.

The posterior predictive distributions of the parameters also express the extent of this heterogeneity by quantifying the uncertainty about the treatment effects in future study sites. Figure **??** shows the posterior predicted distributions for the treatment effects for the categorical logit and the tail shape parameters respectively. The posterior predicted intervals are substantially wider than the posterior intervals from the full-pooling model, displayed for comparison. There is a 75% probability of the next effect inducing a shift in the proportion of households who have positive outcomes for expenditures and revenues, and a 70% chance of inducing a shift in the proportion of households who have positive profits relative to those who make no profit. Yet the converse 25% and 30% respective chances of movement in the other direction represent a tangible risk which should not be disregarded. Overall, however, there is no generalizable prediction of any strong impact on these parameters in future locations.

To understand what these partial pooling results imply for the quantile treatment effects of these variables, it is necessary to recover these from the posterior predicted parameters of the tailored hierarchical PDF models. The posterior predicted quantile results, again computed using the Castellacci (2012) formula, are shown in figure **??** with the full pooling results for comparison. These results show that 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 can be no reliable inference for the next site's quantile effects. The magnitude of the uncertainty intervals on the 95th quantile treatment effects communicate the model's refusal to infer the effect on the tails of these distributions. As before, the full pooling model is highly misleading, displaying unwarranted precision and confidently predicting a positive effect.

The quantile treatment effect results for household business outcomes at the group level shed light on the patterns detected in the results on the means, MAD and variance effects. The distributional treatment effects of microcredit on these outcomes are essentially zero and highly generalizable below the 80th percentile. While the effects might be very large in the upper tail, the extremely sparse tails prevent any generalization above this point in a partial pooling context. The moderate degree of pooling and generalizability detected in the model overall is attributable to the strong pooling on the quantile treatment effects below the 80th quantile. This is also what likely drives the substantial pooling on the mean and MAD. The no-pooling result on the variance is now of dubious importance given that the variances of these distributions are unlikely to provide a reliable or stable characterization of the dispersion.

4.4 Discussion and Policy Implications

The most straightforward interpretation of distributional treatment effects is at the group level, such that the results describe the ways in which expanding access to microcredit changes the cross-sectional distribution of household outcomes. Transporting distributional treatment effects from the group level to the individual level requires much stronger assumptions than those required for average affects, because there is no law of iterated expectations that applies to quantiles or variances. The quantile treatment effect is not the treatment effect for the household at that quantile unless a rank-invariance assumption holds. Such an assumption seems inappropriate for the microcredit setting where there are likely to be heterogeneous effects that do not preserve the households' rank ordering. In any case, causal effects at the group level are still useful objects to estimate and understand for policy purposes, as the shapes of the distributions of these consumption and business outcomes are likely to have welfare implications.

The moderate yet generalizable increase in dispersion in business outcomes across households detected by the MAD model suggests that expanding access to microcredit causes ex-post inequality to increase. Combined with the small or even zero effect observed on the means, this suggests that microcredit could be socially harmful if economic inequality has deleterious impacts on the political or economic system. However, caution is required when interpreting these results: this could equally be produced by households using microcredit to experiment with their business strategies, increasing the volatility of outcomes in the short run but increasing the mean in the longer run. These results could also be generated by households taking on increased risk in their business strategies, which would increase the volatility of their consumption outcomes. Such behavior would contradict the claim that households use microcredit products to smooth consumption and investment; this claim seems unlikely to be generally true given these results.

The results from the quantile treatment effect aggregation allow a more detailed exploration and economic interpretation of the dispersion results. Although the aggregation procedures failed to find any strong evidence of generalizable effects at any quantile, the patterns of uncertainty are not uniform across quantiles. Nor is the evidence uniform across the two mechanisms illuminated in the tailored hierarchical PDF models: there is some weak evidence that microcredit could affect households on the extensive margin, moving them from the spike at zero into the positive tail of profits and revenues, whereas there is no evidence of any tail expansion due to any impact on the intensive margin. If indeed there is some small movement of households from zero profits to positive profits, this could explain the increase in variance through a channel that is generally considered welfare enhancing, although the inequality concerns would still apply. Overall, while the evidence for either economic mechanism is insufficient for policy purposes, further studies on this question - perhaps using methodologies that can target individual treatment effects - could be useful for both researchers and policymakers.

The quantile results have their own economic interpretation: groups of households who receive random expansions in their access to microcredit are generally the same as groups which do not receive this access below the 75th percentile of outcomes, and above this point they may differ, but no general prediction is possible. The tailored hierarchical PDF model detects extremely fat tails in the business outcomes, and the model therefore declines to make any inference at

all in these upper tails. Household consumption outcomes do not seem to have such fat tails, but the prediction on the tail effects to the general case still does not promote any conclusion of a positive impact above the 75th quantile. The precise and generalizable zero effect detected along most of the distribution is not simply a reflection of the presence of the spike at zero, which typically accounts for 15-50% of the distribution, not 75%. The pattern of zero impact along most of the distribution and a noisy, inconclusive result in the upper tail is found in every outcome regardless of the methodology used.

Hence, policymakers considering the likely impact of microcredit in their contexts cannot rely on the prospect of reproducing the positive upper tail effects found in some RCTs: there is no evidence that these effects are generalizable. Yet there is strong and generalizable evidence of zero impact on most of the outcome distribution at the group level, including at the lower tail. This finding effectively rebuts any claim that microcredit generally leads to worse outcomes at the group level. These results leave open the question of whether microcredit may lead to better outcomes in some sites, although such effects are not generalizable and cannot be reliably expected in any future setting. Thus, these results suggest that microcredit interventions are unlikely to be creating winners and losers on a large scale, despite the predictions in the applied theoretical literature (Kaboski and Townsend 2011, Buera, Kaboski and Shin 2012, Buera and Shin 2013). However, microcredit could still have positive and negative effects on some individual households, which quantile regression would not detect as long as they are roughly symmetric and their outcomes remain inside the general support of the control group's distribution. Yet the evidence does show that if such individual effects do exist, they are not large enough in magnitude that they reliably lead to outcomes far outside of those seen in the control group.

The presence of extremely heavy tails in the household business outcomes has several important methodological and economic implications. These tails are so heavy as to impede the functioning of the central limit theorem on this data, so the previous analysis of average treatment effects and variance treatment effects is likely to be inaccurate. For these variables, the ordinary least squares regressions performed in the original randomized controlled trials are likely to perform poorly compared to quantile regression techniques or parametric modelling of the power law. The finding of extremely heavy tails is reasonably common in economics when it is tested (e.g. Bazzi 2016, Pancost 2016, Gabaix 2008, Fama 1965), which suggests that the widespread use of ordinary least squares regression may be problematic. In addition to these methodological concerns, the finding of fat tails has economic implications: in these populations, certain individual households account for large percentages of the total business activity. It may be worth studying these highly productive individuals specifically in order to understand their role in the economy, and should certainly caution against any data analytic techniques which trim the data based on the outcome variable. Trimming the top 1% of households out of the profit distribution in the microcredit data would be akin to studying the US economy after trimming out the top 1% of firms. We will never fully understand the economies we study in developing countries if we trim out the most productive households, and on the contrary, we should emphasise these individuals in both theoretical and empirical work.

The analysis presented here is not exhaustive, and a deeper understanding of the household-level distributional impacts of expanding access to microcredit could be generated by including

more economics knowledge of the contextual microstructure to the analysis. Ideally, such analysis would apply an individual-level structural model to this data, but as there is currently no established methodology for partial pooling on structural parameters, that is beyond the scope of this paper. Regardless, the conclusion of the current analysis remains salient: we can reliably predict that there will be no difference between the treatment and control groups below the 75th quantile in future sites, but we cannot reliably predict the effect of expanding access to microcredit above the 75th percentile in the next site. There is strong evidence that there are no negative effects of microcredit in the general case, but there is also no clear evidence that any positive effects will generalize to future sites.

4.5 Understanding the Remaining Heterogeneity

While the results of the hierarchical aggregation display less heterogeneity across the experiments than the disaggregated results suggested, understanding the remaining heterogeneity could be important from a policy perspective. There are a number of covariates both within and across sites which could be responsible for these differences in the distributional effects of microcredit, or might at least predict the differences. 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 household with these prior businesses, conditioning the analysis on this variable could help to explain the remaining heterogeneity in the causal impact of microcredit. At the site level, there are many covariates that describe differences in economic conditions and study protocols, but as these are plausibly endogenous to the effect of microcredit in the site their predictive power does not necessarily reflect a causal relationship. In addition, with only 7 studies, any analysis of covariates at the site-level is speculative at best and regularization or sparsity estimation will be necessar to avoid severe overfitting: this exercise is described in Appendix **??**. The remainder of this section focuses on covariate analysis within study sites.

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 remarkably across the two types of households. Figures **??** and **??** 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. 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. (2015) and Meager (2016), which performed similar analyses within a single site and for the average effects respectively, and found major 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 no impact on total consumption at any quantile (figure **??**). These experienced households are also responsible for all of the observed activity on the business outcomes - this group produces the large point estimates and the massive uncertainty in the tails of the profit, revenues and expenditures distributions at the general level. However, these 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 **??**). Taken together, this suggests that households who don't have prior businesses may generally use microcredit to change the composition of their consumption bundles; to the extent that they are opening new businesses, they do not spend much on them nor do they receive much in the way of revenues (figure **??**).

Households who do have previous business experience may generally use microcredit to inject new capital into their businesses, and there is some evidence that they bring in considerable revenues, but they do not necessarily see positive increases in their profits. Examining the two potential mechanisms from the parametric PDF models shows that these experienced households are more likely to shift from zero expenditures and revenues into the positive tail, suggesting that their businesses may have been in "maintenance" or hibernation (figure **??**). The fact that there is no corresponding major increase in profits could be due to these households making mistakes or experiencing bad luck, but it is unclear why this would happen en masse. A more plausible explanation is that this result reflects a business strategy involving experimentation or investments with a relatively long maturation horizon. The fact that these households increase their consumption suggests they have some expectation of future increases in profitability. Together, these results raise the possibility that the time horizon chosen for the studies in the literature may have been too short, and that following up with the households in the sample over a longer time period might yield substantial new insights. However, this may also reflect underlying issues with measuring profitability, in which case the consumption results should be emphasized.

4.6 Is Low Take-up Responsible For These Results?

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. There is some 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. Unfortunately, there is no satisfactory way to identify the distributional impact only on those households

who were randomly induced to take up a loan (the "compliers" in the Rubin causal framework), because it is unlikely that the Stable Unit Treatment Value Assumption holds within a village.

This section provides some additional exploratory analysis which provides additional suggestive evidence that take-up patterns alone cannot explain these results. Ideally, the right comparison to make is between those households who took up microcredit due to the random expansion of access, and the same group of households in the control group. But we cannot identify those households in the control group, nor can we separate the compliers from the always-takers in the treatment group, so we cannot estimate this effect. We could compare the outcomes of the treated households who took up the microloans to the outcomes of the control group households who did not take up the loan, but this probably overestimates the effect since many people in the control group would still not have taken up the loans had they been in the treatment. We could compare the outcomes of the treated households who took up the loans to the control households who took up the loans, but a simple selection model with some barriers to entry suggests this would probably underestimate the effect on the compliers. Therefore, computing these two comparisons gives a rough ballpark on either side of the correct but infeasible comparison.

The results show that for almost every outcome variable, the "treatment effect" on the selected sample is almost identical to the intention to treat effect, suggesting no real difference for households who took up loans versus households who did not. 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 similar results as comparing all households, as shown in figure **??**. Consumption is an exception to this trend, and the 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 **??**. 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 be smaller. 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.

5 Conclusion

This paper addresses the challenges of aggregating sets of distributional treatment effects without imposing unwarranted assumptions on the degree of external validity across studies. I develop new Bayesian hierarchical models and associated pooling metrics to estimate the distributional treatment effects of policy interventions and assess the generalizability of the results. I apply these methods to aggregate the impact of expanding access to microcredit on the entire distribution of various household economic outcomes, and find that the analysis can shed light on important aspects of the data occluded by simple average treatment effects analyses. I also find that for the microcredit application, comparatively simple full-pooling methods misleadingly

produce "statistically significant" results unwarranted by the actual evidence for three of the six household outcomes studied. These results illustrate the importance of using partial pooling methodologies for evidence aggregation when the true generalizability of the treatment effects is not known. The models developed in this paper can be used to aggregate the evidence on a wide range of interventions for which distributional effects may be salient, such as policies designed to affect educational outcomes or agricultural productivity.

Moreover, these results highlight the importance of performing meta-analysis on distributional effects rather than simply examining average treatment effects. Previous papers aggregating the average impact of microcredit found generalizable information, but small or even null treatment effects (Meager 2015, Vivalt 2015). Aggregating the effects on the mean absolute deviations (MAD) show moderately sized generalizable increases in the cross-sectional dispersion in outcomes. Aggregating the quantile treatment effects illuminates the pattern underlying both of these results: microcredit generally does not affect the distribution of household outcomes below the 75% quantile, and above this quantile the variation within and across sites is so extensive as to prevent generalization. Although different methods were used to aggregate consumption outcomes and business outcomes, the pattern is detected in both cases. For business outcomes, the models detect such heavy tails as to imply unbounded variances; this indicates how challenging it is to do inference on the upper tails of these distributions, and may explain why the model declines to make a general prediction in the upper tail. This demonstrates the value of using parametric models that can capture high kurtosis when analyzing business outcomes, rather than using Gaussian approximations which may not be reliable.

Overall, it is clear that groups who receive random expansions in their access to microcredit are generally the same as groups which do not receive this access below the 75th percentile of outcomes. Above this quantile there may be some non-zero effect, but there is insufficient evidence to conclude anything about the general case; the effect could manifest differently across different contexts. The economic interpretation of these distributional effects must remain at the group level, unless a rank-invariance assumption can be invoked, which is unlikely to hold for microcredit interventions. While we can reliably predict zero difference between the treatment and control groups below the 75th% quantile in future sites, we cannot reliably predict the effect of expanding access to microcredit above the 75% percentile in the next site. This result contrasts with the prediction of the applied theoretical literature, which was that credit constraint relaxation was likely to produce winners and losers (Kaboski and Townsend 2011, Buera and Shin 2013). The tailored hierarchical PDF models provide some insight into the underlying mechanism, and they show more support for a category-switching effect moving households from making zero profit to positive profit than for any effect on the shape of the tail; however the evidence is weak in both cases. I find suggestive evidence that the zero effect detected for most of the distribution is not simply a reflection of the spike at zero, nor is it likely to be a result of low take-up.

An analysis of the role of household covariates reveals that the majority of the impact of microcredit and the heterogeneity in this impact across sites occurs within the group of households who had previous experience running their own businesses. There is strong evidence that these experienced households increase their consumption above the 75th percentile in the general case,

although there is still little change below this percentile. While households without business experience see essentially zero impact at every quantile for business outcomes and consumption, they do see some movement at the upper tail of consumer durables spending. Taken together, these results suggest that some households use microcredit to change the composition but not the total amount of their consumption, while other households use microcredit to expand their businesses and increase total consumption. Perhaps these experienced households are better positioned to take advantage of microcredit because they are innately more productive, which they signal by having already started a business, or perhaps this is a result of learning by doing. In the latter case, following up with the sampled households from these randomized trials over a longer time horizon could yield substantial new insights into the impact of microcredit. However, as both groups of households exhibit fat tails in the distribution of their business outcomes, there is likely to be important individual-level variation even within these groups. The presence of high kurtosis in these outcomes suggests that it is important to study individual decisions and partial equilibrium effects even if the overall goal is to understand general equilibrium effects or macroeconomic issues, as certain individuals make major contributions to total productivity and output even in rural village economies.

The analysis presented in this paper suggests many avenues for future research on microcredit interventions. A deeper understanding of the household-level distributional impacts of expanding access to microcredit could be generated within a structural model. There is currently no methodology for partial pooling on structural parameters, but developing and applying hierarchical structural models could be the focus of future work. In addition, the current analysis studies only the results of randomized controlled trials, but there may be useful information about microcredit in observational studies. As there is currently no way to rigorously combine these two types of evidence for aggregation purposes, this exercise remains for future research. Regardless, the conclusion of the current analysis remains salient for policy purposes and for decisions about future research. The fact that this conclusion cannot be uncovered by examining simple models on the first or second moments of a distribution using Gaussian approximations demonstrates the importance of quantile regression and of parametric modeling which leverages the underlying economic structure of the data being aggregated.

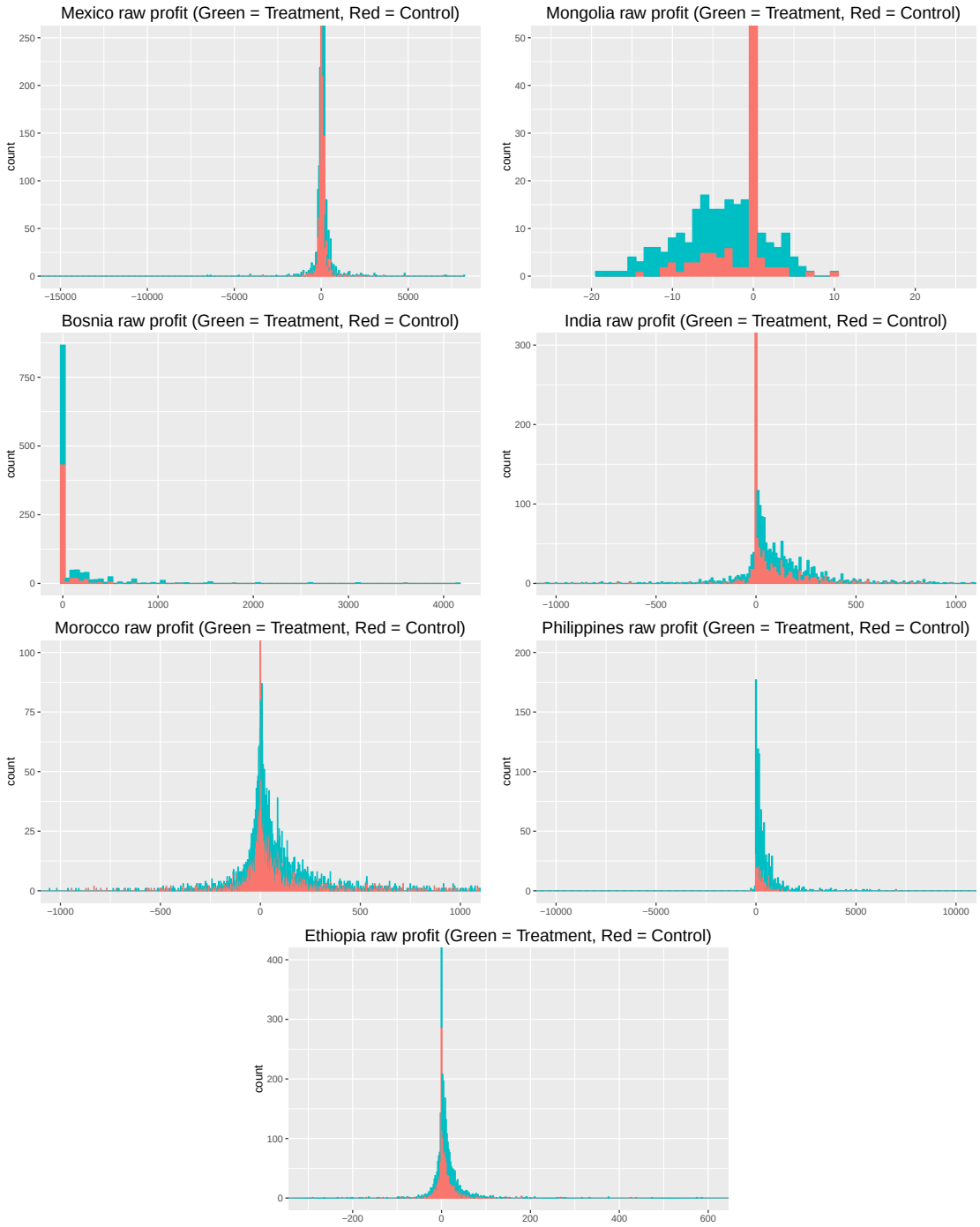

Figure 1: Histograms of the profit data in each site, in USD PPP per 2 weeks. Display truncated both vertically and horizontally in most cases. [Back to main]

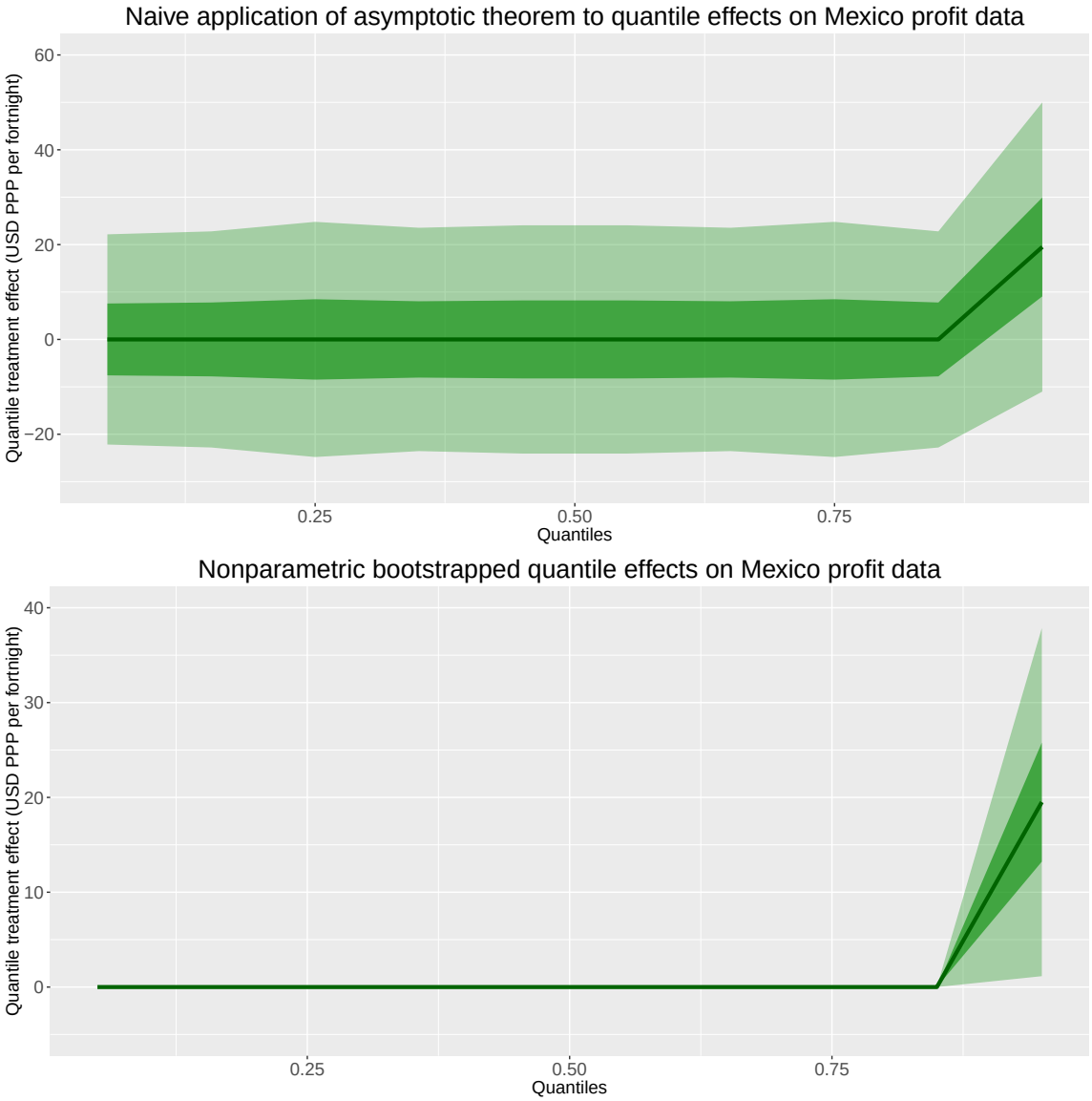

Figure 2: Quantile TEs for the Mexico profit data, Mosteller theorem approximation standard errors (above) and nonparametrically bootstrapped standard errors (below). The green line is the estimated effect, the opaque bands are the central 50% interval, the translucent bands are the central 95% interval. 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]

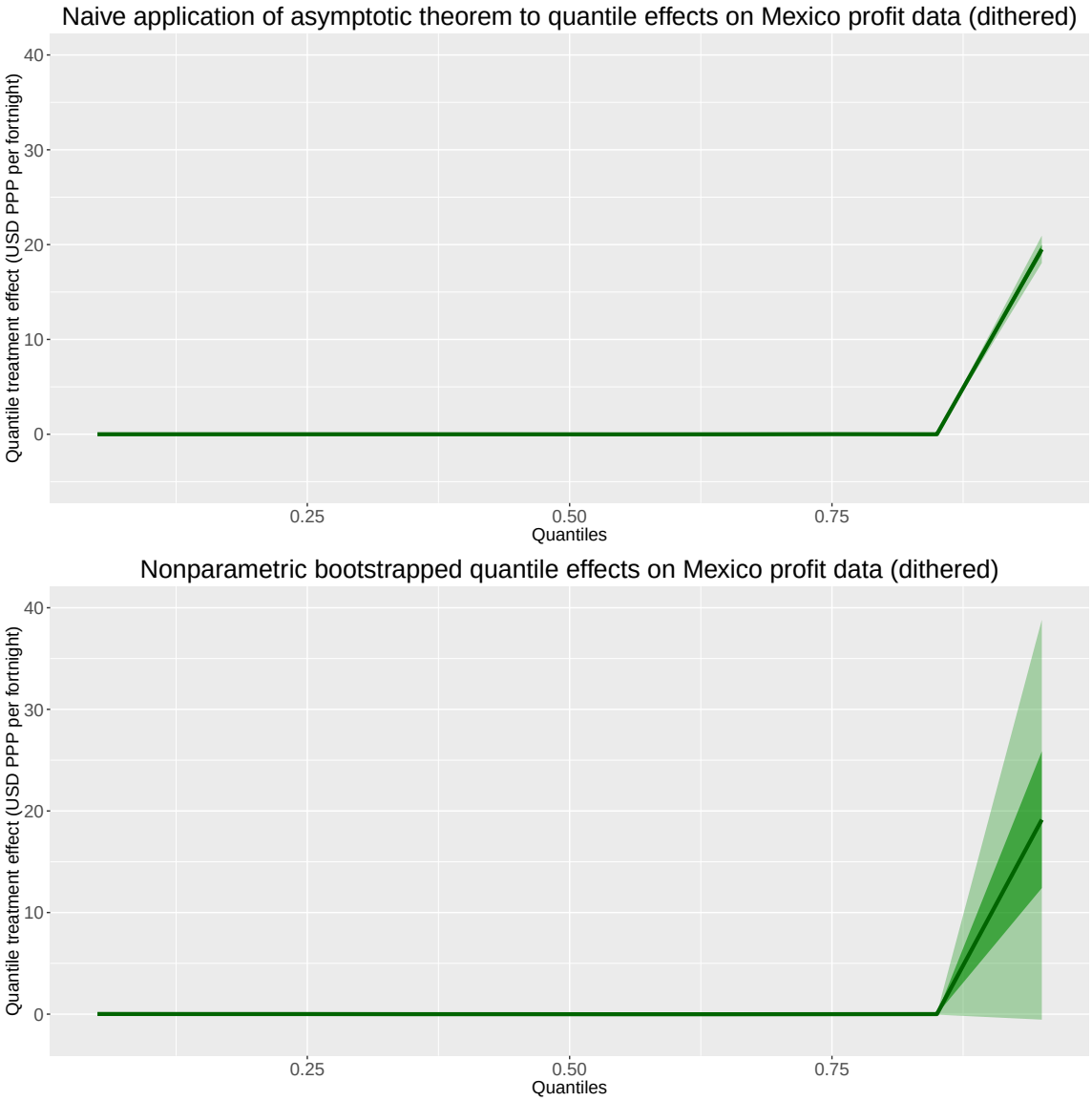

Figure 3: Quantile TEs for the dithered Mexico profit data, Mosteller theorem approximation standard errors (above) and nonparametrically bootstrapped standard errors (below). The green line is the estimated effect, the opaque bands are the central 50% interval, the translucent bands are the central 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 because the discrete spike at zero in this data is too large to be smoothed by a small amount of continuous noise. [Back to main]

Table 1: Dispersion Treatment Effects: Mean Absolute Deviation (effect specified as exp(Γ))

Outcome	Model	Effect Estimate $\tilde{\Gamma}$	SE	Posterior Quantiles			
				2.5th	25th	75th	97.5th
Profit	BHM	0.168	0.079	0.021	0.123	0.210	0.346
	Full Pooling	0.138	0.040	0.061	0.112	0.165	0.216
Expenditures	BHM	0.166	0.073	0.033	0.121	0.206	0.325
	Full Pooling	0.151	0.047	0.060	0.120	0.183	0.243
Revenues	BHM	0.142	0.074	0.013	0.096	0.182	0.306
	Full Pooling	0.113	0.038	0.038	0.087	0.138	0.188
Consumption	BHM	0.064	0.126	-0.165	0.011	0.105	0.351
	Full Pooling	0.044	0.023	-0.001	0.029	0.059	0.089
Consumer Durables	BHM	0.234	0.187	-0.134	0.165	0.307	0.559
	Full Pooling	0.246	0.062	0.123	0.204	0.287	0.368
Temptation Goods	BHM	-0.034	0.056	-0.141	-0.057	-0.012	0.078
	Full Pooling	-0.024	0.016	-0.056	-0.035	-0.013	0.007

Notes: These treatment effects are specified as an exponentiated multiplicative factor on the control group dispersion: if $\tilde{\Gamma} = 0$ the effect is zero, if $\tilde{\Gamma} = 0.7$ the effect is a 100% increase in the dispersion (i.e. the treatment group is twice as dispersed as the control group). [Back to main]

Table 2: Dispersion Treatment Effects: Standard Deviation (effect specified as $\exp(\gamma)$)

Outcome	Model	Effect Estimate $\tilde{\gamma}$	SE	Posterior Quantiles 2.5th	25th	75th	97.5th
Profit	BHM	0.547	0.323	-0.100	0.368	0.732	1.181
	Full Pooling	0.589	0.007	0.575	0.584	0.594	0.604
Expenditures	BHM	0.262	0.229	-0.188	0.137	0.391	0.713
	Full Pooling	0.188	0.007	0.173	0.183	0.192	0.202
Revenues	BHM	0.279	0.280	-0.284	0.119	0.436	0.843
	Full Pooling	0.197	0.007	0.183	0.192	0.202	0.211
Consumption	BHM	0.286	0.346	-0.386	0.123	0.451	0.951
	Full Pooling	0.226	0.008	0.211	0.221	0.231	0.241
Consumer Durables	BHM	0.374	0.367	-0.340	0.219	0.515	1.117
	Full Pooling	-0.003	0.011	-0.025	-0.010	0.005	0.019
Temptation Goods	BHM	0.036	0.361	-0.684	-0.135	0.211	0.744
	Full Pooling	-0.067	0.008	-0.082	-0.072	-0.062	-0.052

Notes: These treatment effects are specified as an exponentiated multiplicative factor on the control group dispersion: if $\tilde{\gamma} = 0$ the effect is zero, if $\tilde{\gamma} = 0.7$ the effect is a 100% increase in the dispersion (i.e. the treatment group is twice as dispersed as the control group). [Back to main]

Table 3: Average Treatment Effect of Microcredit Intervention (τ)

Outcome	Model	Effect Estimate	Posterior Distribution Quantiles			
		$\tilde{\tau}$	2.5th	25th	75th	97.5th
Profit	BHM (Joint)	2.565	-2.923	0.018	4.775	10.235
	BHM (NC)	6.809	-3.029	1.819	10.381	24.492
	Full Pooling	7.245	-1.780	4.139	10.351	16.270
Expenditures	BHM (Joint)	4.177	-0.939	2.021	5.993	11.334
	BHM (NC)	6.717	-2.304	2.565	9.702	22.065
	Full Pooling	13.011	-2.581	7.645	18.376	28.602
Revenues	BHM (Joint)	6.033	-1.521	3.236	8.631	15.056
	BHM (NC)	14.453	-1.397	6.577	19.934	43.527
	Full Pooling	22.481	4.608	16.330	28.631	40.354
Consumption	BHM (Joint)	2.609	-4.303	0.733	4.579	9.255
	BHM (NC)	3.436	-6.275	0.825	5.927	13.211
	Full Pooling	4.626	-1.138	2.642	6.609	10.389
Consumer Durables	BHM (Joint)	1.628	-2.002	0.700	2.490	5.603
	BHM (NC)	1.826	-3.903	0.675	2.880	8.290
	Full Pooling	2.288	-23.916	-6.729	11.306	28.493
Temptation Goods	BHM (Joint)	-0.705	-3.057	-1.150	-0.167	1.151
	BHM (NC)	-0.790	-3.332	-1.263	-0.218	1.279
	Full Pooling	-0.637	-1.065	-0.784	-0.490	-0.209

Notes: All effects are in USD PPP per fortnight. The BHM(Joint) refers to the model that estimates effects on both the mean (location) and dispersion of the outcome distribution, in this case the dispersion is measured by the mean absolute deviations. The BHM (NC) is "non-corrected" as it only estimates effects on the mean and does not adjust for effects on the dispersion. The Full Pooling Model in both papers was computed with Eicker-Huber-White standard errors, which are generally robust to heteroskedasticity but which do not exploit the specific knowledge of the structure of the heteroskedasticity in this problem. [Back to main]

Table 4: Pooling Factors for MAD Effects: Joint Model

Outcome	Treatment Effects			Control Group Means		
	$\omega(\Gamma)$	$\breve{\omega}(\Gamma)$	$\lambda(\Gamma)$	$\omega(\Delta)$	$\breve{\omega}(\Delta)$	$\lambda(\Delta)$
Profit	0.469	0.339	0.705	0.003	0.007	0.005
Expenditures	0.514	0.739	0.817	0.003	0.004	0.004
Revenues	0.459	0.641	0.743	0.002	0.003	0.003
Consumption	0.127	0.267	0.559	0.114	0.277	0.542
Consumer Durables	0.199	0.476	0.838	0.001	0.002	0.002
Temptation Goods	0.314	0.452	0.791	0.005	0.003	0.012

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 Variance Effects: Joint Model

Outcome	Treatment Effects			Control Group Means		
	$\omega(\gamma)$	$\breve{\omega}(\gamma)$	$\lambda(\gamma)$	$\omega(\eta)$	$\breve{\omega}(\eta)$	$\lambda(\eta)$
Profit	0.002	0.002	0.004	0	0.001	0
Expenditures	0.003	0.030	0.007	0	0.001	0
Revenues	0.002	0.007	0.005	0	0	0
Consumption	0.002	0.011	0.006	0.006	0.023	0.020
Consumer Durables	0.002	0.043	0.013	0	0.001	0
Temptation Goods	0.002	0.005	0.006	0	0.005	0.001

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]

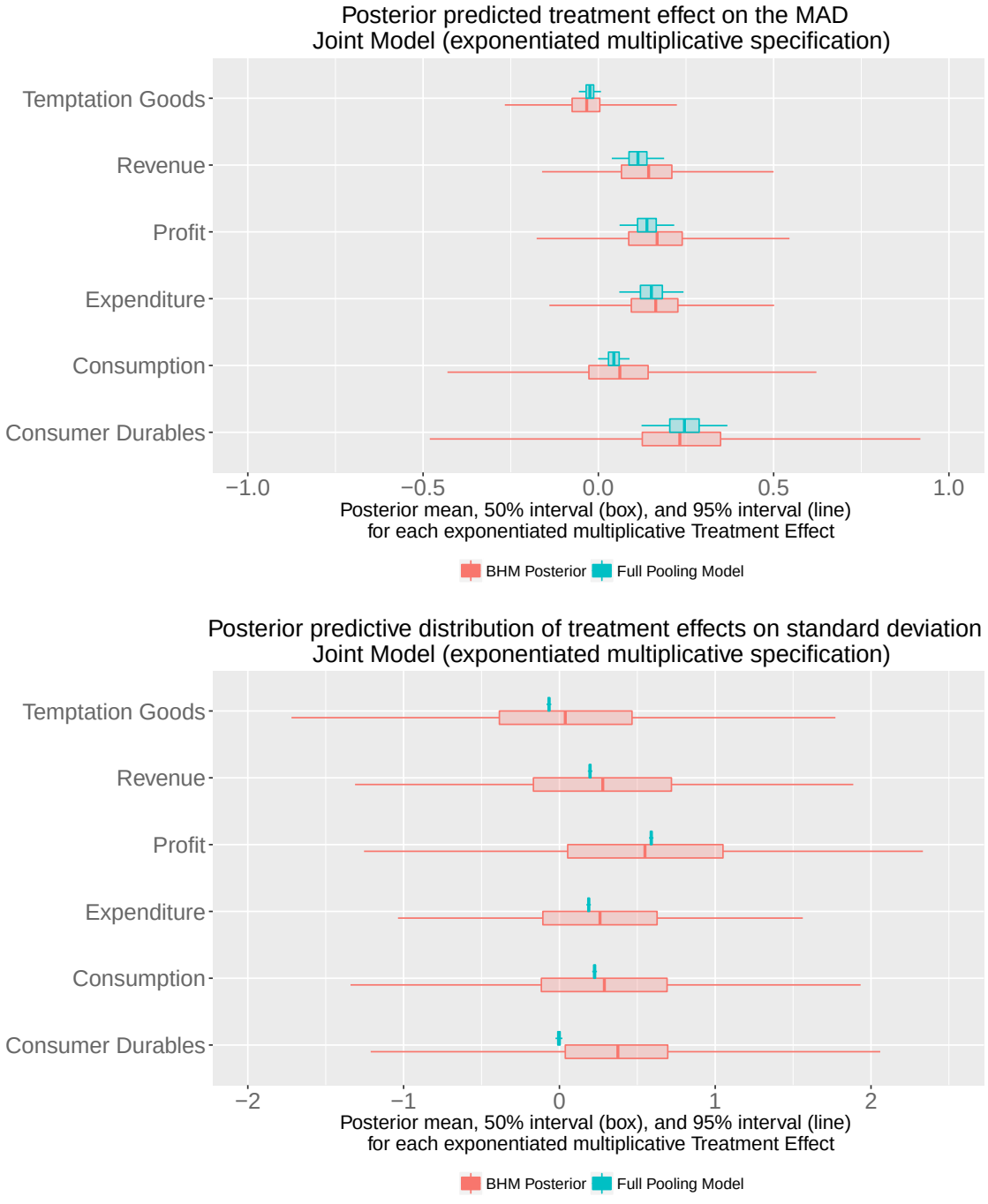

Figure 4: Marginal posterior predictive distribution of Γ_{K+1}, and of γ_{K+1} from the joint model. This is the predicted treatment effect in a future exchangeable study site, with uncertainty intervals that account for the estimated generalizability (or lack of it). [Back to main]

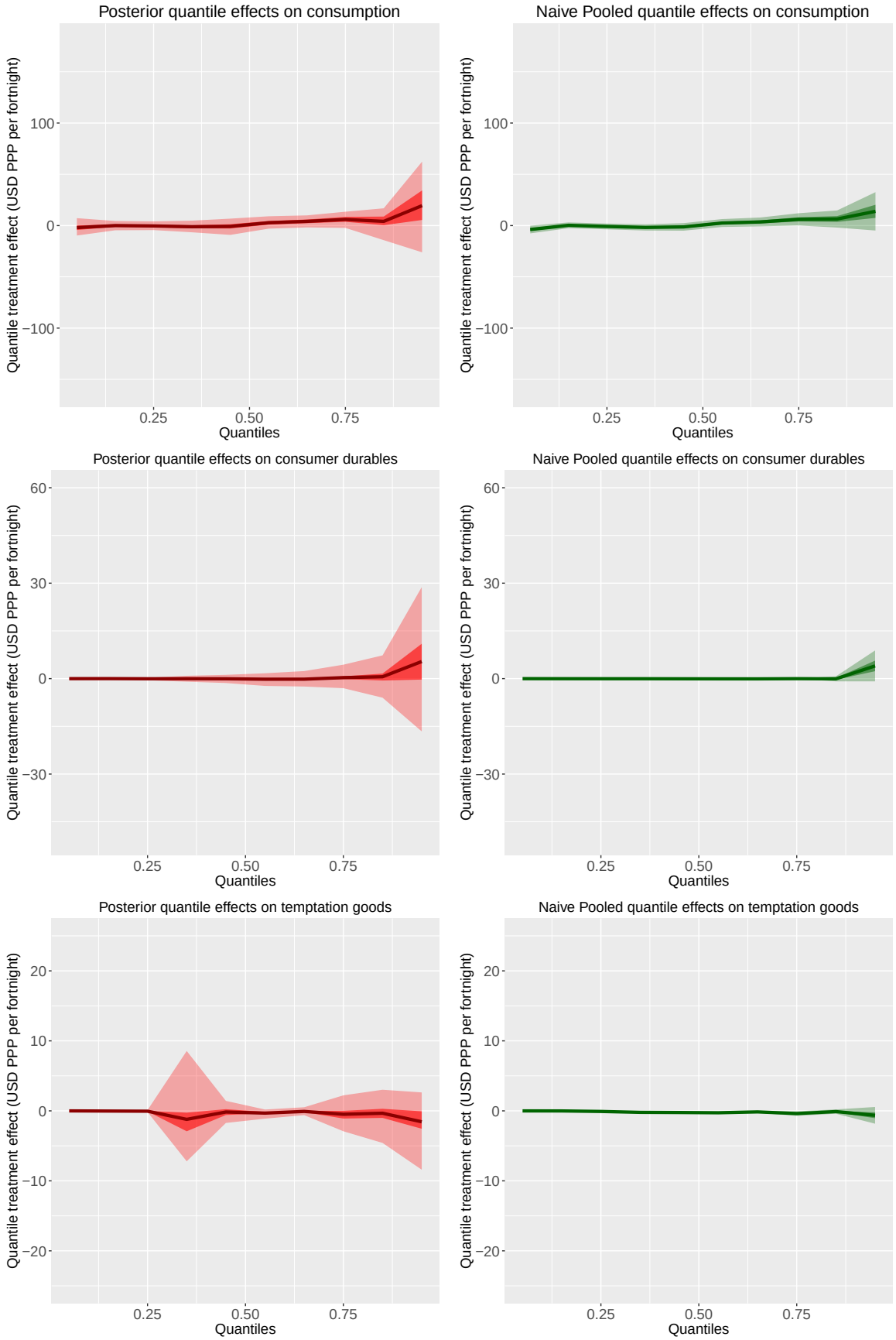

Figure 5: General Quantile Treatment Effect Curves (β_1) for consumption-type variables. The dark line is the posterior mean, the opaque color bands are the central 50% posterior uncertainty interval, the translucent color bands are the central 95% posterior uncertainty interval. [Back to main]

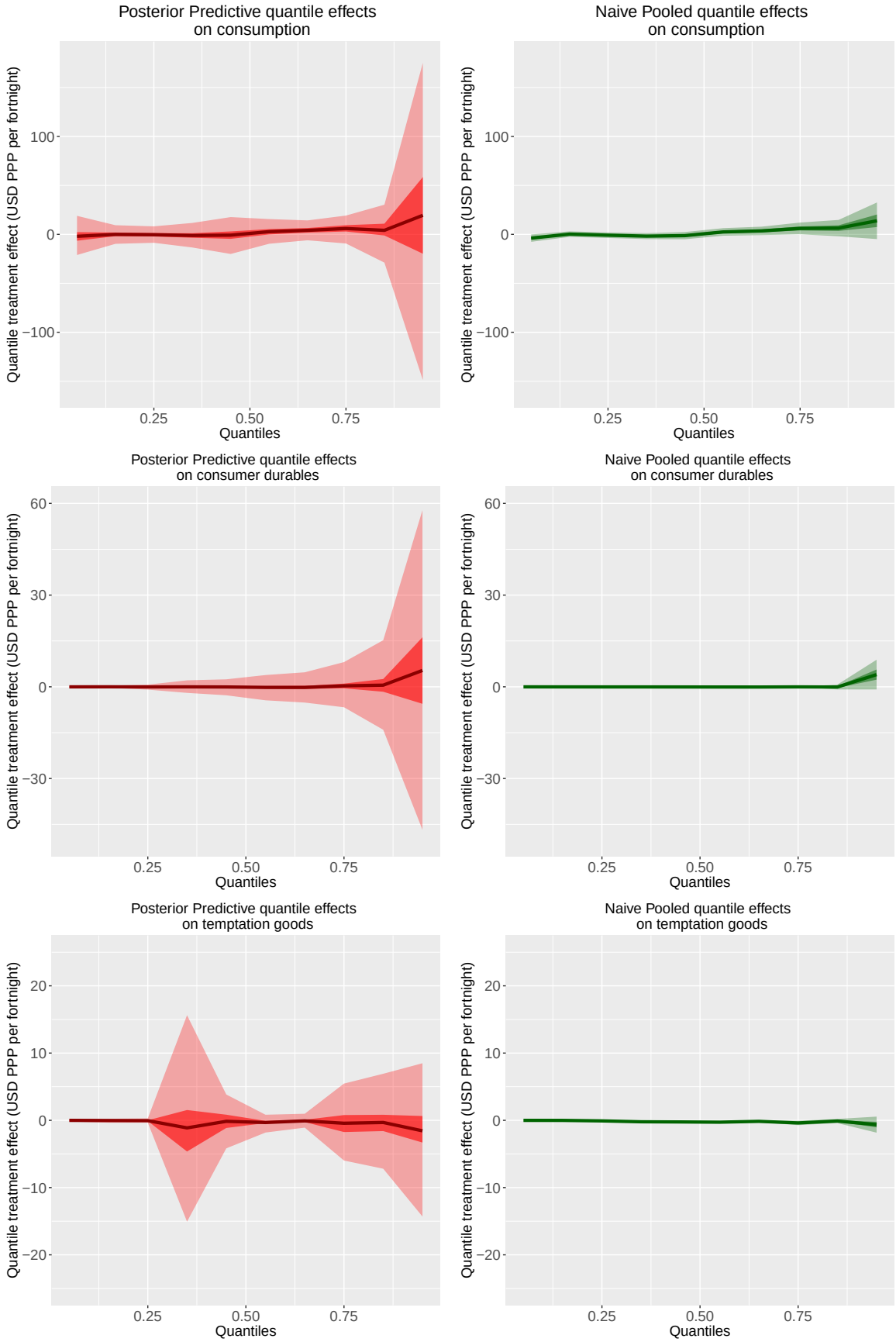

Figure 6: Posterior Predictive Quantile Effect Curves ($\beta_{1,K+1}$) for consumption-type variables.The dark line is the posterior mean, the opaque color bands are the central 50% posterior predictive uncertainty interval, the translucent color bands are the central 95% posterior predictive uncertainty interval. [Back to main]

Table 6: 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]

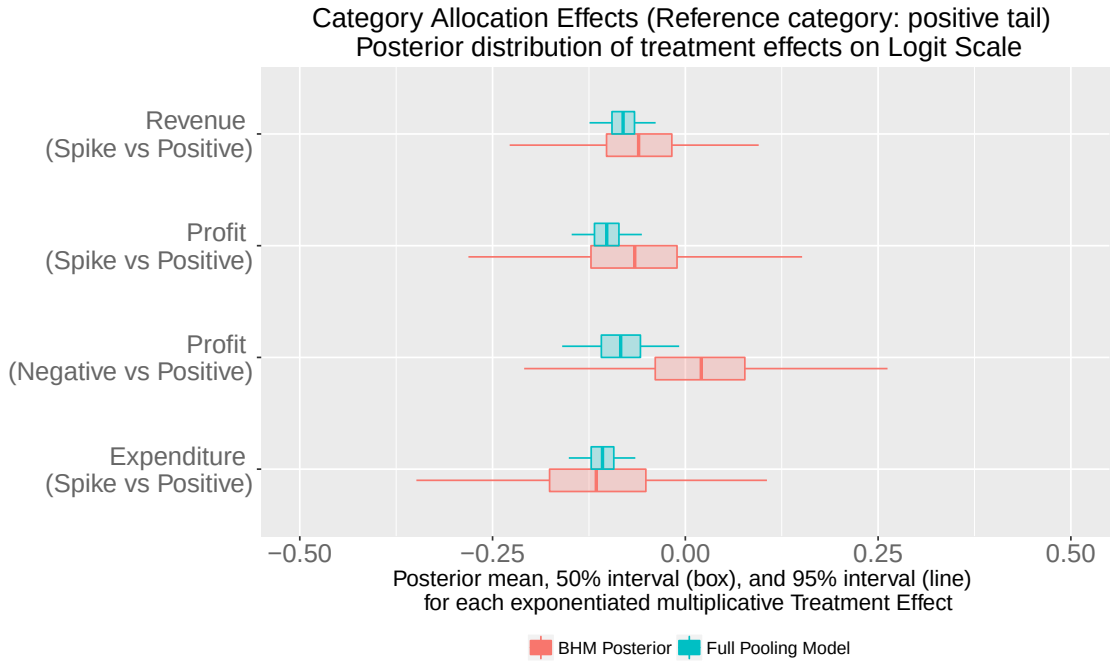

Figure 7: Posterior distributions for the logit treatment effects (π_j) on category assignment. These treatment effects are specified as an exponentiated multiplicative factor on the control group proportion of households in the category: if $\tilde{\pi_j} = 0$ the effect is zero, if $\tilde{\pi_j} < 0$ the treatment increases the proportion of households in the positive tail relative to other categories. [Back to main]

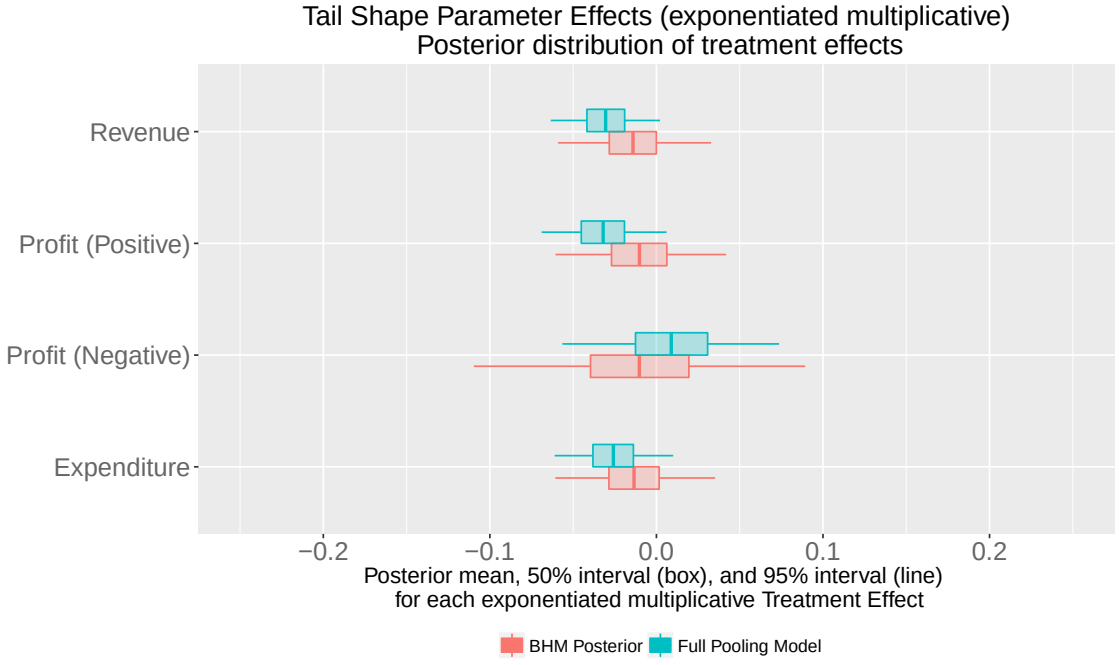

Figure 8: Posterior distributions for the Pareto shape treatment effects (κ_j) in each site. These treatment effects are specified as an exponentiated multiplicative factor on the control group scale parameter: if $\tilde{\kappa_j} = 0$ the effect is zero, if $\tilde{\kappa_j} = 0.7$ the effect is a 100% increase in the scale parameter. [Back to main]

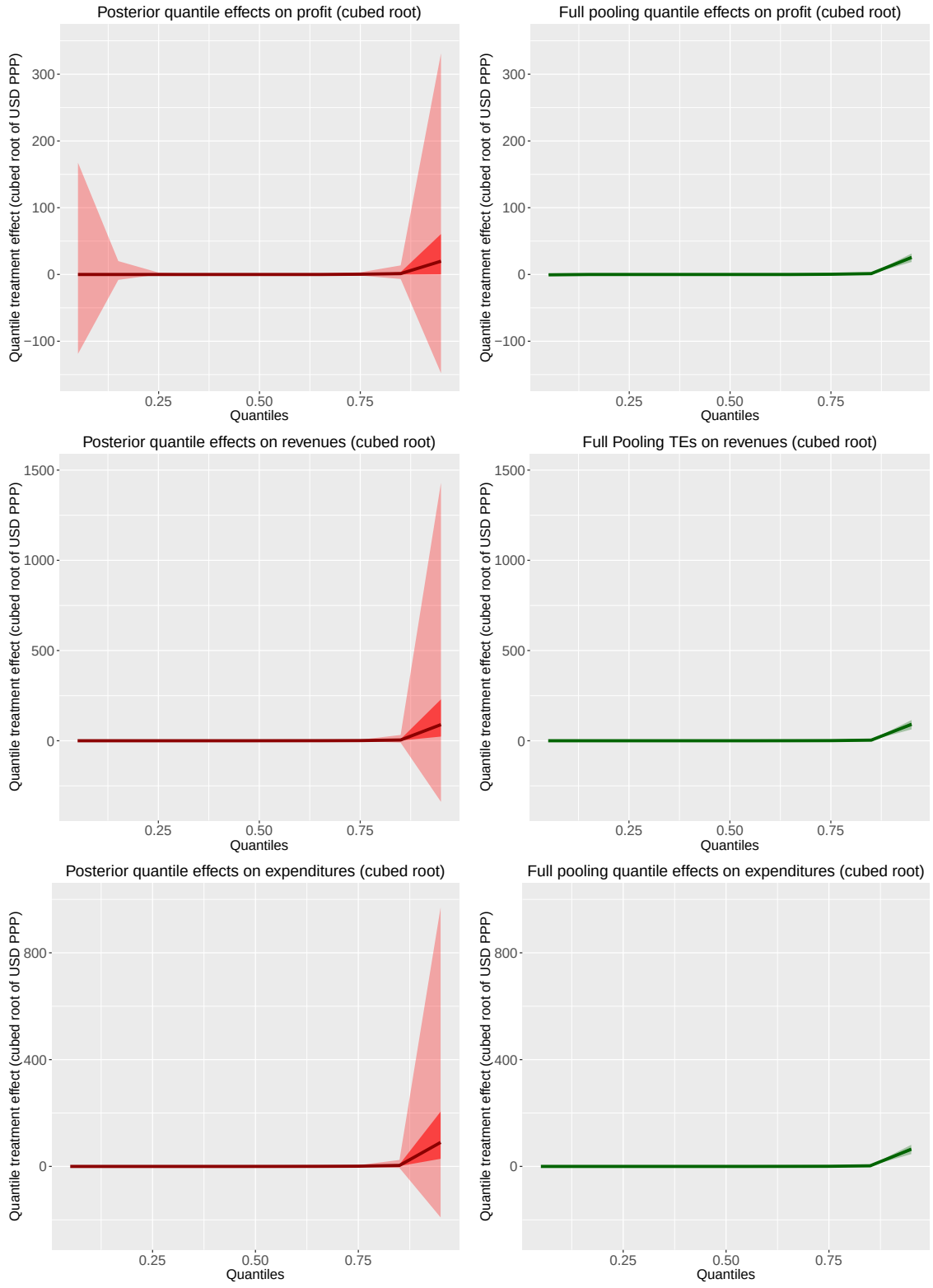

Figure 9: General Quantile Treatment Effect Curves (β_1) for business variables. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in cubed root of USD PPP due to the scale differences in the uncertainty at the right tail versus the rest of the distribution. Results are shown on the raw scale in Appendix C. [Back to main]

Table 7: Pooling Factors for Categorical Logit Parameters (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	0.721	0.907	0.144	0.421	0.240
Profit (Zero vs Positive)	0.137	0.476	0.688	0.013	0.379	0.487
Expenditures (Zero vs Positive)	0.084	0.612	0.783	0.010	0.498	0.570
Revenues (Zero vs Positive)	0.131	0.694	0.881	0.010	0.509	0.562

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 8: Pooling Factors for Tail Shape Parameters

Outcome	Treatment Effects			Control Group Means		
	$\omega(\pi_j)$	$\breve{\omega}(\pi_j)$	$\lambda(\pi_j)$	$\omega(\alpha_j)$	$\breve{\omega}(\alpha_j)$	$\lambda(\alpha_j)$
Profit (Negative Tail)	0.389	0.855	0.991	0.284	0.346	0.494
Profit (Positive Tail)	0.219	0.785	0.988	0.036	0.074	0.089
Expenditures	0.175	0.756	0.987	0.019	0.061	0.050
Revenues	0.169	0.692	0.977	0.014	0.036	0.029

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]

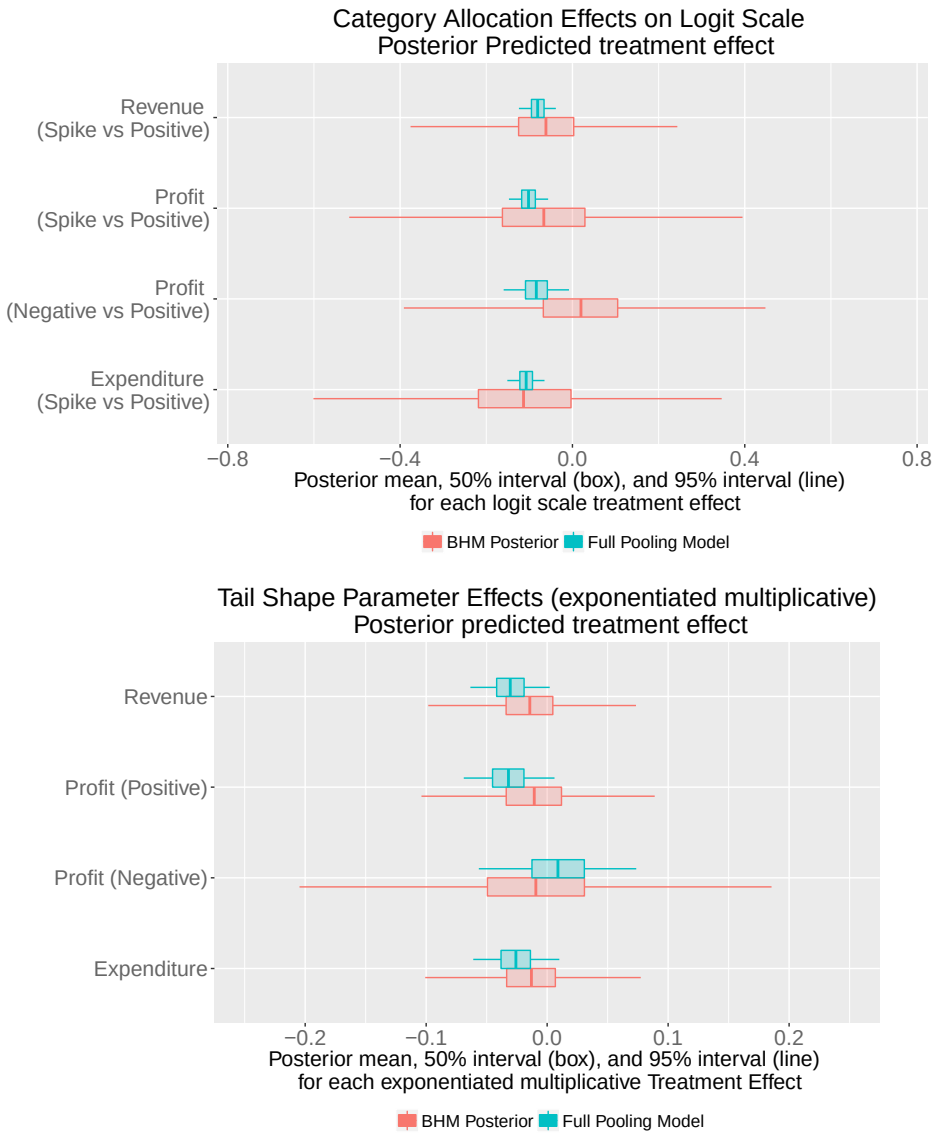

Figure 10: Posterior predicted distributions for the logit treatment effects on category assignment and tail shape effects. In each case this is the predicted treatment effect in a future exchangeable study site, with uncertainty intervals that account for the estimated generalizability (or lack of it). [Back to main]

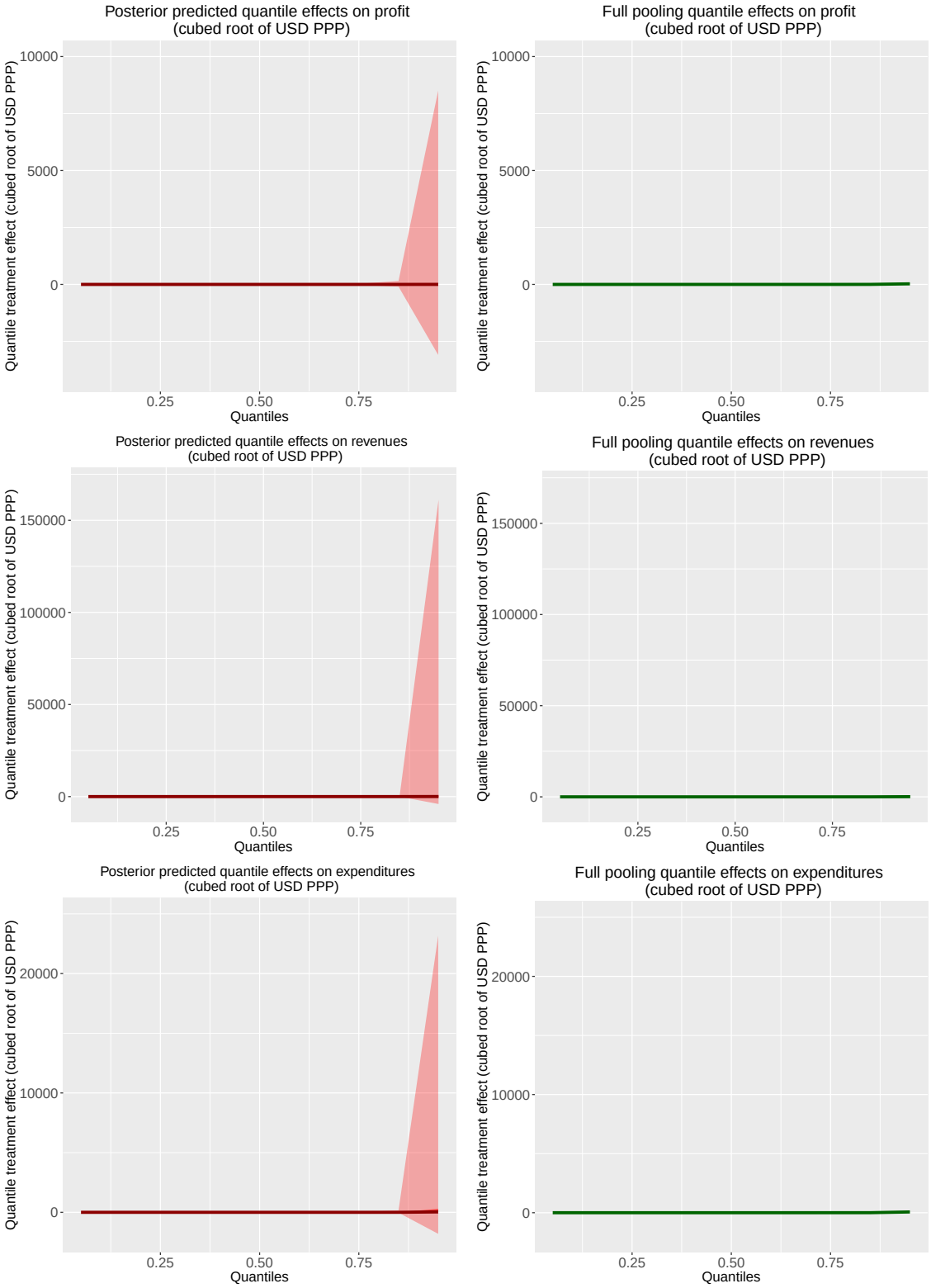

Figure 11: Posterior predicted quantile treatment effect Curves for Business Variables. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in cubed root of USD PPP due to the scale differences in the uncertainty at the right tail versus the rest of the distribution. The results are shown in the raw scale in Appendix C. [Back to main]

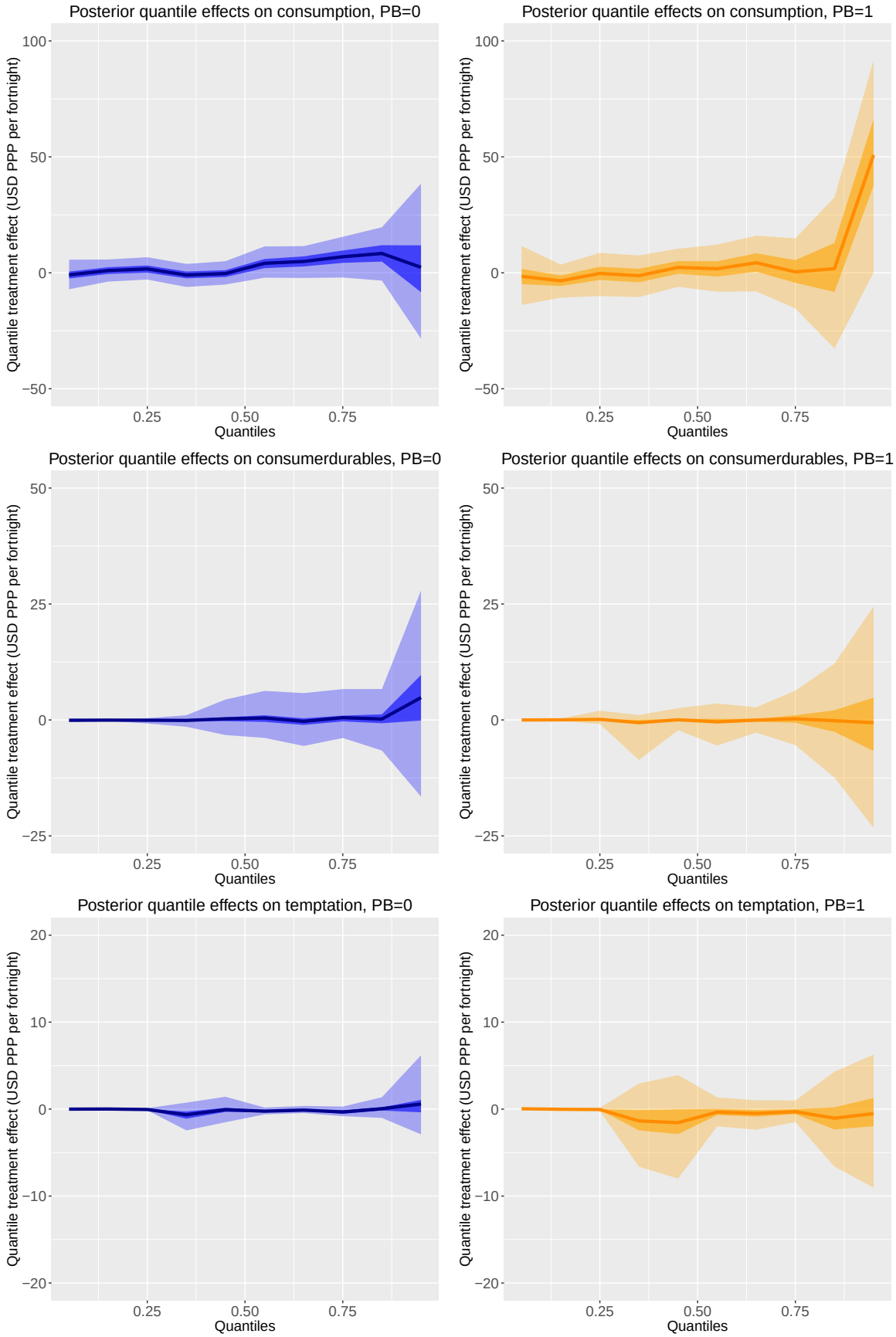

Figure 12: General Quantile Treatment Effect Curves split by prior business ownership (β_1) for consumption-type variables. The dark line is the posterior mean, the opaque color bands are the central 50% posterior uncertainty interval, the translucent color bands are the central 95% posterior uncertainty interval. [Back to main]

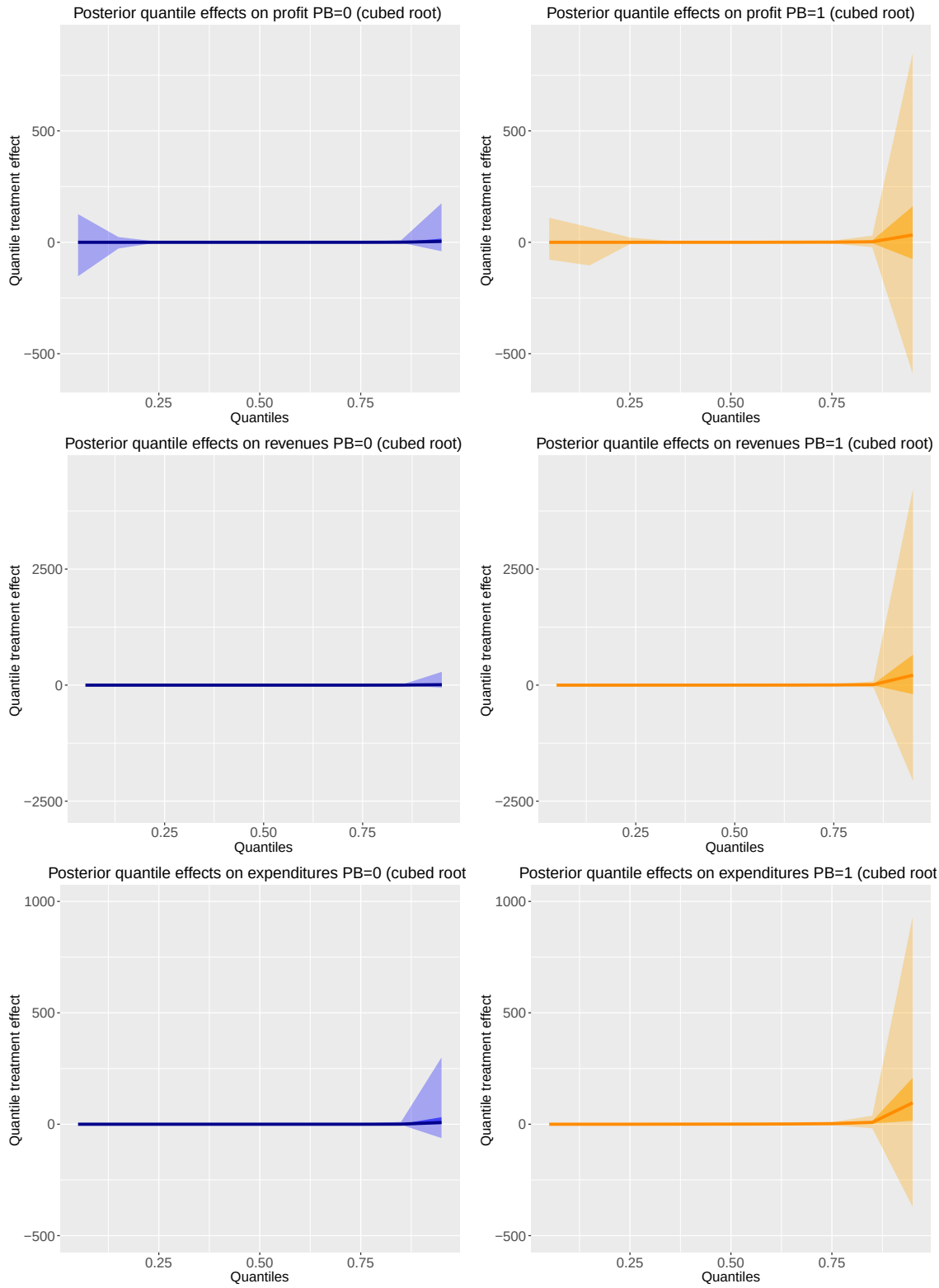

Figure 13: General Quantile Treatment Effect Curves (β_1) for business variables split by prior business ownership. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in cubed root of USD PPP due to the scale differences in the uncertainty at the right tail versus the rest of the distribution. [Back to main]

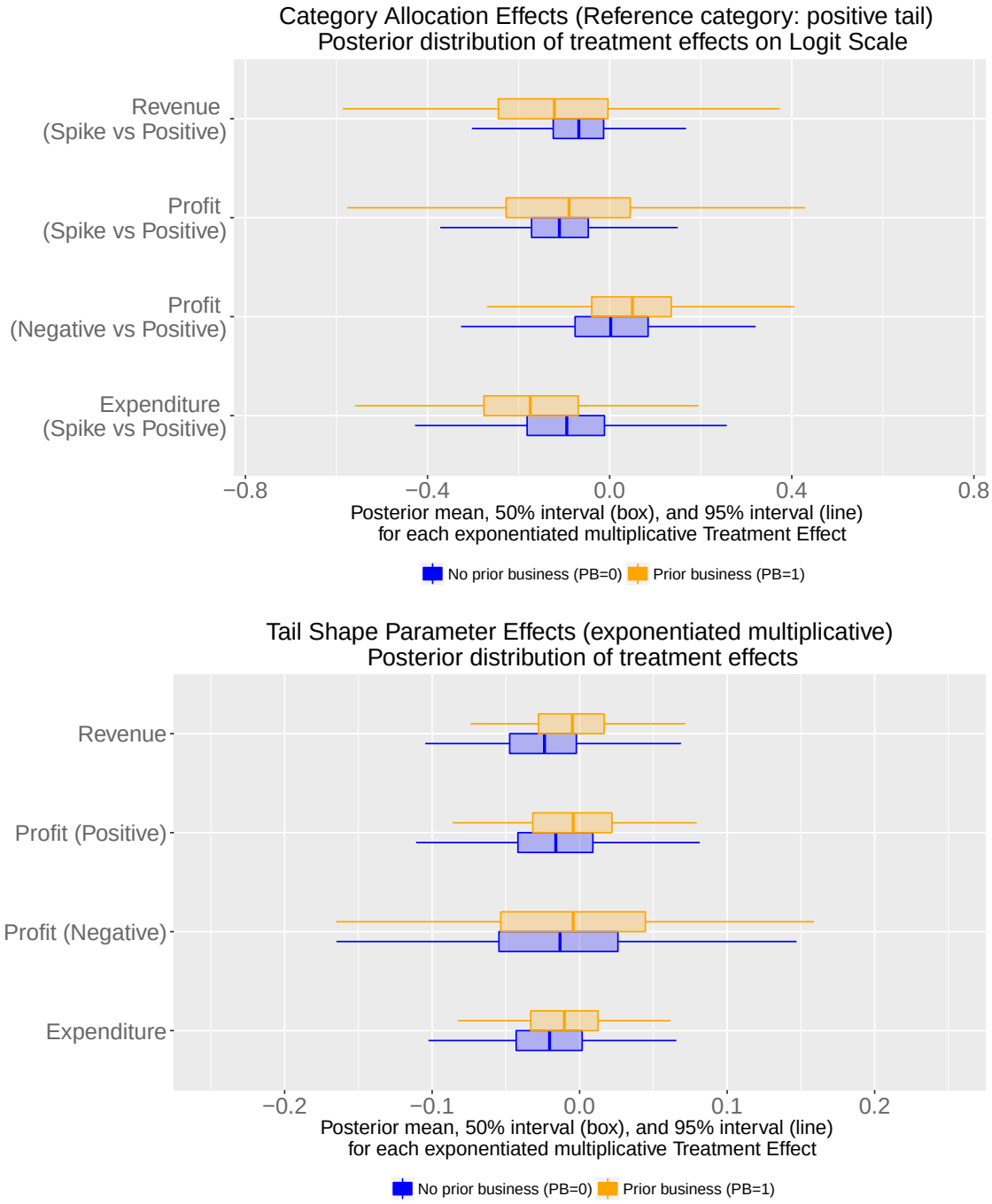

Figure 14: Upper panel: Posterior distributions for the logit treatment effects (π_j) on category assignment split by prior business ownership. These treatment effects are specified as an exponentiated multiplicative factor on the control group proportion of households in the category: if $\tilde{\pi_j} = 0$ the effect is zero, if $\tilde{\pi_j} < 0$ the treatment increases the proportion of households in the positive tail relative to other categories. Lower panel: Posterior distributions for the Pareto shape treatment effects (κ_j) in each site. These treatment effects are specified as an exponentiated multiplicative factor on the control group scale parameter: if $\tilde{\kappa_j} = 0$ the effect is zero, if $\tilde{\kappa_j} = 0.7$ the effect is a 100% increase in the scale parameter. [Back to main]

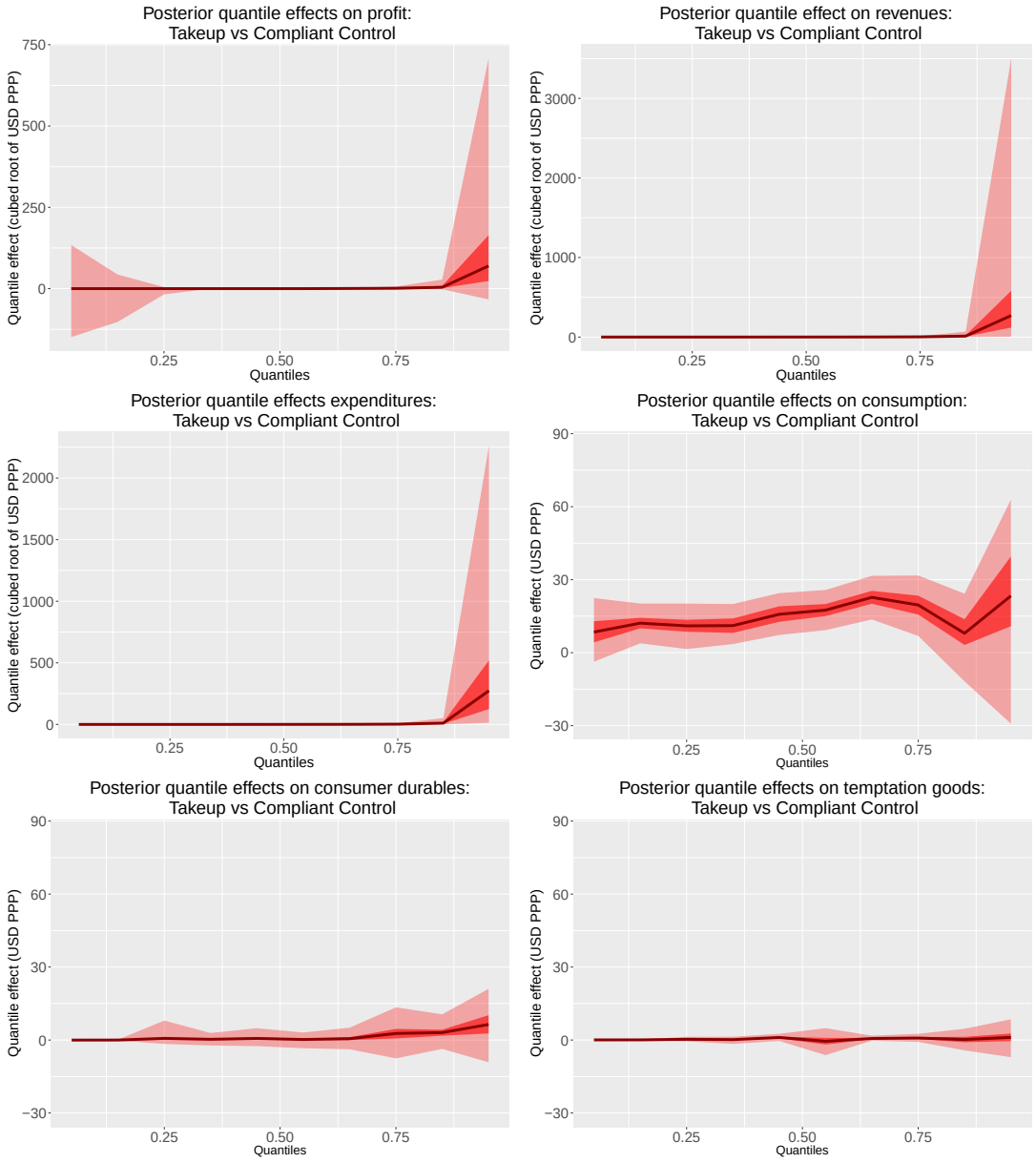

Figure 15: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Compliant control households who did not take up. This effect should overestimate the true impact of microcredit on those who take it up in a simple selection framework. Consumption variables 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]

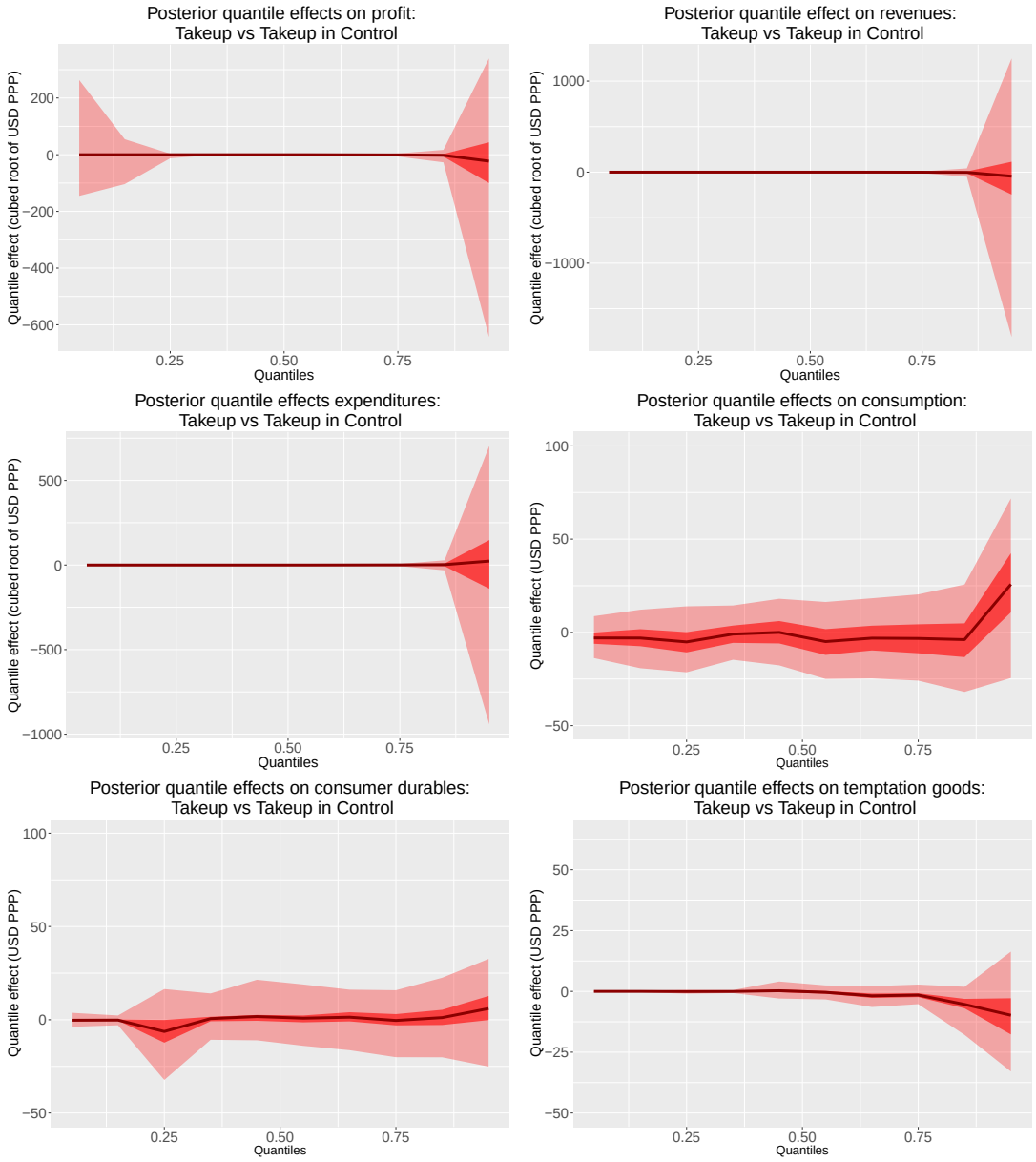

Figure 16: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Control households who took up. This effect should underestimate the true impact of microcredit on those who take it up in a simple selection framework. Consumption variables are in USD PPP per 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]

Appendices

A Variance Effects: Independent Specification

This section provides the specification and results of fitting an independent model to the variance treatment effects, which is not as sensitive to the priors as the joint model. This independent model has the same lower level of the likelihood $f(\mathcal{Y}_k|\theta_k)$ as the joint model:

$$y_{nk} \sim N(\mu_k + \tau_k T_{nk}, ((\exp(\eta_k + \gamma_k T_{nk})))^2) \; \forall \; k. \tag{A.1}$$

The upper level $\psi(\theta_k|\theta)$ is specified independently as follows:

$$\begin{aligned} \tau_k &\sim N(\tau, \sigma_\tau^2) \; \forall \; k \\ \mu_k &\sim N(\mu, \sigma_\mu^2) \; \forall \; k \\ \gamma_k &\sim N(\gamma, \sigma_\gamma^2) \; \forall \; k \\ \eta_k &\sim N(\eta, \sigma_\eta^2) \; \forall \; k. \end{aligned} \tag{A.2}$$

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

$$\begin{aligned} \sigma_\mu, \sigma_\tau &\sim Cauchy(0, 50) \\ \sigma_\eta, \sigma_\gamma &\sim Cauchy(0, 5) \\ \tau &\sim N(0, 1000^2) \\ \gamma &\sim N(0, 100^2) \\ \eta &\sim N(0, 100^2) \\ \mu &\sim N(0, 1000^2). \end{aligned} \tag{A.3}$$

The results of fitting this model are shown in the tables and figures which constitute the remainder of this appendix.

Table 9: Marginal Posteriors for Variance Treatment Effects: Independent Model

Outcome	Model	Effect Estimate	SE	Posterior Quantiles			
				2.5th	25th	75th	97.5th
Profit	BHM	0.549	0.269	0.007	0.402	0.699	1.091
	Full Pooling	0.589	0.007	0.575	0.584	0.594	0.604
Expenditures	BHM	0.265	0.211	-0.156	0.145	0.383	0.683
	Full Pooling	0.188	0.007	0.173	0.183	0.192	0.202
Revenues	BHM	0.279	0.256	-0.219	0.135	0.422	0.781
	Full Pooling	0.197	0.007	0.183	0.192	0.202	0.211
Consumption	BHM	0.295	0.343	-0.396	0.132	0.466	0.949
	Full Pooling	0.226	0.008	0.211	0.221	0.231	0.241
Consumer Durables	BHM	0.359	0.437	-0.407	0.193	0.539	1.178
	Full Pooling	-0.003	0.011	-0.025	-0.010	0.005	0.019
Temptation Goods	BHM	0.039	0.342	-0.617	-0.131	0.203	0.722
	Full Pooling	-0.067	0.008	-0.082	-0.072	-0.062	-0.052

Notes: These treatment effects are specified as an exponentiated multiplicative factor on the control group dispersion: if $\tilde{\gamma} = 0$ the effect is zero, if $\tilde{\gamma} = 0.7$ the effect is a 100% increase in the dispersion (i.e. the treatment group is twice as dispersed as the control group). [Back to main]

Table 10: Pooling Factors for Variance Effects: Independent Model

Outcome	Treatment Effects			Control Group Means		
	$\omega(\gamma)$	$\breve{\omega}(\gamma)$	$\lambda(\gamma)$	$\omega(\eta)$	$\breve{\omega}(\eta)$	$\lambda(\eta)$
Profit	0.002	0.003	0.004	0	0.001	0
Expenditures	0.002	0.006	0.008	0	0.001	0
Revenues	0.002	0.006	0.005	0	0.001	0
Consumption	0.002	0.012	0.006	0.002	0.025	0.020
Consumer Durables	0.002	0.089	0.014	0	0.005	0
Temptation Goods	0.002	0.004	0.006	0.001	0.005	0.001

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]

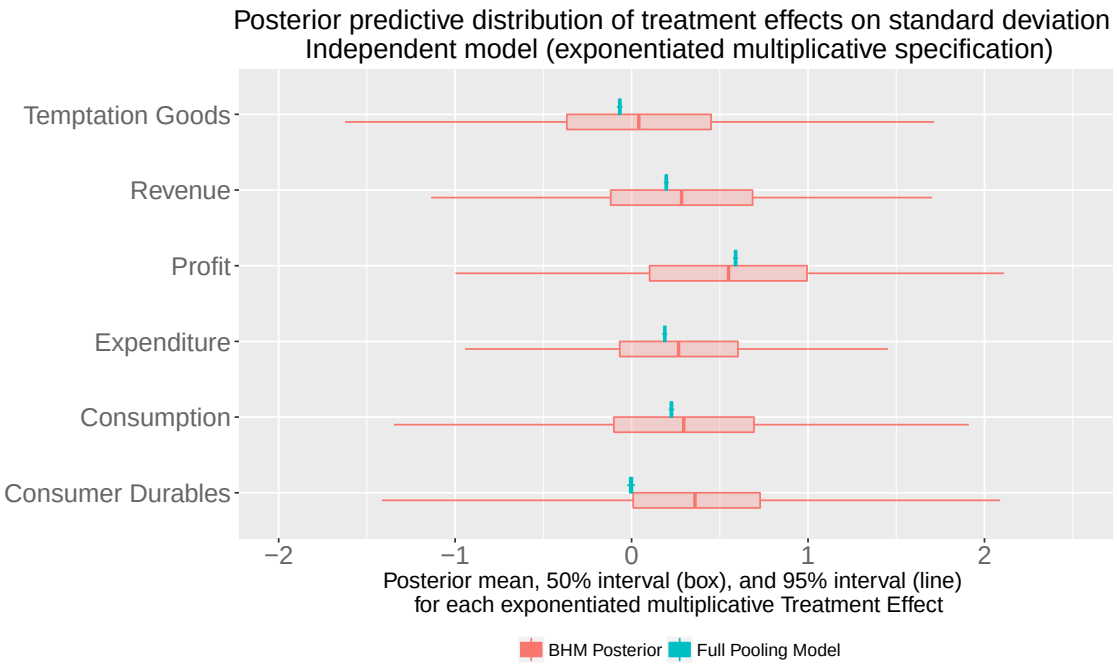

Figure 17: Marginal posterior predictive distribution of γ_{K+1} from the independent model.[Back to main]

B 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.

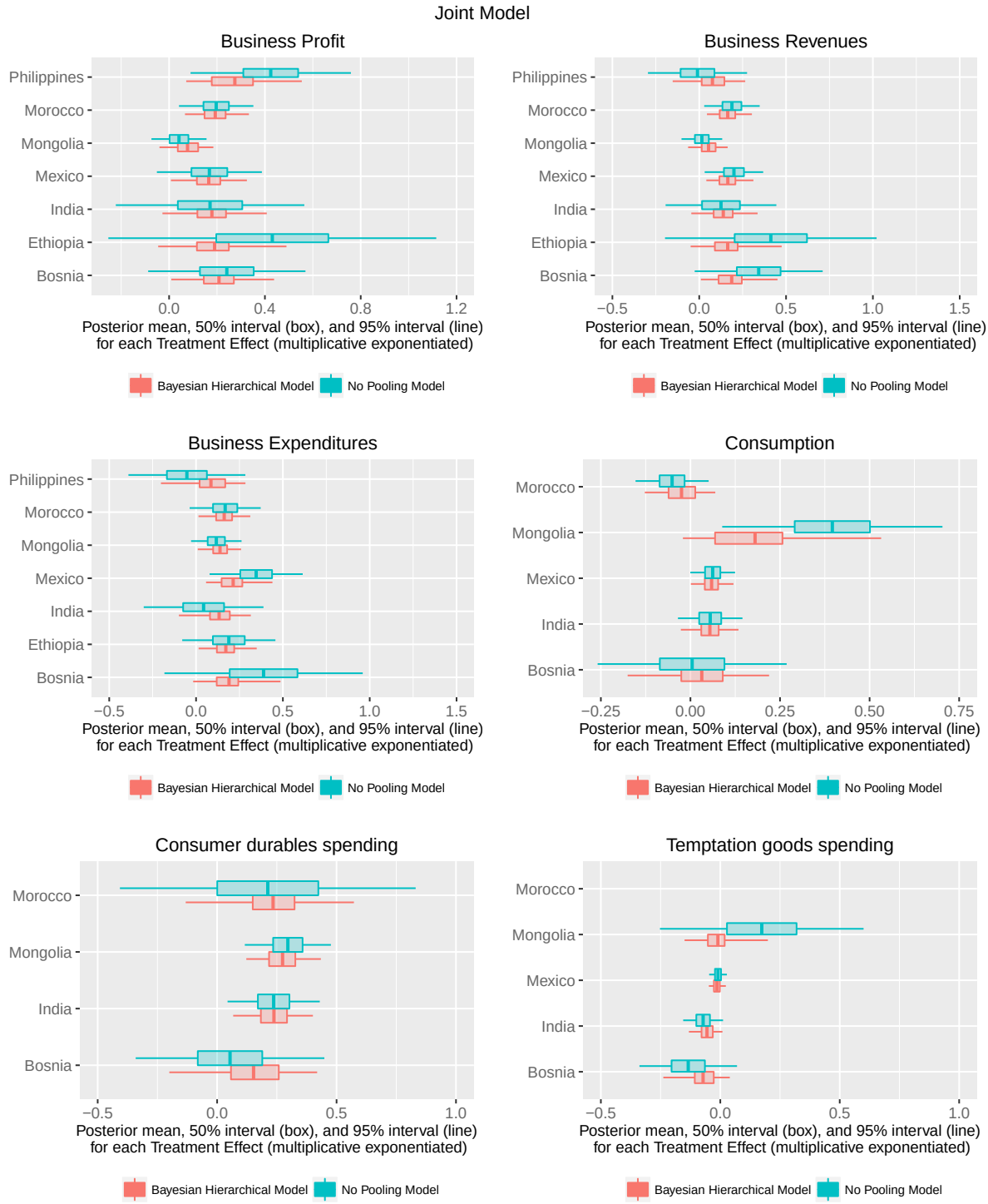

Figure 18: Marginal posterior distribution of Γ_k from the joint model.[Back to main]

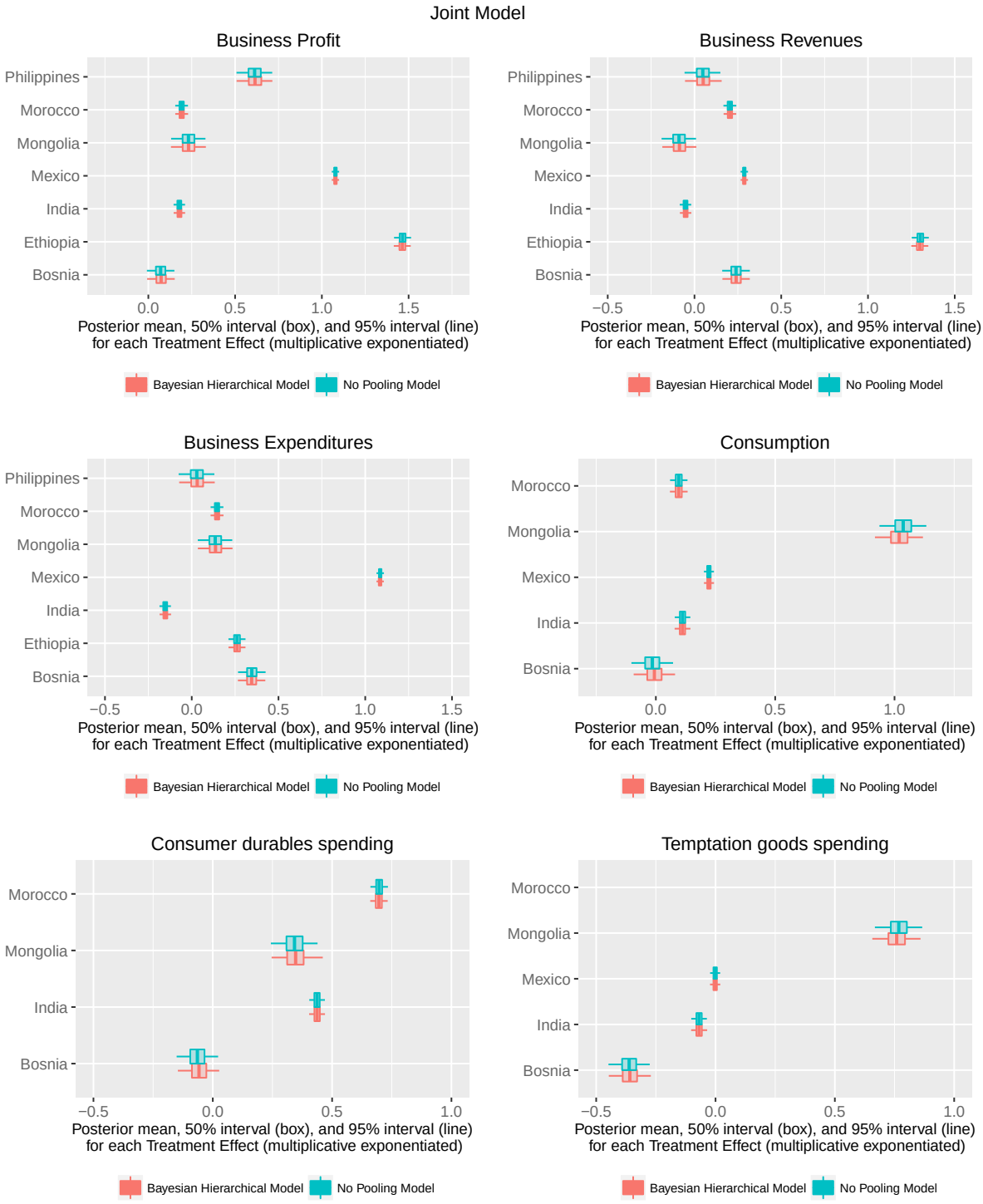

Figure 19: Marginal posterior distribution of γ_k from the joint model.[Back to main]

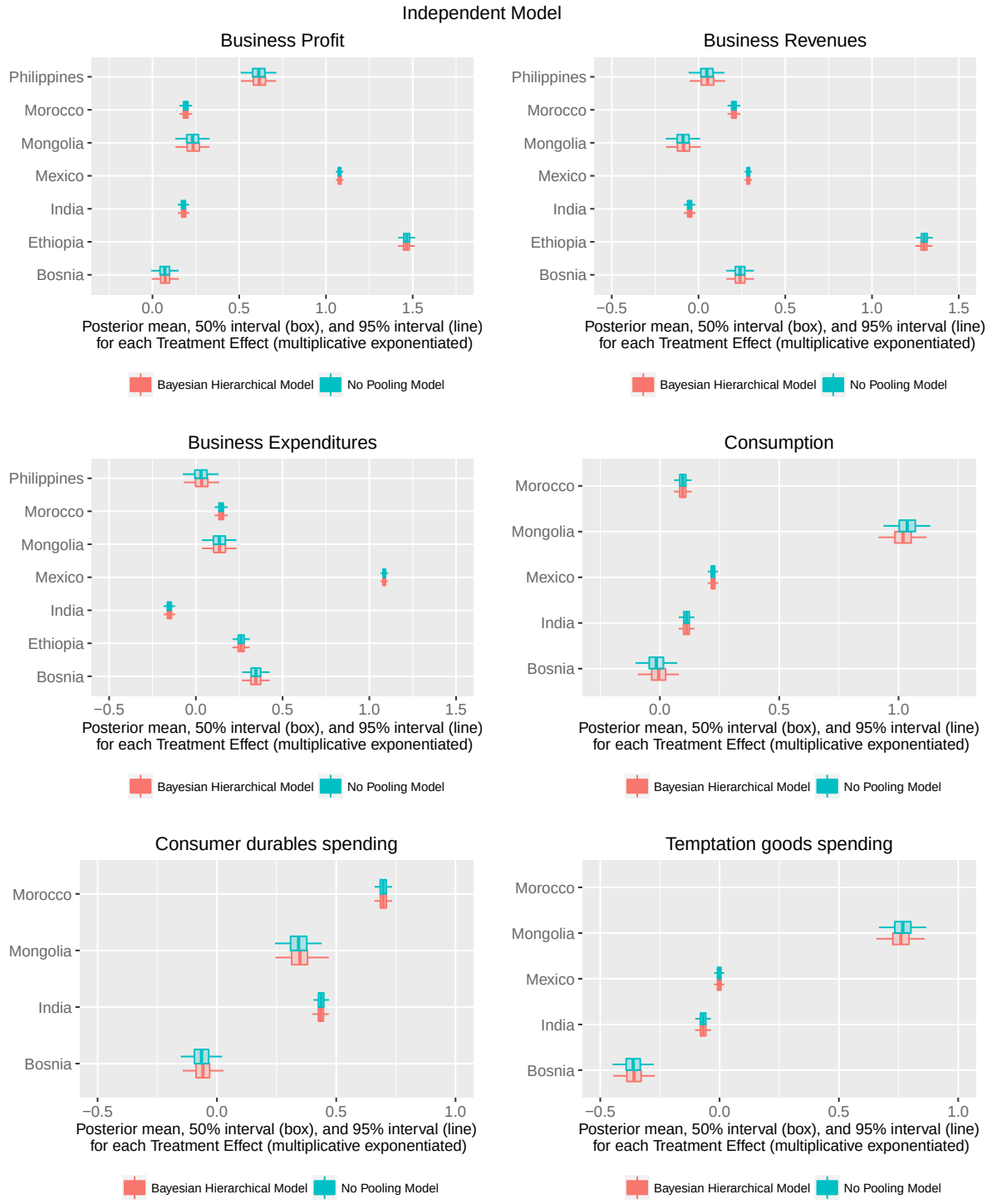

Figure 20: Marginal posterior distribution of γ_k from the independent model [Back to main]

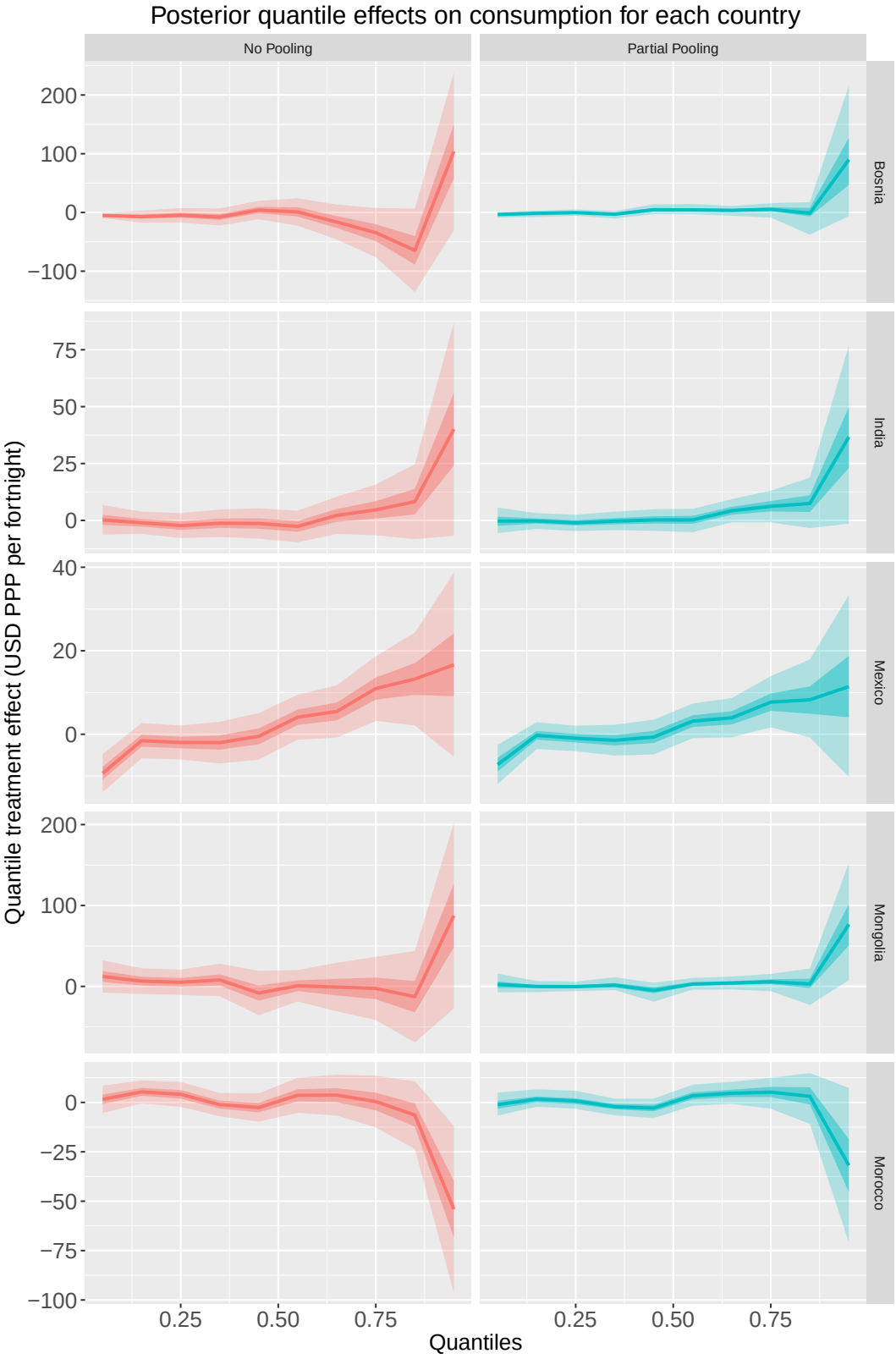

Figure 21: Site by site results for the consumption outcomes. [Back to main]

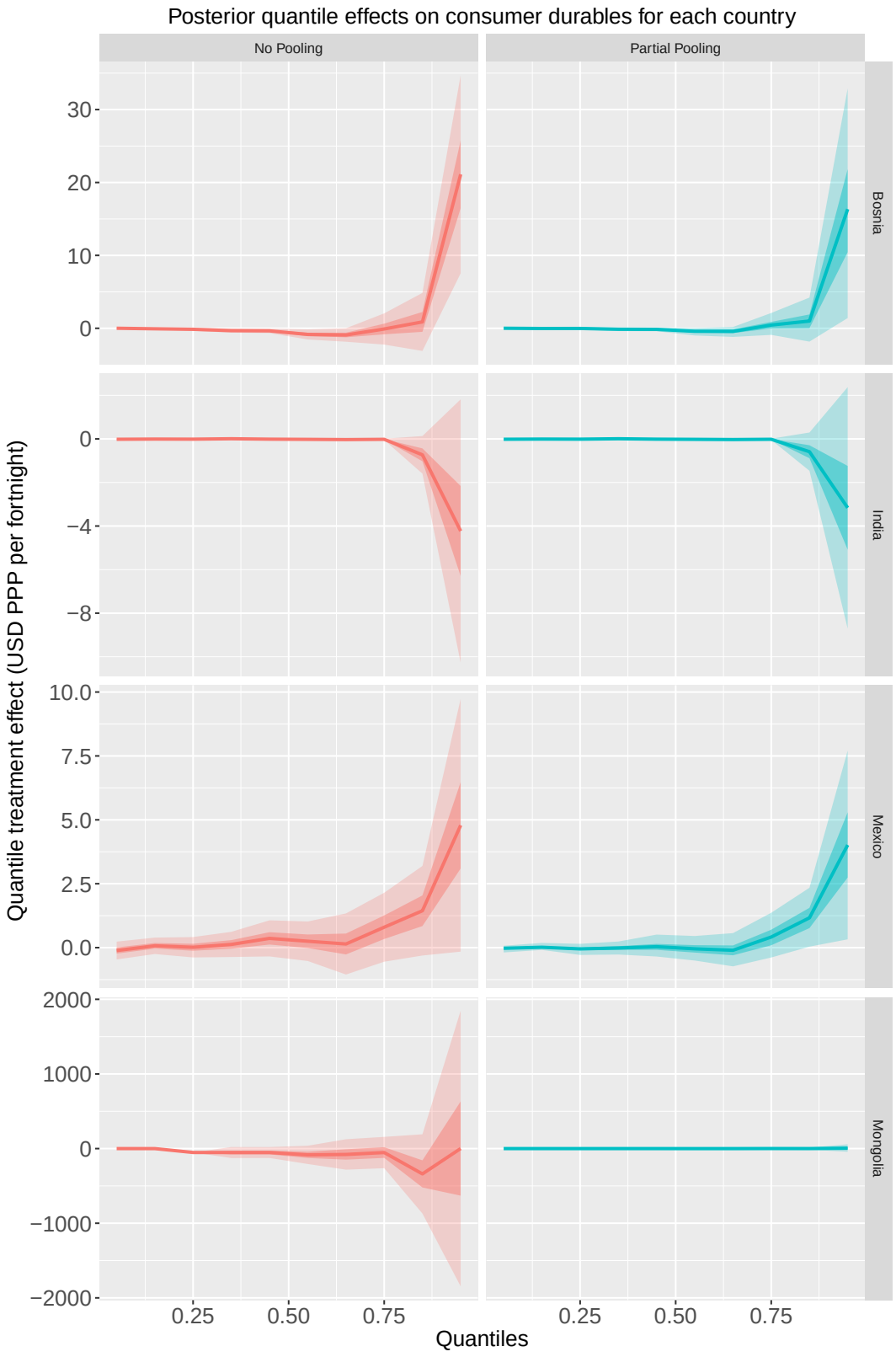

Figure 22: Site by site results for the consumer durables outcomes.[Back to main]

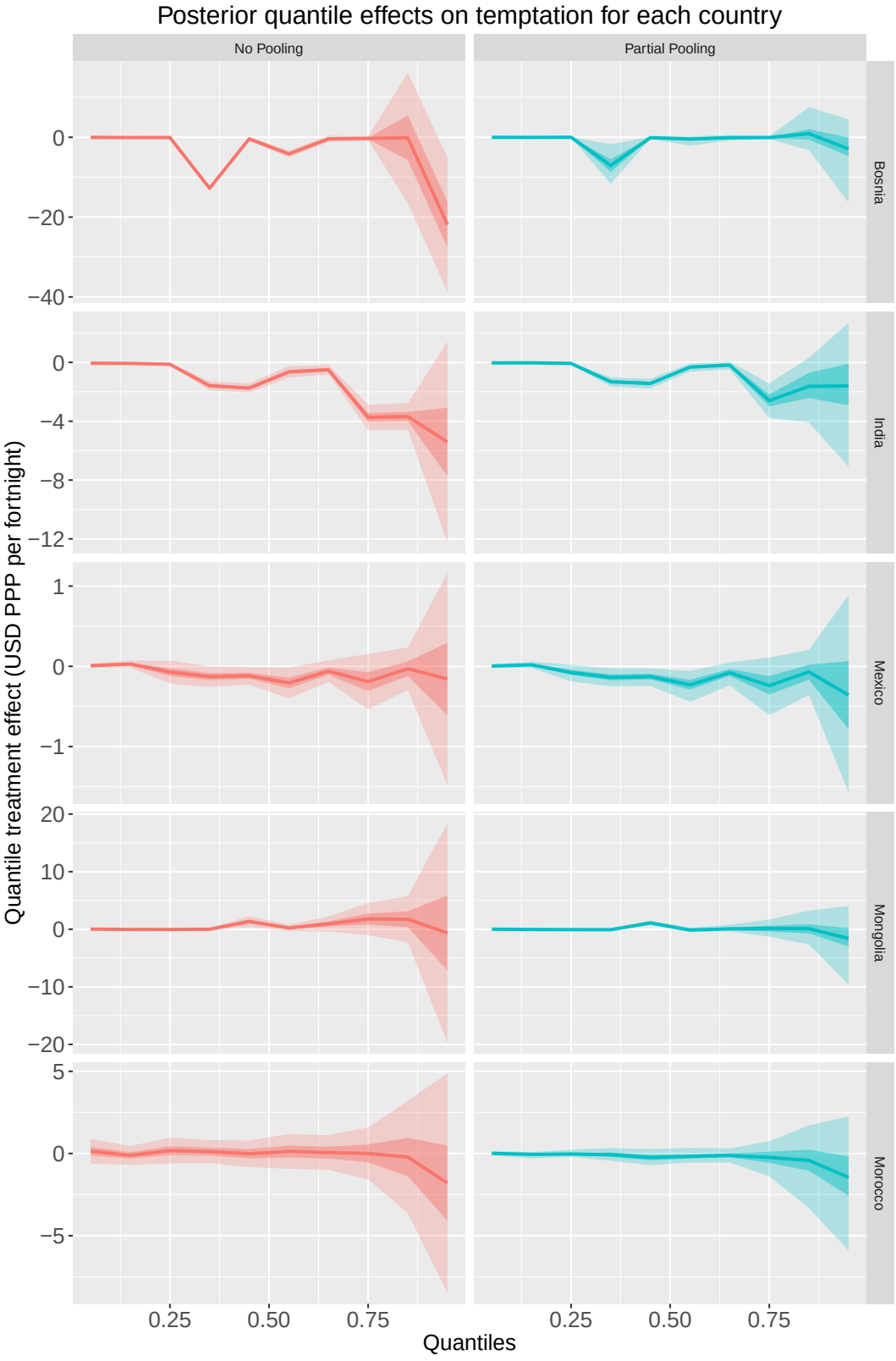

Figure 23: Site by site results for the temptation outcomes. [Back to main]

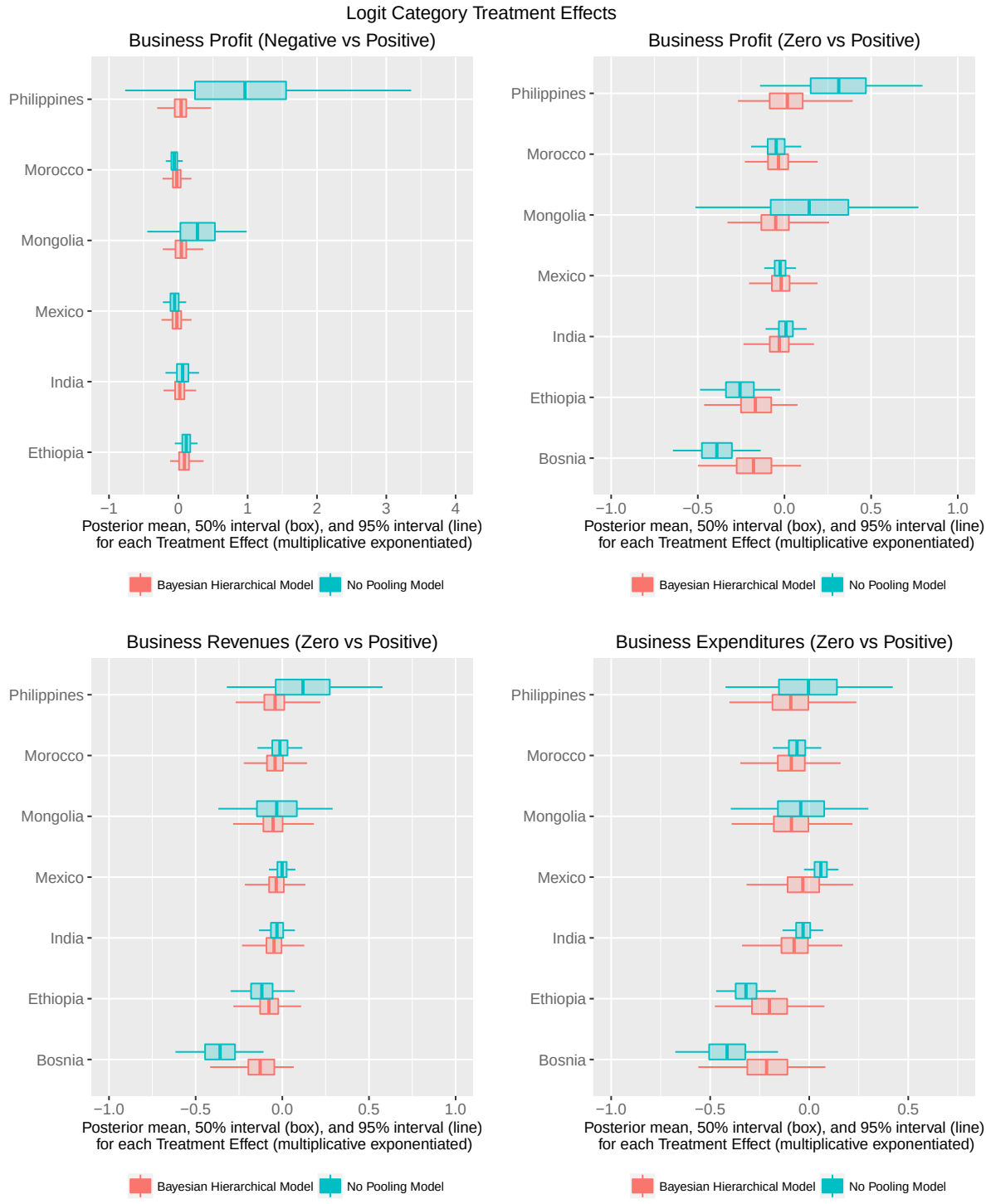

Figure 24: Posterior distributions for the logit treatment effects on category assignment in each site. [Back to main]

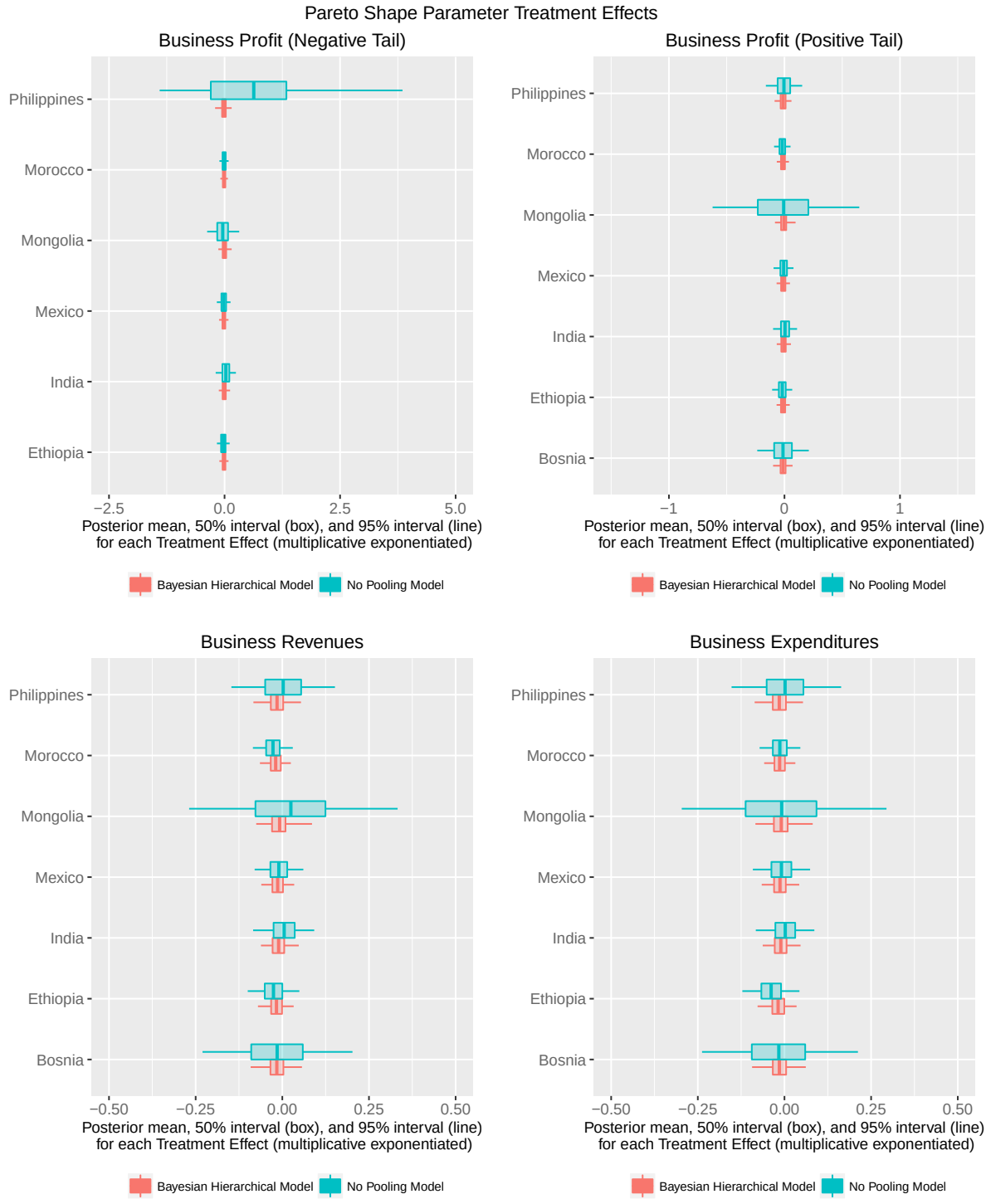

Figure 25: Posterior distributions for the Pareto shape treatment effects on category assignment in each site.[Back to main]

C Raw scale graphics for business variable quantile effects

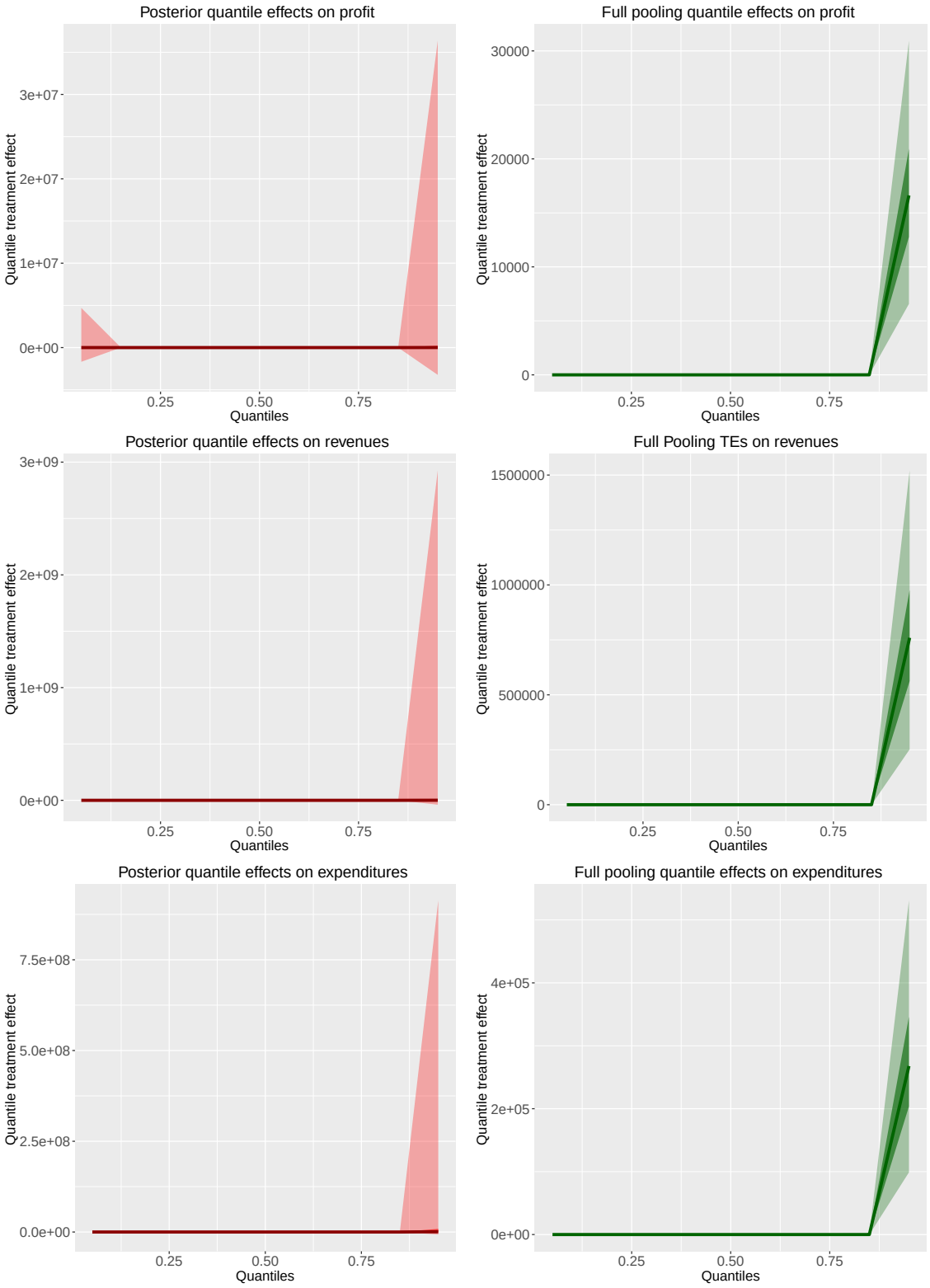

Figure 26: General Quantile Treatment Effect Curves (β_1) for business variables. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in USD PPP per two weeks to show the scale differences in the uncertainty at the right tail versus the rest of the distribution. [Back to main]

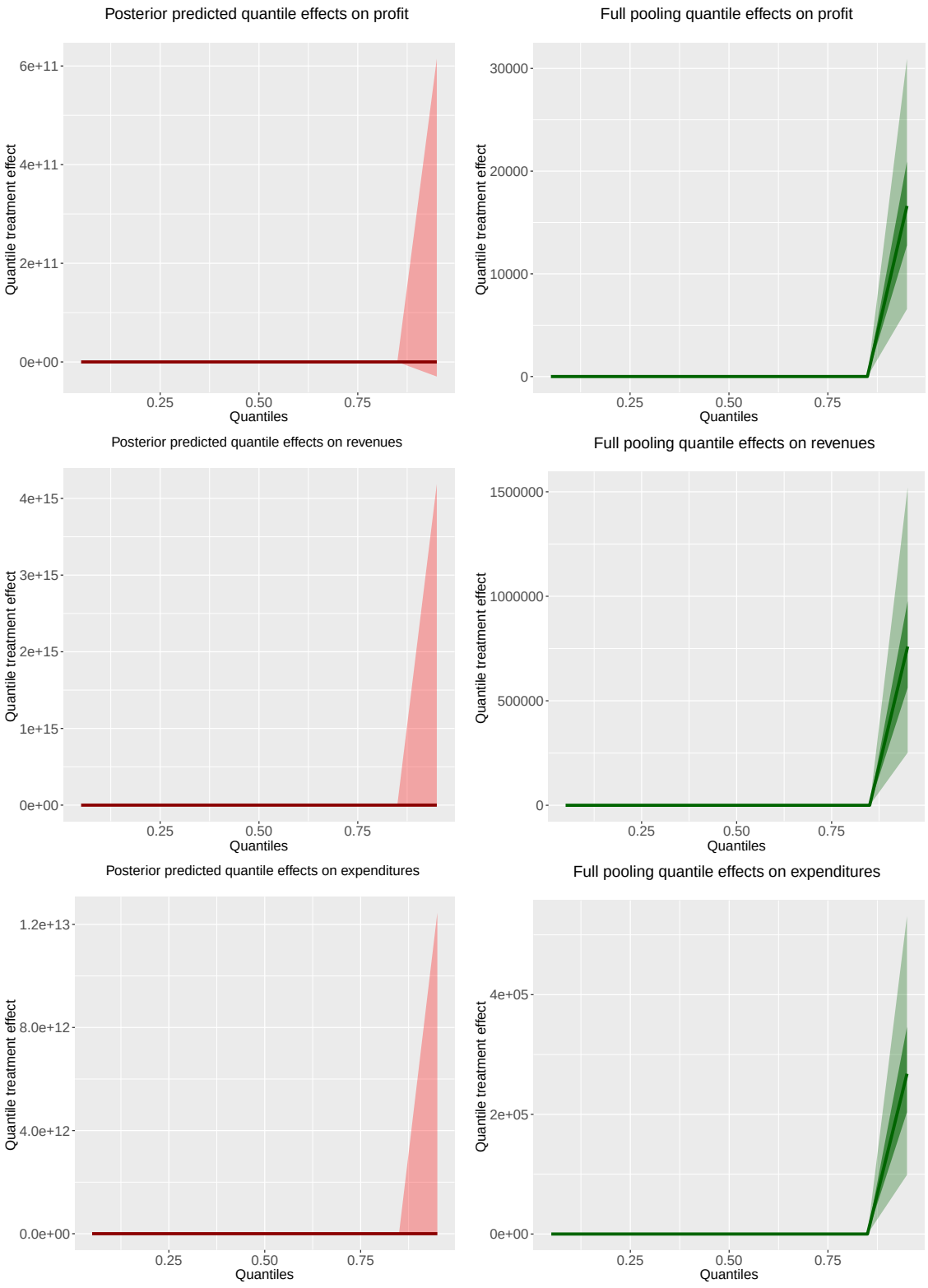

Figure 27: Posterior predicted quantile treatment effect Curves for Business Variables. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in USD PPP to show scale differences in the uncertainty at the right tail versus the rest of the distribution. [Back to main]

D Posterior Inference on Scale Parameters

Tailored Hierarchical PDF Model of Expenditures: Scale Parameters

parameter	mean	SD	quantiles: 2.5%	25%	50%	75%	97.5%
ρ	-2.15	0.10	-2.36	-2.21	-2.16	-2.1	-1.95
κ	-0.01	0.02	-0.06	-0.03	-0.01	0.00	0.04
ρ_1	-2.25	0.02	-2.3	-2.26	-2.25	-2.23	-2.20
ρ_2	-1.92	0.07	-2.07	-1.97	-1.92	-1.87	-1.78
ρ_3	-2.23	0.06	-2.35	-2.27	-2.23	-2.19	-2.12
ρ_4	-2.23	0.03	-2.29	-2.25	-2.23	-2.22	-2.18
ρ_5	-2.20	0.02	-2.23	-2.21	-2.2	-2.18	-2.16
ρ_6	-2.39	0.04	-2.47	-2.41	-2.39	-2.36	-2.3
ρ_7	-1.89	0.03	-1.94	-1.91	-1.89	-1.87	-1.84
κ_1	-0.01	0.03	-0.07	-0.03	-0.01	0.00	0.04
κ_2	-0.01	0.04	-0.08	-0.03	-0.01	0.01	0.08
κ_3	-0.01	0.04	-0.09	-0.03	-0.01	0.00	0.06
κ_4	-0.01	0.03	-0.06	-0.03	-0.01	0.01	0.05
κ_5	-0.01	0.02	-0.06	-0.03	-0.01	0.00	0.03
κ_6	-0.01	0.03	-0.09	-0.03	-0.01	0.00	0.05
κ_7	-0.02	0.03	-0.08	-0.04	-0.02	0.00	0.03
σ_κ	0.03	0.03	0.00	0.01	0.02	0.04	0.09
σ_ρ	0.24	0.10	0.12	0.18	0.22	0.28	0.51

Table 11: All parameters specified at the exponential level, hence, the scale parameter that enters the Pareto distribution is $\exp(\rho)$ and it is this value which cannot be below zero. Distributions for which $\exp(\rho)$ or $\exp(\rho+\kappa)$ are in the interval $[0, 2]$ have infinite variance. Parameter vector elements ordered alphabetically by author surname as follows: 1 = Angelucci et al. 2015 (Mexico), 2 = Attanasio et al. 2015 (Mongolia), 3 = Augsberg et al. 2015 (Bosnia), 4 = Banerjee et al. 2015 (India), 5 = Crepon et al. 2015 (Morocco), 6 = Karlan and Zinman 2011 (Philippines), 7 = Tarozzi et al. 2015 (Ethiopia)). The columns are in order as follows: the posterior mean, standard deviation of the posterior distribution, then the five remaining columns are the $\{2.5, 25, 50, 75, 97.5\}\%$ quantiles of the posterior distribution. All $\hat{R}$ values are less than 1.1 indicating good mixing between chains. [Back to main]

Tailored Hierarchical PDF Model of Revenues: Scale Parameters

parameter	mean	SD	quantiles: 2.5%	25%	50%	75%	97.5%
ρ	-2.12	0.13	-2.38	-2.19	-2.12	-2.05	-1.86
κ	-0.01	0.02	-0.06	-0.03	-0.01	0.00	0.03
ρ_1	-2.18	0.02	-2.23	-2.20	-2.18	-2.17	-2.14
ρ_2	-1.69	0.08	-1.85	-1.74	-1.69	-1.64	-1.55
ρ_3	-2.29	0.06	-2.40	-2.33	-2.29	-2.25	-2.18
ρ_4	-2.24	0.03	-2.30	-2.26	-2.24	-2.22	-2.19
ρ_5	-2.18	0.02	-2.22	-2.20	-2.18	-2.17	-2.15
ρ_6	-2.37	0.04	-2.45	-2.40	-2.37	-2.34	-2.29
ρ_7	-1.90	0.02	-1.94	-1.91	-1.90	-1.88	-1.85
κ_1	-0.01	0.02	-0.06	-0.03	-0.01	0.00	0.03
κ_2	-0.01	0.04	-0.08	-0.03	-0.01	0.01	0.09
κ_3	-0.02	0.04	-0.09	-0.03	-0.02	0.00	0.06
κ_4	-0.01	0.03	-0.06	-0.03	-0.01	0.01	0.05
κ_5	-0.02	0.02	-0.06	-0.03	-0.02	0.00	0.02
κ_6	-0.02	0.03	-0.08	-0.03	-0.02	0.00	0.05
κ_7	-0.02	0.03	-0.07	-0.03	-0.02	0.00	0.03
σ_ρ	0.31	0.13	0.16	0.23	0.28	0.36	0.63
σ_κ	0.03	0.03	0.00	0.01	0.02	0.03	0.09

Table 12: All parameters specified at the exponential level, hence, the scale parameter that enters the Pareto distribution is $\exp(\rho)$ and it is this value which cannot be below zero. Distributions for which $\exp(\rho)$ or $\exp(\rho+\kappa)$ are in the interval $[0, 2]$ have infinite variance. Parameter vector elements ordered alphabetically by author surname as follows: 1 = Angelucci et al. 2015 (Mexico), 2 = Attanasio et al. 2015 (Mongolia), 3 = Augsberg et al. 2015 (Bosnia), 4 = Banerjee et al. 2015 (India), 5 = Crepon et al. 2015 (Morocco), 6 = Karlan and Zinman 2011 (Philippines), 7 = Tarozzi et al. 2015 (Ethiopia)). The columns are in order as follows: the posterior mean, standard deviation of the posterior distribution, then the five remaining columns are the $\{2.5, 25, 50, 75, 97.5\}\%$ quantiles of the posterior distribution. All $\hat{R}$ values are less than 1.1 indicating good mixing between chains. [Back to main]

Tailored Hierarchical PDF Model of Profit: Scale Parameters
(Negative tail (subscript = 1) and Positive tail (subscript = 3)

parameter	mean	SD	quantiles: 2.5%	25%	50%	75%	97.5%
ρ_1	-1.96	0.13	-2.23	-2.03	-1.95	-1.88	-1.71
ρ_3	-2.01	0.12	-2.24	-2.08	-2.02	-1.95	-1.75
κ_1	-0.01	0.05	-0.11	-0.04	-0.01	0.02	0.09
κ_3	-0.01	0.03	-0.06	-0.03	-0.01	0.01	0.04
$\rho_{1,1}$	-2.12	0.04	-2.21	-2.15	-2.12	-2.09	-2.04
$\rho_{1,3}$	-2.09	0.03	-2.14	-2.11	-2.09	-2.07	-2.04
$\rho_{2,1}$	-1.72	0.09	-1.91	-1.79	-1.72	-1.66	-1.55
$\rho_{2,3}$	-1.62	0.16	-1.93	-1.73	-1.62	-1.51	-1.31
$\rho_{3,1}$	-1.96	0.34	-2.66	-2.13	-1.96	-1.79	-1.29
$\rho_{3,3}$	-2.21	0.06	-2.33	-2.25	-2.21	-2.17	-2.1
$\rho_{4,1}$	-2.08	0.06	-2.2	-2.12	-2.08	-2.04	-1.96
$\rho_{4,3}$	-2.10	0.03	-2.16	-2.12	-2.09	-2.07	-2.04
$\rho_{5,1}$	-2.02	0.03	-2.08	-2.04	-2.02	-2.00	-1.95
$\rho_{5,3}$	-2.07	0.02	-2.11	-2.08	-2.07	-2.05	-2.02
$\rho_{6,1}$	-2.04	0.21	-2.49	-2.16	-2.03	-1.9	-1.65
$\rho_{6,3}$	-2.21	0.04	-2.3	-2.24	-2.21	-2.19	-2.13
$\rho_{7,1}$	-1.76	0.04	-1.84	-1.79	-1.76	-1.73	-1.67
$\rho_{7,3}$	-1.80	0.03	-1.86	-1.82	-1.8	-1.78	-1.75
$\kappa_{1,1}$	-0.02	0.05	-0.12	-0.05	-0.01	0.02	0.09
$\kappa_{1,3}$	-0.01	0.03	-0.07	-0.03	-0.01	0.01	0.05
$\kappa_{2,1}$	0.00	0.07	-0.13	-0.04	0.00	0.04	0.15
$\kappa_{2,3}$	0.00	0.04	-0.08	-0.03	-0.01	0.02	0.10
$\kappa_{3,1}$	-0.01	0.09	-0.21	-0.05	-0.01	0.03	0.18
$\kappa_{3,3}$	-0.01	0.04	-0.10	-0.03	-0.01	0.01	0.07
$\kappa_{4,1}$	-0.01	0.06	-0.12	-0.04	-0.01	0.03	0.12
$\kappa_{4,3}$	-0.01	0.03	-0.07	-0.03	-0.01	0.01	0.06
$\kappa_{5,1}$	-0.01	0.04	-0.09	-0.04	-0.01	0.02	0.07
$\kappa_{5,3}$	-0.01	0.03	-0.07	-0.03	-0.01	0.00	0.04
$\kappa_{6,1}$	-0.02	0.08	-0.2	-0.05	-0.01	0.03	0.15
$\kappa_{6,3}$	-0.01	0.04	-0.09	-0.03	-0.01	0.01	0.06
$\kappa_{7,1}$	-0.01	0.05	-0.11	-0.04	-0.01	0.02	0.09
$\kappa_{7,3}$	-0.01	0.03	-0.07	-0.03	-0.01	0.01	0.05
σ_{ρ_1}	0.27	0.14	0.11	0.18	0.23	0.31	0.64
σ_{ρ_3}	0.29	0.13	0.13	0.2	0.26	0.34	0.62
σ_{κ_1}	0.06	0.06	0.00	0.02	0.04	0.07	0.21
σ_{κ_3}	0.03	0.03	0.00	0.01	0.02	0.04	0.10

Table 13: All parameters specified at the exponential level, hence, the scale parameter that enters the Pareto distribution is $\exp(\rho)$ and it is this value which cannot be below zero. Distributions for which $\exp(\rho)$ or $\exp(\rho+\kappa)$ are in the interval $[0, 2]$ have infinite variance. Parameter vector elements ordered alphabetically by author surname as follows: 1 = Angelucci et al. 2015 (Mexico), 2 = Attanasio et al. 2015 (Mongolia), 3 = Augsberg et al. 2015 (Bosnia), 4 = Banerjee et al. 2015 (India), 5 = Crepon et al. 2015 (Morocco), 6 = Karlan and Zinman 2011 (Philippines), 7 = Tarozzi et al. 2015 (Ethiopia)). The columns are in order as follows: the posterior mean, standard deviation of the posterior distribution, then the five remaining columns are the $\{2.5, 25, 50, 75, 97.5\}\%$ quantiles of the posterior distribution. All $\hat{R}$ values are less than 1.1 indicating good mixing between chains. [Back to main]

E Analysis of Site-Level Covariates

This section discusses the role of site-level covariates in predicting the remaining heterogeneity in the impact of microcredit across different studies. For a full discussion of the issues involved in this analysis, see section 5.2 of Meager (2016). I consider a model with many site-level contextual variables, although this is not exhaustive. In the order in which they appear in the X_k vector, they are: the site's average value of the outcome in the control group, a binary indicator on whether the unit of study randomization was individuals or communities, a binary indicator on whether the MFI targeted female borrowers, the interest rate (APR) at which the MFI in the study usually lends, a microcredit market saturation metric taking integer values from 0-3, a binary indicator on whether the MFI promoted the loans to the public in the treatment areas, a binary indicator on whether the loans were supposed to be collateralized, and the loan size as a percentage of the country's average income per capita. Table **??** displays the values taken by each of these variables in each site, although of course they must be standardized for any sparsity estimation procedure:

Contextual Variables (Pre-Standardization)

	Rand unit	Women	APR	Saturation	Promotion	Collateral	Loan size
Mexico (Angelucci)	0	1	100.00	2	1	0	6.00
Mongolia (Attanasio)	0	1	120.00	1	0	1	36.00
Bosnia (Augsburg)	1	0	22.00	2	0	1	9.00
India (Banerjee)	0	1	24.00	3	0	0	22.00
Morocco (Crepon)	0	0	13.50	0	1	0	21.00
Philippines (Karlan)	1	0	63.00	1	0	0	24.10
Ethiopia (Tarozzi)	0	0	12.00	1	0	0	118.00

Table 14: Contextual Variables: Unit of randomization (1 = individual, 0 = community), Women (1= MFI targets women, 0 = otherwise), APR (annual interest rate), Saturation metric (3 = highly saturated, 0 = no other microlenders operate), Promotion (1 = MFI advertised itself in area, 0 = no advertising), Collateral (1 = MFI required collateral, 0 = no collateral required), Loan size (percentage of mean national income). [Back to main]

For unidimensional treatment effects, the protocol is to proceed with a regularized regression of the treatment effect in each site on the standardized covariates as in Meager (2016). But for the multidimensional distributional treatment effects, there is no comparable established procedure to my knowledge. Therefore, the results of this appendix should be interpreted with caution, and future work on this topic is necessary to provide confidence in any of the conclusions presented here. Because the results of the main analysis in the consumption data have shown negligible impact of microcredit except in the right tail, and most notably at the 95th percentile, I have pursued a cross-site covariance analysis strategy that leverages this by performing a standard ridge procedure on the effects at this quantile. Similarly, for the business variables, the main variation across sites occured in the logit coefficients governing the category switching effect, so I focus the site-level covariate analysis on these coefficients.

The results of these selected ridge regressions at the study level are shown in figure **??**, which displays the absolute magnitude of the coefficients on the various contextual variables for each of the 6 outcomes. The larger the magnitude, the more important is the variable as a predictor of the treatment effects for that outcome (Hastie et al, 2009). In this case the

results are not as clear as in Meager (2016), perhaps reflecting weaknesses in the selected ridge analysis strategy employed in this section. However, even here the results appear to favour the economic variables over the study protocol variables. In particular, the logit switching effects are most strongly predicted by the loan size, and collateralisation seems to play a role in most cases. Although the randomization unit is almost as predictive as collateralization for the consumption variables, none of these variables are strongly predictive for these outcomes; note the difference in the absolute magnitude of the ridge coefficients shown in the two panels of the figure. This contrasts to the results of the means analysis in Meager (2016) which typically found the interest rate to have the highest predictive power, followed by the loan size. This may reflect weaknesses in the means analysis, especially in the case of the business variables which we now know to be fat tailed. However, as noted above, it may also reflect methodological issues with the ridge procedure chosen here.

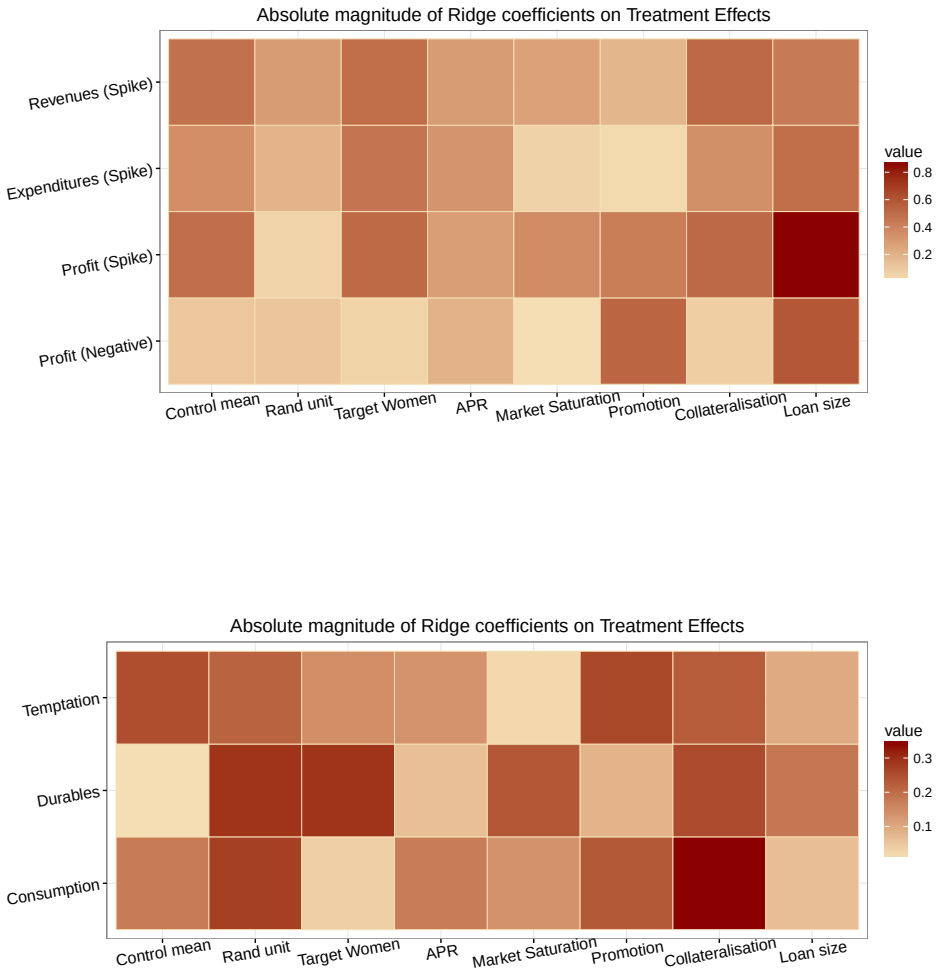

Figure 28: Absolute Magnitude of the Ridge Regression Coefficients for all outcomes and covariates [Back to main]

References

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

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

[3] 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.

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

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

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

[7] 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.

[8] 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.

[9] 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.

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

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

[12] Bazzi, S. (2016) "Wealth Heterogeneity and the Income Elasticity of Migration" American Economic Journal: Applied, Forthcoming 2016, https://www.aeaweb.org/articles?id=10.1257/app.20150548&&from=f

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

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

[15] 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

[16] Chernozhukov, Victor, IvÃ¡n Fernandez-Val, and Alfred Galichon.(2010) "Quantile and probability curves without crossing." Econometrica 78.3 1093-1125.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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

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

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

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

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

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

[29] 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

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

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

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

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

[34] 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

[35] 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.

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

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

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

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

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

[41] 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

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

[43] Meager, R. (2015). "Understanding the Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of 7 Randomised Experiments." Available at SSRN 2620834.

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

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

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

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

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

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

[50] Vivalt, E. (2015) "How much can we generalise from impact evaluations?" Working Paper, NYU

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

[52] Wickham, H. (2009) “ggplot2: elegant graphics for data analysis”. Springer New York, 2009.