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

Rachael Meager*†

July 24, 2018

Abstract

Recent studies of microcredit find evidence of both positive and negative treatment effects at certain quantiles of household outcome distributions, yet the findings differ across contexts. Using a Bayesian hierarchical framework, I develop new models to aggregate the evidence on distributional effects and assess the generalizability of the results. For continuous variables such as consumption, I provide a limited-information model in which information is passed across quantiles and monotonicity is enforced using variable transformation. For partially discrete outcomes like profit for which the sampling behaviour of quantiles is unknown, I aggregate using richly-parameterized mixture models applied to the full data. The results show a precise and generalizable zero effect from the 5th to 75th quantiles, then a large positive effect on the right tail which is more imprecise and heterogeneous across contexts. There are no generalizable negative impacts. A bounding exercise suggests that the precise zero is not due to low take-up, contrary to common belief. Covariate analysis shows that households who had previously operated businesses account for the majority of both the impact and the uncertainty in the tails of the consumption and profit distributions.

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

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

1 Introduction

Financial market expansions in the developing world have the potential to create winners and losers. Increasing access to credit in particular may have heterogeneous effects both because borrowers differ in their investment opportunities and because of general equilibrium dynamics (Banerjee 2013, Kaboski and Townsend 2011). Proponents of financial interventions such as microcredit claim that the positive impact on high-productivity borrowers justifies continued market expansion; detractors claim that the resulting "saturation" of credit markets leads to exploitative lending practices which systematically harm the most vulnerable borrowers (Ahmad 2003, Schicks 2013, Roodman 2012). The debate continues despite decades of research because the microcredit literature has focused on estimating average treatment effects which cannot evince this heterogeneity. Although recent studies have estimated sets of quantile treatment effects in order to detect heterogeneous impacts, the findings differ across contexts, impeding the formulation of any general consensus and leaving open the possibility of cherrypicking results (Banerjee et al 2015a, Crepon et al 2015, Angelucci et al 2015). Existing meta-studies of microcredit ignored these sets of quantile effects due to a lack of methodology to aggregate them (Meager 2018, Vivalt 2017, Banerjee et al 2015b, Duvendack et al. 2014). In this paper I develop these methods and aggregate the evidence on the distributional treatment effects of expanding access to microcredit.

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

Aggregating the evidence on these distributional effects across contexts can provide a more reliable indication of the typical impact that microcredit is likely to have in future policy contexts. Combining information from multiple studies improves power and prevents cherrypicking, that is, undue focus on the most extreme effects in the literature,

which may be the least likely to replicate (Rubin 1981). However, this potential heterogeneity across sites makes it difficult to combine the microcredit studies: pooling the data and performing one analysis is likely to underestimate the true uncertainty (Gelman et al 2004). As is common in empirical social science, the extent of true heterogeneity in the impact of microcredit across settings is unknown, yet it influences the optimal combination of evidence. Thus, aggregation raises the question of external validity: the extent to which the recorded impact of a policy in one setting predicts its impact in another, and the extent to which the average of all effects predicts the next effect (Allcott 2015, Bisbee et al 2016). If microcredit has different distributional impacts in different contexts, it may be imprudent to use results from one context to guide policy another context (Pritchett and Sandefur 2015). It is therefore important to use an aggregation method that does not rely on unwarranted assumptions about the extent of heterogeneity in effects across studies.

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

The hierarchical approach is well established in statistics and is increasingly used for evidence aggregation in economics (Dehejia 2003, Hsiang, Burke & Miguel 2013, Vivalt 2016, Bandiera et al 2017, Meager 2018). The implementation is typically Bayesian due to the potential for improved performance in practice, especially when there are few studies (Rubin 1981, Chung et al 2013). Yet within this framework there are no tools to aggregate distributional effects such as sets of quantile treatment effects. Even outside of the hierarchical framework, the economics literature on external validity and generalizability has focused on different kinds of average effects (Heckman, Tobias, & Vytlacil 2001, Angrist 2004, Angrist & Fernandez-Val 2010, Bertanha & Imbens 2014, Allcott 2015, Dehejia, Pop-Eleches and Samii 2015, Gechter 2015, Athey & Imbens 2016, Andrews and Oster 2018). This necessitates the development of new methods for aggregating evidence on distributional treatment effects in the presence of treatment effect heterogeneity.

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

Aggregation of distributional effects presents new challenges relative to average effects. When the metric of interest is a set of quantile treatment effects, it seems intuitive to build quantile aggregation models based on existing Bayesian hierarchical models for average effects such as Rubin (1981). These limited-information models use the knowledge that the within-study sampling variation of both means and quantiles is often asymptotically Gaussian (Mosteller 1946). However, an aggregation model for sets of quantiles must also pass information across the quantiles at every level of the model because neighbouring quantiles typically contain information about each other and must be monotonically increasing. Failure to incorporate this structure may result in the quantile crossing problem described in Chernozhukov et al 2010.[2] To solve this problem I exploit the fact that Bayesian inference treats unknown objects as random variables: the quantile treatment effects can be constrained to imply monotonic outcome quantiles by transforming the implied unconditional quantiles using functions that only have support on monotonic vectors. The transform passes information across the neighbouring quantiles via the constraint, and all posterior uncertainty is automatically preserved within the Bayesian framework, in contrast to ex-post rearrangements or smoothing strategies (He 1997, Chernozhukov et al 2010).

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

Applying these models to seven randomized trials of expanding access to microcredit, I find a precise zero effect on household outcomes from the 5th to 75th percentiles. Above the 75th percentile, there is substantial probability of a large positive impact on most outcomes, but there is greater uncertainty around this effect due to greater heterogeneity within and across studies. Thus, I find some evidence of the potential for positive effects and no evidence of systematic harm to any group of borrowers, as there are no generalizable

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

negative quantile effects at any part of the distribution. Part of the greater uncertainty in the tails is due to the tail shape of the business variables, which are so heavy that average treatment effect analysis and Gaussian asymptotics are likely to be unreliable on this data (Koenker and Basset 1978, Mosteller 1946). The likelihood of large, economically important increase in household profits and consumption is much greater than the chance of a zero or negative impact, but there is substantial uncertainty about the specific effect size in any context. By contrast, classical full pooling methods applied to the same data declare "statistically significant" effects in the right tail of business variables as a result of ignoring the heterogeneity across sites.

To better understand these results, I pursue a bounding exercise to show that the precise and generalizable impact at zero is unlikely to be due to low take-up of loans. A covariates analysis reveals that both the majority of the right tail impact of microcredit and the heterogeneity across studies occurs within the group of households who had previous business experience. This group of experienced households records an increase in consumption above the 75th percentile in the general case, although there is still a precise zero effect below this point. This overall pattern suggests that most households' lives are not transformed by microcredit access, and this result holds even among those who take up the loans or have previous business experience. Expanding access to credit does create potential improvements in consumption and profit for some of these households, although any such improvements would be accompanied by increases in inequality within the community; the social welfare effects of microcredit are complex.

These results demonstrate the value of analysing and aggregating evidence using appropriate methodology. The models developed in this paper could be applied to study the distributional effects of other financial interventions, trade and innovation policies, educational subsidies and local migration incentives, all of which have social welfare implications (Borusyak and Jaravel 2018, Duflo, Dupas and Kremer 2017, Chetty, Hendren, and Katz 2016, Bryan, Chowdhury and Mobarak 2014, Autor et al 2014, Katz, Kling and Leibman 2001, Autor, Katz and Krueger 1988). Quantile regression and heavy-tailed parametric models can accommodate the heavy tails found in many economic data sets, particularly for earnings and assets (Bazzi 2016, Pancost 2016, Gabaix 2008, Fama 1965). By contrast, common implementations of average treatment effects analysis and even subgroup analysis such as ordinary least squares regression are not robust to heavy tails and typically provide inaccurate results when applied to such distributions (Koenker and Basset 1978). Recent evidence that inequality can persist or even worsen in equilibrium due to spillover effects on social cohesion or even scientific innovation suggests that understanding these distributional effects is a first-order concern (Folgi and Guerrieri 2018, Chetty and Hendren 2018, Bell et al 2017). In these settings, multi-study aggregation of quantile effects using the methods I provide here can deliver inference that is both more informative and more reliable than analyses of average treatment effects alone, or of the heterogeneous effects in any single study.

2 Data and Context

The rapid increase in the scale of microcredit operations over the past 30 years has lead to a large academic literature studying credit market interventions in the developing world (Banerjee, 2013). I consider seven of these studies which meet the following inclusion criteria: the main intervention must be an expansion of access to microcredit either at the community or individual level, the assignment of access must be randomized, and the study must be published before February 2015 (the period of my literature search). The selected studies are: Angelucci et al. 2015, Attanasio et al. 2015, Augsburg et al. 2015, Banerjee et al. 2015b, Crepon et al. 2015, Karlan and Zinman 2011, and Tarozzi et al. 2015, six of which were published in a special issue of the *American Economics Journal: Applied Economics*.[3] I focus on expanding access to microcredit because this is the intervention closest to the policy of subsidizing microfinance institutions (MFIs) or promoting interventions under the general umbrella of "microcredit". I restrict the sample to randomized controlled trials (RCTs) because they typically have high internal validity for estimating causal effects, and because as yet there is no established methodology designed to aggregate both RCTs and observational evidence in a single framework.[4]

I focus on the economic outcomes most directly implicated by the theoretical benefits of microfinance. These include: household business expenditures, business revenues, and business profits, household consumption, consumer durables spending and temptation goods spending. All six of these outcomes are linked to the core claim that offering households more credit on more favourable terms should stimulate entrepreneurship (Morduch 1999, Yunus 2006, Roodman 2012). Because microfinance insitutions (MFIs) offer lower interest rates relative to informal moneylenders, poor entrepreneurs may be able to start new businesses or grow their existing businesses, increasing their business expenditures, revenues and ultimately profits (Yunus 2006). Greater economic prosperity should enable households to increase their consumption in the medium and long run. Yet even households without business investment opportunities may use microloans to shift spending away from temptation goods and towards durable goods (Roodman 2012, Banerjee 2013). Such consumption transformation might arise if access to microcredit increases a household's expectation of escaping poverty in the future, or if microcredit solves a self-control problem (Banerjee and Mullainathan 2010, Banerjee 2013). Thus, these six variables should capture the main channels through which relaxing credit constraints for households in the developing world may have positive consequences.

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

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

Yet even for these outcome variables, there is potential for microcredit to have different effects on different households. When a new MFI enters a community, households differentially select into loan take-up, and those who do take up are likely to experience heterogeneous consequences depending on how they use the loan or whether they experience shocks (Banerjee 2013). Microcredit may not have any impact for certain households because the amount they can borrow may be too small relative to a lumpy investment opportunity, or the terms of the loan may be restrictive and undesirable for investment purposes, or the term to maturity may be too short (Banerjee 2013). But even when borrowers do benefit on average, there may be the potential for winners and losers in general equilibrium due to effects on wages or displacement of informal lending by local savers who now receive lower returns on their savings (Kaboski and Townsend 2011, Morduch 1999). It may be that multiple microlenders into a community can lead to predatory lending practices and "overlending" to households who cannot feasibly repay the loan (Shicks 2013, Ahmad 2003). This concern underlay the "No Paygo" movement against microcredit in Nicaragua and was ostensibly part of why the government of Andhra Pradesh shut microcredit down during the crisis of 2010 (Microfinance Focus 2011, Banerjee 2013). It is therefore plausible that microcredit access could have zero impact, large positive or large negative impacts for different types of households. Even if the groups of households who experience large effects are small, the social welfare consequences could be substantial, particularly if economic inequality across households is affected.

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

Although estimating sets of quantile effects is not the only way to detect heterogeneous effects, it serves as a reasonable first pass at the problem (Banerjee et al 2015). It is often sensible to estimate the quintile or decile effects regardless of the underlying proposed source of the heterogeneity, because this provides an estimate of the causal impact on the entire distribution of outcomes. This approach prevents cherrypicking of subgroups and allows little room for selective reporting. Although the set of quantile treatment effects

does not estimate the quantiles of the distribution of individual treatment effects, heterogeneity in the effects on different quantiles is evidence of heterogeneity in these individual effects. The quantiles approach also permits the detection of heterogeneity not predicted by observables, which is particularly important for microcredit, an intervention for which even households which are closely matched on covariates could have considerably different treatment effects (Kaboski and Townsend, 2011). Moreover, the presence of general equilibrium effects make it important to assess the impact of treatment on the overall shape of the distribution and to characterise heterogeneity based on relative position (ranks or quantiles) rather than absolute values of covariates. The quantile strategy also permits detection of changes to economic inequality within a community, without having to reference baseline data which may not be available (indeed, it is not available for many of the microcredit studies). Sample quantiles also have desirable robustness properties relative to sample means, particularly in the presence of heavy-tailed underlying data (Koenker and Basset 1978).

Despite the restrictive inclusion criteria, the selected studies still differ substantially in their implementations and local contexts (table 1). They cover seven different countries, they have different partner NGOs, offering similar but not identical loan contract structures with different interest rates and loan sizes, and they differ in terms of their randomization units - five randomized at the community level and two at the individual level - with various encouragement and sampling designs. Given this heterogeneity across studies, there is little justification for assuming homogeneous average effects or quantile treatment effects. However, the 95% confidence intervals of the quantile effects do overlap across most of the studies, suggesting there may be meaningful similarities in the underlying effects. This pattern was also observed in the average treatment effects in these studies, which turned out to have only moderate underlying heterogeneity despite these contextual differences (Meager 2018). However, similarities in the average treatment effects may be uninformative about the true generalizability of the effects if indeed these averages are composed of heterogeneous quantile effects. In this context, where the generalizability of the evidence across settings is unclear ex-ante, the Bayesian hierarchical framework is an appropriately cautious way to proceed with evidence aggregation.

The open data policies of the *American Economics Journal: Applied Economics* and *Science* allow me access to the microdata from all of these experiments, such that I can standardize which quantiles I compute across studies and can construct each underlying variable in a uniform manner across studies. The variables were measured in different currencies, in different years, and over different time periods (this matters because these are all flow variables). I standardize all measurements to be USD PPP in 2009 dollars over a two-week period, which is the shortest time period recorded in any study. Business variables require further standardization: to capture the potential for microcredit to allow individuals to open new businesses or to switch to operating any existing seasonal businesses throughout the year, households with no business or missing business data have

profits imputed as zero. This was the decision made by the original authors of many of the seven studies, although not all of them. Because the business creation channel is closely tied to Yunus' claims about the Grameen Bank, I employ this strategy throughout and apply it to business expenditures and revenues as well. Household consumption variables do not require imputation within sites, but unfortunately were not measured in all sites. However, as consumption behaviour is relevant to the welfare impact of microcredit, these variables must be analyzed regardless (Banerjee 2013). The six variables selected here are measured in reasonably comparable ways across sites. While it would be ideal to examine effects on other variables such as income and assets, the measurement and definition of those variables differed across the studies to such an extent that it is unclear how to aggregate them.[5] 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.

It would be useful to understand the role of household and study-level covariates in determining both the observed outcomes and the quantile treatment effects, but there are limitations to pursuing a covariates analysis in this literature. Only three of the microcredit RCTs collected comprehensive individual-level baseline surveys, and many household covariates recorded in the endlines could plausibly have been affected by credit access. However, one pre-treatment variable was recorded at endline in all studies due to its theoretical importance: a binary indicator that a household had previous experience operating a business (Banerjee et al 2015). Although each study also recorded a binary indicator of whether a household takes up a loan, this decision is downstream of loan access and therefore cannot be entered as a simple control (Acharya, Blackwell and Sen 2016). The network links between households, the potential for general equilibrium effects, and the impact of the mere expectation of taking up credit in the future even if one does not take it up today means that the Stable Unit Treatment Value Assumption (SUTVA) is likely to be violated within a community that experiences any increase in access (Banerjee 2013, Kinnan and Townsend 2012, Breza 2012). Hence I analyse the effect of expanding access itself, which was often called the Intention to Treat Effect in the original studies (Banerjee et al 2015b). To investigate the role of take-up in this context, I pursue a bounds analysis explained further in section 4.3. Although covariates at the study level may also predict variation in effects across context, there at least seven such covariates and only seven studies, so conventional regression analysis will be overfitted and misleading.[6] It is still useful to aggregate the evidence without conditioning on covariates, as this permits an understanding how much unconditional heterogeneity there is; if there is little or no variation across settings, further analysis becomes a less pressing concern for future work.

Other than standardizing the definition and construction of variables as much as pos-

[5] This issue was noted in Meager (2018) and in my pre-registration accessible on the OSF website at https://osf.io/tdvc8/ .

[6] In Appendix E I provide the results of a Ridge regression analysis on this question, but caution against interpreting these results too strongly.

sible, in most other respects I have attempted to conform to the decisions made by the original authors themselves. Certain potential issues such as attrition or sample selection were left as they were in the studies themselves and in most cases I have used the entire sample available in the online data sets.[7] I do however analyse the impact of prior business experience on outcomes even in those studies which did not report it, because they all recorded the necessary underlying information; replicating and standardizing this subgroup exploration mitigates the risk of selective analysis leading to false positives. I do not winsorize any of the variables because most of the studies did not do so, and Augsburg et al (2015) found that winsorizing outliers sometimes made results statistically significant when they were not significant in the full sample. If the extreme values do not change the point estimate but increase the uncertainty, winsorising them may underestimate the true uncertainty. As my analysis shows, the behaviour of the upper tails turn out to play an important role in determining the impact of microcredit.

3 Methodology

3.1 Bayesian Hierarchical Models

3.1.1 Hierarchical Models

Consider a body of evidence consisting of K studies indexed by k, each of which provides some k-specific data $\mathcal{Y}_k$ about a given policy intervention. Together, the K data sets contain 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 evidence aggregation is that there are often benefits to conducting inference on all the unknown θ_k parameters together and borrowing information across the K studies (Rubin 1981, James and Stein 196, Stein 1956). This can be expressed by positing the potential existence of some general parameter $\theta \in \Theta$ which is common across study sites at the population level. Typically, θ is specified as the expected value

[7] Ethiopia is the only exception: this study contained a cross-randomized family planning treatment. I use only the pure control and the pure microcredit samples, which is the conservative choice given that we do not know how microcredit interacts with family planning (the study estimates a very imprecise interaction).

of any θ_k in the set of studies before the outcome is known, so $\theta = E[\theta_k]$.[8] One can learn about this θ using the evidence on $\{\theta_k\}_{k=1}^K$, but the optimal learning procedure depends on the heterogeneity or dispersion of $\{\theta_k\}_{k=1}^K$ around θ, denoted Σ_θ (Rubin 1981). 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.[9] 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.

Joint estimation of θ and Σ_θ is the core challenge of aggregation across studies. Before aggregation occurs, the data has been analyzed separately in each study: this constitutes a "no pooling" model, where each effect θ_k is estimated using only the data from its own site, $\mathcal{Y}_k$. The resulting estimates, denoted $\{\hat{\theta}_k\}_{k=1}^K$, are only optimal for the set $\{\theta_k\}_{k=1}^K$ if $K < 3$ and if indeed no general common parameter θ exists.[10] The heterogeneity of $\{\hat{\theta}_k\}_{k=1}^K$ is generally biased upwards for Σ_θ because it includes the sampling variation of each $\hat{\theta}_k$ around its θ_k (Stein 1951, James and Stein 1961). These estimates or the underling data must be combined in some way to estimate θ, Σ_θ and θ_{K+1}. A "full pooling" aggregation method is an estimation procedure for θ which uses all the data $\mathcal{Y}$ and assumes that $\theta_k = \theta_{k'} \; \forall \; k, k'$. This assumption may be made explicitly or implicitly: any estimator that does not leverage the K-site structure nor estimate Σ_θ is a full pooling estimator. A "partial pooling" estimator uses the full data $\mathcal{Y}$ to estimate θ but does not assume $\theta_k = \theta_{k'} \; \forall \; k, k'$. A partial pooling aggregation procedure provides estimates of θ, Σ_θ as well as new estimates of $\{\theta_k\}_{k=1}^K$ produced by transferring some information across sites, denoted $(\tilde{\theta}, \tilde{\Sigma}_\theta, \{\tilde{\theta}_k\}_{k=1}^K)$.

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

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

The "upper level" of the model describes the potential for statistical dependence between

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

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

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

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

While in principle the hierarchical model could be specified with a nonparametric likelihood, a parametric structure is often preferable in low-data environments, such as evidence aggregation with a small or moderate number of studies.[11] Any partial pooling model must impose some structure to determine the extent of the pooling and how the pooling will be informed by the data. If the analyst faces a low-data environment at the cross-study level, this structure must not be too flexible or the model risks overfitting the scarce data that is available. Nonparametric methods often lack the power to deliver reliable inference at the general level. 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

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

parametric upper-level structure only because they impose such strong assumptions about Σ_θ. Parametric hierarchical likelihoods relax the assumptions on Σ_θ without providing too many degrees of freedom relative to the number of studies being aggregated.

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

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

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

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

Since these functional forms reflect underlying knowledge of the data or statistics being studied, hierarchical models based on these structures can more effectively separate the sampling variation from the between-study variation in effects (Rubin 1981). With the local variation specified in a particular way, the choice of the parent distribution that governs the between-study heterogeneity in effects can now be made in view of tractability and performance properties measured by the Mean Squared Error. These considerations typically motivate the use of Gaussian structure at the upper level of the model because they implement beneficial forms of shrinkage across the studies and have been shown to perform well for a variety of problems (McCullough and Neuhaus 2011, Efron and Morris 1975, Gelman et al 2004). In particular, if one is concerned with inference on only location and scale parameters, McCullough and Neuhaus (2011) shows that the Gaussian performs well even if the true underlying distribution is not Gaussian.[12]

Hierarchical models do require that $\{\theta_k\}_{k=1}^K$ be "exchangeable", such that their joint distribution is invariant to permutation of the indices (Diaconis, 1977). This means the analyst must have no knowledge of the ordering or any sub-clustering of the treatment effects *a priori* that is not specified in the model (Rubin 1981). If economic theory demands that a particular covariate should be correlated in a certain way with the treatment effects, that can be translated into *conditional* exchangeability by introducing this covariate into

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

the model. Yet theory and prior knowledge rarely provide certainty about these relationships, and building sufficiently weak structure that still permits inference on the role of covariates is typically challenging in a low-data environment. In the absence of strong prior knowledge about the treatment effects, exchangeability is a reasonable structure to impose (Gelman et al 2004). Any future site for which θ_{K+1} is used to predict the effect must be exchangeable with the sites in the sample for this prediction to be valid, which is generally a requirement for predicting out-of-sample effects (see for example Allcott (2015)).

3.1.1.1 Pooling Metrics for Hierarchical Models

The hierarchical framework also provides several natural metrics to assess the extent of pooling across sites shown in the posterior distribution (Gelman et al. 2004, Gelman and Pardoe 2006). In the context of multi-study aggregation, 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^2_θ. 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):

$$\omega(\theta_k) \equiv \frac{\hat{se}^2_k}{\tilde{\sigma}^2_\theta + \hat{se}^2_k}. \tag{3.6}$$

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 $\tilde{\sigma}^2_\theta$ to the magnitude of $\hat{se}^2_k$, the sampling variation in the no-pooling estimate of the treatment effect from site k. Here, $\omega(\theta_k) > 0.5$ indicates that $\tilde{\sigma}^2_\theta$ is smaller than the sampling variation, indicating substantial pooling of information and a "small" $\tilde{\sigma}^2_\theta$. 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:

$$\breve{\omega}(\theta_k) \equiv \{\omega : \ \tilde{\theta}_k = \omega\tilde{\theta} + (1-\omega)\hat{\theta}_k\}. \tag{3.7}$$

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:

$$\lambda_\theta \equiv 1 - \frac{\frac{1}{K-1}\sum_{k=1}^{K}(E_{post}[\epsilon_k] - \overline{E_{post}[\epsilon_k]})^2}{E_{post}[\frac{1}{K-1}\sum_{k=1}^{K}(\epsilon_k - \bar{\epsilon_k})^2]}. \tag{3.8}$$

The denominator is the posterior average variance of the errors, and the numerator is the variance of the posterior average error across sites. If the numerator is relatively large then there is 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.

3.1.2 Bayesian Implementation

While hierarchical models can be estimated using frequentist methods, in practice Bayesian inference offers several advantages. The major strength 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 (Rubin 1981). This conditioning is required for tractability in the maximum likelihood estimation (MLE) framework as it is commonly implemented, because of the nonlinear interdependencies between $\{\theta_k\}_{k=1}^K$, θ, and Σ_θ.[13] 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:

$$f(\theta|\mathcal{Y}) = \frac{\mathcal{L}(\mathcal{Y}|\theta)\mathcal{P}(\theta)}{\int_\Theta \mathcal{L}(\mathcal{Y}|\theta)\mathcal{P}(\theta)d\theta}. \tag{3.9}$$

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 ensures 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 the details 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 or "weakly informative" (Gelman et al 2008). If there is substantial expert knowledge of the likely values before seeing the data, perhaps from economic theory or previous studies, this can be incorporated using stronger priors. Even if the prior distributions introduce some bias due to incorrect centering, they may still improve the mean squared error of the estimation by reducing the variance: the prior regularizes the estimates (Chung et al. 2013, 2015). In low-data environments such as the cross-study level of the hierarchical model, overfitting and high variance can be the major obstacle to making reasonable inferences or predictions. Here, as in many other statistical problems, regularization towards zero often improves performance (Hastie, Tibshirani and Friedman 2009, section 10.2).

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

Bayesian inference also provides a useful framework for decision-making about policy and future research. The distribution of the treatment effect in a hypothetical future site θ_{K+1} is often the object of most interest for policymakers, but the distribution of this object must be computed accounting for the full joint posterior uncertainty rather than conditioning on a particular point estimate 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 averages over the posterior uncertainty on the unknowns (θ, Σ_θ). Formally, the posterior predictive distribution is:

$$f(\theta_{K+1}|\mathcal{Y}) = \int \psi(\theta_{K+1}|\theta) f(\theta|\mathcal{Y}) \mathrm{d}\theta \tag{3.10}$$

The Bayesian framework is well-suited to providing these objects because the task 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. From a decision-theoretic perspective, if we wish to conduct cost-benefit analyses or make policy accounting for our uncertainty about any of these unknown parameters the correct object to take expectations over is the posterior distribution of the parameters, not the sampling distribution of some chosen estimator.

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

$$\mathcal{L}_{\mathcal{A}_\Theta}(\mathcal{Y}|\theta) = \mathcal{L}(\mathcal{Y}|\theta) \cdot \mathbb{1}\{\theta \in \mathcal{A}_\Theta\}. \tag{3.11}$$

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

$$p(\theta^*) = p_\theta(\mathrm{f}^{-1}(\theta))|det(J_{\mathrm{f}^{-1}}(\theta))|. \tag{3.12}$$

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

implementation ensures coherence because the constraint is imposed on the parameter itself and is thus accounted for in both estimation and inference using the resulting joint posterior of all unknowns.

Tractability issues can arise in Bayesian inference on hierarchical models due to the same issues that lead frequentists to adopt Empirical Bayes, these can often be surmounted by the use of Markov Chain Monte Carlo (MCMC) methods. These methods construct a Markov chain which has the posterior distribution as its invariant distribution, so that in the limit, the draws from the chain are ergodic draws from the posterior. This chain is constructed by drawing from known distributions at each "step" and using a probabilistic accept/reject rule for the draw based on the posterior distribution's value at the draw. While these chains always converge to the correct distribution in the limit, popular algorithms such as the Metropolis-Hastings or Gibbs samplers can be prone to inefficient random walk behavior when the unknowns are correlated, as with hierarchical models. Instead, I use Hamiltonian Monte Carlo (HMC) methods, which 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 (Hoffman and Gelman, 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, 2017).

3.2 Limited Information Asymptotic Quantile Models

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:

$$Q_Y(u) = F_Y^{-1}(u). \tag{3.13}$$

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

$$Q_{y_{nk}|T}(u) = \beta_{0k}(u) + \beta_{1k}(u)T_{nk} \tag{3.14}$$

For a single quantile u, the treatment effect is the univariate parameter $\beta_{1k}(u)$. If there is only one quantile of interest, a univariate Bayesian hierarchical model can be applied, as in Reich et al (2011). But in the microcredit data, researchers estimated a set of 10 quantiles $\mathcal{U} = \{0.05, 0.1, 0.15, ..., 0.95\}$ and interpolated the results to form a "quantile difference curve". This curve is constructed by computing the quantile regression at all

points of interest:

$$Q_{y_{ik}|T} = \{Q_{y_{ik}|T}(u) = \beta_{0k}(u) + \beta_{1k}(u)T_{ik} \ \ \forall \ u \in \mathcal{U}\} \tag{3.15}$$

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:

$$\begin{aligned} \beta_{0k} &= (\beta_{0k}(0.05), \beta_{0k}(0.15), ...\beta_{0k}(0.95)) \\ \beta_{1k} &= (\beta_{1k}(0.05), \beta_{1k}(0.15), ...\beta_{1k}(0.95)) \end{aligned} \tag{3.16}$$

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:

$$\begin{aligned} Q_{0k} &= \{Q_{0k}(u) \ \ \forall \ u \in \mathcal{U}\} \\ Q_{1k} &= \{Q_{1k}(u) \ \ \forall \ u \in \mathcal{U}\}. \end{aligned} \tag{3.17}$$

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

$$\begin{aligned} \beta_{0k} &= Q_{0k} \\ \beta_{1k} &= Q_{1k} - Q_{0k}. \end{aligned} \tag{3.18}$$

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 reported information on quantile difference functions building on the approach of Rubin (1981) and a classical result from Mosteller (1946) about the joint distribution of sets of empirical quantiles. Mosteller shows that if the underlying random variable is continuously distributed, then the asymptotic sampling distribution of a vector of its empirical quantiles is a multivariate Normal centered at the true quantiles and with a known variance-covariance structure. This im-

plies that the difference of the empirical quantile vectors from two independent samples, $\beta_{1k} = (Q_{1k} - Q_{0k})$, is also asymptotically a multivariate Gaussian. The theorem offers a foundation for a hierarchical quantile treatment effect aggregation model using the knowledge that the sampling variation is approximately a multivariate Gaussian, and that as a result modelling the parent distribution as Gaussian will be both tractable and have attractive performance (Rubin 1981, Efron and Morris 1975). The resulting analysis requires only the limited information reported by each study (although it can be fit to the full data) and is applicable to any continuous distribution as long as there is sufficient data in each of the studies to make the asymptotic approximation reasonable.

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

$$\hat{\beta}_{1k} \sim N(\beta_{1k}, \hat{\Xi}_{\beta_1 k}) \; \forall \; k \tag{3.19}$$

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:

$$\beta_{1k} \sim N(\beta_1, \Sigma_1) \; \forall \; k. \tag{3.20}$$

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, consider monotonizing the $|\mathcal{U}|$-dimensional vector β_0, with uth entry denoted $\beta_0[u]$. This is necessary in aggregation because even though unconditional quantiles typically do not exhibit crossing, the partial pooling exercise has the potential to introduce crossing if the constraint is not enforced. Thus, I map β_0 to a new

vector β_0^* as follows:

$$\beta_0^*[u] = \begin{cases} \beta_0[u], & \text{if } u = 1 \\ log(\beta_0[u] - \beta_0[u-1]) & \text{if } 1 < u < |\mathcal{U}| \end{cases} \tag{3.21}$$

Any vector β_0 to which this transform is applied and for which inference is performed in the transformed space will always be monotonically increasing. For the rest of the paper, I denote parameters for which monotonicity has been enforced by performing inference on the transformed object as 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. Frequentist solutions such as rearrangement, smoothing or projection each prevent the violation of the constraint in one specific way chosen a priori according to the analyst's own preferences (He 1997, Chernozhukov et al. 2010). While each strategy performs well in terms of bringing the estimates closer to the estimand (as shown in Chernozhukov et al. 2010) the Bayesian transformation strategy can flexibly borrow from each of the strategies as and when the data supports their use. Imposing the constraint throughout the inference avoids the additional complications of choosing *when* during aggregation one should implement the ex-post fixes proposed in the frequentist literature; for example, in the case of rearrangement, it would be hard to interpret the result of partially pooling information on the 25th quantile only to have some other quantile substituted in for certain studies ex-post.

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

$$\begin{aligned} \hat{Q}_{0k} &\sim N(\beta_{0k}^m, \hat{\Xi}_{0k}) \ \ \forall\ k \\ \hat{Q}_{1k} &\sim N(Q_{1k}^m, \hat{\Xi}_{1k}) \ \ \forall\ k \\ &\text{where } \ Q_{1k} \equiv \beta_{0k}^m + \beta_{1k} \end{aligned} \tag{3.22}$$

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

$$\begin{aligned} \beta_{0k}^m &\sim N(\beta_0^m, \Sigma_0) \ \ \forall\, k \\ \beta_{1k} &\sim N(\beta_1, \Sigma_1) \ \ \forall\, k \\ \text{where} \ \ \beta_1 &\equiv Q_1^m - \beta_0^m \end{aligned} \tag{3.23}$$

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

$$\begin{aligned} \beta_0^m &\sim N(0, 1000 * I_{10}) \\ \beta_1 &\sim N(0, 1000 * I_{10}) \\ \Sigma_0 &\equiv diag(\nu_0)\Omega_0 diag(\nu_0)' \\ \Sigma_1 &\equiv diag(\nu_1)\Omega_1 diag(\nu_1)' \\ \text{where} \ \ \nu_0, \nu_1 &\sim \text{halfCauchy}(0, 20) \text{ and } \Omega_0, \Omega_1 \sim LKJCorr(1). \end{aligned} \tag{3.24}$$

This formulation is convenient as the form of $\hat{\Xi}_{1k}$ is exactly derived in the Mosteller (1946) theorem, though the individual entries need to be estimated. The structure could be modified to take in the empirical quantile treatment effects $\{\hat{\beta}_{1k}\}_{k=1}^K$ and their standard errors instead of $\{\hat{Q}_{1k}\}$ if needed. The model imposes no structure on (Σ, Σ_0), other than the logical requirement of positive semi-definiteness. This complete flexibility is made possible by the discretization of the quantile functions; these matrices could not take unconstrained form if the quantile functions had been modelled as draws from Gaussian Processes.[14] 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.

3.2.0.1 Pooling Metrics for Nonparametric Quantile Treatment Effects

Conventional pooling metrics for hierarchical models are designed to be applied to univariate treatment effects. For the multivariate Normal quantile curve aggregation models, the object that governs the dispersion of β_{1k} around β_1 is the parent variance-covariance matrix Σ_1. The raw size of this matrix is the purest metric of that dispersion, but this can only be measured in terms of a certain matrix norm, and different norms will give different answers. I proceed using a statistical argument to determine the appropriate norm.[15] 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 $\Sigma_1^{1/2}$.

[14] 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.

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

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.

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:

$$\begin{aligned}\boldsymbol{\omega}(\beta) &= \frac{1}{K}\sum_{k=1}^{K}\frac{\text{tr}(\hat{\Xi}_{\beta_1 k})}{\text{tr}(\hat{\Xi}_{\beta_1 k} + \Sigma)} \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{\text{tr}(\hat{\Xi}_{\beta_1 k})}{\text{tr}(\hat{\Xi}_{\beta_1 k}) + \text{tr}(\Sigma)}\end{aligned} \tag{3.25}$$

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:

$$\boldsymbol{\lambda}_{\boldsymbol{\beta_1}} = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{1}{U}\sum_{u=1}^{U}\frac{var(E[\beta_{1k}[u] - \beta_1[u]])}{E[(var(\beta_{1k}[u] - \beta_1[u])]}\right). \tag{3.26}$$

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:

$$\breve{\omega}(\boldsymbol{\beta_1}) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{U}\sum_{u=1}^{U}\frac{\beta[u]_{1k} - \hat{\beta}_{1k}[u]}{\beta_1[u] - \beta_{1k}[u]}\right). \tag{3.27}$$

3.3 Full Information Finite Sample Parametric Quantile Models

The strength of the model based on the Mosteller (1946) theorem is that it works for any continuous outcome variable; its weakness is that it *only* works for continuous variables. In the microcredit data, this approach will work for household consumption, consumer durables spending and temptation goods spending. But household business profit, revenues and expenditures are not continuous because many households either did not own or did not operate their businesses in the month prior to being surveyed and therefore recorded zero for these outcomes. This creates large "spikes" at zero in the distributions, as shown in the histograms of the profit data for the sites (figure 1). 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 2: the first panel is the result of using the Mosteller asymptotic approximation, and the second panel is the result of the nonparametric bootstrap applied to the standard errors on the same data. The former produces the dubious result that the uncertainty on the quantiles in the discrete spike is the same as the uncertainty in the tail; the latter produces the dubious result that the standard errors are exactly zero at most quantiles.

The potential for quantile regression techniques to fail when the underlying data is not continuous is a well-understood problem (Koenker and Hallock 2001; Koenker 2011). In some cases, "dithering" or "jittering" the data by adding a small amount of random noise is sufficient to prevent this failure and reliably recover the underlying parameters (Machado and Santos Silva, 2005).[16] But in the microcredit data, the complications caused by these spikes at zero are not effectively addressed by dithering. The results in figure 3 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. 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.

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 pursue a flexible and richly-parametised approach at the data level using mixtures of distributions in which treatment can affect all

[16] 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.

aspects of the shapes of the component distributions as well as the weights on each of the components themselves.[17] While this requires substantial input from the researcher and the aggregation model must be tailored to each specific case, this method will automatically solve the two problems discussed with naive quantile aggregation. Directly modelling the PDFs as proper densities, which therefore integrate to proper and thus weakly monotonic Cumulative Density Functions, will automatically deliver monotonically increasing vectors of quantiles. The model transfers information across neighbouring quantiles because they are directly linked by the functional form assumptions.

For the household business variables in the microcredit data, there is sufficient contextual economic information to build a parametric model. In this setting, economic theory predicts that these variables should be mixtures of spikes at zero and continuous tails because they are the output of a partially discrete decision process. First, a household has an extensive margin decision to make about whether to operate a business this season or not. This decision may be different at different times of the year depending on the outside options, as many households in these contexts engage in seasonal agricultural labour or intermittent construction labour for part of the year, only operating their businesses during the "lean season". Only those households who decide to open and operate their businesses go on to make an intensive margin decision, the result of which manifests some continuous expenditures, revenues and profit. This explains the spike at zero observed in all three business variables, which is a real feature of the generating process and not an artefact of the data collection. Economic theory and prior research suggest that the continuous portions of business variables such as revenues and profit tend to follow power laws or other fat-tailed laws (Stiglitz 1969, Gabaix 2008, Allen 2014, Bazzi 2016). Hence, the outcome PDF can be modeled as a mixture of three distributions: a lower tail, a spike at zero, and an upper tail. As T_{nk} may affect the mass in the components and the shape of the tail components, I specify treatment effects on all aspects of this mixture PDF. The model can then aggregate effect of the treatment on each of the parameters that govern the distribution, as well as the implied quantile treatment effects.

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

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

to select between them, and I determine what if any inferences are robust to the choice of tail distribution.

Consider the following tailored hierarchical PDF model to aggregate the quantile effects on household business profit. Denote the probability mass in the jth mixture component for a household n with treatment status T_{nk} to be $\Lambda_j(T_{nk})$ for $j = 1, 2, 3$. This dependence can be modeled using a multinomial logit specification, denoting the intercept in site k for mixture component j as α_{jk} and the treatment effect as π_{jk}. For the spike at zero, the Dirac delta function can be used as a distribution, denoted $\delta(x)$ for a point mass at x. If using the Pareto distribution for the continuous component, the tails are 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 have already 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:

$$\begin{aligned} y_{nk}|T_{nk} \sim\ & \Lambda_{1k}(T_{nk})Pareto(-y_{nk}|\iota_{1k}, \exp(\rho_{1k} + \kappa_{1k}T_{nk}) \\ & +\Lambda_{2k}(T_n)\delta_{(0)} \\ & +\Lambda_{3k}(T_n)Pareto(y_{nk}|\iota_{3k}, \exp(\rho_{3k} + \kappa_{3k}T_{nk}) \ \forall\ k \\ \text{where}\ \ \Lambda_{jk}(T_{nk}) =\ & \frac{\exp(\alpha_{jk} + \pi_{jk}T_{nk})}{\sum_{j=1,2,3}\exp(\alpha_{jk} + \pi_{jk}T_{nk})} \end{aligned} \tag{3.28}$$

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

$$(\alpha_{1k}, \alpha_{2k}, \pi_{1k}...)' \equiv \zeta_k \sim N(\zeta, \Upsilon)\ \forall\ k \tag{3.29}$$

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

$$\begin{aligned} \zeta &\sim N(0, 10) \\ \Upsilon &\equiv diag(\nu_\Upsilon)\Omega_\Upsilon diag(\nu_\Upsilon)' \\ \nu_\Upsilon &\sim \text{halfCauchy}(0, 5) \\ \Omega_\Upsilon &= I_{|\zeta|} \\ \alpha_{mk} &\sim N(0, 5) \end{aligned} \tag{3.30}$$

I build a similar using LogNormal tails, which are each governed by a location parameter and a scale parameter. The latter can only be positive valued so I again employ the exponential transform to ensure the support constraint is satisfied. I model the location parameter using a linear regression format in which the value for the control group in site k is μ_k and the value for the treatment group is $\mu_k + \tau_k$. The scale parameter is modelled similarly to the pareto scale parameter with the control group's value being $\exp(\sigma_k^c)$ and the treatment group's value being $\exp(\sigma_k^c + \sigma_k^t)$. This produces the following model:

$$\begin{aligned} y_{nk}|T_{nk} \sim \Lambda_{1k}(T_{nk}) & LogNormal(-y_{nk}|\mu_{1k} + \tau_{1k}T_{nk}, \exp(\sigma_{1k}^c + \sigma_{1k}^t T_{nk})) \\ & + \Lambda_{2k}(T_n)\delta_{(0)} \\ & + \Lambda_{3k}(T_n) LogNormal(y_{nk}|\mu_{3k} + \tau_{3k}T_{nk}, \exp(\sigma_{3k}^c + \sigma_{3k}^t T_{nk}) \; \forall \; k \\ \text{where } \; \Lambda_{jk}(T_{nk}) &= \frac{\exp(\alpha_{jk} + \pi_{jk}T_{nk})}{\sum_{j=1,2,3}\exp(\alpha_{jk} + \pi_{jk}T_{nk}))} \end{aligned} \tag{3.31}$$

The upper level of the model is also specified Gaussian as in equation 3.29 with independence enforced for tractability, and the priors are specified in the same way. The tailored hierarchical PDF aggregation models for revenues and expenditures are constructed as above for both the Pareto and LogNormal cases, but with no negative tail and hence only 2 mixture components.

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

$$\begin{aligned} Q(u) = -\text{Pareto}^{-1}\left(1 - \frac{u}{\Lambda_1(T_n)} \mid \iota_{1k}, \rho_{1k}(\exp(\kappa_{1k}T_n))\right) * \mathbb{1}\{u < \Lambda_1(T_n)\} \\ + 0 * \mathbb{1}\{\Lambda_1(T_n) < u < (\Lambda_1(T_n) + \Lambda_2(T_n)\} \\ + \text{Pareto}^{-1}\left(\frac{u - (1 - \Lambda_3(T_n))}{\Lambda_3(T_n)} \mid \iota_{3k}, \rho_{3k}(\exp(\kappa_{3k}T_n)\right) * \mathbb{1}\{u > (1 - \Lambda_3(T_n))\} \end{aligned} \tag{3.32}$$

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

3.3.0.1 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 modelling the outcomes explicitly means we do not have recourse to a sampling variance-covariance matrix within the model itself. In order to construct a sampling variance-covariance matrix, I fit a no-pooling version of the tailored PDF model, omitting the upper level of the hierarchy. I use the set of no pooling model parameters $\{\hat{\zeta}_k\}_{k=1}^K$ and their accompanying posterior variance-covariance matrix $\hat{\Sigma}_\zeta$ to construct the pooling metrics of interest. Hence, the translation of the conventional pooling metric in this case is

$$\begin{aligned} \boldsymbol{\omega}_V(\beta) &= \frac{1}{K}\sum_{k=1}^{K}\frac{\text{tr}(\hat{\Sigma}_{\zeta k})}{\text{tr}(\hat{\Sigma}_{\zeta k}+V)} \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{\text{tr}(\hat{\Sigma}_{\zeta k})}{\text{tr}(\hat{\Sigma}_{\zeta k})+\text{tr}(V)}. \end{aligned} \tag{3.33}$$

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 Main Results

4.1.1 Consumption Variables

Aggregating the quantile treatment effects for household consumption, consumer durables spending and temptation goods spending using the Mosteller-based model shows that microcredit has a precise, generalizable zero effect below the 75th percentile. Beyond this point there is an imprecise positive effect that exhibits high variance across sites. Figure 4 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

diverge in the upper tail. This difference itself signals the presence of heterogeneous effects across settings: full pooling and partial pooling differ only in the weights they apply to the effects in each site when combining them to deliver an aggregate point estimate, and the result can only differ if the components being weighted differ. In this setting, the full-pooling model typically underestimates the uncertainty, particularly on the upper quantiles, and thus delivers inference with higher precision than is warranted by the evidence.

While the inference on the average quantile effects across all sites are similar for the hierarchical model and the full pooling model, the inference on the predicted effect in the next site is more uncertain in the hierarchical context. Posterior predictive distributions of these consumption variables are shown in figure 5 with the full-pooling model for comparison. The results show considerably more uncertainty about the outcomes, particularly at the right tail, than would be suggested by taking either the full-pooling model or the posterior distribution of β_1 from the partial pooling model. 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. 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 2. There is almost zero pooling of the control group quantiles according to two of the three metrics, and intermediate pooling according to the third metric. This indicates that the control groups are substantially different in across studies, and suggests that the zero impact along most of the distribution is indeed generalizable across heterogeneous contexts.

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

4.1.2 Business Variables

To analyse household business expenditures, revenues and profits, I fit both the Pareto and LogNormal models and use posterior predictive checking to select the structure that fits the data best (Gelman et al 2004). This requires simulating data from the posterior distribution of each model, and then comparing the simulated data to the real data. As table 3 shows, the LogNormal model outperforms the Pareto in terms of predicting the

actual observed control group quantiles, particularly in the right tail. The Pareto shape has too little mass near zero and too much mass in the tail relative to the LogNormal. However, the broad patterns observed in the results of the LogNormal model are also observed in the Pareto model, and are thus robust to choice of tail distribution (see Appendix A). In particular, both models show a precise zero impact at most quantiles, and then the potential for large increases in the right tails.

The quantile treatment effect results from the the Lognormal hierarchical PDF models for all business outcomes are shown in figure 6 with the full pooling results shown for comparison. The models find a precise and generalizable zero effect below the 75th percentile, although the lower tail of profit is an imprecise zero. Above the 75th percentile there is a large positive point estimate, but much less precision and more uncertainty, due to heterogeneity both within and across sites. By contrast, the full pooling models find much larger and more precise "statistically significant" effects in the tails. The difference is dramatic because when the tails are sparse, a little more pooling goes a long way; yet as with consumption, the presence of different aggregate point estimates in the tails is itself a signal of heterogeneity across settings, such that the full pooling assumption is unwarranted in this setting and unlikely to produce reliable inference. In a frequentist sense, the apparently "statistically significant" results in the upper tails "detected" in the full pooling model are eliminated by the application of a hierarchical model. In a Bayesian sense, the full pooling model is misleadingly precise in the upper tail, and the posterior uncertainty we should have about these tail effects is much larger. However, there is more than a 90% probability of a positive effect on the 85th and 95th quantiles of all the distributions, suggesting that microcredit may indeed be affecting these tails in some positive way.

The posterior predicted quantile results for future effects, again computed using the Castellacci (2012) formula, are shown in figure 7 with the full pooling results for comparison. Any detected heterogeneity in the quantile treatment effects on household business outcomes is typically localized above the 85th percentile. Below this point, the effect is zero and reasonably generalizable, but above this point the high variation and sparsity in the tails means that there is great uncertainty about the exact impact microcredit will have on the right tail of the next distribution to which it is applied. As before, the full pooling model displays unwarranted precision and magnitude of impact relative to the more moderate and uncertain prediction made by the hierarchical model.

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 6 and the tail shape effects are shown in table 7. 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.

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

4.2 The role of business experience

While the results of the hierarchical aggregation display less heterogeneity across the experiments than the disaggregated results suggested, understanding the remaining heterogeneity is important. There are a number of covariates both within and across sites which could predict these differences in the distributional effects of microcredit in theory. At the household level, the most relevant pre-treatment covariate is the previous business experience of the households in the sample, as measured by their operation of a business prior to the microcredit intervention. As different study populations had differing prevalence of 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 will be necessary to avoid severe overfitting: this exercise is described in Appendix E. 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 across the two types of households. Figures 8 and 9 show the general distributional impact of microcredit on the six household outcomes of interest for each of the household types. For most outcomes, households with no prior business ownership see negligible impact of microcredit across the entire distribution, leading to a generalizable and precise impact of zero across all quantiles, with only a small increase in the variance in the right tail. Households with prior businesses are responsible for the positive and large point estimates in the right tails, but also for the noise in that tail, suggesting that they are also the source of the heterogeneous effects. This confirms the results of Banerjee et. al. (2015) and Meager (2018), which performed similar analyses within a single site and for the average effects respectively, and found differences in the way households with business experience respond to microcredit relative to those without such experience.

A closer examination of the results yields indirect evidence about the different ways in which these two types of households respond to increased access to microcredit. For households with business experience, there is strong evidence of a positive effect on total consumption at the 95th percentile, whereas households without experience see little impact on total consumption at any quantile (figure 8). 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 8). Taken together, this suggests that some households who don't have prior businesses may generally use microcredit to change the composition of their consumption bundles; but even this smaller effect occurs only in the tail and is imprecisely estimated (figure 9).

4.3 The role of take-up

One concern about the models presented in the main analysis is that they ignore the role of differential take-up in explaining the impact of microcredit. While the results of the analysis stand for themselves as group-level causal impacts, the economic interpretation of the results might differ if we knew, for example, that the zero impact along most of the outcome quantiles was entirely due to lack of take-up by most of the households in the sample. The main results contain suggestive evidence that the lack of impact at most quantiles is not solely due to lack of take-up: the 2 sites that randomized loan access rather than branch access and therefore had almost full take-up (Bosnia and the Philippines) displayed the same trend as all the other sites (Appendix B). However, 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 for individual

households within a village.

I pursue a bounding exercise that provides additional evidence that take-up patterns alone cannot explain the precise zero results along most of the distribution. 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. Even though in the microcredit studies, there are households in the control groups who do manage to access microcredit, presumably these are the "always takers" and not the compliers against which the appropriate comparison can be made.

However, we can 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. In a simple model in which selection into treatment is positively correlated with treatment effects, this probably overestimates the effect since this group contains both compliers and never-takers, the latter of which are usually assumed to have zero treatment effects (Imbens and Rubin 2015). This forms a likely lower upper on the effect for compliers. To find a lower bound, one can compare the outcomes of the treated households who took up the loans to the control households who took up the loans. Consider a simple model in which selection into treatment is positively correlated with treatment effect and households borrow if the benefits outweigh the costs, and suppose that expansion of access to microcredit reduces the costs of taking a loan (if only because one has less far to walk to the MFI branch). In such a world, comparing households who took up in treatment to those who took up in control would most likely underestimate the effect on the compliers. Therefore, computing these two comparisons gives a rough ballpark on either side of the correct but infeasible comparison.[18]

I find that for almost every outcome variable, the "treatment effect" on the selected sample is similar 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 10. 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 11. 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

[18]The potential for SUTVA violations is what prevents me from pursuing the computation of the LATE for those who take up treatment. I should note that potential SUTVA violation would also affect the validity of the levels of the bounds provided here, but the gap between the two bounds should not be affected by violations of SUTVA unless there are differential violations by comparison type, which seems unlikely.

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, yet this would require more structure than the current analysis.

5 Discussion

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

The precise zero effect from the 5th to 75th percentile of most of the household outcomes is a true zero and not a mechanical artefact of the spike at zero nor an economic consequence of the low takeup. Consumption, consumer durables and temptation goods do not exhibit a spike of households who record an outcome equal to zero (this is almost true by definition, since it is hard to survive on nothing), yet microcredit still has a precisely estimated zero effect for most of the distribution. Even for profit, revenues and consumption the spike only accounts for at most 50% of the outcome distribution, yet the zero effect applies to 75% of the distribution. Similarly, Bosnia and the Philippines had over 90% takeup and yet still exhibited zero effects from the 5th to 75th percentile (see Appendix B). The bounding exercise pursued in section 4.3 aggregates all the data on the question of takeup and shows that even the effects on the outcome distribution for those who take up microloans are likely to be zero along most of the distribution.

I do find evidence of large positive effects of microcredit on the right tail of all outcome distributions, although these effects are imprecisely estimated and heterogeneous across contexts. Thus, the quantile analysis effectively decomposes the small and moderately noisy average treatment effect estimates from all the papers, aggregated in Meager (2018), into an imprecise yet large effect on the tail, and a precise zero everywhere else. These tail effects are large enough to be economically important and are typically concentrated among those households who have previous experience operating businesses, for whom we can rule out a zero effect on consumption at the 95th percentile, though the estimate is still

quite imprecise, as shown in figure 8. Thus overall the aggregated distributional analysis provides evidence that microcredit is likely to do some good and no systemic harm. While the models are unable to precisely predict the effect on the right tail, and thus cannot confidently predict the impact in the next location into which microcredit expands, it is more likely to be positive than negative. Of course, for most of the community, it appears that no systematic change is occurring and the majority of the outcome distribution looks the same in both treatment and control.

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

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

The heavy tails (extreme kurtosis) in the household business outcomes has both methodological and economic implications. Ordinary least squares regressions such as those per-

formed in the original randomized controlled trials are likely to perform poorly compared to quantile regression techniques or parametric modelling of the heavy tail of business variables such as profit (Koenker and Basset 1978, Koenker and Hallock 2011). More substantively, heavy tails suggest that in these populations, certain individual households account for large percentages of the total business activity. This suggests that it may be challenging to understand the economies of developing countries if we trim or winsorize the most productive households who make up a large percentage of total economic activity. It might be more useful to study mechanisms that can produce fat-tailed outcomes, such as multiplicative production functions, experimentation or investments with a relatively high risk exposure and long maturation horizons. The fact that households with prior businesses increase their consumption (figure 8) suggests they have some expectation of future increases in profits or earnings. This highlights the potential benefits of studying these households over longer time horizons, or perhaps taking multiple observations of the same households as in the Townsend Thai Data (2018) and as suggested in McKenzie (2012).

My analysis is not exhaustive, and the conclusions I can draw are limited by the constraints of my framework and of the original studies. It may be that if microcredit interventions were studied over a 10 or 20 year horizon, the imprecise tail effects we observe after two years could either become precise or could lead to benefits across the entire distribution. If the studies had a richer set of baseline data, a deeper understanding of the household-level distributional impacts of expanding access to microcredit could be generated by including baseline covariates and perhaps leveraging more economics knowledge of the contextual microstructure to the analysis. It would be informative to apply an individual-level structural model to this data, such that one could infer the distribution of individual-level treatment effects, but there is currently no established methodology for partial pooling on structural parameters. Finally, by restricting the selected set of studies to be RCTs, there is a possibility of a sample selection bias due to the conditions required to perform field experiments; as yet, there is no established method for combining experimental and observational studies in a single aggregation framework. Despite these cautions, the conclusion of the current analysis remains salient: in general, there is likely to be no difference between the treatment and control groups below the 75th quantile in future sites that receive more access to microcredit, and while we cannot reliably predict the effect above the 75th percentile, the aggregated evidence suggests it is likely to be positive.

6 Conclusion

Understanding the distributional impact of microcredit requires confronting the econometric challenge of aggregating sets of quantile effects without imposing unwarranted assumptions on the degree of external validity across studies. I develop new Bayesian hi-

erarchical models and associated pooling metrics to estimate the distributional treatment effects of policy interventions and assess the generalizability of the results. I approach the problem of passing information across quantiles and ensuring quantile monotonicity using variable transformation, and I use richly-parameterised mixture models to aggregate information on quantiles of partially discrete variables. I apply these methods to aggregate the impact of expanding access to microcredit on the distribution of various household economic outcomes, and find that the analysis can reveal aspects of the data occluded by average treatment effects analyses. For the microcredit data, 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 methods for evidence aggregation when the true generalizability of the treatment effects is not known.

Aggregating the evidence from seven RCTs of expanding access to microcredit generates new insights about the general impact of this intervention on different parts of the distribution of household economic outcomes. I find a precise and generalizable zero impact below the 75th percentile on all outcomes, and a large positive impact above this point which is less precisely estimated due to greater heterogeneity both within and across studies in the right tail. Although I used different methods to aggregate consumption outcomes and business outcomes due to the different underlying structures of the variables, this pattern is robust to choice of methodology and to choice of parametric functional form. Thus, although microcredit has the potential to create winners and losers, there is no evidence that any group of households is systematically harmed. There is a high probability of positive effects on the right tail of the distribution, although these effects are less generalizable across studies and may not manifest in all settings. The precise zero along most of the distribution however suggests that most households' lives are not transformed by microcredit access, and this result holds even among those who take up the loans or have previous business experience, for whom the right tail effects are most pronounced. Taken together the pattern suggests a potential improvement for some, although such improvements are likely accompanied by greater inequality across households; the welfare effects of microcredit are likely to be complex.

These results highlight the importance of analysing distributional effects rather than simply examining average treatment effects. Previous work aggregating the average impact of microcredit found generalizable information, but small or even null treatment effects (Meager 2018, Vivalt 2016). Aggregating the quantile treatment effects reveals that this average effect is composed of a precise zero effect for most of the distribution, and a large positive effect in the right tail which is imprecisely measured and heterogeneous across studies. Moreover, the results highlight the inadequacy of average effects estimation in the presence of heavy tails, which are detected in all the business variables to the extent that the median is substantially more efficient than the mean as a measure of location. Quantile analysis and parametric models which allow for the possibility of heavy tails will typically

outperform average effects analysis here, particularly when Gaussian approximations are used for inference on those averages (Koenker and Basset 1978). Heavy-tailed data is reasonably common in economics, particularly for variables such as earnings, profits and wealth (e.g. Bazzi 2016, Pancost 2016, Gabaix 2008, Fama 1965); the widespread use of least squares regression may be problematic. The benefit of using robust methods such as quantile regression apply not only to the aggregation of distributional effects across settings but also to the analysis of economic data in general.

The models developed in this paper can be used to aggregate the evidence on a wide range of interventions likely to have heterogeneous treatment effects across households, such as trade policies, educational subsidies or incentives for social and geographic mobility (Chetty and Hendren 2018, Duflo, Dupas and Kremer 2017, Chetty, Hendren, and Katz 2016, Autor et al 2014, Bryan, Chowdhury, and Mobarak 2014, Katz, Kling and Leibman 2001). The methods developed to aggregate the microcredit data here can be directly applied to the quantile effects on outcomes such as household earnings, consumption and educational outcomes in these literatures, delivering results which should be both more informative and more reliable than the results of any single study. Although it may not always be possible to extrapolate evidence across contexts, a formal assessment of the heterogeneity in effects and the resulting uncertainty surrounding similar interventions in new settings has the potential to improve the way that research is translated into policy in many areas of economics.

Table 1: Lender and Study Attributes by Country

Country	Bosnia & Herzegovina	Ethiopia	India	Mexico	Mongolia	Morocco	The Philippines
Study Citation	Augsburg et al (2015)	Tarozzi et al (2015)	Banerjee et al (2015)	Angelucci et al (2015)	Attanasio et al (2015)	Crepon et al (2015)	Karlan and Zinman (2011)
Treatment	Lend to marginally rejected borrowers	Open branches	Open branches	Open branches, promote loans	Open branches, target likely borrowers	Open branches	Lend to marginal applicants
Randomization Level	Individual	Community	Community	Community	Community	Community	Individual
Urban or Rural?	Both	Rural	Urban	Both	Rural	Rural	Urban
Target Women?	No	No	Yes	Yes	Yes	No	No
MFI already operates locally?	Yes	No	No	No	No	No	Yes
Microloan Liability Type	Individual	Group	Group	Group	Both	Group	Individual
Collateralized?	Yes	Yes	No	No	Yes	No	No
Any other MFIs competing?	Yes	No	Yes	Yes	Yes	No	Yes
Household Panel?	Yes	No	No	Partial	Yes	Yes	No
Interest Rate (Intended on Average)	22% APR	12% APR	24% APR	100% APR	24% APR	13.5% APR	63% APR
Sampling Frame	Marginal Applicants	Random Sample	Households with at least 1 woman age 18-55 of stable residence	Women ages 18-60 who own businesses or wish to start them	Women who registered interest in loans and met eligibility criteria	Random Sample plus Likely Borrowers	Marginal Applicants
Study Duration	14 months	36 months	40 months	16 months	19 months	24 months	36 months

Note: The construction of the interest rates here is different to the construction of Banerjee et al (2015a); they have taken the maximal interest rate, whereas I have taken the average of the intended range specified by the MFI. In practice the differences in these constructions are numerically small. This table was also printed in Meager (2018) which used the same studies.

Table 2: Pooling Factors for Nonparametric Quantile Models on Consumption

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

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

Table 3: Posterior Predictive Comparison of LogNormal and Pareto Models

Control Group Quantiles	5%	15%	25%	35%	45%	55%	65%	75%	85%	95%
Revenues Data	0	0	0	0	0	0	4	41	154	622
Lognormal Prediction	0	0	0	0	0	12	37	77	154	408
Pareto Prediction	0	0	0	0	0	0	0	5	337	2,793,933
Expenditures Data	0	0	0	0	0	0	0	17	85	411
Lognormal Prediction	0	0	0	0	0	0	15	40	93	283
Pareto Prediction	0	0	0	0	0	0	0	1	94	1,172,324
Profit Data	-29	0	0	0	0	0	0	4	49	226
Lognormal Prediction	-2	0	0	0	0	0	4	21	56	173
Pareto Prediction	0	0	0	0	0	0	0	0	21	70,170

Notes: The posterior predictive distributions are generated by drawing samples of data from the likelihood averaged over the posterior probability of the unknown parameters. Because this data is itself fat tailed, I have compared the actual sample quantiles from the fully pooled control group against the posterior predicted median value of each quantile from each model. [Back to main]

Table 4: Pooling Factors for Categorical Logit Effects (Reference Category: Positive)

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

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

Table 5: Pooling Factors for Lognormal Parameters

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

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

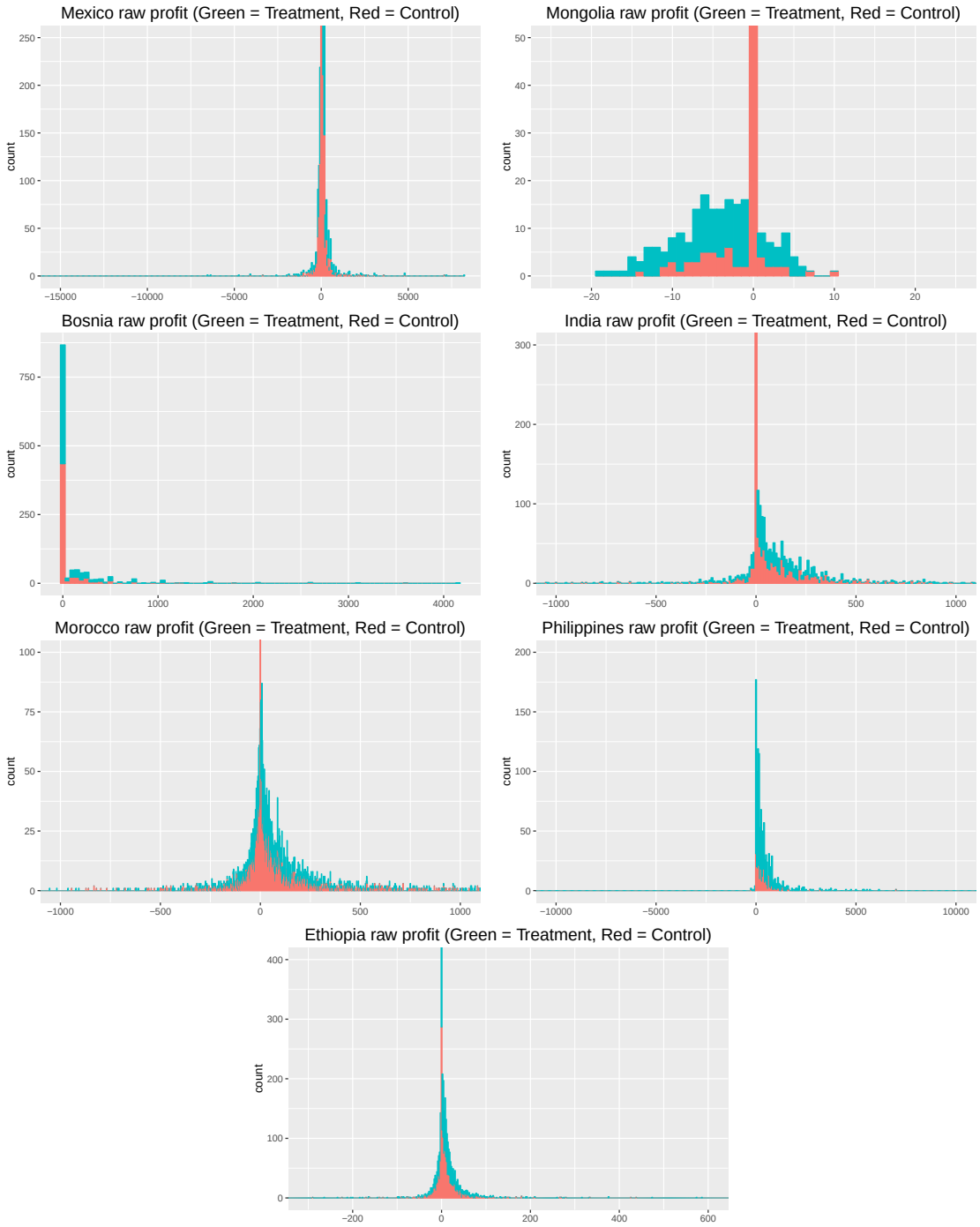

Figure 1: Histograms of the profit data in each site, in USD PPP per 2 weeks. Display truncated both vertically and horizontally in most cases. [Back to main]

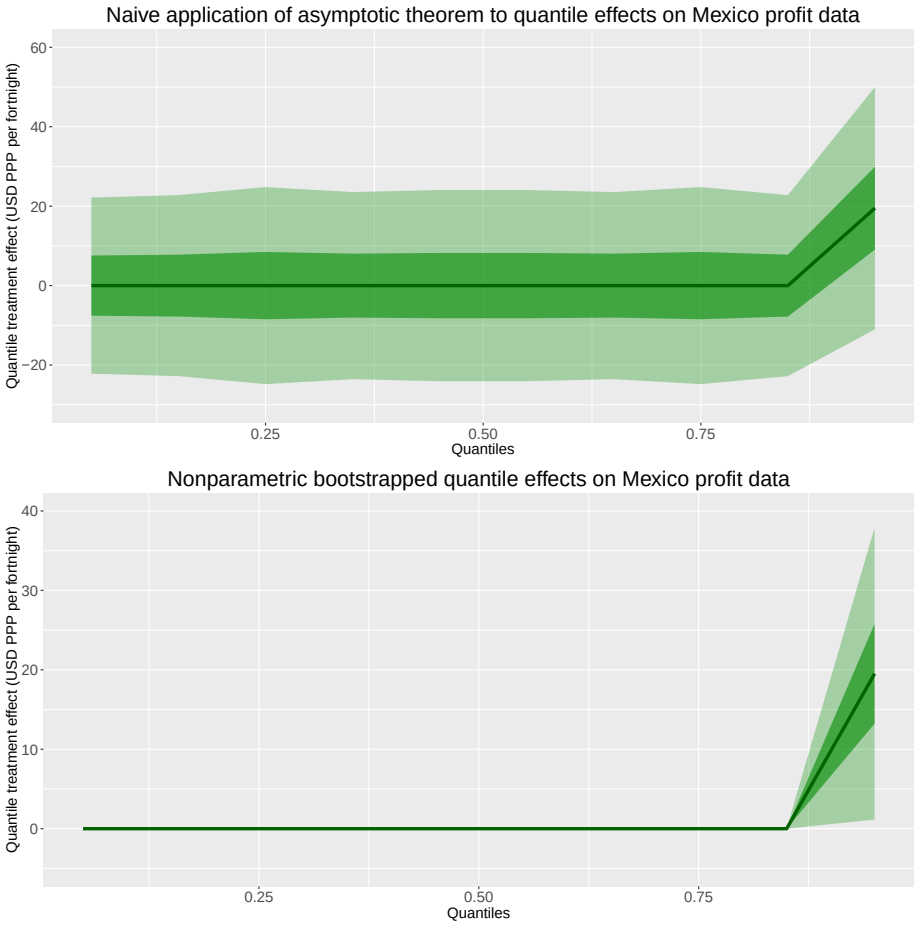

Figure 2: Quantile TEs for the Mexico profit data, Mosteller theorem approximation standard errors (above) and nonparametrically bootstrapped standard errors (below). The green line is the estimated effect, the opaque bands are the central 50% interval, the translucent bands are the central 95% interval.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

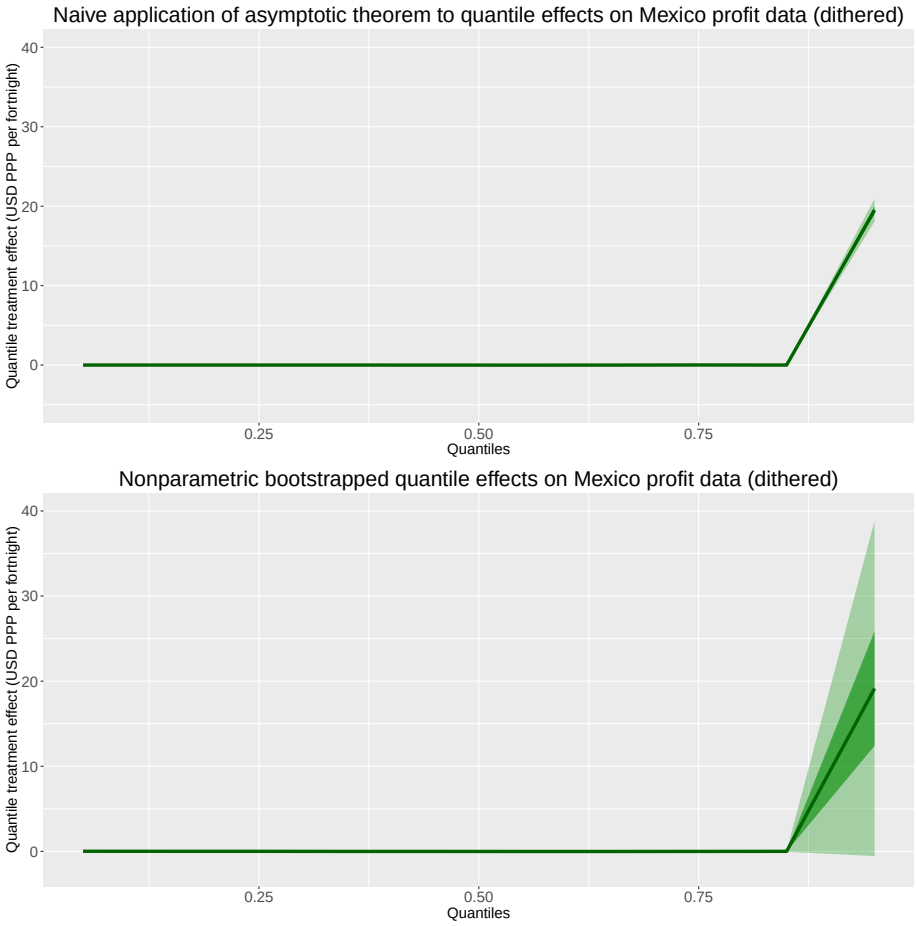

Figure 3: Quantile TEs for the dithered Mexico profit data, Mosteller theorem approximation standard errors (above) and nonparametrically bootstrapped standard errors (below). The green line is the estimated effect, the opaque bands are the central 50% interval, the translucent bands are the central 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. [Back to main]

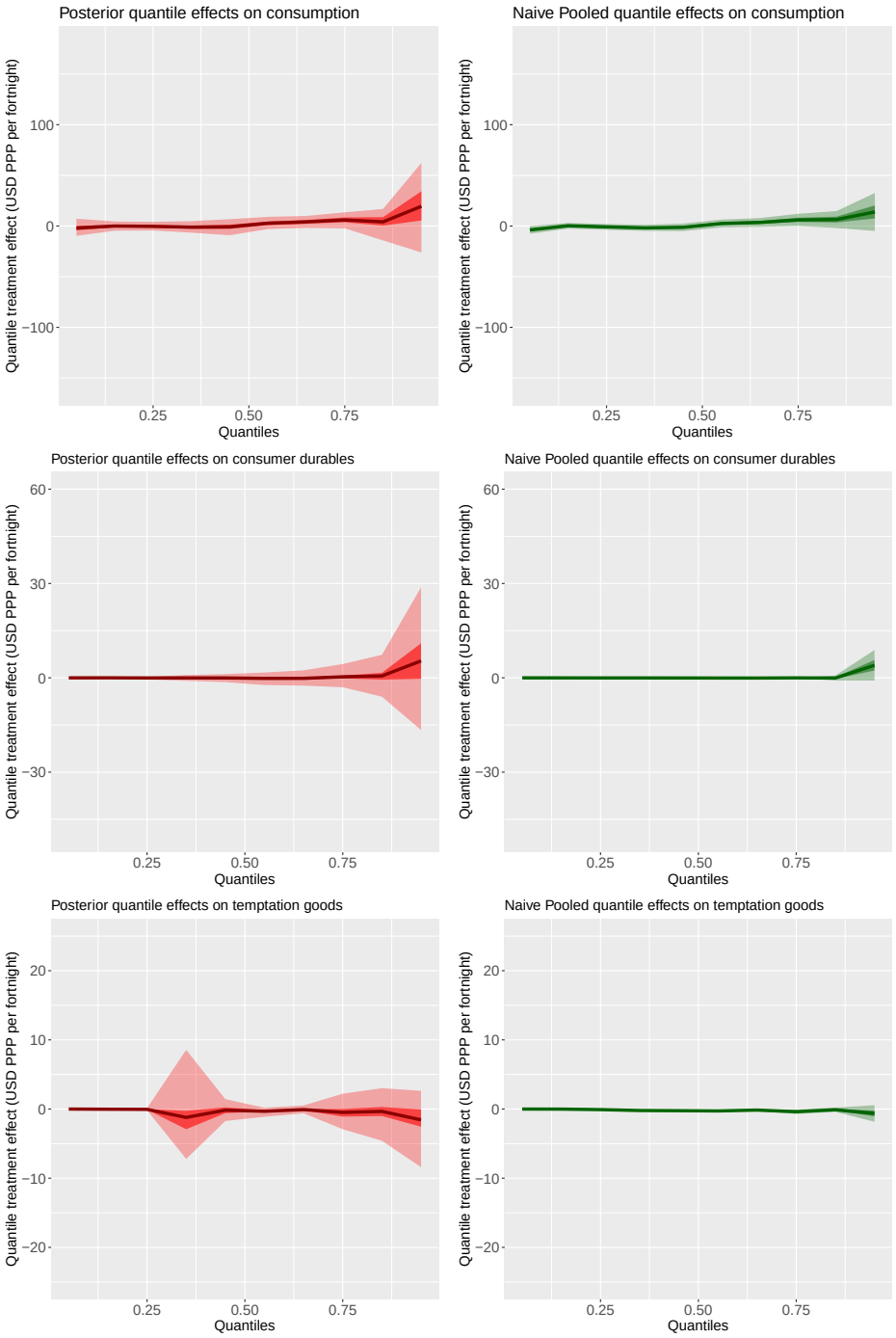

Figure 4: General Quantile Treatment Effect Curves (β_1) for consumption-type variables. The dark line is the posterior mean, the opaque color bands are the central 50% posterior uncertainty interval, the translucent color bands are the central 95% posterior uncertainty interval. [Back to main]

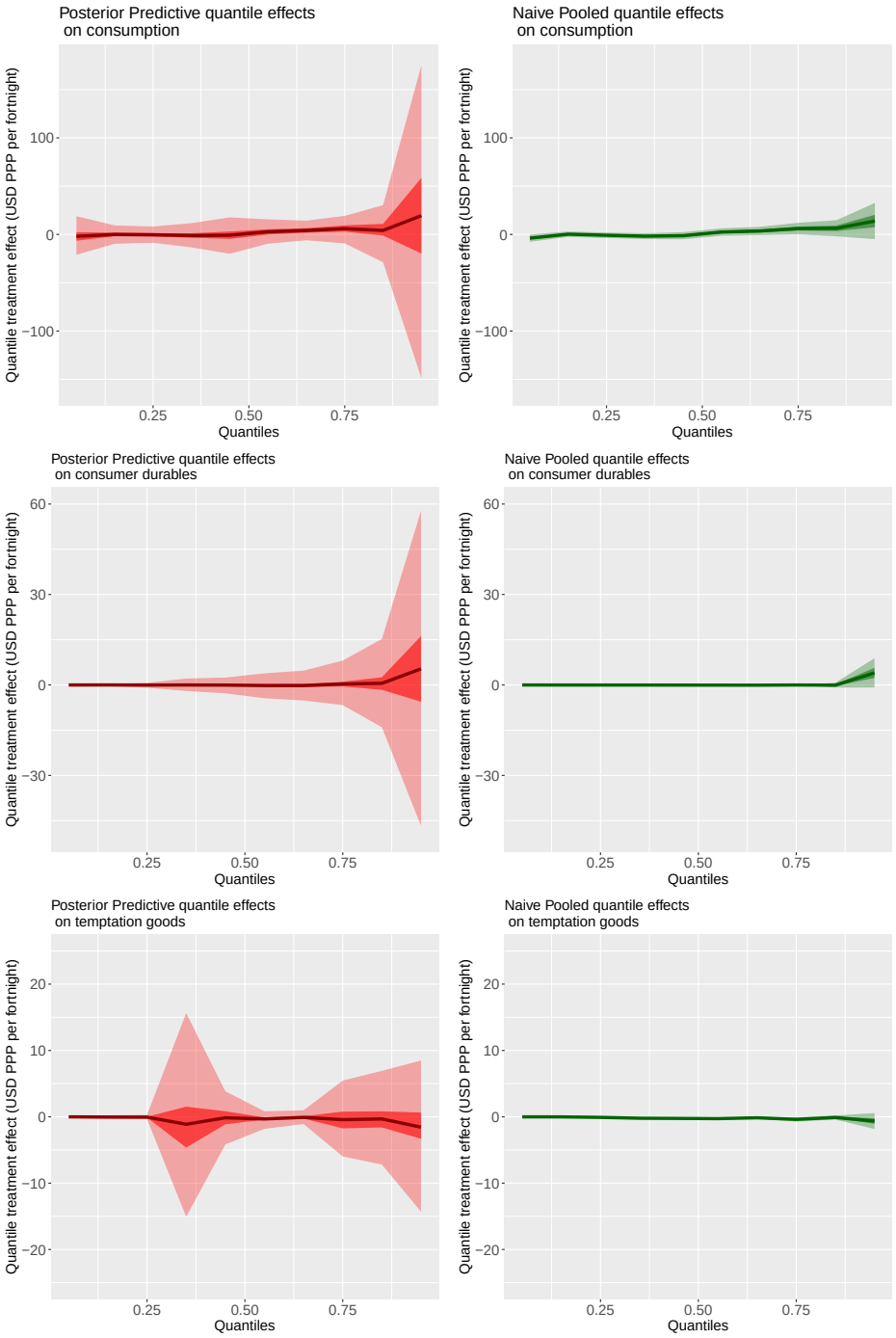

Figure 5: Posterior Predictive Quantile Effect Curves ($\beta_{1,K+1}$) for consumption-type variables.The dark line is the posterior mean, the opaque color bands are the central 50% posterior predictive uncertainty interval, the translucent color bands are the central 95% posterior predictive uncertainty interval. [Back to main]

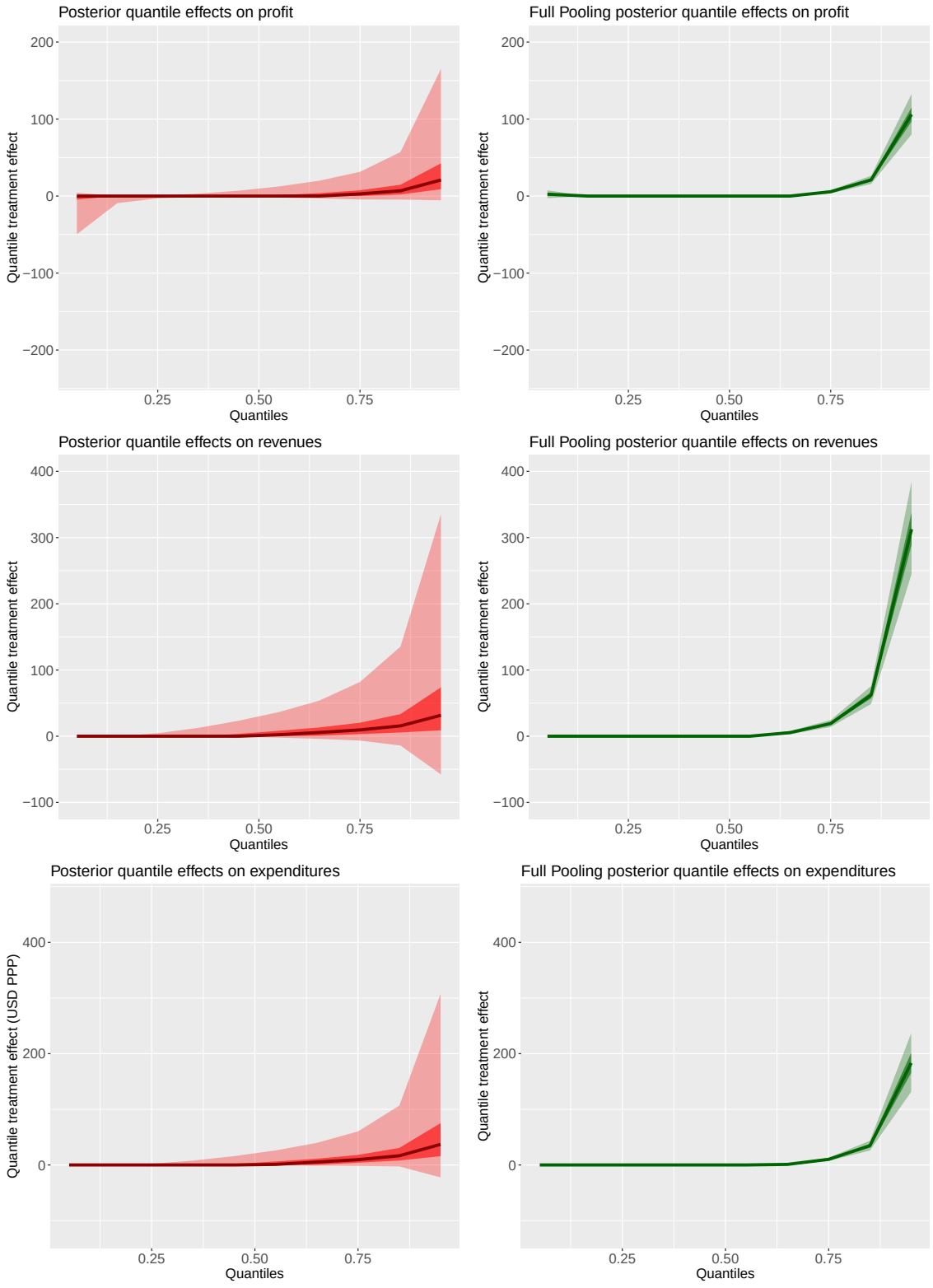

Figure 6: General Quantile Treatment Effect Curves (β_1) for business variables from the LogNormal model. The dark line is the median posterior draw, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. [Back to main]

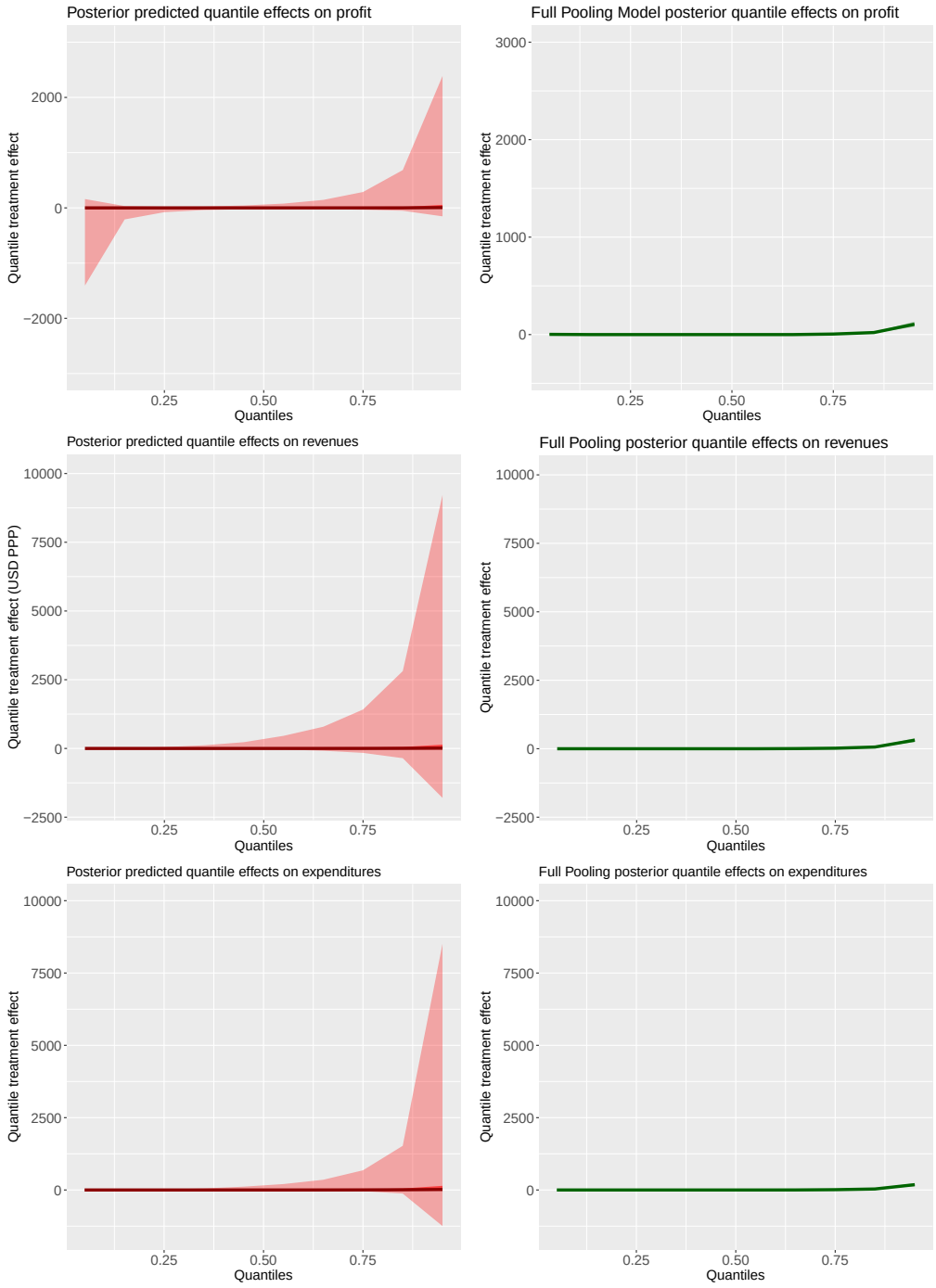

Figure 7: Posterior predicted quantile treatment effect curves for Business Variables from the LogNormal model. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. [Back to main]

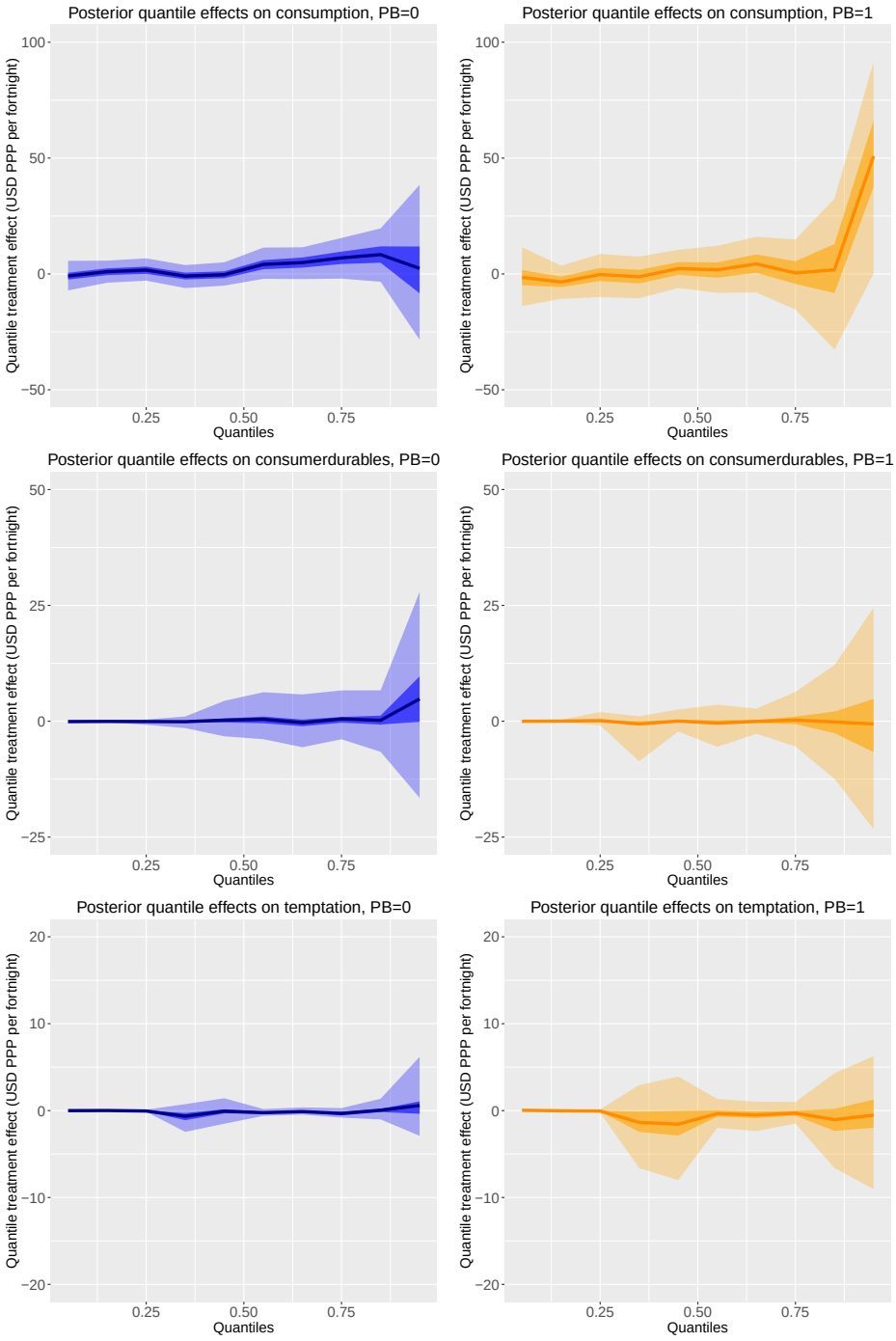

Figure 8: General Quantile Treatment Effect Curves split by prior business ownership (β_1) for consumption-type variables. The dark line is the posterior mean, the opaque color bands are the central 50% posterior uncertainty interval, the translucent color bands are the central 95% posterior uncertainty interval. [Back to main]

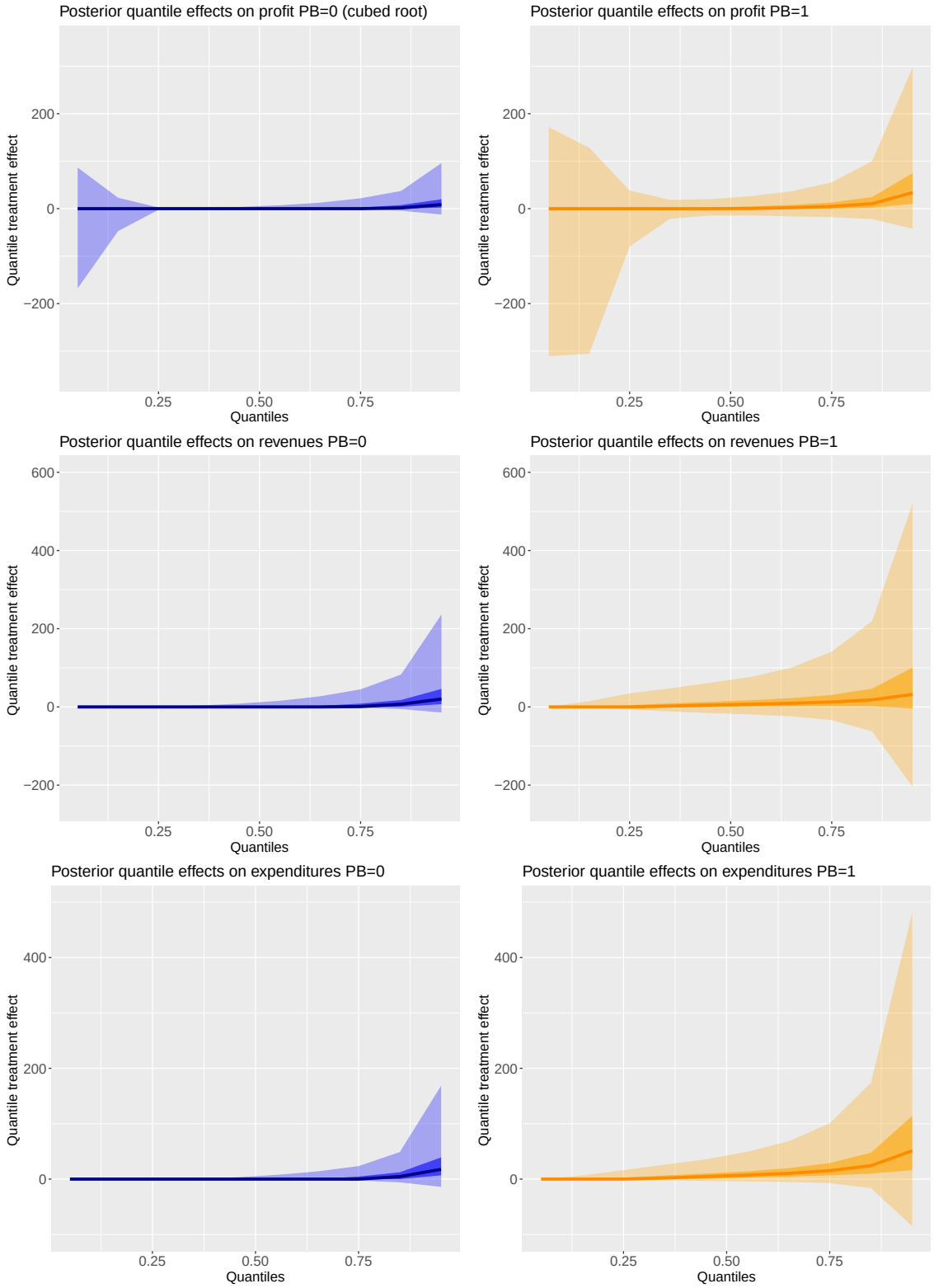

Figure 9: General Quantile Treatment Effect Curves (β_1) for business variables split by prior business ownership. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in cubed root of USD PPP due to the scale differences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

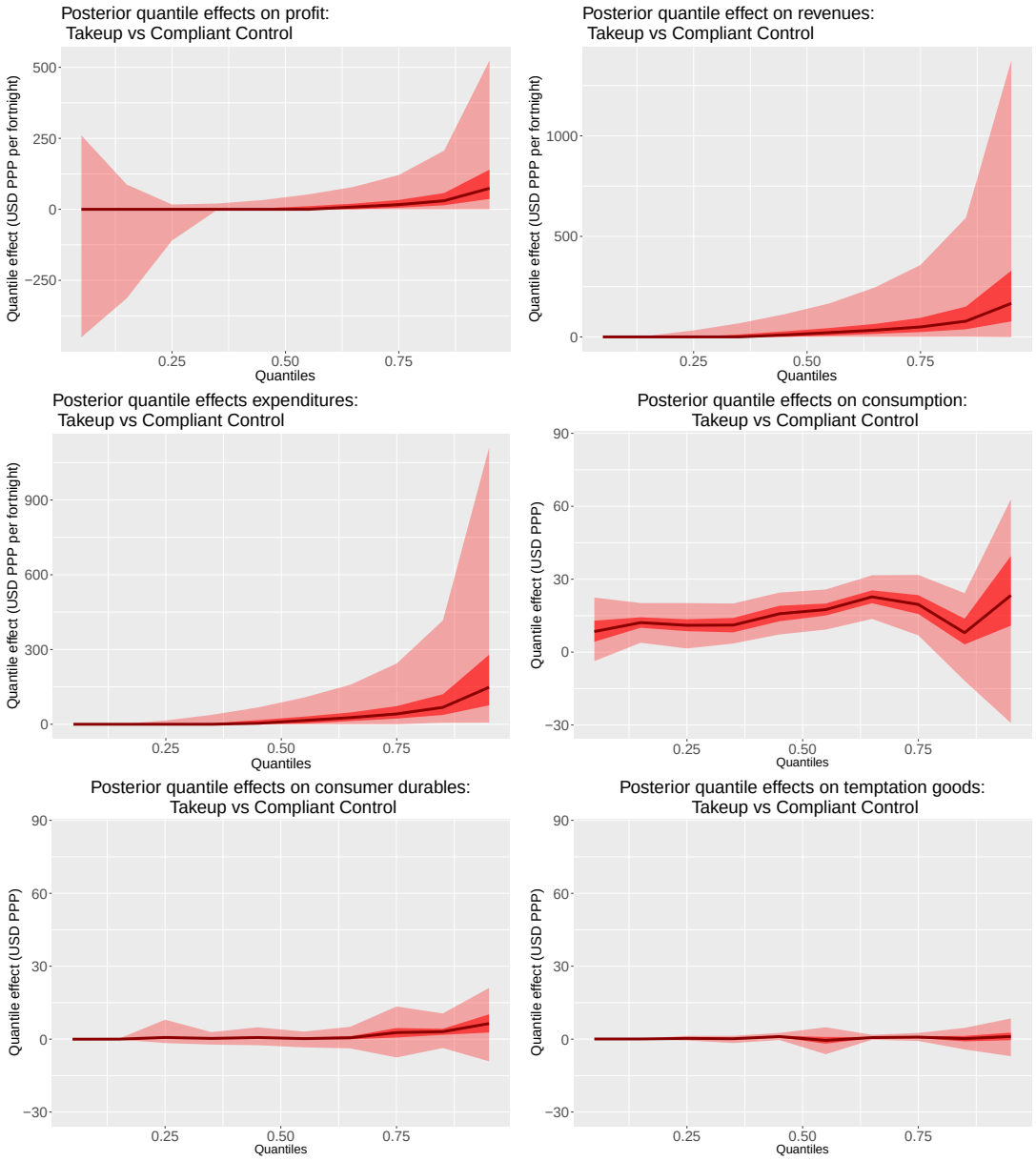

Figure 10: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Compliant control households who did not take up. This effect should overestimate the true impact of microcredit on those who take it up in a simple selection framework. Consumption variables 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]

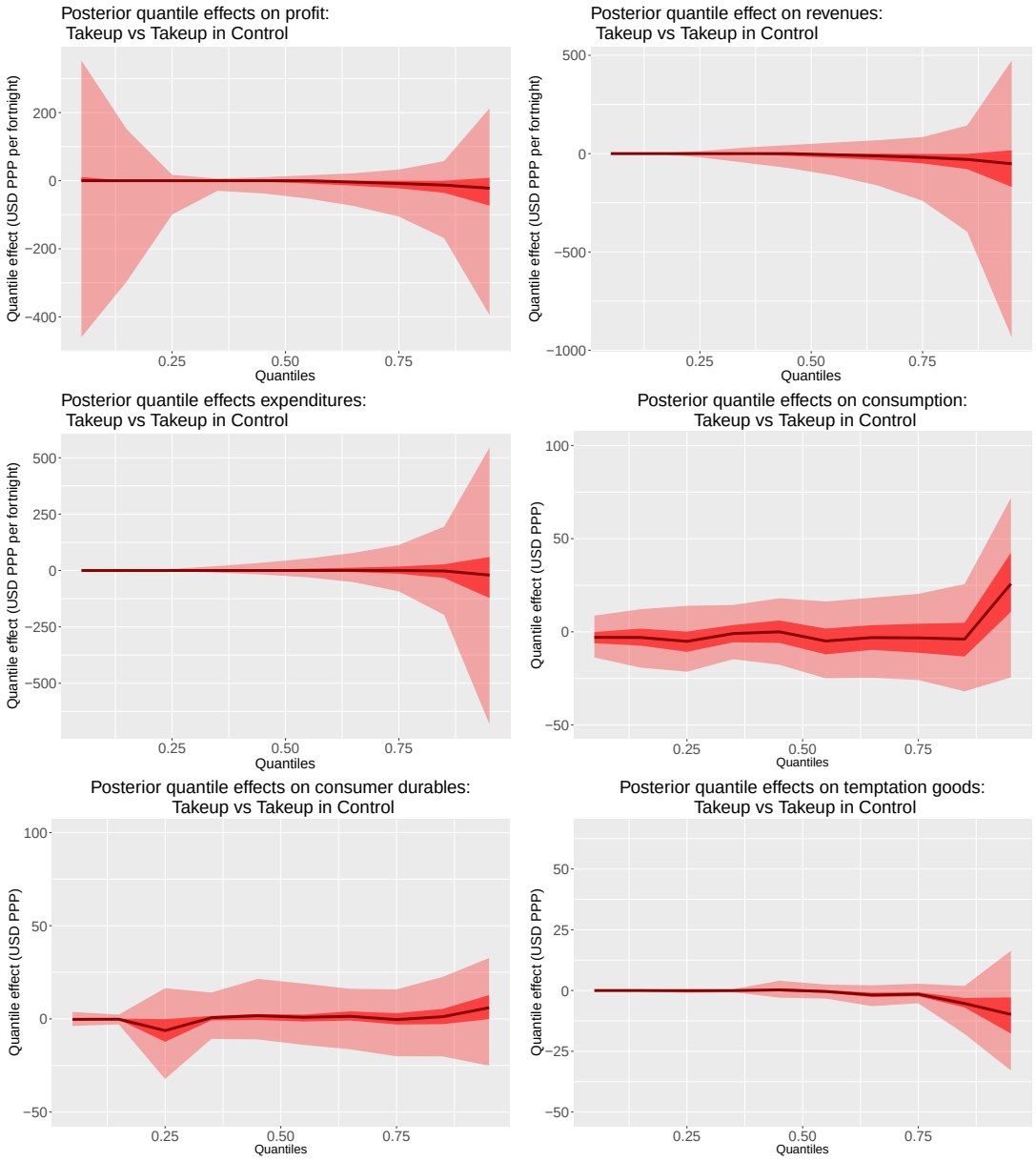

Figure 11: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Control households who took up. This effect should underestimate the true impact of microcredit on those who take it up in a simple selection framework. Consumption variables are in USD PPP per 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 Pareto Aggregation Model Results

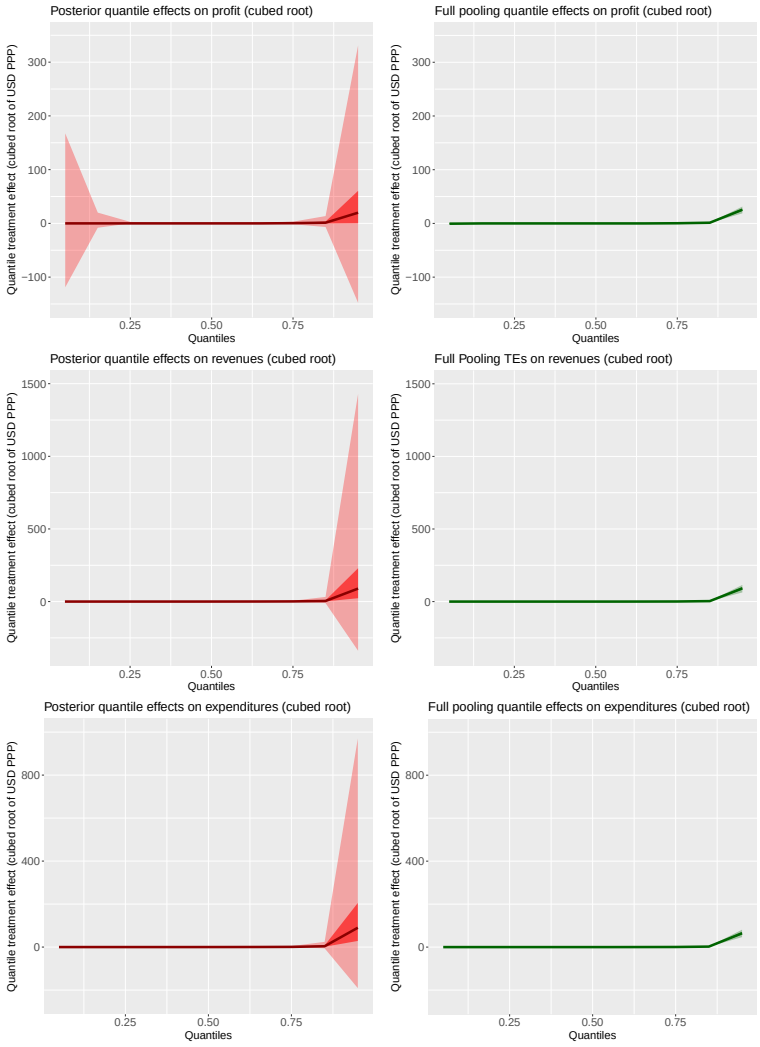

Figure 12: General Quantile Treatment Effect Curves (β_1) for business variables. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in cubed root of USD PPP due to the scale differences in the uncertainty at the right tail versus the rest of the distribution.

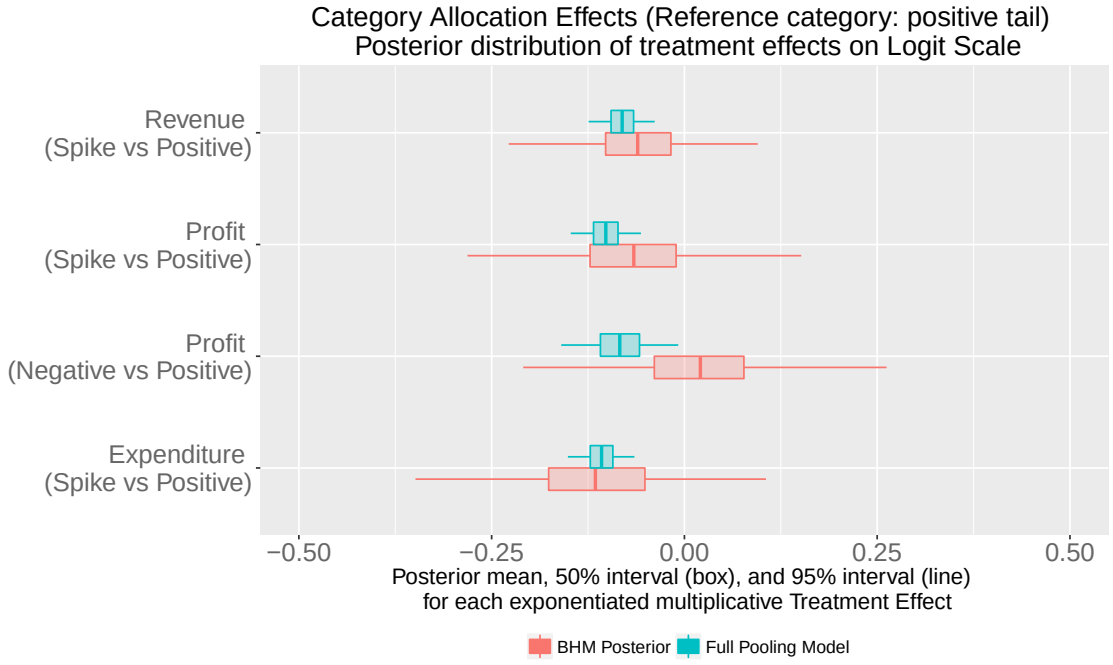

Figure 13: Posterior distributions for the logit treatment effects (π_j) on category assignment. These treatment effects are specified as an exponentiated multiplicative factor on the control group proportion of households in the category: if $\tilde{\pi_j} = 0$ the effect is zero, if $\tilde{\pi_j} < 0$ the treatment increases the proportion of households in the positive tail relative to other categories.

Table 6: 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.

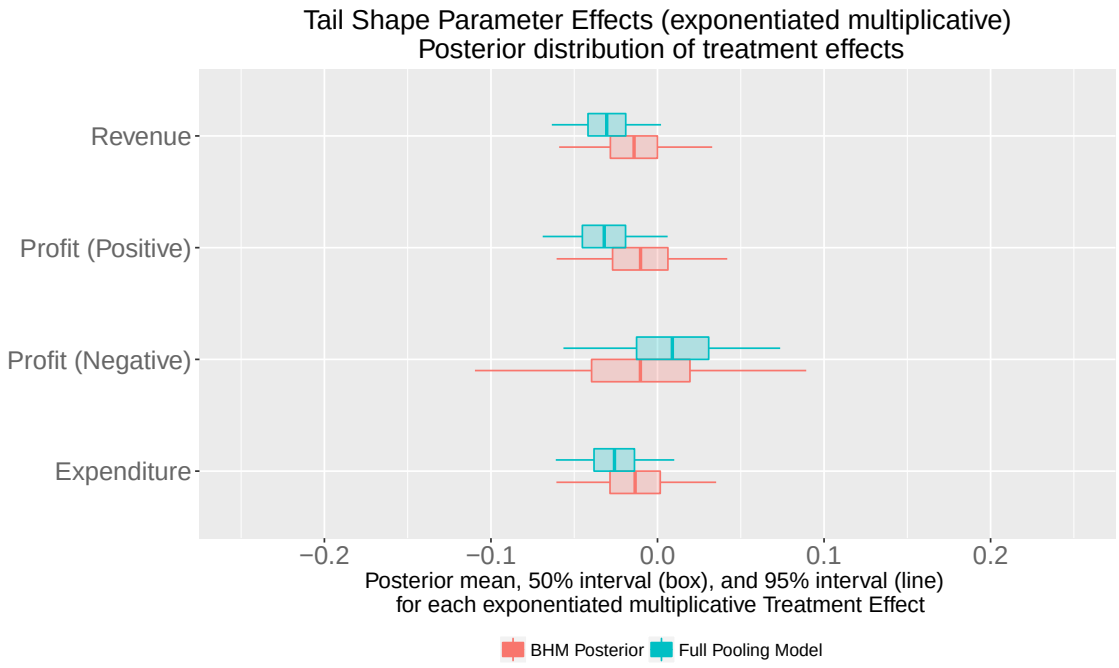

Figure 14: Posterior distributions for the Pareto shape treatment effects (κ_j) in each site. These treatment effects are specified as an exponentiated multiplicative factor on the control group scale parameter: if $\tilde{\kappa_j} = 0$ the effect is zero, if $\tilde{\kappa_j} = 0.7$ the effect is a 100% increase in the scale parameter.

Table 7: 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.

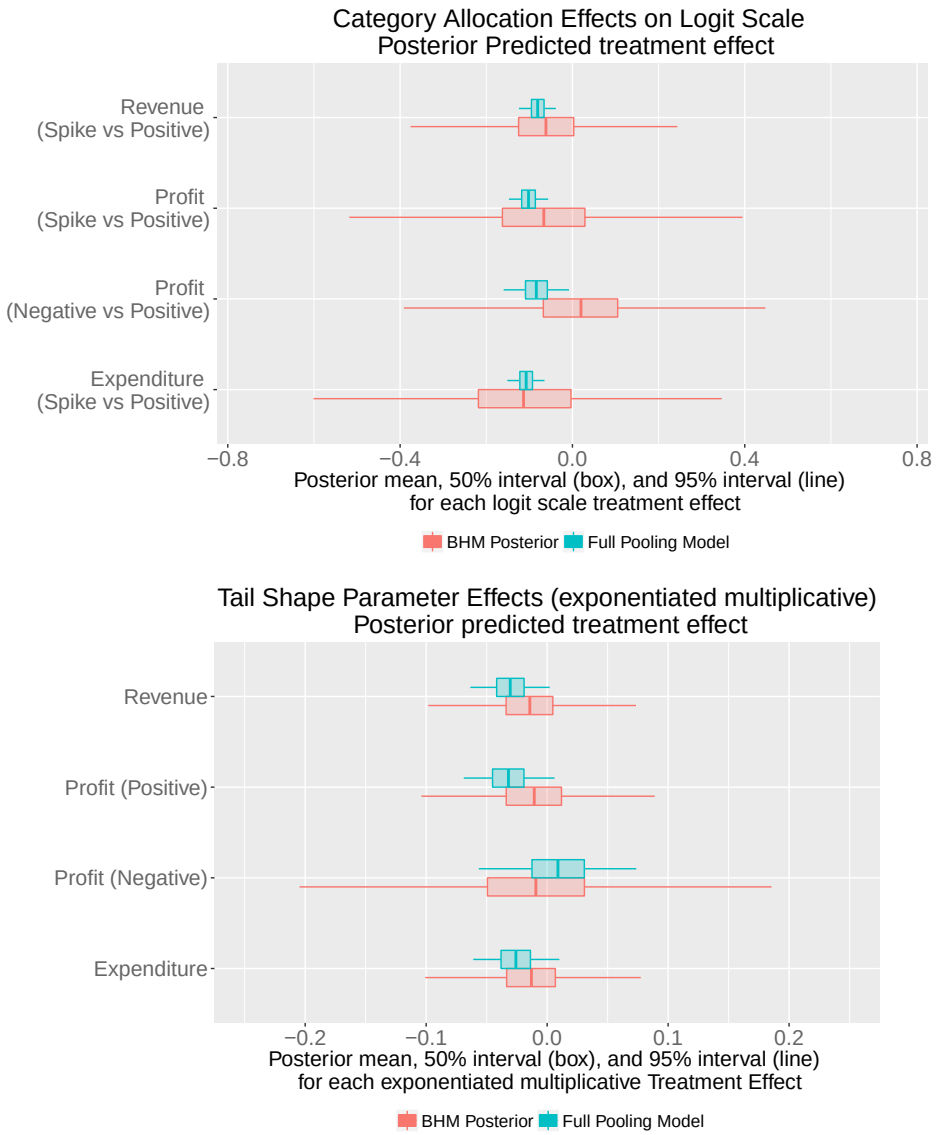

Figure 15: Posterior predicted distributions for the logit treatment effects on category assignment and tail shape effects. In each case this is the predicted treatment effect in a future exchangeable study site, with uncertainty intervals that account for the estimated generalizability (or lack of it).

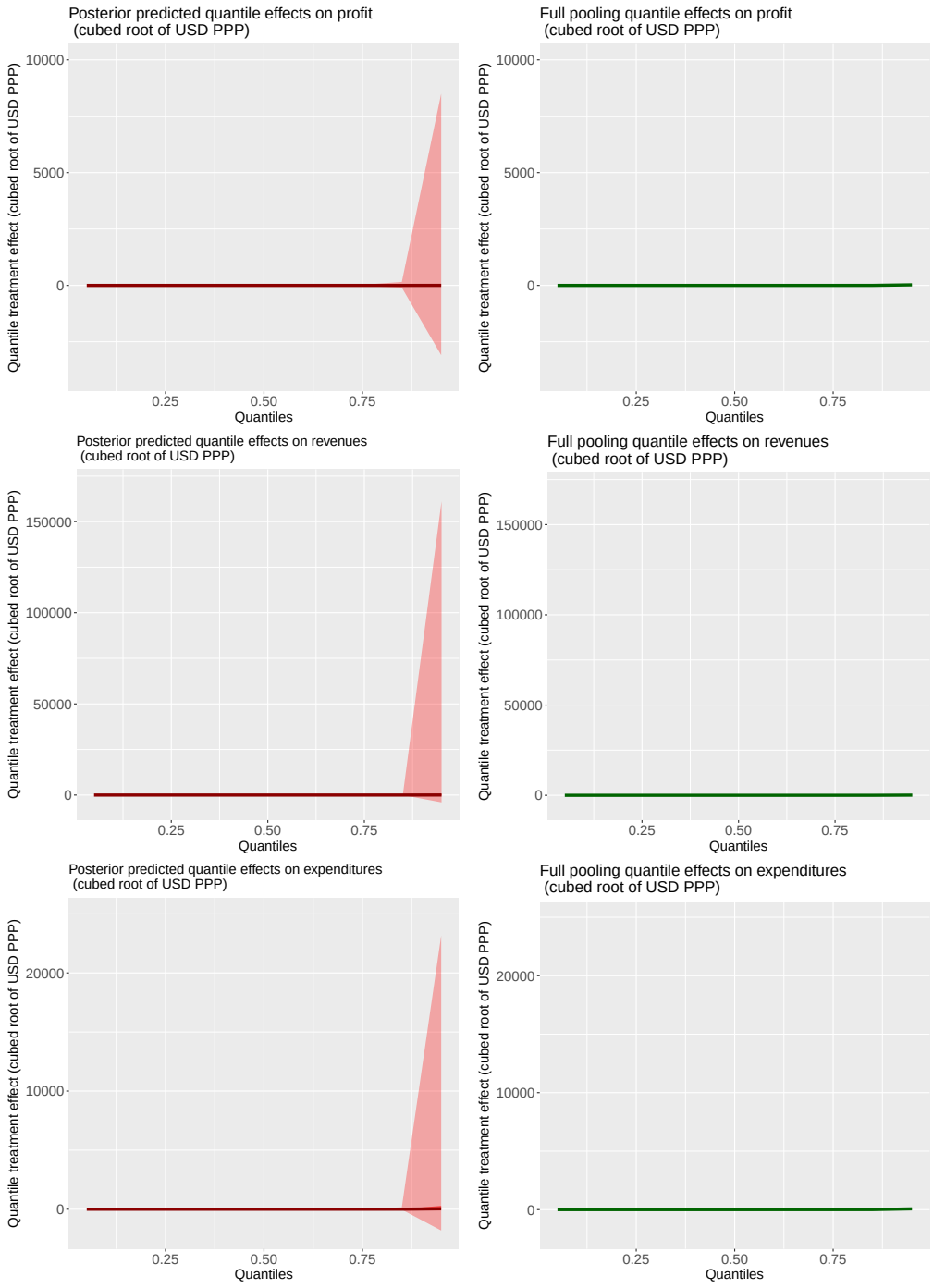

Figure 16: Posterior predicted quantile treatment effect Curves for Business Variables. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in cubed root of USD PPP due to the scale differences in the uncertainty at the right tail versus the rest of the distribution.

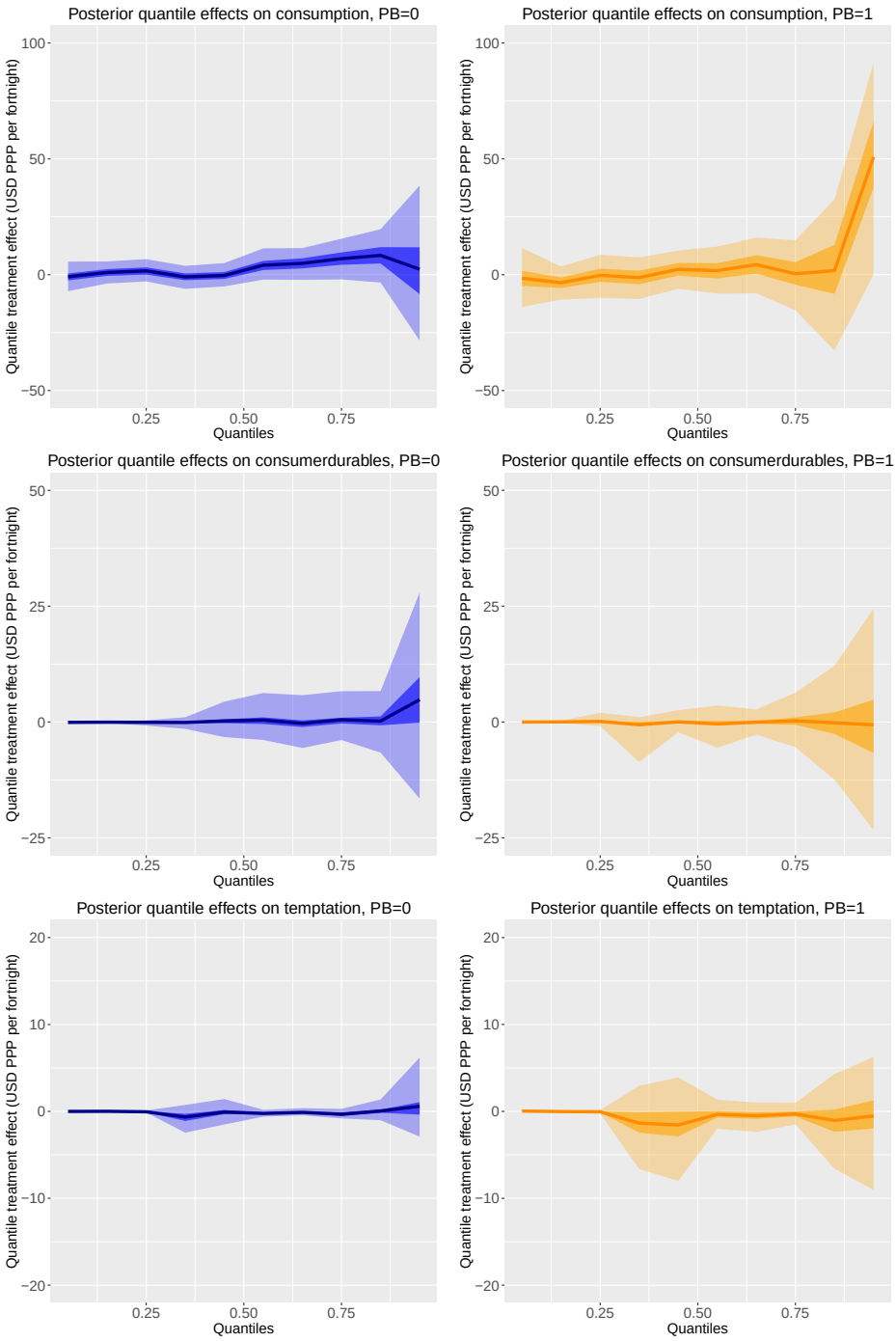

Figure 17: General Quantile Treatment Effect Curves split by prior business ownership (β_1) for consumption-type variables. The dark line is the posterior mean, the opaque color bands are the central 50% posterior uncertainty interval, the translucent color bands are the central 95% posterior uncertainty interval.

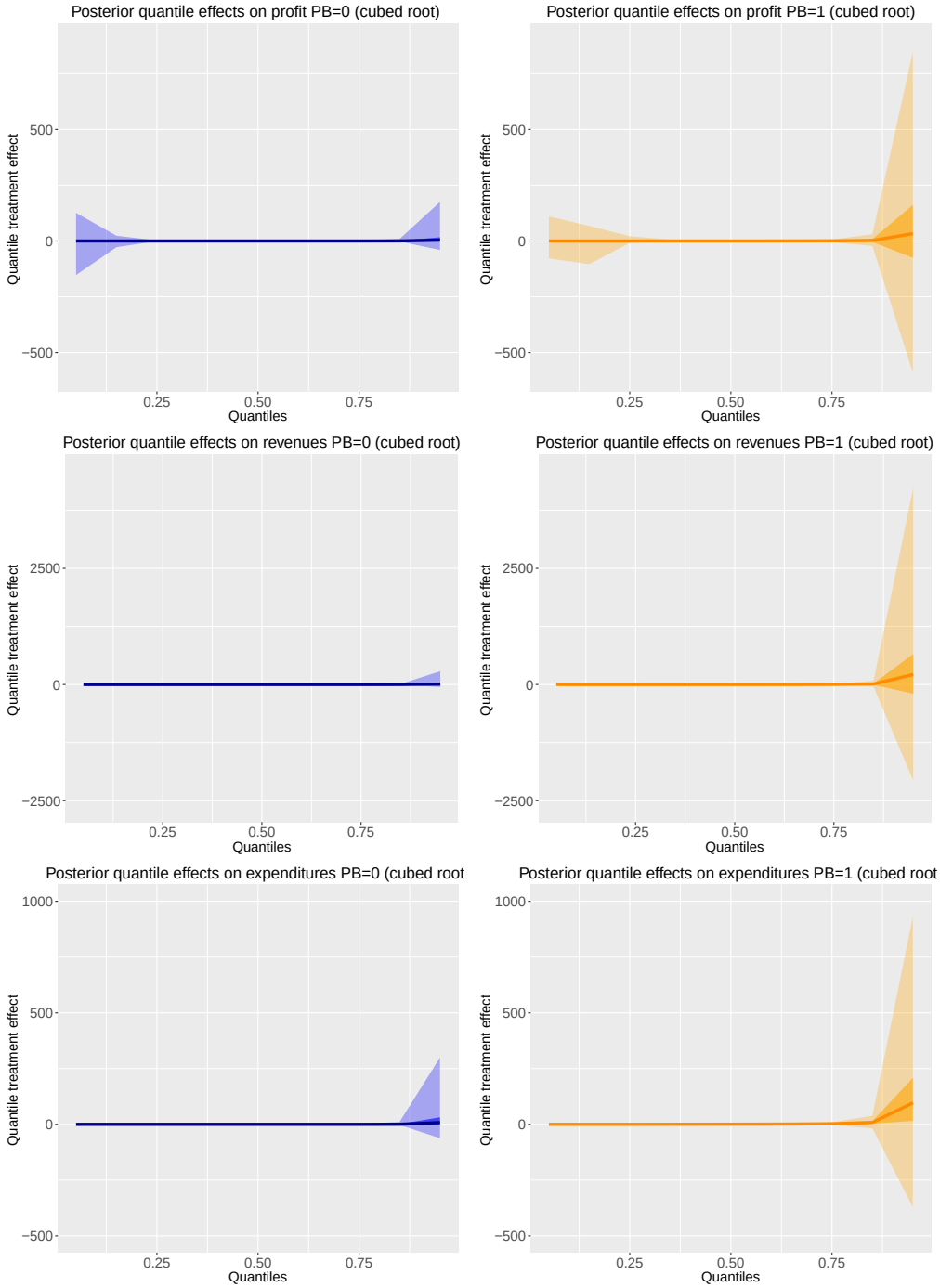

Figure 18: General Quantile Treatment Effect Curves (β_1) for business variables split by prior business ownership. The dark line is the median, the opaque bars are the central 50% interval, the translucent bands are the central 95% interval. Display is in cubed root of USD PPP due to the scale differences in the uncertainty at the right tail versus the rest of the distribution.

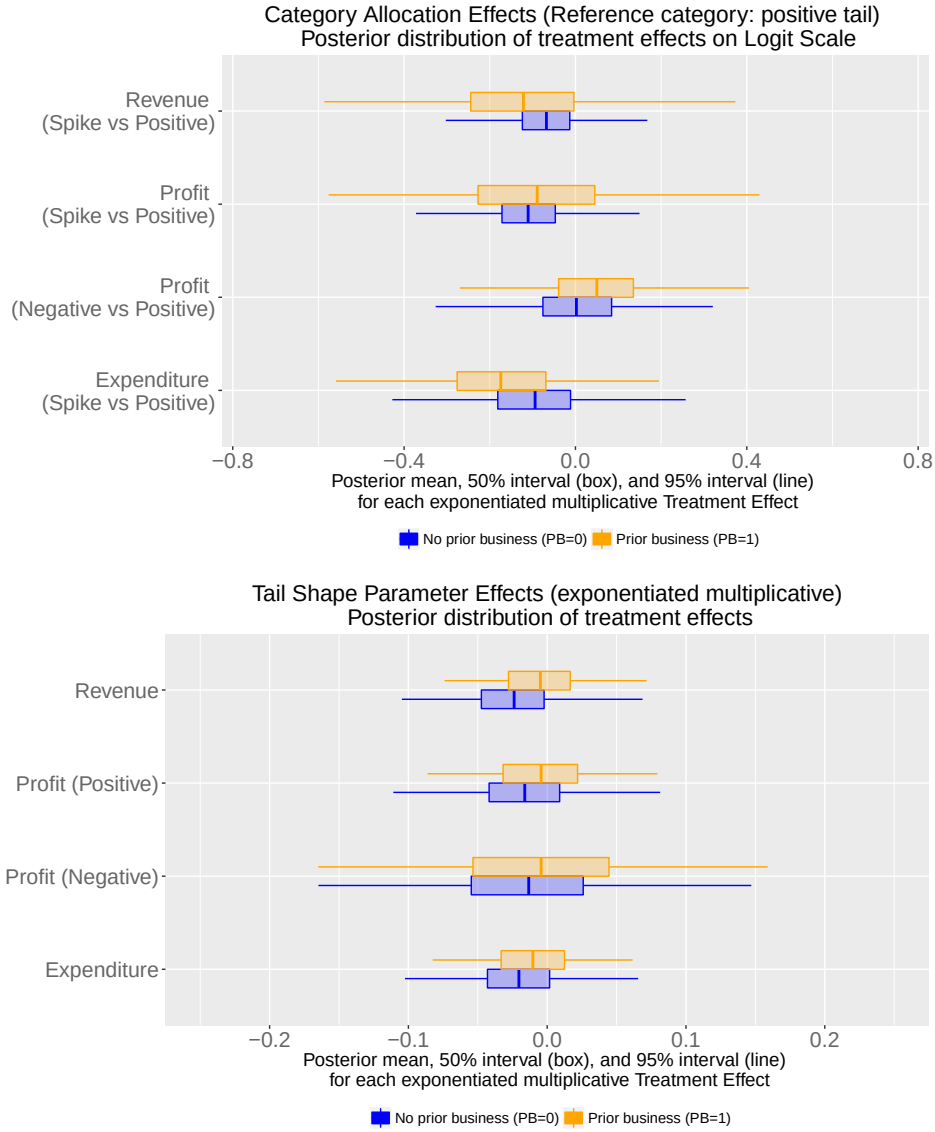

Figure 19: Upper panel: Posterior distributions for the logit treatment effects (π_j) on category assignment split by prior business ownership. These treatment effects are specified as an exponentiated multiplicative factor on the control group proportion of households in the category: if $\tilde{\pi_j} = 0$ 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.

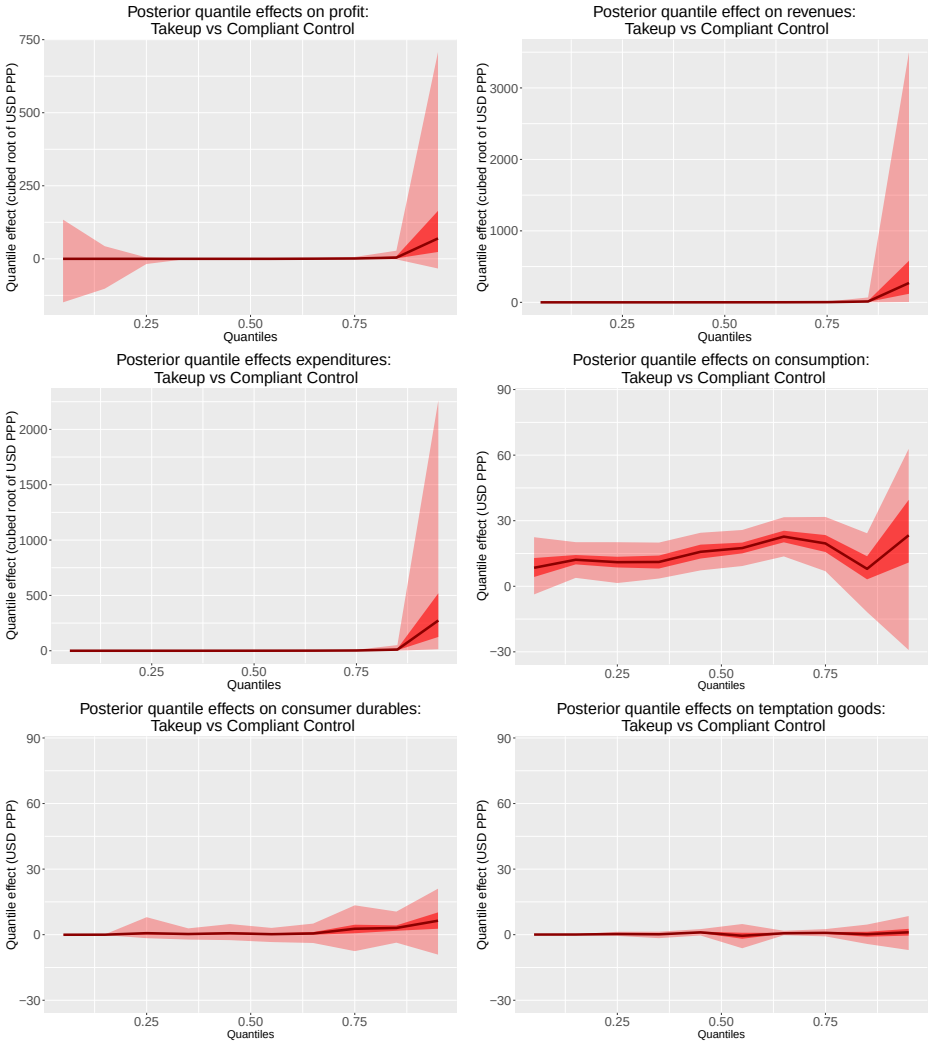

Figure 20: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Compliant control households who did not take up. This effect should overestimate the true impact of microcredit on those who take it up in a simple selection framework. Consumption variables 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.

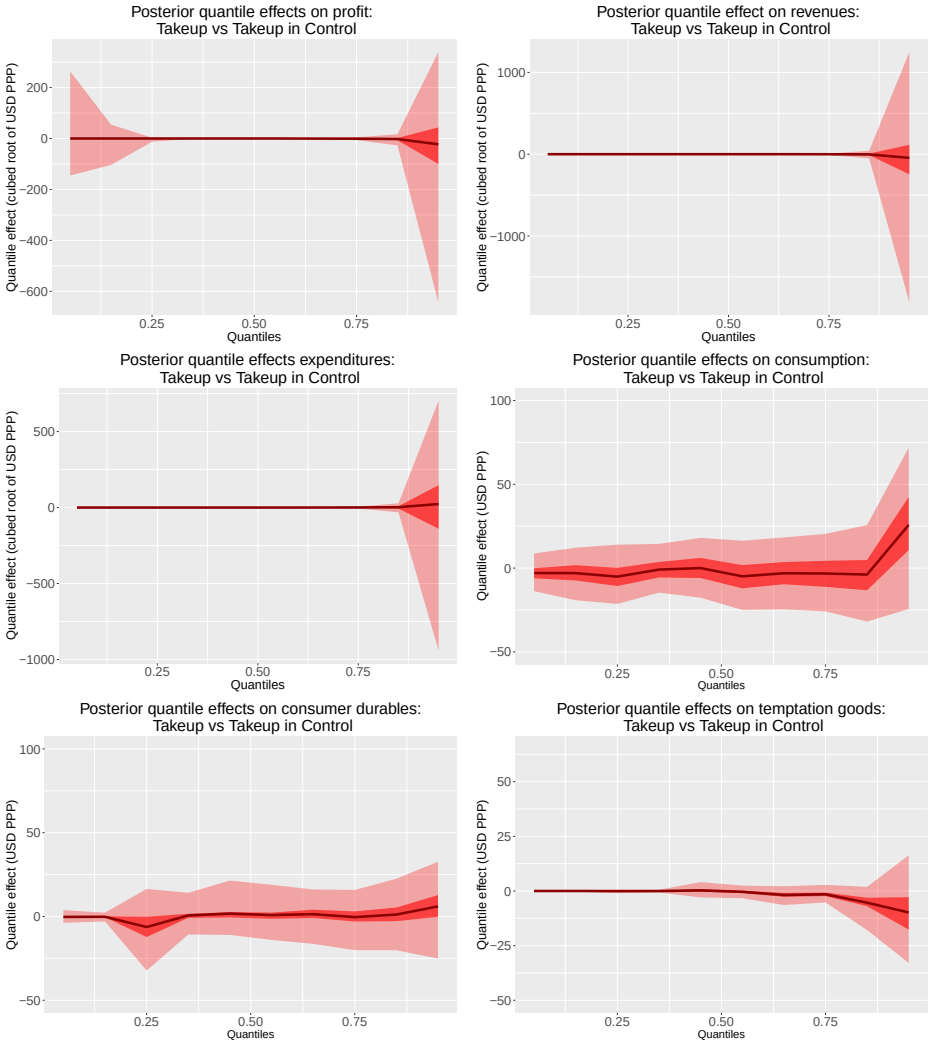

Figure 21: General Quantile Treatment Effect Curves for Business Outcomes: Treated households who took up vs Control households who took up. This effect should underestimate the true impact of microcredit on those who take it up in a simple selection framework. Consumption variables are in USD PPP per 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.

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.

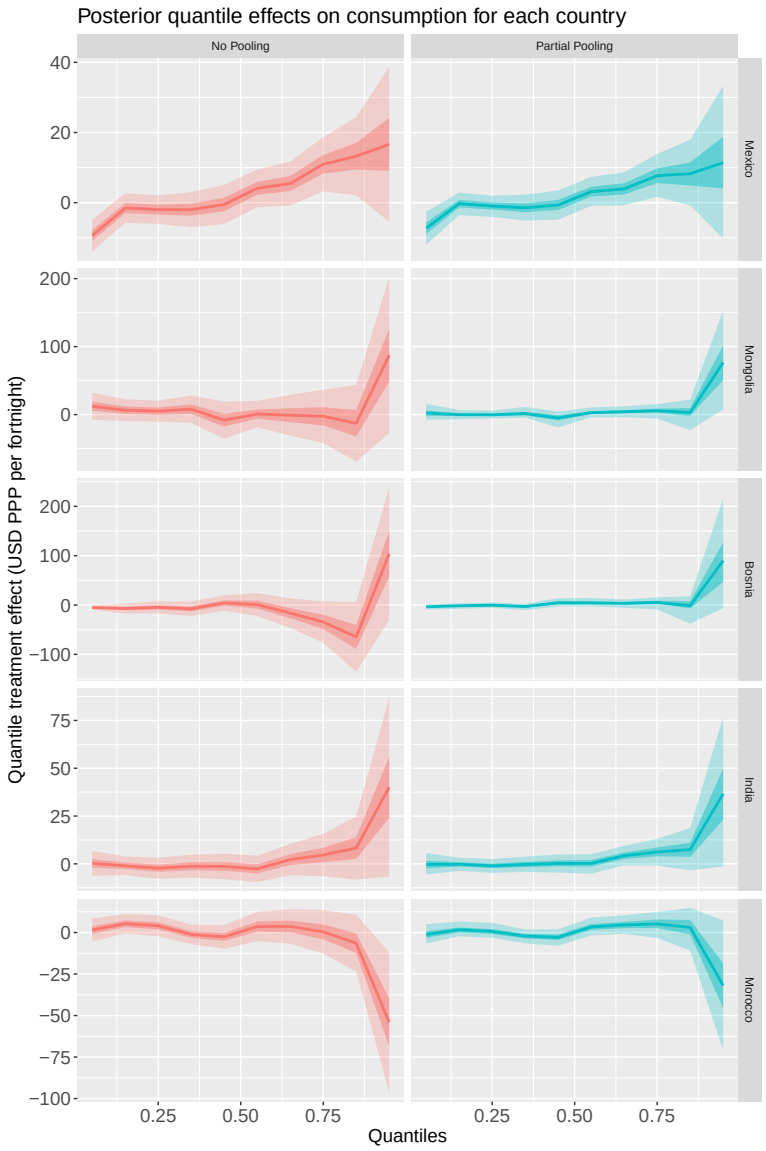

Figure 22: Site by site results for the consumption outcomes. [Back to main]

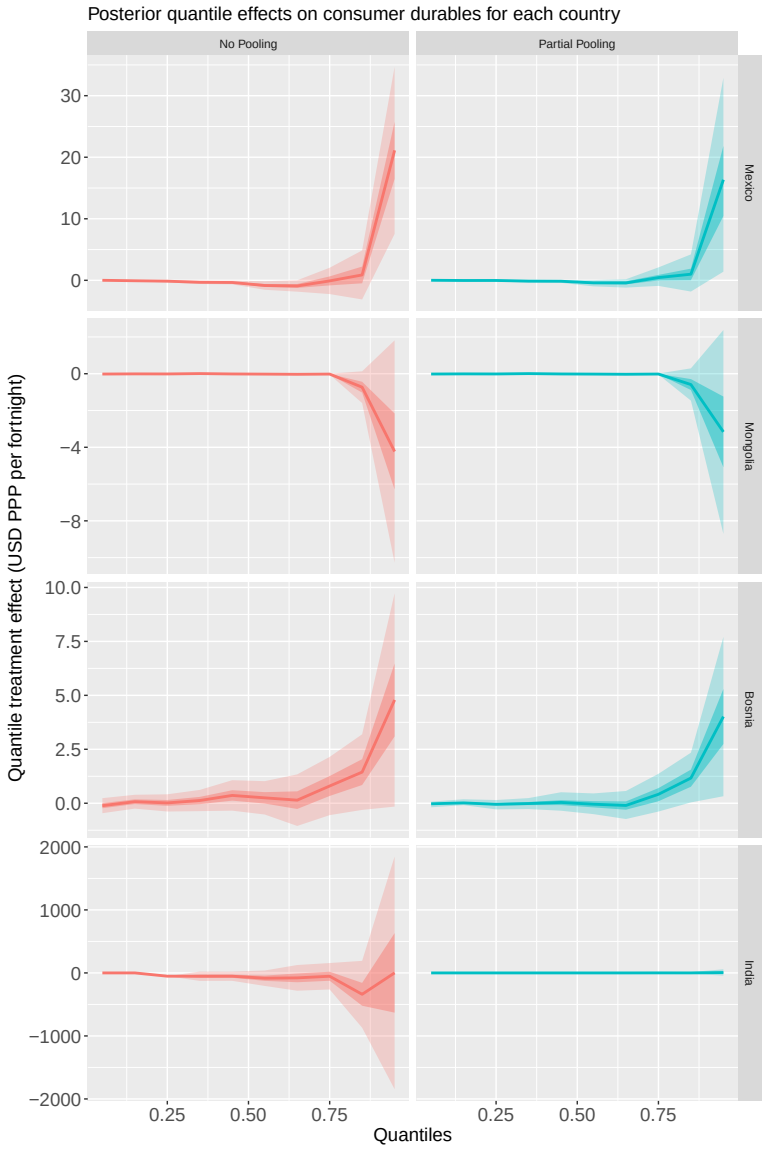

Figure 23: Site by site results for the consumer durables outcomes.[Back to main]

Figure 24: Site by site results for the temptation outcomes. [Back to main]

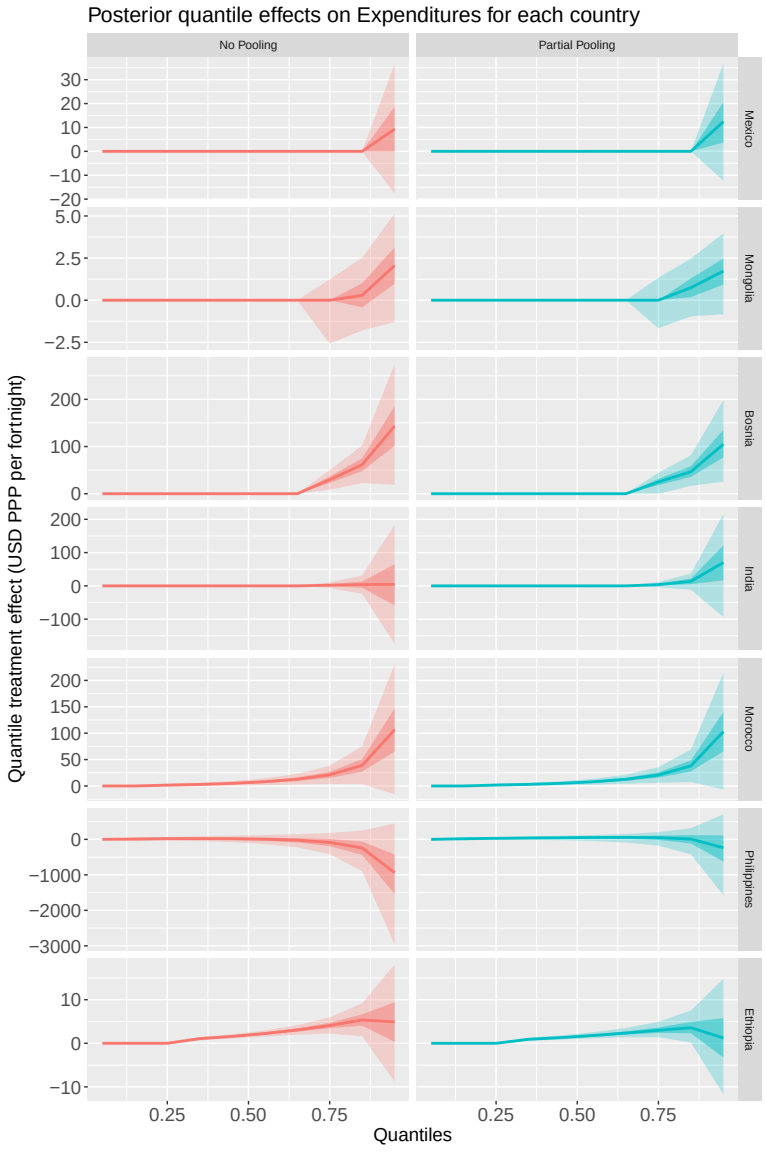

Figure 25: Site by site results for the Profit outcomes. [Back to main]

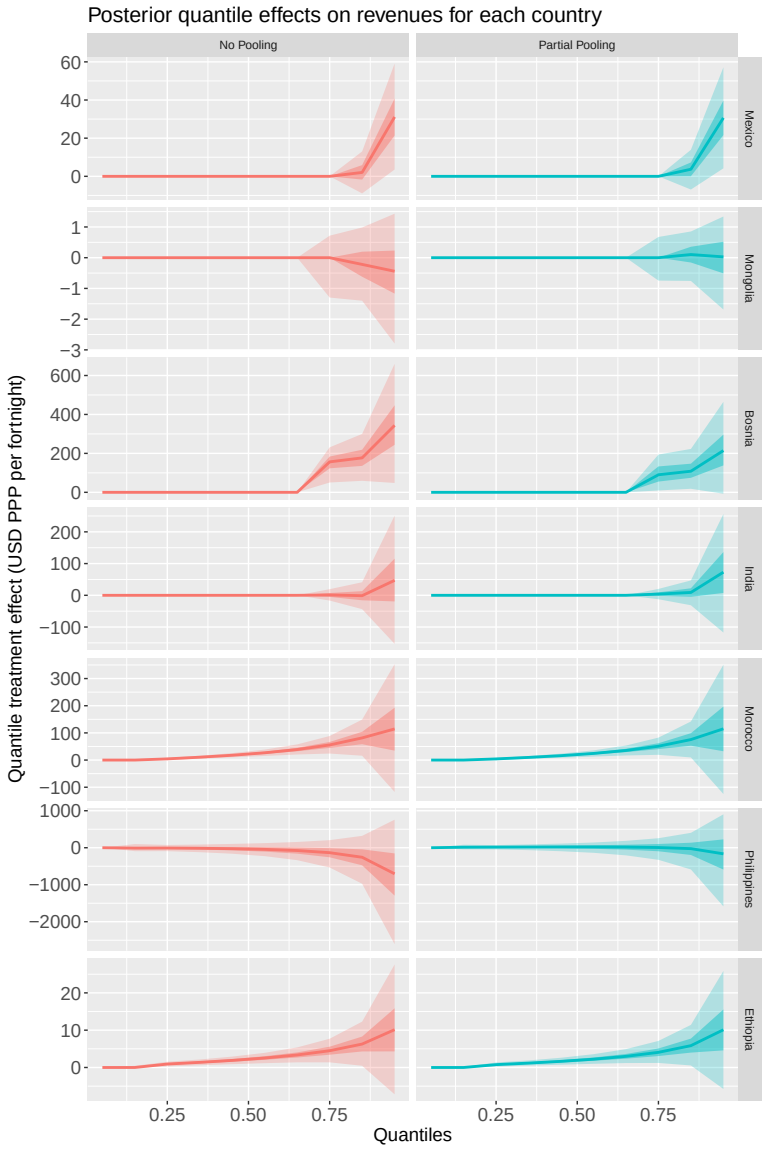

Figure 26: Site by site results for the Profit outcomes. [Back to main]

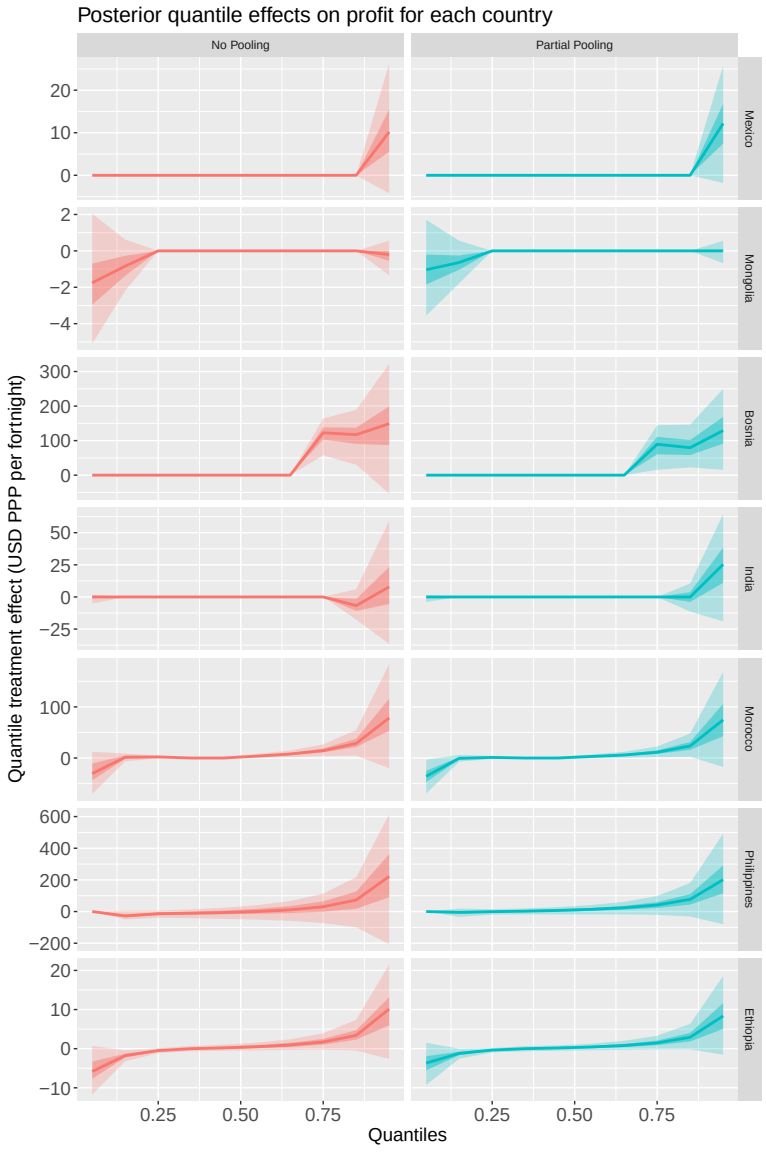

Figure 27: Site by site results for the Profit outcomes. [Back to main]

C Tabular results for Posterior Inference on LogNormal Models

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

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

the excess kurtosis is

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

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

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

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

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

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

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

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

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

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

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

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

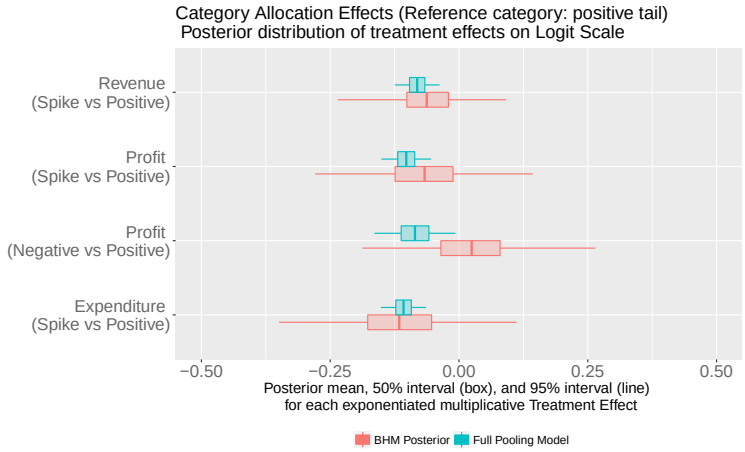

Figure 28: Posterior distributions for the logit treatment effects (π_j) on category assignment. These treatment effects are specified as an exponentiated multiplicative factor on the control group proportion of households in the category: if $\tilde{\pi_j} = 0$ the effect is zero, if $\tilde{\pi_j} < 0$ the treatment increases the proportion of households in the positive tail relative to other categories. [Back to main]

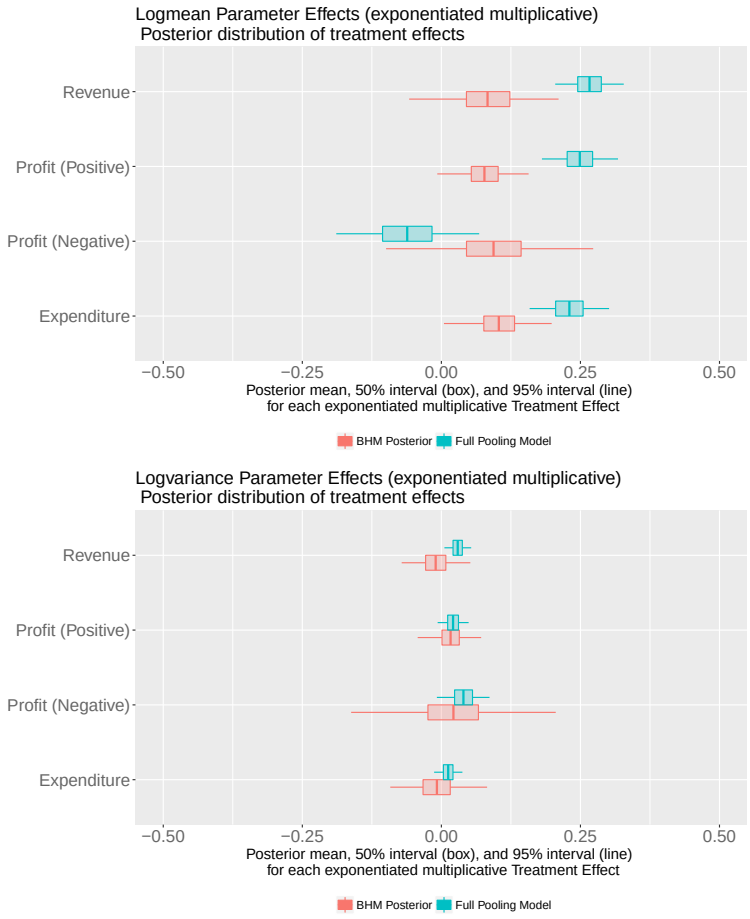

Figure 29: Posterior distributions for the location treatment effects (τ_j) and the scale treatment effects (σ_j^t). [Back to main]

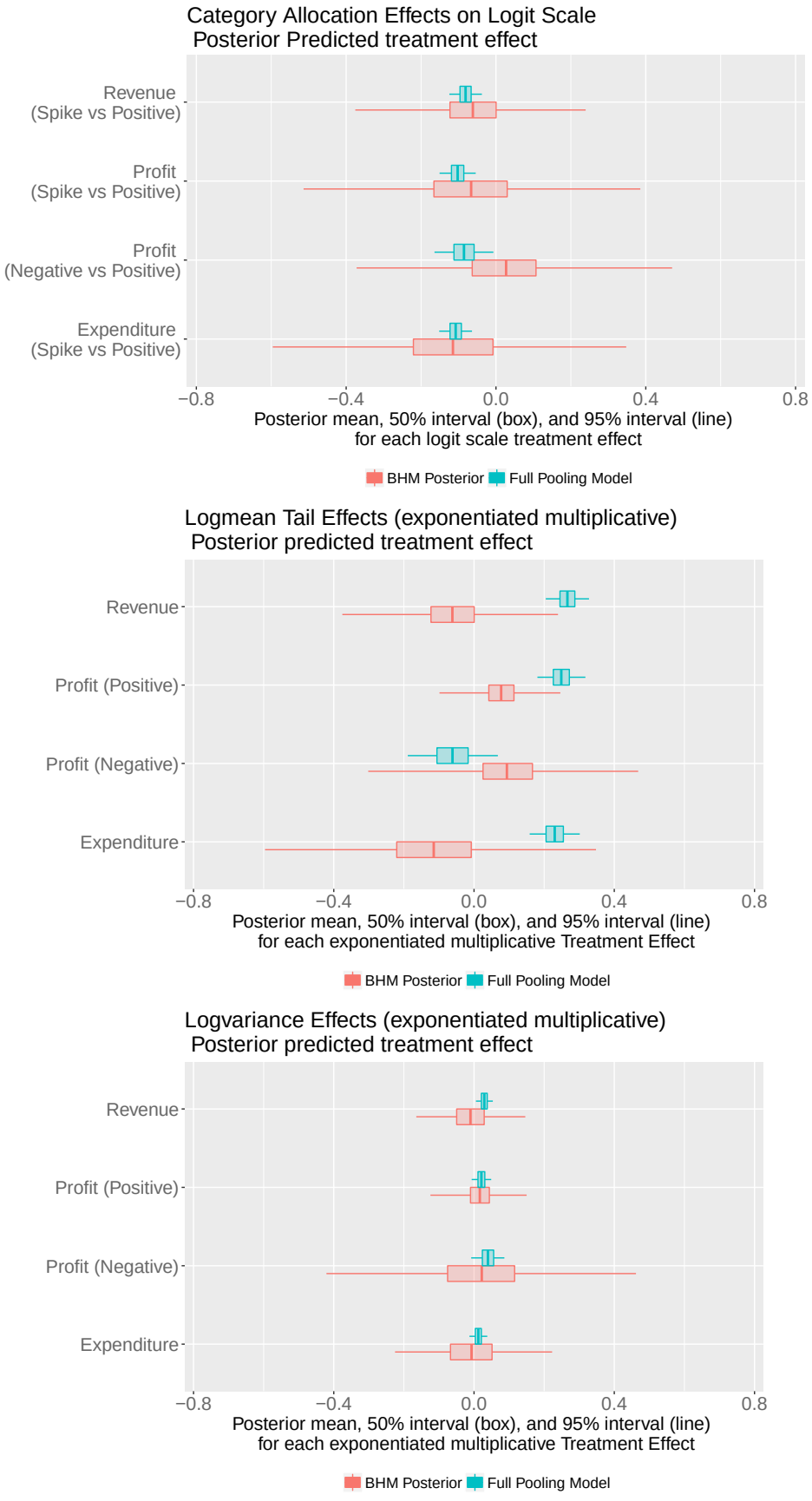

Figure 30: Posterior predicted distributions for the logit treatment effects on category assignment and tail shape effects. [Back to main]

D Variance Effects

D.1 Dispersion Treatment Effects

In a previous version of this paper, and as specified in my pre-analysis plan, I carried out evidence aggregation on the impact of microcredit on the variance of the outcome distributions. However, the detection of extreme kurtosis in the variables rendered the results of this analysis somewhat questionable. As a result, I attempted to fit aggregation models to the Median Absolute Deviations as my alternative measure of dispersion for these distributions. The methods and results of all of these analyses are shown below.

First consider 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.[19] Hence, the standard deviation for a household n in site k with treatment status T_{nk} is:

$$\sigma_{y_k} = \exp(\eta_k + \gamma_k T_{nk}). \tag{D.1}$$

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 the local parameters, is:

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

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.[20] 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:

$$\begin{pmatrix} \mu_k \\ \tau_k \\ \eta_k \\ \gamma_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \tau \\ \eta \\ \gamma \end{pmatrix}, V\right) \; \forall \; k \tag{D.3}$$

Together, equations D.2 and D.3 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

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

[20] 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 $\gamma_k \neq 0$).

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:

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

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, 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 not substantially different.

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

$$MAD(\mathcal{Y}_k) \equiv \frac{1}{N_k} \sum_{n=1}^{N_k} |y_{nk} - \bar{y}_k|. \tag{D.5}$$

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:

$$\begin{aligned}
\hat{\tau}_k &\sim N(\tau_k, \hat{se}_\tau^2) \; \forall \; k \\
\hat{\mu}_k &\sim N(\mu_k, \hat{se}_\mu^2) \; \forall \; k \\
\hat{\Delta}_k &\sim N(\Delta_k, \hat{se}_\Delta^2) \; \forall \; k \\
\hat{\Gamma}_k &\sim N(\Gamma_k, \hat{se}_\Gamma^2) \; \forall \; k.
\end{aligned} \tag{D.6}$$

The upper level of the model is conceptually identical to the full data case, and describes the relationship $\psi(\theta_k|\theta)$

as follows:

$$\begin{pmatrix} \mu_k \\ \tau_k \\ \Gamma_k \\ \Delta_k \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \tau \\ \Delta \\ \Gamma \end{pmatrix}, V\right) \quad \forall\ k \tag{D.7}$$

To complete this model, I use the same priors as specified in equations D.4. 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.

D.2 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 11 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 12). 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 13, which compares the results on the location effect from the joint location/MAD model in equations D.6 and D.7 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.

D.2.1 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 14 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 15). 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 D.2.

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 31. 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 using 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.

Table 11: Dispersion Treatment Effects: Mean Absolute Deviation (effect specified as $\exp(\Gamma)$)

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 12: 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 13: 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 14: 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 15: 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]

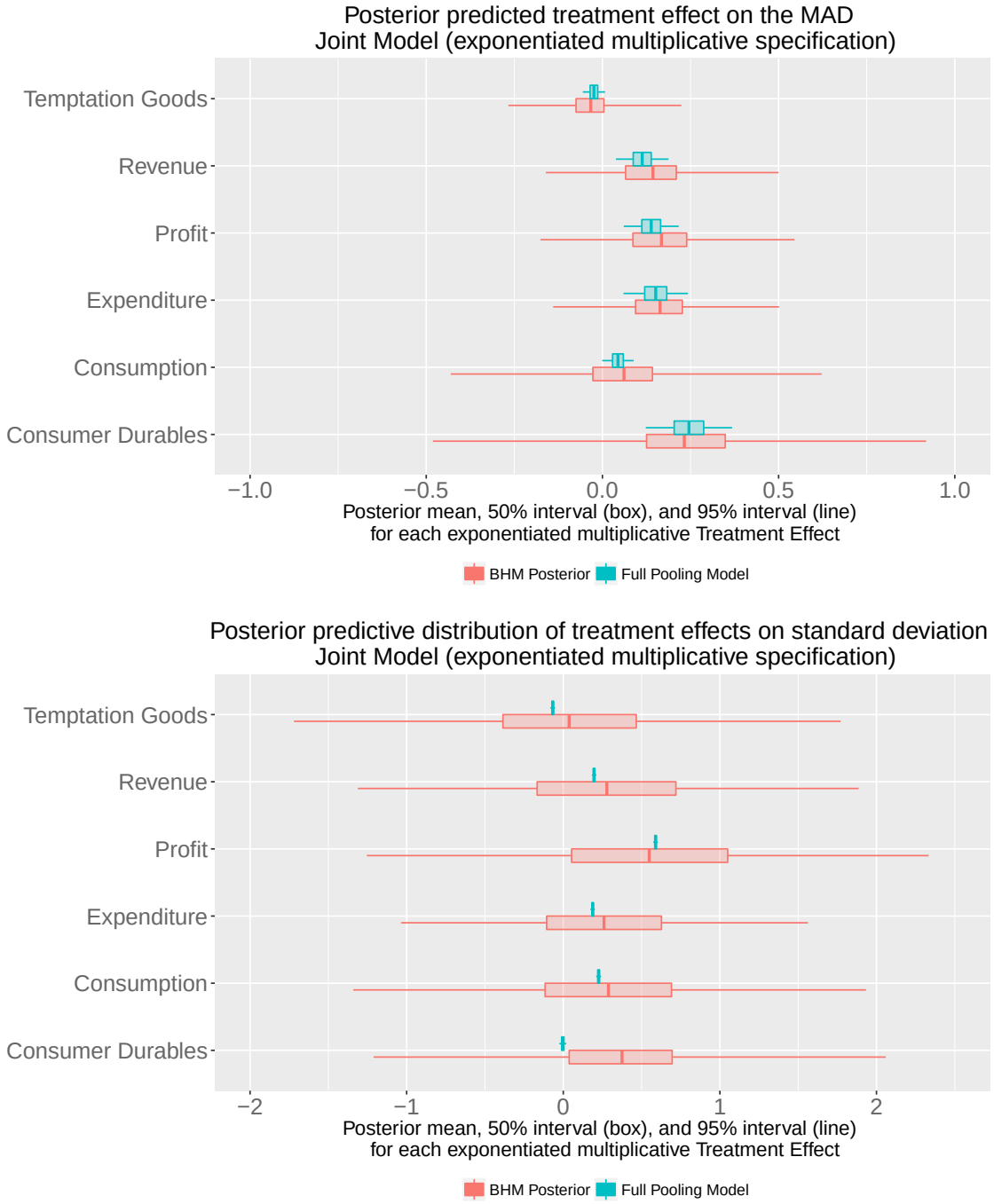

Figure 31: Marginal posterior predictive distribution of Γ_{K+1}, and of γ_{K+1} from the joint model. This is the predicted treatment effect in a future exchangeable study site, with uncertainty intervals that account for the estimated generalizability (or lack of it). [Back to main]

E Examining the Role of Study-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 16 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 16: 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 E, 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.

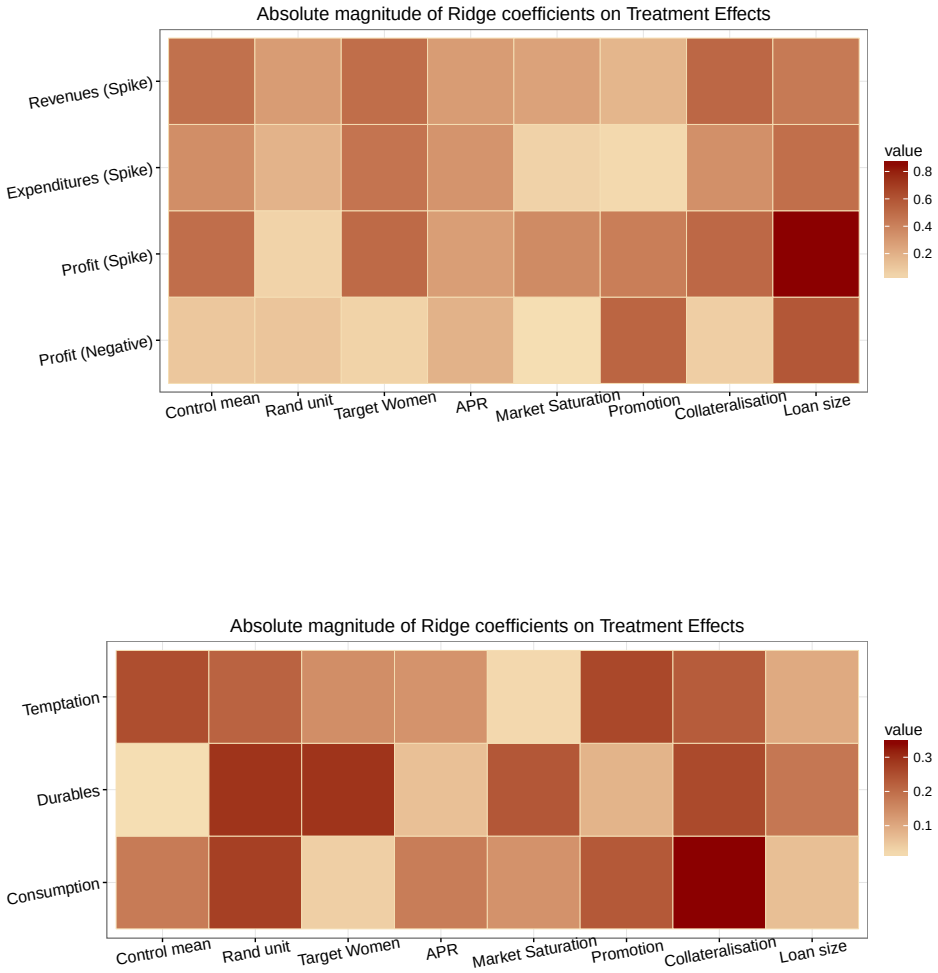

Figure 32: Absolute Magnitude of the Ridge Regression Coefficients for all outcomes and covariates [Back to main]

References

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

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

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

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

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

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

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

[8] 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.

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

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

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

[12] 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.

[13] 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.

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

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

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

[17] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

[31] 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

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

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

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

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

[36] 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.

[37] 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.

[38] 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.

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

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

[41] Duvendack, M., Richard Palmer-Jones & Jos Vaessen (2014) "Meta-analysis of the impact of microcredit on women's control over household decisions: methodological issues and substantive findings", Journal of Development Effectiveness, 6:2, 73-96

[42] 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.

[43] 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.

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

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

[46] 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

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

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

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

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

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

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

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

[54] 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

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

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

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

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

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

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

[61] 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

[62] 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.

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

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

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

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

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

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

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

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

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

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

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

[74] 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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