

Ultra-Poor BRAC Graduation Program: a Bayesian Hierarchical Modeling Impact Evaluation ¹

Uncertainty around external validity

Florian De Bundel

2021

Abstract

What is the average impact of the BRAC Ultra-Poor Graduation program? Despite numerous impact studies, the debate remains partly unsettled because of the important heterogeneity in program effects across different places. In this thesis, I apply Bayesian Hierarchical Modeling on a dataset that includes six Graduation pilots to jointly estimate this heterogeneity and the average effects of the intervention. I find an important true variability in effect across places and use it to estimate the realistic average impact as well as site-specific program effect. I find that the cross-pilots effectiveness of the intervention is not as statistically significant as previously thought. The analysis highlights the limited external validity of the site effect estimates.

¹ I would like to thank Prof. William Parienté for his guidance and support through all the project. I am thankful to Guillaume Dallemagne for his precious help with data management, and to Thomas Keutgen for giving me access to the massive computational power of the platform Support en Méthodologie et Calcul Statistique (SMCS). I am also greatly thankful to Racheal Maeger and Eva Vivalt who pionnied in the field of Bayesian Hierarchical Modeling in economics and have inspired me all along. Finally, thanks to my family and friends, especially Julien De Bundel and Alice Dechamps, for their unfailing support.

1. Introduction

The World Bank estimated that 689 million people found themselves in extreme poverty in 2017 in the world. The COVID-19 is expected to push an additional 100 million individuals below the extreme poverty line in 2020 alone while climate change is already aggravating the situation in the short term ([World Bank, 2020](#)). Many programs are implemented by states, NGOs and international organizations to tackle this massive issue.

Scholars and economists are working to better understand how and why such programs can be effective. They play a crucial role in informing policymakers about the necessary adjustments of programs and their probable impacts when implemented in a new location. In recent years, a lot of Randomized Control Trials (RCTs) evaluated treatment effect of many anti-poverty interventions in different locations with high internal validity. On the one hand, effectiveness of a program widely differs across individuals and contexts. On the other hand, some form of information aggregation is needed to draw conclusion and inform policies. This is directly related to the crucial question of external validity of the Treatment Effect (TE) estimates and the extrapolating power of impact analysis. Bayesian Hierarchical Modeling (BHM) is a new meta-analysis framework in economics that allows to find a middle ground between no aggregation at all (hypothesis of infinite heterogeneity) and total aggregation (hypothesis of no heterogeneity) across different impact evaluation studies.

In this master thesis, I use BHM to raise the question of common effectiveness of a poverty reduction program evaluated in six different countries by taking into account the heterogeneity in treatment effect across sites. Using the data collected by Banerjee et al. ([2015](#)) on BRAC Ultra-Poor Graduation program pilots, I confront my BHM treatment effect estimates and those found through the two extreme hypotheses commonly employed in classical meta-analysis.

I find an important heterogeneity in the treatment effect due to context variation. Especially for the variables the most impacted by the intervention. By taking those heterogeneities into account, I estimated cross site treatment effect means that are similar or higher than those of the original article. However, the uncertainty around those estimates increases for all variables and dramatically rises for some. As a result, I find statistically significant treatment effect for only two main variables two years after the end of the intervention. Whereas eight of the ten main variables present statistically significant TE using Banerjee et al. ([2015](#)) methodology. The BHM approach has better features to estimate cross sites treatment effect in heterogeneous environments. It highlights the relatively limited external validity of each site estimated treatment effect. In addition, the BHM suggests the size of the extrapolation error made in pooled OLS estimation procedure when the treatment effect widely differs across sites.

The next section presents the Ultra-Poor Graduation program and some impact evaluations of its effects. Section 1.2 reviews the literature on Bayesian Hierarchical Modeling as well as on heterogeneity in impact evaluation. In the methodology part, I first detail the data in section 2.1, then present the hierarchical feature of the BHM in section 2.2. Next, I briefly develop the rationales behind Bayesian statistics and section 2.4 formally presents the implementation procedure of BHM for the Graduation pilot dataset. Sections 3.1 to 3.3 compare different settings of BHM and introduce the approximation of treatment effect heterogeneity across sites for each variable. Section 3.4 formally presents the cross-site treatment effect BHM estimates, and section 4 explores possible tracks to explain this heterogeneity.

1.1 The Ultra-Poor Graduation program

The Ultra-Poor BRAC Graduation program (Graduation) is the materialization of an integrated approach to improve the lives of the ultra-poor. Based on six pillars (capital transfers, consumption support, technical skills training, saving encouragement, health education and services, and home visits), it is constructed to provide “the big push” to move the targeted participants beyond the extreme poverty line. The program has been recognized as a valuable tool to achieve the Sustainable Development Goal 1 of the United Nation: “End poverty in all its forms everywhere”. The four main goals of Graduation are social protection, livelihood promotion, financial inclusion, and social empowerment (Coates & MacMillan, 2019).

The program is partly built on the livelihood approach which focuses on self-employment activities. It relies on the importance of developing opportunities for the poor rather than putting an emphasis on the consumption-aspect of the policy approach (Banerjee & Newman, 1993; Balboni et al., 2020). However, time and experience have shown that anti-poverty programs should not be limited to this aspect but rather should provide a holistic set of service packages for a sustainable poverty trap escape (Banerjee et al., 2020). Indeed, ultra-poor are also caught in a self-perpetuating trap of hopelessness (Kharas et al., 2020). They feel invisible, deprived of basic capabilities, with very low sense of worth. Consumption-based and asset transfer programs alone cannot tackle these issues properly. No more than hope and self-esteem alone can buy food and help you find a stable employment. The Graduation approach aims to face both these psychological challenges and material problems. The program targets the “ultra-poor”, the poorest within the poor communities, those living under the extreme poverty line of 1.9 international dollars per day. It tries, within a 24 to 36-month period, to “graduate people out of poverty” (Dharmadasa et al., 2016). Participants are usually selected through village meetings followed by a staff verification (Banerjee et al., 2015). Graduation does not mean end of poverty, rather that participants might find themselves on an upward trajectory to get out of extreme poverty.

The idea behind the Graduation program comes from a large Bangladeshi NGO named BRAC – Building Resources Across Communities - which operates development programs since 1972. BRAC is now considered as the world largest NGO in terms of membership, scope, and budget (MacLennam et al., 2017). To address the question of its adaptability to other contexts than the Bengali context, the Sustainable Livelihoods Project was launched in 2006 by Consultative Group to Assist the Poor (CGAP) and Ford Foundation. The project implemented the approach in ten pilots within eight countries: Ethiopia, Ghana, Haiti, Honduras, India, Pakistan, Peru and Yemen. Six of those were evaluated by researchers from the Abdul Latif Jameel Poverty Action Lab (J-PAL) and from the Innovation for Poverty Action (IPA). The impact evaluation was published in 2015 in the Science Magazine under the name “*A multifaced program causes lasting progress for the very poor: Evidence from six countries*” (Banerjee et al., 2015). As the title suggests, sustainable and significant impacts have been found across countries. I discuss their findings and describe their evaluation method and data in the first section of the methodology part.

Other Graduation programs have also been evaluated by RCTs studies. Bandiera et al. (2017) found significant long-term impacts of a similar BRAC holistic program of assets and skills transfer to poor women in Bangladesh. Banerjee et al. (2016) re-evaluated the pilot site of West Bengal seven years after the beginning of the Graduation implementation and estimated important, sustainable, and even growing impacts in all categories of outcomes. Sulaiman et al. (2016) reviewed 48 different antipoverty programs of different kinds: livelihood development programs, unconditional

cash transfers and Graduation approach. They demonstrated that the former was the most cost-effective (i.e., greatest impact per dollar) with the most lasting impacts when targeting the extremely poor people. More recently Balboni et al. (2020) evaluated the long-term dynamics and found a significant impact of the BRAC program in Bangladesh up to 11 years after the beginning of the program in 2008. Banerjee et al. (2020) reanalyzed the results of Banerjee et al.'s (2015) study of the Ghanaian case and highlighted the superiority of the Graduation approach in comparison to simple asset transfers or access to credit institutions. In addition to the aforementioned studies, several others also pointed out the effectiveness of the Graduation approach (Ahmed et al., 2009; Hulme & Moore, 2008; Mallick, 2009; Martin, et al., 2008; Mukti et al., 2014; Bandiera et al., 2013; Das et al., 2010; Raza et al., 2012; Islam, 2019).

However, it is interesting to note that Bauchet et al. (2015), who evaluated the second Indian pilot in Andhra Pradesh, did not find such a positive impact. Indeed, they observed that the labor market with large wage-employment rate counteracted the program's effects on income and time use. Even though the global impact of Graduation, evaluated by Banerjee et al. (2015), was positive and sustainable, authors pointed out the weaker results in Honduras where a large share of the livestock transfer quickly died. In addition to that, Kidd and Bailey-Athias, in *Debate Graduation* (2017), highlighted the important inclusion error (i.e. participant above the extreme poverty line) in the six pilots evaluated by J-PAL and IPA researchers, especially in Pakistan and Peru. They also emphasized the modest absolute impact on household consumption, even though it may seem important in relative terms. In Bangladesh for example, a year after the end of the program, the average increase in earnings by day was only around 0.06 nominal 2007 USD. Kidd and Bailey-Athias also raised the question of the relevance of the cost-effectiveness analysis of Banerjee et al. (2015), as well as of Bandiera et al.'s (2016) studies. They argue that since the sustainability over time of the program impact is not sufficiently proven (i.e., only evaluated shortly after the interventions ended), the assumptions about the future gains of the program, even though discounted in the cost-benefit analysis, remains disputable.

In the meantime, the reasoning of the Graduation Approach is now being adopted by more than a hundred programs across 40 countries by NGOs, governments, and international institutions (MacLennam, 2017). In 2019, over 1.9 million households had directly benefited from the Graduation programs which represent more than 7 million people (Coates & MacMillan, 2019). It has been adapted to a wide range of countries (i.e. on four continents) and contexts (e.g. urban, rural, mixed), adopted site-specific objectives (e.g. with refugees) and has targeted different groups of people in need (e.g. women, indigenous, youth). For Hashemi et al. (2016), there are key features of the Graduation approach that allow for its scaling up: its design vis-à-vis of potential livelihood opportunities, its transparent targeting, the relatively short period of assistance necessary, the linkages with other public and private assistance programs, and the quality of the local staff who understands and assists individuals in their graduation process.

The idea behind the different pilots launched by CGAP and the Ford Foundation is to try a similar program across different countries, then, check whether its impacts remain cost-effective across contexts. We can expect important inter-site heterogeneity in the TE due to specific contexts and different program implementations. And indeed, Banerjee et al. (2015) rejected the F-test of equality of Ordinary Least Squares (OLS) regression coefficient across sites for almost all pooled outcomes. The BHM methodology could thus use this heterogeneity to better estimate the treatment effect of the intervention across and within sites.

1.2 Related methodological literature

Randomized Control Trial technic is synonymous with high internal validity. RCT studies generate evidence of “what works” in a given context if well implemented. However, evidence from one place have only limited value for any other context. The question of external validity can be raised as “*how does a rigorous estimate of causal impact in one context j affect our beliefs about the causal impact in other contexts k ?*” (Pritchett & Sandefur, 2014, 6). Cartwright and Munro (2010) see external validity as something which has “*to do with whether the result that is established in the study will be true elsewhere*” (Cartwright & Munro, 2010, 262). This external validity issue is particularly problematic for policy recommendation. With the recent rise in popularity of RCT in development economics, policymakers have tried to implement “evidence-based” interventions using internally robust impact evaluation results without taking seriously enough the variation in context into account (Peters et al. 2018; Gechter, 2015; Deaton 2010; Vivalt 2020; Coville & Vivalt, 2019; Banerjee & Duflo, 2009). In many of the experimental evaluation studies published, their external validity is “*mostly a matter of theoretical speculation*” as point out Dehejia et al. (2021, 217).

Modeling the heterogeneity in TE within a study sample is important to understand the individual characteristics that influence program performance. By using those demographic characteristics, it is possible to get better estimates of the program effect in another setting. To further investigate this kind of approach dedicated to the understanding of within sample heterogeneity in TE see Kowalski (2016), Olsen and al. (2013) and Hartman and al. (2015). However, those articles focus on micro-level variation across evaluation samples due to the low number of impact evaluations of the same program. Therefore, it does not reflect the role of contextual, economic or institutional, factors in the effectiveness of the program.

Those contextual settings play an important role in the variability of the TE. Prichett and Sandefur (2014) for instance presented results with a higher heterogeneity due to context-level variables than to individual-level variables in class size effect evaluation. Orr et al. (2019) also examined TE of an education experiment as a function of site-level variables using out-of-sample predictive technic. Raudenbush and Bloom (2015) presented a review of existing conceptual and statistical methodologies for multi-site impact evaluation of intra and inter site heterogeneity of TE. Both within and across sample variations in treatment effect are important to understand how interventions are truly effective. For the external validity question, inter-site variability is particularly problematic and needs to be addressed with caution. Indeed, information about context, program implementation and evaluation process can help the inference process. Taking those considerations into account can potentially lead to better extrapolation and external validity for some other contexts. Heterogeneity can be seen as “*the failure to specify causal models of the processes we are examining*” as mentioned by Deaton (2010, 452), who emphasized the importance of understanding the phenomenon theoretically rather than only taking it into account.

To study these inter-site heterogeneities, a new framework has recently emerged in the field of economics. Indeed, the Bayesian Hierarchical Modeling has been used for decades to evaluate external validity of treatment effect in other disciplines such as medicine and psychometrics. However, only four teams of economists have worked on heterogeneity of TE and external validity of internally valid studies using BHM – at least to my best knowledge. Hierarchical technic is widely used to study site characteristics. But a Bayesian Hierarchical approach allows joint

estimation of the Global Treatment Effect (GTE) and inter-site TE heterogeneity used for Local Treatment Effect (LTE) estimation.

Vivalt (2014 & 2020) used BHM to analyze the heterogeneity of a large set of impact evaluation studies and found important heterogeneity due to contextual factors but failed to find important extrapolating power with only a small set of context variables. She also demonstrated that TE uncertainty can be reduced by taking study and intervention characteristics into account. Bandiera et al. (2016) also applied BHM to evaluate the difference in gender response to performance pay using multiple primary studies. Burke, Hsiang and Miguel (2013 & 2015) reviewed the emerging literature linking climate change and armed conflict using similar framework to discuss variabilities across 55 studies. Meager (2016 & 2019) analysed the average TE and the heterogeneity in seven RCT experiments of microcredit expansion. She highlighted the relatively low TE heterogeneity and pointed out the important role of intervention characteristics in the variation of TE across sites. The Bayesian approach recently gained in popularity with the exponential increase of computational power and with the greater accessibility of implementation software such as Stan.

However, it is important to precise that BHM, even in an ideal situation, cannot answer to all issues linked to external validity of impact evaluation studies. Firstly, it is extremely difficult to evaluate the true variation of TE across sites. Secondly, this framework is not fully robust against four external-validity issues identified by Duflo, Glennerster and Kremer (2008): Hawthorne and John effects, general equilibrium effects, the specific sample problem and the special care problem. Finally, it does not anticipate possible effect due to the scaling up of evaluated intervention (Banerjee and Duflo, 2009).

The BRAC Graduation program evaluated in the six pilots can reveal to which extent the TE is context dependent. By exploiting the inter-site TE heterogeneity, the Bayesian Hierarchical Modeling can lead to more realistic estimations of the global treatment effect and more precise estimates of site-specific TE.

2 Methodology

2.1 Data

I use the dataset from Banerjee et al. (2015) which assesses six of the Graduation pilots launched by CGCP and the Ford Foundation. The program is evaluated in Ethiopia, Honduras, Ghana, India, Pakistan and Peru. The assessment process is made by Randomized Control Trials either at the household level within specific villages or through randomization at both the household and the village level. The total sample size is 10,495 households and varies from 925 in Ethiopia up to 2,606 in Ghana. The randomization procedure was well implemented. Indeed, researchers were not able to reject the test for equality of means across aggregate variables, at baseline, between control and treatment groups. There is little attrition in the samples, with a participation rate of 94% at the end of the program survey and 91% in the follow-up survey a year later. In addition, the compliance rate was at its maximum in all sites, except in West Bengal, India. In this pilot, nearly half of the selected participants did not participate (35% declined the offer and 13% were not finally eligible). Therefore, the results of the evaluation have to be understood as an “intent-to-treat” (ITT) rather than a treatment effect (TE). The details of this evaluation process can be found in the original

article: “*A multifaced program causes lasting progress for the very poor: Evidence from six countries from six countries*” by Banerjee et al. (2015).

All participants from both treatment and control groups were surveyed three times. The first time at the baseline, before the start of the program. The second survey took place at Graduation end (referred to as ‘end’ in the next pages). The follow-up survey took place a year later (abbreviate by ‘fup’), around 36 months after the asset transfers. The surveys have collected data at household and individual levels, for 50 variables around then thematic: consumption, food security, assets, finance, time use, income and revenues, physical health, mental health, political involvement and women’s decision-making.

However, in some countries or surveys, the data was not collected for all variables. If this feature is not significantly disrupting for classical aggregate analysis as the one of Banerjee et al. (2015), it is particularly problematic in case of hierarchical modeling. Indeed, as it will be developed in the next section, BHM needs the information for each variable in each site in order to perform. I, therefore, have to drop those variables from the analysis.

I also compute the thematic indexes, using available variables in all sites in the form of z-scores, like in the original article. I compute those thematic indexes by taking the average of the z-score values of the variables of each thematic-group for each country and then standardize it using the control group within each country and each survey. Many thematic variables are constructed using less household or individual variables than the original article. Therefore, my indexes are not identical to those of Banerjee et al. (2015). In addition, I can neither construct the Mental Health variable for the second survey nor the Women’s Empowerment in the third survey because of missing variables. Household and individual variables used to compute thematic variables are presented in the table 1 in the annex.

2.2 Hierarchical Modeling

We should not expect the same treatment effects in different contexts. There is probably some heterogeneity in the treatment due to different implementation settings and contextual features, to which are added heterogeneity from participants characteristics within each site results. The main challenge is to differentiate between these two kinds of variations, something that can be done with hierarchical modeling (Dehejia, 2014; Meager, 2015; Bandiera et al., 2016). Indeed, by simultaneously considering site-sample variation and dissimilarities in treatment effect across studies, hierarchical modeling can evaluate a relationship between each site results and the inter-site effects of the program.

On the one hand, when the site-level variance is low, results of hierarchical modeling will be close to those of the Full Pooling (FP) model (i.e. when observations are pooled together), meaning that the unexplained heterogeneity across sites is low and external validity is high. On the other hand, if the approximate value of genuine variation across countries is high, it means little external validity. In this case, it will be difficult to extrapolate results in another unstudied context without an additional analysis to explain the factors driving the heterogeneity. The question of Full, Partial and None pooling models will be discussed in the next section.

The heterogeneity parameter can only be evaluated within a hierarchical framework. In addition to the intrinsic value of such estimate, the heterogeneity metric will be used within the hierarchical modeling to get better estimates of the treatment effect. In fact, information about all site effects will be used, to some extent, to infer the local treatment effect of each site. Thus, BHM has two important advantages in comparison with the classical pooled methodology. First, it allows the re-use of available data to analyze the external validity through the question of heterogeneity of impact across sites. Second, since a heterogeneity measure can be interpreted as the degree to which each site contains information about the aggregate results, it is also used in the inference process of the hierarchical modeling by partially pooling the data together.

The formal linkage between the different levels of the hierarchy will be formally presented below. However, the rationale is quite simple. All studies evaluate, in different contexts, a similar program, thus, could be informative one to each other to some extent. Using the words of Bandiera et al. in an article using BHM, “*our beliefs about the treatment effect in any given setting are based not only on the results obtained in that setting but on the results in the other $n-1$ similar settings*” (2018, 18). By formally considering the heterogeneity of treatment effect, BHM may improve the inference process in comparison with simple pooled OLS estimation. The interactions between the levels in the hierarchy are constructed, partially, through the prior specification of the BHM. The point will be detailed in a next section.

The multi-site Graduation effects can be analyzed using the normal-normal hierarchical model (NNHM), also known as the random effect model, which is a widely used framework for meta-analysis (Röver et al., 2021; Borenstein et al., 2010). This kind of model will treat the estimates from the different studies as being from a common normal distribution and will analyze their variance with another normal distribution.

More formally, our model, as well as other Bayesian Hierarchical Models (see Meager, 2016 & 2019; Bandiera et al., 2016; Vivalt, 2015 & 2020; Hsiang et al., 2013 & 2015) is constructed following the relationships presented below (Gelman et al., 2013). The index k is for the site and τ represents treatment effect:

$$\begin{aligned}\widehat{\tau}_k &\sim N(\tau_k; \widehat{se}_k^2) \quad \forall k, \\ \tau_k &\sim N(\tau; \sigma_\tau^2) \quad \forall k.\end{aligned}\tag{eq. 1}$$

At the site level, site estimated treatment effects ($\widehat{\tau}_k$) are assumed to be normally distributed around their true values with a variance estimated by their observed standard deviation around the TE ($\widehat{se}_k^2$). This makes the link between the data and the studies parameters. Since our six sites have been evaluated using RTC techniques, we can assume, by the Central Limit Theorem, that the estimated treatment effect is indeed normally distributed around its true values.

At the population level, the true site treatment effect (τ_k) follows a normal distribution around the treatment effect across sites (τ) with the variance equal to the treatment effect variability between studies (σ_τ^2). This the cross-site variance parameter (σ_τ^2) represents the variability of the TE across sites. Its value can be categorized into 3 classes: it could be zero, infinite, or somewhere in between.

These three possibilities, named as None, Full and Partial pooling of the data, will drive the modeling to distinct directions. The Full Pooling modeling (similar to Banerjee et al., 2015) will not differentiate the countries data and assumes that all study-site approximate a common effect.

In other words, there is a strong assumption that the heterogeneity across sites is zero and that any observed variation in TE comes from intra-study variation (se_k^2). At the other extreme, the None Pooling (NP) scenario will not try to find any commonality across studies and will see each site as totally uninformative in the estimation process of the TE of the other sites. In the former case, no information can be shared between countries and no common effect can be approximated.

Between these two extremes, the Partial Pooling (PP) allows the cross-study estimated heterogeneity to be somewhere between zero and infinity. These two extreme models are particular cases of our main modeling. The PP modeling has an important consequence since it is “*allowing for intermediate degrees of pooling, [...] makes it possible to aggregate evidence across similar but not necessarily identical studies*” (Bandiera et al., 2018, 16).

This rationale can be formalized with the following equation (Gelman and Pardoe, 2006):

$$LTE_k = P \cdot E(\tau)_{K-k} + (1 - P) \cdot \tau_k \quad \text{eq. 2}$$

P represents here the “Pooling factor” (P), the degree to which the estimates are pooled together. This metric is 1 - the average across sites of the ratio of the treatment effect variability to the total variability (sum of between-study variance and average within-study variance of TE):

$$P = 1 - \frac{\sigma_\tau^2}{\sigma_\tau^2 + E(se_k^2)} \quad \text{eq. 3}$$

This Pooling factor is an essential feature of BHM. It allows to estimate the global treatment effect as well as the local treatment effect while taking into account an approximated measure of heterogeneity in program effect across sites (σ_τ^2). Bayesian statistics are more adequate to estimate GTE (τ) and LTE (τ_k) Indeed, BHM makes a joint inference, given the data and the priors, on hyper-parameters: GTE and between-study variance. BHM can estimate simultaneously those parameters. The joint estimation is crucial due to the correlation across different levels of uncertainties: inference on τ and σ_τ^2 depends on estimation of all τ_k and vice-versa (Betancourt & Girolami, 2013). With the Bayesian estimations of the between study variance of treatment effect (σ_τ^2), we can differentiate within the total variation across sites how much is only due to samples idiosyncratic error $E(se_k^2)$.

When the NP model cannot estimate a Global Treatment Effect (GTE), the partial and full pooling models could have different results over the GTE. Indeed, the latter cannot incorporate the heterogeneity of the program effect across countries in its estimates. If there is some heterogeneity in the treatment effect across sites, taking it into account will generate more realistic estimates by increasing the Credible Interval (CI) on both sides of the TE mean.

Meager makes a good point about the approximation of the pooling factor: “*For those concerned about estimating any parameter from a few data points, consider that the alternative is to use zero data points and assume the parameter takes a single known value with certainty.*” (Meager, 2019, 67). Indeed, the measure of heterogeneity of treatment across sites is based on the variation of treatment effect between each site. When we only have a few different sites, like in this case with only six countries, this measure is a rough estimation. This parameter is essential to perform the partial pooling modeling discuss above.

2.3 Bayesian Statistics

Frequentist and Bayesian methods approach statistical inferences from opposite standpoint. Frequentists see parameters fixed and data as random variable whereas Bayesians consider parameters as random and data as fixed. It leads to quite distinct perspectives: when frequentists ask, “*Given a set of parameter values, how likely is it to obtain the values seen in the data?*” Bayesians would rather think in the following way: “*Given that the data are collected, and the experiment will not be repeated, what are the values for the parameters?*” (McNeish, 2016, 2).

Gelman et al. defined Bayesian inference as “*the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations*” (2014, 16). The probability model will be constructed using the hierarchical framework with prior specifications and will infer, for each variable of interest, a posterior distribution for each country (Local Treatment Effect) as well as for the global treatment effect using partial pooling (i.e., considering heterogeneity).

Frequentist methods can be used in hierarchical modeling. However, Bayesian ones have one important advantage. In hierarchical meta-analysis with only a few studies, there is a big issue with the approximation of the inter-study variance. Using prior, Bayesian methods allow to reduce the uncertainty around this approximation by adding information to help the process of simulation and avoid or reduce over-fitting. It is important to note that this process may introduce a bias in the estimates. Frequentist hierarchical modeling would have extremely problematic variance-related issue. In comparison, a potential bias is a minor problem. In addition, the size of the bias, as well as the variance issue, is linked to the informativeness of the prior used.

Bayesian process needs priors on the distribution of the hyper-parameters (treatment effect and its heterogeneity across sites). Two kinds of Bayesian prior are possible: non-informative or informative. On the one hand, the first type is data-driven, meaning that the prior does not add any essential information since it captures its value from the data itself. It is simple to implement, and we do not have to worry about it. On the other hand, informative prior can be particularly useful, especially with a small number of studies, to help the inference given some *a priori* knowledge and to avoid over-fitting. Another key point to make is that there is not unique ‘right’ prior: it only reflects different beliefs of precondition of the BHM. In any way, we can check afterwards the sensitiveness of our results by varying prior specifications.

The Baye Theorem can be implemented using Markov Chain Monte Carlo (MCMC). It is a process of multiple Markov chains which explore, for parameters of interest, the space of possibility given the data and the priors provided. Each chain is constructed by a progression of points linked to each other by a stochastic jump associated with a probability Accept/Reject rule (Gelman et al., 2014). As the chain runs for an extended period, the Markov Chain Monte Carlo estimators satisfy the Central Limit Theorem. Thus, MCMC generates joint posterior distribution conditional on the data and priors provided. This posterior distribution represents all the information and uncertainty about the estimated parameters of interest. Informative priors are quite valuable since they give the possibility to drive the process forward in regions researcher believes to be more relevant. Therefore, it increases the speed and the convergence of the Markov chains (Meager, 2016).

One specific MCMC procedure is the Hamiltonian Monte Carlo (HMC) algorithm. Using Hamiltonian dynamics of the geometry of the region space of high interest, HMC is more efficient than other MCMC algorithms and has higher estimates validity (Betancourt, 2018). As already

developed, the BHM creates a potential problem of correlation between different levels of uncertainty (site and global levels). Indeed, the inference process for the global parameters is dependent on the inferred local parameters of each country, and vice versa. However, the use of HMC algorithm succeeds in overcoming this issue (Meager, 2019). With the help of performing software as Stan, this method has gained popularity and accessibility.

To confirm findings, I run four different chains and compare their results. In this regard, the R^2 statistic from Gelman et al. (2014) is a great diagnosis. It allows to see whether an ensemble chain starting from different points in the parameter space converges to form similar posterior distribution, using each time the same algorithm, data and priors. The MCMC is considered convergent and its estimates as valid when the R^2 does not exceed 1.05; otherwise, the inference process is stopped. This statistic represents the output of a joint test for similarities of the different posterior distributions computed with 5% error possibility (Betancourt, 2017).

Briefly, Bayesian methods are perfectly suited to estimate treatment effect in the context of meta-analysis with only few studies. However, the estimates are sensible to prior distribution specification which drives the algorithm in the parameter space during the stochastic process.

Note that BHM framework requires an important hypothesis to form the joint probability: the exchangeability assumption (Gelman et al., 2014). The model must treat the site-specific treatment effect as i.i.d. given the global mean of effect and its variance. In other words, we should not have additional information about the site treatment effect and its heterogeneity across studies than those in the data itself. However, this assumption is not that difficult to fulfill since if we have additional information, we can simply add it to the data, in the form of additional level of hierarchy or new variables (Bandiera et al., 2018).

2.4 Implementation of BHM for Graduation pilots dataset

Our main modeling, the Rubin model from Rubin (1981), only considers the reported estimates of treatment effects and its standard deviation as inputs in the BHM. Those are found through OLS regression of the variable of interest over the treatment status for each country. The hierarchical model is constructed as a normal-normal likelihood as developed previously. The first hierarchical level assumes that the site-specific estimated treatment effect follows a normal distribution around its true value (which should be correct since the studies used RTC) given the site-specific approximated standard deviation. The second level supposes that the site-specific true value is normally distributed around the general treatment effect across sites given the global variation of the treatment effect across sites (which is a widespread practice in the meta-analysis literature, see Borenstein et al.(2010)).

I use the software R for statistical computing and graphics in collaboration with the RStudio interface for the main data management and computation of this paper. For Hamiltonian algorithms in the Bayesian hierarchical modeling, I use Stan link to RStudio through C++ compiler.

The process starts by extracting raw data, at individual and household levels, from the Banerjee et al. (2015) article in Stata format and formatting it for the present analysis. Alongside visualizing some descriptive statistics, I identify several variables used in the original analysis that will be problematic. Indeed, using a hierarchical framework, I cannot use a variable if it has not been

collected in all sites, especially when working in heterogeneity of treatment effect. I have to drop those variables from the BHM analysis.

To allow comparability between different variables, I present the outcome of this analysis as mean standardized treatment effect for both constructed thematic indexes and variables from the original surveys. By subtracting the mean of the control group of the same survey from the value of each variable, by country, and then dividing it with the standard deviation of the control group still of the same survey and country, I transform the dataset into z-scores.

To perform the Rubin model I first compute, for each observation of each variable, survey and country, a linear regression of the z-scored values of the variable of interest over a dummy representing the treatment status of the individual or household (i.e., 0 for the control group and 1 for the treated group). I then take the mean value by site as the estimated treatment effect ($\hat{\tau}_k$) of the intervention for the variable of interest. I also store the observed intra-site dispersion of the results for each variable and survey (represented by the standard deviation of the regression outputs), as the estimated standard deviation of the variable treatment effect by site ($\widehat{se}_k^2$). I also perform a similar process using the non-standardized values to have a better understanding of the treatment effect when looking at each variable separately.

Using a dataset formed by the $\hat{\tau}_k$ and $\widehat{se}_k^2$ for each variable TE, survey, country, standardized and non-standardized, I start the process of Bayesian inference within the hierarchical framework discuss above. Since I do not have important prior knowledge of the probable value of the hyper-parameters and do not want to introduce too much bias in my estimates, I use very weakly informative priors (Gelman et al., 2014). The prior on the hyper-mean, the average effect across sites, is a normal distribution around a mean of zero and an excessively large variance. This structure assumes that “*that causal effects should not be thought of as large unless data contains evidence to the contrary*” (Wiecek, 2021, 7). The prior for the hyper-standard-deviations, the variance of effects across groups is uniform law from 0 to 10 time the naive standard deviation of across the OLS estimated treatment effect at site level.

$$\tau \sim Normal(0; 100)$$

$$\sigma_\tau \sim Uniform(0; 10.\tilde{\sigma}) \quad \text{eq. 4}$$

For each variable and end line, I estimate the posterior distribution of the global treatment effect using the Partial Pooling with the weakly informative prior specification presented. I also perform the Full (i.e., hypothesis of zero inter-site TE variance) and None (hypothesis of infinite inter-site TE variance) pooling models for comparison. During this process, in addition to the GTE posterior distribution, I also store the posterior distribution for each site separately and extract heterogeneity measures.

To have a better understanding of the robustness of my results, I also perform some predictive information criterion analysis (Gelman et al., 2013). The fit of the model can be assessed by using posterior predictive checks (Rubin, 1981) with the Expected Log Predictive Density (ELPD) for a new data point.

$$ELPD = E_f \left(\log p_{post}(\ddot{y}_i) \right) = \int (\log p_{post}(\ddot{y}_i)) f(\ddot{y}_i) d\ddot{y} \quad \text{eq. 5}$$

Even though several ways are possible to evaluate Bayesian model, one is particularly fit for HBM. The leave-one-out cross validation (LOOCV) technic runs multiple BHM and leaving each time one site out of the analysis to evaluate the predictive capacity of the posterior distribution found over the omitted group (Gelman et al., 2013). The output is a cross-validation information criterion (LOOIC): -2 times the ELPD averaged over the K models run. Closer the LOOIC value is to zero, better the model fits the data. I use this method to compare the accuracy of GTE estimates by my main modeling to its counterpart Full pooling (i.e. which is close to the original article modeling).

Notice that BHM estimation is extremely RAM intensive. I use a machine of computational capacity of 32 cores and 195 gigabytes of RAM, and the calculation took around 15 hours. This is due to the simulation algorithm with the settings of the BHM that I use. Indeed, to increase the accuracy of the estimates and the convergence of the MCMC chains, I parametrize the algorithm with 40 000 iterations by chain, the distance of 0,99 between two points for the stochastic jump and the probabilistic Accept/Reject rule and this for four different chains by variable and by modeling type.

3 Results

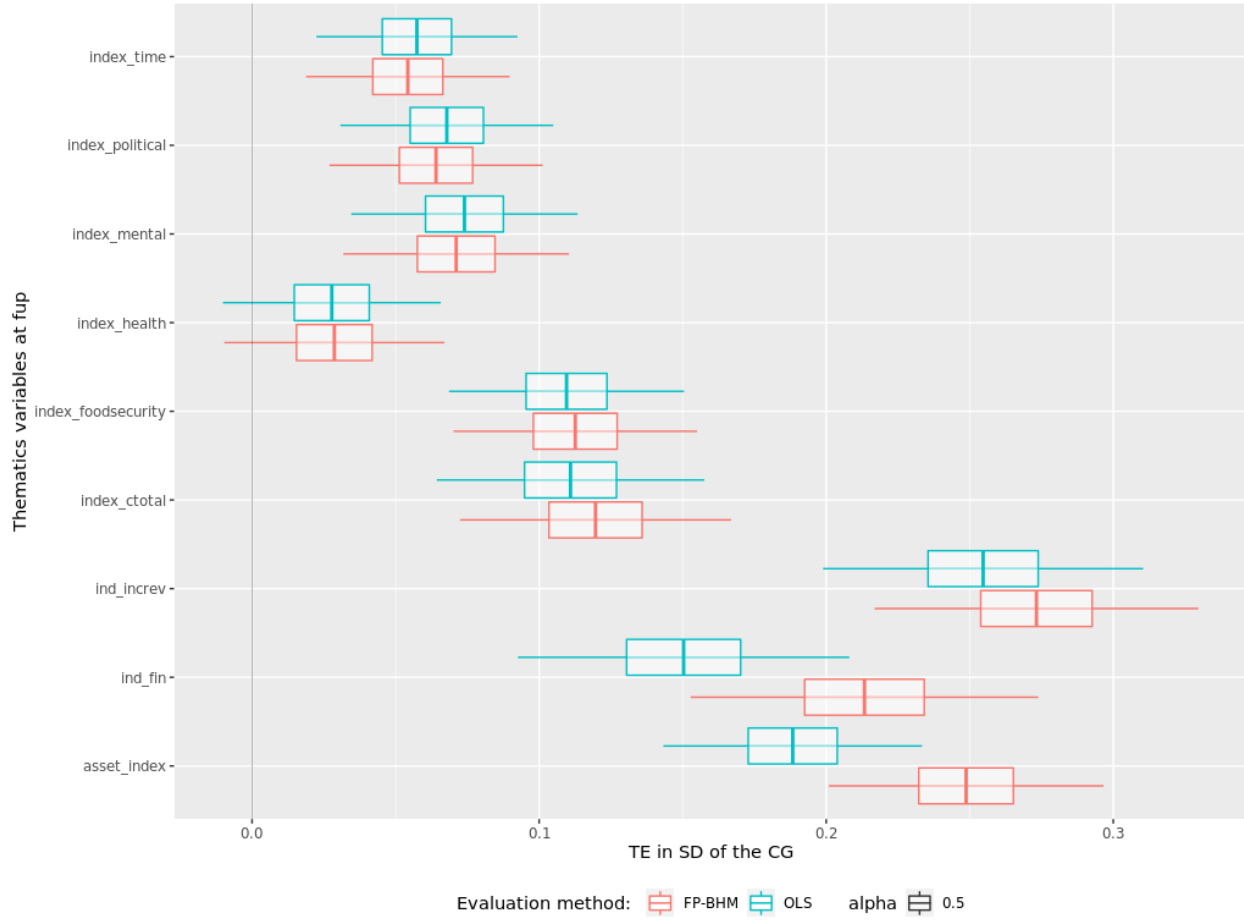
All the following BHM estimates are computed using Rubin modeling developed in the methodology part. As explained, with enough input data and iterations, the chain convergence and the Bayesian estimates are reliable. Using 40.000 iterations and more than 10.000 observations from the Graduation program, all estimates have passed the Gelman R^2 test for resemblance of different chains' posterior distribution and can be considered as valid. Estimates are constructed in z-score values of the treatment effect in order to make possible comparison between variables. The scale of the treatment effect (TE) is thus in standard deviation (SD) from the Control Group (CG). I use the non-information priors presented for the Partial Pooling BHM. The term Treatment Effect is a language misuse. It would be more appropriate to describe those results as intention-to-treat effects of the program. Indeed, as discussed in the data section, the compliance rate of participants in India was problematic.

3.1 Full Pooling Bayesian Hierarchical Modeling

I first compare the outcomes of a frequentist analysis close to Banerjee et al. (2015) and those that I find using BHM with Full Pooling (FP). On the Figure 1a below, I present the mean treatment effect for both methodologies, with their 50% and 95% confidence or credible intervals for the third survey. Similar graphics for the first end line (Figure 1b) as well as for all household and individual variables (Figure 1c and Figure 1d) are available in the annex.

Figure 1a

GTE: OLS and FP-BHM, thematic variables, fup



First of all, frequentist estimates are not the same as the Banerjee et al. (2015) ones. Indeed, due to the hierarchical framework, I have to drop all variables that are not collected in all sites whereas the original study does not have such preoccupation. Therefore, the thematic variables presented here are not exactly constructed as in the original study in order to allow comparability with the BHM estimates.

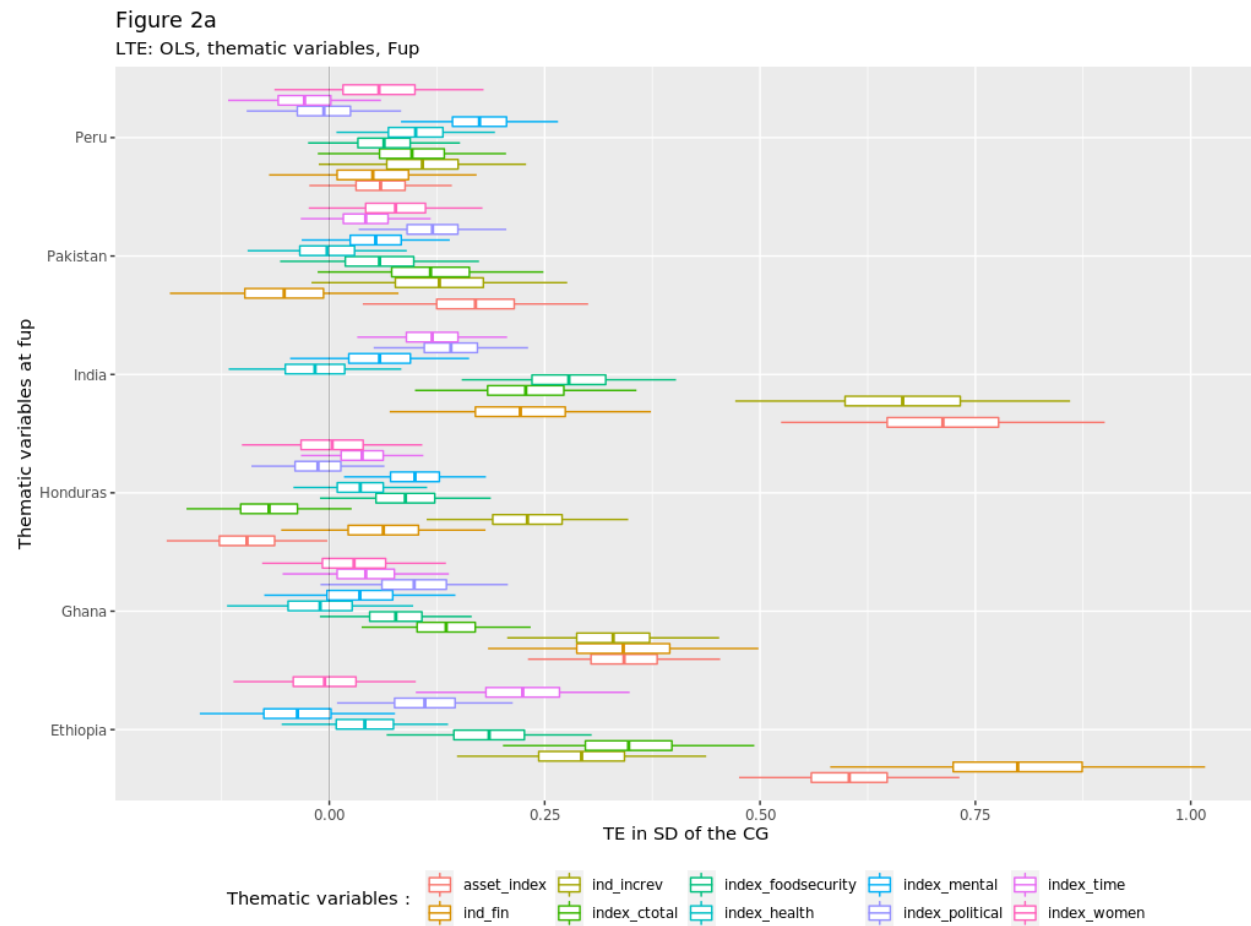
The Full Pooling Bayesian Hierarchical Modeling sets the global variance parameter at zero. Therefore, all $\widehat{\tau}_k$ are used to estimate each country TE. The global treatment effect of the intervention is nothing more than the mean of each identical site-specific treatment effect. Indeed, the prior on the second level of hierarchy stipulates that τ_k are normally distributed around τ and have no inter-site variance in the setting of the full pooling BHM.

The difference in treatment effect between frequentist and FP-BHM estimates are not due to Bayesian or frequentist approaches – which would have nearly identical outcomes in similar settings – but to the modeling hierarchical features and TE heterogeneity across countries, even though not yet formally considered.

Firstly, neither method weights sites in the same way in the GTE estimation procedure. On the one hand, the frequentist methodology, as in Banerjee et al. (2015), makes an OLS regression using all observations from all sites and infers a global cross-site treatment effect. On the other hand, the

Full Pooling BHM also use all information available but to determine each site treatment effect. Afterward, it uses those t_k to infer the global treatment effect across sites. Due to the different number of observations between sites, the hierarchical feature makes the treatment effect estimates different than the simple OLS regression. Indeed, a simple frequentist analysis does not weight the global effect in proportion to each site number of observations. Thus, the global estimates are constructed in way that overweight site effect which has more observations. The table 2 in annex presents sites weighting for both methods and the OLS local treatment effect of the asset index variable as example.

The evaluation samples in Ethiopia and India were much smaller than ones in other countries. Here below, the Figure 2a presents the site-specific (τ_k) treatment effect by thematic index using OLS regression between the observation and a treatment status dummy. We can observe that τ_k for ind_fin, asset_index and ind_increv are especially high for small samples. The global effect is thus reduced for those variables when using simple pooled regression without weighting. Similar graphics, Figure 2b, with estimates a year after the end of the program is in the annex.



Pooled regression without sample size weighting is standard practice in RCT meta-analysis for the authors of the original study. However, a sample size correction is useful when considering heterogeneity of treatment effects across sites and examining external validity of the TE. Indeed, a feature where the global effect is equally driven by all sites is more adequate.

Secondly, FP-BHM estimates also differs from simple pooled OLS due to the structure of the unexplained residual variance across sites (Meager, 2019). The unexplained residual variance is the total variability across group minus the variability due to the differentiated treatment effect (σ_t^2). The FP-BHM imposes that σ_t^2 is zero (as in pooled OLS) but $\widehat{se}_k^2$ has to be identical across sites to match result of the pooled OLS. In Figure 2a, we can see that confidence interval of the same variable across sites (for example `asset_index`) does not have the same size. Local treatment effect CIs reflects the variability of the treatment effect within each group. Variation of CI size across countries are especially important for `ind_fin`, `asset_index` and `ind_increv` variables. On the one hand, Pooled OLS supposes 1° zero heterogeneity in treatment effect across and 2° identical variance of the treatment effect across sites. On the other hand, FP-BHM only imposes 1° but respects the differentiated dispersion of treatment effect within sites.

The fact that the OLS uses individual level data whereas Rubin BHM FP only uses site-level aggregates ($\widehat{\tau}_k, \widehat{se}_k^2$) could be seen as a simplification that could lead to a different outcome. However, I test the Full Pooling setting of the Full Bayesian Hierarchical Modeling (FBHM), which uses individual level data, and leads to very similar results to the ones of Rubin FP-BHM. Basic ideas behind FBHM are introduced in the last part of this paper. FP-BHM and pooled OLS estimation are two possible methods for estimating GTE without formally taking heterogeneity in TE across sites into account. Indeed, in the case of identical intra-site sampling variation ($\widehat{se}_k^2$) in all sites and groups weighted equally, both procedures would lead to identical results. The hierarchical framework is more relevant for the discussion of external validity.

3.2 Partial Pooling Bayesian Hierarchical Modeling

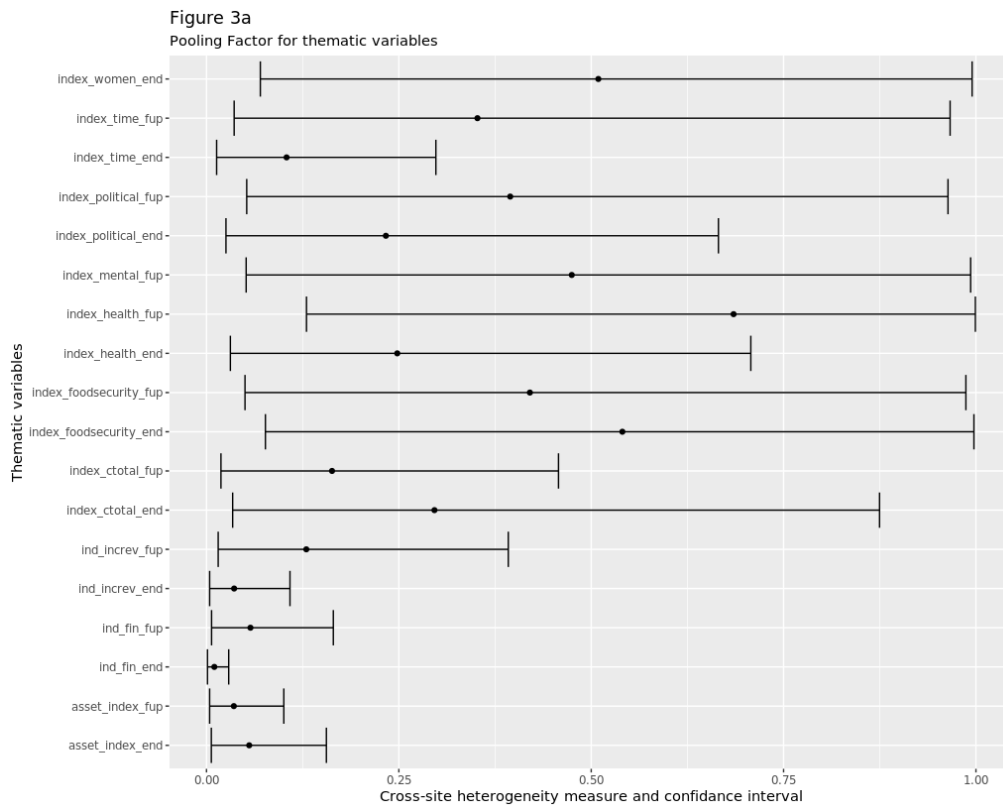
After presenting the rationale behind the Full Pooling BHM and some results, I can introduce the Partial-Pooling Bayesian Hierarchical Modeling (PP-BHM). As already developed, it allows to consider the heterogeneity in the treatment effect across sites and uses it in its estimations of local and global treatment effects.

In order to determine how much pooling is needed, the PP-BHM evaluates, simultaneously to the GTE, the variation between different sites due to treatment effect. If a lot of cross-site heterogeneity is detected, the pooling can only be limited. If treatment effects across countries are similar, some information can be shared between sites to get estimates more precise. To integrate heterogeneity in TE, the PP-BHM needs a pooling metric somewhere 0 (no pooling, i.e. maximum heterogeneity in TE) and 1 (full pooling, i.e. no heterogeneity in TE). The Pooling Factor, as define by equation 3, is 1 minus the proportion of cross-site variability due to treatment effect (σ_t^2) over total variability ($\sigma_t^2 + E(se_k^2)$). Descriptive statistics show the total variability at each site but without additional modeling I cannot differentiate idiosyncratic variation ($E(se_k^2)$) from treatment effect variability across site (σ_t^2).

BHM simultaneously estimates hyper-parameters (τ and σ_t^2). This σ_t^2 found allows to construct the Pooling factor to estimate all local treatment effects using the relation of the equation 2. Hyper-parameter σ_t^2 is the more probable value, out of 40 000 simulated trials, given the data collected and the set of priors introduced in the BHM. It is found using the Hamiltonian Monte Carlo simulation algorithm based on site-level information and restricted with priors on: 1° within-site

treatment effect distribution (eq. 1), 2° across sites treatment effect distribution (eq. 1), 3° hyper-mean distribution (eq. 4) and 4° hyper-standard-deviations distribution (eq. 4).

The Figure 3a represents the Pooling Factor with its 95% credible interval for thematic variables at both end lines. Figures 3b and 3c, in annexe present similar graphs for household and individual variables. The higher is the TE heterogeneity across sites, the lower the pooling factor will be and thus closer is the PP-BHM LTE estimate to the NP-BHM one. All means of the pooling metrics, except three, are below 0.5 which indicates that variance across studies dominates variation within studies. The 95% credible intervals are extremely large for variables with higher pooling factor means. This last feature is due to the low number of sites. Indeed, it is difficult to estimate the TE variation across sites when those variations are quite low, relative to total variation, and only dispose of a few sites in the analysis. On the contrary, *ind_fin* and *asset_index* presents important disparities, relative to the total variation, across places and thus can be estimated with more precision.



CI's are very wide for most of the variables. This is one of the main objections against BHM: the need to estimate the pooling factor from only a few points. Indeed, the estimation is based on the variation of the TE across only six studies and, thus, cannot be precise. However, the only alternative to estimate the GTE is to suppose this parameter at zero as in a cross-site pooled regression.

Table n°3 presents the pooling factors' mean across indexes for thematic, household and individual variables. It can be interpreted as the global treatment effect heterogeneity of the program across countries. Overall, the Partial Pooling BHM is essentially based on site-specific data for the LTE

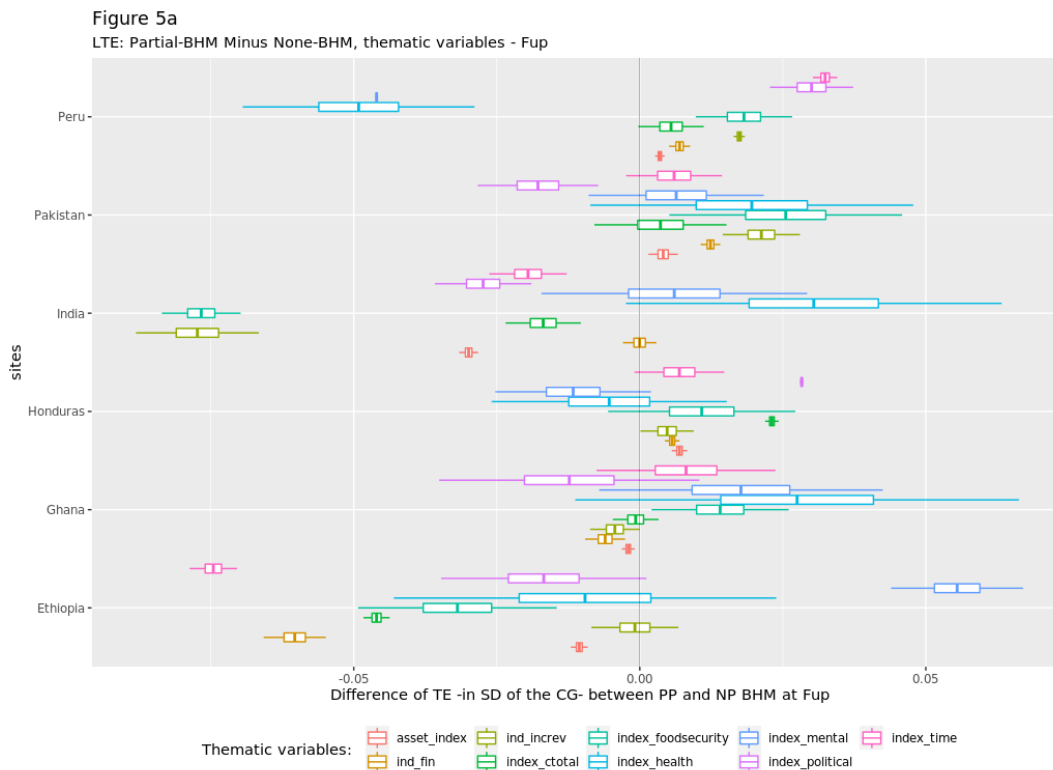
estimation rather than pooling sites information together. Thus, GTE estimates will have higher credible intervals.

Table 3: Pooling metric across variables

	Lower CI	Mean	Higher CI
Thematic level	3,48%	26,36%	60,31%
Household level	4,39%	31,41%	70,30%
Individual level	7,09%	43,66%	80,52%

3.3 Local Treatment Effects

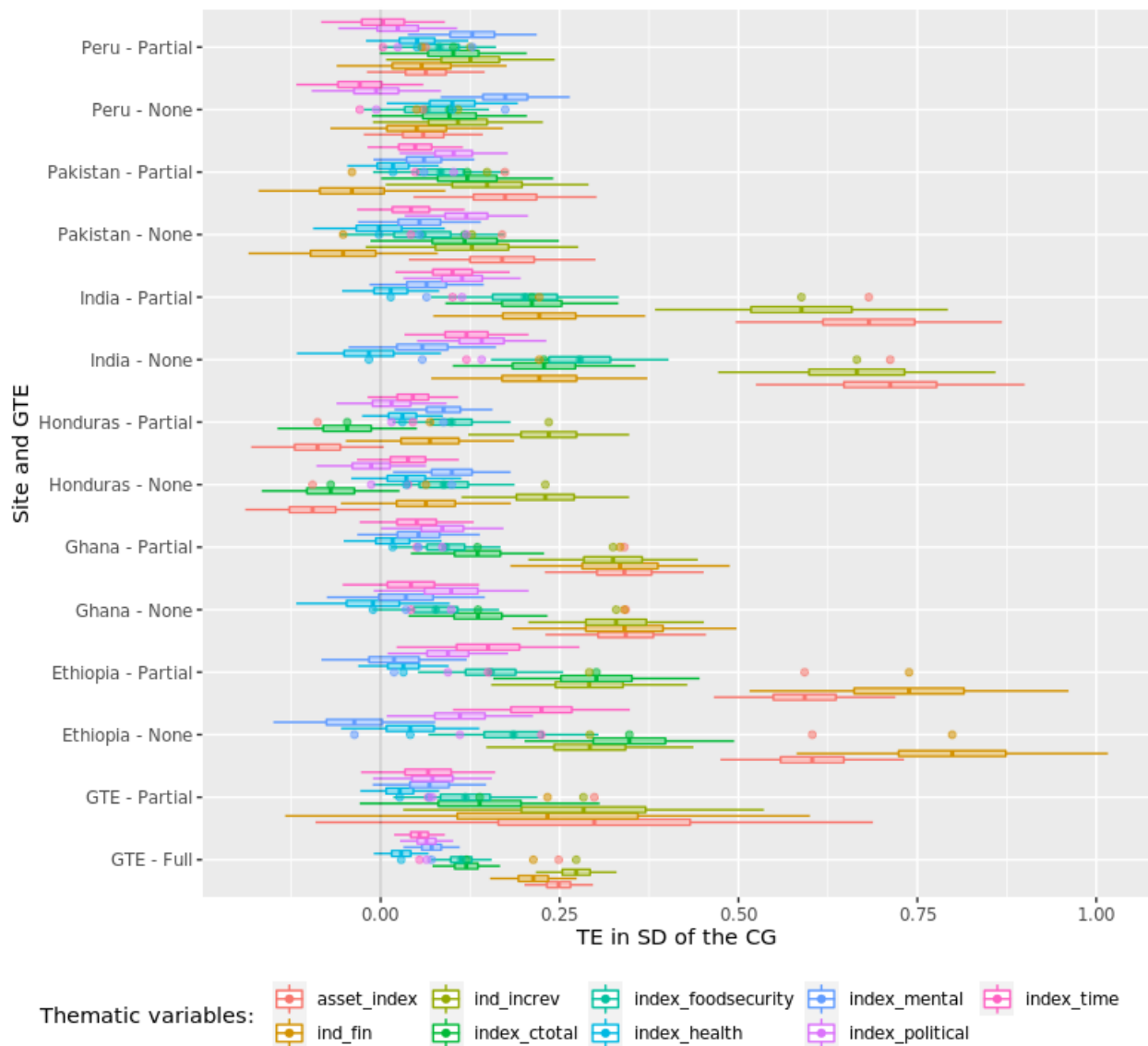
When looking at LTE estimates of PP-BHM and NP-BHM at the Figure 4a to Figure 4c in annex, we can observe that the LTE estimates of those two modeling are close one to each other. The PP-BHM uses the inter-site heterogeneity measure as parameters of how much it could use information about the K-1 sites to help the inference in the site k. Since heterogeneity is quite important between the treatment effect of each site, the PP-BHM does not pool much information thus LTE estimates are close to the NP-BHM. As explained in the methodology section, the NP-BHM considers that the inter-site variability tends toward infinity. In other words, contexts and implementation characteristics differ so much that no information can be pooled. Therefore, LTE are estimated only using site observations. To see the effect of PP-BHM on the LTE, the next graph presents the difference between its estimates (and credible intervals) and those from the NP-BHM for thematic variables at the second end line (the first one is in the Figure 5b in annex).



The rationale behind the correction that the PP-BHM operated is quite simple. For example, the previous graphic presents an important correction for the `index_time` LTE estimate of Ethiopia in the third survey with -0.075 standard deviations from the control group in comparison with NP-BHM estimate. The Bayesian simulation process makes two estimation simultaneously. It evaluates, using its hierarchical feature, that `index_time` has a relatively low heterogeneity across sites. And, at the same time, uses this measure of variability to pool, to some extent, the observation of this variable to estimate the LTE. For `index_time` at fup, observations in Ethiopia present important TE whereas other all other sites show much lower TE with low variability between them (see Figure 6a. 12: `index_time_fup`). The PP-BHM estimate of LTE for this variable is thus reduced for Ethiopia where it is increased for Peru in comparison with NP-BHM. The scale of the correction is proportional to the observed heterogeneity across sites.

Figure 7a

GTE and LTE, PP-BHM vs None and Full BHM, thematic, Fup



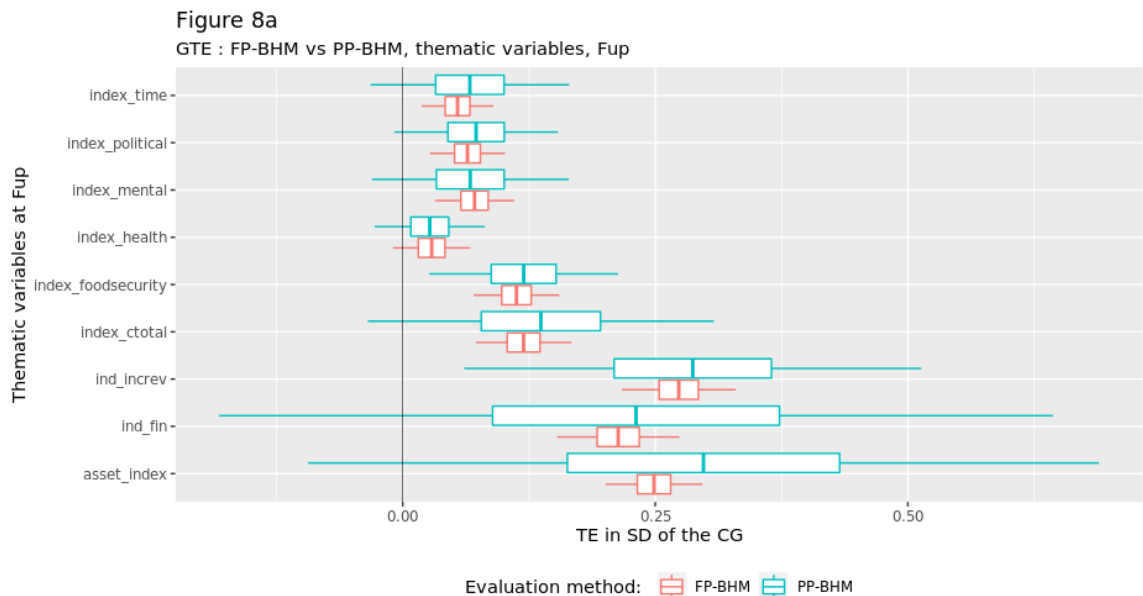
The NP-BHM estimates are thus better for analyzing the LTE when considering that the contexts between sites are too different to be informational one to each other -which is always the case to some extent-. Therefore, NP-BHM cannot produce GTE estimates. PP-BHM can estimate global treatment effect as well as the FP-BHM. The former tries to take the observed cross-site heterogeneity into account in its LTE and GTE estimates whereas the latter does not take it into account at all.

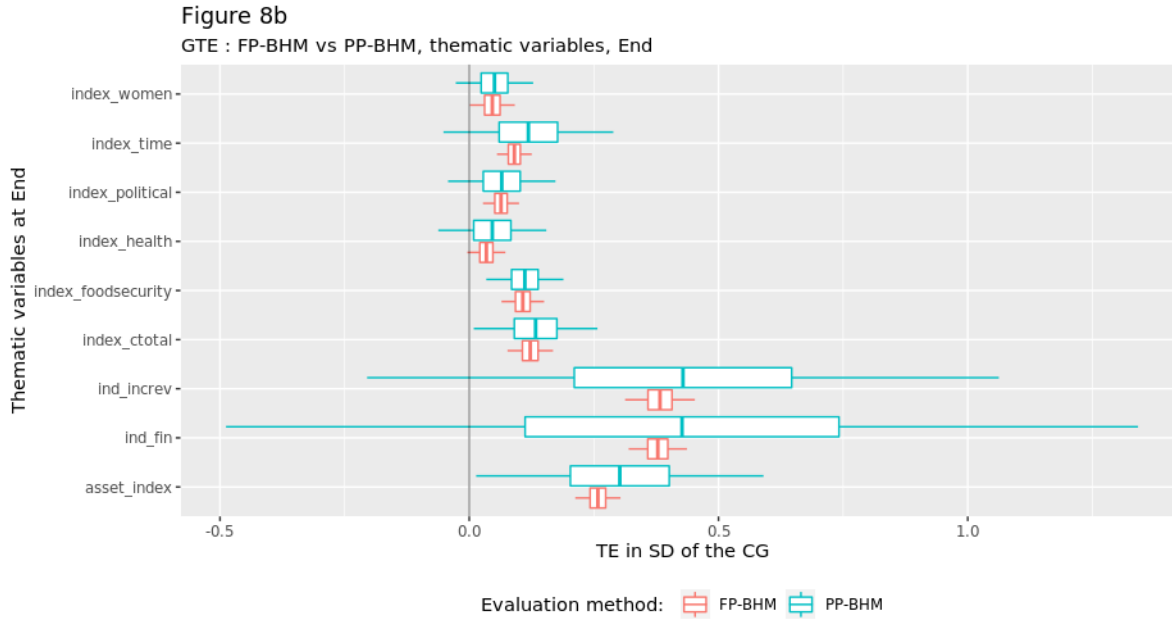
The Figure 7a above presents PP estimates for all thematic variables in comparison with NP for Local Treatment Effect and with FP for Global Treatment Effect (a similar graph for the first end line is in annex Figure 7b). Figure 7a well represents the process of meta-analysis: trying to find trends from a multitude of different categories - in terms of hierarchical levels, kind of variables and type of modeling - of information. For a variable-by-variable estimation of GTE and LTE using NP, PP and FP see Figures 6a for thematic variables and Figures 6b for household and individual level variables.

The two main trends developed can be easily seen in this figure. First, LTE estimations of PP and NP are very close ones to each other's due to low PP correction of estimates. Second, GTE from the PP present much wider credible intervals than those of the FP estimation procedure. The former point is discussed in the next sub-section.

3.4 Global Treatment Effects

The Figures 8a and 8b present the GTE of thematic variable TE for the PP-BHM as well as FP-BHM at both end lines respectively. Figures 8c and 8d in annex present similar graphics for household and individual level variables.





By taking some measure of LTE heterogeneity across sites into account and use it in its estimates, the PP-BHM finds slightly higher means of GTE across thematic indexes. But the interesting part is that the credible intervals increased for all variables and exploded for variables with high cross-site heterogeneity (mainly `ind_increv`, `ind_fin` and `asset_index`). This result was expected and shows the mean feature of the PP-BHM: by considering heterogeneity in the LTE across sites, it increases the uncertainty around the global treatment effect mean in comparison with FP.

At Global Treatment Effect level, FP-BHM estimates will have smaller credible intervals than PP. Indeed, as the pooling factor increases, more information is shared between studies which increases the preciseness of the GTE estimates. In addition, as the pooling factor increases, more the Local Treatment Estimates CI of PP BHM will be reduced in comparison with None Pooling BHM.

In our case, there is too much heterogeneity in treatment effect across countries in proportion to total variability across sites to pool much information for most of the variables. Thus, CI will necessarily increase in comparison with the extreme case where we consider each site equally informative to another site treatment effect (i.e. pooling = 1). The interesting part is the size of the increase.

This feature has important consequences. Using the PP estimations, the GTE of the Graduation program is not statistically significant different to zero (SSDZ), at 95%, for 7 of the 9 thematic variables two years after the end of the program. Whereas only the GTE of `index_health` increased is not considered as statistically significant in the FP approach. In the shorter term, a year after the end of the intervention, the Figure 8b presents a similar pattern but Partial Pooling TE estimates of `index_foodsecurity`, `index_ctotal` and `asset_index` are SSDZ at 95% whereas all are statistically significant using FP. In the annex, Figures 8c and 8d are similar graphs for all individual and household, one and two years after the end of the program, respectively.

For variables with relatively low approximated heterogeneity of LTE (`index_time`, `index_political`, `index_health`, `index_foodsecurity` and `index_mental`), the means GTE estimated are close to the simple pooled OLS regression estimates. However, their 95% credible intervals still double in

comparison. Across those, only `index_foodsecurity` remains SSDZ, for the third survey, at 95%. In the context of the BRAC Graduation program those variables can be seen as capturing the indirect effect of the intervention. Indeed, the multifaced program does not provide directly food, medicine or psychological therapy but rather emphasizes on the livelihood approach by delivering capital transfers, training, consumption support, saving encouragements, etc. Effects of those interventions can be better captured by the other set of variables. Indeed, `ind_increv`, `ind_fin`, `asset_index`, and to some extent, `index_ctotal` display more important estimated GTE means using PP-BHM. However, those estimates also present massive 95% CI which includes zero TE for many. Among those, a year after the end of the program, only TE for `index_ctotal` as well as `asset_index` are SSDZ at 95%. A years later, only the GTE increased of `ind_increv` is statistically significant at this level whereas the effect on `asset_index` and `index_ctotal` are not.

In a nutshell, even if GTE means of thematic variables are similar in PP and FP, statistically informative conclusions widely differ. The Pooling factor of the PP modeling is only a rough approximation of the cross-site heterogeneity due to the low number of sites in the analysis. However, is it more erroneous to estimate it with only six different group or to assume it is zero? To answer this crucial question, I conduct cross predictive checks using Leave-One-Out Cross Validation (LOOCV) technic for PP and FP modelings.

This procedure allows to have some idea of the extrapolating power or external validity of our global estimates. Indeed, LOOCV runs 6 times the model, leaving one site excluded at the time and uses the six Expected Log Predictive Density (ELDP) to check how well it matches with the actual treatment effect of the excluded group. As developed in the methodology part of the paper, the Leave-One-Out Information Criteria (LOOIC) is -2 times the average ELPD across the 6 models run. The closer the measure to zero, the better the model. The next table presents the LOOIC of thematic variables for both models.

Table 4 : LOOIC comparison PP to FP

	End		Fup	
	PP-LOOIC	FP-LOOIC	PP-LOOIC	FP-LOOIC
<code>index_ctotal</code>	-4,10	0,10	3,06	19,18
<code>index_foodsecurity</code>	-10,09	-11,52	-6,41	-7,96
<code>asset_index</code>	5,47	79,85	8,44	155,53
<code>ind_fin</code>	119,52	320,66	14,67	50,95
<code>index_time</code>	-0,85	24,23	-6,72	-7,84
<code>ind_increv</code>	26,35	49,63	8,18	16,68
<code>index_health</code>	-3,18	3,20	-16,31	-18,31
<code>index_political</code>	-6,32	3,70	-9,75	-6,48
<code>index_women</code>	-9,27	-10,31	/	
<code>index_mental</code>	/		-9,39	-10,40

The LOOIC points out that PP-BHM has, overall, a better predictive capacity. For three thematic variables (index_ctotal_end, index_political_end and index_political_fup), the full pooling modeling over-performs its partial counterpart by a relatively small amount. For four other variables (ind_fin_end, ind_increv_end, asset_index_fup and ind_fin_fup), the PP-BHM has drastically better estimates. For the other variables, the partial pooling modeling is more accurate. It is also important to highlight that LOOIC's measures of both models are quite substantial which indicates a limited external validity. In other words, both models still face difficulties to properly predict the LTE distribution density in a site using only other sites' data.

Those results can be easily linked to CI of the pooling metric in Figure 3a. When the heterogeneity across sites is estimated with preciseness (i.e., have small CI), Partial Pooling BHM drastically outperforms Full Pooling BHM. Indeed, as long as the Pooling factor is correctly estimated, PP estimations will always be better than FP. The next table presents means and sums, for thematic variables as well as household and individual, of the LOOIC measure across variables.

Table 5: LOOIC across variables comparison PP and FP				
	Average		Sum	
	PP - LOOIC	FP - LOOIC	PP - LOOIC	FP - LOOIC
Thematic	5,74	36,16	103,30	650,89
Household	-0,82	26,16	-23,92	758,65
Individual	-7,75	8,81	-170,46	193,80

The sums of the prediction errors across thematic indexes are more than six times higher for the FP-BHM. This result is driven by the household level variable with a measure of 758 whereas the PP-BHM presents a value of only -23.9. Individual-level variables LTE are more difficult to predict for both models. On average, the PP-BHM outperforms the FP-BHM across indexes and particularly for household and thematic level variables. The LOOIC tables 4b and 4c, for household and individual level variable, are in the annex.

Based on that information, we can conclude that the heterogeneity across sites TE captured by the PP-BHM plays an important role. The LOOCV procedure examines how well the GTE generated by five of the six sites can correctly predict LTE of the omitted site. The heterogeneity in treatment effect across sites is therefore an important issue when extrapolating about a common global treatment effect. Differences between PP and FP LOOIC' measures reflect the scale of the extrapolation error that can be reduced by Partial Pooling Bayesian Hierarchical Modeling.

4. Discussion

The estimates of GTE using Partial Pooling Bayesian Hierarchical framework highlight the importance of taking the inter-site variability of the program treatment effect into account. The credible intervals are much wider than the confidence intervals of the pooled OLS estimates for all variables. Only two TE estimates remain statistically significant different than zero two years after the end of the intervention. Mean effects are, on the contrary, slightly higher than those found using the classic aggregation technic. However, the estimation of the cross-site heterogeneity is

only a rough approximation of the true inter-study TE variability. Indeed, the LOOIC measure demonstrates that some variables are better estimated using FP-BHM. Overall PP-BHM has more accurate estimates of the aggregate program effects across variables.

Context does matter in the effectiveness of the intervention of some variables and by a significant amount (e.g., see the huge increases of CI from FP to PP in Figure 8a and 8b). The heterogeneity measures estimated are higher for inter-site than for within-site variability levels in almost all variables. It indicates that context information is at least as important as the individual level information to explain the difference of effectiveness of the intervention.

By integrating site-level information into the BHM in the form of covariables, it will improve the understanding of the variability of the GTE of the intervention. For example, we could take into account the value of the asset transfer for each site in our BHM, which varies from 3.75 for Pakistan to 17.14 for Peru, both in terms of local goat prices. By assimilating this context information into the PP-BHM, we could have a better insight of the importance of the asset transfer scale in the treatment effect of the intervention. Using similar models to Meager (2019), M covariates (such as asset transfer value, number of visits or consumption support value), regrouped in a vector W_k , will have predictive power over the variation of TE. β is the M-dimensional vector of coefficients which predicts the variation in LTE for each site. Such model would have the following structure:

$$\begin{aligned}\hat{\tau}_k &\sim N(\tau_k, \widehat{se}_k^2) & \forall k \\ \tau_k &\sim N((\tau + W_k\beta_k), \sigma_\tau) & \forall k\end{aligned}\tag{eq. 6}$$

Bayesian Hierarchical Modeling frameworks are also perfectly fit to take into account additional information at observational level. Whereas the Rubin framework is only Bayesian at the second level of hierarchy (i.e. the simulation procedure is performed using treatment effect at site level), the ‘Full’ Bayesian model can integrate individual-level covariates. For an implementation of such model see Meager (2019) who uses it to explain the heterogeneity in TE of a microcredit access extension by subgrouping observations with household-level information.

By understanding the reason behind heterogeneity in TE, such an analysis can drastically reduce the credible interval around the GTE mean conditional of contextual or observation level covariates. Thus, the external validity could be ameliorated. By pointing out some important characteristics that drive the TE above or below its mean effect on studied site, predictions of treatment effect in another location can be made with more confidence only by analyzing those features.

I have highlighted the importance of taking into account the inter-site heterogeneity in the estimation of Global Treatment Effect means and credible intervals of the BRAC Graduation pilot project. However, much more can be done using these experiments to understand what drives the success of the Graduation approach and what does not. Bayesian Hierarchical Modeling with integration of contextual and household/individual characteristics is a key framework to address this issue. A similar analysis, by Andrew Gelman, Dean Karlan, Chris Udry, Racheal Meager and Witold Wiecek, is in progress. I have no doubt about the potential of their findings and look forward to reading their future study: “*A Multifaceted Approach to Poverty Alleviation in Six Countries: A Bayesian Hierarchical Analysis of the Graduation Program*”.

5. Conclusion

My analysis contrast with the original estimates of Banerjee et al. (2015). These authors present six site-specific analyses and a pooled analysis. The goal of the analyzed pilots is to understand how replicable the BRAC Graduation program success is across contexts. If such program were to be implemented somewhere else, would it succeed, and if it does succeed, to what extent? Since the context, the program, the implementation, the people involved, etc, widely differ from one place to another, the treatment effect has to be heterogeneous across sites. Therefore, site-specific TE have limited external validity and the pooled OLS is computed using a misleading hypothesis of zero TE heterogeneity across sites. There is no clear answer to the problematic of which OLS estimates (site level or pooled) should inform policies about the Graduation treatment effect in another unstudied site. The PP-BHM allows to mixt those two approaches by estimating the variability of treatment effect across sites and using this estimation to improve inference on LTE and GTE.

Thus, the PP-BHM is a compromise between the two approaches used by Banerjee et al. (2015). The approximation of the TE heterogeneity is particularly difficult with only six groups and estimates should be interpreted with caution. However, this approximation allows a middle ground between hypothesis of null and infinite heterogeneity in TE across sites. In addition, the PP-BHM can provide results that are identical to OLS estimation of LTE and GTE, depending on the estimation of the TE heterogeneity across contexts. The data itself determines the choice of pooling. In other words, the data determines how much external validity there is.

A joint estimation of the average impact and heterogeneity across the six pilots BRAC Graduation program suggests that statistically significant impacts are not as certain as estimated by Banerjee et al. (2015). Across the main variables, only approximately 26 percent of observed variability is in fact sampling variation. Overall, the analysis of the intervention has limited external validity. The TE heterogeneity across contexts is stronger for the aspects that are the most impacted by the program.

The methodology also has some drawbacks. Indeed, it is particularly difficult to implement in comparison with OLS estimations. The pooling factor is only a rough approximation of the real treatment effect variability, in particular with only six sites. In addition, it cannot incorporate into the analysis the variables that are not collected in all sites. However, the benefits of the model exceed the limitations. First, a hierarchical structure respects the structure of the data. Second, only BHM can simultaneously estimate both the mean effect across sites and its probable variability. Thus, it allows to partially pool site information together and it provides more precise estimates of LTE and realistic estimations of GTE.

By highlighting the TE heterogeneity across sites, BHM estimates reveal the uncertainty around the probable effect of the program in another location. This uncertainty is important. However, BHM allows the integration of covariates which can potentially reduce the uncertainty of treatment effect conditional on those characteristics introduced. Indeed, much more can be done within BHM framework by using contextual and individual level information to further improve inference and external validity.

References

- Ahmed, A. U., Rabbani, M., Sulaiman, M., & Das, N. C. (2009). The impact of asset transfer on livelihoods of the ultra poor in Bangladesh. *BRAC Res. Monogr. Ser.*, 39, 339-380.
- Balboni, C., Bandiera, O., Burgess, R., Ghatak, M., & Heil, A. (2020). *Why Do People Stay Poor? Evidence on Poverty Traps from Rural Bangladesh*. LSE Working Paper, London School of Economics.
- Bandiera, O., Burgess, R., Das, N., Gulesci, S., Rasul, I., & Sulaiman, M. (2017). Labor markets and poverty in village economies. *The Quarterly Journal of Economics*, 132(2), 811-870.
- Bandiera, O., Burgess, R., Das, N., Gulesci, S., Rasul, I., & Sulaiman, M. (2013). Can basic entrepreneurship transform the economic lives of the poor?.
- Bandiera, O., Fischer, G., Prat, A., & Ytsma, E. (2016). Do women respond less to performance pay? Building evidence from multiple experiments.
- Banerjee, A. V., & Duflo, E. (2009). L'approche expérimentale en économie du développement. *Revue d'économie politique*, 119(5), 691-726.
- Banerjee, A. V., & Newman, A. F. (1993). Occupational choice and the process of development. *Journal of political economy*, 101(2), 274-298.
- Banerjee, A., Duflo, E., & Sharma, G. (2020). *Long-term effects of the targeting the ultra poor program* (No. w28074). National Bureau of Economic Research.
- Banerjee, A., Duflo, E., Chattopadhyay, R., & Shapiro, J. (2016). The long term impacts of a “Graduation” program: Evidence from West Bengal. *Unpublished paper, Massachusetts Institute of Technology, Cambridge, MA*.
- Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (2015). The miracle of microfinance? Evidence from a randomized evaluation. *American economic journal: Applied economics*, 7(1), 22-53.
- Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., ... & Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348(6236).
- Bauchet, J., Morduch, J., & Ravi, S. (2015). Failure vs. displacement: Why an innovative anti-poverty program showed no net impact in South India. *Journal of Development Economics*, 116, 1-16.
- Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, 1(2), 97-111.

- Broderick, T., Giordano, R., & Meager, R. (2020). An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?. *arXiv preprint arXiv:2011.14999*.
- Burke, M., Hsiang, S. M., & Miguel, E. (2015). Climate and conflict. *Annu. Rev. Econ.*, 7(1), 577-617.
- Cartwright, N., & Munro, E. (2010). The limitations of randomized controlled trials in predicting effectiveness. *Journal of evaluation in clinical practice*, 16(2), 260-266.
- Coates, L., & MacMillan, S. (2019). Breaking Out of the Poverty Trap. *Leave No One Behind: Time for Specifics on the Sustainable Development Goals*, 8.
- Das, N. C., & Misha, F. A. (2010). Addressing extreme poverty in a sustainable manner: Evidence from CFPR programme. *BRAC CFPR Work. Pap*, 19.
- Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of economic literature*, 48(2), 424-55.
- Dehejia, R. (2013). *The porous dialectic: Experimental and non-experimental methods in development economics* (No. 2013/011). WIDER Working Paper.
- Dehejia, R. (2003). Was there a Riverside miracle? A hierarchical framework for evaluating programs with grouped data. *Journal of Business & Economic Statistics*, 21(1), 1-11.
- Dehejia, R. (2005). Program evaluation as a decision problem. *Journal of Econometrics*, 125(1-2), 141-173.
- Dehejia, R., Pop-Eleches, C., & Samii, C. (2021). From local to global: External validity in a fertility natural experiment. *Journal of Business & Economic Statistics*, 39(1), 217-243.
- DellaVigna, S., Pope, D., & Vivalt, E. (2019). Predict science to improve science. *Science*, 366(6464), 428-429.
- Devereux, S., & Sabates-Wheeler, R. (2015). Graduating from social protection? Editorial introduction. *ids Bulletin*, 46(2), 1-12.
- Dharmadasa, H., Orton, I., & Whitehead, L. (2016). *Mainstreaming graduation into Social Protection floors* (No. 324).
- Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4, 3895-3962.
- Findley, M. G., Kikuta, K., & Denly, M. (2021). External Validity. *Annual Review of Political Science*, 24, 365-393.
- Gechter, M. (2015). Generalizing the results from social experiments: Theory and evidence from Mexico and India. *manuscript, Pennsylvania State University*.
- Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641-651.
- Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48(2), 241-251.

- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6), 997-1016.
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6), 997-1016.
- Hartman, E., Grieve, R., Ramsahai, R., & Sekhon, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 757-778.
- Hashemi, S., Cawood, S., Razzak, N., Rashid, M. U., & Roy, M. (2016). Can asset transfer promote adaptation amongst the extreme urban poor?: Lessons from the DSK-Shiree programme in Dhaka, Bangladesh. In *Urban Poverty and Climate Change* (pp. 202-220). Routledge.
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1), 1593-1623.
- Hsiang, S. M., Burke, M., & Miguel, E. (2013). Quantifying the influence of climate on human conflict. *Science*, 341(6151).
- Hulme, D., & Moore, K. (2008). Assisting the poorest in bangladesh: Learning from BRAC's 'targeting the ultra-poor' programme. In *Social protection for the poor and poorest* (pp. 194-210). Palgrave Macmillan, London.
- Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1), 443-470.
- Islam, A. (2019). Ultra-poor graduation to eradicate extreme poverty in Bangladesh: a case of BRAC.
- Kharas, H., McArthur, J. W., & Ohno, I. (2019). *Leave no one behind: Time for specifics on the Sustainable Development Goals*. Brookings Institution Press.
- Kerman, J., & Gelman, A. (2006). Bayesian data analysis using R. *R News*, 6(1), 21-24.
- Kowalski, A. E. (2016). *Doing more when you're running late: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments* (No. w22363). National Bureau of Economic Research.
- MacLennan, M., Soares, F. V., & Orton, I. (2017). *Debating graduation* (No. 39). Policy in Focus.
- Mallick, D. (2009). Microfinance and moneylender interest rate: evidence from Bangladesh. Research Monograph Series No. 42. *Research and Evaluation Division, BRAC*, 75.
- Martin, I., Rabbani, M., & Sulaiman, M. (2008). Crafting a graduation pathway for the ultra poor: lessons and evidence from a BRAC programme.
- McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(5), 750-773.

- Meager, R. (2016). Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature. *Manuscript: MIT*.
- Meager, R. (2019). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1), 57-91.
- Mukti, R. F., Samadder, P. D., Emran, A. A., Ahmed, F., Imran, I. B., Malaker, A., & Yeasmin, S. (2014). Score based risk assessment of lung cancer and its evaluation for Bangladeshi people. *Asian Pacific Journal of Cancer Prevention*, 15(17), 7021-7027.
- Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*, 32(1), 107-121.
- Orr, L. L., Olsen, R. B., Bell, S. H., Schmid, I., Shivji, A., & Stuart, E. A. (2019). Using the results from rigorous multisite evaluations to inform local policy decisions. *Journal of Policy Analysis and Management*, 38(4), 978-1003.
- Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, 59-73.
- Peters, J., Langbein, J., & Roberts, G. (2016). Policy evaluation, randomized controlled trials, and external validity—A systematic review. *Economics Letters*, 147, 51-54.
- Peters, J., Langbein, J., & Roberts, G. (2018). Generalization in the tropics—development policy, randomized controlled trials, and external validity. *The World Bank Research Observer*, 33(1), 34-64.
- Pritchett, L., & Sandefur, J. (2014). Context matters for size: why external validity claims and development practice do not mix. *Journal of Globalization and Development*, 4(2), 161-197.
- Raudenbush, S. W., & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, 36(4), 475-499.
- Raza, W. A., Das, N. C., & Misha, F. A. (2012). Can ultra-poverty be sustainably improved? Evidence from BRAC in Bangladesh. *Journal of Development Effectiveness*, 4(2), 257-276.
- Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., ... & Friede, T. (2021). On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Research Synthesis Methods*.
- Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4), 377-401.
- Rubin, D. B. (2013). Bayesian data analysis.
- Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2), 369-386.

- Sulaiman, M., Goldberg, N., Karlan, D., & de Montesquiou, A. (2016). Eliminating extreme poverty: Comparing the cost-effectiveness of livelihood, cash transfer, and graduation approaches. In *Forum. Washington, DC: CGAP. C*.
- Vivalt, E. (2015). Heterogeneous treatment effects in impact evaluation. *American Economic Review*, 105(5), 467-70.
- Vivalt, E. (2020). How much can we generalize from impact evaluations?. *Journal of the European Economic Association*, 18(6), 3045-3089.
- Vivalt, E., & Coville, A. (2019). How Do Policymakers Update?. *Mimeo*.
- World Bank. (2020). *Poverty and shared prosperity 2020: Reversals of fortune*. The World Bank.
- Wiecek (2021), baggr v0.6 R-package
- Zhao, L. (2016). Current trends in Bayesian methodology with applications.

Annexes

Table of content

Name	Description	Page.s
Table 0	Abbreviations	32
Table 1	Thematic variables composition	33
Figure 1	Global Treatment Effect - OLS and Full Pooling Hierarchical Modeling	34-35
Table 2	Observations and weighting across sites	36
Figures 2	Local Treatment Effect – OLS estimates	33
Figures 3	Pooling Factor	37
Figures 4	Local Treatment Effect of thematic variables using FP-BHM or PP-BHM	38-39
Figures 5	Difference between LTE estimates of PP-BHM and FP-BHM	40
Figures 6a	Global and Local Treatment effect pooling comparison for thematic variables	41-46
Figures 6b	Global and Local Treatment effect pooling comparison for household and individual variables	47-63
Figures 7	GTE and LTE estimates - pooling comparison	64
Figures 8	Global Treatment Effect - Full and Partial BHM estimates comparison	65
Tables 4	Leave-one-out information criteria comparison PP to FP	66-67

Table 0 : Abbreviations

BHM	Bayesian Hierarchical Modeling
CG	Control Group
CI	Credible Interval
ELPD	Expected Log Predictive Density
FBHM	Full Bayesian Hierarchical Modeling
FP	Full Pooling
GTE	Global Treatment Effect
HMC	Hamiltonian Monte Carlo
IPA	Innovation for Poverty Action
ITT	Intention-to-treat
LOOCV	Leave-One-Out Cross Validation
LOOIC	Leave-One-Out Information Criterion
LTE	Local Treatment Effect
MCMC	Markov Chain Monte Carlo
NNHM	Normal-Normal Hierarchical Model
NP	None Pooling
OLS	Ordinary Least Squares
PP	Partial Pooling
RCT	Randomized Control Trials
SD	Standard Deviation
SSDZ	Statistically Significant Different to Zero
TE	Treatment Effect

Table 1: Thematic variables composition

Thematic variables	Household and individual variables
Consumption (Index_ctotal): Total consumption per capita per month, standardized	Food purchase (cfood_pcmmonth), non-food purchase (cnonfood_pcmmonth) and durable good purchase (cdurable_pcmmonth), all per capita per month.
Food Security (index_foodsecurity): Everyone gets enough food everyday	No adult skips meal (fs_adultskip), no one went a whole day without food (fs_wholeday), no child skipped meal (fs_childskip) *Miss 1 variable
Assets (asset_index) Total relative value of all assets	Productive assets (asset_prod_index) and of household assets (asset_hh_index)
Financial Inclusion (Ind_fin)	Total amount deposited into savings (loan_totalamt), both formal and informal *Miss 2 variables
Time spent working (index_time) Total time spent working	Time spent on: agriculture (time_agri), livestock (time_animals), business (time_business) and paid work (time_paidlabor)
Income and Revenue (ind_increv)	Revenues from livestock (raminals_month), income from agriculture (iagri_month), income from business (ibusiness_month), income from paid work (ipaidlabor_month) *Miss 1 variable
Physical Health (index_health)	Activities of daily living scores (dailyliving_score) and adult has not missed work due to illness (dday_nomiss). *Miss 1 variable
Mental Health (index_mental)	Index of lack of symptoms of emotional distress (index_stress) and self-perception of life (percep_life) *Miss 1 variable
Political Involvement (index_political)	Voted in last election (pol_vote) *Miss 3 variables
Women's Empowerment (index_women)	Female has the last word on food-related spending decisions (dec_food_major), education expenditures (dec_educ_major) and health-related expenditures (dec_health_major) *Miss 2 variables

Figure 1: Global Treatment Effect – OLS and Full Pooling Hierarchical Modeling

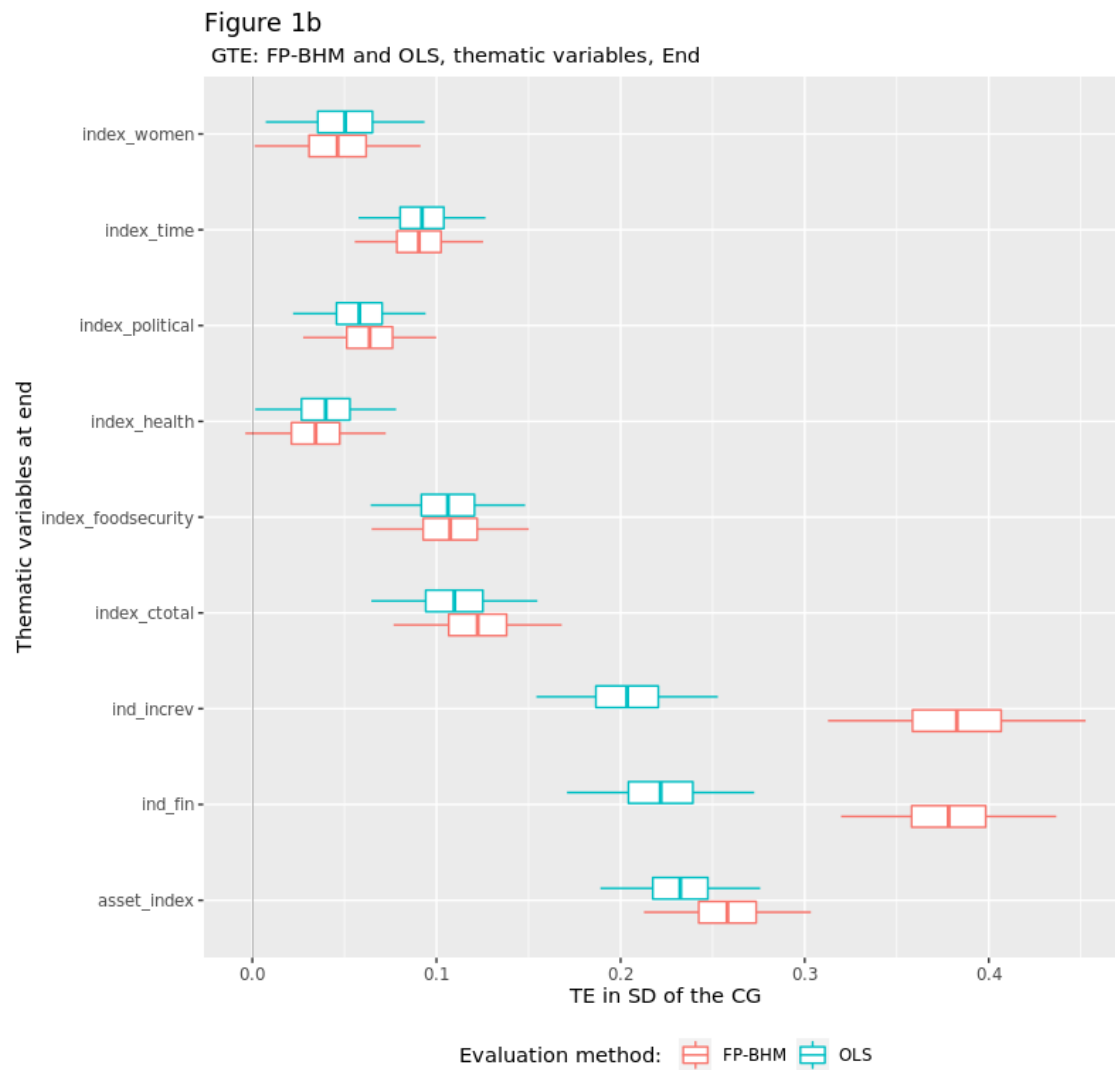


Figure 1c

GTE: OLS and FP-BHM, all variables, Fup

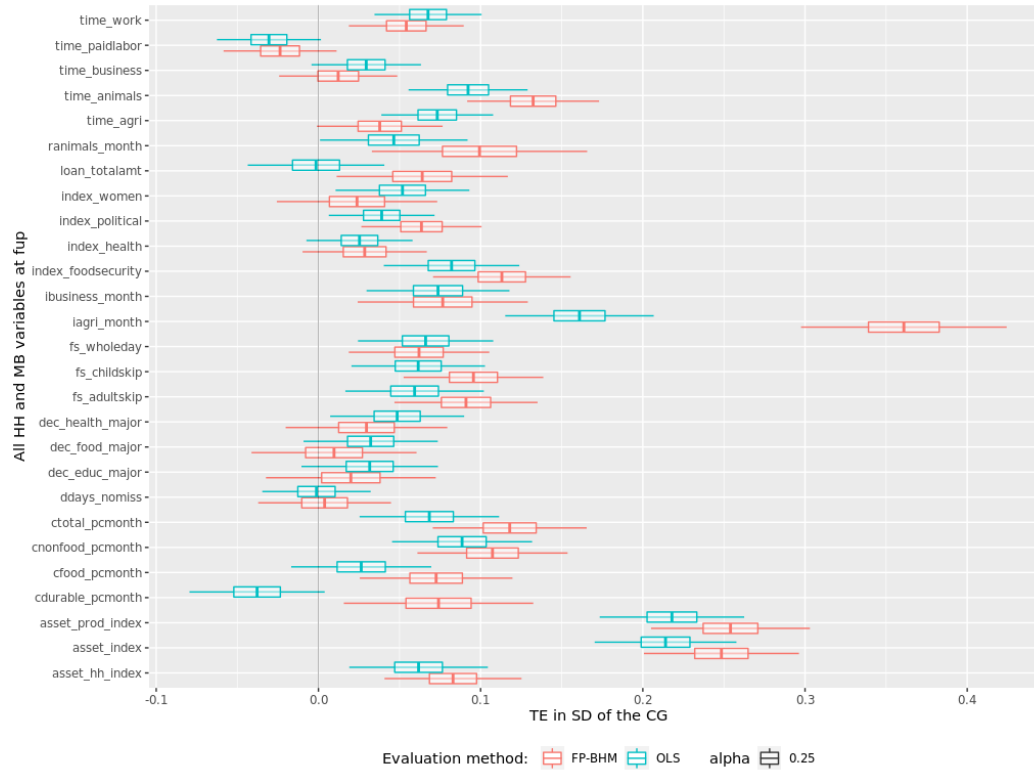


Figure 1d

GTE: FP-BHM and OLS, all variables, End

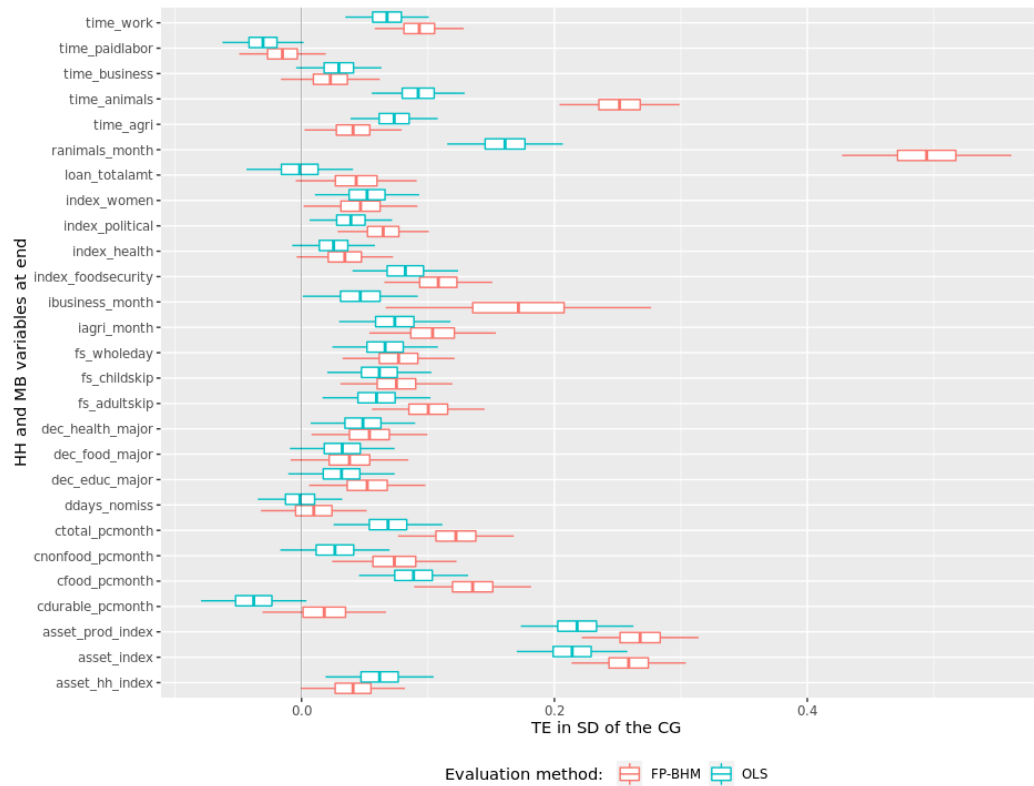


Table 2: Observations and weighting across sites

	Ethiopia	Ghana	Honduras	India	Pakistan	Peru
Number of observation	925	2606	2403	978	1299	2284
Weighting in pooled OLS	8.81%	24.9%	22.89%	9.37%	12.37%	21.76%
Weighting in Hierarchical	16.66%	16.66%	16.66%	16.66%	16.66%	16.66%
E.g.: Local OLS estimated TE of asset_index_fup	0.6035	0.3421	-0.0955	0.7122	0.1696	0.0595

Figure 2: Local Treatment Effect – OLS estimates

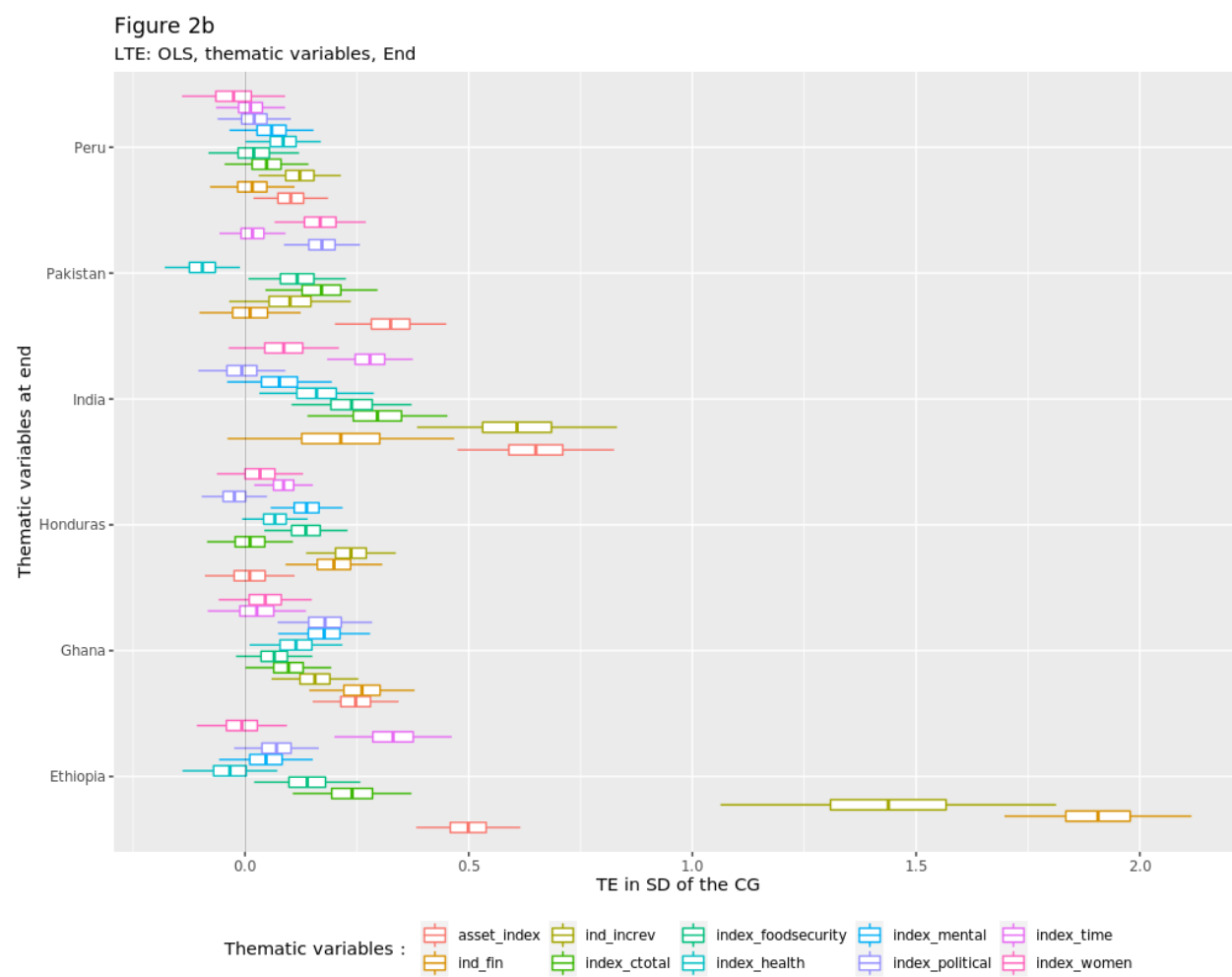


Figure 3: Pooling Factor

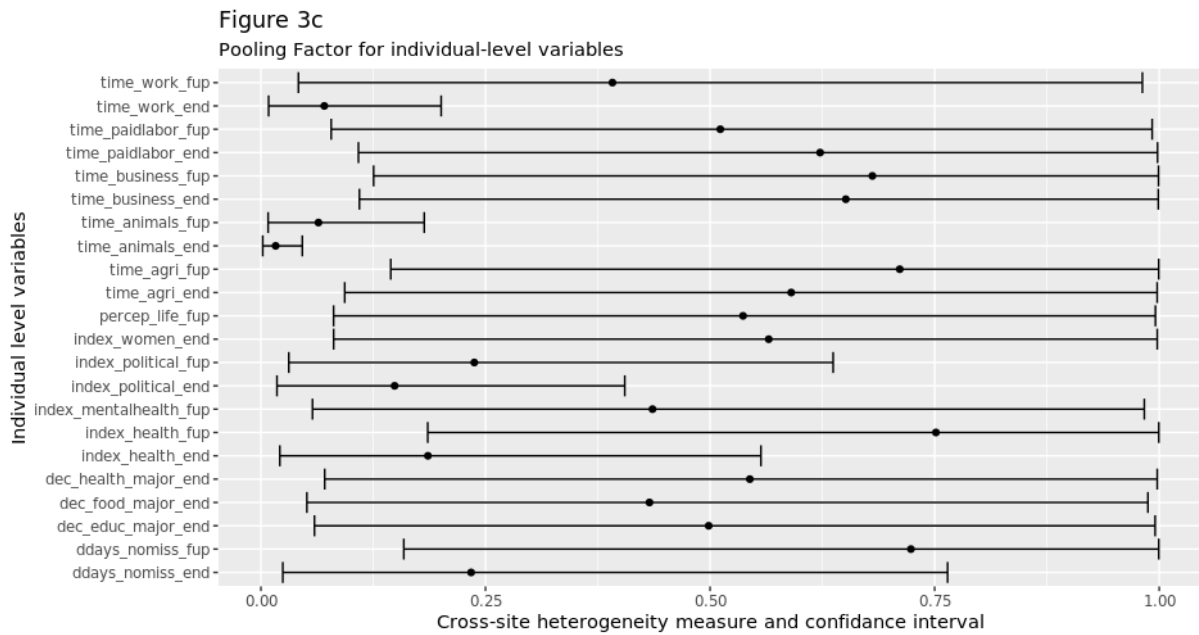
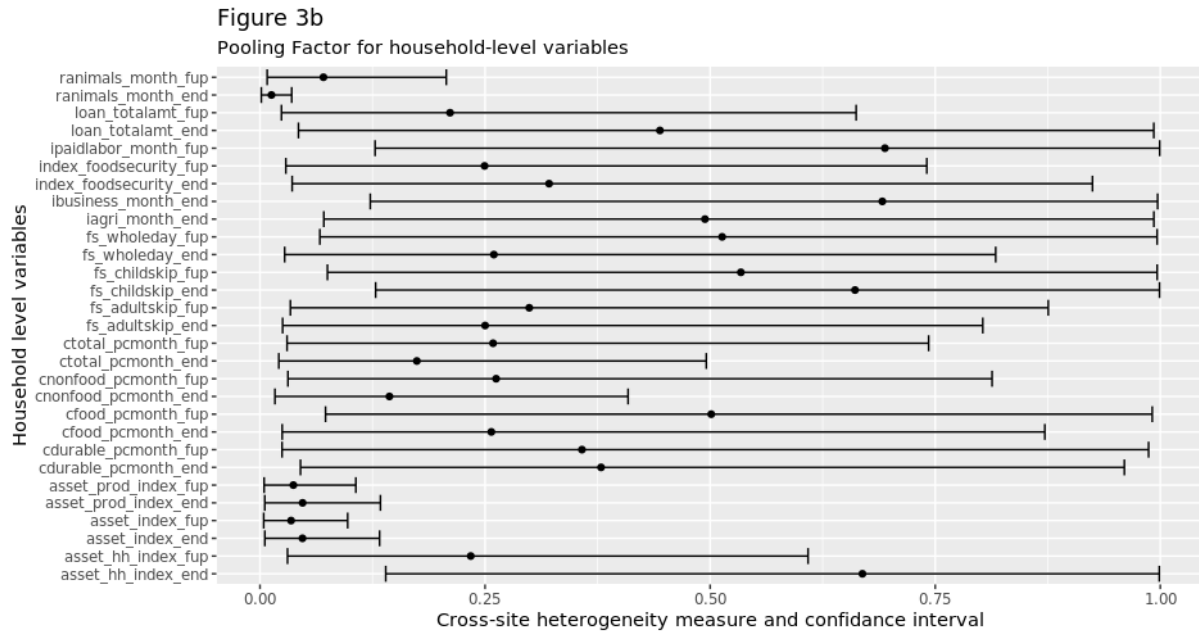


Figure 4 : Local Treatment Effect of thematic variables using FP-BHM or PP-BHM

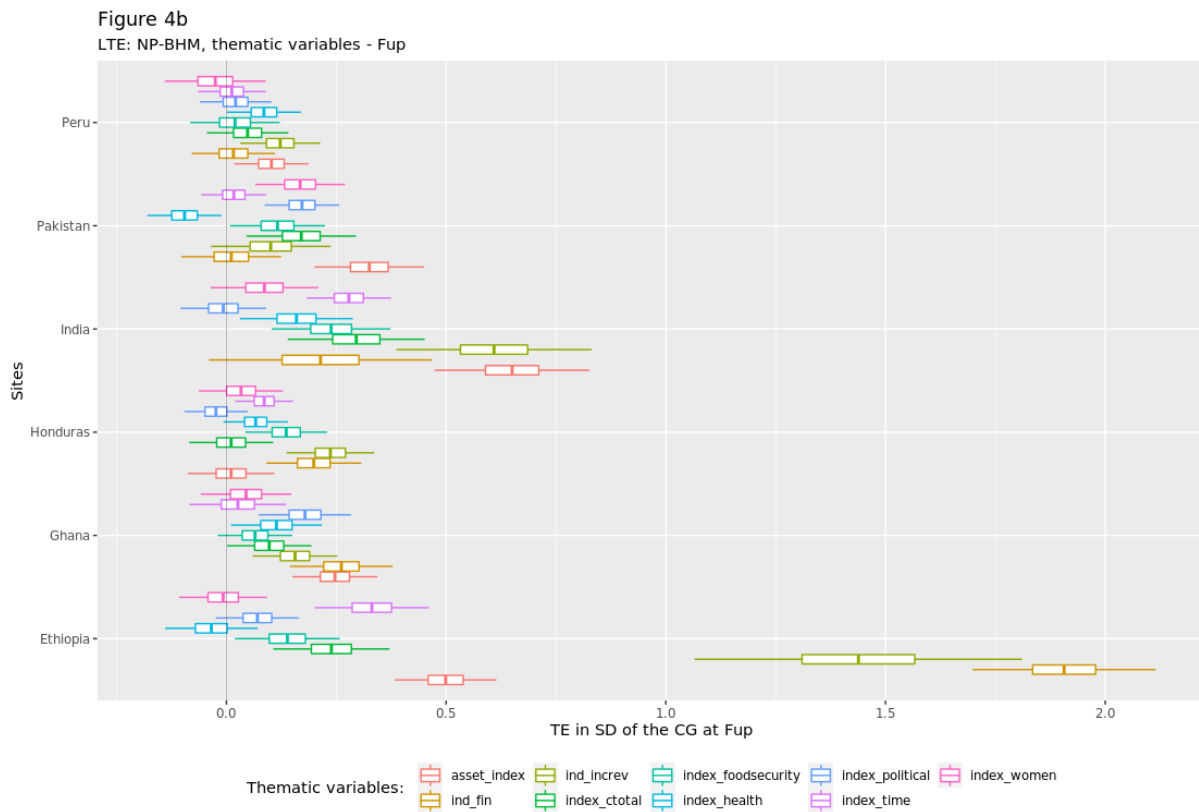
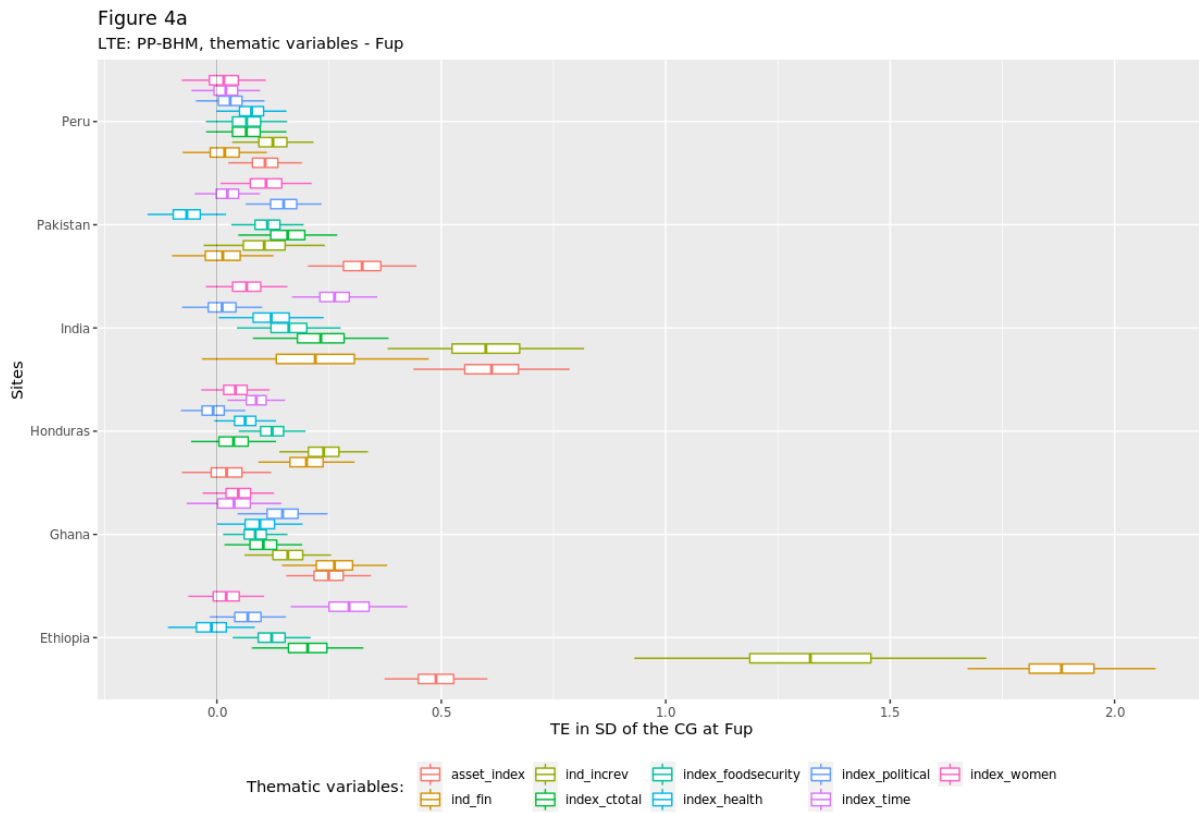


Figure 4c

LTE: PP-BHM, thematic variables - End

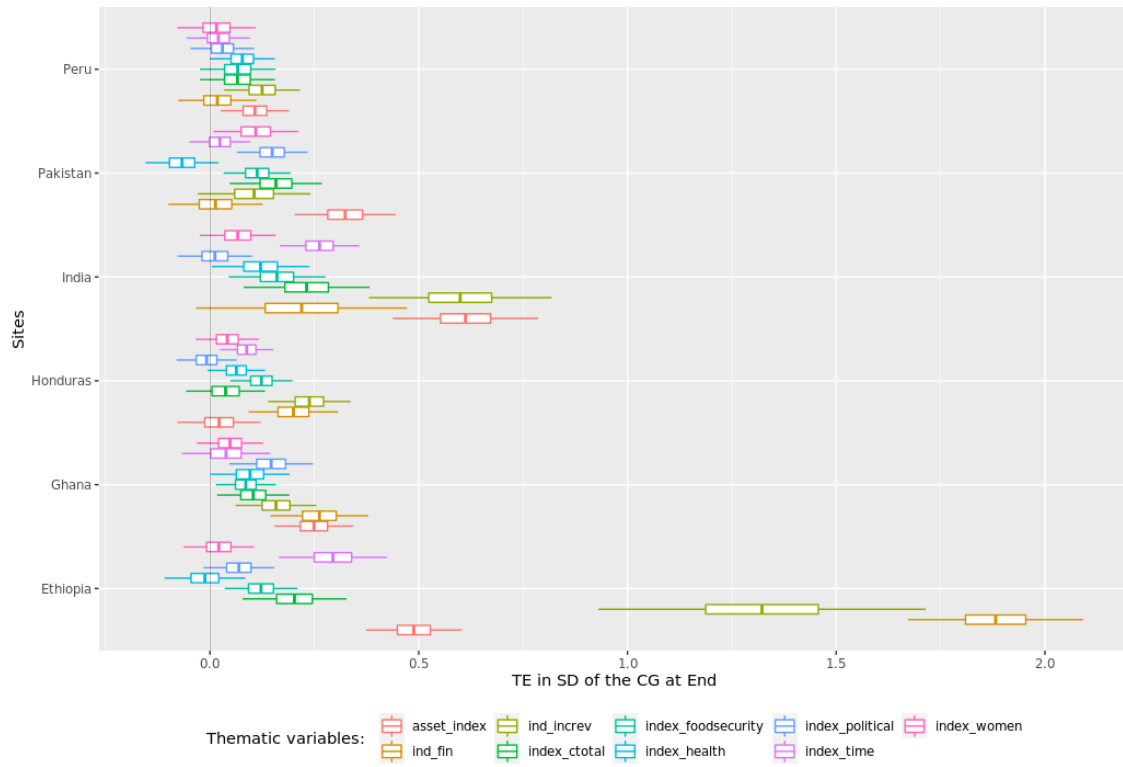


Figure 4d

LTE: NP-BHM, thematic variables - End

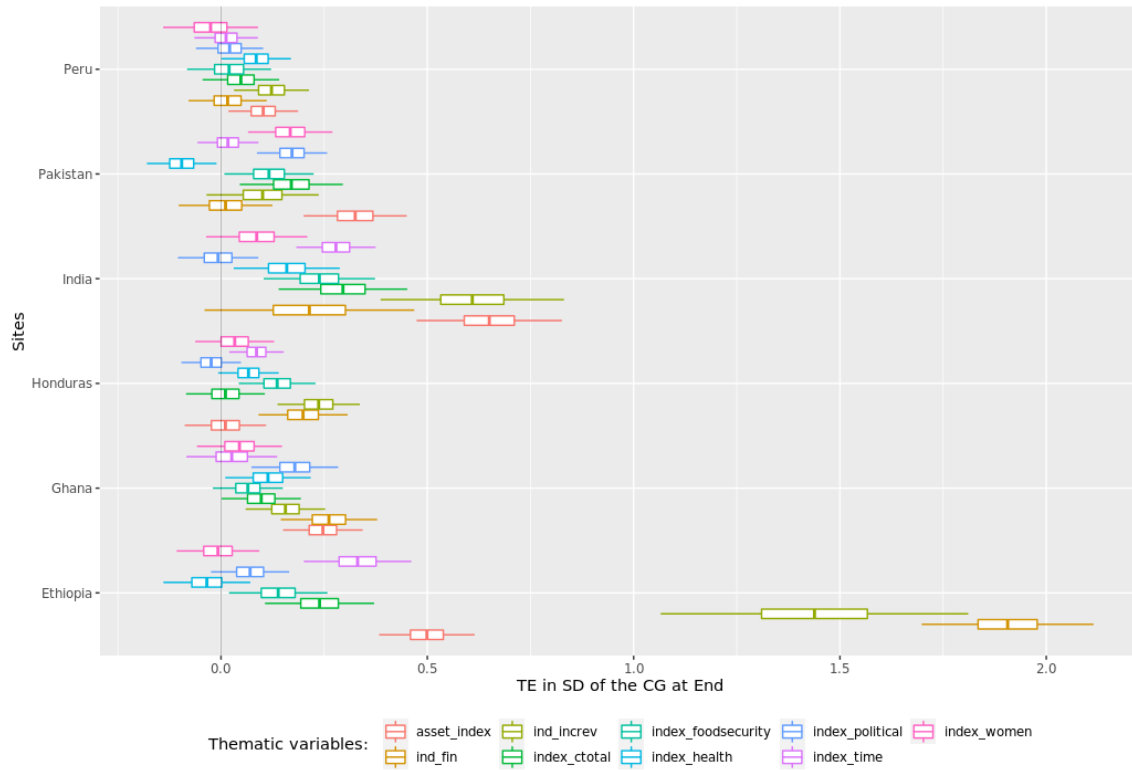


Figure 5: Difference between LTE estimates of PP-BHM and FP-BHM

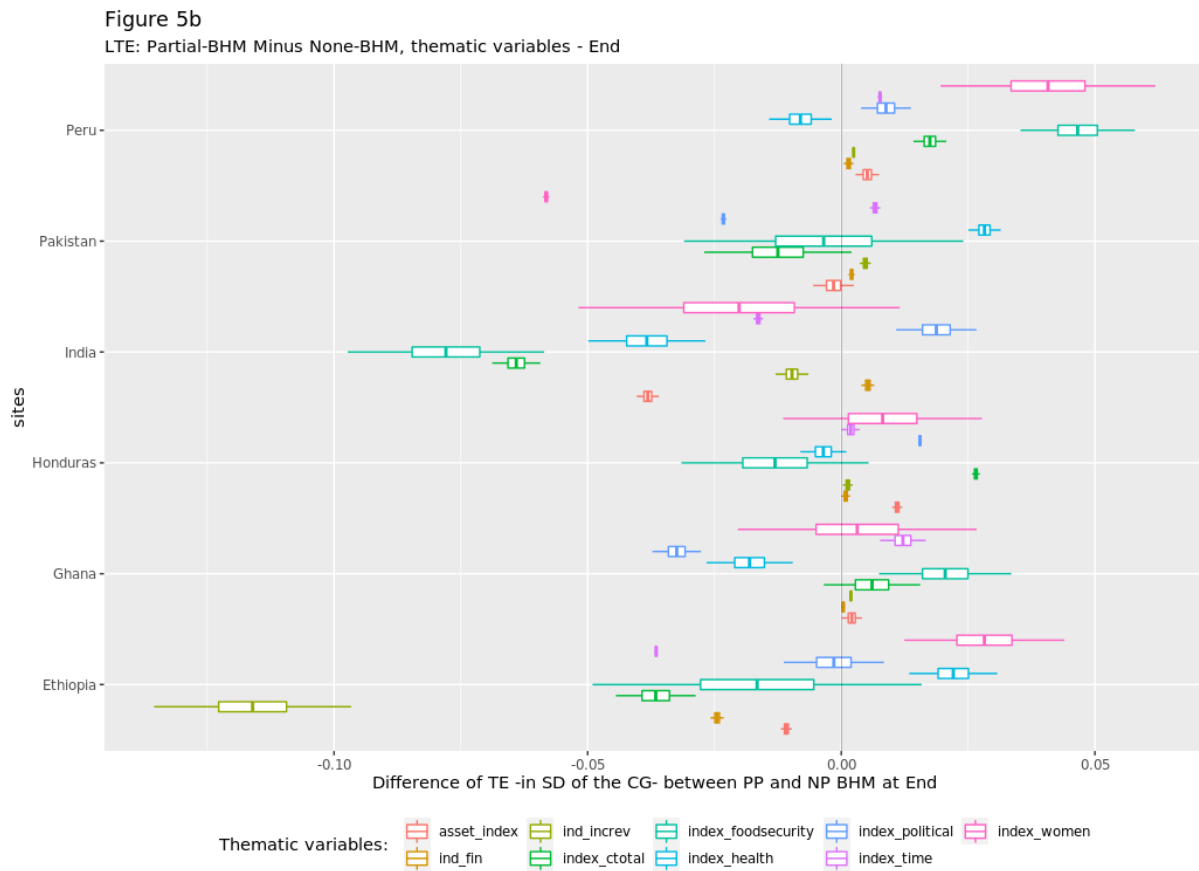


Figure 6a: Global and Local Treatment effect pooling comparison for thematic variables

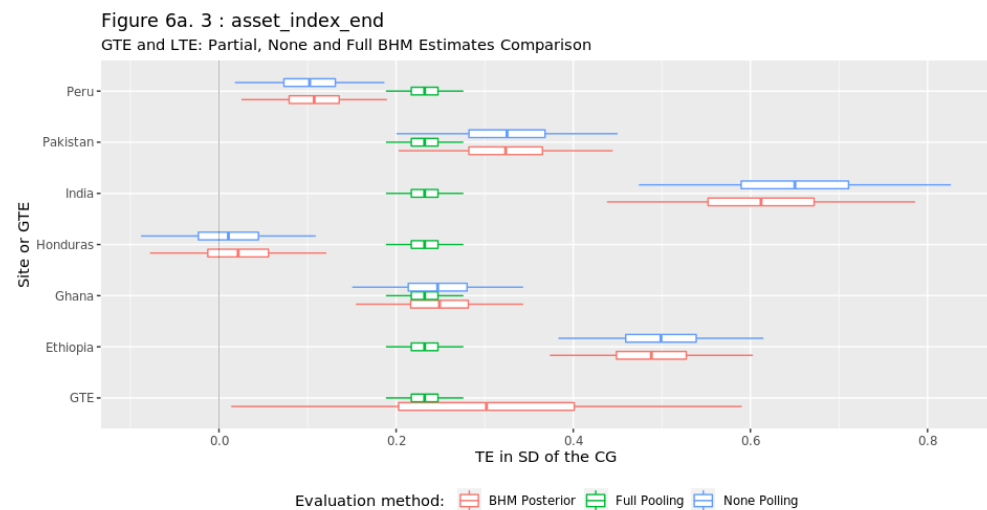
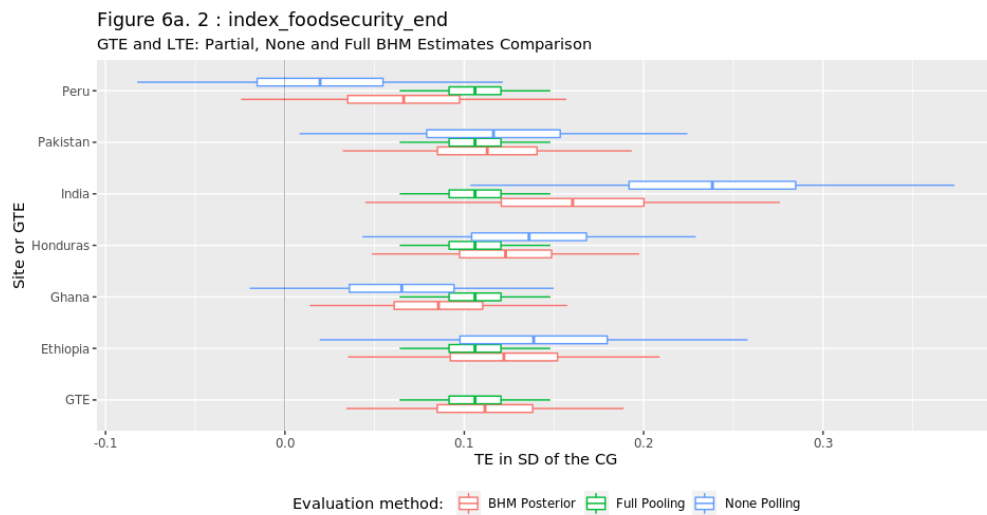
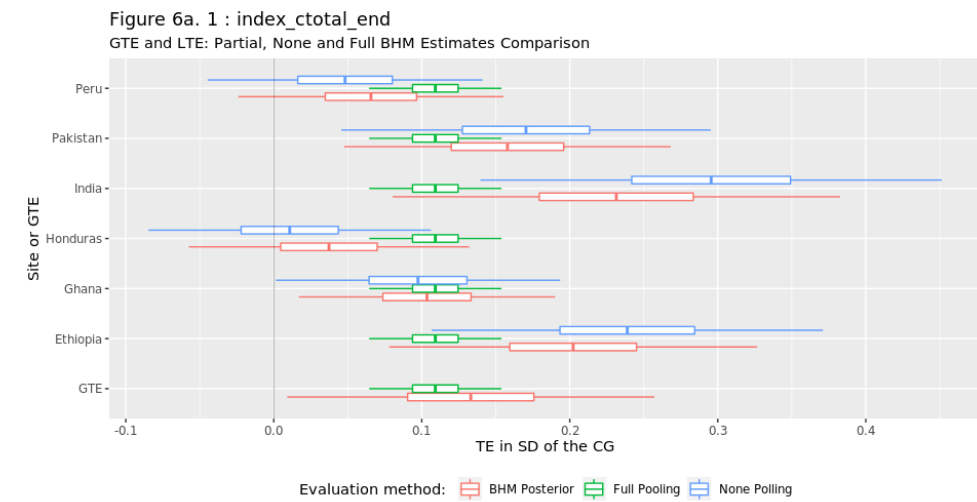


Figure 6a. 4 : ind_fin_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

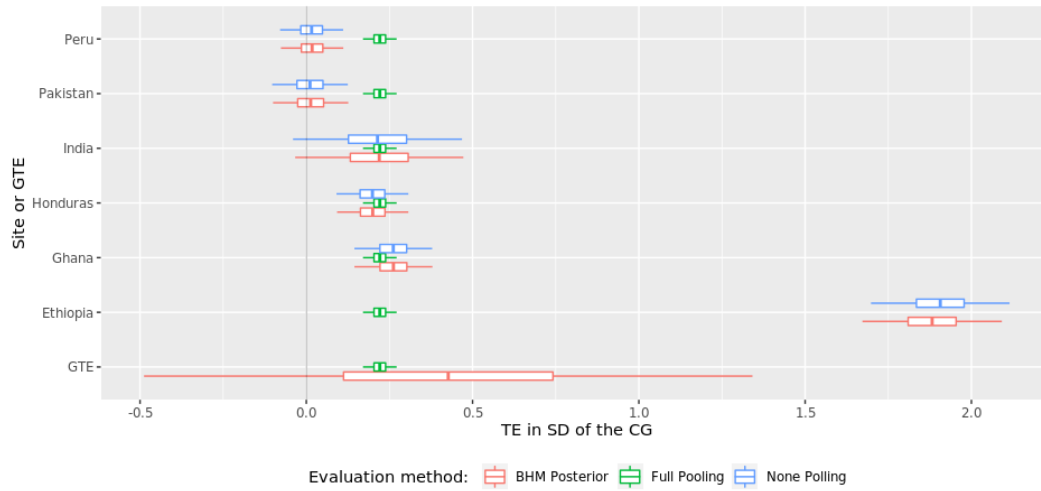


Figure 6a. 5 : index_time_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

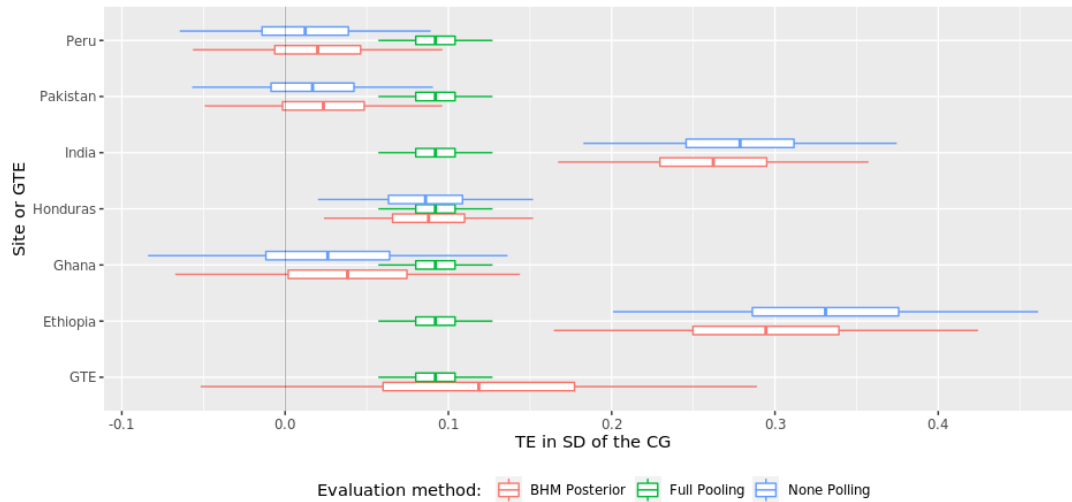


Figure 6a. 6 : ind_increv_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

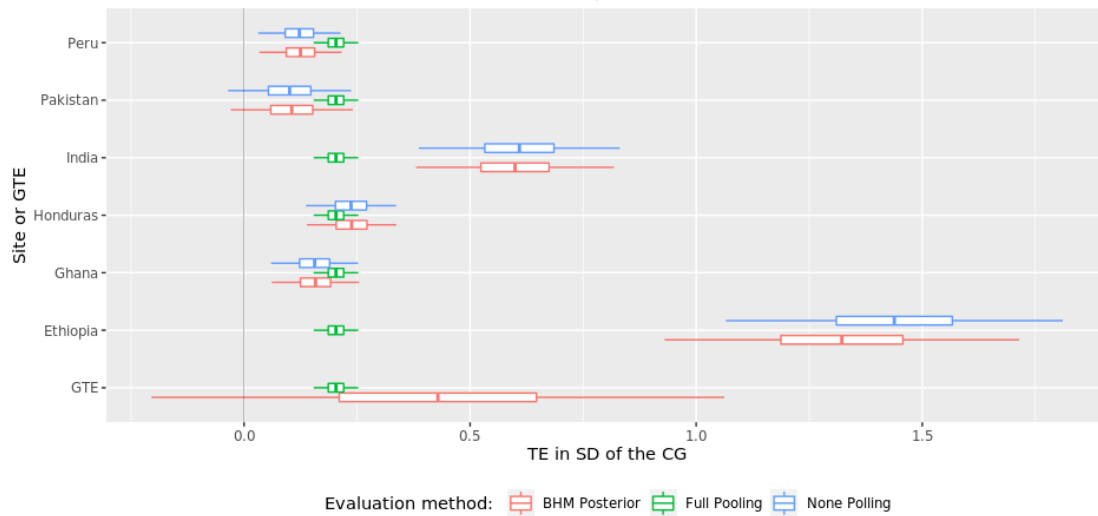


Figure 6a. 7 : index_health_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

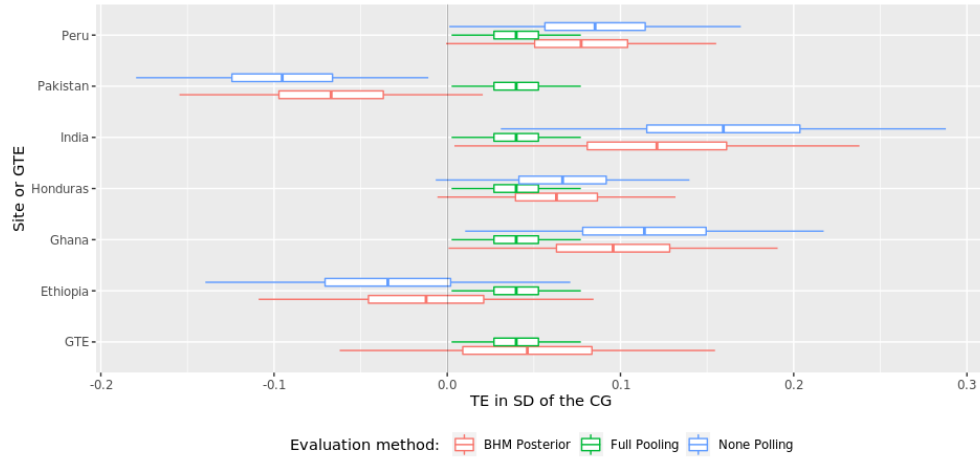


Figure 6a. 8 : index_political_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

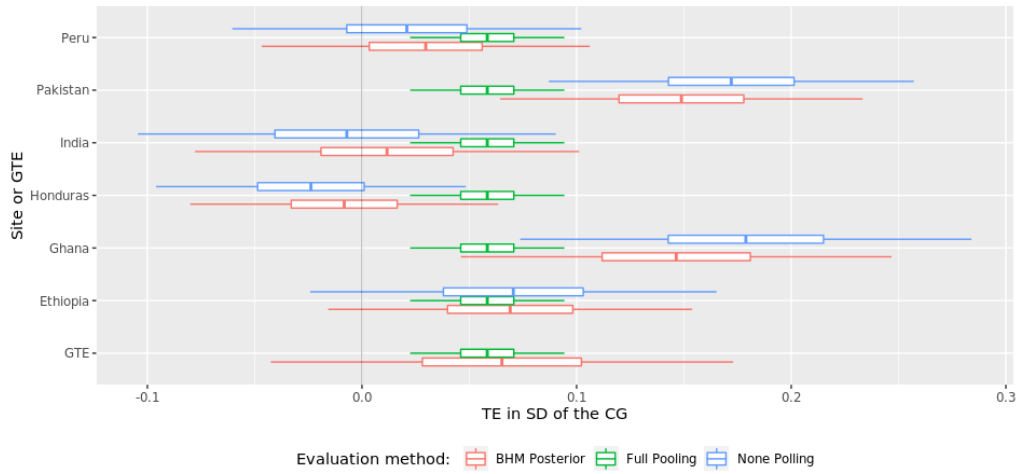


Figure 6a. 9 : index_women_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

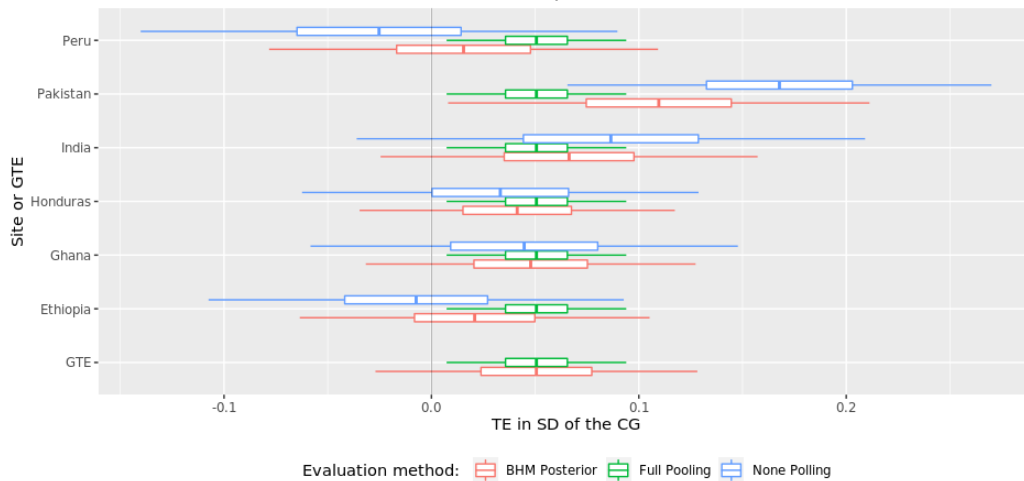


Figure 6a. 10 : index_ctotal_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

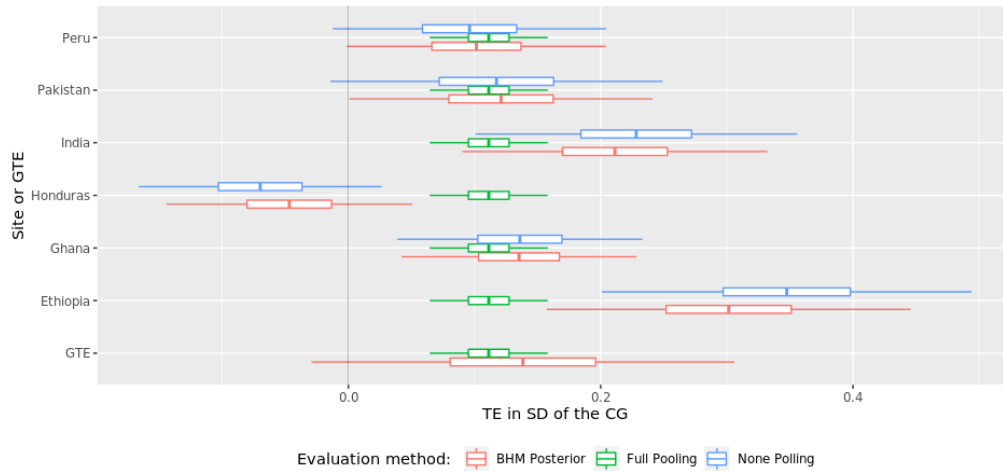


Figure 6a. 11 : index_foodsecurity_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

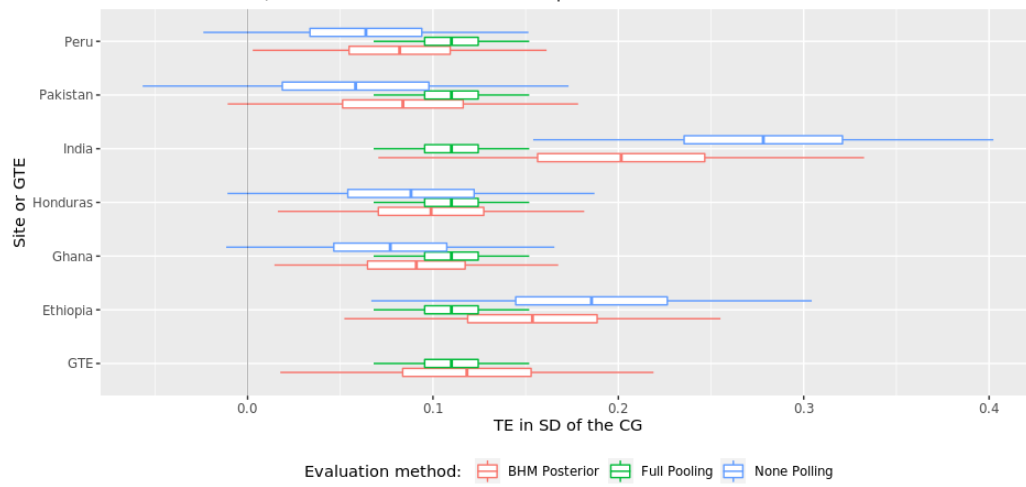


Figure 6a. 12 : asset_index_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

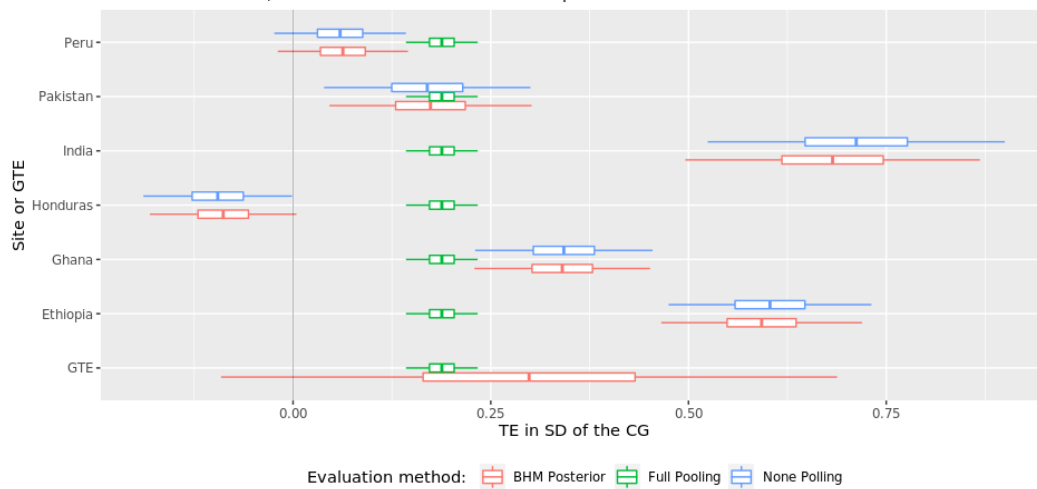


Figure 6a. 13 : ind_fin_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

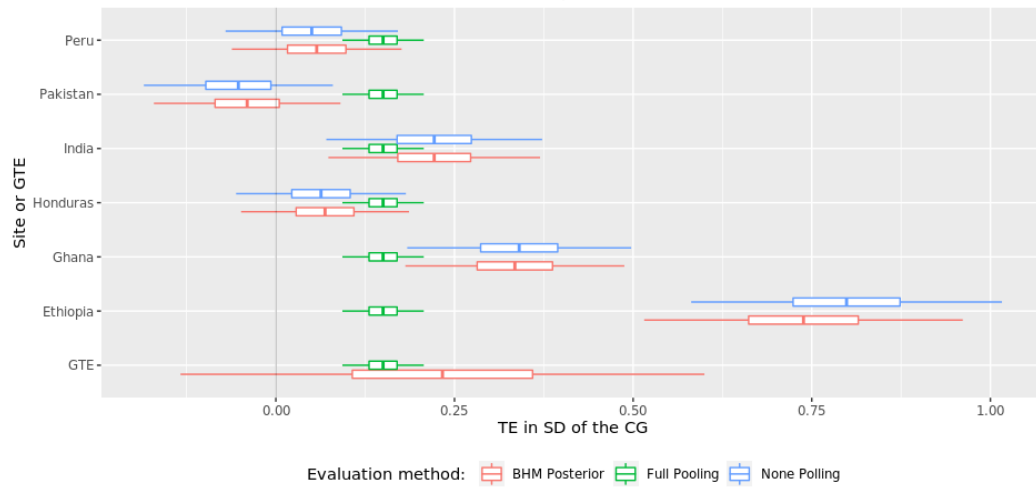


Figure 6a. 14 : index_time_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

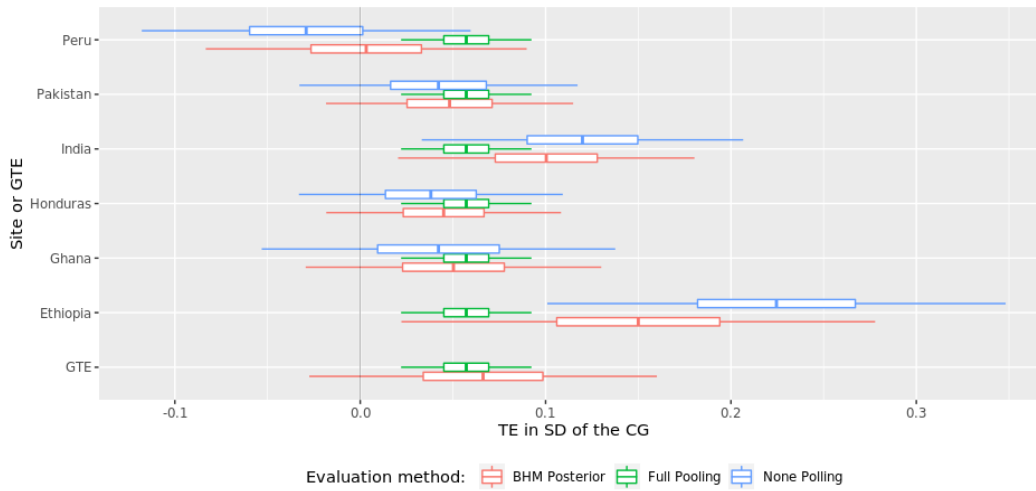


Figure 6a. 15 : ind_increv_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

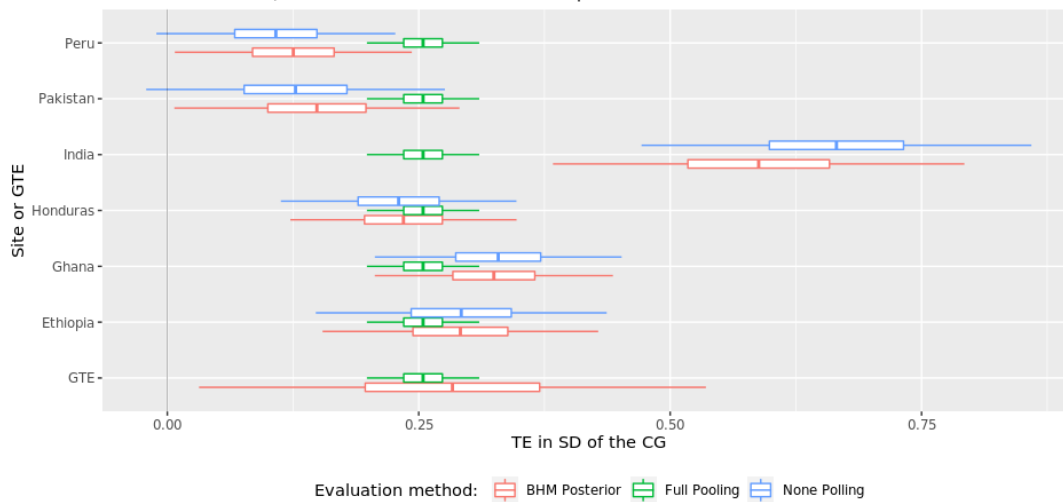


Figure 6a. 16 : index_health_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

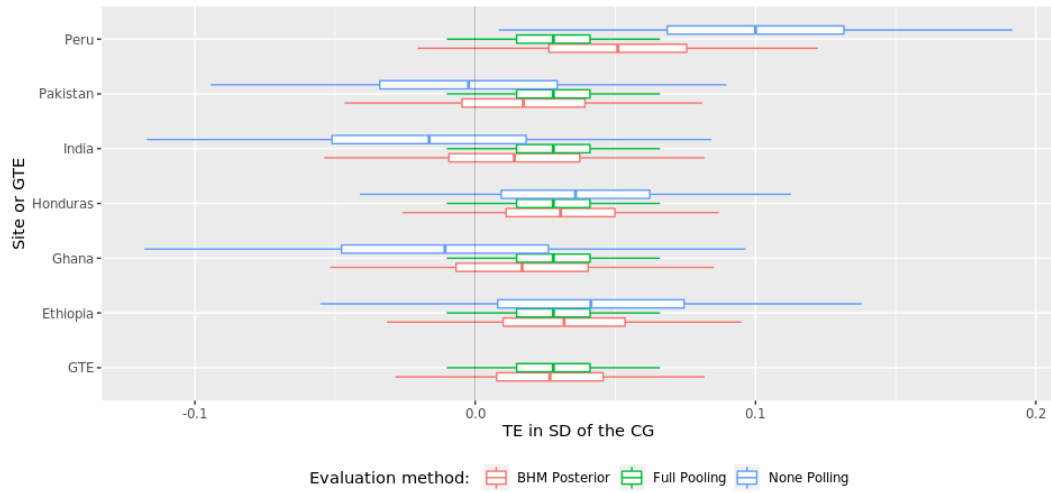


Figure 6a. 17 : index_mental_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

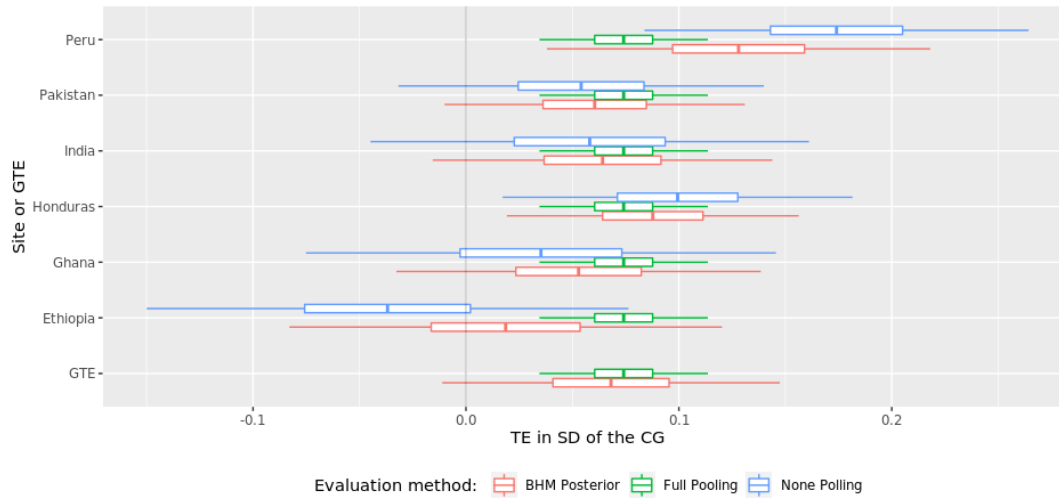


Figure 6a. 18 : index_political_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

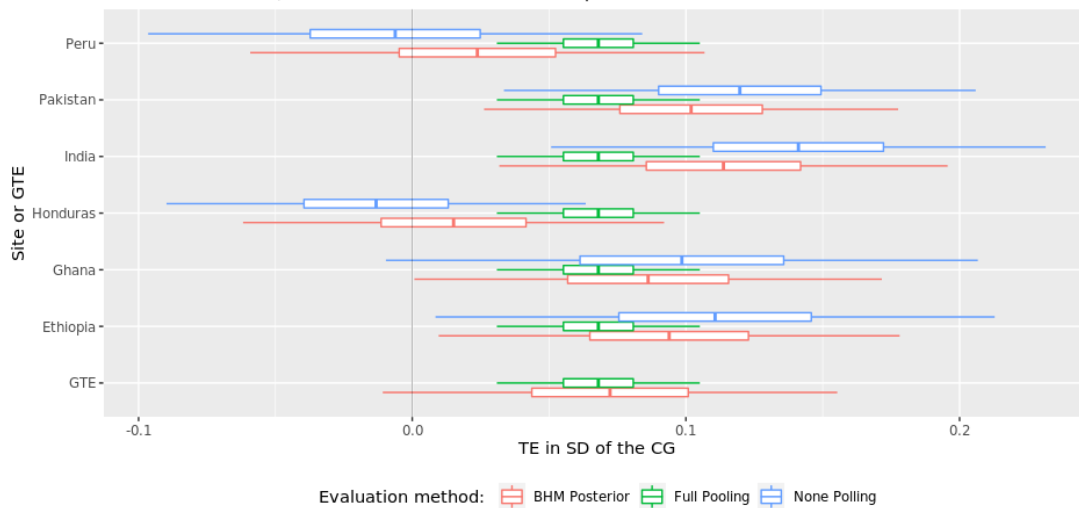


Figure 6b: Global and Local Treatment effect pooling comparison for household and individual variables

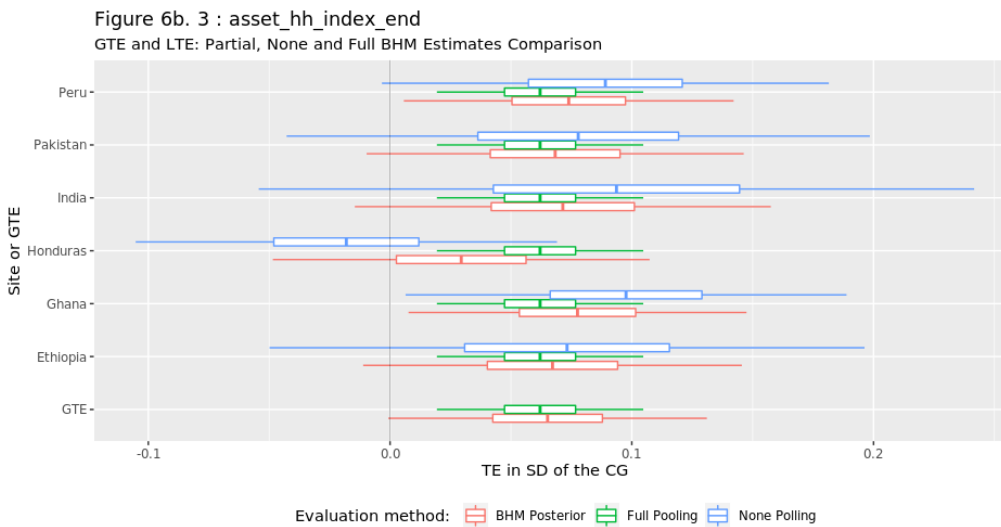
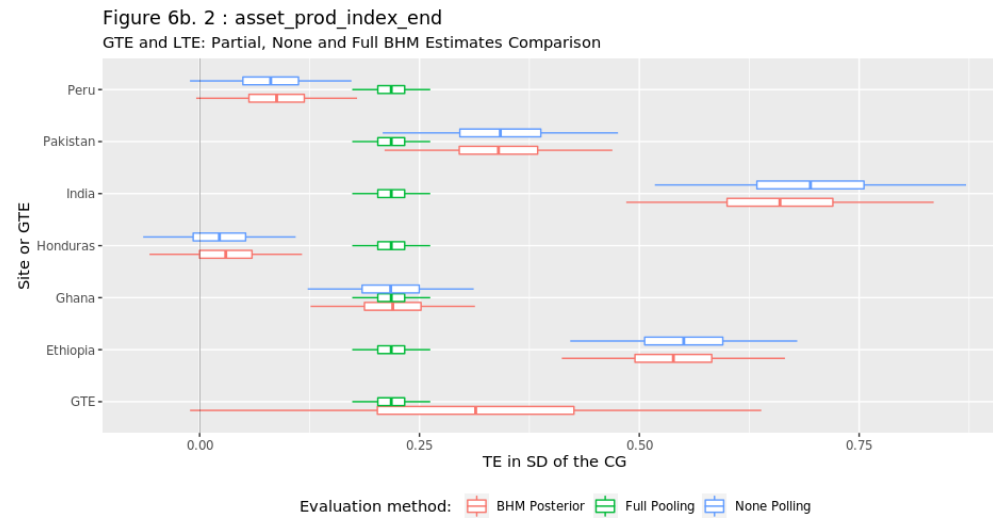
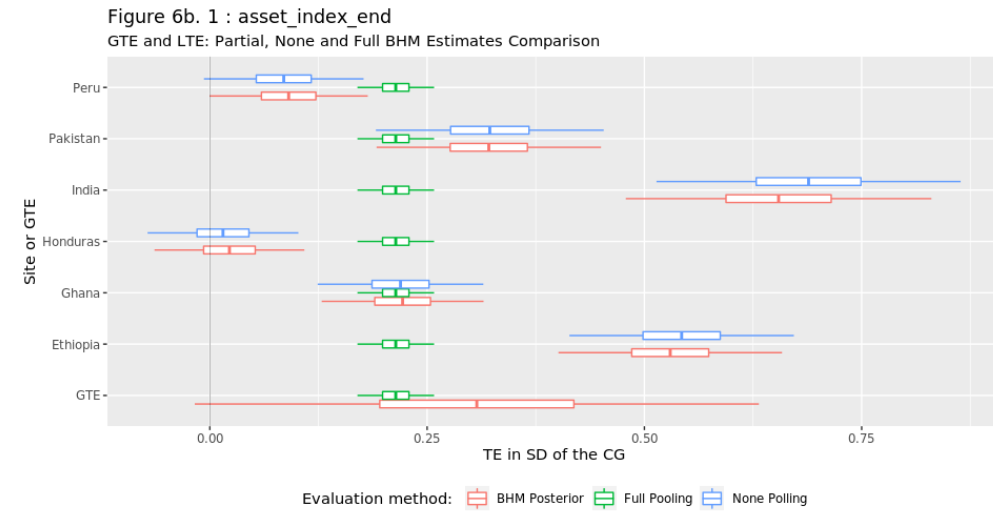


Figure 6b. 4 : iagri_month_end
GTE and LTE: Partial, None and Full BHM Estimates Comparison

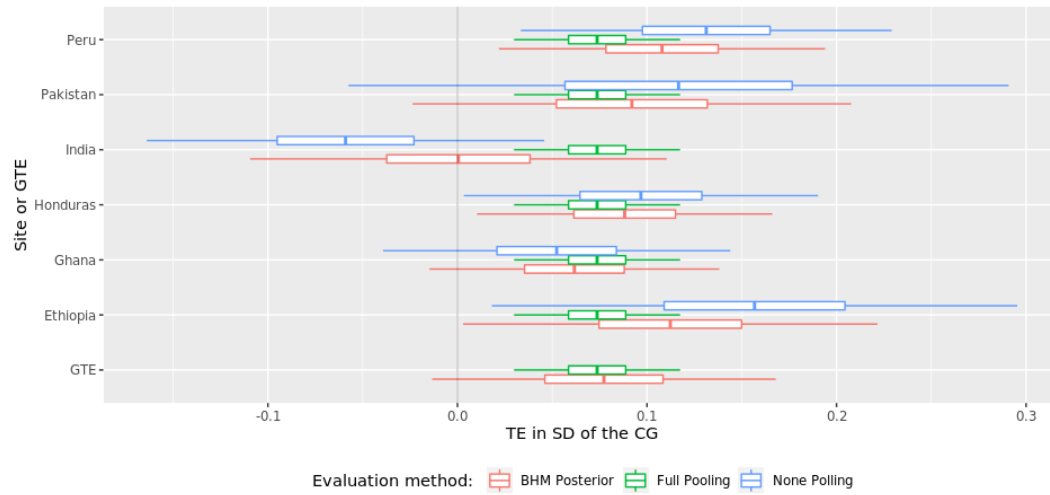


Figure 6b. 5 : ibusiness_month_end
GTE and LTE: Partial, None and Full BHM Estimates Comparison

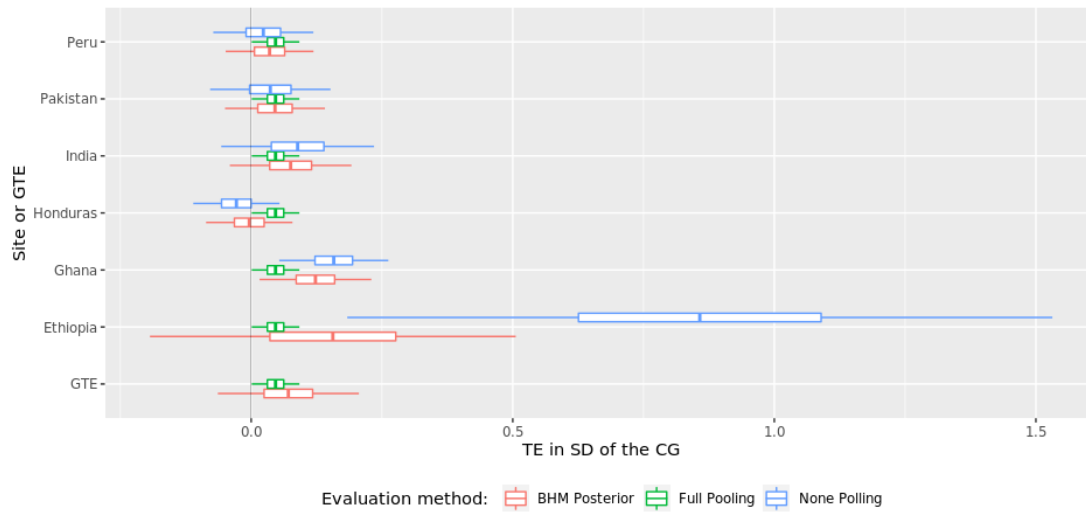


Figure 6b. 6 : ranimals_month_end
GTE and LTE: Partial, None and Full BHM Estimates Comparison

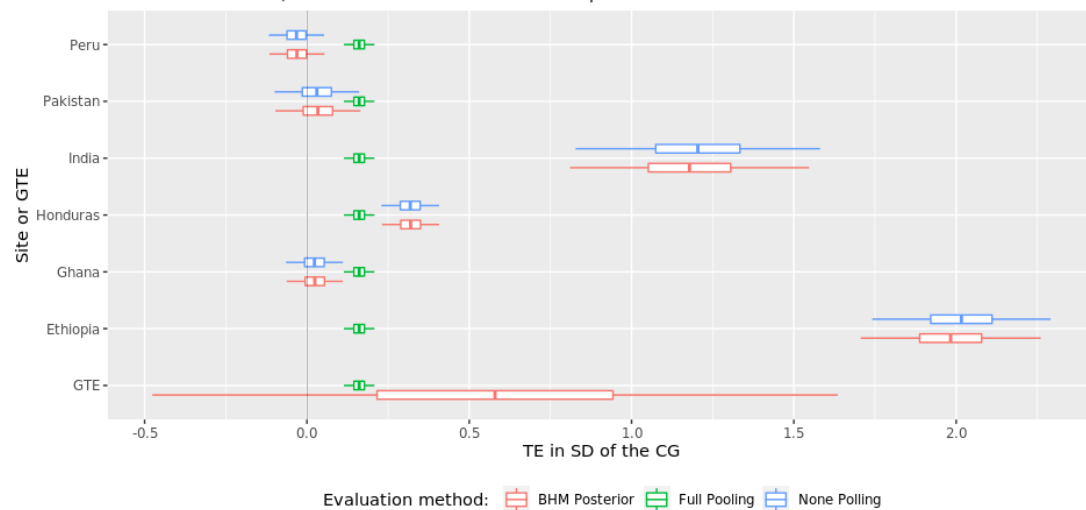


Figure 6b. 7 : ctotal_pcmoenth_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

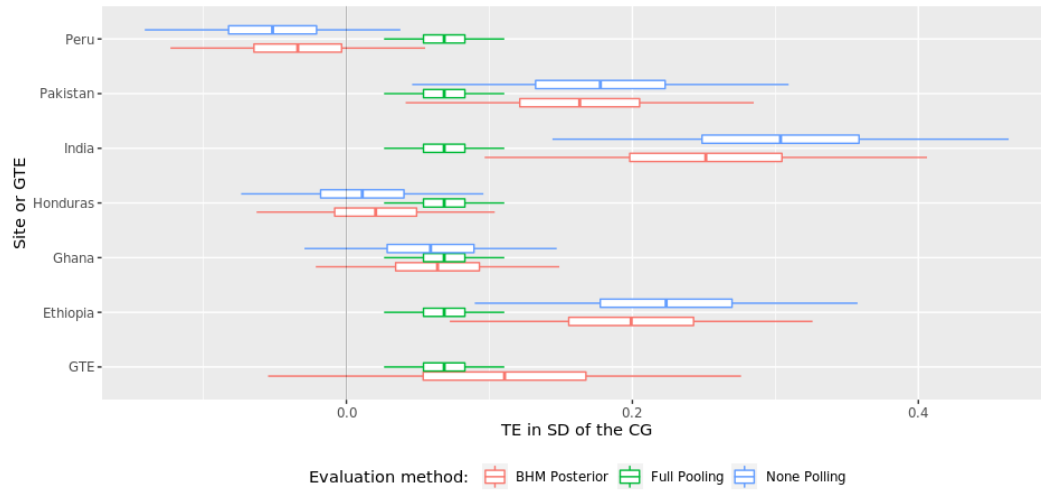


Figure 6b. 8 : cnonfood_pcmoenth_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

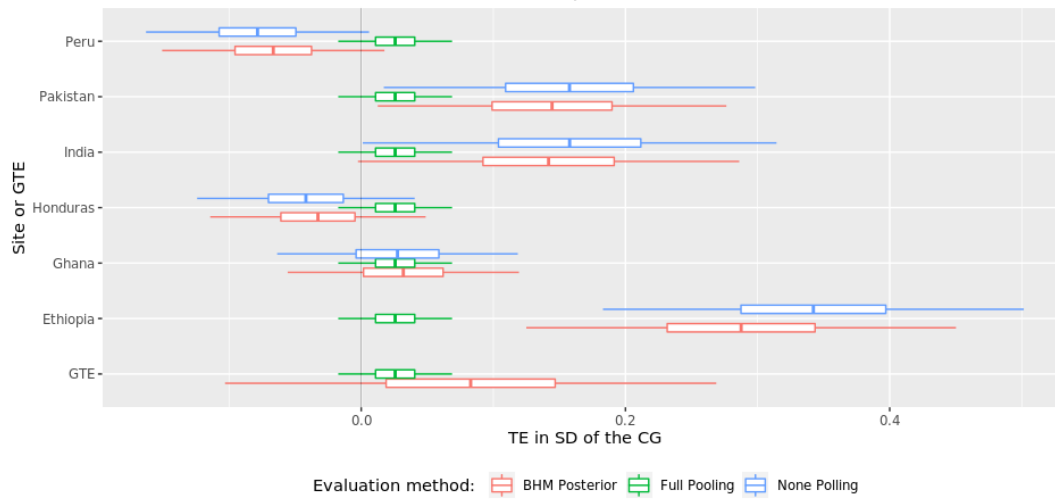


Figure 6b. 9 : cfood_pcmoenth_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

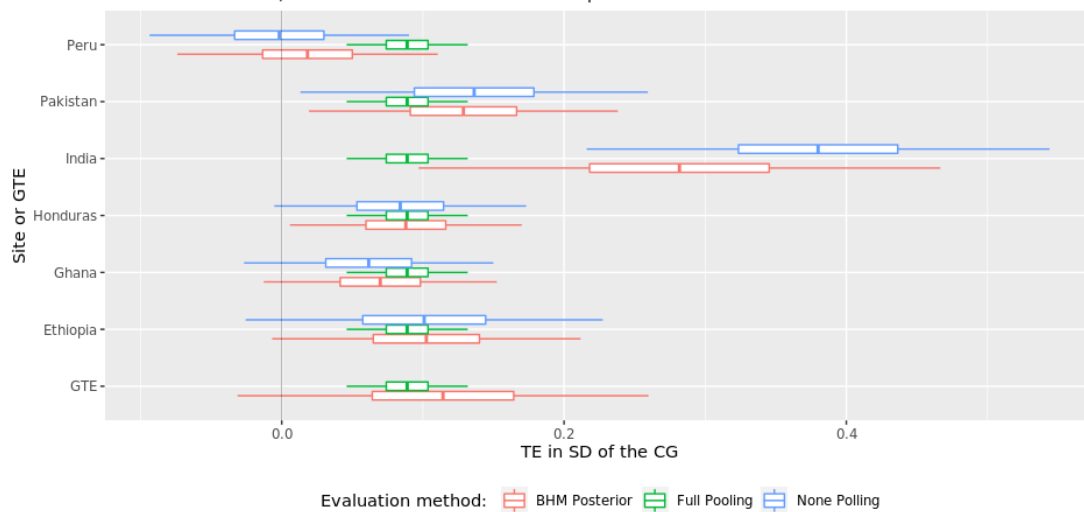


Figure 6b. 10 : cdurable_pcmoenth_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

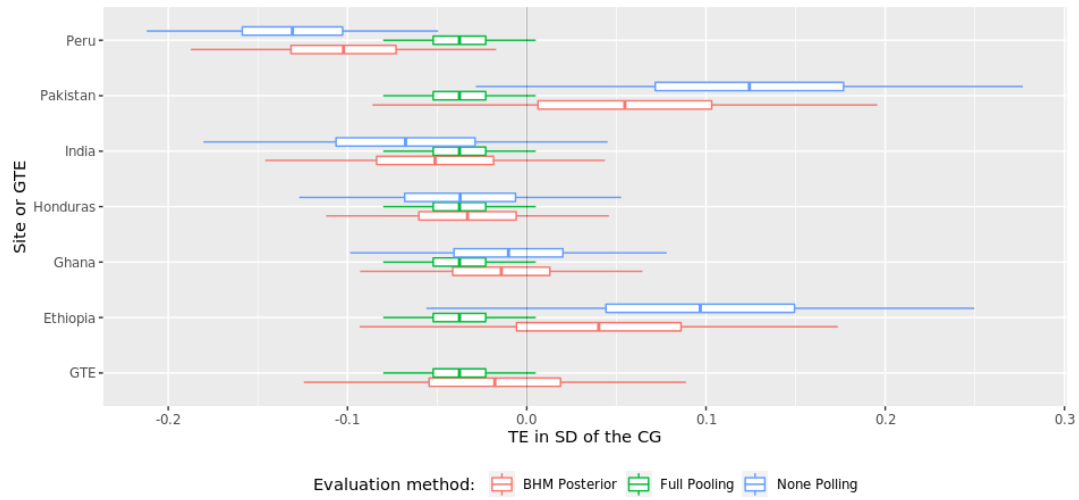


Figure 6b. 11 : index_foodsecurity_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

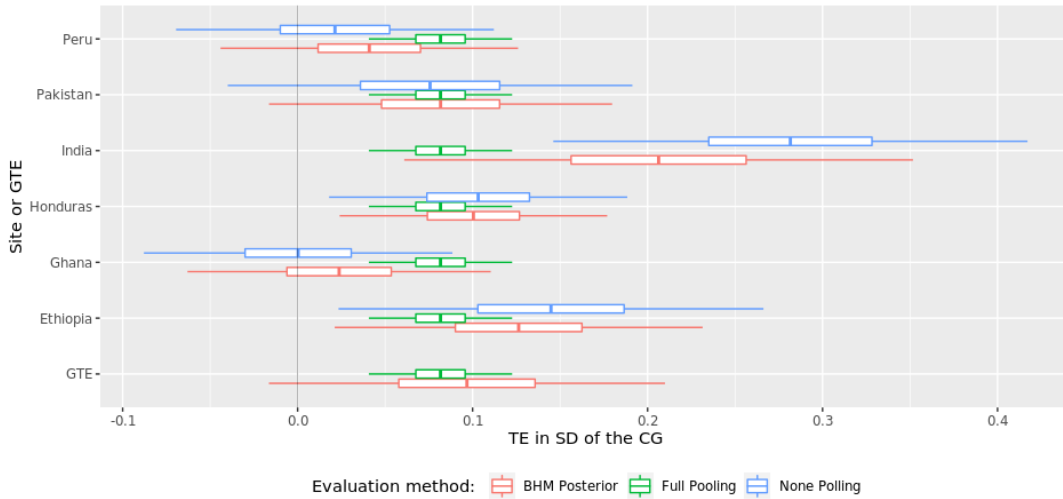


Figure 6b. 12 : fs_adultskip_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

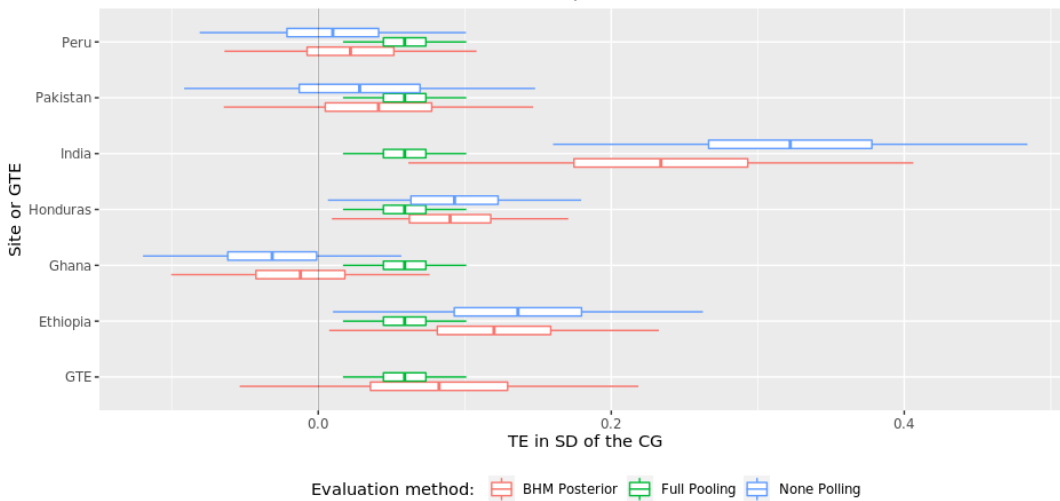


Figure 6b. 13 : fs_wholeday_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

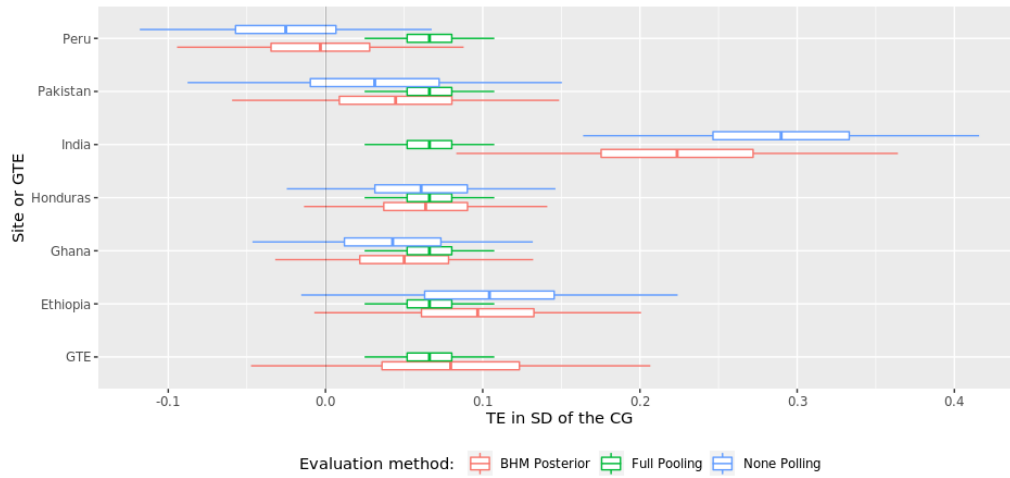


Figure 6b. 14 : fs_childskip_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

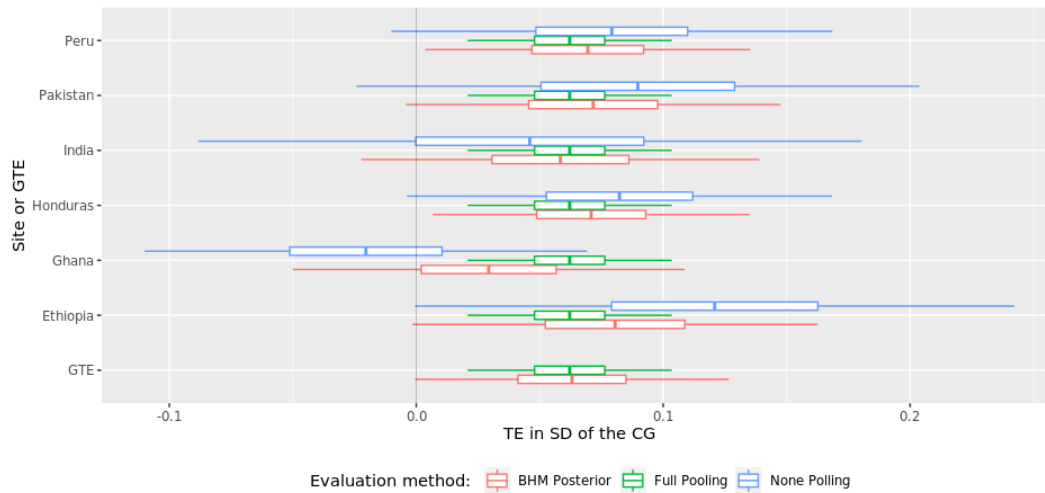


Figure 6b. 15 : loan_totalamt_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

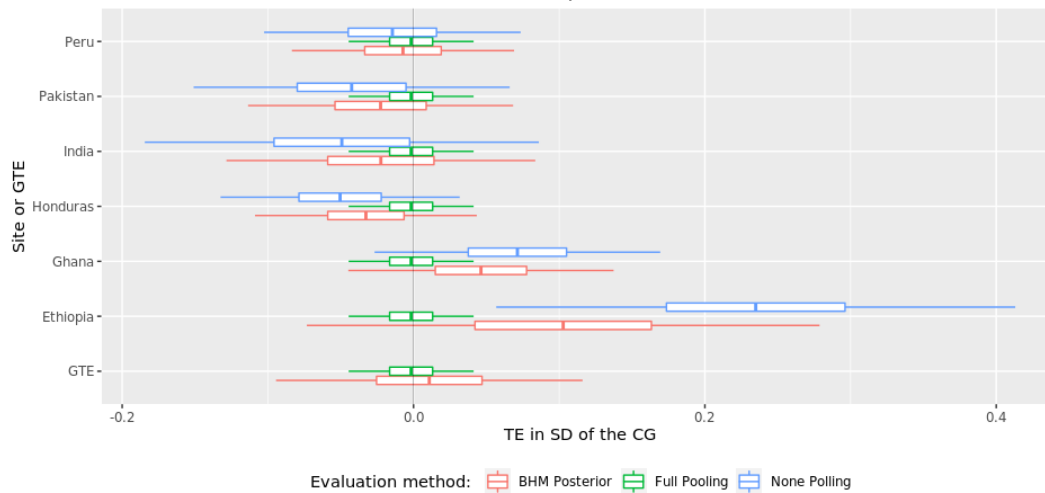


Figure 6b. 16 : index_health_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

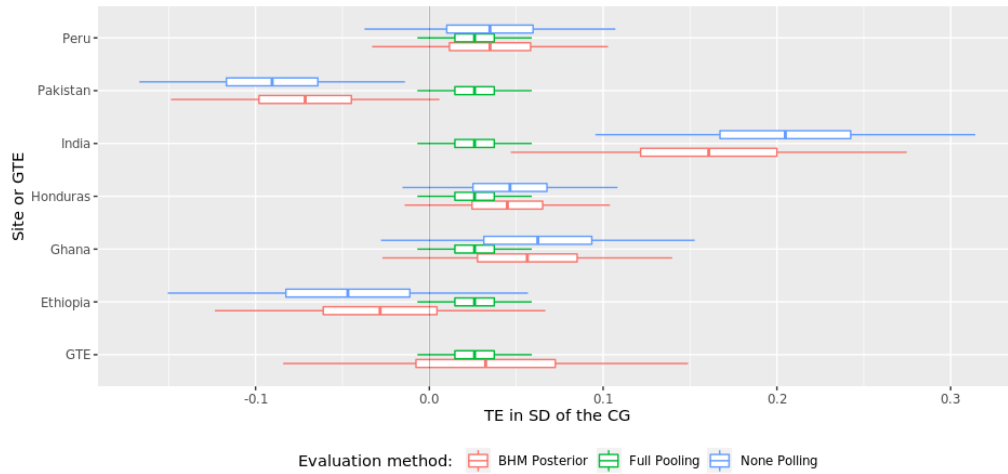


Figure 6b. 17 : ddays_nomiss_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

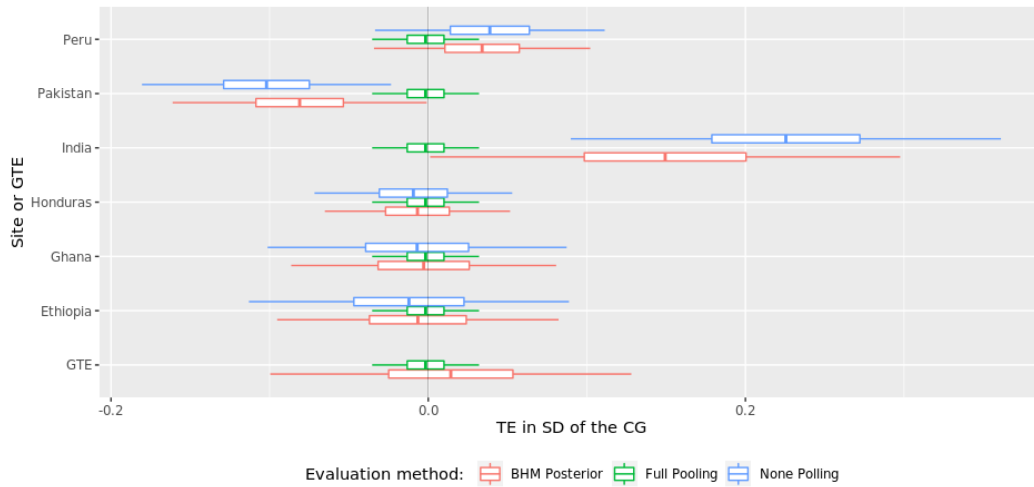


Figure 6b. 18 : time_work_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

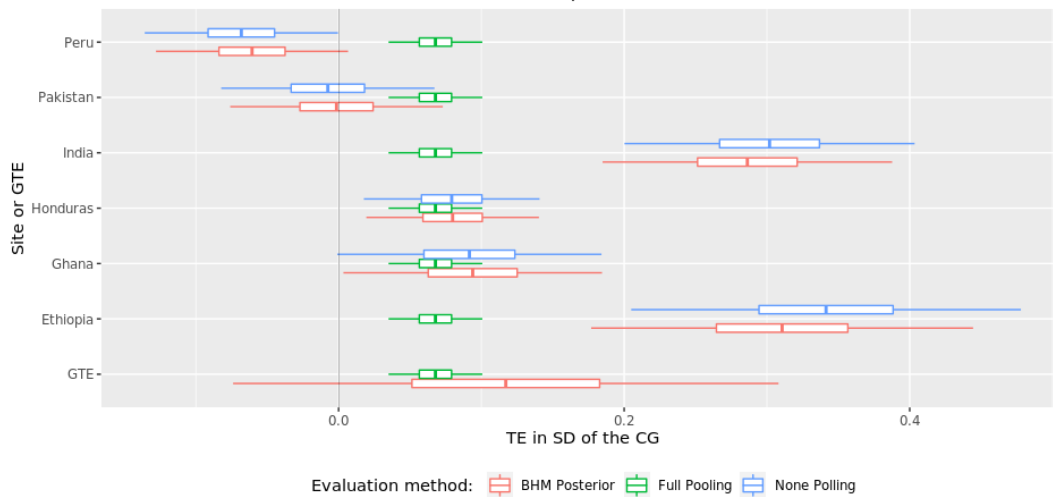


Figure 6b. 19 : time_agri_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

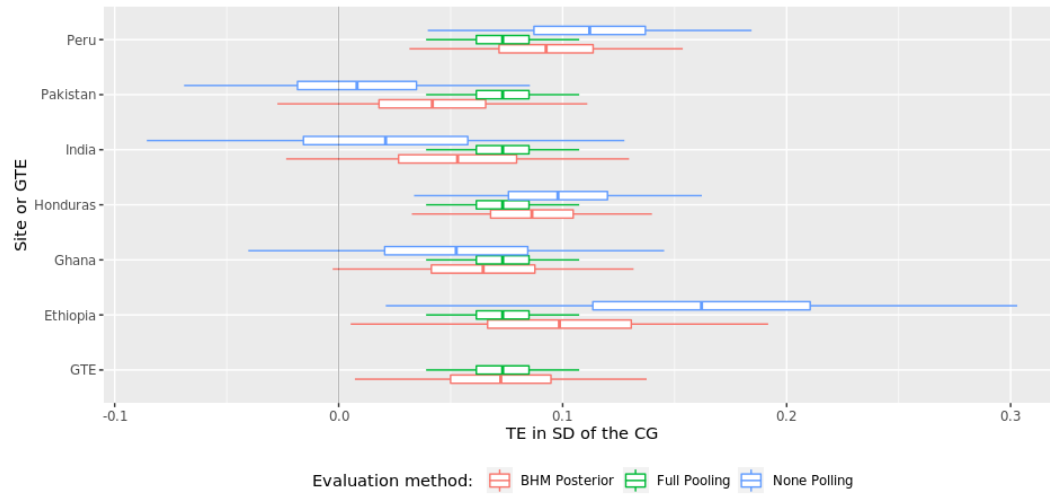


Figure 6b. 20 : time_animals_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

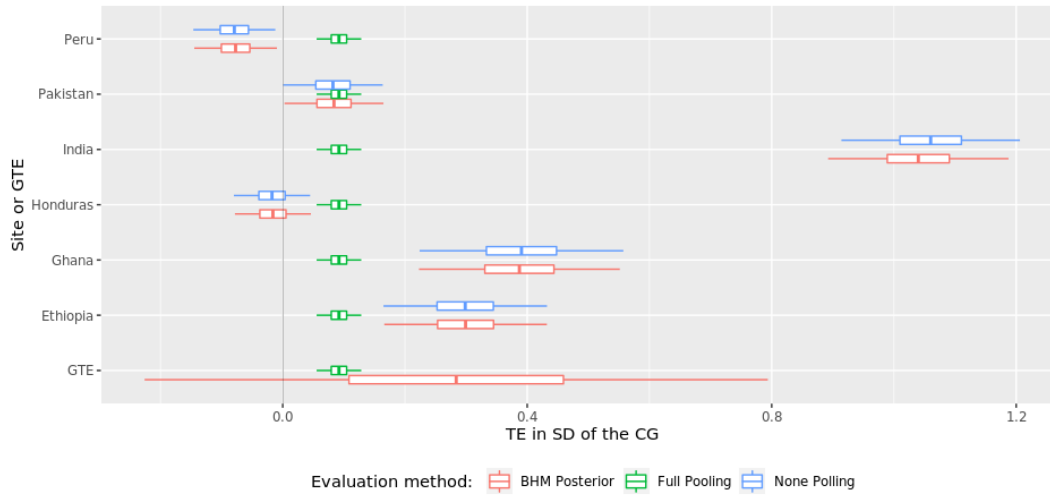


Figure 6b. 21 : time_business_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

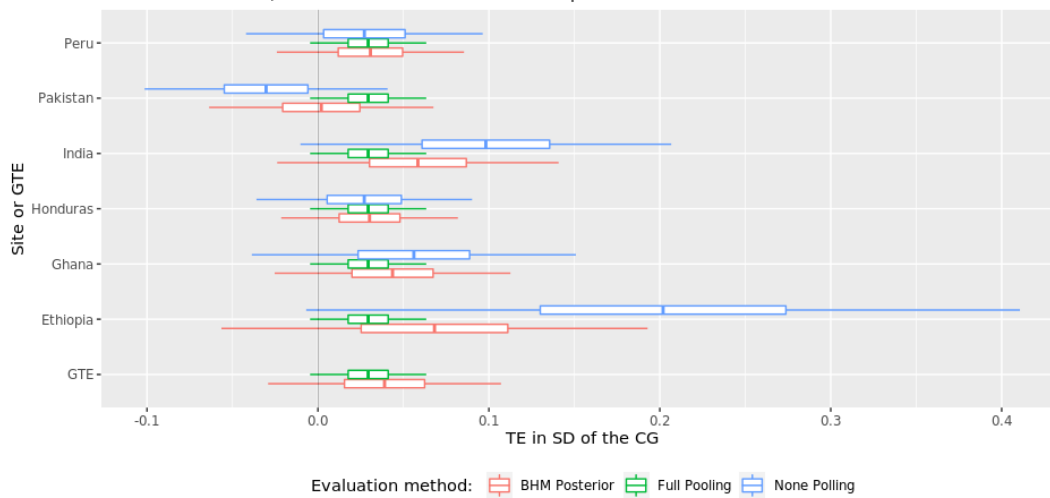


Figure 6b. 22 : time_paidlabor_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

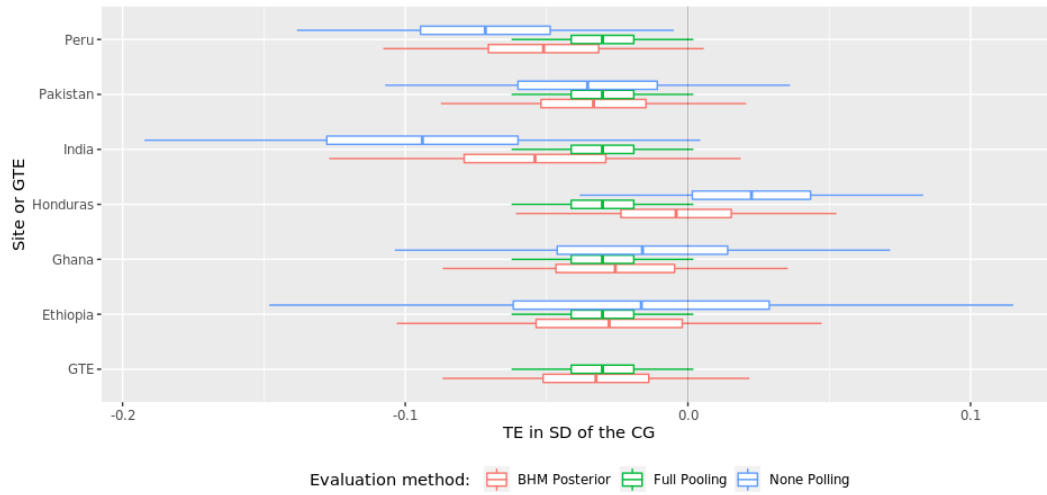


Figure 6b. 23 : index_political_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

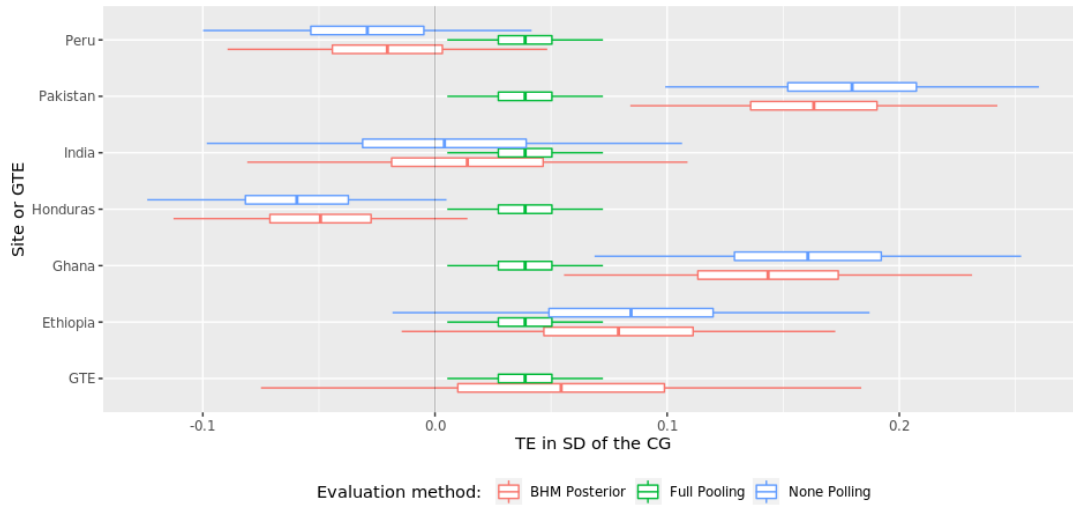


Figure 6b. 24 : index_women_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

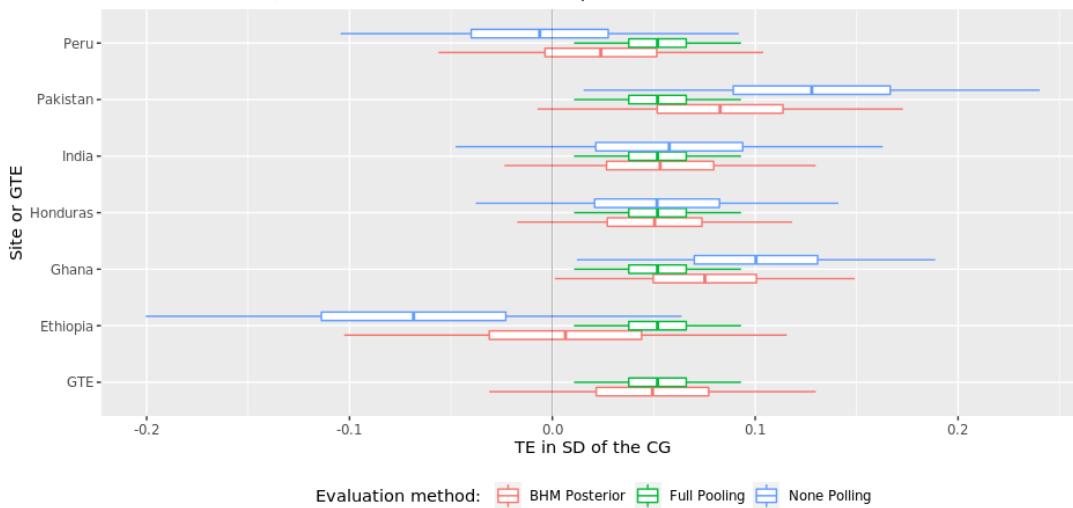


Figure 6b. 25 : dec_food_major_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

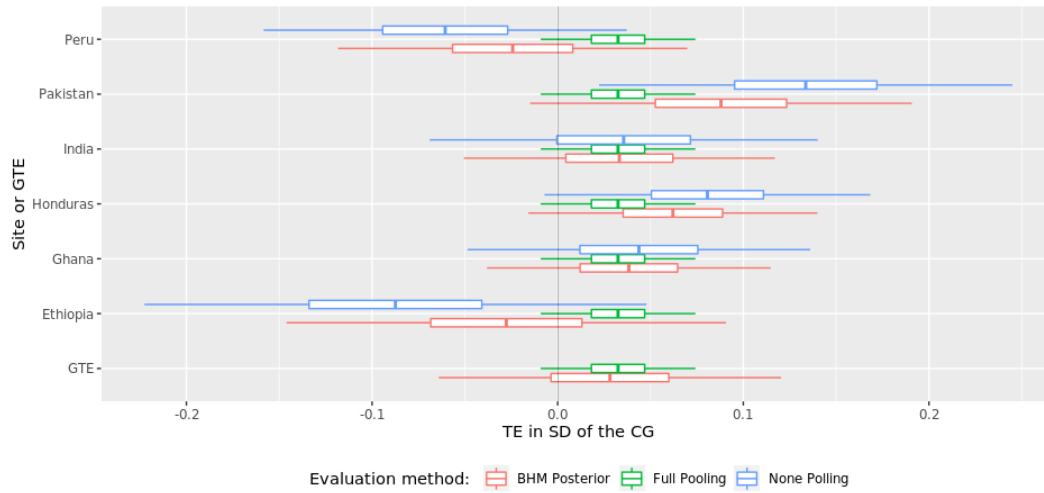


Figure 6b. 26 : dec_educ_major_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

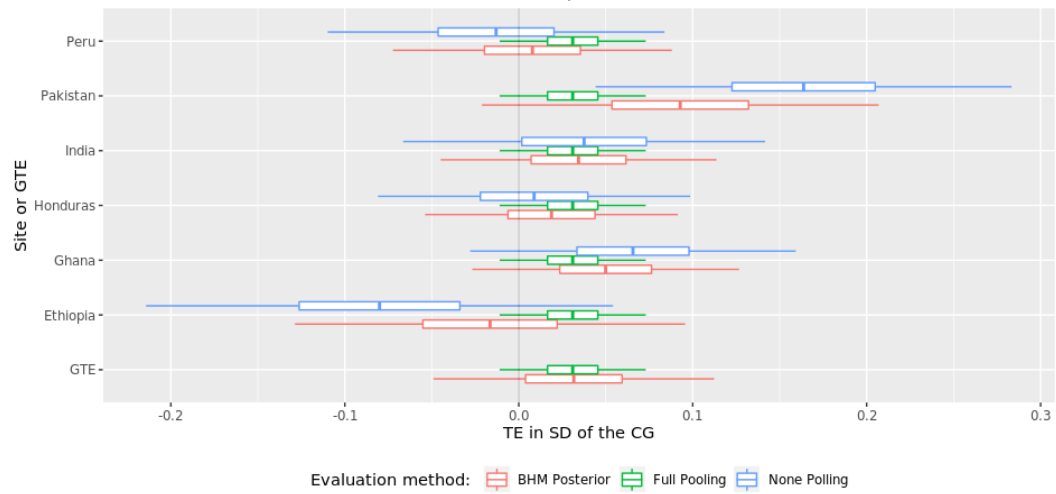


Figure 6b. 27 : dec_health_major_end

GTE and LTE: Partial, None and Full BHM Estimates Comparison

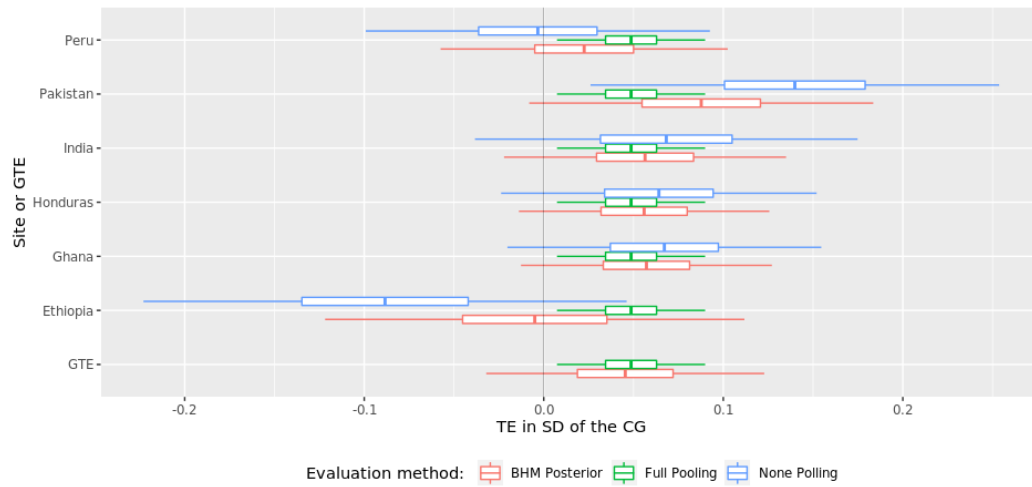


Figure 6b. 28 : asset_index_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

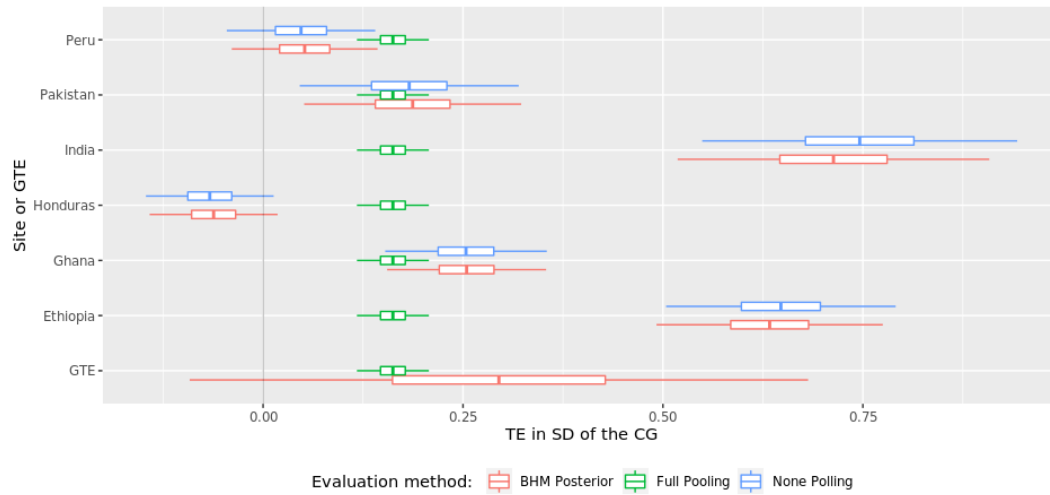


Figure 6b. 29 : asset_prod_index_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

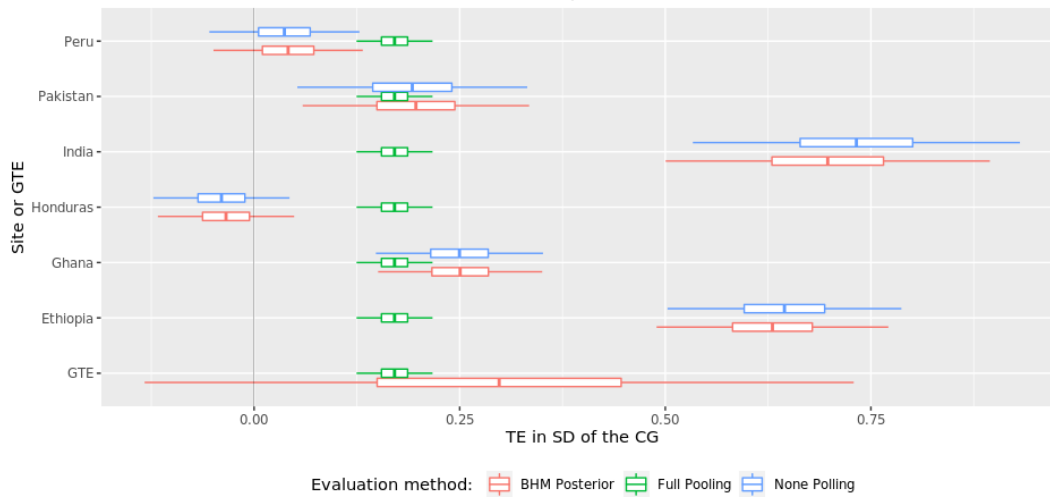


Figure 6b. 30 : asset_hh_index_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

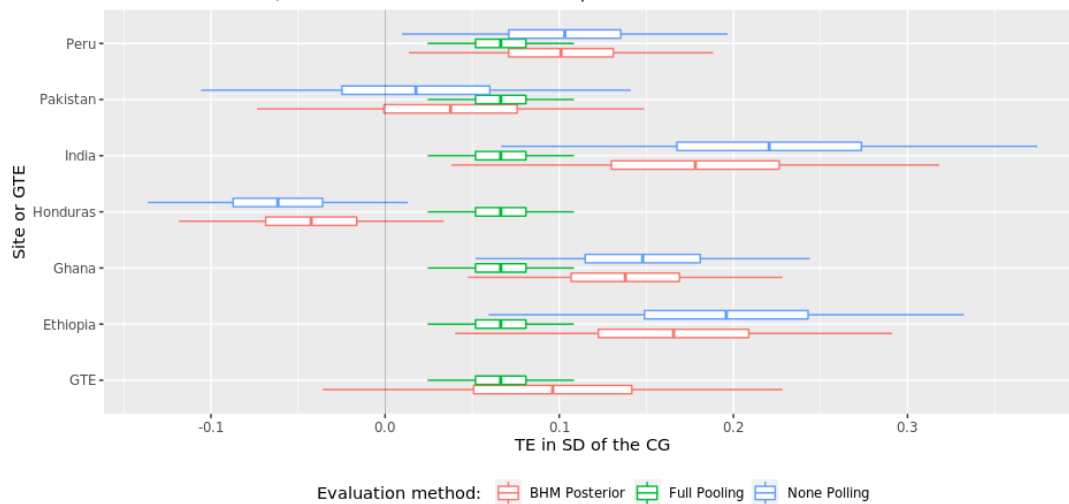


Figure 6b. 31 : ipaidlabor_month_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

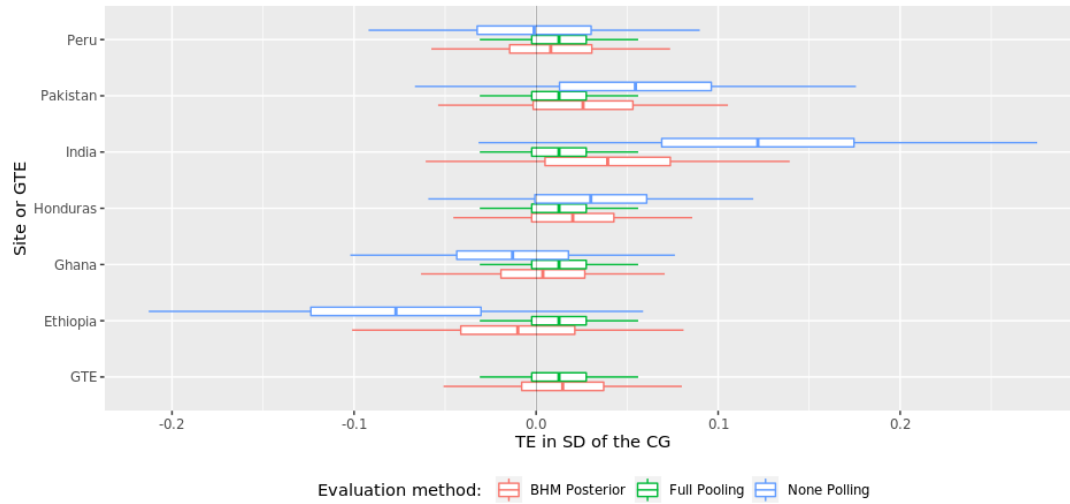


Figure 6b. 32 : ranimals_month_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

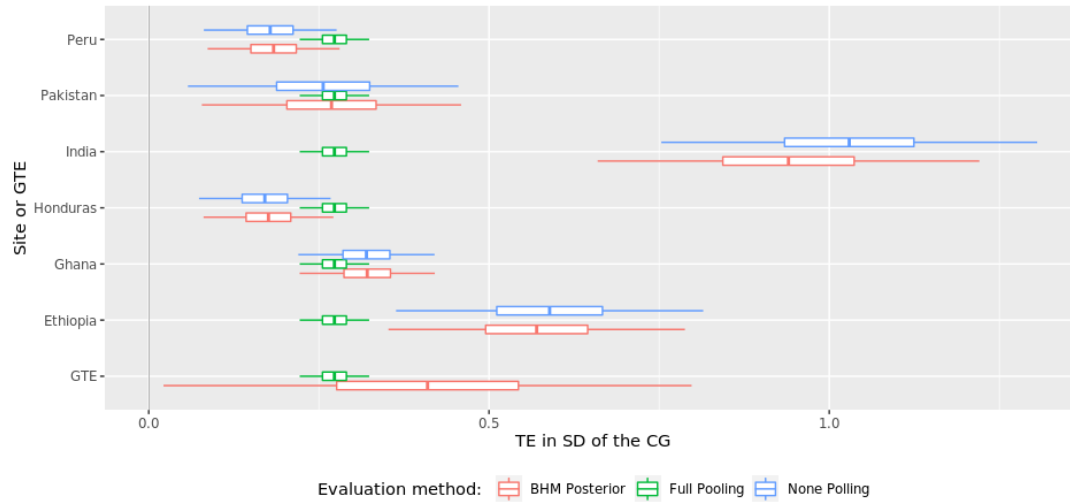


Figure 6b. 33 : ctotat_pcmmonth_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

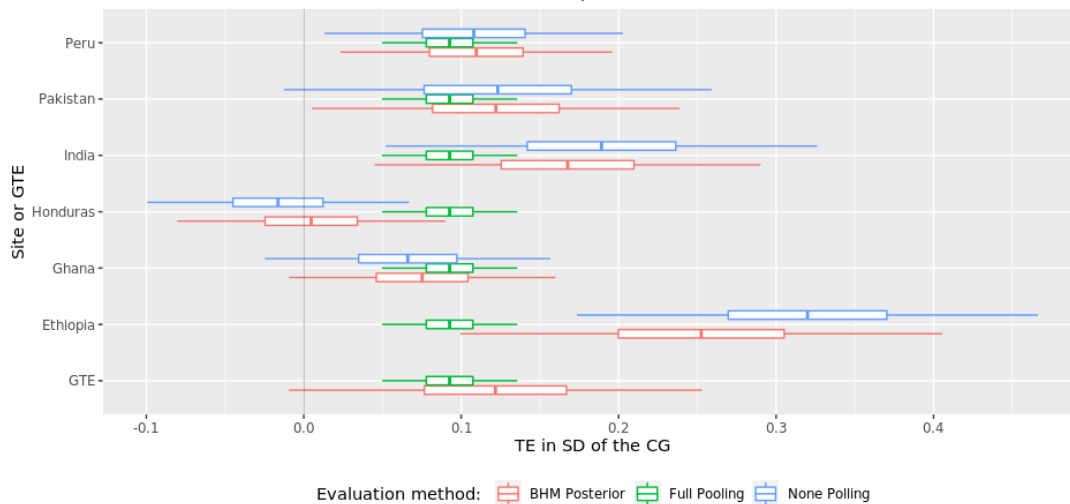


Figure 6b. 34 : cnonfood_pcmmonth_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

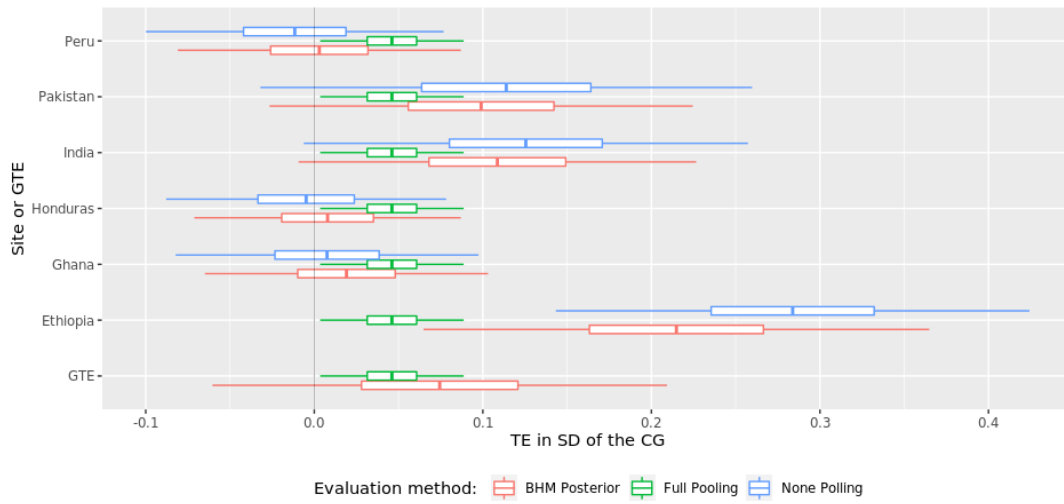


Figure 6b. 35 : cfood_pcmmonth_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

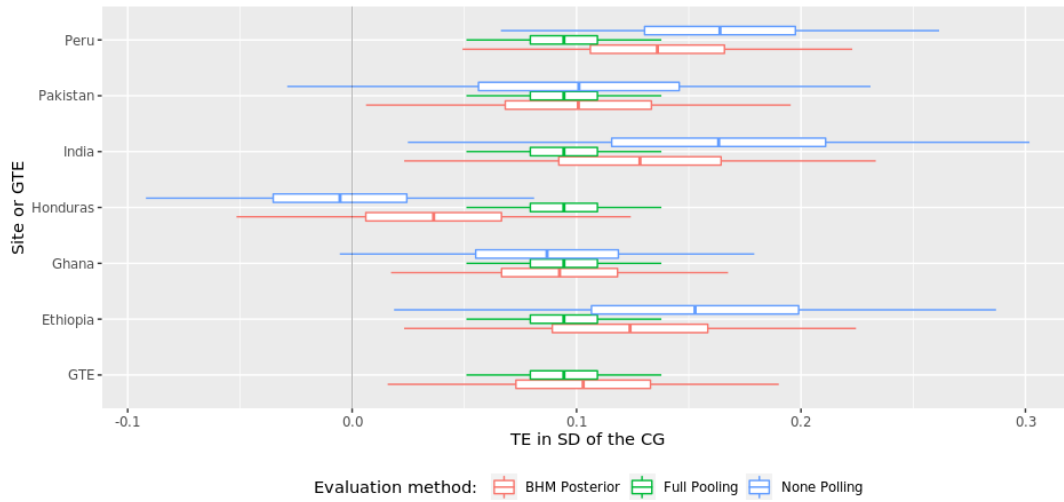


Figure 6b. 36 : cdurable_pcmmonth_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

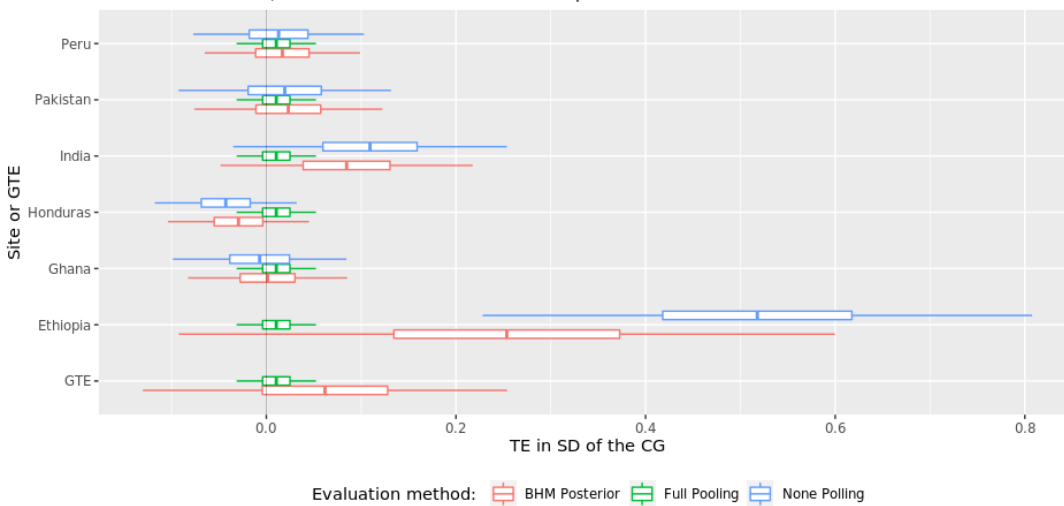


Figure 6b. 37 : index_foodsecurity_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

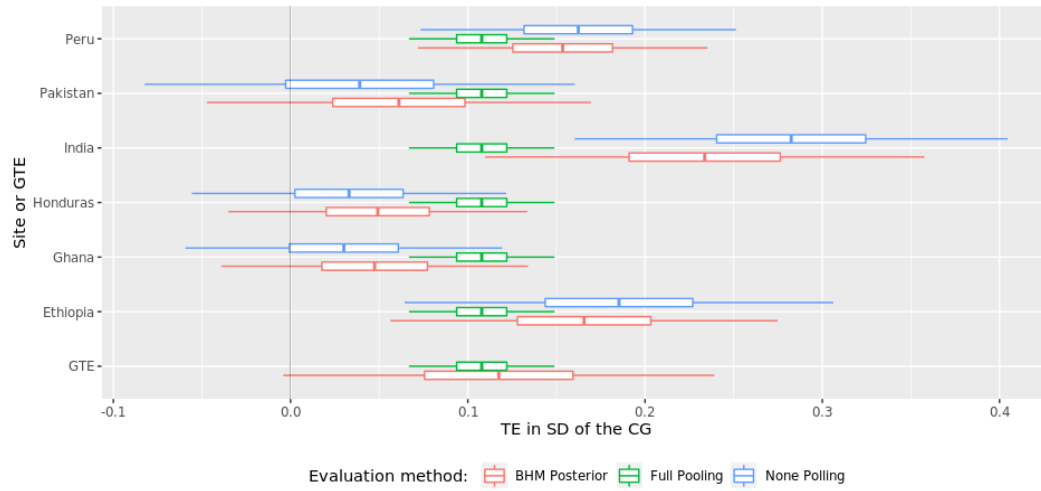


Figure 6b. 38 : fs_adultskip_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

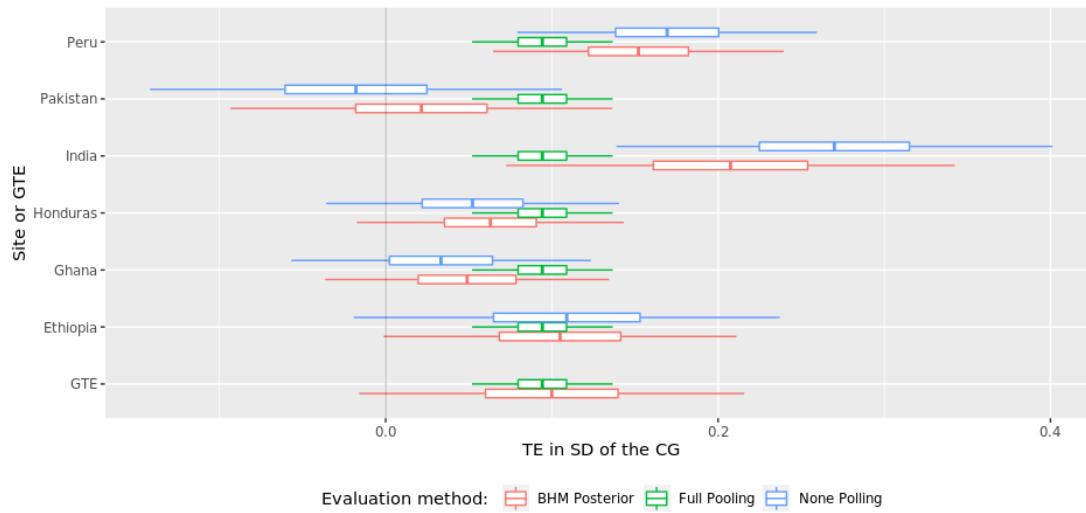


Figure 6b. 39 : fs_wholeday_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

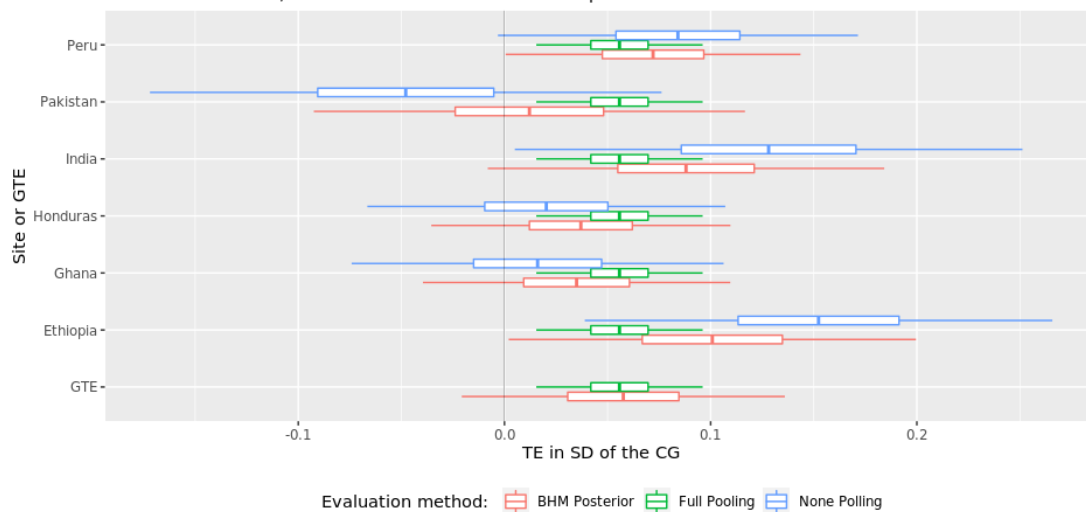


Figure 6b. 40 : fs_childskip_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

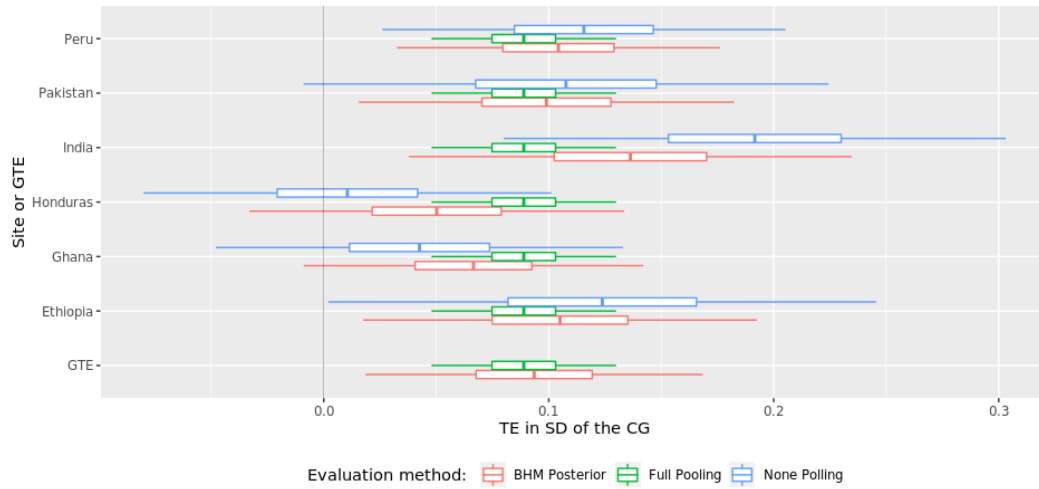


Figure 6b. 41 : loan_totalamt_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

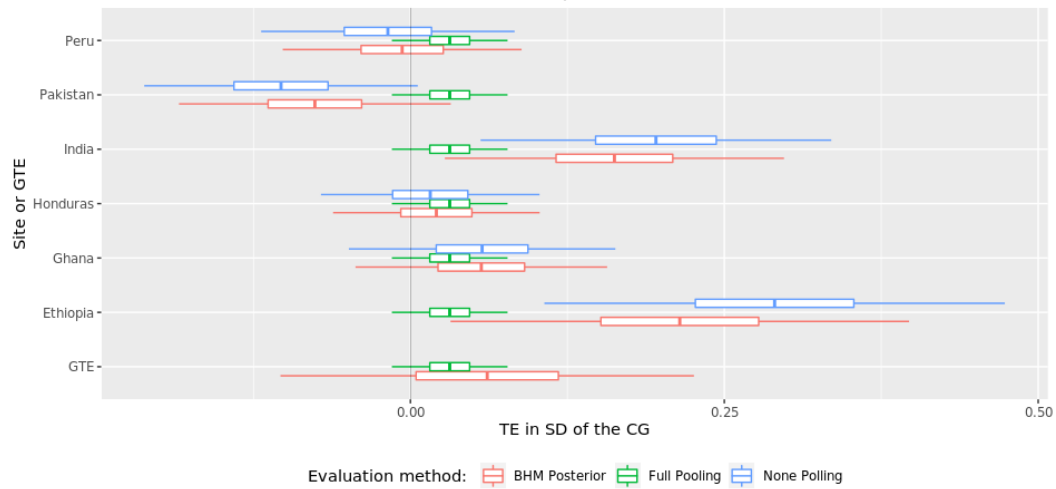


Figure 6b. 42 : index_health_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

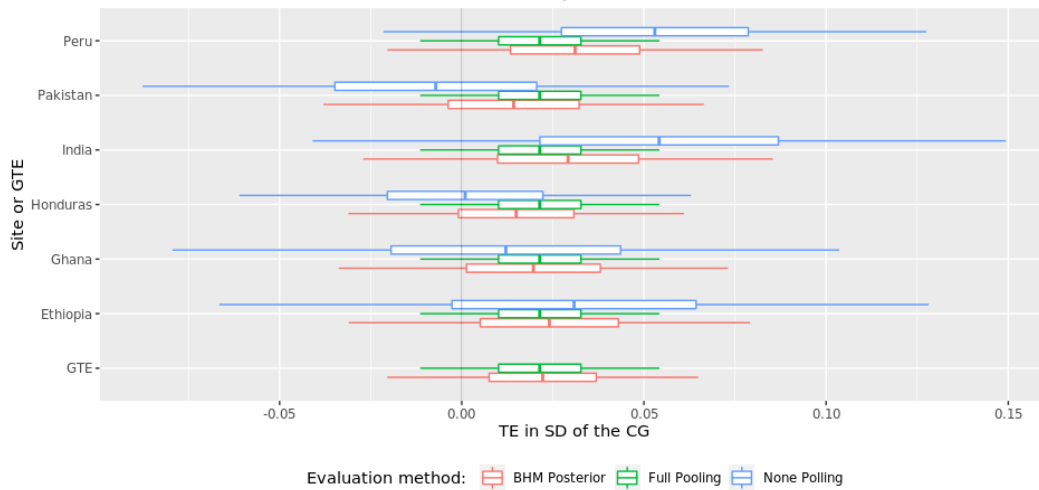


Figure 6b. 43 : ddays_nomiss_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

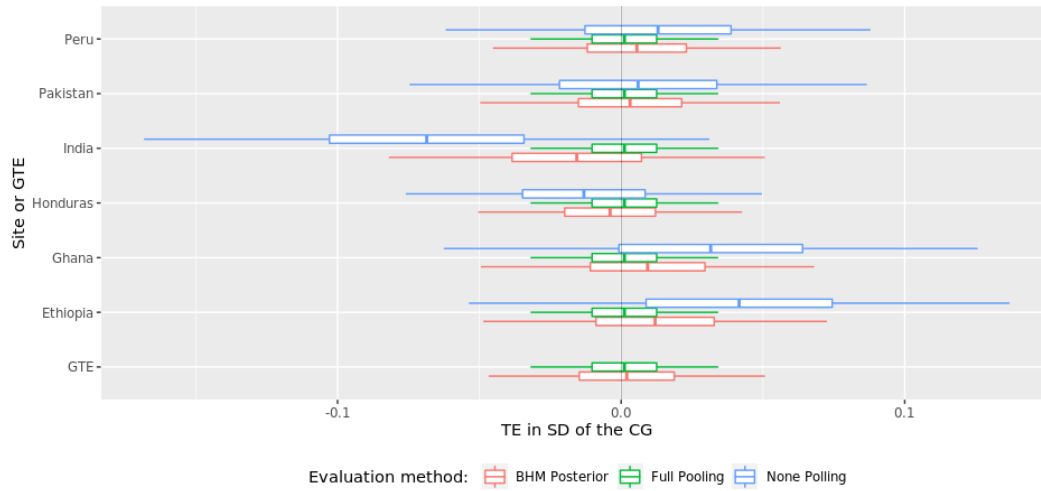


Figure 6b. 44 : index_mentalhealth_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

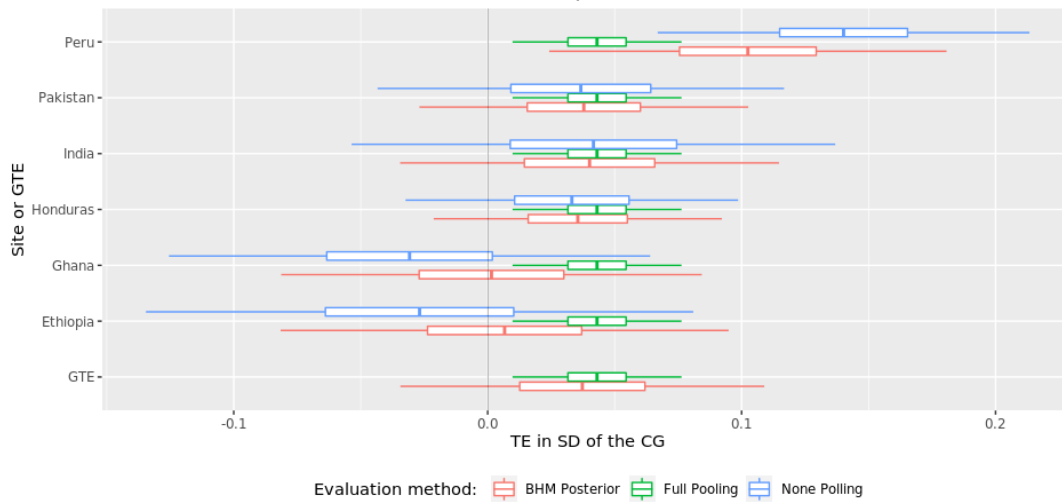


Figure 6b. 45 : percep_life_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

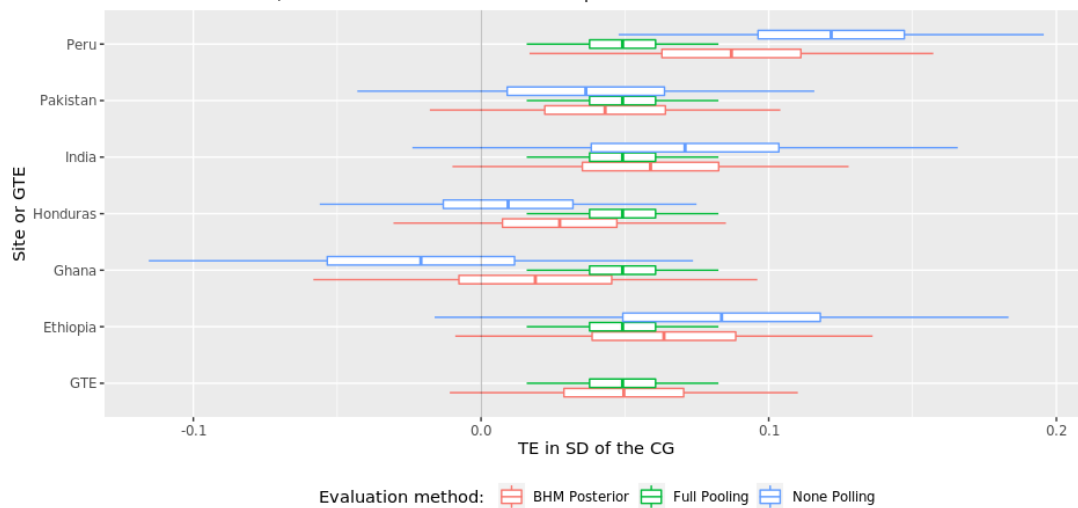


Figure 6b. 46 : time_work_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

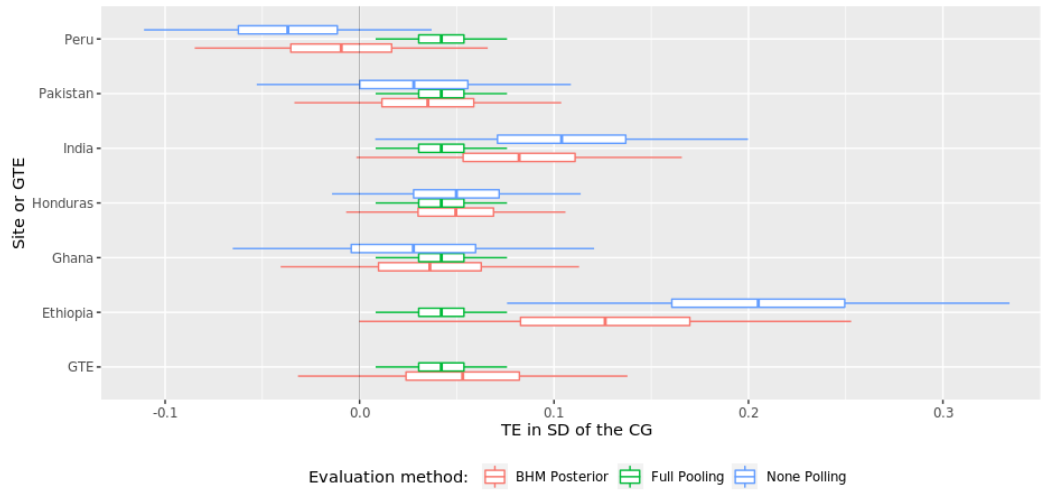


Figure 6b. 47 : time_agri_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

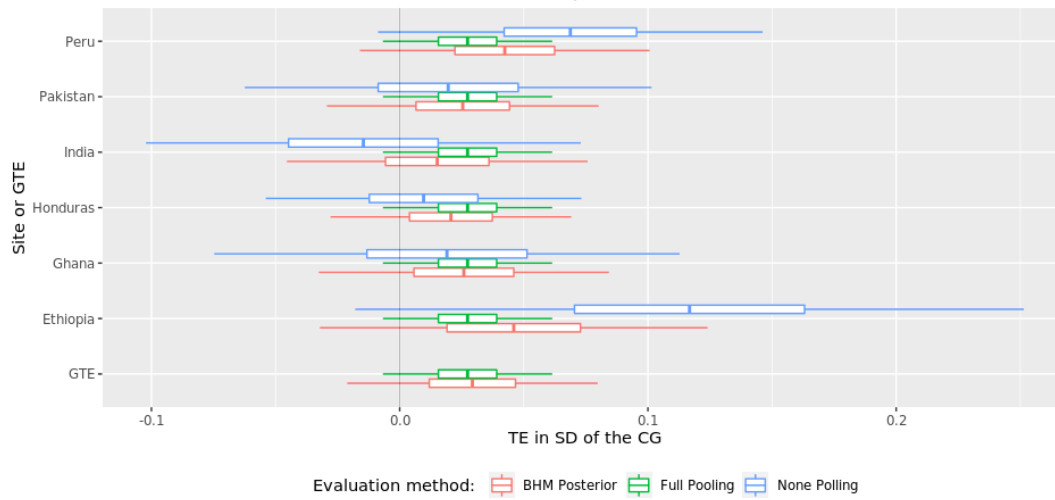


Figure 6b. 48 : time_animals_fup
GTE and LTE: Partial, None and Full BHM Estimates Comparison

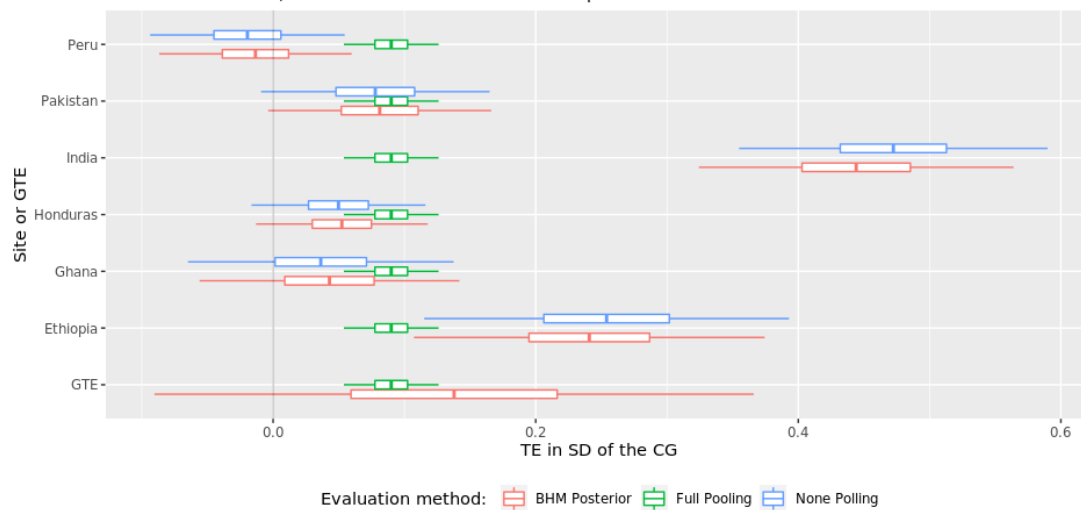


Figure 6b. 49 : time_business_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

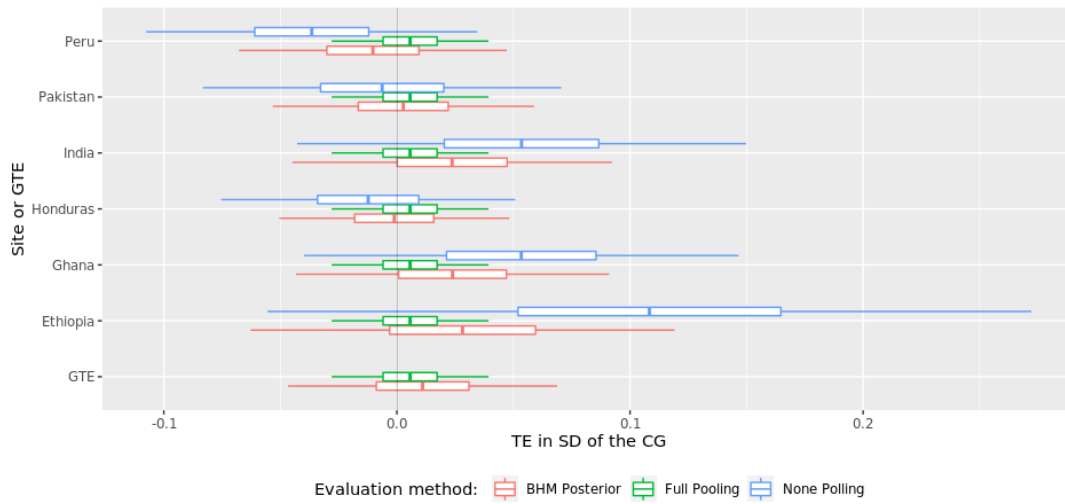


Figure 6b. 50 : time_paidlabor_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

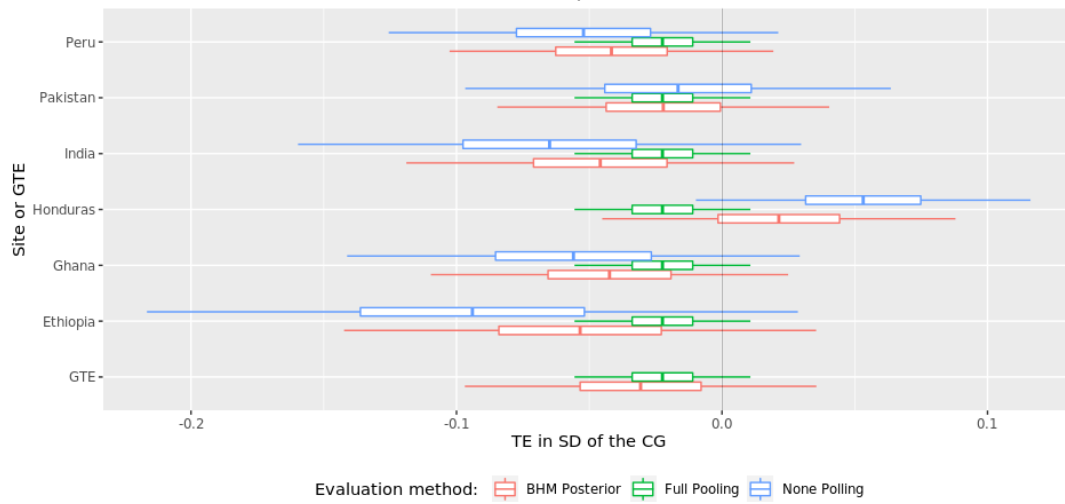


Figure 6b. 51 : index_political_fup

GTE and LTE: Partial, None and Full BHM Estimates Comparison

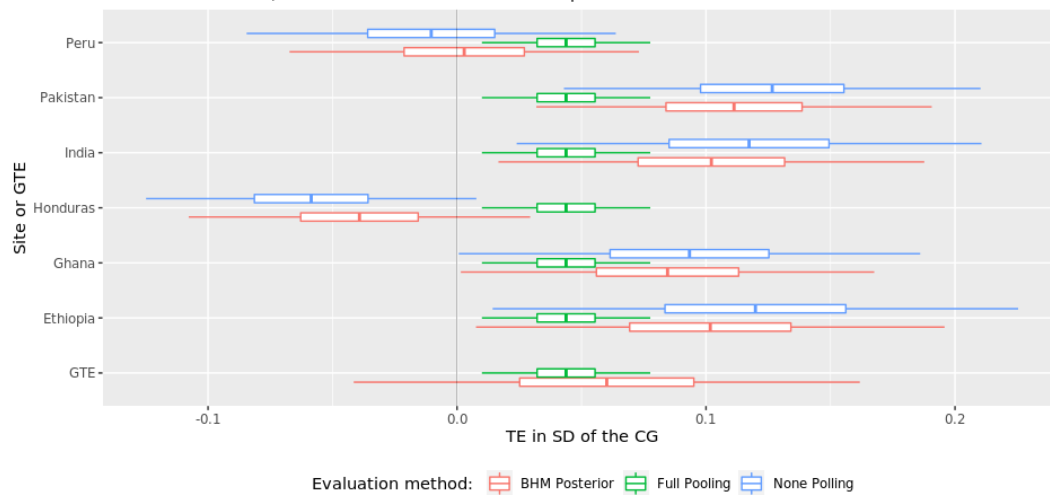


Figure 7: GTE and LTE estimates - pooling comparison

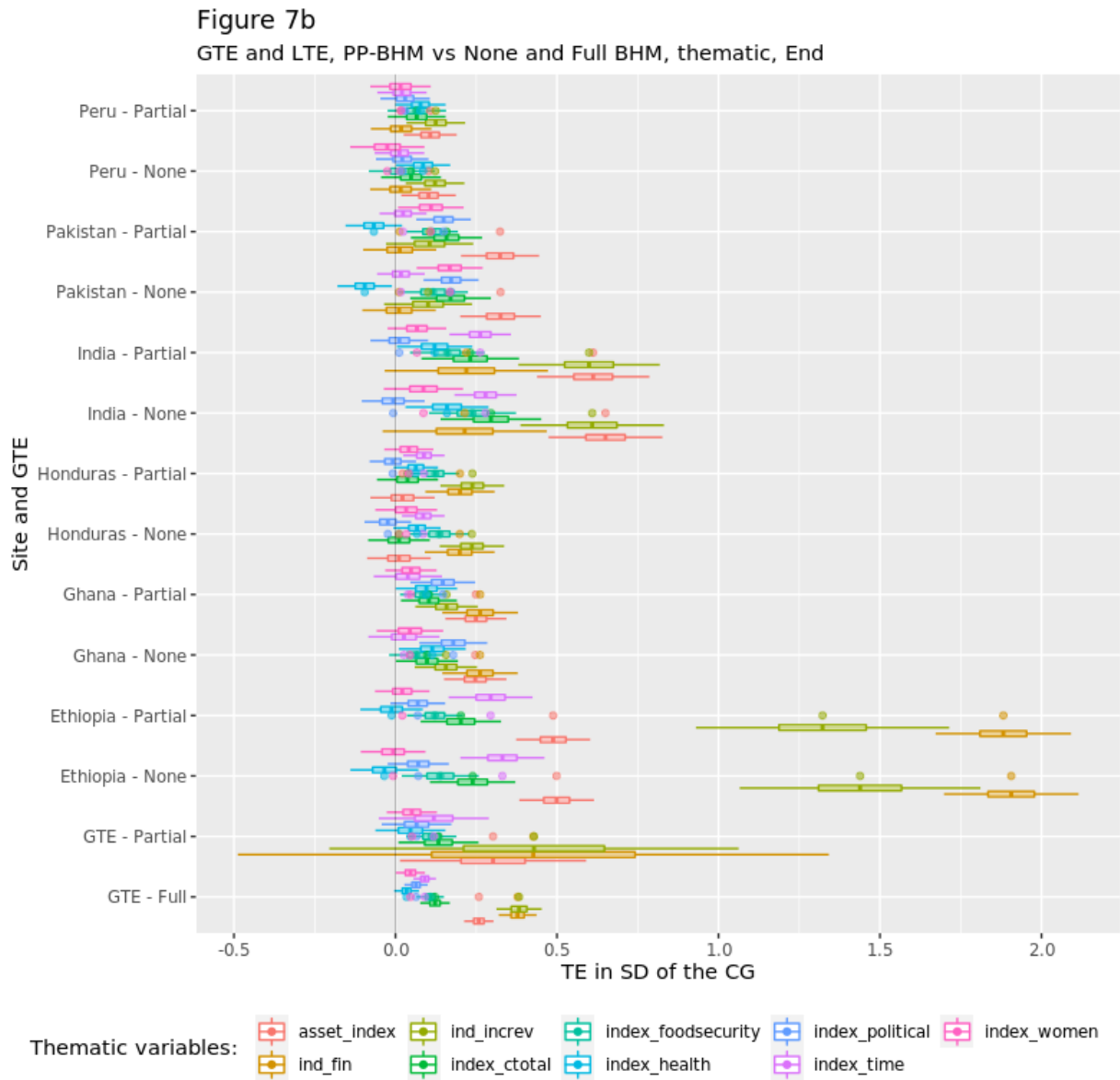
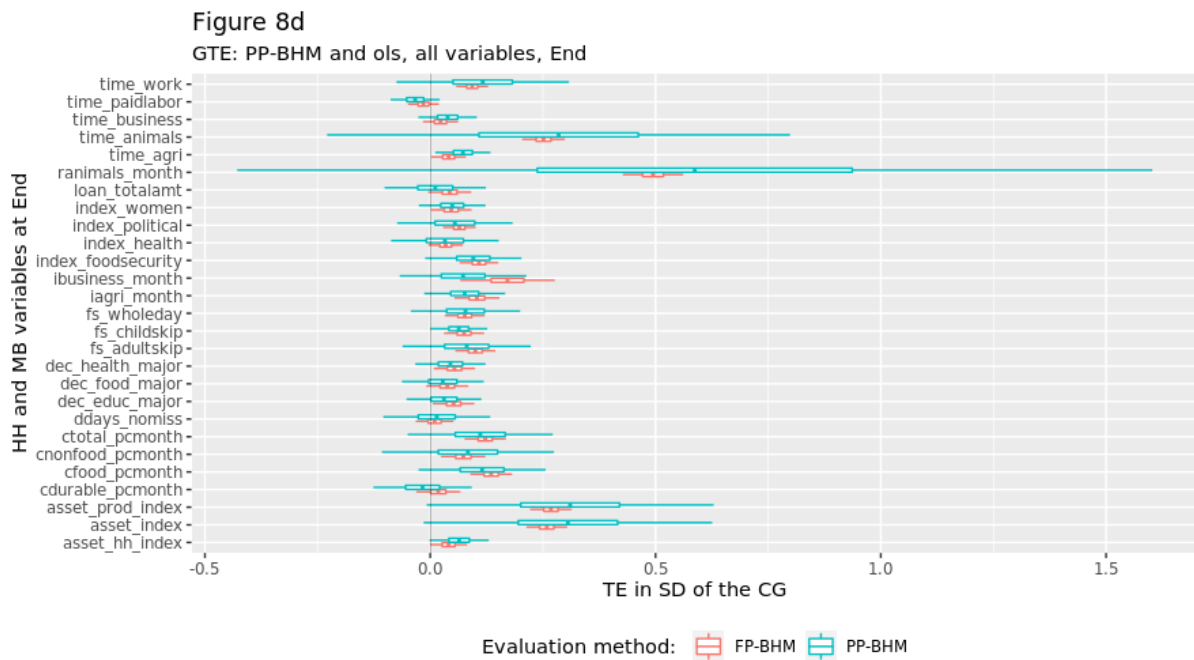
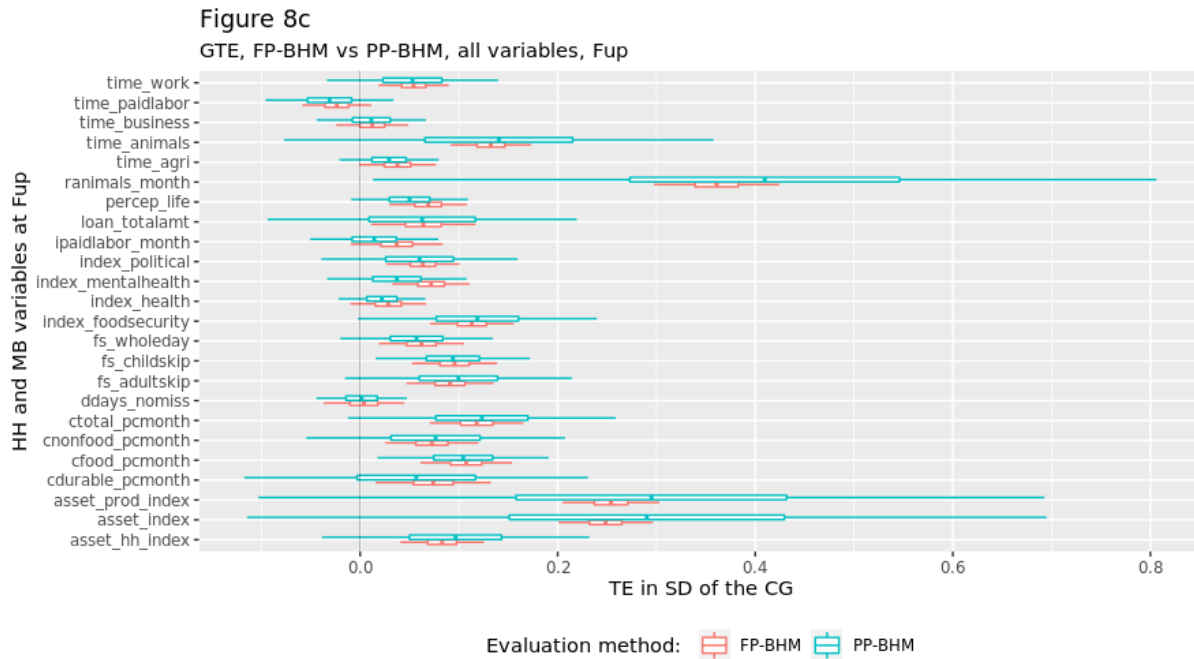


Figure 8: Global Treatment Effect – Full and Partial BHM estimates comparison



Tables 4: Leave-one-out-information-criteria comparison PP to FP

Table 4b : Looic comparison PP to FP		
Individual variables	PP-LOOIC	FP-LOOIC
index_health_end	-2,80	4,38
ddays_nomiss_end	-1,51	-1,21
time_work_end	0,85	50,47
time_agri_end	-13,55	-14,35
time_animals_end	22,73	267,13
time_business_end	-11,98	-14,50
time_paidlabor_end	-15,08	-15,26
index_political_end	-4,94	25,16
index_women_end	-11,31	-12,89
dec_food_major_end	-8,11	-7,89
dec_educ_major_end	-8,78	-10,75
dec_health_major_end	-10,18	-12,45
index_health_fup	-20,50	-22,43
ddays_nomiss_fup	-18,77	-21,16
index_mentalhealth_fup	-9,01	-9,44
percep_life_fup	-13,25	-13,38
time_work_fup	-7,43	-8,42
time_agri_fup	-17,23	-19,40
time_animals_fup	13,13	51,36
time_business_fup	-15,48	-17,56
time_paidlabor_fup	-11,20	-10,62
index_political_fup	-6,06	7,02

Table 4c: Looic comparison PP to FP

Household variables	PP-LOOIC	FP-LOOIC
asset_index_end	6,49	95,06
asset_prod_index_end	6,54	96,09
asset_hh_index_end	-13,62	-14,85
iagri_month_end	-7,29	-8,18
ibusiness_month_end	0,99	0,97
ranimals_month_end	25,03	283,52
ctotal_pcmoenth_end	-1,52	12,29
cnonfood_pcmoenth_end	1,63	18,44
cfood_pcmoenth_end	1,88	-0,43
cdurable_pcmoenth_end	-5,10	-3,57
index_foodsecurity_end	-4,03	-3,30
fs_adultskip_end	-0,78	1,11
fs_wholeday_end	0,36	-0,61
fs_childskip_end	-13,82	-15,56
loan_totalamt_end	-5,28	-7,24
asset_index_fup	8,94	152,81
asset_prod_index_fup	8,58	133,80
asset_hh_index_fup	1,03	17,12
ipaidlabor_month_fup	-13,67	-16,35
ranimals_month_fup	16,03	39,21
ctotal_pcmoenth_fup	-1,37	4,53
cnonfood_pcmoenth_fup	-0,69	0,54
cfood_pcmoenth_fup	-8,21	-7,32
cdurable_pcmoenth_fup	4,73	-2,11
index_foodsecurity_fup	-4,39	2,28
fs_adultskip_fup	-4,68	-1,40
fs_wholeday_fup	-10,35	-11,38
fs_childskip_fup	-10,91	-11,40
loan_totalamt_fup	-0,46	4,59