Bebbo

September 2023

Contents

1 Executive Summary **2**

2 Evaluation Questions **3**
What question does this evaluation answer? . . . 3

3 Study Design **3**
Experiment Design . . . 3
Treatment Condition . . . 4
Control Condition . . . 4
Recruitment . . . 4

4 Descriptives **5**
Respondent Characteristics by Country . . . 5
Construct Variables . . . 5
Pre-Exposure to Bebbo . . . 6
Power Analysis . . . 6
Reliability Analysis . . . 6
Attrition & Survey Behavior . . . 7

5 Results **8**
Regression Model . . . 8
Knowledge and Awareness . . . 8
Confidence and Attitudes . . . 9
Practices . . . 10
Policy Implications of the Results . . . 11

6 User Characteristics Correlated with App Usage **12**

A Baseline Balance **14**

B Additional Plots **15**

C Additional Tables **17**

D Additional Regressions **18**

E Survey Instrument **21**

1 Executive Summary

To support parents to receive timely and quality guidance even when direct contact with service providers is not possible and overcome barriers in access to localized digital solutions with verified content, UNICEF Europe and Central Asia Regional Office (ECARO) developed a mobile parenting app, Bebbo. The mobile application also supports the most vulnerable parents/caregivers with lower education level, in terms of the navigation modalities, off-line operability and selection of the core content. The two main objectives of Bebbo, in line with the UNICEF ECARO Early Childhood Development Theory of Change, are: (1) Improving availability of information for parents on child development, and (2) Supporting parents for responsive caregiving and early intervention. Accordingly, Bebbo app provides users information and interactive tools to help nurture and aid their child's health and development. The launch of Bebbo in 11 countries in the ECA region is a direct response to the identified objective to engage parents and caregivers in nurturing care, positive parenting, stimulating, and learning.

The Context

Parents everywhere are in need of information on various aspects of child development from reliable and validated sources as well as guidance on how to support the health and development of their children. However, services providing this sort of information and support are often non-existent or inaccessible for a lot of parents in many places. Often, service providers, even when accessible, might lack necessary knowledge and skills to respond to the questions and concerns parents might have.

Mobile apps are one of the most convenient and easy ways to access information about child development and parenting. However, parenting apps are mainly in English and provide a limited thematic content without a possibility for parents to familiarize with, track, and support all aspects of their child's health and development. In addition, these apps are, naturally, not adapted to contexts of individual countries. Many apps are not free of charge, which presents a significant barrier, particularly for the most vulnerable families. At the same time, the majority of the existing apps operate only in online mode requiring good internet connectivity that is lacking in remote and rural areas.

The App

[UNICEF TEAM TO PROVIDE MORE INFORMATION ABOUT THE APP IF DESIRED]

Impact Evaluation

We perform a study across two countries, Serbia and Bulgaria, using a randomized encouragement design to compare the impact of encouraging caregivers of young children to use the Bebbo app as compared to a treatment-as-usual (TAU) condition of encouraging them to use a static informational website. By comparing Bebbo to the existing TAU, we are asking the question: "does this new treatment offer something above and beyond the already existing treatments which parents might presumably already be asked to do?"

We measure effects on eight outcomes across three domains: knowledge, attitudes, and practices. A difference-in-difference design is used, with questions asked at both the baseline (before treatment) and the endline (at least 4 weeks after treatment) surveys. Finally, an additional follow-up survey was sent (at least 4 weeks after the endline), to measure impacts of longer-term usage.

Results

We do not find evidence that asking this population to use Bebbo has any impact beyond that of asking them to visit a parenting website. Given the study design, the following facts may have contributed to the lack of evidence of impact:

1. The population was already very "good" in regards to the outcomes of interest. We measured the improvement of parents over time, under treatment, but many could not improve from their baseline scores, which were perfect. This implies that either (i) the outcome questions were not the right questions or (ii) the problem being solved only exists in a subset of the population and measuring the impact on the population as a whole might be difficult.

2. Participants improved from the first questionnaire to the second questionnaire, regardless of treatment arm and regardless of compliance. This seems to imply that the very act of asking the questions improves the way that parents answer them. This implies that reminding parents about the questions, via an informational campaign, may be more important than providing resources to find the answers, which they seem to already have.

3. Very few people complied with treatment and used Bebbo. Of those who were asked to use the app, 28% used the app, 12% used the app more than one day, and only 3% used the app more than three days. Significant pre-exposure to the app in both countries (55% knew about the app and 23% reported having used it before) could have led to the low initial compliance. This implies that the app may only be an effective intervention for a small subset of the population who finds it engaging. [ADD SENTENCE ABOUT APP USAGE ANALYSIS].

We were reasonably powered (70%) to find a small effect size on the population or a medium effect size on the treated with a 28% takeup and no effect on those who don't take up the treatment. Note, this implies that we were not able to detect a smaller effect size or a medium effect size on a smaller takeup group (i.e. the small group of treated users who used the app for multiple days).

2 Evaluation Questions

What question does this evaluation answer?

The design of the study is set up to answer the following question in the positive:

> Is asking parents to use Bebbo an effective policy to improve the parenting knowledge, attitudes, and practices of the general population of caregivers of young children in Bulgaria and Serbia?

Note that the study cannot fully answer the question in the negative, it cannot prove that this intervention is ineffective, it can only fail to measure its effectiveness.

We are only testing the effectivess of "asking" or "inviting" parents to use the app. Alternatively, one might be interested in testing the effectiveness "incentivizing" or "forcing" parents to use the app, but we are not doing that in this evaluation. We consider "an effective policy" one which performs better than the "treatment as usual" case, which we will consider to be an existing, static website. Finally, we are studying the general population of caregivers of young children. No particular care was given to single out any particular subset of the population that might benefit the most (or the least) from Bebbo, nor those who would be most likely to use Bebbo.

3 Study Design

Experiment Design

This study follows a prepost design (Clifford, Sheagley, and Piston 2021) in which we measure the outcomes of interest before treatment (in a baseline survey) and after treatment (in an endline survey). We add an additional survey after the endline, referred to as a follow up, to look for longer-term impacts and test the impact of continued app usage.

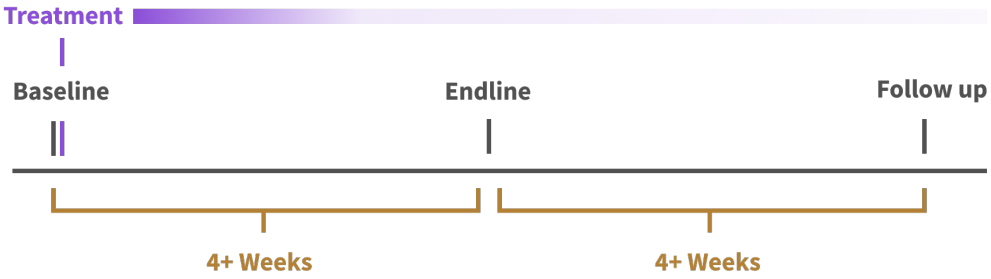


Figure 1: Study Design

Study participants are randomized, from the beginning, to one of two conditions:

1. **Treatment.** Participants in the treatment condition were told that there was one more step to qualify for the study and were then asked to download the app Bebbo and use it regularly, being encouraged that doing so will help them with their parenting.

2. **Control.** Participants in the control condition were told that there was one more step to qualify for the study and were then asked to visit a parenting website and use it regularly, being encouraged that doing so will help them with their parenting.

This follows a randomized encouragement design (Moayyedi and Hunt 2014), as participants were asked to participate in the treatment, but it was not forced, thus leading to takeup that is less than 100%. A randomized encouragement design is used here because:

1. We are interested in the impact of a treatment on a population where individuals can choose whether or not to take the treatment (the "compliers").
2. The compliers and non-compliers might have different reactions to the treatment.

Treatment Condition

Participants were sent the following message at the end of the baseline survey:

> There is just one more step to qualify for the Visa gift card of X. Please download Bebbo, the free parenting app, and discover how it can help you. Using Bebbo regularly can improve your interactions with your children and help you support their development better! You can do so by clicking the link below:

Clicking on the link led them to the Bebbo app page where they were invited to download the app via the app stores (Google or Apple). App usage in the treatment group was tracked via tracking ids sent with the link to the app download page, allowing us to follow the app usage of each individual treatment participant and measure takeup. If someone decided to ignore the page, and instead went on their own to search for and download Bebbo, we would not have data on their usage. Thus, usage data and takeup should be considered a lower bound.

[UNICEF TO ADD MORE INFORMATION ABOUT THE APP AND HOW IT WORKS]

We collected all app usage events. For the sake of our study, we were interested in a subset of events that represented the accessing of content or features that contained information or might impact their behavior.

Control Condition

Participants were sent the following message at the end of the baseline survey:

> There is just one more step to qualify for the Visa gift card of X. Please visit the following free parenting website and discover how it can help you. Using this website regularly your interactions with your children and help you support their development better! You can do so by clicking the link below:

The choice to use a website as a treatment-as-usual (TAU) condition was decided by the evaluation and program team because it represented an alternative (and traditional/existing) way to solve the problem that the Bebbo app was trying to solve. Another option that was considered was to use an alternate parenting app, but the team believed that using a website gave the best chance to detect a difference in the use of an app, rather than the specific implementation of the Bebbo app. Similarly, several websites were considered, and the most basic website was chosen so as to be "static" - without interactive features - so that it acted as an informational resource rather than a web app or platform which would similarly overlap with the concepts behind the Bebbo app.

The website chosen or Bulgaria was 9meseca.bg. Due to a mistake in the implementation, no website was chosen for Serbia and the Bulgarian website was sent to participants in both countries. This implies that for Bulgaria, participants were provided a reasonable alternative to the app. However, in Serbia, they were sent a Bulgarian website, which would be expected to be suboptimal for the particpants. Similarity of results across countries shows that the choice of the treatment-as-usual website did not materially affect results, as discussed in the sections on results.

Recruitment

Participants were recruited to the study with social media ads on the Meta platform (Facebook and Instagram) using the Virtual Lab platform to create and run the recruitment ads. The Virtual Lab platform is used to track and measure the price-per-respondent across multiple strata, solving the core problem of monitoring, computing expectations, and adjusting budget when recruiting via social media platforms. In this study, recruitment was not stratified, due to initial budget pressures when stratifying in the initial pilot.

In exchange for participating in the study, participants were told they could receive gift cards worth up to 12 USD (in their local currency). See figure 2 for examples of the ad material used for recruiting. Recruitment and survey administration was performed on a rolling basis between March and October, 2023. Each individual

participant was treated at the end of the baseline survey and sent the endline survey 4 weeks after completing the baseline survey.

The survey was administered via a chatbot in Facebook Messenger, using the Virtual Lab platform. Respondents who clicked on the advertisements were directed to a Messenger chat with the Virtual Lab Facebook page, which did not contain any content or information related to this study. Consent was provided via chat, as well as all answers to the survey questions and the treatment condition. Gift cards were also provided via chat, using the Tremendous gift card platform to provide Visa international prepaid cards. The Virtual Lab chatbot allowed the researchers to create multi-wave surveys, with independent timing. It additionally allowed the easy provision of gift cards at the end of each wave, which is integrated into the survey directly via the platform.

[TODO: add recruitment stats]

Figure 2: Recruitment Ads

4 Descriptives

Respondent Characteristics by Country

Table 1 provides the baseline characteristics of the respondent population, separated by country.

Generally speaking, most respondents were themselves parents (not grandparents or other caregives), women, under 35 years of age, and spoke the dominant language of the country at home. A little over half had children 0-2, compared to 2-6 years of age. Respondents in Bulgaria were more likely to have a university education (42%) compared to those in Serbia (29%).

Table 1: Baseline Respondent Characteristics

Variable	Value	Bulgaria	Bulgaria %	Serbia	Serbia %
Is Woman	1	1418	0.83	2102	0.80
University Educated	1	725	0.42	748	0.29
Speaks Dominant Lang.	1	1571	0.92	2488	0.95
Is Parent	1	1485	0.87	2374	0.91
Child Age	2-6	935	0.55	1561	0.60
Num. Children	4+	70	0.04	279	0.11
Parent Age	Over 35	365	0.21	550	0.21
Urban Area	1	1059	0.62	941	0.36

Construct Variables

The outcomes of interest consist of eight constructs divided into three domains: knowledge and awareness, confidence and attitudes, and practices. The mapping between the constructs, domains, and questions that make up the constructs are laid out in table 32.

The constructs “Vaccine Knowledge”, “Parenting Confidence”, and “Breastfed” are made up of only one question. The construct “Activities Past 24h” consists of a count of the number of activities, within the

previous 24 hours, that the respondent has done. The construct "Child Dev. Knowledge" consists of a series of true/false questions, which are averaged based on whether or not the respondent answered correctly. The rest of the constructs are created by averaging of a set of likert variables.

Descriptive statistics regarding the baseline responses for the outcomes are shown in table 2. Note that many of the constructs have quite high means and medians and some have a high proportion of respondents with the max score. In particular, 73% and 72% of respondents scored perfectly on the knowledge questions. This is problematic, as knowledge is often considered the easiest to change quickly and was a core outcome of interest for the team. Additionally, knowledge questions seem to be heavily impacted by the repeated survey effect, as discussed further down.

Table 2: Outcome Construct Descriptives Pooled Baseline

name	mean	median	min	max	sd	prop_max	prop_na
Activities Past 24h	4.92	5.0	0	6	1.23	0.41	0.00
Parenting Confidence	3.34	3.5	1	4	0.65	0.37	0.00
Positive Practices	3.20	3.5	1	4	0.75	0.27	0.00
Attitude to Phys. Punishment	3.13	3.0	1	4	0.84	0.36	0.00
Hostile Practices	3.04	3.0	1	4	0.69	0.15	0.00
Child Dev. Knowledge	0.86	1.0	0	1	0.28	0.73	0.00
Vaccine Knowledge	0.72	1.0	0	1	0.45	0.72	0.58
Breastfed	0.37	0.0	0	1	0.48	0.37	0.58

Pre-Exposure to Bebbo

This study recruited Serbian and Bulgarian caregivers online, via social media ads, and invited half of them to download the app Bebbo. What if some people were already familiar with the app? Or had already downloaded and used it before?

If someone had already downloaded the app and still had it on their phone, we would not be able to track their usage and they would be considered "non-compliant" in this design. This is desireable from an analysis perspective, as these people are "always-takers" (Imbens and Rubin 2015) who would have the app regardless of whether they were assigned the treatment or control condition.

Many other people, however, might have decided not to download the app because they had heard about it before or tried it out before and deleted it. This is a concern for the study because these people might have already gotten use out of Bebbo: they could have used it and learned everything there is to learn from the app already.

To check for such "pre-exposure," we ask control group users, at the end of the final follow up survey, if they have ever heard of Bebbo or used Bebbo.

55% of respondents said that they had heard about the app Bebbo and 23% said that they had downloaded and used the app Bebbo. It's worth noting that there might be some social desireability bias or acquiesence bias (Stantcheva 2023) in these responses and we do not have a good way to detect that in this instance. However, despite those potential biases, this is strong suggestive evidence that there was pre-exposure to the treatment in our sample.

Power Analysis

Post-hoc power analysis was performed to see the ability to detect an effect, in terms of standardized deviations (corresponding to Coen's D effect sizes), in the datasets analyzed. To create the effect size, the standardized different is multiplied by the empircal takeup of 28%, which was the percentage of participants that had at least one learning event in the treatment group.

The results show that the study is reasonably powered (70%) to detect a medium effect size on the 28% takeup at a significance level of 1.25%, the equivalent of 10% when controlling for multiple testing (8 outcomes) with a Bonferroni correction. See figure 5.

Reliability Analysis

The outcomes consist of "constructs," some of which combine the answers to multiple questions into one value. The theory is that these questions are measuring the same underlying construct and that the reliability of the construct is increased by combining multiple answers.

We test this assumption, that they are measuring the same underlying construct, by looking for internal consistency using Chronbach's alpha within the variables associated with each construct. Note that all constructs are composed of either Likert scale variables or Binary scale variables and not both, making this analysis valid.

This technique was used to finalize the construct/variable mapping after the data collection completed but before the analysis began, as some of the constructs had a lower internal consistency than hoped. Table 3 summarizes raw and standardized alpha of each construct as used in the final analysis, along with the number of variables in it.

Constructs witha reliability above 0.70 are considered internally consistent. All the constructs were modified to ensure higher reliability (sometimes dropping to one variable). The construct created from Activities in the Past 24 hours had the lowest reliability, possibly because some of the activities might be negatively correlated. Because of this, we decided to use the sum of the variables rather than the mean, removing any concern of internal consistency and rendering the low Chronbach's alpha irrelevant.

Table 3: Reliability: Alpha Matrix

construct	variable count	raw.alpha	std.alpha
Vaccine Knowledge	1		
Child Dev. Knowledge	4	0.82	0.82
Parenting Confidence	2	0.75	0.76
Attitude to Phys. Punishment	1		
Breastfed	1		
Activities Past 24h	6	0.53	0.53
Positive Practices	4	0.8	0.8
Hostile Practices	4	0.72	0.72

Attrition & Survey Behavior

As an online sudy, attrition was generally high: 52% of those who started the survey dropped off before completing it and 54% never came back from the baseline to complete the endline. Table **??** summarizes attrition by stage and treatment condition. It's worth noting that attrition was consistently higher among the treatment group, possibly related to the increased number of questions in the endline survey for that group (additional questions about app usage were added for the treated).

Attrition was particularly high between endline and follow-up survey (66%) but that was primarily driven by a problem in the implementation: due to a mistake in survey coding, the questions asked to the control and treatment group was switched at endline, which informed the control group about the existence of the Bebbo app, potentially contaminating them as a pure control. While high pre-existing awareness was discovered in all groups, even those without this mixup, we have removed all cohorts who experienced the mixup from the follow-up survey analysis to avoid any potential issues.

Table 4: Attrition: Pooled

stage	count	attrition	treated_attrition	control_attrition	attrition_dif
Started Baseline	8994				
Finished Baseline	4321	0.52	0.52	0.52	0.00
Started Endline	1968	0.54	0.54	0.55	0.00
Finished Endline	1679	0.15	0.17	0.13	0.04
Started Followup	569	0.66	0.66	0.66	0.00
Finished Followup	412	0.28	0.29	0.26	0.03

Note that respondents should have been contacted 4 weeks after each wave in order to take the subsequent wave. However, two factors may lead to them not always started the wave after exactly 4 weeks: (i) there were some technical issues which caused the notification to be delayed in some cases and (ii) not everyone begins the survey immediately when notified and maybe need to be reminded several times, or may remember on their own, significantly later.

To improve consistency of the study, we removed anyone who took the endline or followup surveys more than 9 weeks after their previous survey, ensuring that all respondents were responding in a gap between 4-9

weeks. Table 5 summarizes the distribution of this time gap. As you can see, the vast majority (more than 80%) took the survey after 4-5 weeks of the previous survey.

Table 5: Time Gap Descriptives

Time Gap	min	quantile_05	median	quantile_80	max
Baseline - Endline	26 days	28 days	29 days	35 days	63 days
Endline - Followup	28 days	28 days	29 days	33 days	62 days

5 Results

Regression Model

We run the following regression model to measure the intent-to-treat effect (ITT) of assignment to the treatment arm:

$$y_i - y_i^b = \gamma_1 + \beta T_i + \gamma_2 X_i + \epsilon_i$$

Where y_i represents the outcome of interest for individual i measured after treatment, T_i represents the random treatment assignment, X_i a set of control variables and y_i^b represents the outcome of interest measured before treatment. The parameter of interest will be the treatment effect, β.

Note that due to the relatively large number of sepearate outcomes (8), we adjust p-values of the treatment variable to control the false discovery rate (FDR), using Benjamini-Hochberg, reported as the "Adjusted Treatment p-value."

Of interest in this analysis is also the treatment effect on the treated (ToT), which can be estimated using an instrumental variable model given the monotonicity assumption of treatment (Imbens and Rubin 2015), which assumes that people are not less likely to download and use the Bebbo app in the treatment group. To estimate our instrumental variable model, we use 2-stage least squares:

$$\begin{aligned} y_i - y_i^b &= \gamma_1 + \beta \hat{z}_i + \gamma_2 X_i + \epsilon_i \\ z_i &= \gamma_3 + \gamma_4 T_i + \gamma_5 X_i + \delta_i \end{aligned}$$

Where z_i is a binary indicator of takeup based on the recorded app-usage data and $\hat{z}_i$ the predicted takeup based on the first stage regression. Once again, parameter of interest is β. It's worth noting that, given that we cannot say we measured any impact (results are not significant from zero) in the ITT model, the exact values of the ToT model are of less interest and the associated tables can be found in the appendix.

We also run the regression for two separate time periods: endline and follow-up. However, due to large attrition in the follow-up survey and the low long-term app takeup, we are underpowered in our analysis. The results of the follow up regressions can also be found in the appendix.

One of the dangers of a prepost design is that you are priming your respondents with the first survey and that priming may impact how they answer the questions in the post-treatment survey(s) (Stantcheva 2023). Given this particular study design, where our control is a "treatment as usual" (TAU) that involved sharing a website and we do not have data regarding the takeup, or usage, of the website, it is difficult to isolate a priming effect.

We will also plot raw charts showing mean scores at baseline and endline for three groups for each variable: control, treatment with takeup (those who we know downloaded and used the app), treatment without takeup (those for whom we have no data showing they downloaded or used the app). These plots can provide suggestive evidence of priming effects by showing the shift in mean between baseline and endline across all three groups.

Knowledge and Awareness

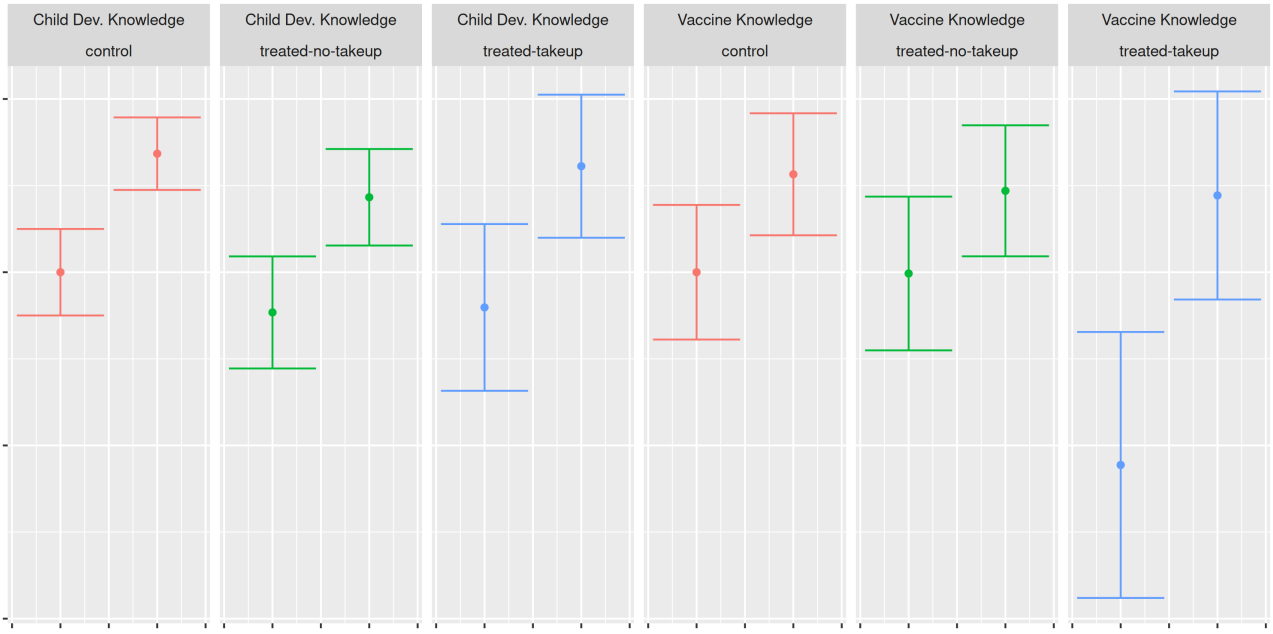
Regression analysis of these outcome constructs show no significant result of treatment:

These two constructs, Vaccine Knowledge and Child Development Knowledge, both suffered from ceiling effects in the baseline survey (72% and 73% respectively). On top of those ceiling effets, they both potentially suffered from priming effects, as evidenced by the consistent improvement in the endline survey for all groups.

Note that there is some suggestive evidence that those with less vaccine knowledge were more likely to download the app, indicating that takeup might be biased towards those who need it the most. That might be driving the small and statistically insignificant positive measured impact on Vaccine Knowledge in the regression. Unfortunately, the study was not designed for subgroup analysis on a small group such as the 28% who failed the vaccine knowledge question at baseline.

Table 6: Pooled: OLS - Endline - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Treatment	0.04	−0.001
	(0.03)	(0.01)
Adjusted Treatment p-value	0.467	0.957
Observations	667	1,811
R^2	0.01	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01



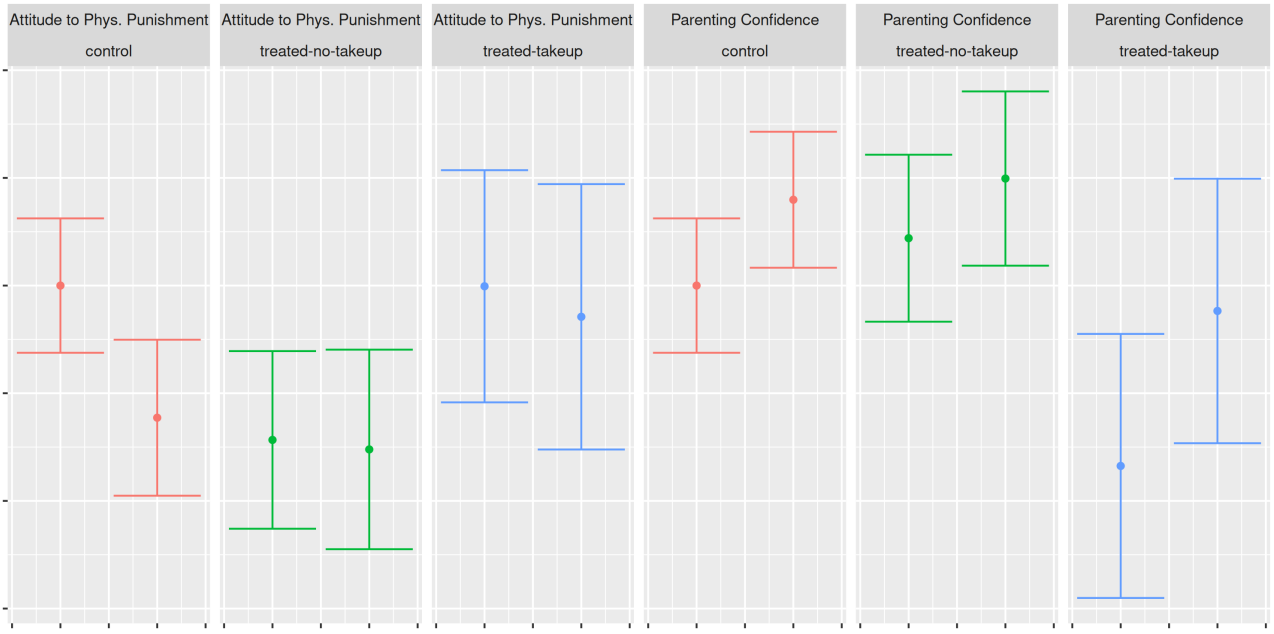
Confidence and Attitudes

Attitude Towards Physical punishment is a single question which asks if the parent believes the child needs to be physically punished. While there might seem to be some suggestive evidence from the coefficients of the regression model, the raw data shows that the positive coefficient is indicative of the fact that the control group got worse over time! They were more supportive of phyisical punishment in the endline survey. While there might be a story to that, it could also be the exact kind of statistical anomoly that multiple testing correction is designed to help us avoid when checking so many outcomes.

Parenting Confidence shows no significant impact in the regression analysis. The raw data shows suggestive evidence that those with lower confidence might be more likely to take up the treatment. The lack of a positive coefficient in the regression, however, might indicate that those in the control group were equally likely to take up either the control website or seek out information on their own in order to improve by endline.

Table 7: Pooled: OLS - Endline - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Treatment	−0.003	0.09
	(0.03)	(0.04)
Adjusted Treatment p-value	0.957	0.116
Observations	1,788	1,777
R^2	0.004	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01



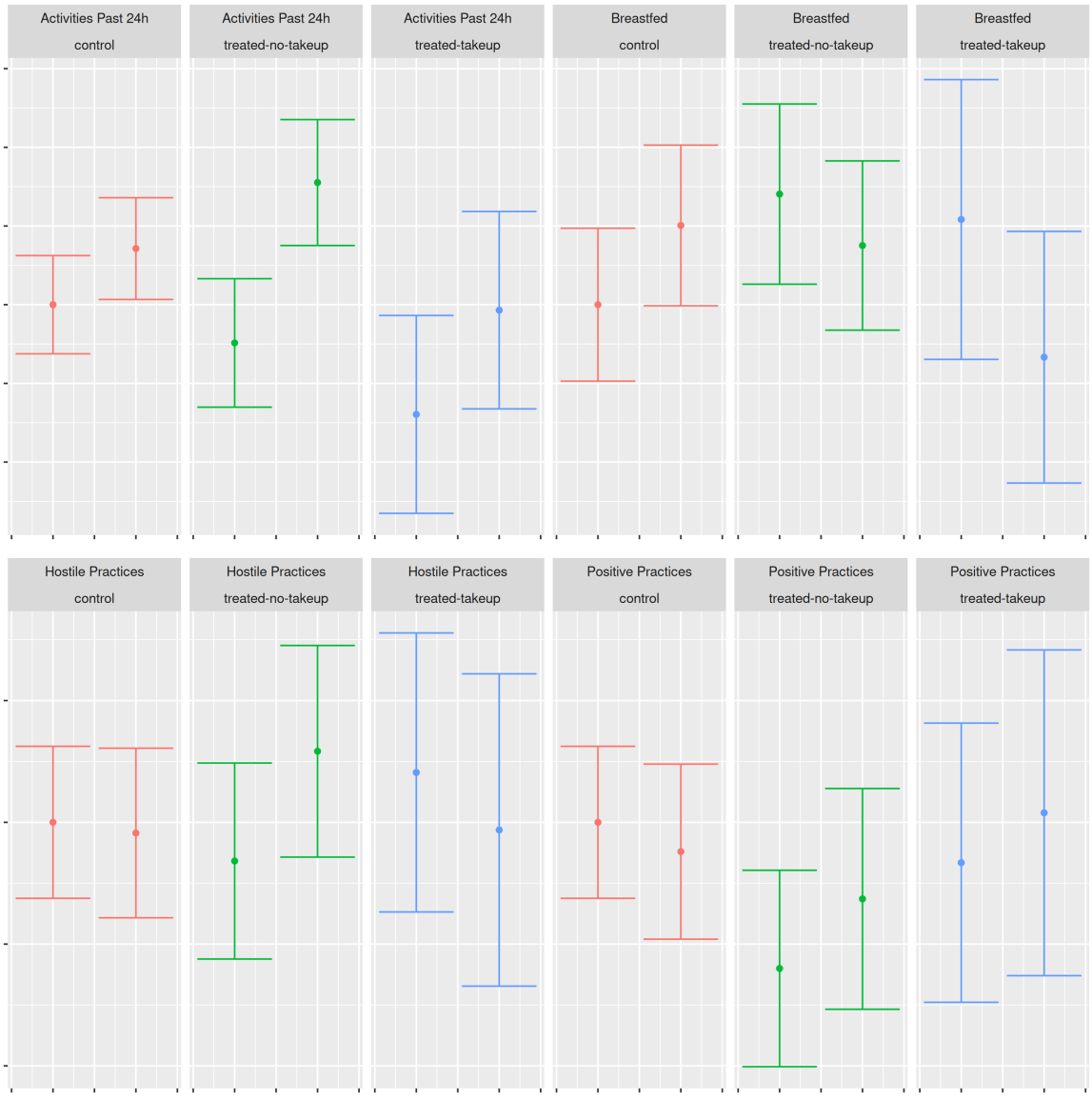
Practices

These four constructs all relate to practices and behaviors of the parent. No significant effect was found for any of the behaviors and there is not much suggestive evidence of selective takeup either. The raw regression results suggest that Activities Past 24h show suggestive evidence of impact, but the raw data shows that the much of the improvement is driven by those in the treated group who did not takeup the treatment, which gives credence to the assumption that this could be statistical noise and is why we have corrected for multiple testing.

Table 8: Pooled: OLS - Endline - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Treatment	−0.03	0.12	0.04	0.04
	(0.03)	(0.06)	(0.03)	(0.03)
Adjusted Treatment p-value	0.505	0.116	0.467	0.467
Observations	630	1,720	1,721	1,717
R^2	0.02	0.01	0.01	0.01

Note: *p<0.1; **p<0.05; ***p<0.01



Policy Implications of the Results

We do not find any significant effect of the use of Bebbo on any of the outcome constructs of interest.

Three reasons, shown in the descriptive data as well as the raw pre-post data might explain why that is the case:

1. The presence of ceiling effects, where much of the population scored high in the baseline and could not improve in the endline.
2. Priming effects led to participants improving from the first questionnaire to the second questionnaire, regardless of treatment arm and regardless of compliance.
3. Low app usage. While takeup defined as "had at least one learning event" was 28%, which would be enough to measure impacts, it's reasonable to believe that in order to have an impact on these outcomes, especially behaviors and attitudes, participants would need to use the app continuously. Especially if we consider the advantage of an app over a static website or informational fly, the advantage comes through continued usage (it is available on your home screen, can send you push notifications, etc.). Given that only 3% used the app more than three days, we would not expect to see much of an impact of this app on the population.

Ceiling effects might be a failure in the creation of the survey instrument. They could also be an example in the bias of the sample population (they are all better-than-average caregivers). But there could be a policy implication as well: it could indicate that most caregivers are quite good already at these outcomes, which is important to consider in the means of addressing the problem. In particular: it could indicate the importance of learning about and focusing effort on subgroups that are worse off. Towards that end, we will perform an analysis to determine the characteristics of the "worse" caregivers.

Priming effects are a result of the study design, however, they indicate potential policy implications as well. In particular: if asking people questions ("Do you know which vaccine your child needs to take next") has such a powerful effect on their knowledge, awareness campaigns might be enough to drive results on these outcomes. Knowledge about vaccines and knowledge about child development both seem like good candidates for such an intervention, given this study.

Finally, low app usage implies that either (i) any app must go through extensive testing and improvement before it will be expected to make an impact measurable on a population level or (ii) apps might not be the most effective method of engaging parents. Like any intervention: the implementation matters and each app can be very different. One app failing to engage does not mean that all apps will fail to engage, however, it does leave the possibility open.

6 User Characteristics Correlated with App Usage

Analysis Design

In this analysis, we attempt to answer the questions -

1. Who are the respondents who used the app?
2. Do more knowledgeable parents use the app more?

We do so by regressing respondents' app usage activity against their characteristics.

Independent variables :

- **Baseline characteristics** - respondents' scores on the construct variables. We only use the construct variables that have sufficiently high internal consistency. We drop the constructs that are highly correlated with other constructs. We find *parent_knw* to be correlated with *caregiver_well_being*. We drop the construct with lower reliability *caregiver_well_being*.
- Demographics variables - parents' age flag (categorical), university flag (binary), gender (categorical), and number of children (numeric)
- Survey response variables - survey duration, start week, country flag

Dependent variables :

- Home opened - binary variable indicating whether the respondent had a home opened event logged
- Home opens - continuous count variable indicating the respondents' count of total home opens

Download Analysis

For respondents who were treated, i.e., asked to download the Bebbo app, we fit a logistic regression model to whether the respondent had a home opened event logged using the respondents' baseline characteristics, demographic variables, and their survey response variables (Table **??**). The goodness of fit measure used is the percentage improvement in deviance over the null deviance (pseudo R^2). The pseudo R^2 for this model is 0.03.

Table 9: home opened (binary) predicted by baseline characteristics

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.11	0.52	2.15	0.03
start_week	-0.04	0.01	-3.87	0
dev_knw_recog	0.31	0.17	1.84	0.06
confidence	-0.15	0.08	-1.89	0.06
attitude	0.01	0.06	0.22	0.82
caregiver_well_being	-0.20	0.09	-2.12	0.03
practices_24	-0.62	0.22	-2.77	0.01
practices_agree	0.13	0.07	1.84	0.07
practices_hostility	-0.10	0.07	-1.44	0.15
parent_age	-0.02	0.01	-3.71	0
number_children	0	0	0.24	0.81
parent_genderWoman	0.31	0.13	2.47	0.01
survey_duration	0.06	0.01	4.59	0
education	0.11	0.10	1.12	0.26
age_flag2-6	0.11	0.10	1.15	0.25
countryserbia	0.20	0.16	1.29	0.20

Continued Usage Analysis

For respondents who downloaded the app, we regress their number of home opens against their baseline characteristics, demographic variables, and survey response variables. The R^2 for this model is 0.065. Note that the country flag is not significant which means that app usage does not differ significantly across the two countries after holding constant user characteristics (Table 9).

Table 10: home opens predicted by baseline characteristics

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.35	1.79	5.78	0
start_week	-0.15	0.03	-4.35	0
dev_knw_recog	0.36	0.59	0.60	0.55
confidence	-0.31	0.28	-1.08	0.28
attitude	-0.07	0.22	-0.30	0.76
caregiver_well_being	-0.90	0.34	-2.64	0.01
practices_24	-2.23	0.78	-2.85	0.005
practices_agree	0.31	0.26	1.19	0.23
practices_hostility	0.31	0.26	1.19	0.24
parent_age	-0.01	0.02	-0.82	0.41
number_children	0	0	0.36	0.72
parent_genderWoman	0.63	0.45	1.41	0.16
survey_duration	0.09	0.05	1.82	0.07
education	0.02	0.36	0.04	0.96
age_flag2-6	-1.22	0.34	-3.55	0
countryserbia	-0.09	0.58	-0.15	0.88

7 Conclusions and Reccomendations

References

[1] Scott Clifford, Geoffrey Sheagley, and Spencer Piston. "Increasing Precision without Altering Treatment Effects: Repeated Measures Designs in Survey Experiments". In: *American Political Science Review* 115.3 (2021), pp. 1048–1065. ISSN: 15375943. DOI: 10.1017/S0003055421000241.

[2] Guido W. Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Apr. 2015. ISBN: 9780521885881. DOI: 10.1017/CBO9781139025751. URL: https://www.cambridge.org/core/product/identifier/CBO9781139025751A535/type/book_part%20https://www.cambridge.org/core/product/identifier/9781139025751/type/book.

[3] Paul Moayyedi and Richard H. Hunt. "Randomized Controlled Trials". In: *GI Epidemiology: Diseases and Clinical Methodology: Second Edition* 7 (2014), pp. 113–118. DOI: 10.1002/9781118727072.ch12.

[4] Stefanie Stantcheva. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible". In: *Annual Review of Economics* 15 (2023), pp. 205–234. ISSN: 19411391. DOI: 10.1146/annurev-economics-091622-010157.

A Baseline Balance

To test for balance between our randomly assigned treatment and control groups, we run an omnibus test, following Hansen and Bowers (2008), to observe standardized differences at baseline and the associated omnibus p-value. Results are reported separately for each country and found in tables 11 and 12. Following **Altman2014**, we do not change our analysis plan based on these results, but it is worth noting that the Bulgaria data does seem to suffer from slight unusual differences between treatment and control condition and the p-value of the omnibus test is significantly low. All the analysis is also reported for only those respondents in Serbia as well, which serves as a robustness check against any concerns that Bulgarians were randomizes into unlucky groups for our analysis.

Table 11: Baseline Balance Serbia

	control_mean	treatment_mean	standardized_diff	z_score
health_knw	0.78	0.75	-0.07	-1.16
dev_knw_recog	0.87	0.88	0.02	0.62
confidence	3.38	3.39	0.02	0.57
attitude	3.08	3.08	-0.01	-0.17
was_breastfed	0.39	0.42	0.07	1.19
practices_24	5.09	4.99	-0.09	-2.39
practices_agree	2.95	2.94	-0.02	-0.42
practices_hostility	3.07	3.05	-0.03	-0.75
(health_knw)	0.40	0.41	0.02	0.63

Overall P-Value: 0.307

Table 12: Baseline Balance Bulgaria

	control_mean	treatment_mean	standardized_diff	z_score
health_knw	0.67	0.65	-0.04	-0.59
dev_knw_recog	0.85	0.81	-0.13	-2.70
confidence	3.27	3.26	-0.01	-0.19
attitude	3.24	3.17	-0.08	-1.64
was_breastfed	0.30	0.35	0.10	1.32
practices_24	4.77	4.70	-0.05	-1.13
practices_agree	3.59	3.59	0.02	0.39
practices_hostility	2.97	3.01	0.05	1.02
(health_knw)	0.45	0.46	0.02	0.33
(was_breastfed)	0.45	0.45	0.01	0.23

Overall P-Value: 0.037

B Additional Plots

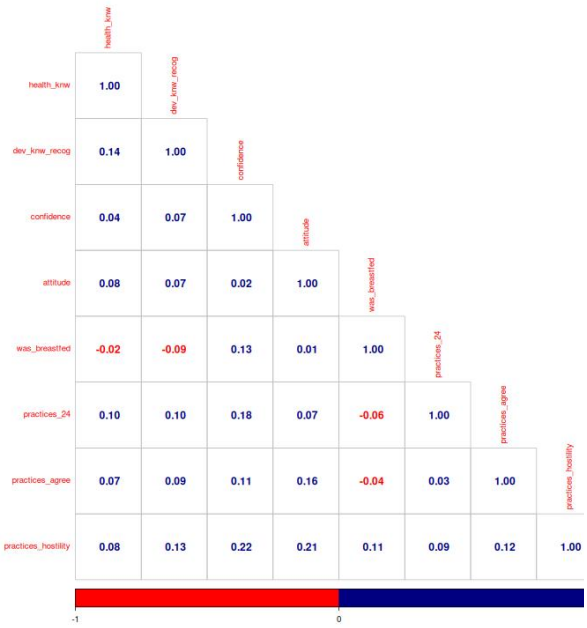


Figure 3: Construct Correlations - Serbia

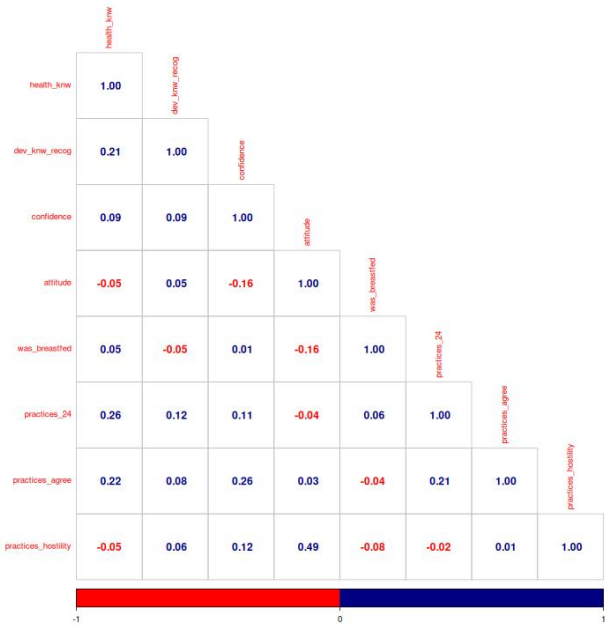


Figure 4: Construct Correlations - Bulgaria

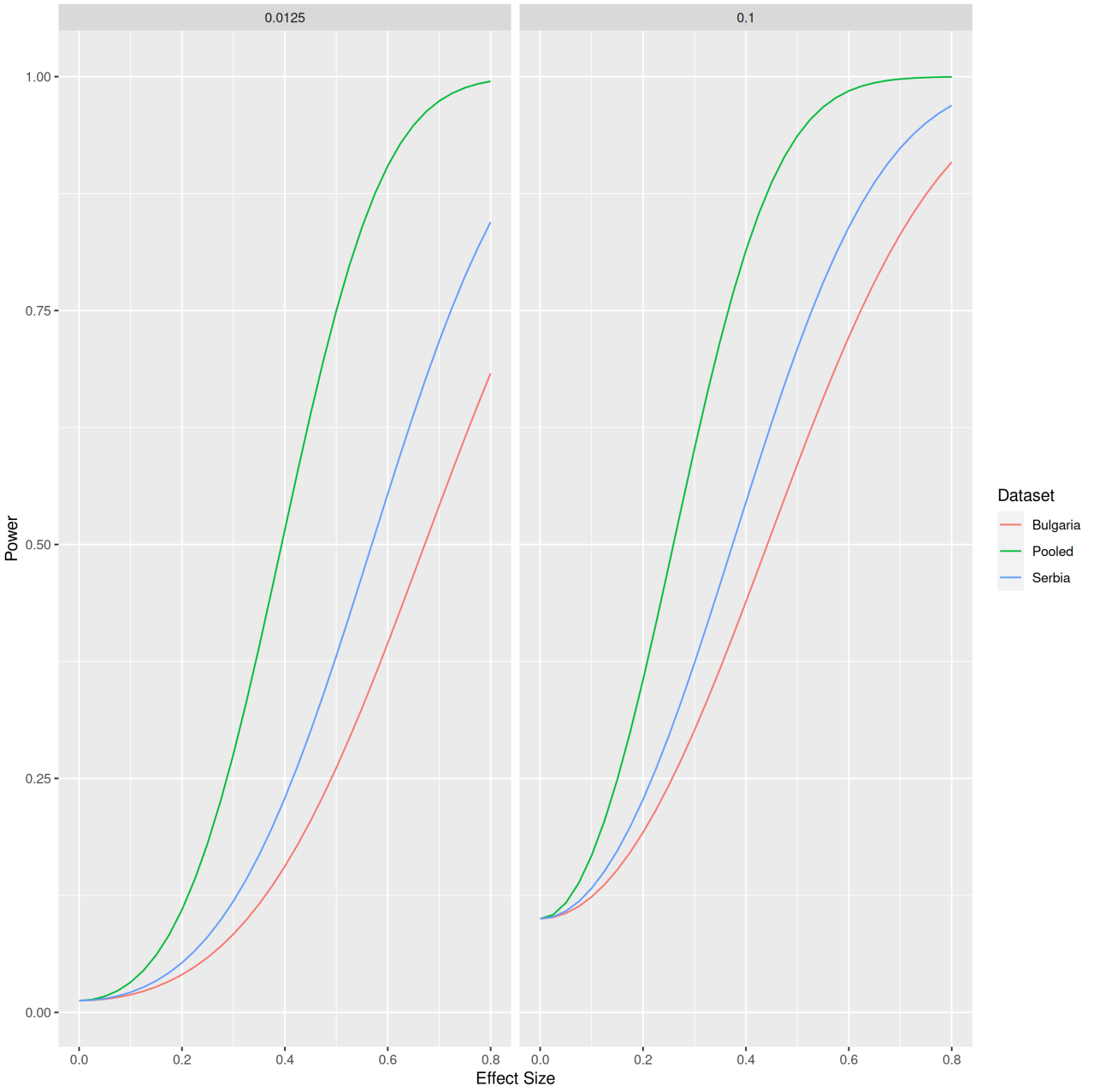


Figure 5: Power Analysis at 28% Takeup

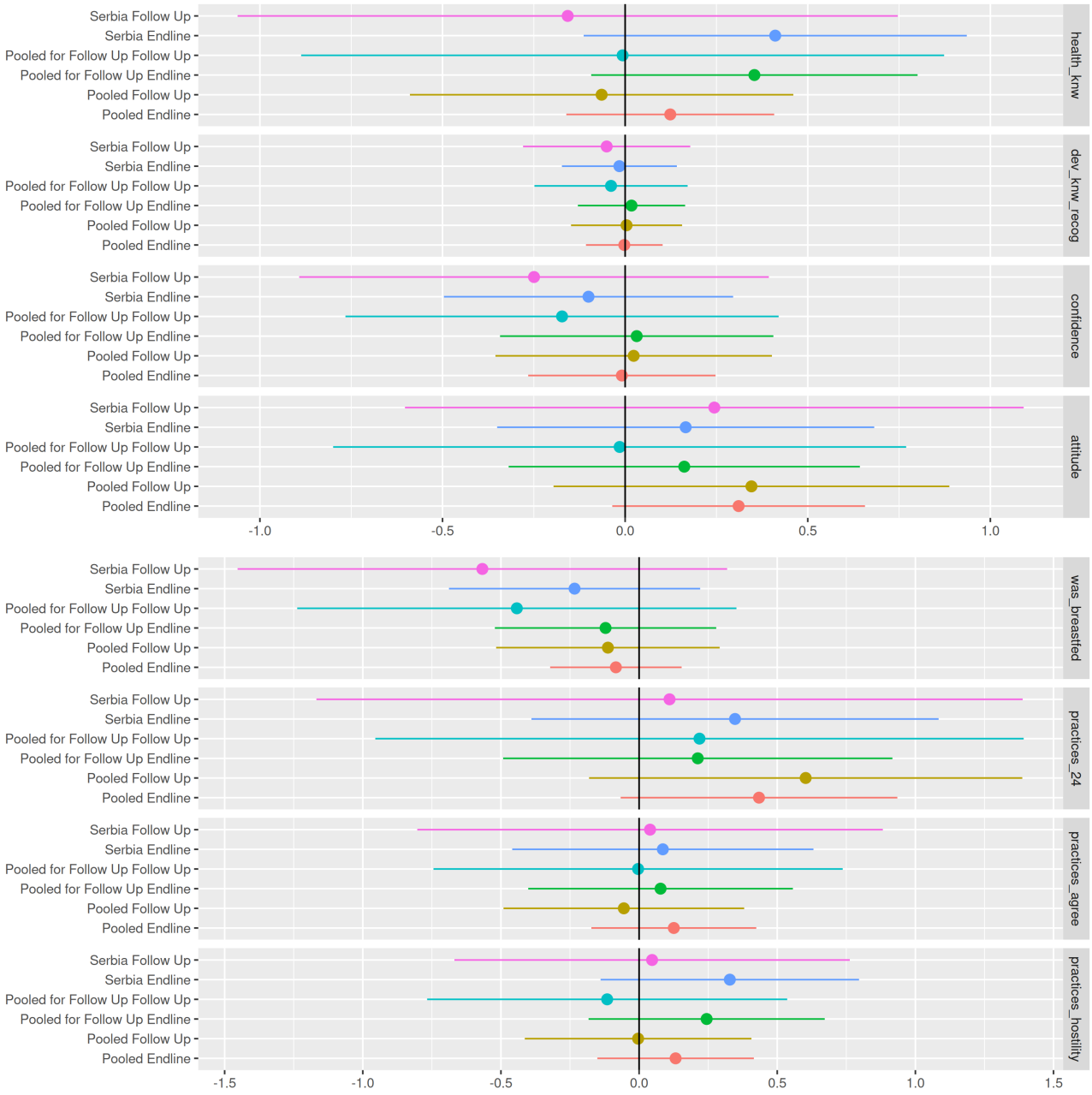


Figure 6: Adjusted Coefficient Plots of 2SLS in Pooled Dataset

C Additional Tables

Table 13: Outcome Construct Descriptives Serbia Baseline

name	mean	median	min	max	sd	prop_max	prop_na
Activities Past 24h	5.06	5.0	0	6	1.16	0.46	0.00
Parenting Confidence	3.41	3.5	1	4	0.65	0.43	0.00
Hostile Practices	3.06	3.0	1	4	0.67	0.18	0.00
Attitude to Phys. Punishment	3.05	3.0	1	4	0.82	0.31	0.00
Positive Practices	2.93	3.0	1	4	0.81	0.24	0.00
Child Dev. Knowledge	0.87	1.0	0	1	0.27	0.76	0.00
Vaccine Knowledge	0.76	1.0	0	1	0.43	0.76	0.61
Breastfed	0.41	0.0	0	1	0.49	0.41	0.61

Table 14: Outcome Construct Descriptives Bulgaria Baseline

name	mean	median	min	max	sd	prop_max	prop_na
Activities Past 24h	4.71	5.00	0.0	6	1.33	0.36	0.00
Positive Practices	3.59	3.75	1.5	4	0.40	0.32	0.00
Parenting Confidence	3.32	3.50	1.0	4	0.63	0.34	0.00
Attitude to Phys. Punishment	3.14	3.00	1.0	4	0.93	0.41	0.00
Hostile Practices	2.97	3.00	1.0	4	0.74	0.11	0.00
Child Dev. Knowledge	0.84	1.00	0.0	1	0.30	0.70	0.00
Vaccine Knowledge	0.67	1.00	0.0	1	0.47	0.67	0.55
Breastfed	0.33	0.00	0.0	1	0.47	0.33	0.55

Table 15: Attrition: Serbia

stage	count	attrition	treated_attrition	control_attrition	attrition_dif
Started Baseline	4847				
Finished Baseline	2615	0.46	0.47	0.45	0.01
Started Endline	1237	0.53	0.51	0.54	-0.03
Finished Endline	1057	0.15	0.17	0.12	0.04
Started Followup	533	0.50	0.51	0.48	0.02
Finished Followup	377	0.29	0.30	0.28	0.02

Table 16: Attrition: Bulgaria

stage	count	attrition	treated_attrition	control_attrition	attrition_dif
Started Baseline	4147				
Finished Baseline	1706	0.59	0.59	0.59	-0.01
Started Endline	731	0.57	0.59	0.56	0.03
Finished Endline	622	0.15	0.17	0.13	0.04
Started Followup	36	0.94	0.95	0.94	0.02
Finished Followup	35	0.03	0.07	0.00	0.07

D Additional Regressions

Table 17: Pooled: 2SLS - Endline - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Used App	0.12	−0.002
	(0.11)	(0.04)
Adjusted Treatment p-value	0.468	0.957
Weak instruments p-value	1.65e-27	1.4e-73
Wu-Hausman p-value	0.869	0.652
Observations	667	1,811
R^2	0.01	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 18: Pooled: 2SLS - Endline - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Used App	−0.01	0.31
	(0.10)	(0.14)
Adjusted Treatment p-value	0.957	0.121
Weak instruments p-value	3.53e-72	1.51e-72
Wu-Hausman p-value	0.711	0.0299
Observations	1,788	1,777
R^2	0.003	−0.01
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 19: Pooled: 2SLS - Endline - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Used App	−0.08	0.43	0.13	0.13
	(0.10)	(0.20)	(0.12)	(0.11)
Adjusted Treatment p-value	0.503	0.121	0.468	0.468
Weak instruments p-value	9.48e-27	7.76e-71	4.5e-71	3.42e-71
Wu-Hausman p-value	0.995	0.0262	0.377	0.0668
Observations	630	1,720	1,721	1,717
R^2	0.03	−0.01	0.003	0.001
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 20: Serbia: OLS - Endline - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Treatment	0.10	−0.004
	(0.05)	(0.02)
Adjusted Treatment p-value	0.298	0.802
Observations	264	799
R^2	0.03	0.02
Note:		*p<0.1; **p<0.05; ***p<0.01

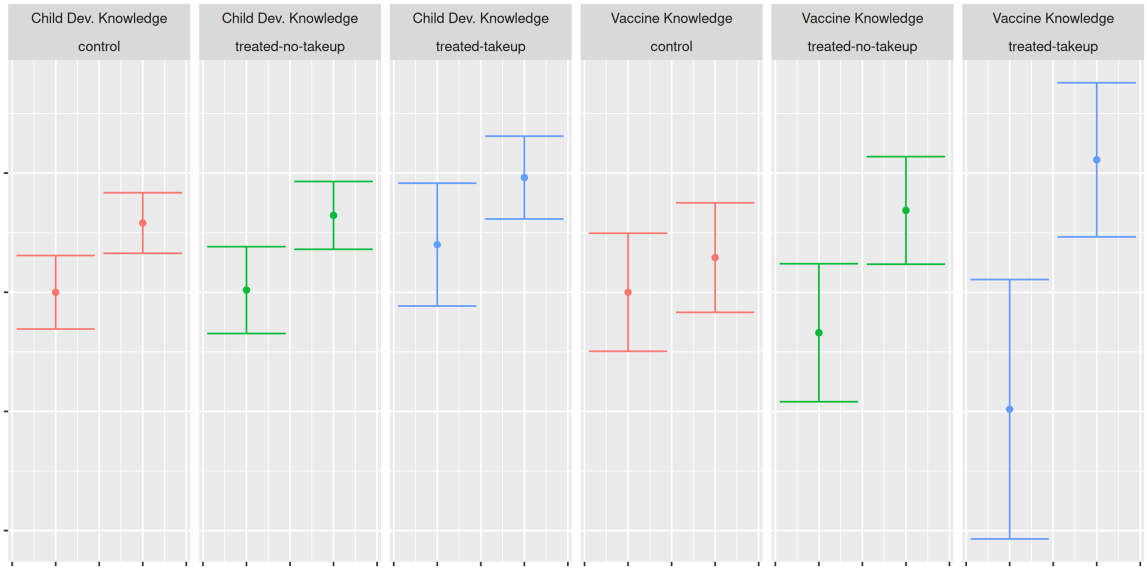


Table 21: Serbia: OLS - Endline - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Treatment	−0.03	0.05
	(0.04)	(0.06)
Adjusted Treatment p-value	0.7	0.673
Observations	789	788
R^2	0.02	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01

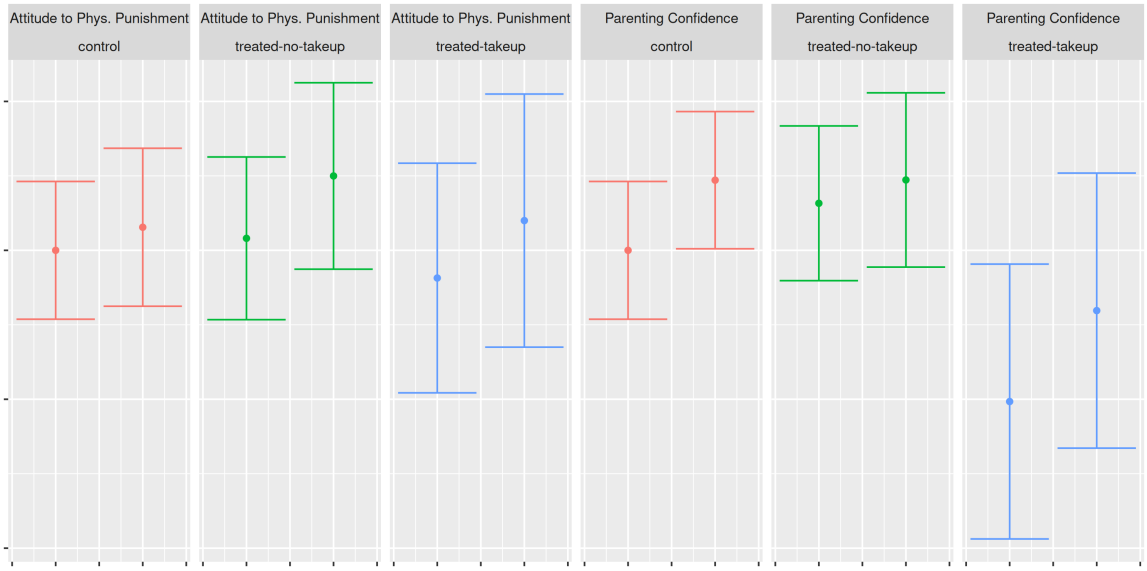
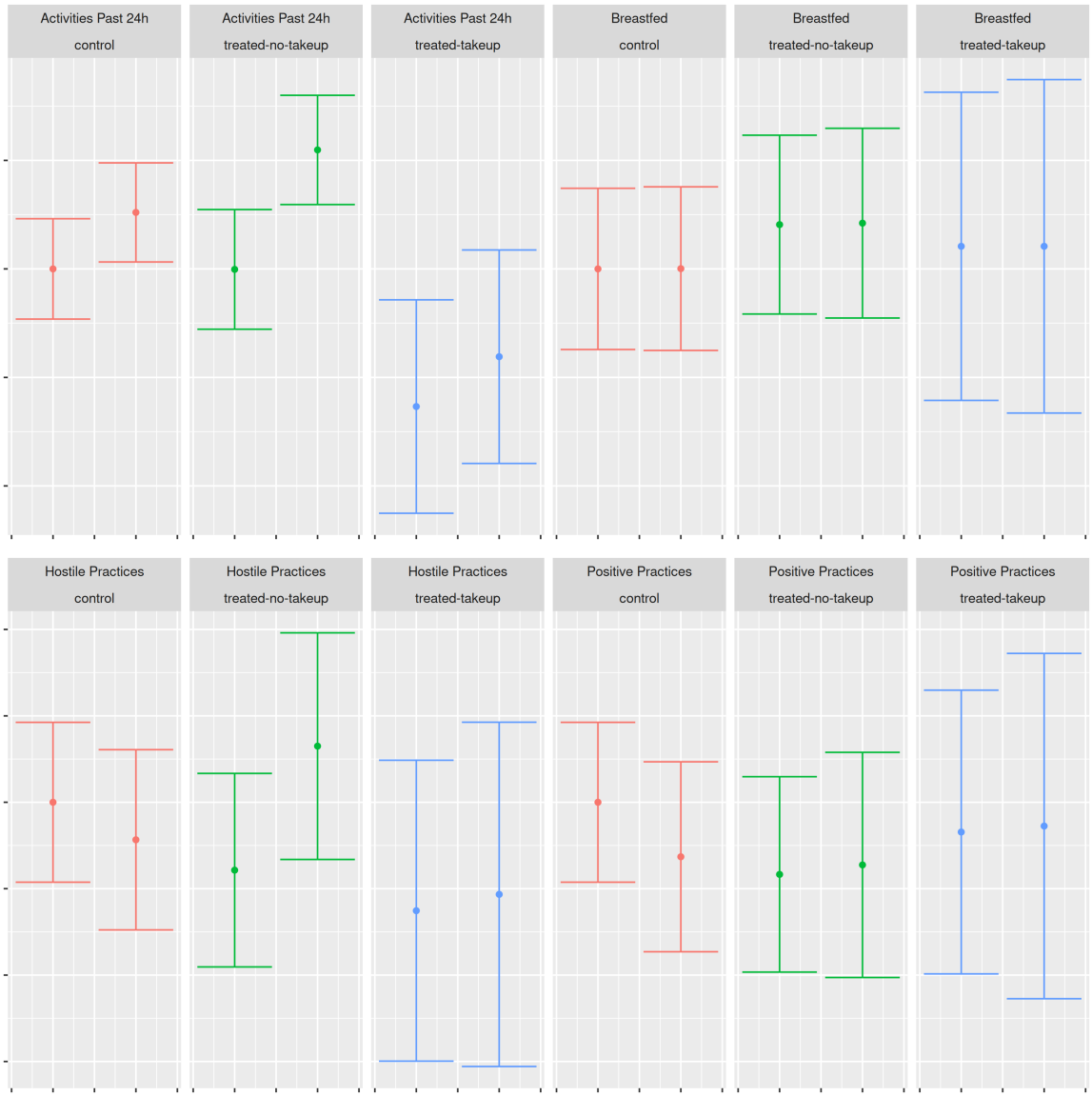


Table 22: Serbia: OLS - Endline - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Treatment	−0.06	0.10	0.02	0.09
	(0.05)	(0.08)	(0.06)	(0.05)
Adjusted Treatment p-value	0.472	0.472	0.792	0.298
Observations	252	765	766	764
R^2	0.05	0.01	0.005	0.02

Note: *p<0.1; **p<0.05; ***p<0.01



E Survey Instrument

Table 23: Serbia: 2SLS - Endline - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Used App	0.41	−0.02
	(0.21)	(0.06)
Adjusted Treatment p-value	0.317	0.802
Weak instruments p-value	3.02e-09	5.63e-32
Wu-Hausman p-value	0.241	0.831
Observations	264	799
R^2	0.01	0.02
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 24: Serbia: 2SLS - Endline - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Used App	−0.10	0.17
	(0.16)	(0.21)
Adjusted Treatment p-value	0.702	0.675
Weak instruments p-value	1.19e-31	1.02e-31
Wu-Hausman p-value	0.467	0.443
Observations	789	788
R^2	0.01	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 25: Serbia: 2SLS - Endline - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Used App	−0.23	0.35	0.09	0.33
	(0.18)	(0.29)	(0.22)	(0.19)
Adjusted Treatment p-value	0.477	0.477	0.792	0.317
Weak instruments p-value	1.18e-09	1.94e-31	1.17e-31	1.05e-31
Wu-Hausman p-value	0.541	0.232	0.88	0.0382
Observations	252	765	766	764
R^2	0.04	−0.004	0.01	−0.01
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 26: Pooled for Follow Up: OLS - Follow Up - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Treatment	−0.001	−0.01
	(0.07)	(0.02)
Adjusted Treatment p-value	0.99	0.99
Observations	133	417
R^2	0.03	0.04
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 27: Pooled for Follow Up: OLS - Follow Up - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Treatment	−0.05	−0.004
	(0.06)	(0.08)
Adjusted Treatment p-value	0.99	0.99
Observations	416	416
R^2	0.04	0.04
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 28: Pooled for Follow Up: OLS - Follow Up - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Treatment	−0.09	0.06	−0.001	−0.03
	(0.07)	(0.12)	(0.08)	(0.07)
Adjusted Treatment p-value	0.99	0.99	0.99	0.99
Observations	132	414	414	413
R^2	0.03	0.004	0.03	0.02
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 29: Pooled for Follow Up: 2SLS - Follow Up - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Used App	−0.01	−0.04
	(0.35)	(0.08)
Adjusted Treatment p-value	0.99	0.99
Weak instruments p-value	5.05e-05	1.62e-17
Wu-Hausman p-value	0.454	0.523
Observations	133	417
R^2	0.03	0.03
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 30: Pooled for Follow Up: 2SLS - Follow Up - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Used App	−0.17	−0.02
	(0.24)	(0.31)
Adjusted Treatment p-value	0.99	0.99
Weak instruments p-value	3.94e-17	3.94e-17
Wu-Hausman p-value	0.582	0.926
Observations	416	416
R^2	0.03	0.04
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 31: Pooled for Follow Up: 2SLS - Follow Up - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Used App	−0.44	0.22	−0.004	−0.12
	(0.31)	(0.47)	(0.30)	(0.26)
Adjusted Treatment p-value	0.99	0.99	0.99	0.99
Weak instruments p-value	4.16e-05	1.03e-16	1.03e-16	8.65e-17
Wu-Hausman p-value	0.472	0.885	0.663	0.966
Observations	132	414	414	413
R^2	0.02	0.01	0.03	0.02
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 32: Construct Variable Mapping

Domain	construct_variable	question
Knowledge and awareness	health_knw	I know which vaccine {{field:child_name}} needs to take next.
Knowledge and awareness	dev_knw_recog	I would be able to recognize if {{field:child_name}} lags behind in social-emotional development (expressing and recognizing feelings and emotions, engaging in interactions, etc.).
Knowledge and awareness	dev_knw_recog	I would be able to recognize if {{field:child_name}} lags behind in cognitive development (mental development, intellectual development).
Knowledge and awareness	dev_knw_recog	I would be able to recognize if {{field:child_name}} lags behind in physical development.
Knowledge and awareness	dev_knw_recog	I would be able to recognize if {{field:child_name}} lags behind in language development.
Confidence and attitudes	confidence	How confident do you feel in your ability to deal with {{field:child_name}}'s emotions?
Confidence and attitudes	confidence	How confident do you feel in your ability to respond properly when {{field:child_name}} misbehaves?
Confidence and attitudes	attitude	Do you agree that in order to bring up, raise, or educate a child properly, the child needs to be physically punished?
Confidence and attitudes	caregiver_well_being	How often can you handle stressful parenting situations successfully?
Practices	was_breastfed	Has {{field:child_name}} been breastfed in the last 24 hours?
Practices	practices_24	In the past 24 hours, did you read books or look at picture books with {{field:child_name}}?
Practices	practices_24	In the past 24 hours, did you tell stories with {{field:child_name}}?
Practices	practices_24	In the past 24 hours, did you sing songs (including lullabies) to or with {{field:child_name}}?
Practices	practices_24	In the past 24 hours, did you take {{field:child_name}} outside the home?
Practices	practices_24	In the past 24 hours, did you play with {{field:child_name}}?
Practices	practices_24	In the past 24 hours, did you name, count or draw things with or for {{field:child_name}}?
Practices	practices_agree	When {{field:child_name}} and I play together, we laugh a lot.
Practices	practices_agree	I joke around with {{field:child_name}}.
Practices	practices_agree	I often smile when I'm around {{field:child_name}}.
Practices	practices_agree	{{field:child_name}} and I play together on the floor.
Practices	practices_hostility	I snap at {{field:child_name}} when he/she gets on my nerves.
Practices	practices_hostility	When {{field:child_name}} upsets me, I lose my patience and punish him/her more severely than I really mean to.
Practices	practices_hostility	When {{field:child_name}} does something wrong, I sometimes threaten him/her.
Practices	practices_hostility	I sometimes make fun of {{field:child_name}}.