Bebbo

February 2024

Contents

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, and stimulating 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

The Bebbo app is designed so that parents can begin using the app by creating a "profile" of their baby, entering basic information such as when the baby was born. After this, the parent is shown relevant content, asked to track milestones, and offered suggestions for content reading or games to play with their child.

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. This study follows a randomized, difference-in-difference design, also known as a prepost design (Clifford, Sheagley, and Piston 2021). Survey questions measuring the outcomes are 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 static parenting website. Given the study design, however, we can make some additional inferences and policy recommendations:

1. The majority of the caregiving population in these countries is already very "good" in regards to the majority of the outcomes of interest. This implies that one may not need a "swiss-army knife" tool that is meant to work on all outcomes across all types of people. It might be more efficient to develop targeted interventions that target desired behaviors directly or target populations or communities directly.

2. Awareness is in and of itself an effective intervention for some outcomes of interest, including knowledge of vaccine requirements. We see this because participants improved from the first questionnaire to the second questionnaire, regardless of treatment arm and regardless of compliance.

3. Most people choose not to use Bebbo after downloading it. Those who do choose to use the app are neither more or less likely to be the people who need support. The app cannot, therefore, be expected to have a population-level impact unless it improves its engagement metrics. In our study, only 24% of respondents who downloaded the app ended up using it beyond the first day, and only 5% used it more than three days. When analyzing app data from Serbia and Bulgaria outside of the study population, we find similar and slightly worse engagement metrics. 70% of Bebbo downloaders never fully complete their profile and 80% never use the app after the first day.

These three facts highlight that promoting Bebbo not only has no measurable impact on our study population, but we would also not expect it to have a large impact on the measured outcomes in the general population in these countries. In particular, this study indicates that raising awareness alone might be a more effective intervention. We recommend that the product team must improve the basic engagement, retention, and churn numbers if they expect the app to have a population-level impact on parenting outcomes.

At the same time, while only about 5% of app downloaders become regular users (use it more than 3 days in the first 30 days), that does still mean that there are over 2000 users in Serbia and Bulgaria who downloaded the app in 2023 and ended up using it at least four times. That usage alone indicates that they see some value. Unfortunately, our analysis does not reveal any trend that suggests those who download are either those who need or do not need the app according to our measures. We recommend relying on qualitative research about those users to understand the true value of the app among those who do choose to download it and end up using it regularly.

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 promoting 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 (see Study Design for details). 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. Given the relatively low engagement seen with the app among study participants, it might be interesting to focus development of the app on the subset of caregivers who are most likely to need Bebbo, given that it has not been effective at engaging the general populaiton, and repeat a more focused study on that subset of the population.

Alternatively, it might be useful to increase the engagement of the app among a general population of caregivers, before testing its impact again on a general population.

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.

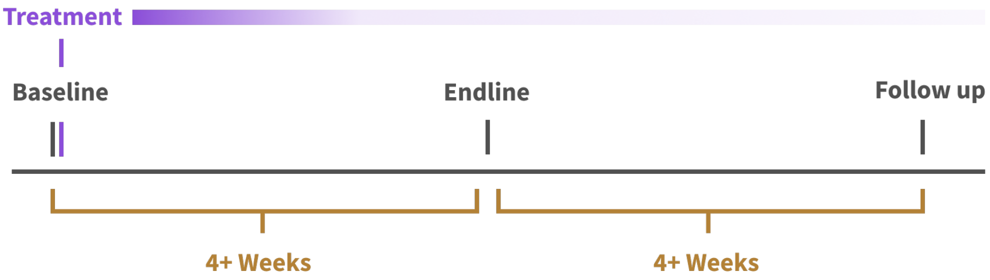


Figure 1: Study Design

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.

The Bebbo app is designed so that parents can begin using the app by creating a "profile" of their baby, entering basic information such as when the baby was born. After this, the parent is shown relevant content, asked to track milestones, and offered suggestions for content reading or games to play with their child.

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. A set of (6) events were determined to fit the bill, namely: advise_details_opened, game_details_opened, child_milestone_tracked, child_measurement_entered, child_vaccine_entered, child_health_checkup_entered.

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 (9meseca.bg) was chosen so as to be "static" (showing the same content to everyone rather than allowing user to create individualized profiles) 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 downside with choosing a treatment-as-usual condition is that if one does not find significant effects of the treatment, one cannot differentiate between the following scenarios:

1. The control is effective and the treatment equally effective.

2. Neither the control nor the treatment are effective.

The assumption with this impact evaluation, from a policy perspective, is that the two are equally important. If the development and promotion of a new app does not improve parenting knowledge beyond what already exists in the market, then it is not an effective investment for public funds. That being said, there is some evidence in this study that allows us to differentiate between the two scenarios and determine that it is likely the latter. In particular, we conclude that neither control nor treatment condition make any significant impact on the general population above and beyond the priming effect of the baseline survey itself (to be discussed in depth in the section with results).

Recruitment

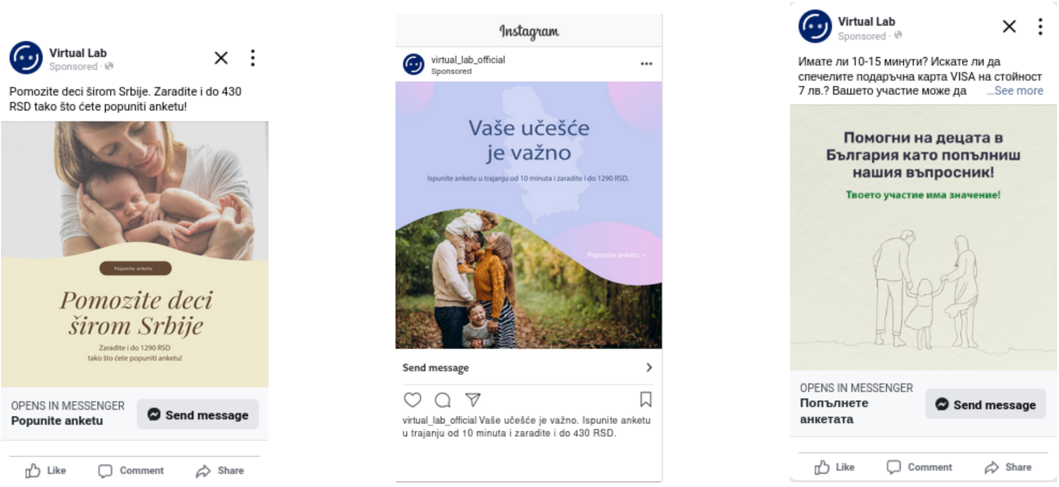
Participants were recruited to the study with social media ads on the Meta platform (Facebook and Instagram) using the Virtual Lab platform to create and run the recruitment ads. The Virtual Lab platform is used to track and measure the price-per-respondent across multiple strata, solving the core problem of monitoring, computing expectations, and adjusting budget when recruiting samples via social media platforms that are representative across desired and measured characteristics.

An initial pilot study was run in Bulgaria to determine the cost effectiveness of the recruitment strategy, together with the incentive amount and mechanism. Results from the pilot indicated that it would not be possible to stratify according to the original plan and stay on budget. The decision was made to move forward with the overall plan, but drop the stratification, in order not to rethink too many parts of the study or fall behind schedule.

One important takeaway and recommendation of this study is to pilot more extensively if the initial pilot shows poor results. When an initial pilot does not provide the desired results, the timeline should be adjusted to reflect that more time is needed to redesign the study and pilot again before launching. Due to time constraints, the study was launched with learnings from the pilot and budget constraints were again run into, limiting the ability to recruit as large of a sample as originally desired in each country. Similarly, attrition between baseline and endline, which was not piloted, was worse than expected. One possible explanation is that the incentive strategy and communication was not attractive enough to respondents, leading to both high recruitment costs, long recruitment time, and higher-than-hoped-for attrition.

In exchange for participating in the study, participants were told they could receive gift cards worth up to 12 USD (in their local currency). These gift cards were delivered as $4 visa international gift cards, which could be spent online but had to be spent for a purchase under $4. 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.



Pomozite deci širom Srbije. Zaradite i do 430 RSD tako što ćete popuniti anketu!

virtual_lab_official Vaše učešće je važno. Ispunite anketu u trajanju od 10 minuta i zaradite i do 430 RSD.

Имате ли 10-15 минути? Искате ли да спечелите подаръчна карта VISA на стойност 7 лв.? Вашето участие може да ...See more

Figure 2: Recruitment Ads

4 Descriptives

Respondent Characteristics by Country

Table 1 provides the baseline characteristics of the respondent population, separated by country. Note that these are also the control variables we will use in all regressions.

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	1420	0.83	2709	0.81
University Educated	1	731	0.43	863	0.26
Speaks Dominant Lang.	1	1572	0.92	3162	0.95
Is Parent	1	1492	0.87	3082	0.92
Child Age	2-6	937	0.55	2110	0.63
Num. Children	4+	71	0.04	316	0.09
Parent Age	Over 35	366	0.21	616	0.18
Urban Area	1	1066	0.62	1375	0.41

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 H.1.

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	5.05	5.00	0	6	1.21	0.48	0.0
Parenting Confidence	3.36	3.50	1	4	0.64	0.38	0.0
Positive Practices	3.10	3.25	1	4	0.80	0.26	0.0
Attitude to Phys. Punishment	3.08	3.00	1	4	0.89	0.37	0.0
Hostile Practices	3.08	3.00	1	4	0.69	0.17	0.0
Child Dev. Knowledge	0.87	1.00	0	1	0.27	0.75	0.0
Vaccine Knowledge	0.74	1.00	0	1	0.44	0.74	0.6
Breastfed	0.41	0.00	0	1	0.49	0.41	0.6

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. In all cases, each variable is attempting to measure a unidimensional construct on the same scale, implying that Chronbach's alpha is a reasonable measure (Tavakol and Dennick 2011).

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 with a reliability above 0.70 are considered internally consistent. After initial reliability analysis, the evaluation team iteratively dropped variables or modified constructs to ensure high reliability. In all cases, the variables that were dropped were clearly not measuring the same construct or measuring it on the same scale. The construct created from Activities in the Past 24 hours initially 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. This was not the original intention of the survey creators, however, it was decided that it was more intentional to measure the construct in this fashion and there was precendence in UNICEF work ([TODO: add reference]).

Table 3: Reliability: Pooled 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	0.56	0.57
Positive Practices	4	0.84	0.84
Hostile Practices	4	0.74	0.74

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? This would not impact the internal validity of the study, however, it has implications for the external validity: what it is we are studying exactly.

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.

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

Ex-post analysis is provided to show 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 well powered (above 80%) to detect a medium effect size (0.5 standard deviations) even when that effect is entirely limited to the 28% takeup group at a significance level of 5%. We also provide plots showing the power at 1.25%, to show the equivalent of a 10% significance level after controlling for multiple testing (8 outcomes) with a Bonferroni correction. See figure B.3.

Thus, the study is well-powered to determine whether or not the 28% who used the app were impacted.

Attrition & Survey Behavior

About 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 4 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 includes not only participants who chose not to return for the follow-up, but also those that were disqualified due to an error in survey coding for a portion of early respondents: 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	9715				
Finished Baseline	5077	0.48	0.48	0.48	0.00
Started Endline	2061	0.59	0.59	0.59	0.00
Finished Endline	1894	0.08	0.10	0.07	0.03
Started Followup	569	0.70	0.70	0.69	0.01
Finished Followup	555	0.02	0.04	0.01	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

min	quantile_05	median	quantile_95	max	Time Gap
2 days	28 days	29 days	57 days	277 days	Baseline - Endline
28 days	28 days	29 days	51 days	173 days	Endline - Followup

App Usage Characteristics

To measure takeup, we pick how many distinct days the participant used the app, as measured by one of the predefined set of "learning events" that include looking at material or entering milestones for their child. Note that in order to generate a "learning event," one must get past the "welcome" screens and create a profile for their child.

Figure 3 shows a histogram of the amount of days that respondents in the treatment group used the Bebbo app. Table 6 shows some takeup numbers for different intensities of usage. Very few treated respondents used the app more than three days (less than 3%) and only about 12% used it more than once. The maximum group of 8 people, less than 1% of the population treated, used the app on more than 5 days in the 4-6 week period.

Additionally, table 7 shows app engagement for those that did download the app, including data from app usage in Serbia and Bulgaria that was not associated with the study. Note that, of those who did download the app in the study (and therefore were complying with the reccomendations of the study) only about 55% actually finished creating a profile for their child and only 24% used the app past the first day and 5% used it for more than 3 days.

The numbers are worse for those outside of the study. Note that the engagement numbers in table 7 were restricted to 30 days after downloading, to compare with the timeframe between baseline and endline within the study. For additional context, we provide table 8 to show a full retention funnel for all users who downloaded in the first eight months of 2023 outside of our study.

These numbers show that less than 30% of downloaders actually complete a profile and access content. Of those, about half never come back to the app in the first 30 days and 30% never come back to the app at all. These are important results. 20% retention after the first day indicates that the app is not engaging or attractive to the majority of caregivers of young children who agree to download it. This is true both in our study population as well as for those who download the app in the wild.

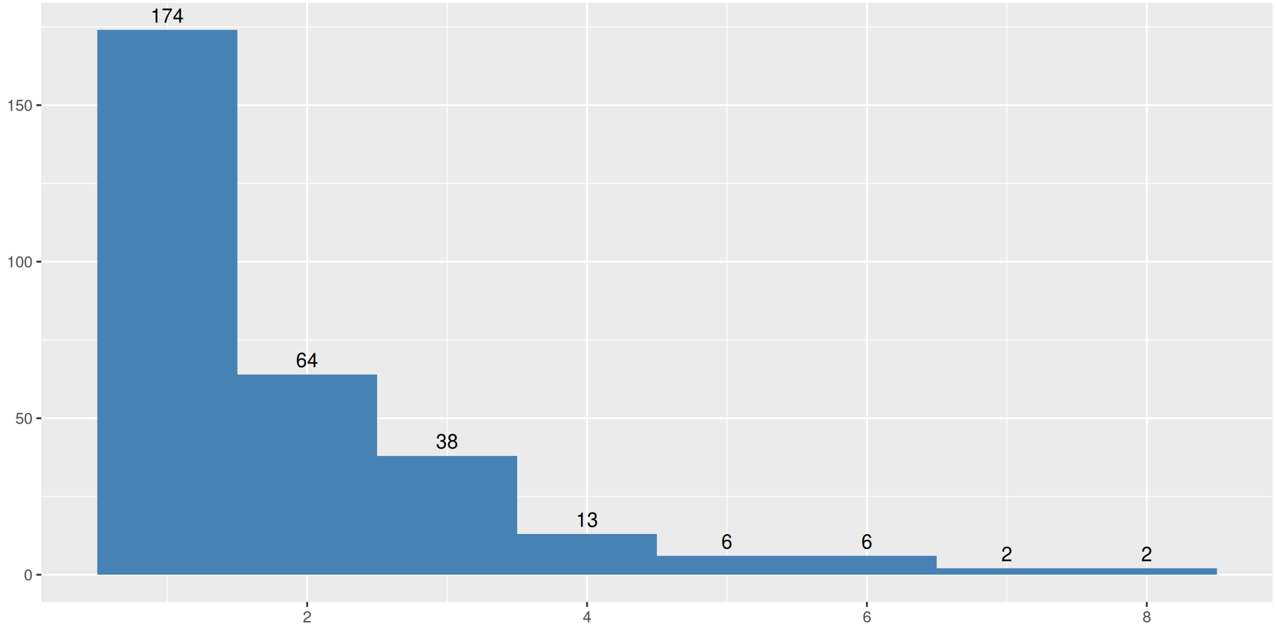


Figure 3: Days with Learning Event (Baseline - Endline)

Table 6: Treatment Takeup

Dataset	Treated	Downloaded	Used	>1 Day	>3 Days	Used (%)	>1 Day (%)	>3 Days (%)
Serbia	678	379	189	85	18	27.9%	12.5%	2.7%
Bulgaria	342	182	116	46	11	33.9%	13.5%	3.2%
Pooled	1020	561	305	131	29	29.9%	12.8%	2.8%

Table 7: App Engagement (30 days)

Dataset	Downloaded	Used (%)	>1 Day (%)	>3 Days (%)
Serbia	379	49.9%	22.4%	4.7%
Bulgaria	182	63.7%	25.3%	6%
Pooled	561	54.4%	23.4%	5.2%
Non-Study	40111	26.7%	13.6%	5.1%

Table 8: App Retention Funnel (Non Study)

Downloaded	Used (%)	After 1 day (%)	After 30 days (%)	After 60 days (%)	After 90 days (%)
40111	29.3%	69.3%	80.9%	85.7%	86.2%

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."

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, the decision was made to rescope this evaluation to focus on the endline. Additional tables are available 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.

Table 1, with the baseline respondent characteristics, shows all the control variables used. Note that several outcomes only applied to parents with children in a certain age group: "Breastfed" and "Vaccine Knowledge" applied to those with children ages 0-2. Thus, the binary control variable representing the age of the child (0-2 or 2-6) was removed in those regressions.

Knowledge and Awareness

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

Table 9: Pooled: OLS - Endline - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Treatment	0.05 (0.03)	−0.003 (0.01)
Adjusted Treatment p-value	0.337	0.797
Observations	719	1,996
R^2	0.01	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01

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 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. This can be seen in the pre-post plots (figure 4) but it can be more formally seen in the regressions in the section App Usage, where the regressions with parents of children 0-2 also show evidence that vaccine knowledge is a predicter of continued app usage, in the inverse (those with less knowledge to begin with are more likely to keep using the app).

Unfortunately, we do not see this same impact on the subset of people even if we perform a subgroup analysis regression, which can be seen in the appendix in table D.1. This indicates that while the app is more likely to be used by that group of people, it would seem that the control subgroup also improved their vaccine knowledge significantly thus canceling out any impact from the app usage.

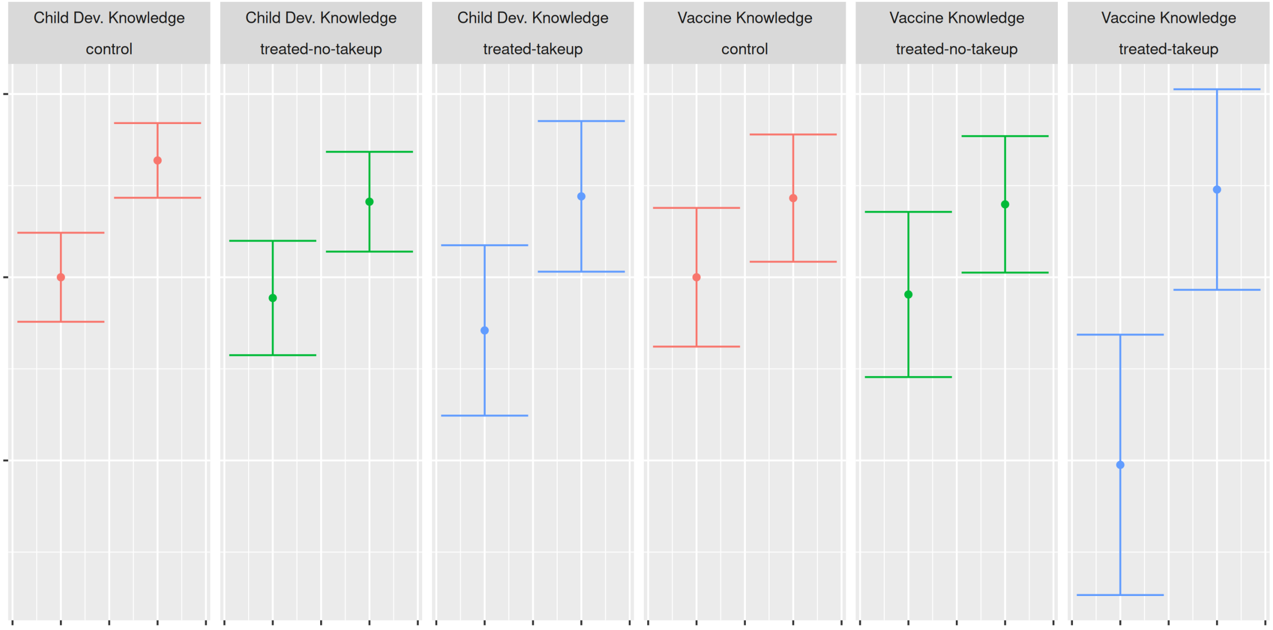


Figure 4: Knowledge and Awareness

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 anomaly 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 10: Pooled: OLS - Endline - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Treatment	0.01 (0.03)	0.08 (0.04)
Adjusted Treatment p-value	0.797	0.157
Observations	1,972	1,961
R^2	0.01	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01

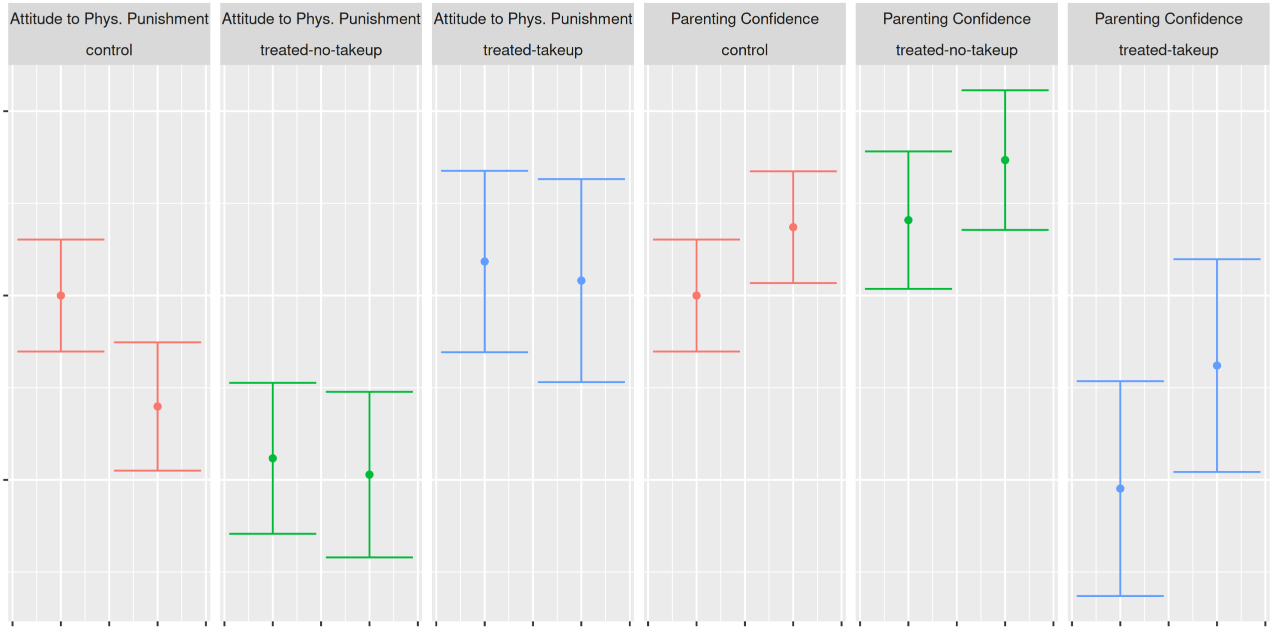


Figure 5: Confidence and Attitudes

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 11: Pooled: OLS - Endline - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Treatment	−0.02 (0.03)	0.12 (0.05)	0.03 (0.03)	0.03 (0.03)
Adjusted Treatment p-value	0.714	0.157	0.618	0.618
Observations	682	1,903	1,904	1,900
R^2	0.02	0.01	0.005	0.01
Note:				*p<0.1; **p<0.05; ***p<0.01

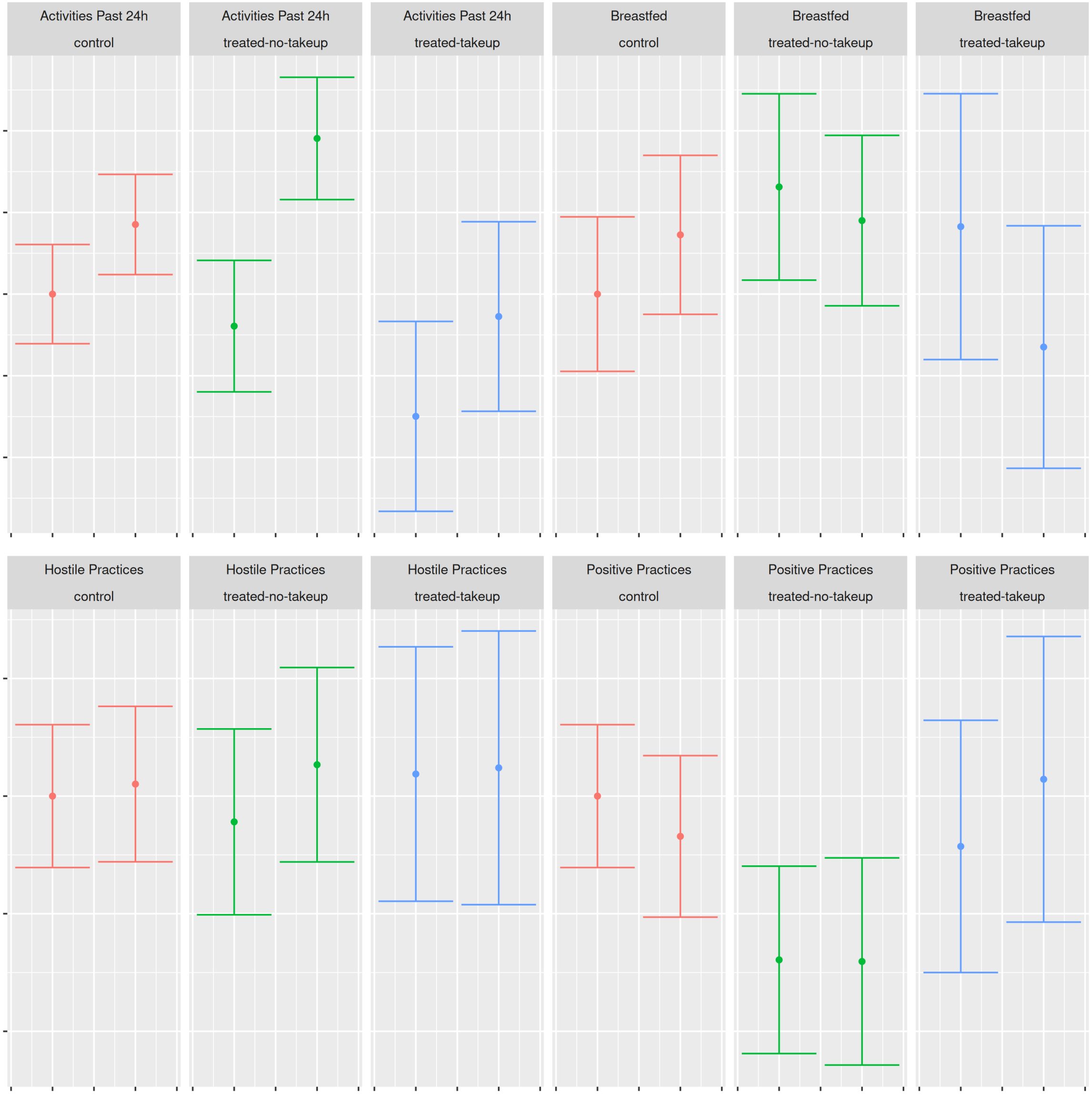


Figure 6: Practices

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

Given that so few caregivers used the app, it seems important to ask the question: "who are the respondents who end up as app users?" We do so by regressing respondents' app usage activity between baseline and endline against their characteristics at baseline. In particular, we will pick two binary outcomes: those who had at least one learning event, who we will say "Used the App," those who had learning events on at least two days, or "Used More Than 1 Day," and finally, those two used the app more than three days.

The results can be found in table **??**. The regression is formulated as a simple linear regression for ease of interperability. The most notable predictor is whether or not the participant is themselves the parent of the child. The other notable epredictor is "Activities Past 24h" which is aligned with what we see in the raw data (under Results), it does seem that in our sample, those who reported doing fewer activities together with their child are more likely to download and use the app. One possible interpretation could be that there is a set of people who are not likely to spend time with their children but are likely to download and use apps. Unfortunately, we do not see a positive impact of this app on those people spending more time with their children, but potentially on a narrow subgroup there might be a positive impact that we do not detect. This could indicate that for "screen parents" that are not currently spending time with their children, an app is an effective way to get in front of them. The open question is whether or not it improves their behavior.

There is some additional evidence that those with young children or more likely to use the app, along with those who have less knowledge of child development, speak the dominant language, and are university educated.

Table 12: App Usage (All)

	Dependent variable:		
	Used the App	Used More Than 1 Day	Used More Than 3 Days
	(1)	(2)	(3)
Is Woman	0.05 (0.04)	0.01 (0.03)	−0.02 (0.01)
University Educated	0.07** (0.03)	0.04 (0.02)	0.003 (0.01)
Speaks Dominant Lang.	0.06 (0.07)	−0.01 (0.05)	0.01 (0.03)
Is Parent	0.23*** (0.06)	0.13*** (0.04)	0.04* (0.02)
Child Age	−0.04 (0.03)	−0.05** (0.02)	−0.02* (0.01)
Num. Children	0.01 (0.06)	−0.01 (0.04)	−0.01 (0.02)
Parent Age	0.05 (0.04)	0.03 (0.03)	0.003 (0.01)
Urban Area	−0.04 (0.03)	−0.003 (0.02)	0.01 (0.01)
Child Dev. Knowledge	0.01 (0.07)	−0.04 (0.05)	0.03 (0.03)
Parenting Confidence	−0.05** (0.02)	−0.04** (0.02)	0.01 (0.01)
Attitude to Phys. Punishment	0.03 (0.02)	0.004 (0.01)	0.01 (0.01)
Activities Past 24h	−0.05*** (0.01)	−0.04*** (0.01)	−0.02*** (0.01)
Positive Practices	0.04* (0.02)	−0.001 (0.01)	0.005 (0.01)
Hostile Practices	−0.01 (0.02)	0.02 (0.02)	0.004 (0.01)
Constant	0.20 (0.14)	0.29*** (0.10)	−0.02 (0.05)
Observations	924	924	924
R^2	0.06	0.05	0.03
Adjusted R^2	0.05	0.03	0.02
Residual Std. Error (df = 909)	0.44	0.33	0.17
F Statistic (df = 14; 909)	4.21***	3.36***	2.14***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 13: App Usage (With Children 0-2)

	Dependent variable:		
	Used the App	Used More Than 1 Day	Used More Than 3 Days
	(1)	(2)	(3)
Is Woman	0.09 (0.07)	0.05 (0.06)	−0.01 (0.03)
University Educated	0.08 (0.05)	0.04 (0.04)	−0.01 (0.03)
Speaks Dominant Lang.	0.16 (0.11)	0.06 (0.09)	−0.03 (0.05)
Is Parent	0.12 (0.11)	0.09 (0.09)	0.06 (0.05)
Num. Children	−0.17 (0.11)	−0.13 (0.08)	−0.04 (0.05)
Parent Age	0.09 (0.06)	0.06 (0.05)	−0.03 (0.03)
Urban Area	0.003 (0.05)	0.02 (0.04)	0.02 (0.02)
Vaccine Knowledge	0.03 (0.07)	−0.02 (0.05)	−0.07** (0.03)
Child Dev. Knowledge	−0.22* (0.12)	−0.23** (0.10)	0.07 (0.06)
Parenting Confidence	0.02 (0.05)	0.001 (0.04)	0.02 (0.02)
Attitude to Phys. Punishment	0.03 (0.03)	0.005 (0.03)	0.02 (0.02)
Breastfed	−0.02 (0.05)	0.04 (0.04)	0.05** (0.02)
Activities Past 24h	−0.07*** (0.02)	−0.05*** (0.02)	−0.01 (0.01)
Positive Practices	0.04 (0.03)	0.03 (0.03)	0.02 (0.02)
Hostile Practices	−0.01 (0.04)	0.01 (0.03)	0.01 (0.02)
Constant	0.22 (0.25)	0.29 (0.20)	−0.18 (0.12)
Observations	367	367	367
R^2	0.09	0.08	0.06
Adjusted R^2	0.05	0.04	0.02
Residual Std. Error (df = 351)	0.46	0.37	0.22
F Statistic (df = 15; 351)	2.25***	2.03**	1.52*
Note:			*p<0.1; **p<0.05; ***p<0.01

7 Conclusions and Reccomendations

By construction, mobile apps are meant to be used more than once. The promise of a mobile app is that users will continually engage with it. On average, Europeans spent more than two hours each day on apps (Statista - “Average daily time spent by users in Europe on mobile apps from October 2020 to March 2021”).

In order for the promotion of an app to be effective as a population-level intervention for parenting, it needs to be engaging enough that a large proportion of the target population will use it and continue to use it. By construction, this implies using it more than one day. Unfortunately, in our study population, 76% of users who complied with the suggestion to download Bebbo abandoned it after the first day and never returned before the endline survey. This is inline with app usage data outside of our study, which shows that over 80% of users abandon it after one day. Tracking app retention and engagement rates is important and can be done long before an app gets promoted to a broad population. It is not clear, a priori, what kind of app will engage caregivers, but that is a problem that should be solved before investing in promotion and scale, not after.

We were not able to predict app usage accurately from baseline survey characteristics, although it does seem as though parents who are less likely to do activities with their kids are more likely to use a parenting app.

We were not able to detect any statistically significant impact of promoting the app Bebbo on a general population of caregivers with children 0-6 years of age. Given the lack of engagement in the Bebbo app, this is not surprising. That being said, there were other factors that might have contributed to the lack of discovery of an impact that are in-and-of-themselves interesting. In particular:

1. There seemed to be a priming effect of the baseline survey, especially for questions related to knowledge, such as “when is your child’s next vaccination due.” That implies that asking the question has an impact on caregivers and an awareness campaign might be a cheap and effective way of improving routine vaccination rates. People can find the information, they just need to be reminded to look.

2. The majority of respondents scored either very good or perfectly on the baseline assessment. If the respondents were representative of the general population, this would imply that most caregivers in these countries are already knowledgeable and following many good practices. As opposed to a “general parenting” intervention, it might be best to focus interventions on particular practices which are still lagging (i.e. breastfeeding) or focus on particular groups or communities where all practices are lagging.

Taken together, these results lead us to recommend that:

1. To understand the impact of the Bebbo app among the approximately 5% of downloaders who end up as regularly users, we recommend relying on qualitative research with the existing users.

2. In its current state, Bebbo cannot be expected to move the needle on a population and must become more engaging to a broader range of caregivers before it can be expected to have significant population-level impact.

3. Awareness campaigns might have a significant impact on many of the outcomes of interest to this study and should be investigated further as potential policies that may be more effective at making a significant population-level impact.

References

[1] Douglas G Altman. “Comparability of randomised groups”. In: 34.1 (2014), pp. 125–136.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] Mohsen Tavakol and Reg Dennick. “Making sense of Cronbach’s alpha”. In: *International journal of medical education* 2 (2011), pp. 53–55. ISSN: 20426372. DOI: 10.5116/ijme.4dfb.8dfd.

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 A.1 and A.2. Following Altman 2014, 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 A.1: Baseline Balance Serbia

	control_mean	treatment_mean	standardized_diff	z_score
health_knw	0.80	0.76	-0.08	-1.47
dev_knw_recog	0.89	0.89	0.04	1.02
confidence	3.40	3.42	0.02	0.63
attitude	3.03	3.02	-0.01	-0.33
was_breastfed	0.45	0.47	0.04	0.67
practices_24	5.24	5.16	-0.07	-2.01
practices_agree	2.86	2.83	-0.04	-1.01
practices_hostility	3.13	3.13	-0.005	-0.14
(health_knw)	0.37	0.37	-0.01	-0.27
(attitude)	1	1.00	-0.03	-1.00
(was_breastfed)	0.37	0.37	-0.01	-0.27

Overall P-Value: 0.354

Table A.2: Baseline Balance Bulgaria

	control_mean	treatment_mean	standardized_diff	z_score
health_knw	0.67	0.65	-0.04	-0.56
dev_knw_recog	0.85	0.81	-0.13	-2.71
confidence	3.27	3.26	-0.01	-0.21
attitude	3.23	3.16	-0.08	-1.57
was_breastfed	0.31	0.35	0.10	1.37
practices_24	4.78	4.71	-0.06	-1.15
practices_agree	3.59	3.59	0.02	0.37
practices_hostility	2.97	3.00	0.05	1.06
(health_knw)	0.45	0.46	0.02	0.40
(was_breastfed)	0.45	0.46	0.01	0.30

Overall P-Value: 0.034

B Additional Plots

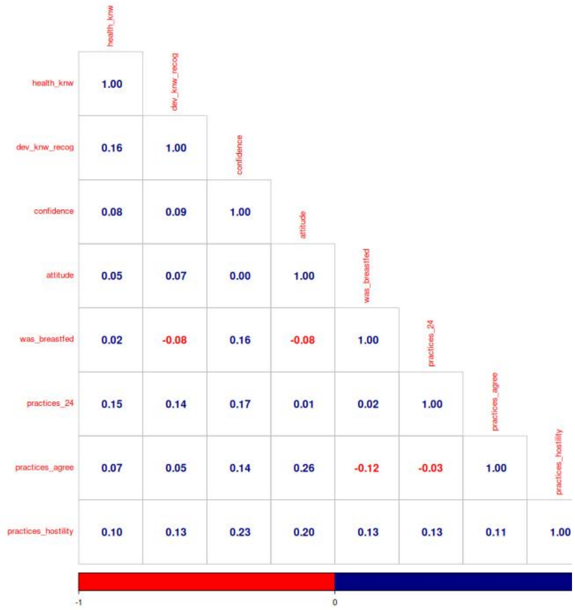


Figure B.1: Construct Correlations - Serbia

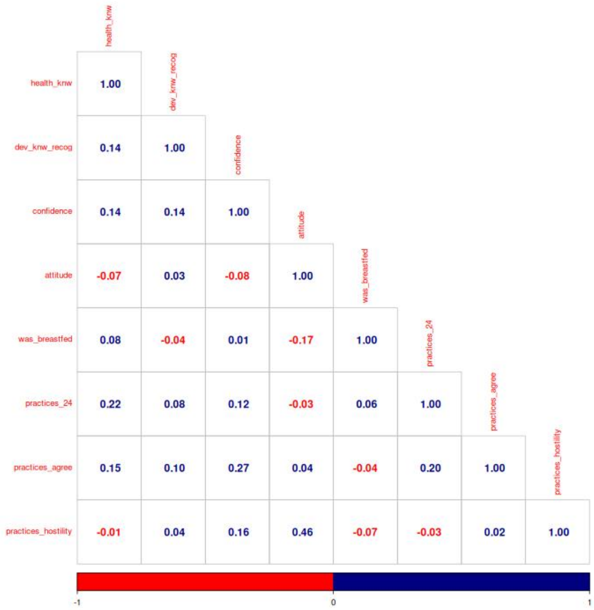


Figure B.2: Construct Correlations - Bulgaria

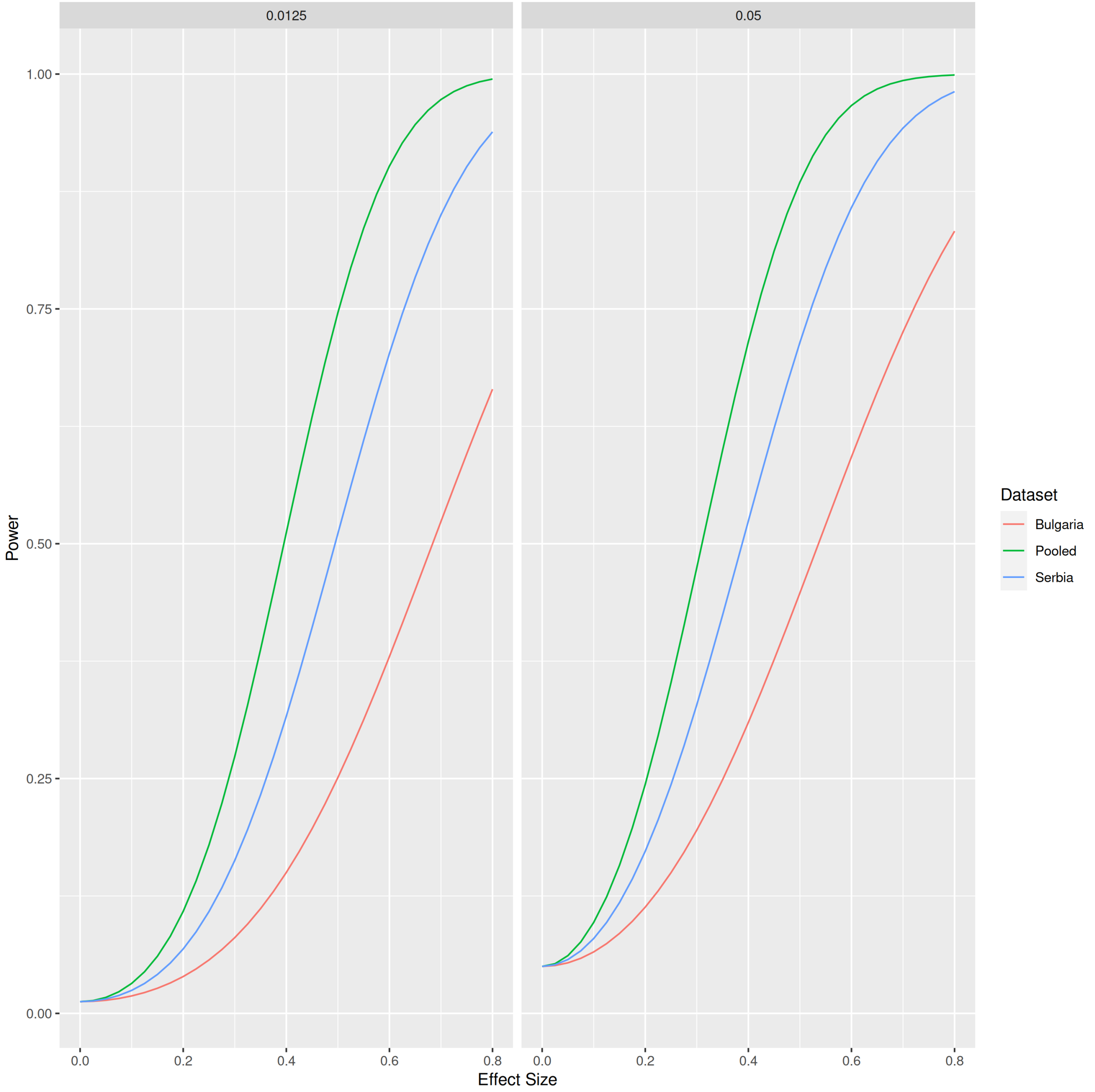


Figure B.3: Power Analysis at 28% Takeup

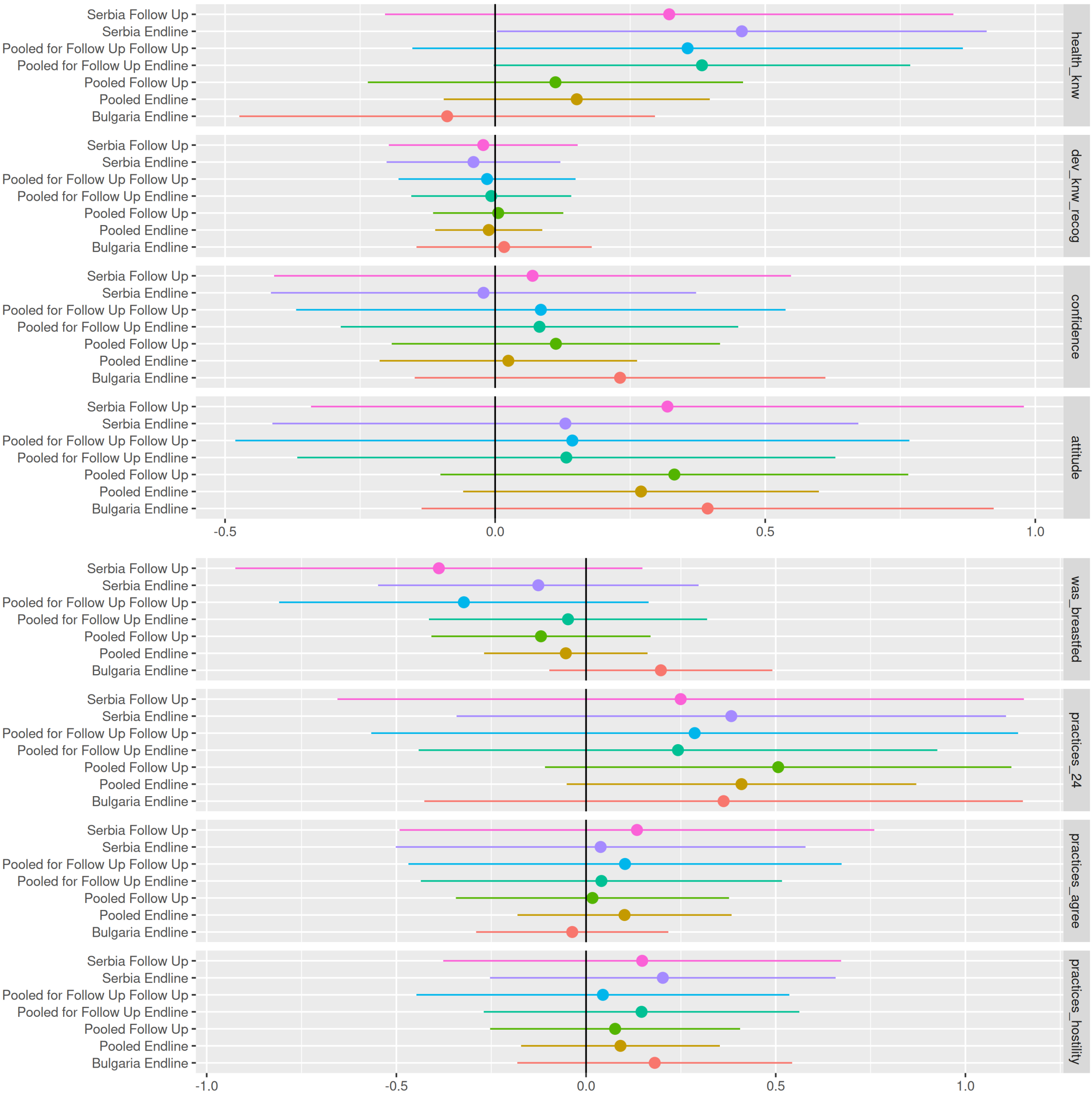


Figure B.4: Adjusted Coefficient Plots of 2SLS in Pooled Dataset

C Additional Tables

Table C.1: Outcome Construct Descriptives Serbia Baseline

name	mean	median	min	max	sd	prop_max	prop_na
Activities Past 24h	5.19	6.00	0	6	1.13	0.54	0.37
Parenting Confidence	3.41	3.50	1	4	0.63	0.42	0.34
Hostile Practices	3.12	3.00	1	4	0.66	0.20	0.37
Attitude to Phys. Punishment	3.01	3.00	1	4	0.89	0.32	0.35
Positive Practices	2.85	2.75	1	4	0.84	0.22	0.37
Child Dev. Knowledge	0.88	1.00	0	1	0.27	0.77	0.34
Vaccine Knowledge	0.77	1.00	0	1	0.42	0.77	0.75
Breastfed	0.46	0.00	0	1	0.50	0.46	0.76

Table C.2: Outcome Construct Descriptives Bulgaria Baseline

name	mean	median	min	max	sd	prop_max	prop_na
Activities Past 24h	4.72	5.00	0.0	6	1.33	0.36	0.55
Positive Practices	3.59	3.75	1.5	4	0.40	0.32	0.55
Parenting Confidence	3.27	3.50	1.0	4	0.64	0.31	0.51
Attitude to Phys. Punishment	3.20	3.00	1.0	4	0.88	0.43	0.52
Hostile Practices	2.99	3.00	1.0	4	0.72	0.11	0.55
Child Dev. Knowledge	0.83	1.00	0.0	1	0.30	0.68	0.50
Vaccine Knowledge	0.65	1.00	0.0	1	0.48	0.65	0.77
Breastfed	0.33	0.00	0.0	1	0.47	0.33	0.79

Table C.3: Attrition: Serbia

stage	count	attrition	treated_attrition	control_attrition	attrition_dif
Started Baseline	5561				
Finished Baseline	3352	0.40	0.40	0.40	0.00
Started Endline	1330	0.60	0.60	0.61	-0.02
Finished Endline	1270	0.05	0.06	0.03	0.03
Started Followup	533	0.58	0.59	0.57	0.03
Finished Followup	520	0.02	0.04	0.01	0.03

Table C.4: Attrition: Bulgaria

stage	count	attrition	treated_attrition	control_attrition	attrition_dif
Started Baseline	4154				
Finished Baseline	1725	0.58	0.58	0.59	-0.01
Started Endline	731	0.58	0.59	0.56	0.03
Finished Endline	624	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 Subgroup Regressions

While the study was not designed for subgroup analysis and thus they are very subject to false positives from multiple testing, it is instructive to run some subgroup regressions.

In particular, we will run the regression on the subset that failed the vaccine knowledge question in the first wave (for children aged 0-2).

Table D.1: Vaccine Failures Subgroup OLS - Endline

	Dependent variable:
	health_knw
treatmenttreated	0.05 (0.08)
Observations	190
R^2	0.05
Note:	*p<0.1; **p<0.05; ***p<0.01

E ToT Regressions

Given that there is no significant impact measured in our evaluation regressions, we do not estimate the impact to be anything other than zero. That being said, the point estimates might still be suggestive evidence and one effective way to measure the point estimates of the impact is through an analysis of the treatment effect on the treated (ToT).

We estimate this 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.

Table E.1: Pooled: 2SLS - Endline - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Used App	0.15 (0.10)	−0.01 (0.04)
Adjusted Treatment p-value	0.333	0.797
Weak instruments p-value	1.2e-33	4.9e-84
Wu-Hausman p-value	0.78	0.529
Observations	719	1,996
R^2	0.02	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01

Table E.2: Pooled: 2SLS - Endline - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Used App	0.02 (0.10)	0.27 (0.13)
Adjusted Treatment p-value	0.797	0.162
Weak instruments p-value	1.54e-82	1.36e-82
Wu-Hausman p-value	0.967	0.0465
Observations	1,972	1,961
R^2	0.01	−0.005
Note:		*p<0.1; **p<0.05; ***p<0.01

F Country Regressions

For robustness checks, we run regressions on the individual countries.

G Follow-up Regressions

As discussed in the text, due to low usage and high attrition, we restricted our primary analysis to the endline survey. That being said, we provide the regressions for the follow-up survey as well in this section.

Table E.3: Pooled: 2SLS - Endline - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Used App	−0.05	0.41	0.10	0.09
	(0.09)	(0.18)	(0.11)	(0.10)
Adjusted Treatment p-value	0.714	0.162	0.619	0.619
Weak instruments p-value	3.75e-32	1.17e-80	6.72e-81	4.82e-81
Wu-Hausman p-value	0.963	0.0115	0.639	0.258
Observations	682	1,903	1,904	1,900
R^2	0.02	−0.01	0.004	0.005
Note:				*p<0.1; **p<0.05; ***p<0.01

Table F.1: Serbia: OLS - Endline - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Treatment	−0.03	0.10	0.01	0.05
	(0.04)	(0.07)	(0.05)	(0.05)
Adjusted Treatment p-value	0.731	0.704	0.891	0.704
Observations	304	949	950	948
R^2	0.04	0.01	0.01	0.02
Note:				*p<0.1; **p<0.05; ***p<0.01

Table F.2: Serbia: OLS - Endline - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Treatment	0.12*	−0.01
	(0.05)	(0.02)
Adjusted Treatment p-value	0.0835	0.731
Observations	316	984
R^2	0.03	0.02
Note:		*p<0.1; **p<0.05; ***p<0.01

Table F.3: Serbia: OLS - Endline - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Treatment	−0.01	0.03
	(0.04)	(0.05)
Adjusted Treatment p-value	0.891	0.731
Observations	974	973
R^2	0.02	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01

Table F.4: Bulgaria: OLS - Endline - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Treatment	−0.04	0.01
	(0.06)	(0.02)
Adjusted Treatment p-value	0.746	0.797
Observations	256	678
R^2	0.04	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01

Table F.5: Bulgaria: OLS - Endline - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Treatment	0.07	0.13
	(0.05)	(0.07)
Adjusted Treatment p-value	0.338	0.336
Observations	665	657
R^2	0.01	0.02
Note:		*p<0.1; **p<0.05; ***p<0.01

Table F.6: Bulgaria: OLS - Endline - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Treatment	0.08	0.11	−0.01	0.06
	(0.04)	(0.10)	(0.03)	(0.05)
Adjusted Treatment p-value	0.336	0.396	0.797	0.396
Observations	234	629	629	627
R^2	0.10	0.03	0.05	0.01
Note:				*p<0.1; **p<0.05; ***p<0.01

Table F.7: Serbia: 2SLS - Endline - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Used App	0.46*	−0.04
	(0.18)	(0.06)
Adjusted Treatment p-value	0.0949	0.732
Weak instruments p-value	1.89e-11	5.96e-35
Wu-Hausman p-value	0.128	0.466
Observations	316	984
R^2	−0.001	0.02
Note:		*p<0.1; **p<0.05; ***p<0.01

Table F.8: Serbia: 2SLS - Endline - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Used App	−0.02	0.13
	(0.16)	(0.22)
Adjusted Treatment p-value	0.891	0.732
Weak instruments p-value	1.37e-34	2.49e-34
Wu-Hausman p-value	0.711	0.543
Observations	974	973
R^2	0.02	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01

Table F.9: Serbia: 2SLS - Endline - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Used App	−0.13	0.38	0.04	0.20
	(0.17)	(0.29)	(0.22)	(0.18)
Adjusted Treatment p-value	0.732	0.711	0.891	0.711
Weak instruments p-value	1.95e-11	5.01e-34	2.94e-34	2.51e-34
Wu-Hausman p-value	0.609	0.118	0.941	0.223
Observations	304	949	950	948
R^2	0.04	−0.01	0.01	0.01
Note:				*p<0.1; **p<0.05; ***p<0.01

Table F.10: Bulgaria: 2SLS - Endline - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Used App	−0.09	0.02
	(0.15)	(0.06)
Adjusted Treatment p-value	0.749	0.797
Weak instruments p-value	1.72e-16	8.09e-35
Wu-Hausman p-value	0.296	0.594
Observations	256	678
R^2	0.03	0.01
Note:		*p<0.1; **p<0.05; ***p<0.01

Table F.11: Bulgaria: 2SLS - Endline - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Used App	0.23	0.39
	(0.15)	(0.21)
Adjusted Treatment p-value	0.342	0.342
Weak instruments p-value	1.16e-33	5.23e-34
Wu-Hausman p-value	0.299	0.126
Observations	665	657
R^2	0.01	0.003
Note:		*p<0.1; **p<0.05; ***p<0.01

Table F.12: Bulgaria: 2SLS - Endline - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Used App	0.20	0.36	−0.04	0.18
	(0.12)	(0.31)	(0.10)	(0.14)
Adjusted Treatment p-value	0.342	0.4	0.797	0.4
Weak instruments p-value	6.97e-15	3.6e-32	3.6e-32	3.19e-32
Wu-Hausman p-value	0.0807	0.24	0.812	0.129
Observations	234	629	629	627
R^2	0.05	0.02	0.06	−0.003
Note:				*p<0.1; **p<0.05; ***p<0.01

Table G.1: Pooled for Follow Up: OLS - Follow Up - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Treatment	0.11	−0.005
	(0.06)	(0.02)
Adjusted Treatment p-value	0.348	0.821
Observations	189	562
R^2	0.03	0.02
Note:		*p<0.1; **p<0.05; ***p<0.01

Table G.2: Pooled for Follow Up: OLS - Follow Up - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Treatment	0.03	0.04
	(0.06)	(0.08)
Adjusted Treatment p-value	0.821	0.821
Observations	561	560
R^2	0.03	0.02
Note:		*p<0.1; **p<0.05; ***p<0.01

Table G.3: Pooled for Follow Up: OLS - Follow Up - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Treatment	−0.10	0.09	0.03	0.01
	(0.06)	(0.10)	(0.07)	(0.06)
Adjusted Treatment p-value	0.348	0.821	0.821	0.821
Observations	188	558	558	557
R^2	0.04	0.01	0.02	0.01
Note:				*p<0.1; **p<0.05; ***p<0.01

Table G.4: Pooled for Follow Up: 2SLS - Follow Up - Knowledge and Awareness

	Dependent variable:	
	Vaccine Knowledge	Child Dev. Knowledge
	(1)	(2)
Used App	0.36	−0.01
	(0.20)	(0.07)
Adjusted Treatment p-value	0.388	0.821
Weak instruments p-value	5.14e-10	2.66e-28
Wu-Hausman p-value	0.346	0.406
Observations	189	562
R^2	0.02	0.02
Note:		*p<0.1; **p<0.05; ***p<0.01

Table G.5: Pooled for Follow Up: 2SLS - Follow Up - Confidence and Attitudes

	Dependent variable:	
	Parenting Confidence	Attitude to Phys. Punishment
	(1)	(2)
Used App	0.08	0.14
	(0.18)	(0.25)
Adjusted Treatment p-value	0.821	0.821
Weak instruments p-value	6.39e-28	7.23e-28
Wu-Hausman p-value	0.699	0.643
Observations	561	560
R^2	0.03	0.03
Note:		*p<0.1; **p<0.05; ***p<0.01

Table G.6: Pooled for Follow Up: 2SLS - Follow Up - Practices

	Dependent variable:			
	Breastfed	Activities Past 24h	Positive Practices	Hostile Practices
	(1)	(2)	(3)	(4)
Used App	−0.32	0.29	0.10	0.04
	(0.19)	(0.34)	(0.23)	(0.20)
Adjusted Treatment p-value	0.388	0.821	0.821	0.821
Weak instruments p-value	3.64e-10	1.87e-27	1.87e-27	1.38e-27
Wu-Hausman p-value	0.136	0.401	0.745	0.59
Observations	188	558	558	557
R^2	−0.03	0.002	0.03	0.004
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

H Survey Instrument

Table H.1: Construct Variable Mapping

Domain	construct_variable	variable
Knowledge and awareness	health_knw	know_which_vaccine
Knowledge and awareness	dev_knw_recog	know_social_emotional_dev
Knowledge and awareness	dev_knw_recog	know_cog_dev
Knowledge and awareness	dev_knw_recog	know_phys_dev
Knowledge and awareness	dev_knw_recog	know_lang_dev
Confidence and attitudes	confidence	confidence_deal_emotions
Confidence and attitudes	confidence	confidence_respond_misbehave
Confidence and attitudes	attitude	physical_punishment
Confidence and attitudes	caregiver_well_being	parenting_stress_2
Practices	was_breastfed	breastfed
Practices	practices_24	past_24h_read
Practices	practices_24	past_24h_stories
Practices	practices_24	past_24h_sing
Practices	practices_24	past_24h_outside
Practices	practices_24	past_24h_play
Practices	practices_24	past_24h_draw
Practices	practices_agree	laugh_together
Practices	practices_agree	joke_with_child
Practices	practices_agree	smile_around_child
Practices	practices_agree	play_on_floor
Practices	practices_hostility	snap_at_child
Practices	practices_hostility	lose_patience_punish
Practices	practices_hostility	threaten
Practices	practices_hostility	make_fun_of