This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Online Appendix

O.1. Additional details on related research focused on parental engagement in LMICs for improved educational outcomes

In this section, we provide a comprehensive review of all the existing research, to the best of our knowledge, on the effectiveness of nudging interventions for enhancing parental engagement in LMICs. Current research finds mixed evidence on the effectiveness of nudging interventions, highlighting the importance of accounting for context to identify effective strategies.

Latin America: In Uruguay, Ajzenman et al. (2022) found that behavioral nudges to parents had no significant effect on school attendance or child development. In Brazil, Bettinger et al. (2021) demonstrated that tailored messages significantly improved educational outcomes by emphasizing the importance of school attendance. In Chile, Berlinski et al. (2022) shows that regular information updates to parents can improve math scores and attendance. Similarly, Santana et al. (2019) reported a substantial increase in students' math GPA through SMS nudges to engage in educational activities.

Africa: In Botswana, Angrist et al. (2020) and Angrist (2022) explored the impact of SMS and direct phone calls, and show that these interventions significantly enhance student learning outcomes and parental engagement. In contrast, in Ivory Coast, Lichand and Wolf (2021) and Wolf and Lichand (2023) found that interventions targeting both teachers and parents did not yield significant improvements, suggesting a potential dilution of impact when multiple stakeholders are involved simultaneously. In Sierra Leone, Crawfurd et al. (2021) noted that SMS reminders paired with phone tutorials did not significantly affect test scores but increased educational engagement. Ome and Menendez (2022) in Zambia showed that SMS stories enhanced reading skills through increased family reading activities, emphasizing the role of parental engagement.

Asia: In Bangladesh, Hassan et al. (2023) and Islam et al. (2022) showed substantial improvements in learning outcomes through telementoring and SMS reminders. However, Beam et al. (2022) noted that benefits were more pronounced for higher socio-economic status households. In Pakistan, Geven et al. (2021) used SMS-based nudges to support girls' education during COVID-19, leading to modest improvements in educational expectations and engagement. In Indonesia, Cerdan-Infantes and Filmer (2015) observed positive impacts on parental knowledge and involvement in school management from SMS messages and school meetings.

Central America and Caribbean: In El Salvador, Amaral et al. (2021) analyzed stress-management and caregiving techniques delivered via SMS, revealing gender-specific effects on mental health and caregiving behaviors. In Jamaica, Diaz et al. (2023) found that SMS nudges positively influenced caregivers' attitudes towards non-violent parenting.

These studies collectively underscore the complex interplay between the method of intervention, the socio-economic background of the recipients, and the specific educational outcomes targeted, revealing how

varying contexts can influence the effectiveness of educational interventions. We contribute to this literature in three distinct ways: (i) We present the first large-scale study focused on early childhood education and associated parental engagement in the LMIC context; (ii) We conduct the first comparative analysis of two different types of information salience nudges (peer and self comparison) and demonstrate their effectiveness in increasing engagement; (iii) We are the first to provide evidence of the value of machine learning-based targeted nudging in the ECE context through an RCT. Our findings suggest substantial gains from carefully designed and implemented targeting policies.

O.2. Field visit to Anganwadi Centers

We visited multiple government run day-care schools (Anganwadi centers) with the collaborators in India. The visits included semi-structured interviews and were primarily aimed at: (i) understanding the operational context of Early Childhood Education in India; (ii) collecting perspectives on how Rocket Learning's digital communities are being perceived by both Anganwadi teachers as well as the parents. In total, the team visited eight Anganwadi centers in the northern states of Rajasthan and Uttar Pradesh and interacted with 10 Anganwadi teachers and more than 40 parents. We summarize the key insights from our visits below.

Anganwadi teacher interviews: Each interview with the Anganwadi teacher lasted for approximately 15 minutes. Teachers were interviewed individually at their respective day-care schools. These interviews revealed that Anganwadi teachers are primarily women, and they are responsible for multiple tasks at the center in addition to teaching. In particular, Anganwadi workers are responsible for providing services related to health, food, and nutrition in addition to pre-school education. The teachers noted that both nutrition (providing food ration) and health services are important drivers of children's attendance in the pre-school. Several of the teachers noted the administrative burden of their jobs which sometimes came in the way of fully devoting their time to improved learning. Teachers overwhelmingly supported Rocket Learning's online communities and found that it had an overall positive effect on the parent engagement, as well as the kids' responses to in-class activities. Teachers noted that they primarily checked the groups to monitor parents who were completing the tasks regularly. The teachers also noted that most families who send their kids to the Anganwadi centers were highly resource constrained. Furthermore, many families shared a single smart-phone. However, according to the teachers, in spite of these constraints, most parents perceived Rocket Learning's digital interventions very positively and wanted to finish all the activities assigned on the groups. Overall, these insights highlighted the positive outlook of Anganwadi teachers towards Rocket Learning and underscored the value of designing digital interventions that could assist teachers in increasing engagement on the digital groups.

Parent interviews: Each interview with the parent lasted approximately 10 minutes. The interviews with parents were conducted either individually or in groups when the parents were visiting their respective Anganwadi center. Rocket Learning's digital interventions were perceived very positively by the parents who noted that they completed different activities and tasks with their kids regularly. Parents frequently mentioned how Rocket Learning's daily activities were easy to understand and fun to do with their kids. As a result, many parents noted regularly checking the group for assigned activities. Major reasons for not being active on the group were either a lack of availability of the phone recharge or a lack of time. Finally, parents were competitive and noted that behavioral interventions such as group-based weekly report cards encouraged them to finish activities regularly.

O.3. Additional details on different treatment interventions and sample report card

In this section, we provide more details on the three nudges that we designed for the experiment.

Peer comparison: Peer comparison nudges are generated by informing parents of their rank based on their activity in the previous week. In particular, consider a focal parent *i* in group g in week *j* of the experiment. Also, let $Y_{i,j-1} \in [0,6]$ denote the number of days, the parent finished any activity. Finally, let $\mathcal{I}_g$ be the set of all parents in group g. Then, for all $i \in \mathcal{I}_g$, we sort the corresponding parent level activity scores ($Y_{i,j-1}$), in a decreasing order. The rank of parent i is simply their position in this sorted ordering.[1] In each week *j*, all parents in the group are sent this ranking in the form of a leader-board (Figure 3a). The top two parents are identified in the leaderboard, and the leaderboard is sent to each parent via a one-on-one WhatsApp message. Furthermore, the leader-board is personalized by showing the rank of the focal parent along with the parent who was ranked right behind the focal parent. Ranks of all parents are not disclosed to everyone to ensure that the leaderboard is not long. The leaderboard is also accompanied by a message that encourages parents to be at the top of the ranking by finishing up more activities in the subsequent week. Parents at the top of the ranking are congratulated via an additional text that encourages them to continue finishing activities and remain on the top of the leaderboard.

Self comparison: Self comparison nudges are also constructed by collecting information on the focal parent's activity in the past week. Consider a focal parent *i* in group g in week *j* of the experiment. Recall that $Y_{i,j-1} \in [0,6]$ denotes the number of days the parent finished any activity in week $j-1$. Then, the personalized feedback discloses $Y_{i,j-1}$ to the focal parent and then elicits a commitment for the following week by asking the parent to choose from a set $\mathcal{S}_{i,j}$ of available targets (goals). The set $\mathcal{S}$ is constructed using the following logic:

[1] We resolve ties arbitrarily in this case.

$$\mathcal{S}_{i,j} = \begin{cases} \{2,3,4\} \text{ if } Y_{i,j-1} \in \{0,1\}, \\ \{4,5,6\} \text{ if } Y_{i,j-1} \in \{2,3\} \text{ and}, \\ \{5,6\} \text{ if } Y_{i,j-1} \in \{4,5,6\}. \end{cases}$$

The set of available targets contains options at least as high as the number of days they were active in the previous week. Furthermore, as before, parents are also rewarded based on whether they accomplish their target. In Figure O.1, we show three digital certificates that were sent from the second week of the experiment. The certificate on the left was sent to all parents who accomplished their target. Similarly, the certificate on the right was shown to all parents who were active for six days. They were encouraged to continue their performance and were promised a digital Trophy if they continued to get another week with 6 activities. Finally, parents who remained very active for more than two consecutive weeks were sent an additional text message to create a "hat-trick" of very active weeks. Notice that the focal parent's activity drives the information feedback, as well as the target that they can choose from in the subsequent week. Therefore, this nudge is fully personalized.

Figure O.1 Self Comparison Nudges

Note. Completion certificates were sent to parents who were assigned the self-comparison nudge. On the left, the certificate is sent to parents who meet their target for the first week. The text in the certificate translates to "Congratulations! You have received one star on achieving your target. Definitely share this achievement on the group." The middle (right) certificate is sent to parents who consecutively achieve their targets for two (three) weeks during the experiment.

Group-wise report cards: Rocket Learning also sends weekly report cards on the groups to inform participating parents of their activity in the week. Figure O.2 contains a mock-up of these report cards. Each row corresponds to a parent in the group and each column corresponds to different days of the week. Rocket Learning sends daily activities on all days of the week except on Sunday. Activity completion on a given day is represented with a smiley emoji. These report cards are used to ensure that parents are actively tracking the activity on the group and to encourage them to complete activities every day of the week. Note that the report card is not personalized and it is sent on the group, instead of being sent as a one-on-one message.

Figure O.2 Report Cards For Weekly Updates

Note. Rocket Learning sends weekly report cards on the groups to update all parents on days when they completed the activity. Note that parents are not rank-ordered and the report card is sent on the group directly and not as a 1-1 message.

O.3.1. Details on Video Testimonials arm

In addition to the peer and self comparison nudges, we also tested the value of non-personalized video testimonials. In what follows, we provide additional details on the video-testimonial nudges.

Video testimonials: To test the impact of non-personalized nudges based on social comparison, we recorded 6 videos, each of comparable length, featuring different parents on rocket learning's platform. In each video, a parent (not part of the experiment) discusses the positive impact that completing Rocket Learning's activities has had on their children. Finally, the video ends with the parent encouraging others to also complete activities. The videos are accompanied by a short text message encouraging parents to watch and learn from other parents. Note that this nudge is non-personalized since the focal parent's information is not used to personalize the testimonials. Below, we present the translated transcripts of two such sample videos along with their total duration.

Transcript 1 (45 seconds): *Let's find out about the experience of parents who are part of the digital groups. For 3 years my kids are connected with the group. The digital group is very useful. My kids have learnt how to color and they also enjoy playing. They also study for an hour or so and I also enjoy spending time with them. The teacher also sends kudos when I finish the task. My kids also ask about the activities of the day and then I sit with them to finish the activity and they are also happy to finish them.*

Transcript 2 (43 seconds): *Let's find out about the experience of parents who are part of the digital groups. My daughter is 4 years old and she has been part of the digital classroom for 2 years now. She has*

learnt a lot in the group including coloring and learning alphabets. I want all parents to join these digital classrooms so that their kids can also learn like my daughter. The digital classrooms also have special sunday activities. For example, last time they taught us Yoga and now my daughter has started doing yoga everyday. Hence, I would like to ask that you also participate in these activities and learn from them.

Empirical analysis: In contrast to peer comparison and self comparison arms, we do not find any evidence of statistically significant impact from video testimonials, as shown in the results in Table 1. While we cannot causally identify the underlying mechanism behind the inefficacy of video testimonials, our field interactions suggest that there may be three potential reasons behind these results. First, in contrast to the other two interventions, video testimonials used significantly more data for users. Current literature (Ramdas and Sungu 2022) suggests that data usage can affect digital choices of low-income households specially in resource-constrained settings. Second, in contrast to the other two interventions that were personalized, video testimonials were not personalized and a standard message was recorded in each of the variants. Finally, the operational implementation of the video testimonials unintentionally included a black screen as the video thumbnail. This black screen may have confused some users and discouraged them from clicking on the video, thereby rendering the nudge ineffective. These results suggest that while nudges can indeed prove effective, a cautious and context-specific approach is warranted to ensure the ultimate success of such interventions. Further, this result underscores the significance of designing initiatives that are resource-efficient and align with the realities of the target audience.

O.4. Additional details on the first RCT

O.4.1. Survey to understand reasons for parental disengagement

To better understand the reasons behind parents' reduced activity on Rocket Learning's digital classrooms, we rely on data from an internal survey done by Rocket Learning. Nearly 1000 users who were (i) active for at least 4 weeks in their lifetime and (ii) inactive for the preceding two weeks were randomly selected for a phone survey. Out of the 948 users who were called, 479 users picked up the call and finally 274 completed the survey.

The survey was divided into two parts: (i) baseline demographic data and (ii) more details on reasons for reduced engagement. The following demographic data was collected for each user: (i) Who is the primary educator? (mother/father/siblings) (ii) What is their education level? (upto 10th/upto 12th/graduation/post-graduation) (iii) Does the primary educator have a smartphone (jio or non-jio phone)? (iv) Do you have the number of your Anganwadi teacher saved in your phone in case you have any doubts? (yes/no). We ask the following questions to understand the perceived value that users get from Rocket Learning's groups and details on reduced engagement: (i) Do you remember getting added to a Anganwadi Whatsapp group? (ii) Have you received any communication from teacher on program overview / benefits? (yes/no) (iii) Do you remember seeing anything on the group? (yes/no) (iv) If yes, then what did you see? (v) What according to

Table O.1 Summary Statistics of Different Features

	Control	Peer Comparison	Self Comparison	Testimonials
Active Users	5.47	5.50	5.48	5.37
	(4.95)	(4.98)	(5.02)	(4.91)
Total Users in the Group	21.95	22.24	21.57	21.28
	(17.28)	(18.49)	(16.06)	(15.64)
Avg. Daily Active Users	0.66	0.66	0.66	0.65
	(0.81)	(0.78)	(0.80)	(0.82)
Avg. parent Msgs	1.74	1.68	1.70	1.68
	(2.54)	(2.33)	(2.40)	(2.45)
Avg. Daily Teacher Msgs	0.21	0.27	0.21	0.22
	(0.64)	(1.71)	(0.61)	(0.58)
Degree Centrality	0.37	0.37	0.37	0.38
	0.27	0.26	0.26	0.27
Closeness Centrality	1.39	1.22	1.31	1.30
	(4.62)	(3.47)	(3.69)	(3.95)
Betweenness Centrality	0.07	0.08	0.07	0.07
	(0.08)	(0.08)	(0.08)	(0.08)
Category Age 4-6	0.40	0.41	0.40	0.40
	(0.49)	(0.49)	(0.49)	(0.49)

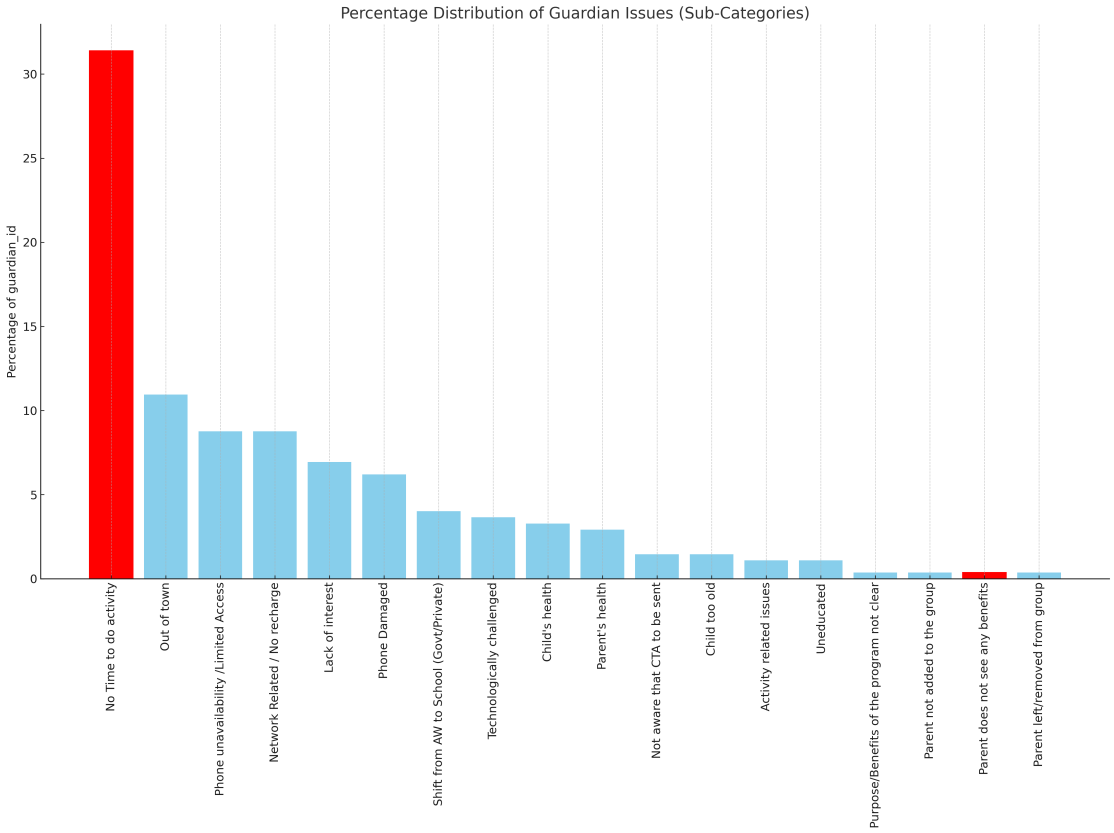
Note: We give a brief description of each feature. Active users: number of unique parents that interacted in their group at least once; Avg: daily active users: daily average of the number of parent that interact in a given group; Avg. parent msgs: average number of messages parents send in a day in a given group; Avg. daily teacher msgs: average number of messages teachers send in a day in a given group; category Age 4-6: whether a given group belongs to the younger age group; the other features are network measures of the graph representation of a given group, where parents are nodes, and edge weights are higher among parents when they interact sequentially in the group. More details are given in the Appendix O.6.2. Notice that there is no significant statistical difference between the features in different arms confirming that the samples are balanced across four arms.

you is the purpose of the group and its benefit? (Open-ended) (vi) Dropoff reasons (drop-down with options as described in Figure O.4.1) and (vii) Call summary (open-ended). In Figure O.4.1 we focus on the reasons for drop-off and find that users overwhelmingly select time constraints and other operational challenges for dropping off. Interestingly, only 1 out of 274 users who completed the survey said that they did not see any value in the tasks assigned on the digital group. This is also aligned with the findings from our field visits where an overwhelming majority of users perceived Rocket Learning's digital tasks positively.

O.4.2. Robustness Checks

In this section, we show that our empirical results are consistent under a series of robustness checks. First, table O.2 presents the first set of robustness checks under multiple specifications, where the outcome variable is either active users or total number of parent messages. We next turn to a task-level analysis to see whether our nudges actually translate into completed activities. Specifically, we construct a refined task-

Figure O.3 Reasons For Reduced Engagement of Parents Identified from A Large Survey



Note. Results from survey responses of 274 users. Most users (32%) note that their primary reason for reduced engagement on Rocket Learning's groups is lack of time. This is followed by other operational issues related to network connectivity, phone availability etc.). Interestingly, only 1 out of the 274 survey respondents said that they reduced engagement since they did not find any benefits in engaging on the digital groups.

completion metric from parents' non-chat messages. In what follows, we first define this task-completion measure and then present the corresponding regression results.

We start by detailing how we construct the task completion outcome. Because image or video submissions are not explicitly labeled as responses to specific tasks, there is an inherent *attribution problem*, hence we rely on temporal proximity to approximate the correct matching of messages and activities. In particular, to construct a *group-day level task completion rate*, we use a simple matching algorithm: for each non-chat message (typically image or video) sent by a parent, we attempt to match it to the most recent activity in the preceding week shared in the group. We clarify that since our outcome of interest is the total number of interactions matched to some task, rather than which specific task receives credit, unique matching is not central in our setting.

Table O.2 Estimated Impact of Nudges on Active Users

Dependent Variables:	Active Users					Total Messages	Active Users	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	OLS	Two-way Clustering	DID	Zero-inflated	Ancova		4-6 years	6-8 years
Variables								
Peer Comparison	0.1055***	0.1349***	0.1360***	0.0893***	0.160***	0.1390***	0.1053***	0.0837***
	(0.0226)	(0.0250)	(0.0241)	(0.0235)	(0.016)	(0.0301)	(0.0311)	(0.0311)
Self Comparison	0.0673***	0.0799***	0.0744***	0.0478***	0.084***	0.0801**	0.0644**	0.0578*
	(0.0226)	(0.0263)	(0.0244)	(0.0248)	(0.008)	(0.0319)	(0.0305)	(0.0312)
Fixed-effects								
Other Controls	Yes	Yes		Yes	Yes	Yes	Yes	Yes
Day FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Group FEs			Yes					
Fit statistics								
Observations	194,208	194,208	316,001	194,208	194,208	194,208	78,652	115,556
Pseudo R^2	0.1440	0.2584	0.3661	0.2206	0.269	0.3241	0.1450	0.1502

Column (1) presents results from estimating Equation 1 with OLS. Column (2) presents results from estimating Equation 1 with two-way clustering at group-day level. Column (3) presents results from estimating the following DID model by also including pre-treatment data. Note that coefficient estimates of $Peer \times Post$ ($Self \times Post$) are included in the column. Column (4) provides the results of the Poisson regression analysis with a zero-inflation adjustment, designed to handle excess zeros in the dataset. The reported R^2 metric is the McFadden's R^2. Column (5) provides the results of the main Poisson regression estimate with ANCOVA adjustment, which basically adds the pre-treatment periods average outcome as a control. (6) presents results with Total Messages as the dependent variable. Column (7) and (8) present results with a subsample of groups with parents in the age group 4-6 years and 6-8 years respectively. Other Controls include controls for age groups of children, district fixed-effects, total users in the group, avg. number of daily teacher messages pre-implementation, betweenness centrality, degree centrality and closeness centrality. Clustered (by group) standard-errors in parentheses. *Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Then, we estimate the *Poisson regression* analogue of our main specification[2] with the task completion based outcome. Results from this analysis, as presented in Table O.3 mirror our main findings. Both self and peer comparison nudges significantly increase the likelihood that parents complete tasks, with peer nudges showing a larger effect on average, as observed previously. These results offer complementary evidence that the nudge intervention not only drives messaging activity but also translates into higher task engagement, reinforcing its pedagogical value. Importantly, for robustness, we also set the feasible matching window to 3 and 5 days instead of 7, and results in Table O.3 remain consistent. To test robustness with respect to how interactions are matched to tasks, we also compare our baseline matching approach to an alternative rule that matches each interaction to the oldest feasible task within the admissible window, and the results still remain consistent (see column 4 of Table O.3)

[2] We additionally note that although the outcome in this case is the average number of completed activities per parent, which is continuous since it is an average of count data. As such, it remains appropriate to estimate a Poisson pseudo-maximum likelihood (PPML) model. Importantly, it still provides consistent estimates of the conditional mean even if the dependent variable is not an integer count, as long as the conditional mean is correctly specified as $\mathbb{E}[Y_i|X_i] \propto e^{\beta^T X_i}$ (Gourieroux et al. 1984, Silva and Tenreyro 2006).

Table O.3 Poisson Regression Estimates for Completion Rate

Dependent Variable:	Average Completion Rate						
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Variables							
Peer Comparison	0.097***	0.097***	0.099***	0.098***	0.104***	0.099***	0.021***
	(0.031)	(0.031)	(0.029)	(0.029)	(0.029)	(0.029)	(0.007)
Self Comparison	0.081**	0.081**	0.070**	0.070**	0.074**	0.070**	0.016**
	(0.032)	(0.032)	(0.030)	(0.030)	(0.030)	(0.030)	(0.007)
Fixed Effects							
Day FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ECE & District FEs	No	No	Yes	Yes	Yes	Yes	Yes
Model Type	Poisson	Poisson	Poisson	Poisson	Poisson	Poisson	OLS
Window size (days)	7	7	7	7	3	5	7
Matching Type	Earliest	Earliest	Earliest	Oldest	Earliest	Earliest	Earliest
Fit Statistics							
Observations	194,208	194,208	194,208	194,208	194,208	194,208	194,208
Pseudo R^2	0.00035	0.02754	0.0600	0.06073	0.06448	0.06150	0.08794

Notes: Other controls include age groups of children, district fixed-effects, total users in the group, average number of daily teacher messages pre-implementation, betweenness centrality, degree centrality, and closeness centrality. Clustered (by group) standard errors in parentheses. Significance codes: ***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.1$.

O.4.3. Do the effects of behavioral nudges persist?

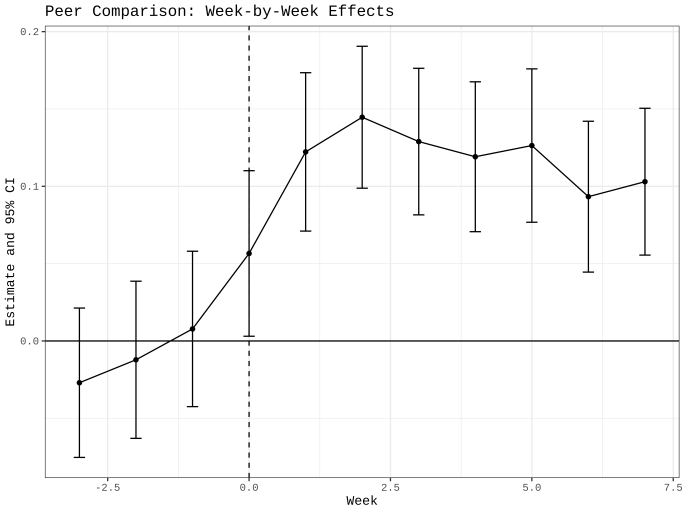
In order to further investigate the long-term temporal effects of behavioral nudges from the experiment, we collect additional data for one month post-implementation and estimate the following Poisson regression model:

$$\log(active_{gt}) = \sum_{w=-3}^{8} \Big(\alpha_{pe,w}\text{Peer}_g + \alpha_{se,w}\text{Self}_g\Big) \times \text{Week}_w + \gamma X_g + \omega_t + \epsilon_{gt}, \tag{O.1}$$

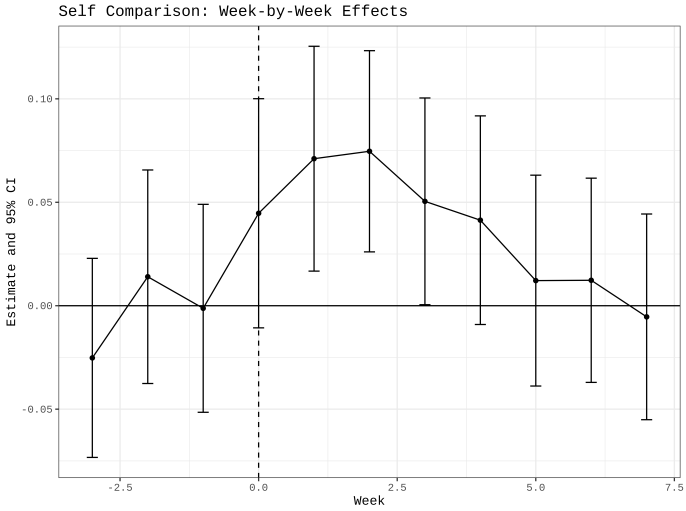
The variable Week_w is an indicator variable that is 1 if the observation from day t is w weeks after the launch of the experiment and 0 otherwise. The reference week in this specification is week 0, one week prior to the implementation. Thus, $\alpha_{pe,w}$ captures the difference between the peer feedback arm and the control arm in week w as compared with the baseline week 0. Figure O.4 illustrates the results and plots the coefficients for the three treatment arms. We find that the effect of peer feedback is statistically significant and sustained both during and after the experiment. However, the effect of self comparison is statistically significant only during the implementation and not after the treatment is stopped. We note that our nudging strategy was of low intensity since each parent was nudged only once per week. Hence, to ensure that the nudges remain effective, one could continue the low-intensity nudges for longer periods. However, long-term nudging even at low intensity might have diminishing effects, and a complete analysis of such a strategy is out of the scope of the current paper. Nevertheless, we consider it an important direction for future research.

Figure O.4 Temporal Effects of Nudges

(a) Impact of Peer Comparison



(b) Impact of Self Comparison



Note. The difference between control and treatment arms in pre-treatment weeks is not significant. The analysis suggests that peer feedback continued to have positive and statistically significant impact both during and after the treatment was taken away. In contrast, self comparison had a statistically significant and positive impact till the fourth week while video testimonials did not have a statistically significant positive impact.

O.5. Details on the M&E survey data and analysis

We extend our analysis beyond engagement activity by linking parental WhatsApp activity to children's actual educational gains. Using Rocket Learning's Monitoring & Evaluation survey of 450 children (ages 3–6), which measures cognitive, pre-literacy, motor, socio-emotional, and executive-function skills at baseline and endline, we show that higher parental engagement on WhatsApp groups is positively and significantly associated with improvements in standardized early childhood learning outcomes. In what follows, we detail the M&E survey dataset and its construction.

Survey design and actual learning outcomes: Rocket Learning's Monitoring and Evaluation (M&E) team conducts annual surveys to assess the impact of Rocket Learning's digital tool on children's actual learning outcomes. These surveys are targeted for children aged between 3-6 and are used assessing development domains like pre-numeracy, cognitive, pre-literacy, motor, and socio-emotional skills. These surveys use standard and well-established standardized assessments from Annual Status of Education Report (ASER) ASER Centre (2025), The International Development and Early Learning Assessment (IDELA) Save the Children (2023), Center for Early Childhood Education and Development (CECED) Center for Early Childhood Health and Development (2025), and state curricula.

Each child is assessed through a series of questions related to different development domains. Survey responses from each child are then aggregated (composite of cognitive, pre-literacy, executive function, socio-emotional, and motor skills subcategories) into an aggregate overall development score. For instance,

pre-literacy domain is assessed based on performance on four sub-tasks (listening comprehension (5 questions), letter identification (10 questions), initial sound recognition (4 questions) and word recognition (3 questions)). Then, a sub-task weighted score is calculated for each domain. The final overall development score is the average score across all the developmental domains. Children usually take 45 minutes to an hour to complete the survey. And on average, a surveyor covers around 5 children a day.

We leverage survey data collected from nearly 450 children from a representative district in the state of Maharashtra, which was collected as part of the 2024-25 survey cycle. The survey is administered to children in households who sent at least one rich media (image/video) message on the WhatsApp group in the prior two months before the baseline survey date. The survey also follows a Pre-Post evaluation design, and uses both baseline and endline responses to estimate changes over time. The dataset contains baseline and endline ECE learning scores (composite of cognitive, pre-literacy, executive function, socio-emotional, and motor skills subcategories) for each child in the sample; child age group (3-4 years, 4-5 years, 5-6 years) and child gender; father's and mother's highest education level (below 10th grade, 10th grade, 12th grade or graduate/post-graduate), and finally, the baseline and endline survey dates for each unit. On average, children scored 45 out of 100 on the overall ECE score at baseline, and 65 at the endline, representing a 44% improvement over the survey period. The average interval between the baseline and the endline surveys was approximately 243 days.

In order to identify whether engagement on Rocket Learning's platform was positively associated with improvements in learning outcomes, we augment this survey data with the interaction data from Rocket Learning's WhatsApp groups during the period between baseline and endline surveys. Using this data, we estimate the following OLS specification:

$$EndlineScore_i = \beta PctActiveDays_i + \gamma BaselineScore_i + \theta^T \mathbf{X}_i + \epsilon_i,$$

where $EndlineScore_i$ ($BaselineScore_i$) is the score of child i in the Endline (Baseline) Survey, $PctActiveDays_i$ is the fraction of days in this period when a non-chat (image/video) message was sent on the group by guardian of child i. Finally, $\mathbf{X}_i$ is the following set of time-invariant child-level features: age, gender, father and mother's education level.We also cluster standard errors at the group level to account for correlation across observations from the same group.

The results from this specification are presented in Table O.4. The results show that increased parental activity has a positive and statistically significant impact on learning outcomes. To further test the robustness of this result, we construct two alternate engagement metrics. First (Column 2), we calculate the average number of active days per week (where a day is considered active if at least one nonchat message is sent in the group). Second (Column 3), we calculate the average number of non-chat messages per day on the group. For both these alternate metrics, the coefficient remains positive and statistically significant. These

findings provide empirical evidence that an increase in engagement on the groups is positively associated with improvements in children's learning outcomes.

It is important to note that although the preceding results are based on a selected sample restricted to households with prior rich-media engagement at baseline, the mechanism linking parental engagement to children's learning outcomes is likely relevant for the broader population. Notably, our experimental results show that nudges are more effective for parents with low pre-treatment engagement. Taken together, these findings suggest that the implications of the M&E analysis for the interventions are not limited to already-engaged parents.

Table O.4 Comparing Engagement Metrics and Standardized ECE Endline Scores

Dependent Variable:	ECE Endline Score		
Model:	(1)	(2)	(3)
Pct. Active Days	7.581**		
	(3.699)		
Avg. Active Days per Week		1.152**	
		(0.586)	
Avg. Nonchat Msgs per Day			4.896**
			(2.502)
Baseline ECE Overall Score	0.222***	0.220***	0.218***
	(0.045)	(0.051)	(0.048)
Observations	456	456	456
Adjusted R^2	0.280	0.288	0.288

Notes: Each column reports results from a separate OLS regression. Clustered (by group) standard-errors in parentheses. The dependent variable is the standardized overall ECE endline score. Each model includes a distinct parent engagement metric as the main predictor. Controls include child age, gender, parental education, and baseline score.

Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

O.6. Details on the nudge targeting policy

O.6.1. Candidate policy generation

In this section, we provide more details on the candidate policy generation. We use the framework developed by Fernández-Loría et al. (2023) to discuss how we use either outcome prediction or treatment effect estimation to estimate different candidate targeting policies. For ease of exposition, we redefine an optimal targeted nudging policy $\pi : \mathbb{R}^d \to [K]$ which is a mapping function that maps a d-dimensional feature vector to an action that optimizes some outcome of interest Y based on the selected action. Let $Y(k|X = x)$ denote

the outcome of interest when an individual (or sub-population) with features x is targeted with the nudge k. Then the optimal targeted nudging policy $\pi^*(x)$ is given by

$$\pi^*(x) = \arg\max_{k \in [K]} \mathbb{E}[Y(k|X=x)]\,. \tag{O.2}$$

Targeting using outcome predictions: Consider Problem (O.2) and let $\hat{\mu}(x,j) = \hat{\mathbb{E}}[Y(j|X=x)]$ denote the expected predicted outcome of interest, when a sub-population with features x is sent nudge j. Then, a targeted nudging policy can be simply defined as

$$\hat{\pi}_\mu(x) = \arg\max_{k \in [K]} \hat{\mu}(x,j)\,. \tag{O.3}$$

Note that $\hat{\mu}(x,j)$ can be estimated using different predictive ML methods and a targeted policy in this case is a two-step procedure where in step (i) we use predictive ML methods for estimating $\hat{\mu}(x,j)$ [3] and in step (ii) we use these estimated outcomes to select the nudge that maximizes the expected outcome of interest.

Targeting using treatment effect predictions: Consider Problem (3) and let $\hat{\tau}(x,j) = \hat{\mathbb{E}}[Y(j|X=x)] - \hat{\mathbb{E}}[Y(0|X=x)]$ denote the Conditional Average Treatment Effect (CATE) when a sub-population with features x is sent nudge j. Note that treatment 0 corresponds to the baseline control arm where no-nudge is sent. Then, a targeted nudging policy can be simply defined as

$$\hat{\pi}_\tau(x) = \arg\max_{k \in [K]} \hat{\tau}(x,j)\,. \tag{O.4}$$

$\hat{\tau}(x,j)$ can be estimated using various machine learning models. We used Doubly Robust Trees and Doubly Robust Forests (Athey et al. 2019, Wager and Athey 2018, Oprescu et al. 2019, Battocchi et al. 2019). A two-step procedure, as previously discussed, can then be used to construct a targeting policy.

Note that both classes of policies have their advantages and are popular in practice. For example, outcome predictions are easy to use and can also directly be leveraged to make (often biased) treatment effect estimations. Treatment effect estimation is often used to estimate the *marginal* change in the outcome of interest from assigning a particular treatment.

O.6.2. Feature engineering

In this section, we provide additional details on the feature engineering we performed to create different features to train the different targeting policies.

[3] For a comprehensive list of ML regression methods we implemented, see §O.6.3

Raw interaction data: We collect detailed data from interactions on the group. This data can be categorized into two primary interaction types: (i) parent interaction, and (ii) teacher interactions, each described as follows:

1. Parent interactions: Within this category, each entry represents an individual interaction of a specific parent within their respective group. Every interaction is associated with a message type, which can be a chat message, video message, audio, or document. Parents are uniquely identified by their parent ID, while their respective groups are uniquely identified by a group ID. For chat messages, the message text is also known, as well as the duration of audio messages. Additionally, the timestamp of each interaction is recorded.
2. Teacher interactions: Within this category, each entry represents an individual interaction of a specific teacher within their respective group. For each teacher interaction, the message type (chat message, video message, audio, or document), timestamp, chat message text, and audio message duration are recorded.

We also collect information on the district where the group is located and whether it corresponds to the age group of 4-6 years or 6-8 years. Based on this data, we construct group-level and parent-level features to estimate different targeting policies. Note that all the features were constructed based on pre-treatment data (8 weeks prior to the start of the experiment).

Group-level features: These features can be briefly summarized in two categories: group interaction-related features and group network measures.

- Active users pre-implementation: Number of unique parents that interact in the group at least once during the pre-implementation period.
- Average daily active parents pre-implementation: Average number of unique parents that interact daily in a specific group over the pre-implementation period.
- Average non-chat parent messages pre-implementation: Average number of non-chat messages parents send daily in a specific group over the pre-implementation period. Non-chat messages, such as videos or images showing attempts at activities, are particularly relevant because responses to Rocket Learning's daily activities are in this format.
- Average daily teacher messages pre-implementation: Average number of total teacher messages (including chat and non-chat) sent daily in a group over the pre-implementation period.
- Age indicators: Indicator of whether the group corresponds to the 4 to 6 age group or 6-8 age group.

To calculate group network measures, we represent group interactions as a weighted directed graph where edges between parents in the group are directed. This choice is predicated on the idea that parent i can influence parent j, but parent j does not necessarily influence parent i. we establish a directed edge from parent i to parent j if parent j interacts after parent i in the group on the same day during the last month of interaction. The weights of these edges correspond to the number of times parent i responds

after parent j throughout the entire month. Using this graph representation, we create an extensive set of network measures (weighted average degree centrality, average closeness centrality, average betweenness centrality, maximum in-degree and out-degree centrality). For a detailed discussion on the graph measure, we refer the interested readers to Newman (2010). Finally, at the individual parent level, we use two sets of features: (i) the average number of non-chat messages sent by that parent in the group in a day over the pre-implementation period and (ii) Rocket Learning's internal classification of individual parents based on their long-term activity on groups (inactive/active/frequent/super-active users).

O.6.3. Hyperparameter Optimization

We train an extensive set of ML predictors using the features described above. The primary outcome of interest is the number of days an individual parent is active. The ML models were trained on an 80% split of the data using a 3-fold cross-validation with grid random search to choose hyperparameters, and performance was reported on a holdout test set of 20% of the data.

Particularly for outcome prediction based targeting policies, we train the following regression models: Decision Tree, Adaboost, Gradient Boosting, Random Forest, Lasso, and Linear Regression, and we use the following grid of hyperparameters for each of these models:

- **Decision Tree**:
  - `max_depth`: Maximum depth of the tree, range from 1 to 100.
  - `min_samples_split`: Minimum number of samples required to split an internal node, values 2, 5, and 10.
  - `min_samples_leaf`: Minimum number of samples required to be at a leaf node, values 1, 2, and 4.
  - `max_features`: Number of features to consider at each split, options include 'auto', 'sqrt', 'log2', and None.
- **Random Forest**:
  - `n_estimators`: Number of trees in the forest, ranging from 100 to 500 with 10 steps.
  - `max_features`: Number of features to consider at each split, options include 'auto', 'sqrt', and 'log2'.
  - `max_depth`: Maximum depth of the tree, with values from 2 to 7.
  - `min_samples_split`: Minimum number of samples required to split an internal node, values 2, 5, and 10.
  - `min_samples_leaf`: Minimum number of samples required to be at a leaf node, values 1, 2, and 4.
  - `bootstrap`: Whether to bootstrap samples when building trees, options include 'True' and 'False'.

- **Gradient Boosting**:
  - `n_estimators`: Number of boosting stages (trees), options include 50, 100, 200, and 300.
  - `learning_rate`: Step size that shrinks the contribution of each tree, with a value of 0.5.
  - `max_depth`: Maximum depth of individual trees, options include 3, 5, 6, 7, 8, and 10.
  - `min_samples_split`: Minimum number of samples required to split an internal node, options include 450, 600, 800, and 1200.
  - `min_samples_leaf`: Minimum number of samples required to be at a leaf node, options include 100, 300, and 500.
  - `max_features`: Number of features to consider at each split, options include 'sqrt' and 'log2'.
  - `subsample`: Fraction of samples used for fitting the trees, with a value of 0.8.
- **Adaboost**:
  - `n_estimators`: Number of boosting stages (base estimators), options include 50, 100, and 200.
  - `learning_rate`: Step size that shrinks the contribution of each base estimator, options include 0.01, 0.1, 0.2, and 0.3.
  - `loss`: The loss function to use, options include 'linear', 'square', and 'exponential'.
- **Lasso**:
  - `alpha`: Regularization strength, options include 0.001, 0.01, 0.1, 1.0, and 10.0.
  - `max_iter`: Maximum number of iterations, options include 100, 500, 1000, and 2000.
- **DR Learners**: For the doubly robust learners (Doubly Robust Forest and Doubly Robust Tree), we use the standard cross-validation procedure from the EconML (Battocchi et al. 2019) package implementation, with the same number of folds as the other previous methods.

Model re-training to account for covariate shift from first to second experiment: To account for the covariate shift observed between the first and second experiments, we re-trained the random forest targeting policy by implementing a weighted version of the model. This approach follows the methodology outlined in Sugiyama et al. (2007), which adjusts for the differences in feature distributions between the train and test datasets by an importance weighting scheme. We utilized training datasets from both the first and second experiments, denoted as $\mathcal{D}_1 = \{(\mathbf{x}_{1i}, y_{1i})\}_{i=1}^{n_1}$ and $\mathcal{D}_2 = \{(\mathbf{x}_{2i}, y_{2i})\}_{i=1}^{n_2}$, respectively. Here, $\mathbf{x}_{ij}$ and y_{ij} represent the vector of covariates and the outcome for the j-th observation in the i-th experiment dataset, and n_1, n_2 are the number of observations in the first and second experiments, respectively. The specific features are consistent across both datasets. Then, we estimate a logistic regression where the dependent

variable $z_i := 1$ if the observation is from the first experiment and $z_i = 0$ otherwise. The predicted probabilities $p_{train}(\mathbf{x}_i) = 1 - p_{test}(\mathbf{x}_i) := \mathbb{P}(z_i = 1 \mid \mathbf{x}_i)$ are used to compute the importance weights for each observation, $w_i(\mathbf{x}_i)$. The final targeting policy we employ in practice is obtained by re-training the Random Forest predictor, with observations being weighted according to w_i.

Table O.5 Group features summary statistics in second RCT data

	Control	Target	Uniform
active users	5.26	5.07	5.25
	(4.64)	(4.33)	(4.62)
daily active users	1.86	1.80	1.83
	(1.34)	(1.20)	(1.28)
total users in the group	14.02	14.13	13.96
	(9.05)	(9.20)	(8.80)
parent msgs	2.63	2.52	2.54
	(3.56)	(3.12)	(3.25)
daily teacher msgs	1.11	0.71	1.23
	(9.52)	(2.93)	(11.93)
category ECE	0.47	0.48	0.48
	(0.50)	(0.50)	(0.50)
degree centrality	0.46	0.46	0.47
	(0.27)	(0.27)	(0.27)
closeness centrality	1.58	1.60	1.62
	(2.87)	(3.17)	(3.41)
betweenness centrality	0.09	0.10	0.09
	(0.09)	(0.09)	(0.09)

Note: See Table O.1 for a brief description of each feature. Notice that there is no significant statistical difference between the features in different arms confirming that the samples are balanced across four arms.

O.6.4. Additional Details on the Value of Targeted Nudging

Conservative heuristic summary of statistical separation between the targeting and uniform policies: In this section, we estimate the degree of statistical separation between the targeting and uniform policies using a conservative heuristic, relying only on data from the first RCT. This heuristic is intended to capture the implied chance that targeting outperforms the uniform baseline policy, based on confidence intervals for the two policies. The analysis reveals that, everything else equal, the targeted policy had an approximately 56% chance of outperforming the uniform baseline policy (described in more detail below). This modest, yet slightly higher evidence in favor of targeting motivated us to run the second experiment and test it in the field. We note that, while this quantity should not be read as a literal calibrated probability, it provides a conservative summary of how plausible targeting outperformance is based on the off-policy analysis.

Let $\bar{Y}_{\text{target}}$ and $\bar{Y}_{\text{peer}}$ denote the population mean outcomes under the targeted and uniform–peer policies, respectively. Fix $\alpha \in (0,1)$ and suppose we have valid $(1-\alpha)$-level confidence intervals $\text{CI}_{\text{target},1-\alpha}$ and $\text{CI}_{\text{peer},1-\alpha}$ that are disjoint and ordered with the targeted interval strictly above the peer interval. Let $E_{\text{t}} := \{\bar{Y}_{\text{target}} \in \text{CI}_{\text{target},1-\alpha}\}$, $E_{\text{p}} := \{\bar{Y}_{\text{peer}} \in \text{CI}_{\text{peer},1-\alpha}\}$. Since $\mathbb{P}(E_{\text{t}}) \geq 1-\alpha$ and $\mathbb{P}(E_{\text{p}}) \geq 1-\alpha$, the union bound gives

$$\mathbb{P}(E_{\text{t}} \cap E_{\text{p}}) = 1 - \mathbb{P}(E_{\text{t}}^{\text{c}} \cup E_{\text{p}}^{\text{c}}) \;\geq\; 1 - \left[\mathbb{P}(E_{\text{t}}^{\text{c}}) + \mathbb{P}(E_{\text{p}}^{\text{c}})\right] \;\geq\; 1 - 2\alpha.$$

Whenever $E_{\text{t}} \cap E_{\text{p}}$ occurs, the non–overlap implies $\bar{Y}_{\text{target}} > \bar{Y}_{\text{peer}}$; hence $\Pr\big(\bar{Y}_{\text{target}} > \bar{Y}_{\text{peer}}\big) \;\geq\; 1-2\alpha$. To operationalize the condition, let $LB_{\text{target}}(1-\alpha)$ and $UB_{\text{peer}}(1-\alpha)$ denote the lower and upper endpoints of the respective $(1-\alpha)$-level intervals, and define

$$\alpha^{\star} \;=\; \inf\left\{\alpha \in (0,1) :\; LB_{\text{target}}(1-\alpha) > UB_{\text{peer}}(1-\alpha)\right\}.$$

This is the smallest level at which the intervals are strictly separated in favor of targeting, yielding the bound $\Pr\big(\bar{Y}_{\text{target}} > \bar{Y}_{\text{peer}}\big) \;\geq\; 1-2\alpha^{\star}$. Empirically, we fit a random-forest model 20 times on distinct train/test splits. For each run, we predict engagement under self and under peer, compute uplift (self - peer), construct a personalized policy that assigns the self nudge only to parents with positive predicted uplift (others remain on peer), and evaluate both the personalized and the uniform peer policies on the held-out fold via inverse-probability weighting (IPW). Using these 20 runs' average performance evaluations, we construct confidence intervals for the average outcome under each policy and, from these, derive a conservative lower bound on the probability that the targeted policy outperforms the uniform benchmark. We find that $\alpha^{\star} \approx 0.2217$ which yields $1-2\alpha^{\star} \approx 0.5566$ (about 56%). We interpret this quantity as a conservative heuristic summary of the off-policy evidence in favor of targeting. In particular, we can conclude that the smallest confidence level at which the intervals separate implies only modest statistical separation between the two policies, with slight evidence in favor of targeting outperformance.

To contextualize this conservative guarantee and make the heterogeneity gains explicit, we complement the scalar bound with a Qini-style curve (Radcliffe 2007) adapted to personalization budgets. Ranking parents by predicted uplift, we define for each budget $x \in [0,1]$ a policy that assigns the top $x\%$ their predicted optimal nudge (self if uplift > 0, otherwise peer) and evaluate its expected outcome on held-out data via IPW. Figure O.5 plots the estimated outcome as a function of x, with pointwise confidence bands at level $1-\alpha^\star = 77.83\%$ aggregated over the 20 replications. The curve peaks around $x \approx 0.40-0.45$ at roughly $3.98-4.00$, compared with the uniform–peer baseline near 3.67; this corresponds to an absolute gain of about 0.30 (approximately $8-9\%$ of the uniform peer mean engagement). For small budgets ($x \leq 0.10$), the improvement is modest (about $0.05-0.10$); beyond $x \approx 0.45$ the curve flattens, indicating that only $\sim 45\%$ of parents have positive predicted uplift, so additional personalization yields no further gains. The "all self" baseline lies near 3.47, about 0.20 below the uniform–peer level, underscoring that indiscriminate self messages are inferior and that benefits arise from selectively assigning self to those predicted to respond. The targeted band lies above the peer band for moderate budgets and at $x = 1$, consistent with the non-overlap that underpins the bound $1-2\alpha^\star$. Overall, the shape and level of the curve indicate heterogeneity that is meaningful but not large, an interpretation aligned with the conservative heuristic analysis.

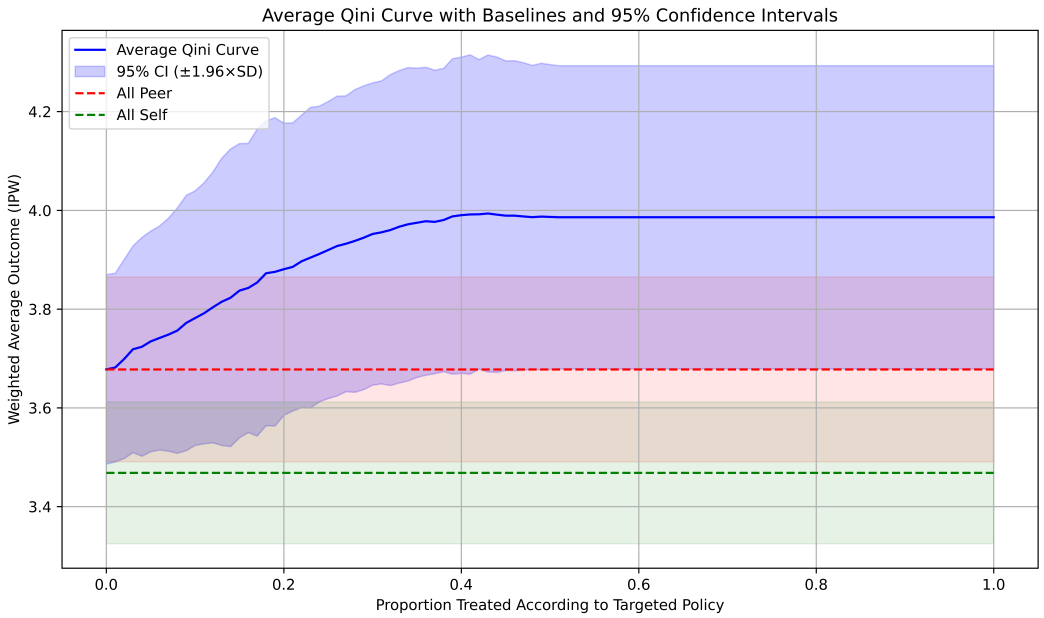


Figure O.5 Qini curve comparing targeted and uniform nudging strategies. The curve plots the predicted average outcome (estimated via inverse probability weighting) as a function of the fraction x of the population receiving their predicted optimal treatment. The shaded bands show 77.83% confidence intervals across 20 replications since the smallest $\alpha \in (0,1)$ at which the targeting policy's lower bound of $(1-\alpha)$ confidence interval exceeds the upper bound of $(1-\alpha)$ confidence interval of the best uniform policy (peer comparison) is approximately $\alpha = 0.2217$.

Additional discussion on personalization margin and covariate-shift: Next, we discuss reasons that could help explain the relatively modest average value of targeting over the uniform arm that we observed in our off-policy and on-policy evaluation.

First, the personalization margin is intrinsically narrow, since we have only two treatment arms, peer and self comparison and by the fact that our uniform benchmark assigns the peer comparison nudge rather than a pure control. Personalization therefore operates on a single margin: reallocating a subset of parents from peer to self. Once all parents with positive predicted uplift for self have been reassigned, personalization saturates and no further heterogeneity can be leveraged. In contrast, a richer library of nudges could create additional margins along which targeted assignment might realize larger gains. This structural constraint mechanically caps the attainable advantage of targeting over the uniform peer policy.

Second, a substantial subset of parents lie near the model's decision boundary, i.e. their predicted outcomes under peer and self nudges are similar, so modest estimation errors can flip their optimal assignment. As shown in Figure O.6, roughly 50% of parents have an average predicted difference in total engagement days of below 0.5 days (out of 28), when we compare the peer and self comparison nudges. This proximity inflates variance in our targeting off-policy evaluation, thus widening the confidence bands and, by construction, lowering our conservative probability bound. Importantly, the same prevalence of near-ties helps explain why, in the second on-policy RCT, the targeted and uniform policies delivered statistically indistinguishable average outcomes: small assignment flips across many borderline parents are unlikely to cumulate into large average differences.

Third, a plausible explanation is *covariate shift* between the off-policy evaluation and policy-learning data and the RCT population. To mitigate distributional differences in covariates, we re-trained our random-forest targeting model using importance weighting as in Sugiyama et al. (2007), as detailed in our manuscript. While this approach formally corrects for shifts in the feature distributions, it also introduces additional estimation variability: the reweighted training objective can substantially reduce the effective sample size (Polo and Vicente 2020), especially when a subset of observations receives large weights, leading to noisier parameter estimates and wider confidence intervals in off-policy evaluation.

In order to quantify the covariate shift impact, we computed the effective sample size (ESS)(Polo and Vicente 2020) associated with the importance weights used in our reweighted random-forest training approach (Sugiyama et al. 2007). Let $\{w_i(\mathbf{x}_i)\}_{i=1}^n$ denote the covariate-shift weight assigned to observation i. The ESS is defined as

$$ESS := \frac{(\sum_{i=1}^n w_i(\mathbf{x}_i))^2}{n \sum_{i=1}^n w_i(\mathbf{x}_i)^2}$$

which intuitively captures the percentage of effective samples after adjusting for the covariate shift. Using the weights employed in our model training, we obtain $ESS \approx 0.3$, indicating that covariate shift exerts a considerable impact on the precision of our estimates, highlighting the practical challenge of learning policies under distributional mismatch.

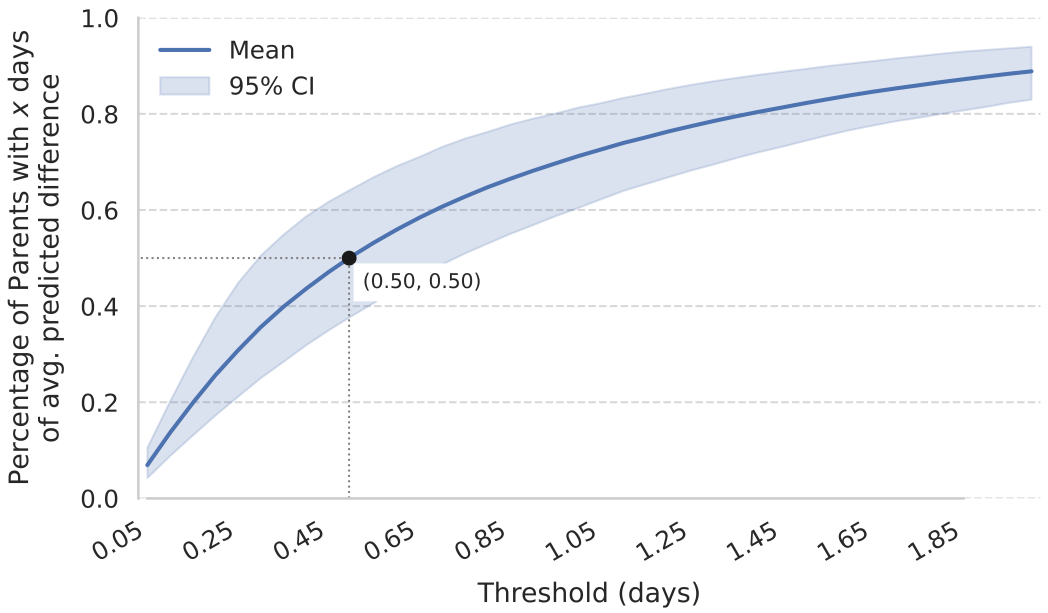


Figure O.6 For each candidate threshold x (x-axis), we compute the fraction of parents i in the held-out test set whose predicted difference in active days between self and peer comparison nudges is less than x days (i.e. $|\widehat{Y}_{i,\text{self}} - \widehat{Y}_{i,\text{peer}}| < x$, where $\widehat{Y}_{i,\text{peer}}, \widehat{Y}_{i,\text{self}}$ are the predicted outcomes under peer and self comparison nudges for parent i, respectively). The solid line plots the mean coverage rate on a held-out test set over 20 independent runs, and the shaded ribbon shows the corresponding 95% confidence interval across runs. As x increases, a larger share of parents are effectively "indifferent" between the two nudges, rising from approximately 7% at $x = 0.05$ days to over 90% at $x = 2$ days. The horizontal dashed line marks the 50% coverage level, and the vertical dashed line identifies the smallest threshold (≈ 0.5 days) at which at least half of the parents lie within the indifference set.

O.6.5. Further discussion on observed differences on the performance of targeted nudging

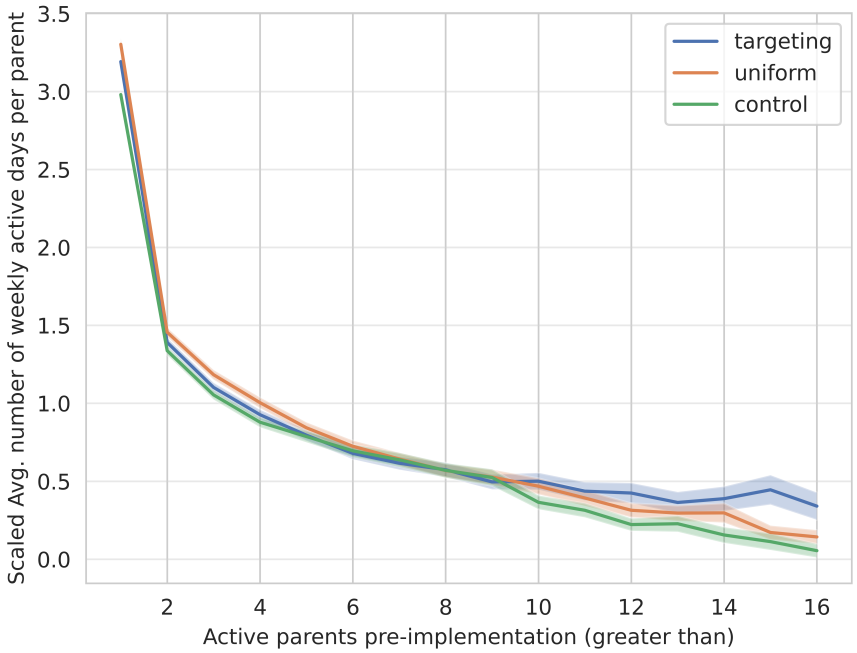
In §5.2 we discussed how targeted nudging is more effective on parents who belong to large groups with many active parents. Given the substantial gains generated from targeting in this sub-population, we next analyze what drives these differences in the effect of targeted nudging for larger groups. Recall that targeted nudging differs from uniform nudging by selecting a subset of users to nudge with self-comparison instead of peer comparison. Furthermore, peer comparison nudges rank users within the same group based on their activity in the group. Naturally, peer comparison in larger groups with many active users would lead to more users potentially receiving lower absolute ranks in comparison to smaller groups even with the same amount of activity.[4] Hence, peer comparison becomes less effective when users are presented with lower absolute ranks for themselves.

[4] For example, users in a group of 20 members are ranked from 1 to 20 while users in a group of 5 members are ranked from 1 to 5. Hence, 15 members in the larger group receive absolute ranks lower than 5 while no one in the smaller group receives an absolute rank lower than 5.

To further analyze this effect, we leverage the rank data from the second experiment. In particular, for any focal parent j, we calculate the lowest absolute rank (denoted by $\texttt{lar}$) that this parent could have received in any treatment week based on their activity and the activity of other parents in the same group.[5] As before, we define $\mathcal{G}_z^{\texttt{lar}}$ to be the parental sub-population with an absolute minimum rank of at least z. In Figure O.7, we plot the conditional mean of the outcomes for different sub-populations $\bar{Y}_{\mathcal{G}_z^{\texttt{lar}}}(k),\ k \in \{\texttt{t}, \texttt{u}, \texttt{c}\}$, $\forall z \in \{1, \ldots, 16\}$. For instance, $\bar{Y}_{\mathcal{G}_5^{\texttt{lar}}}(\texttt{t})$ denotes the average number of active days per week for all parents who received an absolute minimum rank of at least 5 in the targeting arm. As hypothesized, we find that targeting is more effective for parents with higher absolute minimum ranks (who would be part of larger groups with more active parents). In fact, targeted nudging consistently outperforms both uniform nudging as well as control across all users with an absolute minimum rank of 10 and above. This provides further evidence of the differential effect that peer comparison nudges (and in turn targeted nudging) can have based on group-level features such as the number of active users pre-implementation.

[5] For instance, in a group of four parents where two parents interacted for three days and the other two for one day in a given week, the first set of parents (with three days of interaction) would be tied for first place, giving them a best possible rank of 1. The second set of parents (with one day of interaction) would be tied for third place, resulting in a best possible rank of 3. We repeat this calculation for each week to then determine the minimum rank for each parent across all treatment weeks, thus identifying the lowest rank a parent could achieve during the entire treatment period.

Figure O.7 Comparison of average active days among targeted vs. non-targeted parents with respect to activity ranks during treatment



Note. For each arm, we calculate the average number of active days in a week during treatment conditional on guardians' lowest absolute rank being above a threshold. More precisely, we report $\bar{Y}_{\mathcal{G}_z^{\texttt{lar}}}(k),\ k \in \{\texttt{t}, \texttt{u}, \texttt{c}\}$, for a range of values $z \in \{1, \cdots, 16\}$, and their 95% confidence intervals. Notice that when we only include guardians with large enough lowest absolute ranks, there is an inversion in the best-performing arm in both rank metrics, and the gap between the targeting and uniform/control arms widens as we increase this threshold for the minimum rank metric. We highlight that $\mathcal{G}_{16}^{\texttt{lar}}$ includes all parents who fall above the 99.5th percentile. Additionally, to maintain confidentiality, we divided the reported values by a normalizing constant obtained by sampling from an uniform distribution on $(0, 1)$.