Paid to Learn: Piloting a New Type of Social Protection for Developing Countries

Erlend Berg, D Rajasekhar and R Manjula*

November 2023

Abstract

Public-works programmes are widespread in developing countries, in part because this form of social protection is potentially self-targeting. However, public works can also be administratively demanding, costly and corrupt. We pilot a proposed welfare scheme designed to retain the self-targeting aspect of public works while improving accessibility and reducing operational complexity and fraud: beneficiaries are paid to solve educational micro-tasks on a smartphone. Hence, while participants in traditional public-works schemes typically build or maintain physical infrastructure, participants in the piloted scheme would build their own human capital. In a randomised-controlled trial in Karnataka, South India, we recruited 114 illiterate workers with experience of India's flagship public-works scheme. The treatment group was given access to the pilot scheme for a week. The intervention increased the number of characters recognised from the local script by 65%, and more than half the test users reported preferring the piloted scheme over the existing public-works programme. We also report on an earlier pilot that had no effect on character recognition, probably due to an overly ambitious 'syllabus'.

JEL: I38, O15, O17

Key words: Social protection; self-targeting; public works; NREGA; literacy; India

1 Introduction

Many governments use means tests to target welfare benefits to those who need them. Typically, means testing in a rich country involves using tax data to establish the income, wealth or employment status of potential beneficiaries, which is then compared to a welfare scheme's eligibility criteria. However, more than 60% of the world's working population are in the informal economy (ILO, 2018), and most of these are in developing

*Berg (corresponding author): University of Bristol, School of Economics, Priory Road Complex, Bristol BS8 1TU, United Kingdom. erlend.berg@bristol.ac.uk. Rajasekhar: Institute for Social and Economic Change. Manjula: Institute for Social and Economic Change. We are grateful to Dylan Holden-Sim for excellent software development and support. Ethical approval was obtained from the University of Bristol. The project was funded by the University of Bristol and the Institute for Social and Economic Change. Declarations of interest: none. American Economic Association RCT registry identifier: AEARCTR-0008155.

countries. As these workers do not appear in administrative records, governments in developing countries need alternative targeting strategies.[1]

Public-works schemes are available in many developing countries. Rather than paying unemployment allowance, in these schemes the government acts as 'employer of last resort'. If the combination of the type of work offered (typically unskilled manual labour, such as building or repairing roads) and remuneration is relatively unattractive compared to market wages, so that only those in need apply, then this form of social protection can effectively be self-targeting. However, public-works programmes are administratively demanding and prone to inefficiency and corruption. Moreover, to provide effective social protection such schemes should provide employment and income as and when needed, but in practice there are often delays and inefficiencies in both the provision of work and the payment of wages.

This paper presents pilot results for a new form of self-targeting welfare scheme in which beneficiaries are paid to engage with a smartphone app. The app can be made available on demand, avoiding delays in access. Work is instantly verified and payments can be made directly from the centre to the end user, cutting out 'middle men' and thus reducing the scope for corruption. Administrative costs are transparent and likely to be far lower.

Various types of work could be offered in the app. In the pilot described here, the tasks were educational and aimed at improving character recognition, a key pre-literacy skill. Therefore, while participants in traditional public works typically build or maintain physical capital, beneficiaries of the piloted scheme build their own human capital.[2]

We present the results of a randomised-controlled trial in South India in which 114 illiterate adults with experience of the government's flagship public-works programme, the Mahatma Gandhi National Rural Employment Guarantee (NREG), were recruited. Participants in the treated group were each provided with a smartphone on which the educational app was pre-installed. They had access to the app for a week, and could work on it as little or as much as they wanted, within set working hours. They accrued earnings per second of engagement with the app, and their identity was verified with fingerprints to prevent them from 'farming out' the work to others. After a week's exposure, the participants recognised 1.7 more characters of the local Kannada script, an increase of 65% over the control group at baseline. At least some of the effect appears to come from the extensive margin, that is, participants who went from recognising zero characters at baseline to recognising at least one character at follow-up. Analysis of data from the app

[1] *Proxy* means testing, whereby a household's income or wealth is estimated using observables such as TV ownership and roof construction material, is widely used but prone to measurement error. Fiszbein and Schady (2009) document the use of proxy means tests in conditional cash transfer programmes.

[2] Note that introducing the type of scheme proposed here would not necessarily lead to a decline in infrastructure investment: any funds saved from reduced administration costs and fraud could be used to improve public infrastructure more efficiently and transparently, using commercial providers and skilled labour rather than the unskilled labour of welfare beneficiaries.

shows that the number of attempts required to recognise characters correctly declined steadily over time.

After the intervention, the treated participants were asked to compare the piloted scheme with the existing public-works programme, NREG, along a number of dimensions. A majority of users find the piloted scheme less physically and mentally tiring, more dignified, more suitable during a pandemic and less likely to be associated with fraud and corruption. After just a week of exposure, 53% of treated participants would choose to work with the app given a choice between that and the existing scheme. However, traditional public works are deemed by a majority to be more suitable for healthy, working-age men and women, and to be better for the overall development of the village. (The latter is probably due to the fact that the traditional scheme typically improves local infrastructure, while the benefits of the piloted scheme are primarily private.)

The paper also reports the results of an earlier pilot in which no significant learning results were found (Appendix A). In the earlier pilot, the app exposed test users to all the 52 basic characters of the Kannada script, plus 15 vowel diacritics. Given the low starting level of the illiterate test users (the average number of characters recognised at baseline was about 5) and that they only had access to the app for a week, attempting to teach all 52+15 characters was probably too ambitious. A decision was made to re-pilot with a focus on 10 commonly used characters. The second pilot is the focus of the paper, but we report results from the earlier pilot for transparency.

We have previously field-tested an SMS-based prototype of a self-targeting welfare scheme using basic mobile phones (Berg et al., 2017). In that experiment, the micro-tasks were pure 'ordeals', that is, they were non-productive rather than educational, and users tested the prototype for two–three hours only and in a supervised setting. To our knowledge, this paper is the first to discuss and present experimental results on using educational smartphone-based micro-tasks as the basis for a self-targeting social protection scheme.

The pilots were designed to test the scheme's educational potential and user acceptability in the short run. While promising, a trial of longer duration and with more advanced educational content is required to test effects on measures of targeting, accessibility, administration costs, fraud and labour market choices.

The idea of self-targeting social protection has been explored previously. Nichols and Zeckhauser (1982) discusses imposition of deadweight losses ('ordeals'), including menial work, as a targeting mechanism for social transfers. Ravallion (1991) studies the merits of several South Asian public-works programmes in providing self-targeting welfare. Besley and Coate (1992) explore theoretically how a work requirement on beneficiaries can act as a screening mechanism when income is unobservable. Coady et al. (2004) review and assess self-targeting in social protection schemes across a range of developing countries. Kleven and Kopczuk (2011) set out a model in which the complexity of the application

process can be varied; greater complexity improves precision in targeting but imposes a cost on applicants. Alatas et al. (2016) discuss self-targeting social protection schemes in the context of an Indonesian conditional cash transfer programme, finding that requiring beneficiaries to apply, as opposed to being automatically enrolled, improves targeting as those who are unlikely to pass the asset test do not bother to apply.

India's Mahatma Gandhi National Rural Employment Guarantee (NREG) is the world's largest public-works scheme.[3] However, it is marred by rationing (not being available on demand, as it is supposed to), complex and costly administration, widespread fraud, and delays and non-payment of wages. Sukhtankar (2017) provides an overview of the scheme and surveys related research. Using nationally representative survey data, Dutta et al. (2012) find that, across India, only 56% of those who wanted to work under the scheme were able to do so. Ravallion (2019) discusses the roles of local implementation (non-wage) costs and corruption in explaining rationing in the scheme. The challenges of scheme administration are highlighted in an ethnographic study of NREG in Bihar: 'the state government lacks the capacity to run projects as documented... Limitations on state capacity are complex, ranging from inadequate staffing, training and salaries to an inability of officials to navigate panchayat politics and the entrenched opposition of landowners.'[4] Imbert and Papp (2011, 2014) find evidence of large-scale leakage in NREG budgets by comparing representative household survey data on employment with official scheme reports.[5] Banerjee et al. (2020) find that 47% of control-group GP leaders in their sample in Bihar agree that corruption in the administration is a major constraint in NREG implementation. Delays and non-payment of wages are also widely reported (e.g., Aggarwal, 2017).

Rodriguez-Segura (2022) reviews evidence on the use of educational technology ('EdTech') in developing countries. Ksoll et al. (2014) evaluates an SMS-based adult literacy programme in California and finds it highly effective in improving reading skills.

2 The app

Under the proposed social protection scheme, beneficiaries earn money by engaging with a smartphone app. Figure 1 shows the main screens of the piloted app, which aims to teach illiterate beneficiaries to recognise the characters of the Kannada script. On the main screen, the task screen, a sound is played, corresponding to the pronunciation

[3] According to official records (https://nrega.nic.in, accessed on 29 September 2022), it generated 3.6 billion days of paid work for 106 million beneficiaries in the 2021–22 financial year.

[4] Witsoe (2014) as quoted by Sukhtankar (2017).

[5] They find that the proportion of reported scheme labour-days provided that could be verified through household surveys were 51% in 2007–08, 71% in 2009-10 and 80% in 2011-12. Note that their leakage measures are conservative since they compare officially reported NREG labour days to survey-based measures of days in *any* public works, not just NREG.

of a randomly selected Kannada character. This character, along with three randomly selected incorrect ('distraction') characters, are shown as 'buttons' on the screen, in random order. The user's task is to tap the character that corresponds to the sound played. If an incorrect character is selected, the user's choice is momentarily highlighted in red but the screen does not otherwise change, thus permitting further attempts. When the correct character is pressed, this button is momentarily highlighted in green before the app moves to the next screen. The task screen also has a 'sound' button which allows the user to replay the sound of the task character.

Once the task is answered correctly, the app moves to the attention check screen. A green dot is displayed in a random location on the screen. The user just needs to touch the dot in order to proceed. The purpose of this is to make sure that the user is actually looking at the screen, as opposed to interacting with the app without looking. Once the attention check is passed, the app displays a fingerprint logo, prompting the user to touch the phone's fingerprint reader. The aim of authenticating the registered user is to prevent the work being 'farmed out' to someone else. Once passed, the cycle is complete and the app displays the next task.

Earnings accrue at a fixed rate per second whenever the beneficiary is engaging with the app. The user is deemed to be engaged when the app registers a 'valid action': Either pressing a character button, or pressing the green dot on the attention screen dot, or providing the correct fingerprint at the fingerprint screen. Whenever a valid action is registered, earnings accrue for five seconds, or until the next valid action is registered. In this way, earnings accrue continuously as long as a valid action is registered at least every five seconds. Users can pause at any time, for a shorter or longer period, and their earnings are preserved but will not increase.

In order for the proposed social protection scheme to be equitable and not favour those with pre-existing knowledge or skill, the app is designed to reward engagement rather than task performance. For this reason, users are rewarded equally for tapping correct and incorrect answers to the tasks. However, in order to prevent users from earning money simply by pressing the same incorrect character button again and again, without cycling through the screens, earnings accrue only once per character button per task.

The balance earned is visible at the top of each of the main screens. The balance is shown in green and increases every second whenever earnings are accruing (for five seconds after any valid user action), and in grey when engagement is paused. A user's earnings never decreases.

To prevent overwork, and for comparability with NREG, tasks were only available to beneficiaries between 9am and 5pm, Monday through Friday. Outside of these hours, a 'sleep' screen was displayed, informing the user of when tasks would next be available. Within the eight-hour 'working day', users were free to engage as little or as much as they

wanted, and to take breaks whenever they wanted, and for however long they wanted. There is thus no requirement to put in a full working day, or to work continuously, in order to benefit.

There was a maximum daily earnings limit of 289 rupees, which was selected to match the daily wage rate under NREG in Karnataka at the time. The per-second rate in the app was set so that the maximum daily earnings would be reached after six hours of engagement. That is, in order to attain the maximum daily earnings, users would need to engage with the app for six hours a day within the eight-hour window, but could take their two hours of breaks whenever they wanted. If the maximum daily earnings limit was reached (after six hours of engagement), or in any case at 5pm, the sleep screen would be displayed until 9am the next working day.

The app selects tasks and answer options from a subset of ten characters from the Kannada script, as an earlier pilot suggests that teaching all 52 basic characters over one week is overly ambitious. The ten taught characters, and the remaining 42, are shown in Figure 2. The ten characters comprise six consonants and four vowels and were selected on the basis of being commonly used and easy to combine into simple words.

Clearly, the piloted app is very basic. The educational aspect would need to be more sophisticated in order for the system to be viable as a social protection scheme at scale. It would need to have more content, and more variation in content. It should probably adapt to the user's pre-existing skill level and progress to more advanced topics as the user learns.

3 Experimental design and data

The fieldwork took place between November 2021 and March 2022. Participants were recruited from four villages in the Ramanagara district of Karnataka, India.[6] In each village, enumerators went door to door to enlist participants. The criteria were age (18+), being a native speaker of Kannada, being self-declared illiterate and having completed no more than two years of schooling, and having participated in NREG public works in the past three years. Those who met these criteria and gave their consent to take part in the study were enlisted.

In all, 114 participants were recruited. After completing the baseline survey and literacy test, they were randomly allocated to treatment and control arms. Randomisation was individual and stratified by village, sex and age. Table 1 presents participant summary statistics by experimental arm. The average age of participants was 57, and 81% were female. On average they had 0.24 years of schooling (86% had completed

[6]Three of the villages were in Magadi taluk and one was in Channapatna taluk. The villages were selected on the basis of a high proportion of native Kannada speakers in the population and a high participation rate in NREG public works, according to NREG statistics for the period 2016–2021.

no schooling), and 74% were married as opposed to unmarried, divorced or widowed. Most participants were lower-caste ('Other Backward Castes'), while the proportion of scheduled castes and tribes (SC/ST) was 12%. The mean number of Kannada characters identified correctly at baseline (out of the ten focus characters) was 2.57. The normalised difference between the arms, shown in the last column, does not exceed the conventional threshold of 0.25 for any of the variables.

Basic literacy tests were conducted before and after the intervention in both experimental arms. Participants were shown a piece of paper with all the main Kannada characters, and asked if they could identify any of them. There were ten versions of the sheet, differing only in the random ordering of the characters, and participants were randomly assigned to different test sheets at baseline and follow-up. The ten characters taught in the app were presented separately, as in Figure 2, though still in random order. Enumerators made a note of characters identified correctly by the participants, as well as characters attempted but incorrectly identified.

The tests also included nine words, varying in length between two and four characters and composed exclusively of the ten taught characters. All participants were presented with the same nine words, in random order, before and after the intervention.

Treated participants were given access to the app for one week. To avoid selection on phone ownership, all treated participants were lent dedicated smartphones. These were all of the same make and model (Nokia 5.3) and locked to the literacy app so that they could not be used for any other purpose.[7] The app was only available on these phones, that is, it could not be installed on any other device. The participants' fingerprints were registered on the devices as they were handed out.[8] Participants were told that they could use the app as much or as little as they wanted in the week they had access to it, within the constraints on working hours described above. All treated participants returned the phone at the end of the week, when they were also fielded an endline survey and character recognition test, and paid.

All participants received 100 rupees at the end of the week for taking part in the study. Treated participants received their accrued app earnings in addition.

Four treatment and three control participants dropped out between baseline and follow-up surveys, so the final panel consists of 53 treated and 54 control participants. The baseline characteristics of those observed at follow-up are also balanced across the experimental arms (not shown).

[7]This was done to avoid confounding the results with any more general effects of having access to a smart phone.

[8]Fingerprint registration and authentication were local to the device, so biometric information was not transmitted over the internet or mobile network, and it was deleted after the intervention.

4 Results

4.1 Learning outcomes

Figure 3 presents the main results visually. The bars represent the mean number of characters recognised, out of the ten characters taught in the app, by experimental arm before and after the intervention. Treated participants correctly identify about two more characters after the intervention, while the increase in the control group is small. The figure also confirms that test scores were balanced across the experimental groups at baseline.

These results are confirmed by regression analysis in Table 2. The basic difference-in-difference specification, without fixed effects, is shown in column 1. The number of taught characters identified correctly by treated individuals increases by 1.767 more than the corresponding increase in the control group, and the difference is statistically significant. Standard errors are clustered at the participant level (the level of randomisation) here and in all panel regressions presented.

Our preferred specification, which includes participant fixed effects, is presented in column 2. The main coefficient suggests that the number of characters recognised by treated individuals increased by 1.684 more than the control group after one week of exposure to the programme, and the coefficient is statistically significant. This represents an increase of 65% relative to the number of characters recognised in the control-group mean at baseline.

Recall that the app exposed users to only 10 out of 52 basic characters in the Kannada script. In column 3, the outcome variable is the number of 'non-taught' characters recognised. This can be thought of as a placebo check: if the main effect on literacy in column 2 stemmed from the treated group generally improving their pre-literacy skills, rather than learning within the app, then one might expect them to learn characters beyond the 10 taught in the app. However, there is no significant learning effect for the characters not taught in the app.

In column 4, the outcome variable is the number of words read correctly, out of the nine presented. Recall that each of these words are composed of two, three or four of the taught characters. The coefficient of interest is small and not statistically significant, suggesting that the improvement in character recognition did not imply an improved ability to read words based on the taught characters. Reading words may be a separate skill ('blending') which needs to be learnt separately from character recognition.[9]

One may wonder whether the average improvement in mean character recognition score was driven by those who already knew a few characters, or none. The distributional effect of the intervention on the recognition of taught characters is explored in Figure 4.

[9]Kannada is a phonetic language, meaning that each character has a fixed pronunciation wherever it occurs.

Visually comparing the distributions of test scores, the clearest effects of the intervention seem to be a reduction in the number of participants recognising zero characters, along with a more general rightward shift of the distribution. Thus it seems that the intervention worked at least in part on the extensive margin: it taught some participants who started off recognising zero characters to recognise at least some characters.

The regressions presented in Table 3 analyses the number of *attempted* characters and words in the literacy tests, irrespective of whether the character or word was identified correctly or not. This could be informative of whether the intervention gave participants more confidence to attempt more characters or words, whether they truly knew them or not. The results suggest that treated participants attempted more of the taught characters, but increases in attempted other characters, and words, are not statistically significant. It seems, therefore, that the increased confidence of treated participants was restricted to the characters that they had actually learnt.

'Hit rates', understood as the proportions of characters or words identified correctly, out of those attempted, are analysed in Table 4. None of interaction coefficients are statistically significant. This again suggests that the intervention did not cause 'false confidence' among the participants.

Table 5 looks at whether learning in the app is related to basic demographics. The outcome variable, 'Characters learnt', is constructed as the individual increase in the number of characters recognised from baseline to endline. Learning is not predicated by sex, age, marital status or caste. But participants with some schooling learn about 1.9 fewer characters on average. Recall that most participants have no schooling, but 14% have one or two years. One possible interpretation is that those who never had a chance to go to school are able and/or motivated to learn, while those who did enrol but dropped out early are selected on being less able and/or motivated.

4.2 App usage

There was considerable variation in how much time users spent on the app over the week they had access to it, ranging from just over a minute to almost 21 hours, with a mean of 5.15 and a median of 3.41 hours for the week. Since earnings are directly proportional to time spent in the app, and maximum earnings would require $5 \times 6 = 30$ hours of engagement over the week, even the heaviest user earned only 70% of the maximum possible, and the median user earned only 11% of the maximum possible over the week.[10]

Figure 5 is a graphical representation of each test user's interaction with the app over time. Horizontal lines represent individual users, and dots show the day and time of each completed task. The vertical lines demarcate the beginning (9am) and end (5pm) of each

[10]When asked why they did not spend more time on the app, the most common response from users was that they were 'busy elsewhere'. One interpretation of this is that the scheme functioned as a fallback option, as intended.

'working day'. A small proportion (0.06%) of tasks are recorded as completed outside of these hours due occasional to delays in the system or network.

Figure 6 confirms that there is a strong positive relationship between the time spent engaging with the app and the total number of tasks a user solved.

Figure 7 shows the average number of attempts required by a user to solve a task correctly, as a function of the number of tasks solved previously. The downward trend is indicative of learning, as users require fewer attempts to solve each task as they gain experience.[11] A similar downward trend is apparent when the vertical axis represents the average time in seconds required to solve a task, rather than the number of attempts per task (not shown).

In the endline survey, the average treated user reports spending 9.68 hours on the app over the week.[12] It is of interest to look at where this time is taken from. Table 6 analyses data from a time use module in which all participants were asked, before and after the intervention, how much time they had spent on a range of activities over the past seven days. The specification is otherwise the same as those in Tables 2, 3 and 4. There is no significant effect of the intervention on the time spent on any of the measured categories, including traditional public works, private paid work, work on own land or business, domestic work or education. There is also no significant impact on the sum of time spent on all these activities (not shown). In itself this might lend support to the idea that the time spent on the app was taken out of leisure time, which was not explicitly captured in the survey. However, several of the time-use coefficients are imprecisely measured. In particular, the not significant coefficient of -5.44 hours on paid work should be seen in conjunction with the negative and significant effect on private work found in the earlier pilot (Table A3).

There is also no significant impact on the time spent at school or on homework among children in the household, or on participants' self-reported physical or mental health (not shown).

4.3 User satisfaction

As part of the follow-up survey, test users—the treated partipants—were asked to compare 'app work' to NREG work along a number of dimensions. (Recall that all partici-

[11] The minimum achievable is 1, when every task is solved correctly at the first attempt. The curve was smoothed by binning: the number of attempts required per task was averaged over the first 100 tasks, the next 100 tasks, and so on. Only the ten users who solved at least 5000 tasks are included here. Including all users would give a graph with an even steeper drop in the number of attempts, but would be distorted as not all points on the curve would be based on data from the same users (a form of survivorship bias).

[12] This is almost double the mean time engaged as measured in the app. Clearly it is possible to spend time with the app without being recorded as engaged with it, since continuous recorded engagement required the user to perform a valid action every five seconds. The difference may also arise from recall error. Perhaps there was also desire to please the enumerators by exaggerating time spent on the app.

pants were selected on the basis of having had experience of NREG public works in the past three years.) For each of these questions, the answer options were 'NREG work', 'App work', 'No difference' and 'Cannot say'. Table 7 shows the proportion of users answering 'App work' to each question, out of those who stated a preference, that is, out of those who answered either 'App work' or 'NREG work'.

Along most dimensions, the average test user compared the piloted scheme favourably to public works. The average user found the piloted scheme less physically tiring (98%), less mentally tiring (65%), more dignified (74%), better for building self-confidence (52%), more suitable during a pandemic (98%), more likely to teach them something new (98%), more suitable for pregnant women (100%), women with young children (98%), the elderly (93%), those in poor health (100%), the disabled (100%) and the illiterate (92%), more likely to pay out the correct amount of benefits (52%),[13] more likely to pay benefits on time (92%) and less likely to be associated with fraud and corruption (94%).

There were also a few dimensions along which users expressed a preference for traditional public works over the piloted scheme. Public works were thought to provide better experience for taking up other employment (69%), probably because the type of work that the participants would normally be offered — manual labour — would be more similar to that offered under the existing public works programme. Most users also considered public works to be more suitable for healthy, working-age men (100%) and women (78%), possibly in contrast to the other demographic categories considered. This suggests that the piloted scheme may be better thought of as a possible complement to, rather than a substitute for, traditional public works.

Ninety-eight per cent of test users thought public works would be better for the overall development of the village. This is most likely because of the focus of NREG on the construction and repair of communal infrastructure, whilst the piloted scheme is intended to build the human capital of individual beneficiaries. (The enumerators did not mention to respondents the possibility of spending any savings associated with the piloted scheme on infrastructure.) Ninety per cent of test users thought public works would be more likely to be available when needed. Anecdotally, for some users this reflected the fact that the proposed scheme, as piloted, was available to treated users for one specific week only. If rolled out at scale, however, the proposed scheme may well be deemed more easily accessible than NREG, as the latter is demand-led in principle but often not in reality.

[13]This proportion might have been expected to be even higher, given that all test users were paid the correct amount without delay according to our records, and no complaints were registered on the helpline. Several users linked their answer to actually having to work in the app, while in NREG they are sometimes paid without working. In effect NREG beneficiaries sometimes get paid part of the proceeds from fraudulently reported work, often without being fully aware of what is going on. One test user said, 'In NREG, wages are sometimes paid without working also. But in the app work, wages are paid only when we work on it.' Still, in the earlier pilot 88% of test users thought 'app work' was better for receiving the correct amount of pay.

Finally, test users were asked, if they could choose only NREG or the app, which they would prefer. Fifty-three per cent of users with a preference answered that they would prefer the proposed scheme. Given the brief exposure the users had to the scheme, and the crudeness of the piloted app, it is noteworthy that more than half of them state that they would prefer it over NREG given a choice.

Table 8 presents a regression of whether test users prefer the piloted scheme ('app work') over traditional public works for themselves, on basic demographic characteristics. Those who responded 'No difference' or 'Cannot say' are coded with missing values and thus excluded from the analysis. No significant pattern across sex, age, schooling, marital status or caste group is detected. This suggests that while the appeal of the piloted scheme is not universal among test users, it is broad.

When earnings in the pilot are added to the right-hand side of the above regression, the coefficient is positive and significant (not shown). This may suggest that those who earn more in the pilot are more likely to prefer the scheme, though clearly earnings are potentially endogenous to scheme preference.

5 Conclusion

We pilot a form of social protection in which beneficiaries are paid, by the second, to engage with an educational smartphone app. In a sample of illiterate adults with experience of the existing Indian public-works programme, NREG, we randomly allocate access to the smartphone-based scheme for a week. We find large and significant effects on the ability of participants to recognise characters of the local script, a key pre-literacy skill, when they are exposed to a subset of 10 characters. However, there was no effect on character recognition in earlier pilot where participants were exposed to all 52 characters, arguably because the 'syllabus' was overly ambitious.

There is some evidence that the piloted scheme crowds out paid work, although it is not clear that this would persist if the intervention lasted longer as users may then test the scheme before making decisions about how to spend their time. If crowding out did persist, and if this were deemed undesirable, it would be straightforward to adjust the rate of pay in the scheme so as to make it less attractive.

Across both pilots, a majority of test users find the piloted scheme more dignified and less likely to be associated with fraud than the existing scheme, and about half would prefer it for themselves given a choice. Preferring the piloted scheme is not predicted by basic demographic characteristics, suggesting the scheme would have broad appeal.

The scale of the pilots discussed here was too small, and the intervention too short, to study the potential for reductions in fraud, costs and administrative delays compared to traditional public-works schemes, but this is something we hope to do in future work.

Overall, the results of the pilot suggest that the type of scheme proposed here could

be explored as a complement to existing public-works schemes, appealing to a different and broader set of demographics while potentially reducing costs.

References

Aggarwal A., 2017. Ten Ways MGNREGA Workers Do Not Get Paid. Economic And Political Weekly 52.

Alatas V., Purnamasari R., Wai-Poi M., Banerjee A., Olken B.A., Hanna R., 2016. Self-targeting: Evidence from a field experiment in Indonesia. Journal of Political Economy 124, 371–427.

Banerjee A., Duflo E., Imbert C., Mathew S., Pande R., 2020. E-governance, Accountability, and Leakage in Public Programs: Experimental Evidence from a Financial Management Reform in India. American Economic Journal: Applied Economics 12, 39–72. URL `https://doi.org/10.1257/app.20180302`.

Berg E., Rajasekhar D., Manjula R., 2017. Cellfare: Delivering self-targeted welfare using mobile phones. Centre for the Study of African Economies Working Paper 2017-14.

Besley T., Coate S., 1992. Workfare versus welfare: Incentive arguments for work requirements in poverty-alleviation programs. The American Economic Review 82, 249–261.

Coady D., Grosh M., Hoddinott J., 2004. Targeting of Transfers in Developing Countries: Review of Lessons and Experience. Washington, DC: World Bank.

Dutta P., Murgai R., Ravallion M., van de Walle D., 2012. Does India's Employment Guarantee Scheme Guarantee Employment? Economic and Political Weekly 47, 55–64. URL `https://www.jstor.org/stable/23214599`.

Fiszbein A., Schady N.R., 2009. Conditional Cash Transfers: Reducing Present and Future Poverty. World Bank Publications.

ILO, 2018. Women and men in the informal economy: A statistical picture. Geneva: International Labour Office, 3rd edition.

Imbert C., Papp J., 2011. Estimating Leakages in India's Employment Guarantee. In: R. Khera (Ed.), Battle for Employment Guarantee, 269–78. Oxford: Oxford University Press.

Imbert C., Papp J., 2014. Estimating leakages in India's Employment Guarantee: An Update. Background paper for the Social Protection and Labour India Team, World Bank. Technical report.

Kleven H.J., Kopczuk W., 2011. Transfer program complexity and the take-up of social benefits. American Economic Journal: Economic Policy 3, 54–90.

Ksoll C., Aker J.C., Miller D., Perez-Mendoza K.C., Smalley S.L., 2014. Learning Without Teachers? A Randomized Experiment of a Mobile Phone-Based Adult Education Program in Los Angeles. Center for Global Development Working Paper 368.

Nichols A.L., Zeckhauser R.J., 1982. Targeting transfers through restrictions on recipients. The American Economic Review 72, 372–377.

Ravallion M., 1991. Reaching the rural poor through public employment: Arguments, evidence, and lessons from South Asia. The World Bank Research Observer 6, 153–175.

Ravallion M., 2019. Is a Decentralized Right-to-Work Policy Feasible? In: K. Basu, M. Ghatak, K. Kletzer, S. Mundle, E. Verhoogen (Eds.), Development, Distribution, and Markets. New Delhi: Oxford University Press.

Rodriguez-Segura D., 2022. EdTech in developing countries: A review of the evidence. The World Bank Research Observer 37, 171–203.

Sukhtankar S., 2017. India's National Rural Employment Guarantee Scheme: What Do We Really Know about the World's Largest Workfare Program? Brookings-NCAER India Policy Forum 13, 231–286.

Witsoe J., 2014. The practice of development: An ethnographic examination of the national rural employment guarantee act in Bihar. Technical report, Mimeo, Union College.

Table 1: Summary statistics and balance check

	Control	Treatment	Normalized difference
Age	56.246	57.368	0.088
	(9.703)	(8.310)	
Female	0.807	0.807	0.000
	(0.398)	(0.398)	
Years of schooling	0.246	0.228	-0.020
	(0.635)	(0.598)	
Married	0.667	0.807	0.221
	(0.476)	(0.398)	
SC/ST	0.123	0.123	0.000
	(0.331)	(0.331)	
Taught characters identified correctly	2.596	2.544	-0.011
	(3.278)	(3.333)	
Observations	57	57	

Note: Summary statistics by experimental arm at baseline. Normalized differences in the third column. Standard deviations in parentheses.

Table 2: Main results

	Taught characters identified correctly (1)	Taught characters identified correctly (2)	Other characters identified correctly (3)	Words read correctly (4)
Treated	-0.0526			
	(0.621)			
After	0.255	0.241	0.796	0.352
	(0.270)	(0.392)	(0.590)	(0.310)
Treated × After	1.767***	1.684**	-0.853	-0.144
	(0.443)	(0.643)	(1.000)	(0.449)
Individual fixed effects	No	Yes	Yes	Yes
Control-group mean at baseline	2.596	2.596	3.088	0.561
Observations	221	221	221	221
Individuals observed twice	107	107	107	107

Note: OLS regressions. In columns 1 and 2, the dependent variable is the number of taught characters identified correctly (0–10), in column 3 it is the number of 'other' (not taught) characters identified correctly and in column 4 it is the number of words read correctly. Standard errors, clustered at the individual level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Attempts

	Taught characters attempted	Other characters attempted	Words attempted
After	0.889	0.0370	0.222
	(0.618)	(1.409)	(0.380)
Treated × After	2.507**	0.755	0.193
	(0.980)	(2.023)	(0.666)
Individual fixed effects	Yes	Yes	Yes
Observations	221	221	221
Individuals observed twice	107	107	107

Note: OLS regressions. The dependent variables are the number of taught characters, 'other' (not taught) characters and words attempted, whether identified correctly or not. Standard errors, clustered at the individual level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Hit rates

	Taught characters hit rate	Other characters hit rate	Words hit rate
After	-0.0404	0.0361	0.344
	(0.0989)	(0.210)	(0.244)
Treated × After	-0.0145	-0.0702	-0.343
	(0.120)	(0.233)	(0.264)
Individual fixed effects	Yes	Yes	Yes
Observations	160	91	50
Individuals observed twice	63	28	16

Note: OLS regressions. The dependent variables are proportions of taught characters, 'other' (not taught) characters and words identified correctly, out of those attempted. Standard errors, clustered at the individual level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Who learns?

	Characters learnt
Female	0.239
	(0.814)
Age	-0.494
	(0.373)
Age, squared	0.00470
	(0.00332)
Any schooling	-1.904***
	(0.666)
Married	0.403
	(1.117)
SC/ST	0.400
	(0.807)
Observations	53

Note: OLS regression. The dependent variable is the difference between the number of taught characters identified correctly at follow-up and at baseline. Only the treatment group is considered. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Time use

	NREG (public works)	Private paid work	Work on own land or business	Domestic work	Education
After	0.444	0.648	-1.926	0.667	-0.111
	(0.448)	(3.322)	(2.821)	(2.726)	(0.160)
Treated × After	-1.331	-5.441	-1.206	-2.082	-0.0776
	(1.001)	(4.226)	(4.239)	(4.451)	(0.315)
Individual fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	221	221	221	221	221

Note: OLS regressions. The dependent variables are the number of hours spent on each activity in the last seven days. Standard errors, clustered at the individual level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Comparing 'app work' to NREG work: Participant responses

Question	Proportion answering 'app work'
Which type of work	
... is more physically tiring?	2%
... is more mentally tiring?	35%
... is more dignified?	74%
... is better for building self-confidence?	52%
... provides better experience to take up other employment?	31%
... is more suitable during a pandemic such as Covid-19?	98%
... would help you to learn something new?	98%
... is more suitable for healthy, working-age men?	0%
... is more suitable for healthy, working-age women?	22%
... is more suitable for women who are pregnant?	100%
... is more suitable for women who have young children?	98%
... is more suitable for the elderly?	93%
... is more suitable for people in poor health?	100%
... is more suitable for the disabled?	100%
... is more suitable for the illiterate?	92%
... is better for the overall development of the village?	2%
... is better for receiving the correct amount of pay due to each worker?	52%
... is better for better for receiving payment on time?	92%
... is more likely to be available whenever the participant needs it?	10%
... is likely to have less fraud and corruption associated with it?	94%
If you could choose to work only with NREG or with the app, which would you choose personally?	53%

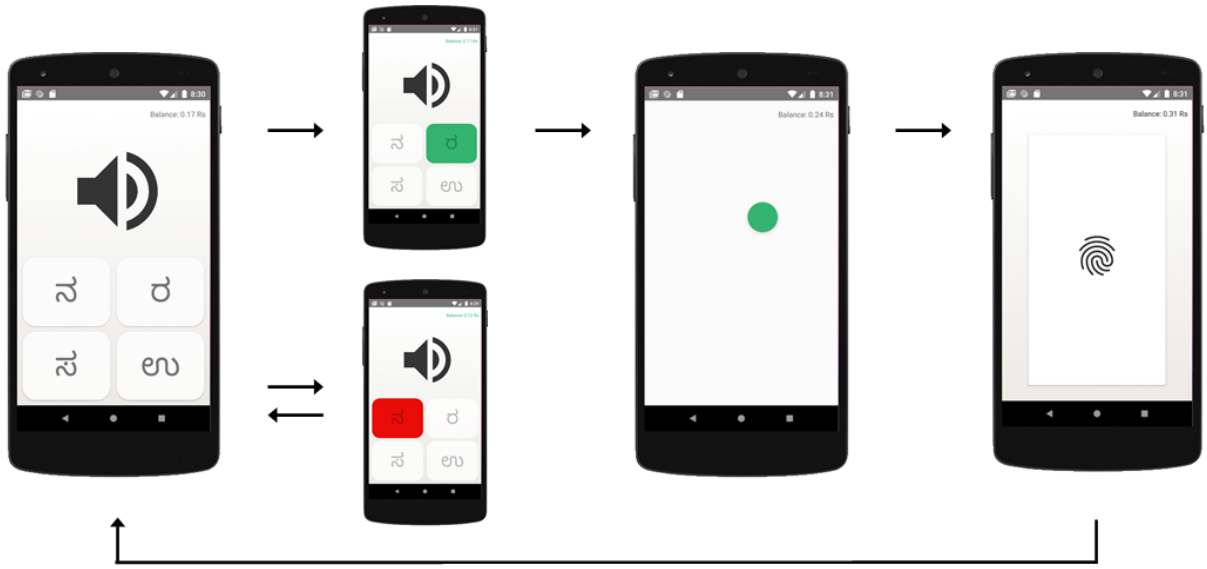
Note: The four answer options for all questions listed here were 'App work', 'NREG work', 'No difference' and 'Cannot say'. The second column shows the proportion who responded 'App work' to the question, out of those who selected either 'NREG work' or 'App work'. These questions were only asked of the treatment group, as only these have experience with both types of work.

Table 8: Who prefers the piloted scheme over traditional public works?

	Prefers piloted scheme
Female	-0.187
	(0.195)
Age	-0.0280
	(0.0803)
Age, squared	0.000325
	(0.000704)
Any schooling	0.198
	(0.220)
Married	-0.0983
	(0.205)
SC/ST	-0.233
	(0.252)
Observations	49

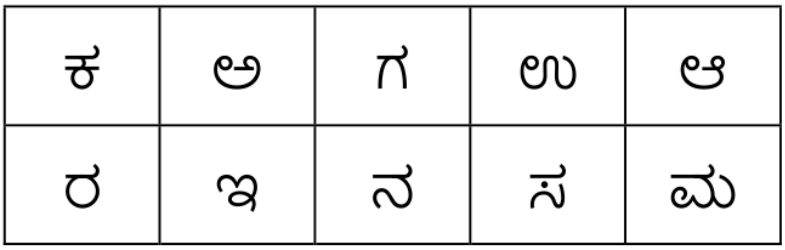
Note: OLS regression. The dependent variable is a binary variable indicating whether the participant would prefer 'app work' over NREG work for themselves, given the choice. Those who responded 'No difference' or 'Cannot say' are coded as missing. Only the treatment group is considered. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 1: Main app screens



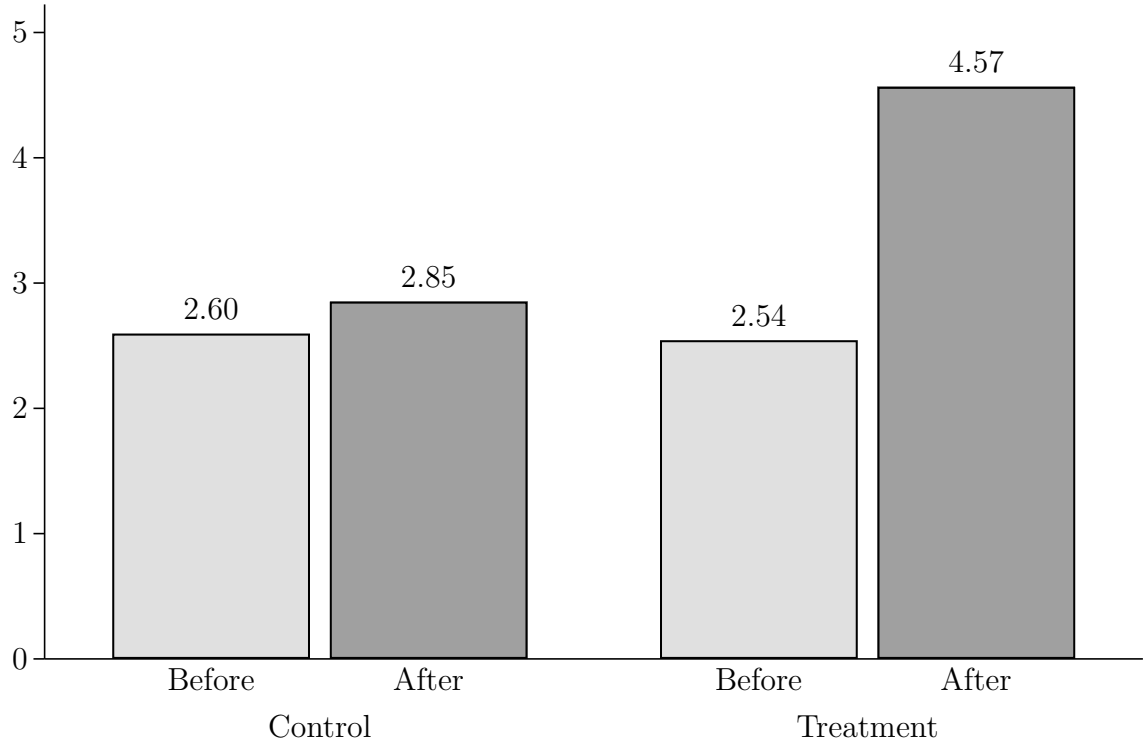
Note: The figure shows the main screens of the app. The task screen shows four randomly selected Kannada characters while playing the sound of one of them. The sound can be repeated by pressing the sound icon. The task screen is displayed until the correct button is pressed. On the attention check screen, the user needs to press a green dot that appears at a random location. The finger print screen prompts the reader to authenticate biometrically, before the cycle starts again.

Figure 2: Kannada characters



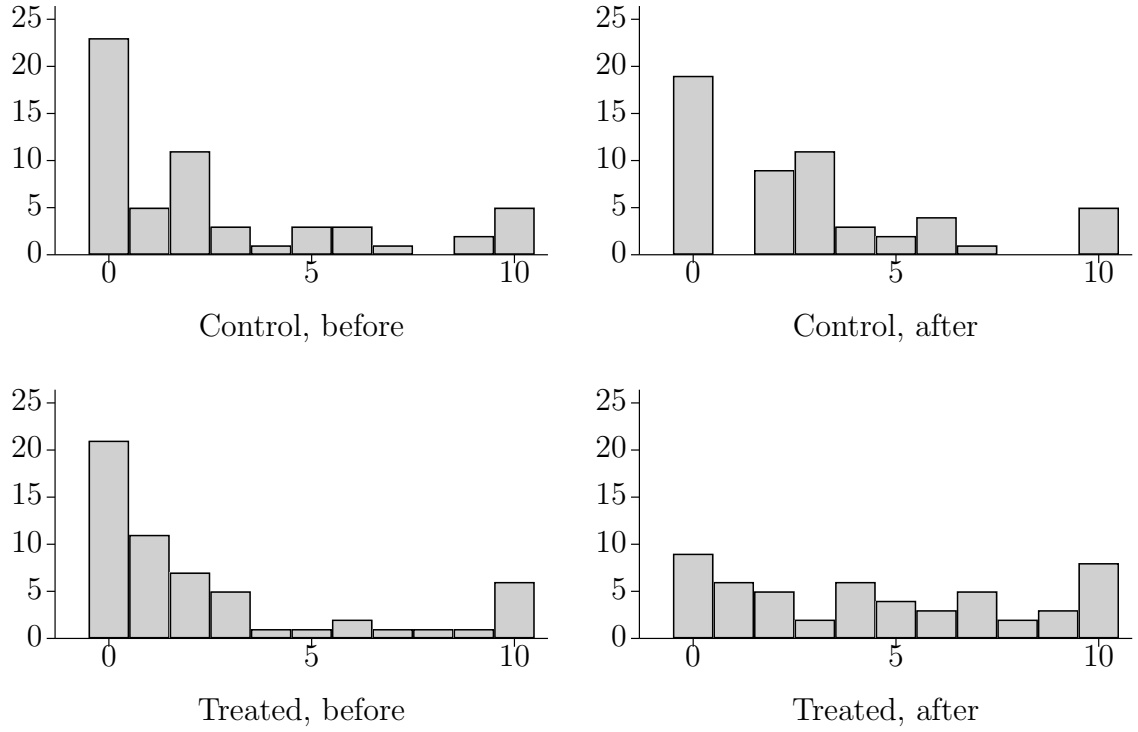
Note: The figure shows the basic characters of the Kannada script, highlighting the ten characters taught in the app.

Figure 3: Number of characters recognised before and after intervention, by experimental arm



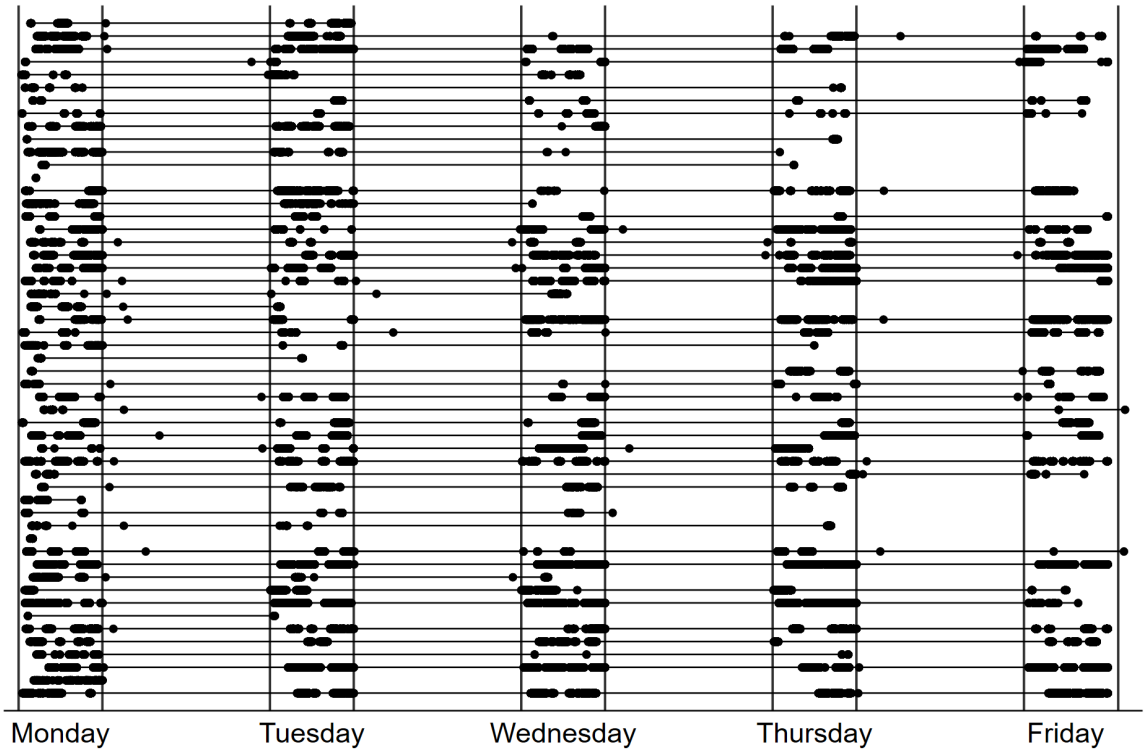
Note: The graph shows the mean number of characters correctly identified in the literacy test, out of the ten characters taught, before and after the intervention and by experimental arm.

Figure 4: Number of characters recognised before and after intervention, by experimental arm



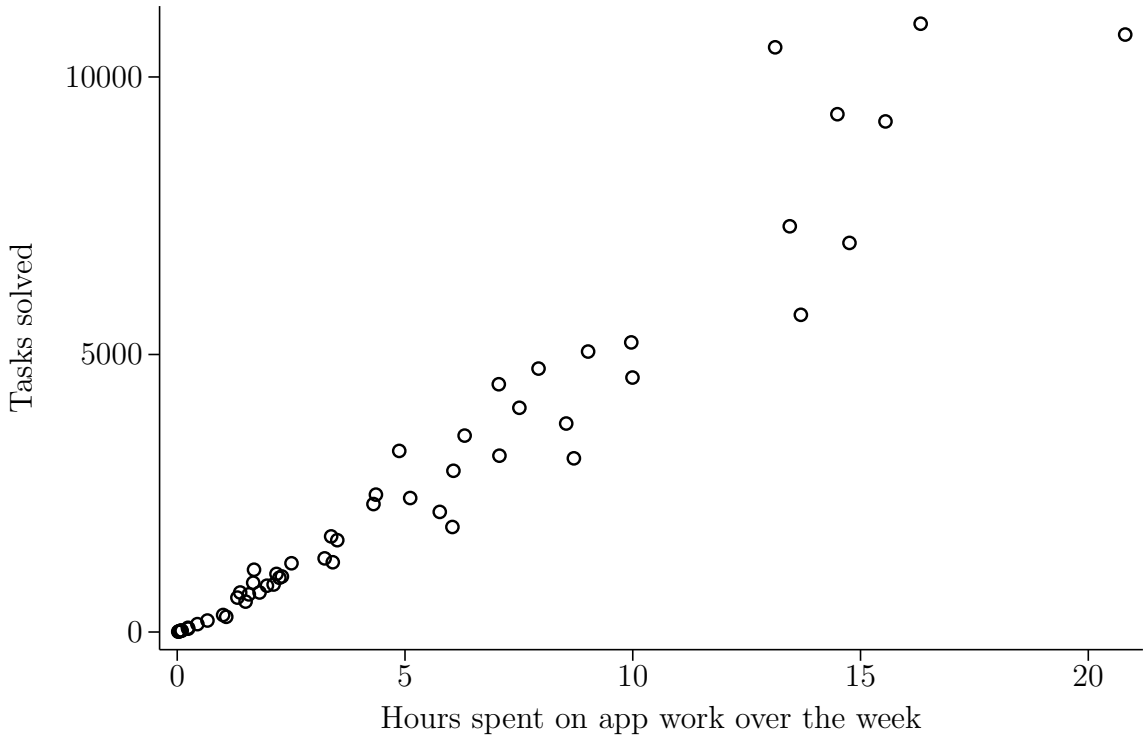
Note: Histograms showing the number of characters correctly identified, out of the ten characters taught, before and after the intervention and by experimental arm.

Figure 5: App usage



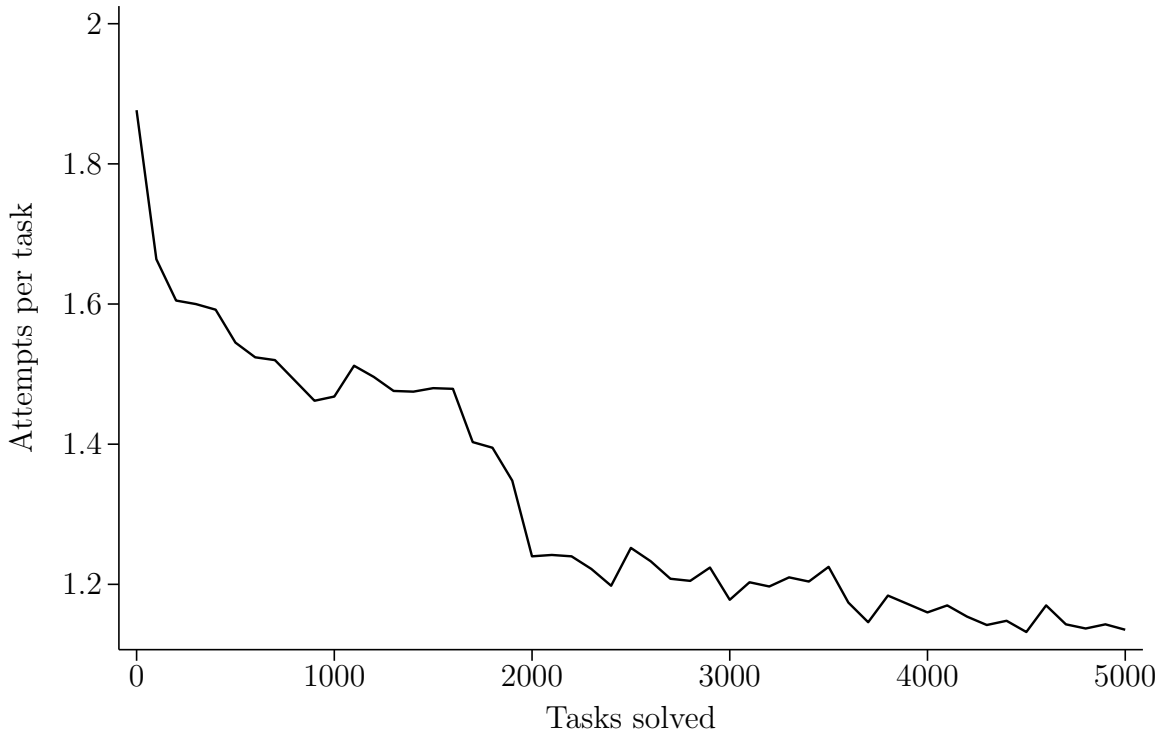
Note: The figure shows each user's engagement with the app over the week they had access to it. Horizontal lines represent individual users, and each black dot indicates the day and time that the user solved a task correctly. Each working day is demarcated by vertical lines at 9am and 5pm. The small number of tasks registered as completed outside of these times relate to system or network delays.

Figure 6: The number of tasks solved plotted against time spent engaging with the app



Note: The figure shows, for each treated participant, the total number of tasks solved plotted against the time spent engaging with the app over the week.

Figure 7: Evolution of the number of attempts required to solve a task correctly



Note: The figure shows the average number of attempts required by a user to solve a task correctly, as a function of the number of tasks solved previously. The downward trend is indicative of learning, as users require fewer attempts to solve each task correctly. The minimum achievable is 1, when every task is solved correctly at the first attempt. The curve has been smoothed by averaging the number of attempts required for the first 100 tasks, the next 100 tasks, and so on. Only the ten users who solved at least 5000 tasks are included, as otherwise the graph would be distorted by survivorship bias.

A The earlier pilot

The results of an earlier pilot are reported here. The main difference between the two pilots was that in the earlier instance, treated users were exposed to all 52 characters of the Kannada script, plus 15 vowel diacritics, over the one week they had access to the app. This contrasts with the second pilot in which, as explained above, participants were only exposed to a subset of 10 characters. Another difference was that while characters were drawn uniformly for each task for the pilot described above, in the earlier pilot the algorithm exposed users with higher frequency to characters they had previously got wrong. A further difference is that participants with up to four years of education were recruited in the first pilot, compared to a maximum of two years in the second pilot, as long as they self-identified as illiterate. Still, in both pilots a large majority of participants had no schooling.

In the earlier pilot, 117 illiterate adults with experience of NREG were recruited in Karnataka, South India, and randomly assigned to treatment and control arms.[14] Baseline summary statistics, by treatment arm, are presented in Table A1. None of the normalized differences exceed the conventional critical value of 0.25. Three intended treatment and two intended control participants dropped out between baseline and endline. Compared to the second pilot, the average participant was about six years younger, and more likely to belong to a scheduled caste or tribe.

Treated participants had access to the app for one week. Character recognition was measured before and after the intervention among all participants. Literacy results are presented in Table A2. There is no significant effect on the number of characters recognised correctly, or on the 'hit rate' — the proportion of characters recognised out of those attempted. The coefficient for the number of words read correctly is large relative to the control-group mean, but not significant at the 5% level.

In hindsight it seems unsurprising that there is little evidence of learning in the earlier pilot. Exposing illiterate participants to the full set of characters in one week was probably overly ambitious. With an intervention of longer duration it would probably be a good idea to start with a small set of characters and gradually add to it as the participants learn.

Self-reported time use analysis for the earlier pilot is presented in Table A3. It appears that the intervention made the treated group spend 13.4 fewer hours on paid work in the intervention week, indicating a substitution away from normal paid activities. While noteworthy, it is not necessarily the case that this effect would persist if the intervention duration were longer, as users may then try out the piloted scheme initially, only to return to their regular activities after a while. For comparison, the self-reported average number

[14]The fieldwork took place in the period July–September 2021 in four villages Madhugiri taluk, Tumkur district. Allocation to experimental arms was individual and stratified by village, sex and age.

of hours spent on the app in the treated group was 16.4. The effect on education is not identified since no participant (all adults) reported spending any time on this activity.

The results of the user poll are largely similar across the two pilots, with a few exceptions. Table A4 presents the proportion of treated participants who agreed with each statement, out of those that had an opinion. One difference is that more participants in the earlier pilot find the 'app work' more mentally tiring, perhaps because of the additional mental load of being exposed to the entire character set. Still, 90% thinks that app work is more dignified than NREG. It is still the case that the app work is thought to be less suitable for healthy, working-age men and women and more suitable for pregnant women, the disabled and those in poor health. App work is still thought to be associated with less fraud and delays in payment, but less good for the overall development of the village. Out of the 46 treated respondents who stated a preference, 66% would personally prefer app work over NREG work if they could only choose one.

Table A5 reveals that none of the basic demographic characteristics considered are associated (at the 5% significance level) with preferring the piloted scheme over traditional public works, again lending support to the idea that the appeal is broad.

Table A1: Summary statistics and balance check for the earlier pilot

	Control	Treatment	Normalized difference
Age	50.707	52.542	0.126
	(10.79)	(9.520)	
Female	0.603	0.678	0.109
	(0.493)	(0.471)	
Years of schooling	0.603	0.407	-0.129
	(1.138)	(1.002)	
Married	0.845	0.847	0.005
	(0.365)	(0.363)	
SC/ST	0.414	0.322	-0.133
	(0.497)	(0.471)	
Taught characters identified correctly	5.50	4.57	-0.088
	(7.13)	(7.77)	
Observations	58	59	

Note: Summary statistics by experimental arm at baseline. Normalized differences in the third column. Standard deviations in parentheses.

Table A2: Literacy results for earlier pilot

	Characters identified correctly (1)	Characters identified correctly (2)	Character hit rate (3)	Words read correctly (4)
Treated	-0.929			
	(1.413)			
After	3.286*	3.286	0.00995	-0.357
	(1.749)	(2.473)	(0.135)	(0.330)
Treated * After	3.411	3.411	0.101	0.804*
	(2.671)	(3.778)	(0.173)	(0.438)
Individual fixed effects	No	Yes	Yes	Yes
Control-group mean at baseline	5.50	5.50	10.1	1.02
Observations	224	224	150	224
Individuals observed twice	112	112	53	112

Note: OLS regressions. In columns 1 and 2, the dependent variable is the number of taught characters identified correctly (0–10), in column 3 it is the number of 'other' (not taught) characters identified correctly and in column 4 it is the number of words read correctly. Standard errors, clustered at the individual level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Time use in the earlier pilot

	NREG (public works)	Private paid work	Work on own land or business	Domestic work	Education
After	-0.321	2.875	-3.625	0.911	0
	(0.456)	(4.222)	(4.151)	(3.477)	(.)
Treated × After	-1.107	-13.43**	-3.964	2.143	0
	(1.598)	(5.806)	(7.022)	(5.188)	(.)
Individual fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	224	224	224	224	224

Note: OLS regressions. The dependent variables are the number of hours spent on each activity in the last seven days. The coefficients for education are not identified since none of the participants reported spending any time on this activity. Standard errors, clustered at the individual level, are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Comparing 'app work' to NREG work in the earlier pilot: Participant responses

Question	Proportion answering 'app work'
Which type of work	
... is more physically tiring?	8%
... is more mentally tiring?	64%
... is more dignified?	90%
... is better for building self-confidence?	90%
... provides better experience to take up other employment?	50%
... is more suitable during a pandemic such as Covid-19?	100%
... would help you to learn something new?	100%
... is more suitable for healthy, working-age men?	2%
... is more suitable for healthy, working-age women?	9%
... is more suitable for women who are pregnant?	100%
... is more suitable for women who have young children?	100%
... is more suitable for the elderly?	100%
... is more suitable for people in poor health?	100%
... is more suitable for the disabled?	100%
... is more suitable for the illiterate?	81%
... is better for the overall development of the village?	29%
... is better for receiving the correct amount of pay due to each worker?	88%
... is better for better for receiving payment on time?	92%
... is more likely to be available whenever the participant needs it?	13%
... is likely to have less fraud and corruption associated with it?	93%
If you could choose to work only with NREG or with the app, which would you choose personally?	66%

Note: The four answer options for all questions listed here were 'App work', 'NREG work', 'No difference' and 'Cannot say'. The second column shows the proportion who responded 'App work' to the question, out of those who selected either 'NREG work' or 'App work'. These questions were only asked of the treatment group, as only these have experience with both types of work.

Table A5: Who prefers the earlier piloted scheme over traditional public works?

	Prefers piloted scheme
Female	0.189
	(0.150)
Age	0.160*
	(0.0914)
Age squared	-0.00159*
	(0.000858)
Any schooling	0.153
	(0.182)
Married	0.292
	(0.213)
SC/ST	0.101
	(0.145)
Observations	47

Note: OLS regression. The dependent variable is a binary variable indicating whether the participant would prefer 'app work' over NREG work for themselves, given the choice. Those who responded 'No difference' or 'Cannot say' are coded as missing. Only the treatment group is considered. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.