

Evaluation

Fabian Gunzinger*
Warwick Business School

August 24, 2022

Contents

1	Introduction	2
2	Methods	2
2.1	Data	2
2.2	Estimation	5
2.3	Variables	7
2.4	Code access	10
3	Results	10
3.1	Main results	10
3.2	Intensive and extensive margins	12
4	Discussion	13
A	Robustness checks	15
A.1	Relaxing anticipation assumption	15
A.2	Unbalanced aggregation	15
A.3	Inflows and outflows	16

*fabian.gunzinger@warwick.ac.uk

Abstract

Neat and succinct abstract right here...

1 Introduction

2 Methods

2.1 Data

Dataset description: I use data from a UK-based financial management app that allows users to link accounts from different banks to obtain an integrated view of their finances. The complete dataset contains more than 500 million transactions made between 2012 and June 2020 by about 250,000 users, and provides information such as date, amount, and description about the transaction as well as account and user-level information. Crucially for this paper, the app can access up to three years of historic data for each linked account.

The main advantage of the data for the study of consumer financial behaviour is that we can observe user behaviour at the transaction level across all their accounts, and that the data is automatically collected rather than collected through a survey.

The main limitation is the non-representativeness of the sample relative to the population as a whole. Financial management apps are known to be used disproportionately by men, younger people, and people of higher socioeconomic status (Carlin et al. 2019). Also, as pointed out in Gelman et al. (2014), a willingness to share financial information with a third party might not only select on demographic characteristics, but also for an increased need for financial management or a higher degree of financial sophistication. Because our analysis does not rely on representativeness, we do not address this.¹

Cleaning: I use the dataset described above for a number of projects, and perform a number of steps to create a minimally cleaned version of the dataset that is the basis for all such projects. These steps are performed in a dedicated data repository and not run as part of this project, but the module with all cleaning functions is available in the project directory.²

Here, I briefly describe the main cleaning steps and their rationale. I drop all transactions with a missing description string because these cannot be categorised, and all transactions that are not automatically categorised by the app. Dropping these transactions makes is likely that we will underestimate amounts spent and saved, but minimises the risk of incorrectly classified transactions. I group transactions into transaction, spend, and income subgroups. Spend subgroups are defined following Muggleton et al. (2020); income subgroups, following Hacıoglu et al. (2020).³ Finally, I classify as duplicates and drop transactions with identical user ID, account ID, date, amount, and transaction description. This will drop some genuine transactions, such as a user buying two identical cups of coffees at the same coffee shop on the same day. However, data inspection suggests that in most cases, we remove genuine duplicates.

¹For an example of how re-weighting can be used to mitigate the non-representative issue, see Bourquin et al. (2020).

²Link to cleaning functions: [🔗](#)

³Link to classification file: [🔗](#)

Table 1: Sample selection

	Users	User-months	Txns	Txns (m£)
Raw sample	271,856	7,948,520	662,112,975	124,573
Drop test users	270,782	7,878,398	656,047,534	122,887
App signup after March 2017	88,368	2,320,421	202,580,838	38,816
At least one savings account	50,226	1,334,328	125,841,337	26,645
At least one current account	48,794	1,303,164	123,468,715	26,263
At least £5,000 of annual income	20,647	541,746	55,857,451	11,760
At least 10 txns each month	14,229	369,944	40,662,904	8,529
At least £200 of monthly spend	10,438	272,228	31,529,498	6,837
No more than 10 active accounts	9,788	248,975	27,589,696	5,426
Complete demographic information	7,720	202,633	22,671,753	4,378
Working age	7,568	197,951	22,279,279	4,213
Final sample	7,568	197,951	22,279,279	4,213

Notes: Number of users, user-months, transactions, and transaction volume in millions of British Pounds left in our sample after each sample selection step. Link to sample selection code: [🔗](#).

Sample selection: We select our sample so as to include users for whom we can be reasonably certain that we observe all relevant financial transactions, and do so for at least six months before and after they sign up to the app. In addition to that, we exclude users who might use the app for business purposes as well as pensioners, whose financial objectives might be different.

Table 1 lists the precise conditions we applied to implement these criteria and their effect on sample size. We remove the first and last month of data for all users because we are unlikely to observe all transactions for these months. We also drop test users, since their objectives for app use might have been different from ordinary users.⁴

To ensure that we observe users for at least 12 months around app signup, we require 6 months of data before the signup month, and another five months after the signup month. Our main outcome variable is netflows into a user’s savings accounts. It is thus critical that we observe enough historical data for these savings accounts to ensure that we observe all transactions during our 12 month period of interest. This is complicated by the fact that we cannot see when an account was opened at the bank, but only when it was added to the app. While cases where a user adds an account to the app as soon as it was opened are unproblematic, users will often add accounts after they were opened, either because they have accounts that they opened before signing up to the app, or because they opened new accounts after signup but add them to the app with a delay. In such cases, it is critical that, once the account is added, we observe the complete historical data up to 6 months before signup or up to the month in which the account was opened, whichever happened later. To see why this is critical, imagine a scenario where a user opens an account 10 months before they sign up to the app, makes a monthly transfer to the account of £100, adds the account to the app on signup, but we observe only 3 months of historical data. In this case we would observe that the user saved £300 before signup and £600 after, and erroneously conclude that post signup savings were twice as high. The most extreme case we need to cover is that of a user opening a savings account more than six months before

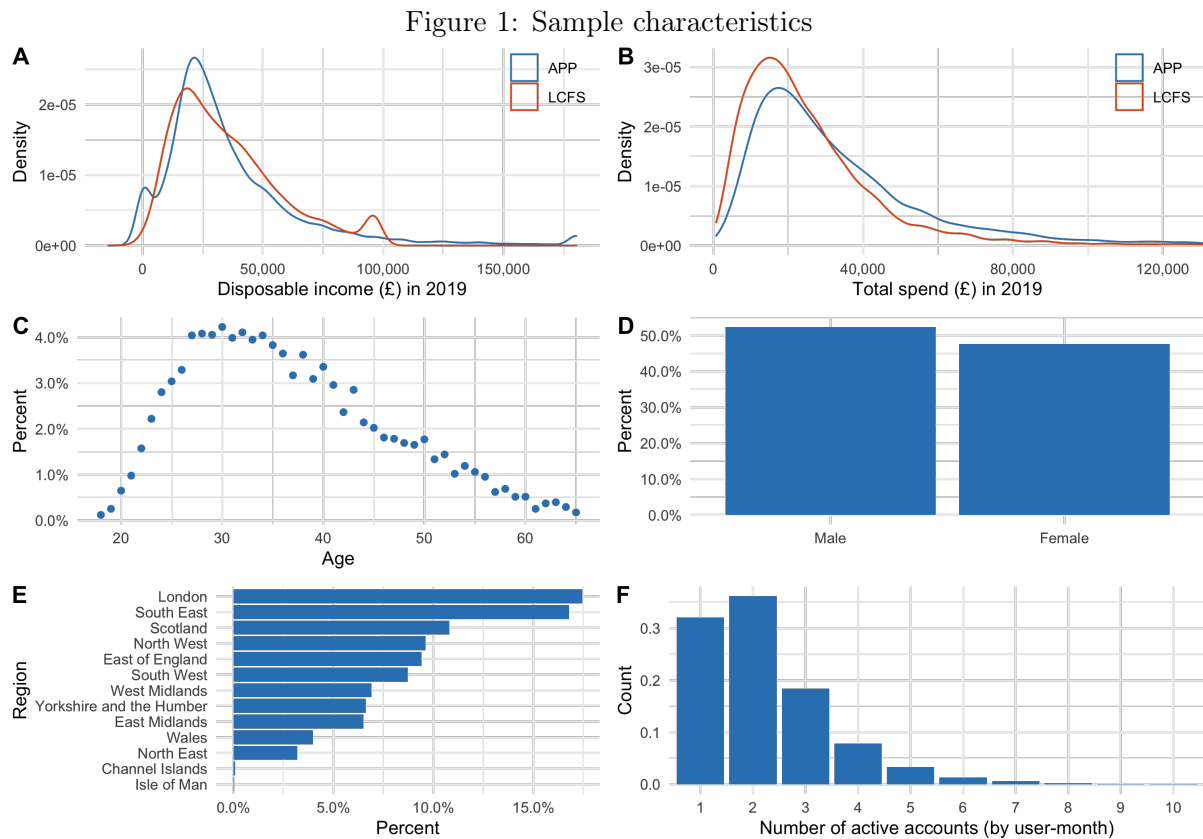
⁴We cannot identify test users precisely, but drop users who signed up prior to or during the first year the app was in operation.

signup and adding the account to the app five months after signup, in which case we need to be sure to observe 12 months of historical data. As shown in Appendix ??, all major banks started providing 12 months of historical data for current and savings accounts from April 2017 onwards, which is why we restrict our sample to users who signed up in or after that month.

To ensure that we can be reasonably certain to observe users have added all their financial accounts to the app, we restrict our sample to users with at least one savings and current account, with an annual income of at least £5,000, and a minimum of 10 transactions and a spend of £200 every month. To remove users who might use the app for business purposes, we drop users with more than 10 active accounts in any given month. Finally, we remove users for whom we cannot observe all demographic information we use as covariates in our analysis, and users who are not between the ages of 18 and 65, as their financial objectives are plausibly different.

Data transformations: To minimise the influence of outliers, we winsorise spend, income, and savings accounts flow variables at the 1 percent level or – if we winsorise on both ends of the distribution – at the 0.5 percent level.

Summary statistics: Figure 1 describes the sample.



Notes: Panels A and B show the distribution of disposable income and total spending in 2019, respectively, benchmarked against the 2018/19 wave of the ONS Living Cost and Food Survey (LCFS). The remaining panels show the data distributions of age, gender, region, and the number of active accounts.

Table 2 provides summary statistics.

Table 2: Summary statistics

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Txn count	111.9	59.7	10	70	101	142	327
Month income	2,853.5	2,495.2	0.0	1,407.2	2,217.4	3,595.9	15,027.5
Savings account inflows	747.2	2,448.3	0.0	0.0	0.0	400.0	18,809.5
Savings account outflows	762.4	2,422.9	0.0	0.0	0.0	400.0	18,099.4
Savings account netflows	-7.4	2,883.7	-20,000.0	0.0	0.0	50.0	21,675.4
Month spend	2,760.4	2,609.9	200.0	1,225.4	2,016.7	3,318.9	17,092.2
Age	37.5	10.0	18	30	36	44	65
Female dummy	0.4	0.5	0	0	0	1	1
Urban dummy	0.8	0.4	0	1	1	1	1
Discretionary spend	860.6	736.1	0.0	369.3	663.0	1,118.3	4,181.7
Active accounts	3.1	1.7	1	2	3	4	10

We use data from the 2018-2019 wave of the Office of National Statistics’ Living Costs and Food Survey (LCFS).⁵ Data covers the period between April 2018 and March 2019.

2.2 Estimation

We want to estimate the effect of app use over time. Given our data, a natural way to do this would be to use a dynamic two-way fixed effects model that includes user and year-month fixed effects and dummies indicating time since app signup. The estimated coefficients on these dummies are then conventionally interpreted as dynamic treatment effects. However, a series of recent papers have documented that while such an approach is frequently used in applied research, the parameter estimates from dynamic two-way fixed effects models are not valid estimators of dynamic treatment effect in most settings. In particular, in settings with staggered treatment assignment, where units are first exposed to treatment at different points in time (as is the case in our setting), dynamic two-way fixed effects are valid only if there is homogeneity in treatment effects across treatment adoption cohorts. In most settings, including our own, this is a very strong assumption.⁶

Because of this, I use a new estimator proposed by Callaway and Sant’Anna (2021), which allows for arbitrary treatment effect heterogeneity across treatment adoption cohorts and time, and allows for the incorporation of a parallel trends assumption conditional on covariates. In describing the estimator, I follow the approach of Callaway and Sant’Anna (2021) of first defining the causal parameter of interest, and then discussing identification, estimation, and inference.

The basic building block of the framework, and the causal effect of interest, is the group-time average treatment effect: the average treatment effect at time t for the group of individuals first treated at time g , defined as:

$$ATT(g, t) = \mathbb{E}[Y_{i,t}(g) - Y_{i,t}(0) | G_i = g], \quad (1)$$

where $Y_{i,t}(g)$ is the potential outcome in time period t of an individual i in group g , and $Y_{i,t}(0)$ is the (counterfactual) potential outcome of that same individual if they had remained untreated.

⁵We accessed the data via the UK Data Service at the following url: <https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8686>.

⁶Two excellent reviews of this new literature are Roth et al. (2022) and Baker et al. (2022).

These effects are identified if two main assumptions hold: if there is limited and known anticipation of treatment, and if the assumption of parallel trends between treatment and comparison groups holds either unconditionally or conditionally on a set of covariates.⁷ Because the purpose of this section is to convey the core idea of the estimation approach, I keep things as simple as possible and discuss only the case with no anticipation effects and where the parallel trend assumption holds unconditionally. Callaway and Sant’Anna (2021) show that the same overall approach also works when allowing for known anticipation and conditional parallel trends.⁸

Given these two assumptions, Callaway and Sant’Anna (2021) show that $ATT(g, t)$ is identified by comparing the expected change of group g between periods t and $g - 1$ with that of a comparison group that is not yet treated at time t :

$$ATT(g, t) = \mathbb{E}[Y_{i,t} - Y_{i,g-1} | G_i = g] - \mathbb{E}[Y_{i,t} - Y_{i,g-1} | G_i = g'], \text{ for any } g' > t \quad (2)$$

As this holds for all g' that are not yet treated at time t , it also holds for an average over all groups in the set \mathcal{G} containing all groups g' for which $g' > t$,

$$ATT(g, t) = \mathbb{E}[Y_{i,t} - Y_{i,g-1} | G_i = g] - \mathbb{E}[Y_{i,t} - Y_{i,g-1} | G_i \in \mathcal{G}]. \quad (3)$$

This equation encapsulates two main results: that the period just before treatment, $g - 1$, is a valid reference period, and that the group of all individuals that have not yet been treated at time t are a valid comparison group for estimating treatment effects in time t .⁹ As a result of this, units who are treated in the very first period in the dataset are dropped from the sample, since there exists no possible control group based on which to identify their treatment effect, and since they are not useful as a control group themselves. Similarly, unit treated in the very last period in the data are also dropped, since there exists no “not-yet-treated“ group that could serve as a comparison group for them.

$ATT(g, t)$ can then be estimated by replacing expectations with their sample analogues,

⁷Additional assumptions are (i) that the treatment is absorbing in the sense that once an individual is treated they will remain treated forever, (ii) that individuals in the data are randomly and independently drawn from a larger population, and (iii) an overlap condition that ensures that there is a positive number of users that is first exposed to the treatment at any period and that – under the conditional parallel trends assumption – propensity scores for initial treatment times based on covariates are bounded away from zero. The first assumption could be violated only if a user closes their account on the app and then signed up again later on, all within the roughly within the two-year data periods I use. This cannot be more than a tiny minority of users. The second assumption holds less trivially. One way to think of a super population from which the users in the dataset are drawn is to think of knowledge about the app as partially random, and about the super population of all individuals who would have signed up had they learned about the apps existence.

⁸Intuitively, known anticipation merely shifts the reference period from the period immediately before treatment to the period before anticipation of treatment begins. When relying on the conditional parallel trends assumption, the overall the group-time average treatment effect $ATT(g, t)$ is the average of unconditional group-time average treatment effects for each value of the covariate vector X_i . As discussed in Roth et al. (2022), estimation is challenging when X_i is continuous or can take on a large number of values, since then we will typically lack data to estimate unconditional group-time treatment effects for each value of X_i . There are different semi- and non-parametric approaches that can be used in such cases. Below, I use the doubly-robust estimator, which is the default in the ‘did’ package. See Callaway and Sant’Anna (2021) and Roth et al. (2022) for more details.

⁹If a group of never-treated individuals is available, then these could also serve as a comparison group. But because my dataset does not contain such individuals, I do not discuss this case.

$$\widehat{ATT}(g, t) = \frac{1}{N_g} \sum_{i:G_i=g} [Y_{i,t} - Y_{i,g-1}] - \frac{1}{N_g} \sum_{i:G_i \in \mathcal{G}} [Y_{i,t} - Y_{i,g-1}] \quad (4)$$

Once these building blocks are estimated, aggregating them to event-study type treatment effects that provide the (weighted) average treatment effect l periods away from treatment adoption across different adoption groups can be achieved by simply calculating

$$ATT_l = \sum_g w_g ATT(g, g + l), \quad (5)$$

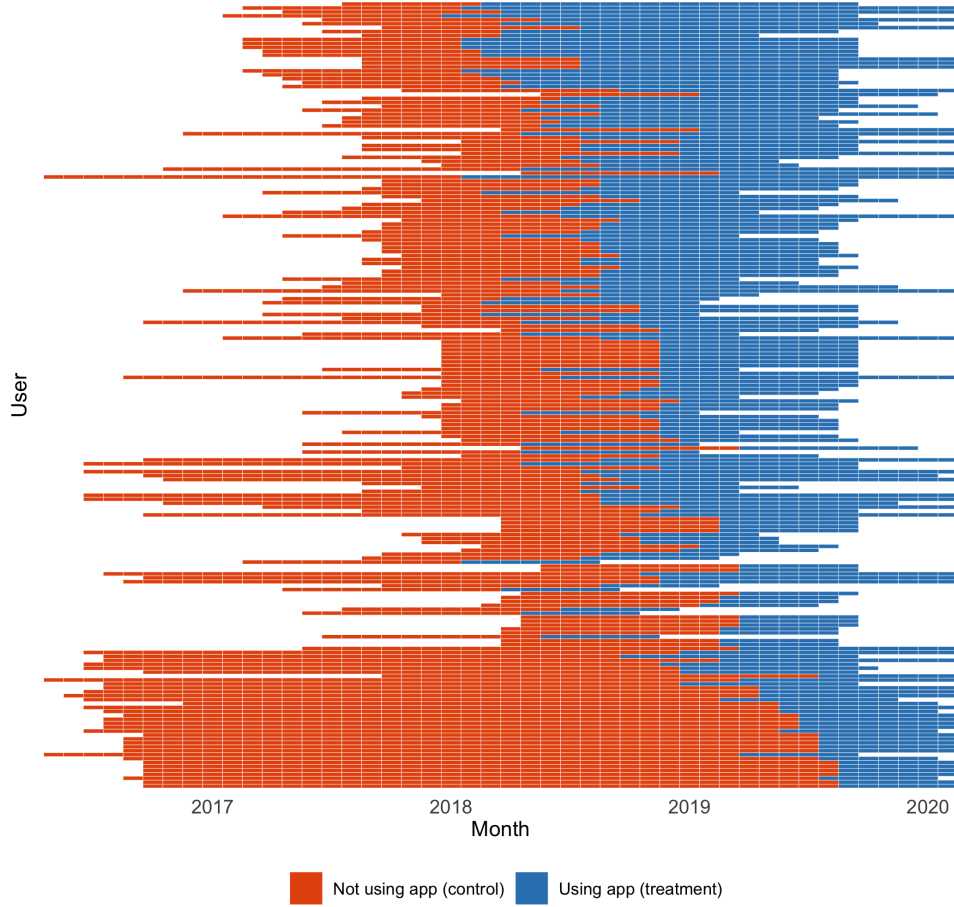
where the group weights w_g are the groups relative frequencies in the treated population. When calculating these event study parameters, I use a panel balanced in event times, with all units being observed for at least 5 treatment periods. This avoids the ATT_l being influenced by different group compositions at different periods l .¹⁰

2.3 Variables

Treatment A user changes treatment status from untreated to treated when they start using the app. Figure 2 shows the treatment history for 200 randomly selected users.

¹⁰See Section 3.1.1 in Callaway and Sant’Anna (2021) for a more detailed discussion.

Figure 2: Treatment assignment plot



Notes: Each horizontal line shows the observed pre and post signup periods in blue and red, respectively, for one of 200 randomly selected users. The faint vertical white lines indicate month borders, whitespace indicates periods in which we do not observe the user. To the left of the observed period, this is because the app cannot access data before that point when the user signs up; to the right, because they have stopped using the app.

Outcomes Savings... see Table 3 for details.

For a more nuanced understanding of how app use affects savings we also consider net-savings – total savings account inflows minus outflows – as a proportion of monthly income to see whether a willingness to save more might be offset by a (later) need to withdraw funds, and a dummy variable for whether a user has any savings account inflows in a given month to see whether the app helps users save at all. To investigate possible channels, we consider total spend, highly discretionary spend, banking charges, the total amount of borrowing, as well as payday borrowing, all as proportion of monthly income.

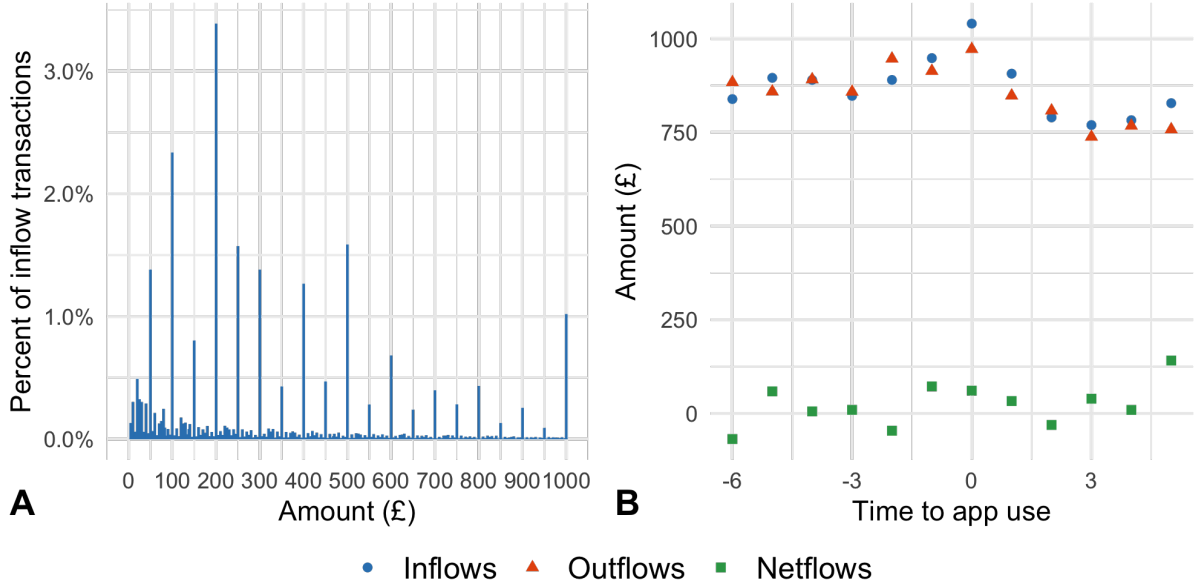
Net savings (*netflows_norm*) Inflows into minus outflows out of all of a user’s savings accounts divided by monthly income. To capture only “user-generated” flows, we exclude interest and “save the change” transactions, as well as transactions of less than £5 in absolute value. Monthly income and raw inflows and outflows are winsorised at the 1 percent level. We focus on net inflows to capture effective savings.

Positive net savings dummy (*has_pos_netflows*) Dummy equal to 1 if there were positive net savings (as defined above). Captures extensive margin of savings (change in number of months

with positive net deposits)

Positive net savings (*pos_netflows*) Equal to net savings if there were positive net savings. Captures intensive margin of savings (change in deposit amount in months with positive net deposits)

Figure 3: Savings patterns



Notes: Panel A shows distribution of savings account inflow amounts, making clear that most transactions are the kinds of round amounts we would expect savings transactions to be. The data is truncated at £1000. Panel B shows inflows, outflows, and netflows into savings accounts for six months before and five months after app use.

Covariates We control for baseline behaviour, events, and personal characteristics that, to various degrees, capture a person’s need, capacity, motivation, and awareness to save. Table 3 lists all covariates used together with their definition and the rationale for including them. For all variables, we include contemporaneous values as well as lags for up to 6 periods. In addition, we control for the previous six months of savings to capture time-invariant unobserved drivers of savings behaviour (in specifications without fixed effects) as well as a possible signal for a higher or lower need for future savings.

Following VanderWeele (2019) we include covariates that affect either outcomes or the propensity for treatment or both, exclude from this set of variables those that are instruments (affect the outcome only through their effect on treatment propensity) and add to it proxies for unobserved variables that are a common cause of both outcomes and treatment propensity.¹¹


The table below describes the construction and rationale for including of all variables used. The code used to construct the variables is available on [GitHub](#).

¹¹VanderWeele (2019) calls this the “modified disjunctive cause criterion” for covariate selection, as it includes the set of variables that are causally related to either outcomes, or treatment propensity, or both, but modified to account for potential bias by excluding instruments and including proxies of unobserved causes of both outcomes and treatment.

Table 3: Covariates

Variable (name in dataset)	Definition	Rationale
Primary outcome		
Covariates		
New loan dummy (<i>new_loan</i>)	Dummy variable equal to 1 if user takes out a new loan. Calculated positive inflows of funds tagged as “loan”.	Might increase (additional funds) or decrease (need to repay) propensity to save in month of takeout and lower propensity to save in the future due to need to repay.
Unemployment benefits dummy (<i>unemp_benefits</i>)	Dummy variable equal to 1 if user has inflow of funds tagged as “job seeker benefits”.	Might lower a user’s ability to save but increase their need for a money management app.
Monthly income (<i>month_income</i>)	Average monthly income in a calendar year, calculated as the sum of all credits tagged income payments in said year divided by 12.	Income may alter the need and ability to save and correlate with cognitive characteristics that alter a person’s propensity to use a money management app.

2.4 Code access

We provide links to code that creates key elements of the paper such as variable definitions and sample selection directly in the relevant places in the paper so they can be accessed conveniently. The links are indicated with the GitHub logo, . The hope is that this helps the curious reader clarify questions about subtleties they might have while reading the paper. The complete projects GitHub repo is at https://github.com/fabiangunzinger/mdb_eval.

3 Results

3.1 Main results

- Figure 4 shows the effect of app use on monthly discretionary spend (top row) and monthly net-inflows into savings accounts (bottom row) under the unconditional (left column) and conditional (right column) parallel trends assumptions.
- Estimates are group-time average treatment effects aggregated by time since treatment exposure.
- All results are presented with a uniform 95% confidence band, based on bootstrapped standard errors clustered at the user level that also account for autocorrelation in the data.¹²
- Conditional results use the doubly-robust estimator discussed in the methods section.
- We can see that discretionary spend falls by between £100 and £150 per month once users start using the app, depending on the parallel trends assumption used. Given that average monthly discretionary spend is about £860 (see Table 2), this corresponds to a drop in discretionary spend of about 11-17 percent, which is substantial.

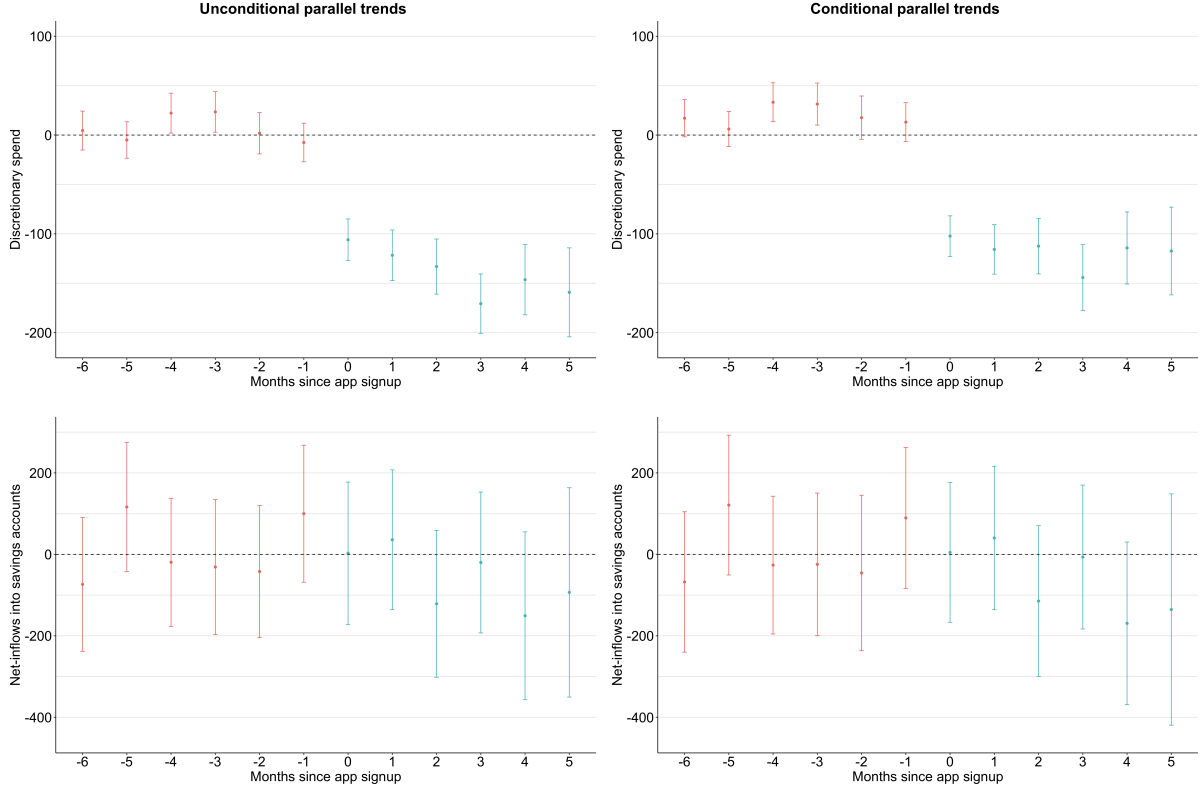
¹²A uniform 95% confidence band accounts for multiple hypothesis testing in that it is constructed such that *all* shown coefficients cover their corresponding true value 95 percent of the time. In contrast, a more commonly used pointwise 95% confidence band is constructed such that the confidence interval for each parameter covers the true parameter 95 percent of the time.

- Conditional parallel trends are important in contexts when (i) there are covariate specific trends in outcome paths and (ii) the distribution of covariates differs between groups. (E.g. people who sign up to job training differ from those who don't and job outcomes depend on these covariates)

Actually, for my here, cov distr between treatment and control differs only if early users are different from later ones.

- Given that our comparison group is the set of “not-yet-treated“ users rather than a set of “never-treated“ users, the relatively small difference in results is as expected.
- At the same time, we would have expected there to be some difference. If we discretionary spend is a constant fraction of income and total spend, then we would expect parallel trends to hold only for groups with the same income and total monthly spend. Similarly, if we think that the average spend per account observed is constant, then parallel trends hold only for users with the same number of observed accounts (if we think of active accounts as observed accounts).
- In contrast to discretionary spend, net-inflows into savings accounts do not change once users start using the app. The wide confidence bands reflect the large variation in net-inflows already seen in Table 2.
- Results indicate that parallel trend assumption might not hold. So we should interpret these result with some caution.

Figure 4: Main results

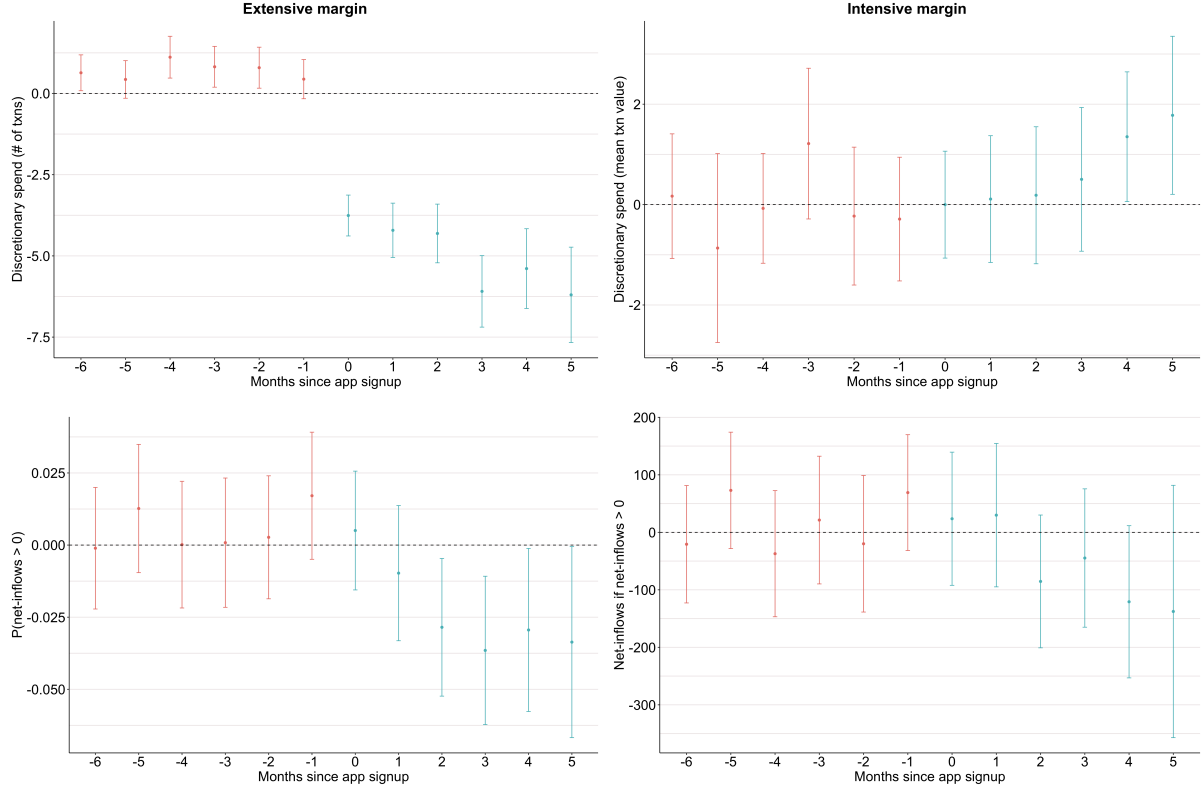


Notes: The effect of app use on monthly discretionary spending (top row) and monthly net-inflows into savings accounts (bottom row) under the unconditional (left column) and conditional (right column) parallel trends assumption. Point estimates represent group-time average treatment effects aggregated to periods since treatment exposure, as defined in Section 2.2. Red lines represent point estimates and uniform 95% confidence bands for pre-treatment periods allowing for clustering at the user level. If the null hypothesis that parallel trends hold in all periods is correct, these should be equal to zero. Blue lines provide similar information for post-treatment periods.

3.2 Intensive and extensive margins

- Figure 5 shows the effect of app use on monthly discretionary spend (top row) and net-inflows into savings accounts (bottom row) disaggregated into the effect on the extensive (left column) and intensive (right column) margins. For discretionary spend, the extensive margin is the number of discretionary transactions per month, and the intensive margin is the average value of a discretionary spend transaction. For net-inflows into savings accounts, the extensive margin is the probability that net-inflows are positive in a given month, and the intensive margin is the value of net-inflows if net-inflows are positive.
- We can see that the reduction of discretionary spend seen in Figure 4 is driven by changes on the extensive margin: users make about five fewer discretionary purchases once they start using the app, while the average amount of each purchase stays unchanged (and actually increases slightly over time). The mean value of a discretionary spend transaction in our data is about £25, so that five transactions account for the effect shown in the main results.
- As in the aggregated effects above, app use has no effect on savings behaviour.

Figure 5: Intensive and extensive margins



Notes: The effect of app use on monthly discretionary spend (top row) and net-inflows into savings accounts (bottom row) disaggregated into the effect on the extensive (left column) and intensive (right column) margins. For discretionary spend, the extensive margin is the number of discretionary transactions per month, and the intensive margin is the average value of a discretionary spend transaction. For net-inflows into savings accounts, the extensive margin is the probability that net-inflows are positive in a given month, and the intensive margin is the value of net-inflows if net-inflows are positive. Point estimates represent group-time average treatment effects aggregated to periods since treatment exposure, as defined in Section 2.2. Red lines represent point estimates and uniform 95% confidence bands for pre-treatment periods allowing for clustering at the user level. If the null hypothesis that parallel trends hold in all periods is correct, these should be equal to zero. Blue lines provide similar information for post-treatment periods.

4 Discussion

Limitations:

- Can't say whether increase in savings was achieved by going into debt elsewhere
- Limitations: We have more data for users that signed up later. So average user in the study is not the average MDB user. If time of signup is mainly driven by financial savyness, then study sample is closer to overall population than MDB sample (if we rank groups as early joiners > late joiners > never joiners in terms of financial sophistication). If, however, signup reflects something like openness to newness, then it's not necessarily correlated with financial savyness. Either way, we might ignore it for now. We could test whether behaviour differs between early or late adopters, but that doesn't seem important enough.

References

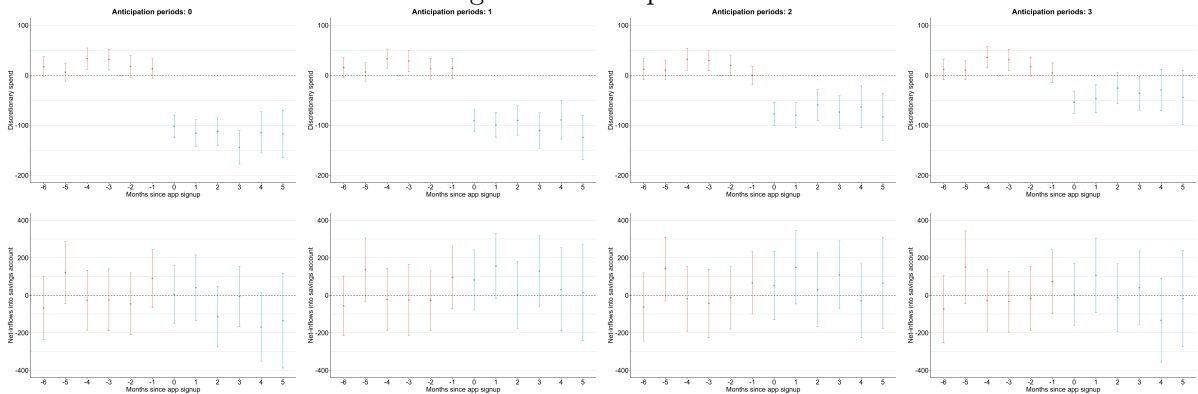
- Baker, Andrew C, David F Larcker, and Charles CY Wang (2022). “How much should we trust staggered difference-in-differences estimates?” In: *Journal of Financial Economics* 144.2, pp. 370–395.
- Bourquin, Pascale, Isaac Delestre, Robert Joyce, Imran Rasul, and Tom Walters (2020). “The effects of coronavirus on household finances and financial distress”. In: *IFS Briefing Note BN298*.
- Callaway, Brantly and Pedro HC Sant’Anna (2021). “Difference-in-differences with multiple time periods”. In: *Journal of Econometrics* 225.2, pp. 200–230.
- Carlin, Bruce, Arna Olafsson, and Michaela Pagel (2019). “Generational Differences in Managing Personal Finances”. In: *AEA Papers and Proceedings*. Vol. 109, pp. 54–59.
- Gelman, Michael, Shachar Kariv, Matthew D Shapiro, Dan Silverman, and Steven Tadelis (2014). “Harnessing naturally occurring data to measure the response of spending to income”. In: *Science* 345.6193, pp. 212–215.
- Hacioglu, Sinem, Diego Känzig, and Paolo Surico (2020). “The Distributional Impact of the Pandemic”. In:
- Muggleton, Naomi K, Edika G Quispe-Torreblanca, David Leake, John Gathergood, and Neil Stewart (2020). “Evidence from mass-transactional data that chaotic spending behaviour precedes consumer financial distress”. Tech. rep. DOI: [10.31234/osf.io/qabgm](https://doi.org/10.31234/osf.io/qabgm). URL: psyarxiv.com/qabgm.
- Roth, Jonathan, Pedro HC Sant’Anna, Alyssa Bilinski, and John Poe (2022). “What’s Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature”. In: *arXiv preprint arXiv:2201.01194*.
- VanderWeele, Tyler J (2019). “Principles of confounder selection”. In: *European journal of epidemiology* 34.3, pp. 211–219.

A Robustness checks

A.1 Relaxing anticipation assumption

- In our setting where users self-select into signing app to the app, it is possible that our results are influenced by anticipation effects.
- In particular, it is possible that users started to cut back on discretionary spend a few months before signing up to the app. In this case, our estimates in Figure 4 would underestimate
- Three reasonable scenarios: users just deciding to sign up for reasons unrelated to trajectory of d spend, decidign to cut back a few months earlier and wanting additional help, looking for tool to help them curb increase in d spend.
- Raw data is consistent with third story.
- Anticipation doesn't seem to be an issue.
- Raw data also explains why increase of anticipation window reduces effect: spend increased month by month. Anticipation moves reference period back. The further back reference periods, the closer d spend is to post signup.
- As Callaway and Sant'Anna (2021) point out in Remark 1, the parallel trend assumption becomes stronger as we increase delta, since parallel trends are now required to hold also in periods prior to actual treatment.
- In our setup, this is not the case, due to self-selection: pre-treatment periods differ for treated and untreated units because treated units tend to experience higher d spend before signup, which might be what causes them to sign up.

Figure 6: Anticipation ...



Notes: ...

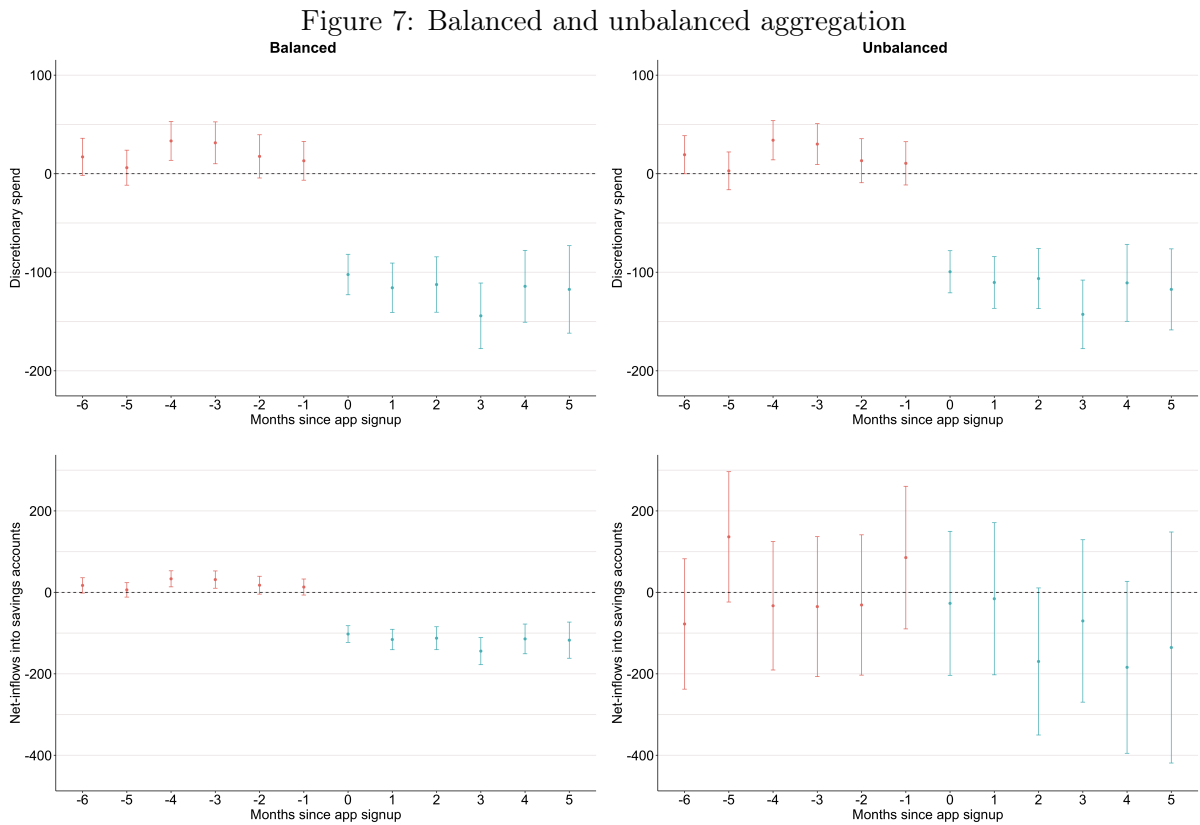
A.2 Unbalanced aggregation

The baseline specification relies on a panel balanced in event time and thus only includes groups that have been exposed to treatment for at least 5 periods. Figure 7 reproduces the baseline

results in the left panel and compares it with results based on the full sample.

As discussed in Callaway and Sant’Anna (2021) (section 3.1.1), these two approaches entail a trade-off. When using the full sample, the aggregated parameters are a function of the weighted average treatment effects for each group e periods after treatment (which is what we want) as well as compositional changes due to different groups being included for different periods e and different weights attached to these groups. While parameters aggregated using a panel balanced in event time do not suffer from compositional and weighting changes, but are calculated based on a smaller number of groups.

As expected, using the full data reduces the size of the confidence intervals. But the results are otherwise very similar.

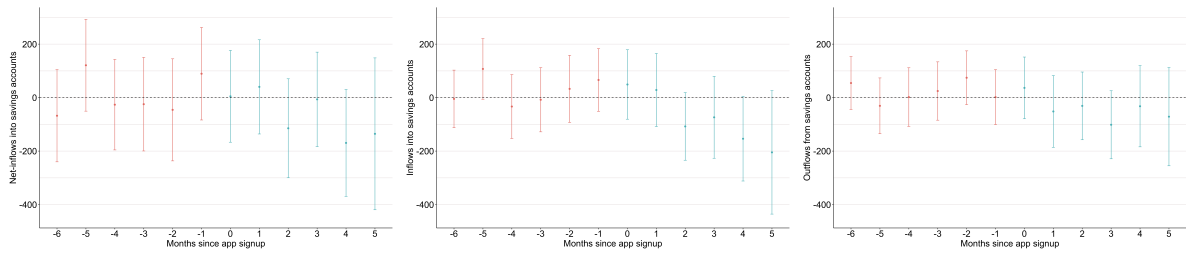


Notes: ...

A.3 Inflows and outflows

- Netflows are unchanged because effects on inflows and outflows closely mirror each other.

Figure 8: Inflows and outflows



Notes: ...