

Evaluation*

Fabian Gunzinger

Neil Stewart

May 30, 2022

Contents

1	Introduction	2
2	Methods	2
2.1	Dataset	2
2.2	Data preprocessing	2
2.3	Summary statistics	3
2.4	Treatment	3
2.5	Outcomes	5
2.6	Covariates	6
2.7	Estimation	6
2.8	Code access	10
3	Results	10
3.1	Subgroups	10
4	Discussion	11
A	Money Dashboard application	13
B	Variable construction	13
C	Alternative specifications	13
C.1	Static TWFE	13
C.2	Dynamic TWFE	13
C.3	Alternative window lengths	15
C.4	Alternative matching method	15
C.5	Post treatment periods as control	15
C.6	Synthetic controls	15
C.7	Classic two-way fixed-effects model	15

*Fabian Gunzinger, Warwick Business School, fabian.gunzinger@warwick.ac.uk; Neil Stewart, Warwick Business School, neil.stewart@wbs.ac.uk.

Abstract

Neat and succinct abstract right here...

1 Introduction

2 Methods

While we were unable to pre-register the analysis because we have had access to and been working with the Money Dashboard data for months, we proceeded in the same spirit: we first wrote a draft of the paper in the form of a pre-analysis plan, following Olken (2015), then tested the entire code base – data pre-processing, balance checks, main analysis, and extensions – with a 1 percent sample, and finally ran the entire analysis.

2.1 Dataset

- Money Dashboard can access up to three years of historic data for each account a user links to their account.
- Each user for whom we have sufficient data thus serves as both a treatment unit and a potential control unit.
- 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.

2.2 Data preprocessing

Sample selection To assess the impact of app use on users' financial behaviour we need to observe their relevant financial history for a sufficiently long period of time prior to and after signup. For our purpose here, "relevant financial history" includes the complete set of spending transactions and all savings account inflows and outflows, and "sufficiently long period" is a period of 6 months prior to and after signup, with the month of signup being the first month of the latter period.¹

Table 1 provides an overview of the precise conditions we applied to implement these criteria and their effect on the sample size. The set of functions that implement each condition can be found on [path to github](#).

¹In Appendix C.4 we show results with different window lengths. **The results are unchanged.**

Table 1: Sample selection

	Users	User-months	Txns	Txns (m£)
Raw sample	136,390	4,003,444	333,024,789	62,853
Annual income of at least £5,000	46,586	1,280,760	121,994,097	22,263
At least one savings account	27,485	826,675	83,429,511	16,240
At least 12 months of data	24,036	806,604	82,016,111	16,008
At least 6 months of pre-signup data	15,260	423,614	44,198,505	8,991
Monthly spend of at least £200	8,080	222,482	25,729,961	5,329
At least 10 txns each month	7,565	206,958	24,295,873	4,926
Complete demographic information	5,995	169,892	20,078,357	3,967
Working age	5,884	166,249	19,735,591	3,817
Final sample	5,884	166,249	19,735,591	3,817

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: [🔗](#).

Data transformations To minimise the influence of outliers, we winsorise spend, income, and savings accounts flow variables at the 5 percent level or – if we winsorise on both ends of the distribution – at the 2.5 percent level.

Question: how to determine winsor level? Currently using 5 percent because 1 percent still leaves very large values in the data: 20k spend, 15k income, 20k sa inflows/outflows, all per user-month

2.3 Summary statistics

Figure 1 describes the sample.

Table 2 provides summary statistics.

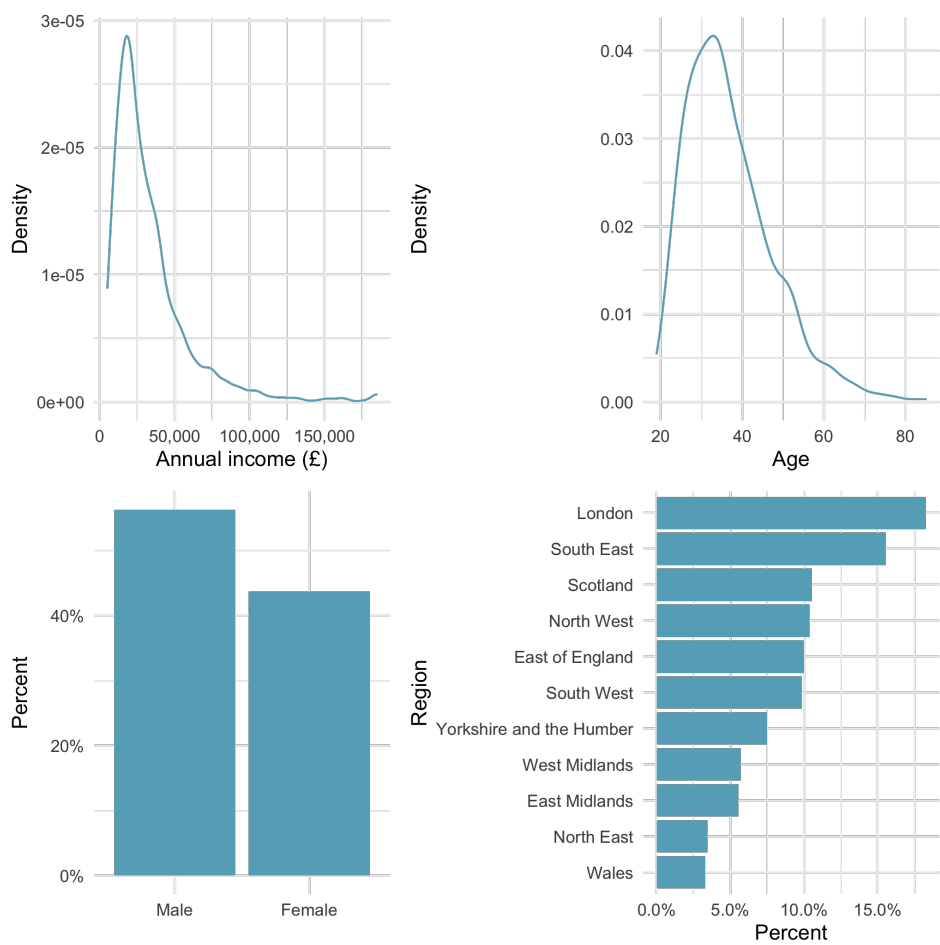
Table 2: Summary statistics

Statistic	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
user_id	478,986.800	59,525.180	338,943	428,938	478,224	527,472	589,752
ym	584.283	11.403	547	576	585	593	607
ymn	201,830.300	98.569	201,508	201,801	201,810	201,906	202,008
txns_count	114.518	57.550	10	72	104	147	247
txns_volume	17,061.590	17,019.070	202.950	6,072.015	10,698.600	20,419.510	68,686.270
month_income	2,911.936	1,933.084	417.155	1,526.060	2,310.784	3,732.883	8,001.687
inflows	512.565	1,028.315	0.000	0.000	0.000	452.425	4,000.000
outflows	539.061	1,123.001	0.000	0.000	0.000	450.000	4,319.829
netflows	-23.267	1,203.212	-4,532.847	0.000	0.000	70.000	3,996.807
netflows_norm	-0.009	0.442	-1.651	0.000	0.000	0.028	1.416
inflows_norm	0.205	0.400	0.000	0.000	0.000	0.192	1.527
outflows_norm	0.216	0.435	0.000	0.000	0.000	0.193	1.654
has_pos_netflows	0.298	0.458	0	0	0	1	1
pos_netflows	250.086	748.606	0.000	0.000	0.000	70.000	3,996.807
t	0.428	0.495	0	0	0	1	1
tt	-2.190	12.977	-37	-10	-2	6	54
month_spend	2,742.312	2,067.423	200.000	1,263.615	2,110.460	3,526.060	8,435.741
age	37.286	9.970	18	29	36	44	65
is_female	0.422	0.494	0	0	0	1	1
is_urban	0.845	0.362	0	1	1	1	1
has_sa_account	1.000	0.000	1	1	1	1	1
generation_code	2.533	0.661	1	2	3	3	4
new_loan	0.032	0.175	0	0	0	0	1
prop_credit	0.098	0.200	0.000	0.000	0.000	0.080	1.000
disc_spend	0.364	0.290	0.000	0.162	0.284	0.472	1.156

2.4 Treatment

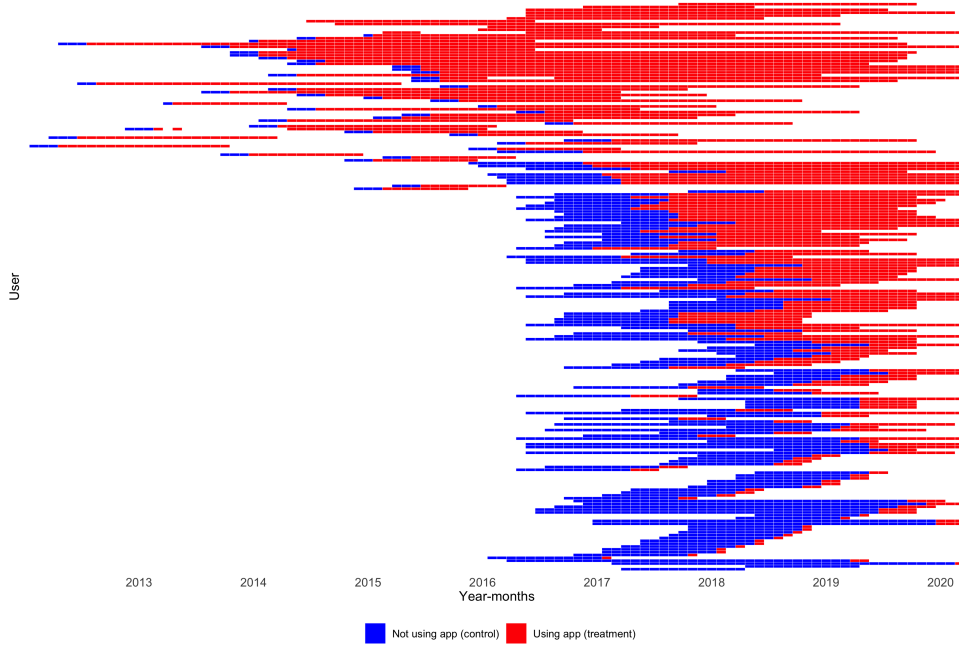
A user changes treatment status from untreated to treated when they start using the app. Figure 2 shows the treatment history for 200 randomly selected users.

Figure 1: Sample description



Notes:

Figure 2: Treatment assignment plot



Notes: Each horizontal line shows for one of 200 randomly selected users the observed pre and post signup periods in blue and red, respectively. 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.

2.5 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)

Adjusting for multiple hypothesis tests We think of our secondary outcomes as exploratory and do not make any adjustments for multiple hypothesis testing.² An alternative approach, based on Anderson (2008), would be to group outcomes into “savings”, “spending”, “borrowing”, and “fees”, and consider them as different dimensions of a latent variable of interest which we might call “financial management skills”. We do not do that for two reasons: first and foremost, because we think it is natural to think of the amount saved as the ultimate outcome and of other outcomes as providing a more nuanced understanding of savings behaviour or as suggesting possible channels through which app use affects savings. Thinking of savings as the main goal is also reflected in Money Dashboard’s main promise, which is to help users spend less and save more, as shown in Figure 7. Second, as pointed out in Carlin et al. (2017), incurring overdraft fees is not an unambiguous sign of a financial mistake, as the opportunity to go into overdraft confers a benefit to the consumer.³

2.6 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.⁴

2.7 Estimation

Control group design:

- We only have data for a self-selected group of people who choose to use the app. This has a couple consequences:
- By virtue of signing up to an app that helps them manage their money, these users are different from those who don’t sign up. As a result, we are unable to answer the question of whether app use helps the average person in the population as a whole save more.⁵

²For a recent game-theoretically motivated discussion of when and how to correct for multiple hypothesis testing, see Viviano et al. (2021).

³For further discussions on fees, see Jørring (2020) and Stango and Zinman (2009).

⁴VanderWeele (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.

⁵One way to get closer to that answer is to re-weight our sample on observable demographic variables so as to match the UK population as a whole. But our sample differs from the population as a whole both in ways that are observable (demographic variables) and unobservable (self-awareness that they need help managing their money, cognitive resources to engage with the app, motivation to do so). Re-weighting would only help us deal with the

- Even among people who do eventually sign up to the app, the decision when to do so is unlikely to be random – *something* makes them sign up at the particular point in time they do and not before or after. If we think of this factor as “motivation to save more”, then said motivation is inextricably linked with the decision to sign up so that we cannot differentiate between the two.
- Hence: due to the first point above, we cannot estimate an ATE (effect of app use on the average citizen), and due to the second point we also cannot estimate a pure ATT (effect of app use on users). Instead, our estimated effect of app use captures the effect of being motivated to save more and using the app to do so.⁶
- This is true for both our matched DiD and our TWFE design. While these two approaches use a different counterfactual to estimate the effect of app use (behaviour of a matched control in the case of DiD and extrapolating within-user pre-signup behaviour in the case of TWFE), neither can help us with the fact that the decision to sign up is likely correlated with the time-varying unobserved effect “motivation to save more”.

DiD:

- We use a difference-in-differences design to estimate the effect of app use. Because we do not have a separate control group, we use the per-signup data of Money Dashboard users as control periods and use matching to find comparable control user for each treatment user.
- To do this, we use the matching estimator for panel data proposed by Imai, Kim, and Wang (2021). Following paper, we conduct the following steps:
- For each treated observation, we find a set of control observations with that share the same treatment history for a period of L periods before the treatment and F periods after the treatment. In our baseline specification, we rely on a year’s worth of data around the treatment period and set $L = 6$ and $F = 0, 1, 2, 3, 4, 5$.
- Identification assumption is that potential outcomes only depend on treatment status of the past L periods. In general, this means that if treatment has a cumulative effect over time, the full effect is reached after L periods. In our context, this means that any effect on savings behaviour from using the app is fully realised after L periods. (I think this means that if we look at the treatment effect for F periods forward, the effect should not become stronger after $F = L$).

DiD identification assumptions:

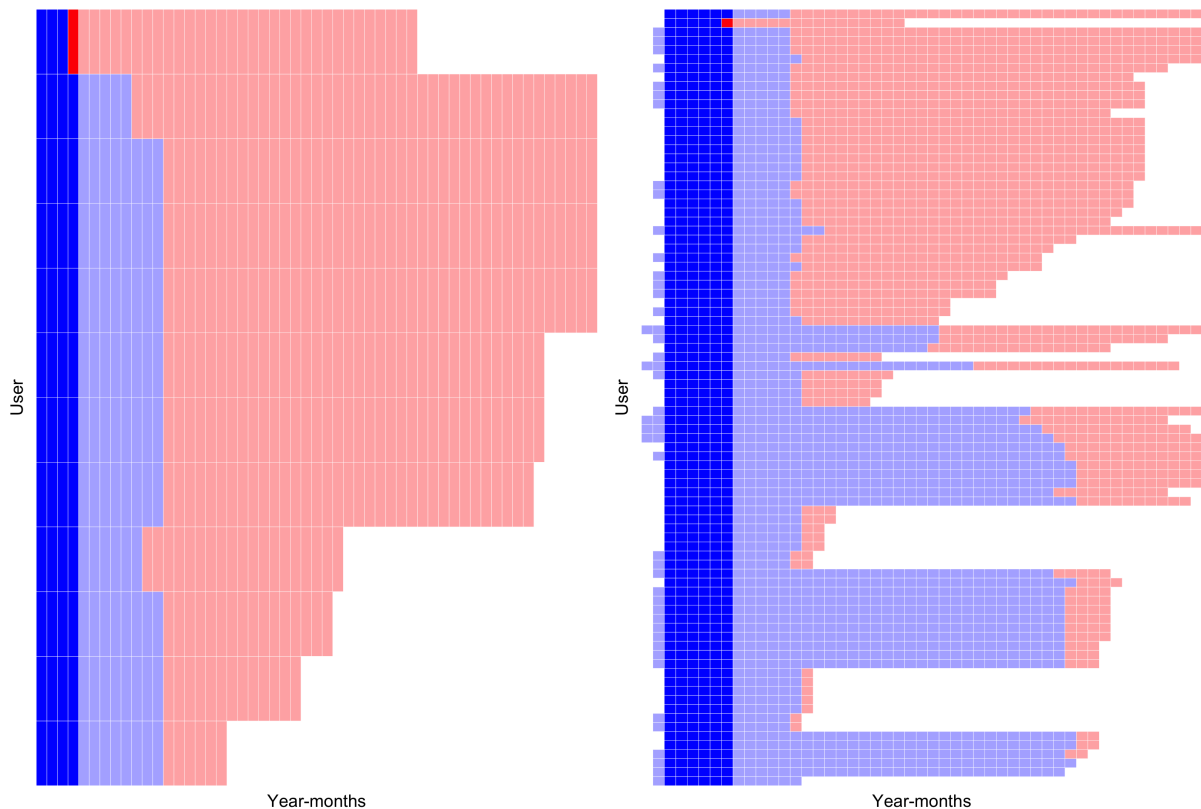
- No spillover effects: the potential outcome of unit i at time t is independent of the treatment status of other units. This is violated if a user’s partner or friends also use the app and, through sharing their experiences or motivations, influence the user’s savings behaviour.

first of these.

⁶One way to get a step closer to ATT would be to find a variable that correlates with “motivation to save” and control for it / match on it.

- Carryover effects no longer than L periods: a user’s potential outcome in period t is independent of treatment status in periods more than L periods ago. Given that we are dealing with an absorbing treatment, this is not a very strong assumption in our context, and we choose L based on what we think is an informative number of periods to observe pre-app use behaviour.⁷.
- Parallel trends: the spending trajectory between treated and control units would have continued to be parallel if the treated unit hadn’t started using the app. This is violated whenever an intended change in savings behaviour also provided the impetus for the user to start using the app, which is likely to occur frequently. To the extent this is the case, we have an omitted variable “motivation to save more”, which both changes the user’s savings behaviour and their treatment status. Because of this, what we are measuring is not a pure ATT of app use – the effect of app use on savings over and above the change precipitated by a change motivation – but the effect of app use for users motivated to save more.

Figure 3: Match set examples

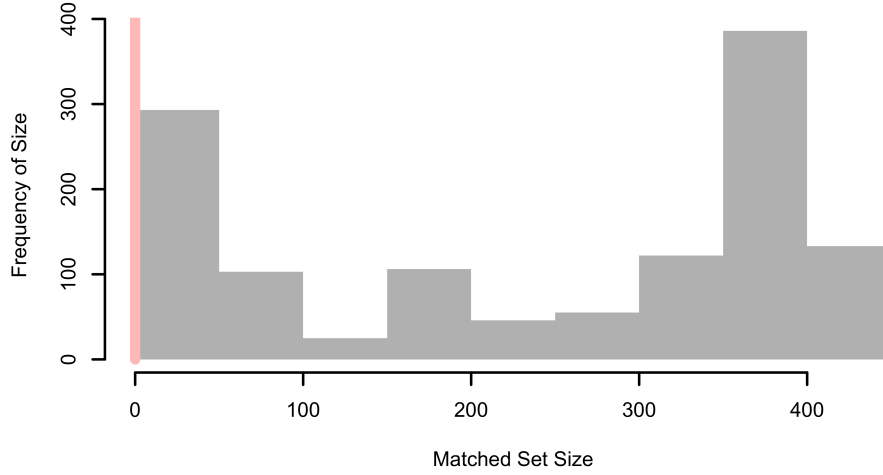


Notes:

Is estimate causal?

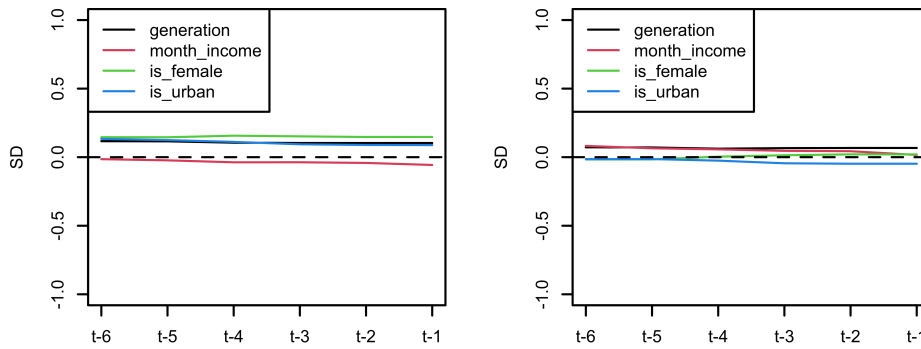
⁷An absorbing treatment is one that cannot be reversed, and hence we only change from untreated to treated once.

Figure 4: Distribution of size of matched control units



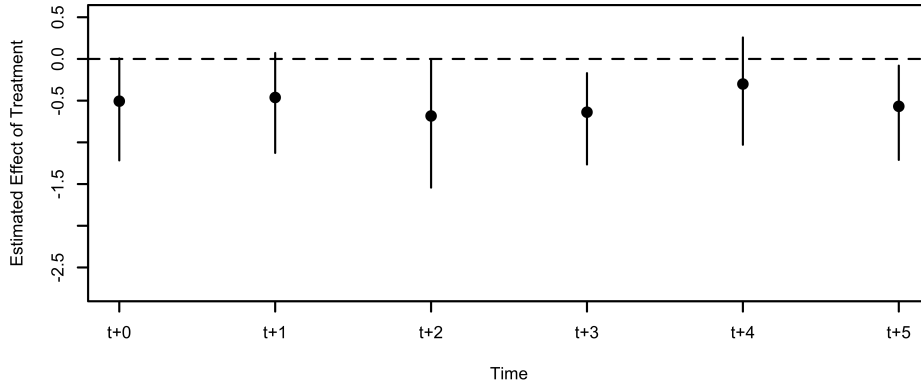
Notes: In the first step of the matching procedure, each user gets assigned a set of potential control users that share the same treatment history for a specified number of periods before the user signs up to the app (6 months, in our baseline specification), but that do not sign up themselves for a specified number of periods after the treatment user has signed up (another 6 months, in our baseline specification). The figure shows the distribution of the sizes of these sets of potential control users. The pink vertical bar on the left shows the count of users for whom no control users could be found.

Figure 5: Covariance balance



Notes: Average covariate standard deviation between treatment and control units for each pre-treatment period using the entire set of potential controls on the left and, on the right, the refined set of controls, which, in our baseline specification, consists of the nearest neighbour match based on the propensity score.


Figure 6: Matching estimates



Notes:

- King and Zeng (2006) show that there are four sources of bias (omitted variable, post-treatment, interpolation, extrapolation).
- Discuss each in turn to argue that effect is causal (for our population of interest, which are people signing up to MDB).

2.8 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 Subgroups

To analyse which groups benefit most from adopting Money Dashboard, we split our sample by gender, generation, income quartiles, and pre-adoption savings behaviour.

We define generations as follows: boomers were born between 1946 and 1964, Gen X between 1965 and 1980, Millennials between 1981 and 1996, and Gen Z after 1997.⁸

Subgroup analysis: same Fig an Tab as in main analysis, but with line for each subgroup. One figure for each of: gender, generations, income terciles, per-adoption average savings tercile (inspired by Carlin et al. (2017), see Fig 5 and Table 4).

See also section 6 in Gargano and Rossi (2021)

⁸Based on age ranges provides by [Beresford Research](#).

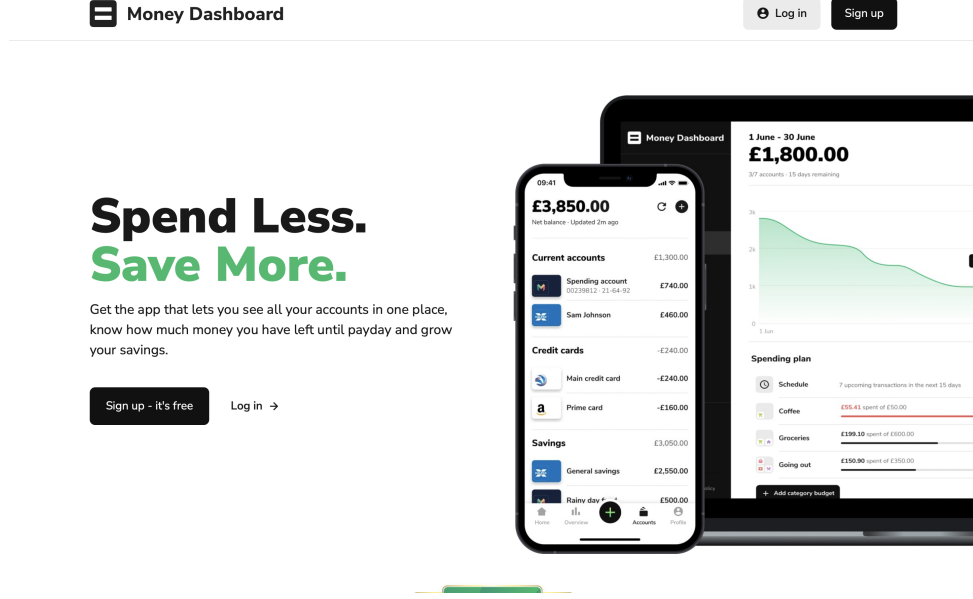
4 Discussion

References

- Abadie, Alberto and Jérémy L’Hour (2021). “A penalized synthetic control estimator for disaggregated data”. In: *Journal of the American Statistical Association* 116.536, pp. 1817–1834.
- Anderson, Michael L (2008). “Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects”. In: *Journal of the American statistical Association* 103.484, pp. 1481–1495.
- Carlin, Bruce, Arna Olafsson, and Michaela Pagel (2017). “Fintech adoption across generations: Financial fitness in the information age”. Tech. rep. National Bureau of Economic Research.
- Gargano, Antonio and Alberto G Rossi (2021). “Goal Setting and Saving in the FinTech Era”. In:
- Imai, Kosuke and In Song Kim (2021). “On the use of two-way fixed effects regression models for causal inference with panel data”. In: *Political Analysis* 29.3, pp. 405–415.
- Imai, Kosuke, In Song Kim, and Erik H Wang (2021). “Matching Methods for Causal Inference with Time-Series Cross-Sectional Data”. In: *American Journal of Political Science*.
- Jørring, Adam (2020). “Financial sophistication and consumer spending”. Tech. rep. Working Paper.
- King, Gary and Langche Zeng (2006). “The dangers of extreme counterfactuals”. In: *Political analysis* 14.2, pp. 131–159.
- Olken, Benjamin A (2015). “Promises and perils of pre-analysis plans”. In: *Journal of Economic Perspectives* 29.3, pp. 61–80.
- Stango, Victor and Jonathan Zinman (2009). “What do consumers really pay on their checking and credit card accounts? Explicit, implicit, and avoidable costs”. In: *American Economic Review* 99.2, pp. 424–29.
- Sun, Liyang and Sarah Abraham (2021). “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects”. In: *Journal of Econometrics* 225.2, pp. 175–199.
- VanderWeele, Tyler J (2019). “Principles of confounder selection”. In: *European journal of epidemiology* 34.3, pp. 211–219.
- Viviano, Davide, Kaspar Wuthrich, and Paul Niehaus (2021). “(When) should you adjust inferences for multiple hypothesis testing?” In: *arXiv preprint arXiv:2104.13367*.

A Money Dashboard application

Figure 7: Money Dashboard website screenshot



Notes: Screenshot from the top of the Money Dashboard website, at moneydashboard.com, accessed on 29 April 2022.

B Variable construction

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](https://github.com).

C Alternative specifications

C.1 Static TWFE

$$y_{it} = \alpha_i + \lambda_t + \beta T_{it} + \gamma X_{it} + \epsilon_{it} \quad (1)$$

- Comparison: pre vs post signup within each individual.
- Assumption: there are no time-varying unobserved effects that affect both y and T.
- Discussion: there is something that made the individual sign up in the first place, and it might well be an individual level shock that we don't observe (unexpected large expense, loss of job, etc.)
- See Imai and Kim (2021) for problems with twfe

C.2 Dynamic TWFE

$$y_{it} = \alpha_i + \lambda_t + \sum_{s=-6}^5 \beta_s T_{its} + \gamma X_{it} + \epsilon_{it} \quad (2)$$

Table 3: Variable construction

Variable (name in dataset)	Definition	Rationale
Primary outcome		
Net savings (<i>sa_netflows_norm</i>)	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.
Secondary outcomes		
Positive net savings dummy (<i>has_pos_sa_netflows</i>)	Dummy equal to 1 if there were positive net savings (as defined above).	Captures extensive margin of savings (change in number of deposits).
Positive net savings (<i>pos_sa_netflows</i>)	Equal to net savings if there were positive net savings.	Captures intensive margin of savings (change in deposit amount).
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.

- See Sun and Abraham (2021) for problems with that and compare to their approach as implemented in fixest.

C.3 Alternative window lengths

Show results for 12 months on either end.

C.4 Alternative matching method

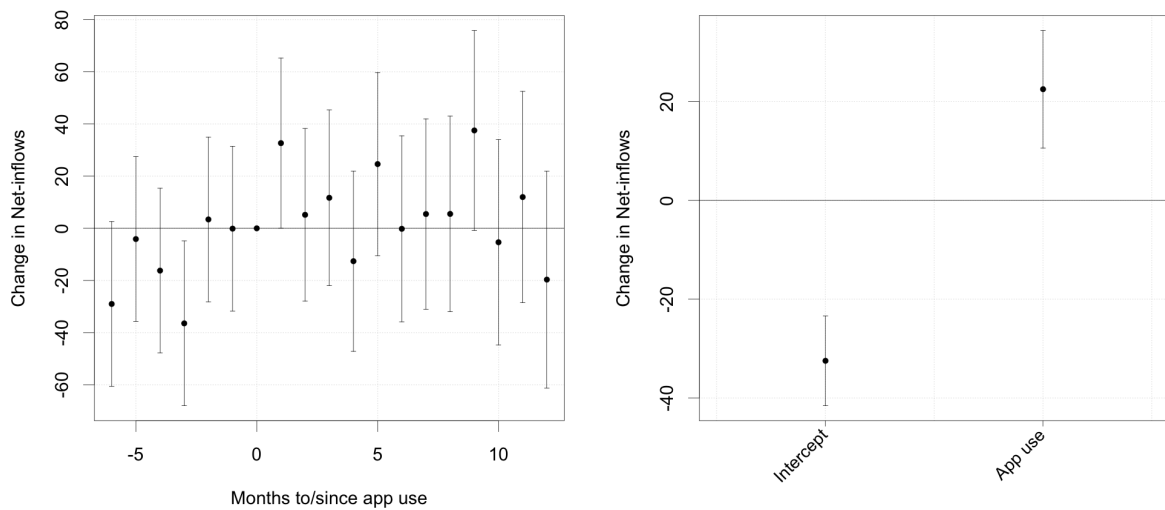
C.5 Post treatment periods as control

C.6 Synthetic controls

See Abadie and L'Hour (2021) for how to use synthetic controls for disaggregated data.

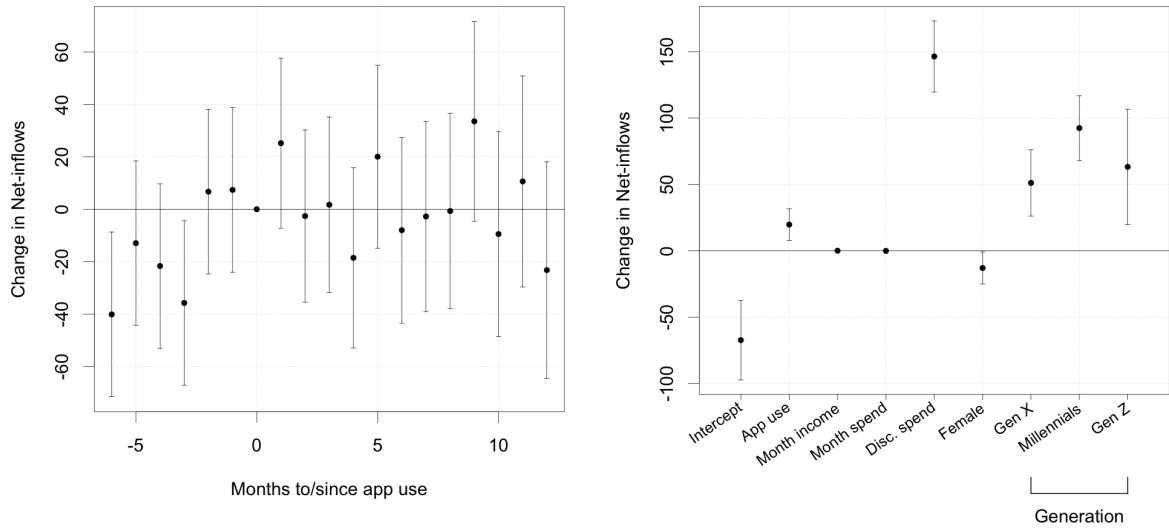
C.7 Classic two-way fixed-effects model

Figure 8: Naive pre-post results



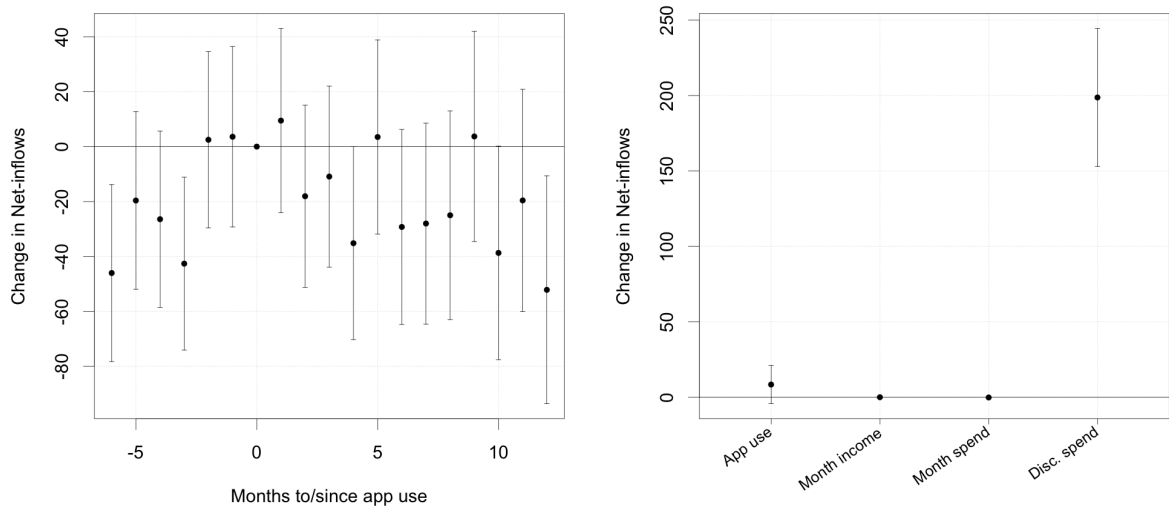
Notes: Notes: ...

Figure 9: Pre-post with controls



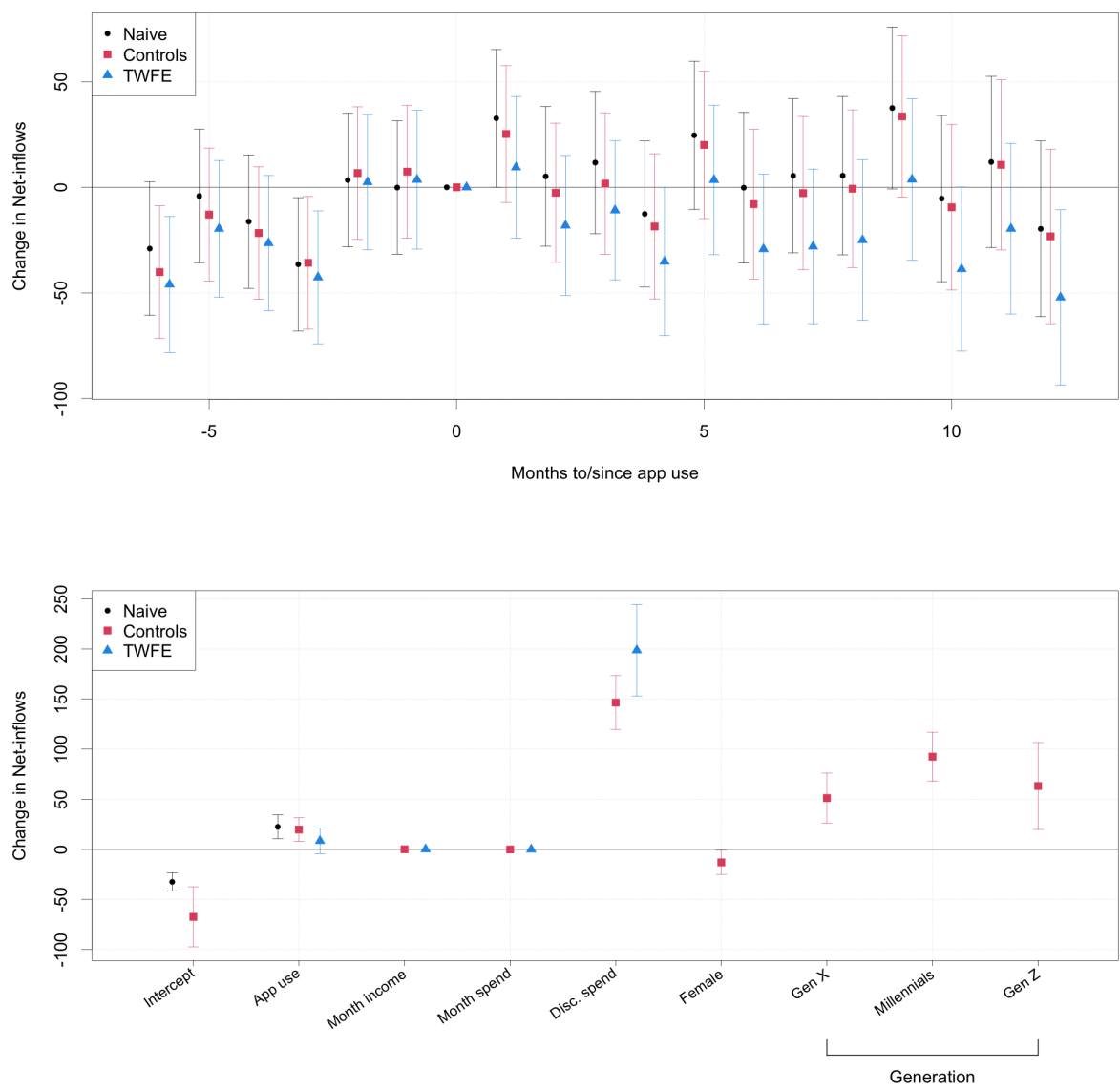
Notes: Notes: ...

Figure 10: TWFE



Notes: Notes: ...

Figure 11: Comparison



Notes: Notes: ...