Does Money Dashboard help its users spend less and save more?

Fabian Gunzinger *
Warwick Business School

September 2, 2022

Contents

*fabian.gunzinger@warwick.ac.uk. I thank Neil Stewart for invaluable support and encouragement, and Redzo Mujcic and Zvi Safra for helpful comments.

Abstract

Neat and succinct abstract right here...

1 Introduction

This paper evaluates whether Money Dashboard, a UK-based financial aggregator app, helps its users reduce their discretionary spend and increase their "rainy-day savings".

The question is important because a large number of adults in the UK and the US do not have enough savings to cover unexpected expenses like car or medical bills: in the UK, 25 percent of adults would be unable to cover an unexpected bill of £300 (Philipps et al. 2021), while in the US, about 30 percent would be unable to cover a $400 bill (Governors of the Federal Reserve System 2022). But while there is a large body of research that studies reasons for why savings are low, little is known about what could help people save more.[1]

FinTech aggregator apps like Money Dashboard are promising because they provide easy access to financial information that make it easier to monitor ones spending and saving, and often also offer tools such as budgeting and the setting of spending goals. Financial information that is more easily accessible and is aggregated in ways that help people keep track of their goals might be beneficial because rational inattention theory predicts that it makes people more likely to access that information, which, in turn, might lead to better consumption decisions. Similarly, tools that help with budgeting and with setting spending goals have the potential to help users make consumption decisions more in line with their intentions because such tools can act as commitment devices that – if users experience disutility from falling short of their goals – introduce a cognitive cost to overspending or undersaving.[2]

The nascent literature that studies the effect of FinTech apps on financial outcomes suggests that these apps can indeed lead to improved financial outcomes: they have been found to reduce spending by providing users with information about their spending relative to peers (D'Acunto et al. 2020) and offering budgeting options (Lukas and Howard 2022), to increase savings by offering budgeting options (Gargano and Rossi 2021), and to reduce non-sufficient fund fees by facilitating access to information (Carlin et al. 2022).

In this paper, I specifically test whether using Money Dashboard is associated with a reduction in discretionary spending and an increase in "rainy-day savings". I use a new estimator proposed by Callaway and Sant'Anna (2021) that corrects for recently identified problems in two-way fixed effects estimates.

[1] Well-documented behavioural biases that help explain undersaving are, among others, present bias (Laibson 1997, Laibson and Marzilli-Ericson 2019), inertia (Madrian and Shea 2001), over-extrapolation (Choi et al. 2009), and limited self-control and willpower (Thaler and Shefrin 1981, Benhabib and Bisin 2005, Fudenberg and Levine 2006, Loewenstein and O'Donoghue 2004, Gul and Pesendorfer 2001). One danger of viewing low savings mainly as a result of behavioural biases is that while these biases likely do play some role and designing environments and tools to help correct them are thus part of the solution, it is at least conceivable that this is an area where the focus on behaviour-level solutions distracts from an effort to find more effective society-level solutions, a danger inherent in behavioural science research convincingly highlighted in Chater and Loewenstein (2022): if the main problem is that many people are unable to earn enough to save, then the effectiveness of helping them manage their low incomes more effectively pales in comparison with efforts to help them earn more.

[2] On rational inattention theory, see, for instance, Brunnermeier and Nagel (2008), DellaVigna (2009), and Sims (2003). On commitment devices see, among others, Thaler and Shefrin (1981), Laibson (1997), and O'Donoghue and Rabin (1999) for theoretical foundations, and Beshears, Milkman, et al. (2016) and Hsiaw (2013) for a discussion of soft commitment devices.

I find...

There are two main limitations to the approach. First, the data is not generated by a randomised experiment. The gold-standard to evaluate whether use of Money Dashboard improves financial outcomes would be a randomised controlled-trial, where out of a sample of potential users (ideally random and representative of the UK population), we would randomly grant access to the app to some users and then compare outcomes of those treated users with the control group of users who did not have access. Instead, the data I have access to only contains data for individuals who self-selected into using the app. Individuals will choose to do so for a number of different reasons, all of which are unobserved in the data, and at least some of which would probably have changed their financial outcomes even if they had not signed up to the app. Any changes in financial outcomes we observe are thus "aggregate" or "net" effects of these unobservables and the "pure" causal effect of app use.

To see this, think of the net effect as $\textit{net effect} = \textit{causal effect} + \textit{"need"}(\downarrow) + \textit{"motivation"}(\uparrow)$, where the arrows indicate the direction of the bias, and consider three cases that illustrate three stylised but plausible scenarios for signup. First, consider a user who signs up in the hope that the app will help them reign in discretionary spending that has has gotten out of hand. If it takes the user some time to fully adjust their spending, then even if the app does help them make these adjustments, the estimated positive effect of app use will be biased downward. Next, consider a user who decides to start bringing their own lunch to work instead of eating out in an effort to save for a new car and signs up to MDB in the hope that the app will help them keep track of their spending. Such a user would probably have reduced their discretionary spend even if they had not signed up to the app, thus creating an upward bias on our estimated net effect. Finally, consider a user who signs up to MDB purely because they happened to see an advert for the app on the Bus and got curious. In this case, we can think of signup being close to random – almost as if the user had been allocated to the treatment group in our ideal experiment – and the estimated net effect will closely resemble the causal effect of app use. Hence, under the weak assumption that at least some users sign up for reasons that are not as good as random, our estimated effects will be biased upwards or downwards depending on the relative proportion of users whose unobservable reasons for signup create an upward and downward bias.

The second limitation is that even if I were able to isolate the effect of the app, I am not able to differentiate between the contributions of different features of the app such as improved access to information and budgeting.

However, despite these limitations, the results tell us... suggest that further research is worthwhile

My work mainly contributes to three strands of the literature. The first, is the aforementioned recent literature that studies the effect of FinTech apps on financial outcomes. The second, is the very recent literature on studying interventions to help increase "rainy-day savings" an area of household finances that has until recently had no attention. In addition to studies testing the effect of FinTech apps on savings, there is a strand of research that studies the use of auto-enrolment into employer-sponsored savings accounts – similar to the ones used to increase pension savings (Thaler and Benartzi 2004, Choi et al. 2004, Choukhmane 2019) – and finds an increase in both participation (relative to opt-in accounts) and account balances (Beshears, Choi, et al. 2020, Berk et al. 2022). Finally, my work also contributes to a rapidly growing literature of

using financial-transaction data from banks or financial aggregator apps to understand consumer financial behaviour. As already mentioned, Kuchler and Pagel (2020) use data from a financial aggregator app to estimate time preferences. Similar data has been used to show that consumer spending varies across the pay cycle (Gelman et al. 2014, Olafsson and Pagel 2018), to test the consumer spending response to exogenous shocks (S. R. Baker 2018, Baugh et al. 2014), and to better understand the generational differences in financial platform usage patterns (Carlin et al. 2019). Some researchers use transaction-data directly provided by banks. Ganong and Noel (2019) show that consumer spending drops sharply after the predictable income drop from exhausting unemployment insurance benefits, Meyer and Pagel (2018) analyse how individuals reinvest realised capital gains and losses, and Muggleton et al. (2020) show that chaotic spending behaviour is a harbinger of financial distress.

The remainder of this paper is organised as follows: Section 2 introduces the dataset used, discusses preprocessing and presents summary statistics; Section 3 introduces the empirical approach used in the analysis; Section 4 presents the results; and Section 5 concludes. To make it easier for interested readers to clarify questions about details and subtleties of data preprocessing and analysis steps, I provide links to the scripts that implement the steps discussed in the text in the relevant places throughout the text.[3]

2 Data

2.1 Dataset description

I use data from Money Dashboard (MDB), a UK-based financial aggregation 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 270,000 users, and provides information such as date, amount, and description of the transaction as well as account and user-level information. Crucially, for this paper, MDB can access up to three years of historic data for each linked account.

The data's main advantage for the study of consumer financial behaviour is that it allows us to observe all savings and spending transactions for users who linked all their financial accounts. This means that for such users, we can be sure that any reduction in spending we observe is not offset by an increase in spending in unliked accounts. Furthermore, data is collected automatically and in real-time rather than through surveys that collect data with a time-lag and often rely on consumers's ability and willingness to provide accurate information.

The data's main limitation is that because users's self select into using MDB, the sample is not representative of the wider UK population: it is well documented that FinTech app users are more likely to be male, younder, and higher-income earners than the average person 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. However, because, as discussed in the introduction, the aim of this paper is to assess the effect of MDB on people

[3]The projects GitHub repo that contains all files used to produce the results can be found at https://github.com/fabiangunzinger/mdb_eval.

who choose to use it, the lack of representativeness is not an issue.[4] A second limitation is that while we can observe user's complete financial behaviour if they add all their financial accounts to the app, it is not trivial to distinguish between users who do and do not do that. I address this challenge in the sample selection process documented below. A third issue is that while the app is able to classify many transactions into types, it misclassifies some transactions and cannot classify others altogether. I address this as part of the cleaning process documented below.

2.2 Preprocessing

Data 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.[5] Here, I briefly describe the main cleaning steps and their rationale. I drop all transactions with a missing description string because these cannot be categoriesed, and all transactions that are not automatically categoriesed by the app. Dropping these transactions makes it likely that I will underestimate amounts spent and saved, but minimises the risk of incorrectly classified transactions. I also group transactions into transfer, spend, and income subgroups, following Muggleton et al. (2020) to define spend subgroups and Hacıoğlu-Hoke et al. (2021) to define income subgroups.[6] 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 when a user buys 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.

To minimise the influence of outliers, I winsorise all variables at the 1 percent level or – if we winsorise on both ends of the distribution – at the 0.5 percent level.[7] I rely on winsorisation (replacing top values with percentile values) instead of trimming (replacing top values with missing values) because data inspection suggests that in most cases, very large (absolute) values are not the result of data errors, which would call for trimming, but reflect genuine outcomes, which makes winsorising appropriate because it leaves these observations in the data while lowering their leverage to influence results.

Sample selection: The three main goals of sample selection are to select a sample of users for whom I can be reasonably certain to observe all relevant financial accounts,[8] account histories of at least 12 months, and who are not using MDB for business purposes. Table 1 lists the precise conditions I apply to implement these criteria and their effect on sample size.[9]

In my main analysis, I show effects of app use for the 12-months period from 6 months before and 5 months after MDB signup, treating the month of signup as period 0. To ensure that results

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

[5] The module with all cleaning functions is available on Github.

[6] The precise list used to classify transactions is available on Github.

[7] The code that performs the winsorisation are available on Github.

[8] Relevant accounts include all current, savings, and credit-card accounts, but exclude long-term savings and investment accounts.

[9] The code that implements the selection criteria is available on Github.

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.

are not affected by the number of accounts we observe for an individual, it is thus critical that we can observe at least 12 months of history for all accounts a user adds to the platform. This ensures that in the extreme case where a user adds an account they had used for some time in the fifth month after they signed up, I observe all transactions on that account throughout the 12-month period of interest.[10] Data exploration suggests that all major banks start providing 12 months of historical data from April 2017 onwards, which is why include only users who sign up after that data.[11]

To ensure that I can be reasonably certain to observe users have added all their financial accounts to the app, I 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, I drop users with more than 10 active accounts in any given month.

Finally, I drop test users and users that are not of working age (younger than 18 or older than 65) because these groups might have objectives other than to reduce spending and increase savings.[12] And I retain only users for whom we can observe the complete set of demographic covariates (gender, age, region) to ensure all users can be used throughout the analysis. These steps do impact the sample size significantly.

[10]For example: if a user has made monthly payments of £100 into a savings account for two years by the time they sign up to MDB but links that account five months after joining, we can only observe the historical payments if MDB can access at least 12 months of history. If this is not the case, and MDB can only, say, access 6 months of history, we would erroneously conclude that the user started saving £100 more starting in the month they signed up.

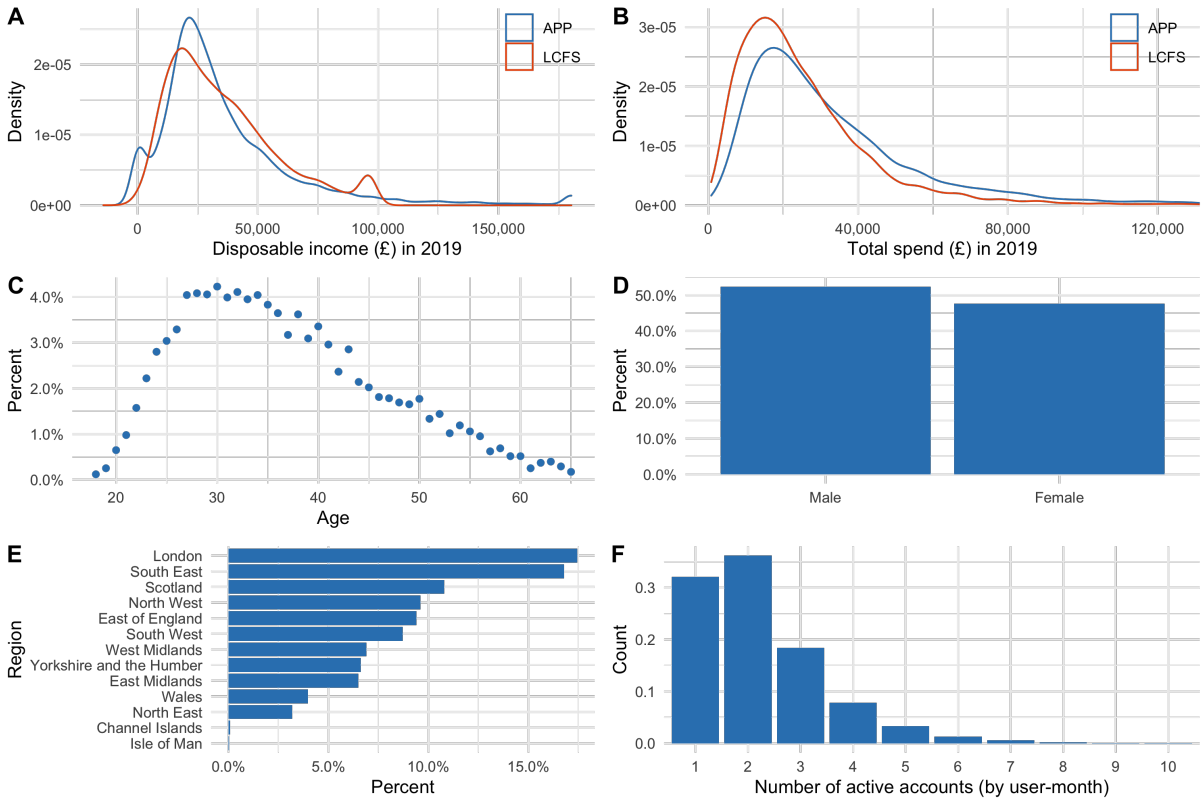
[11]The Jupyter Notebook containing the data exporation is available on Github.

[12]We cannot identify test users precisely, but drop users who signed up prior to or during 2011, the first year the app was in operation.

2.3 Summary statistics

Figure 1 provides a view of some salient sample characteristics. Panels A and B show that while distributions of disposable income and total spend in 2019 broadly mirror that of the ONS Living Cost and Food Survey (LCFS) data, the MDB data tends to slightly underestimate incomes and overestimate spending.[13] Given that our sample is likely biased towards high-earners, as discussed above, we would expect both income and spend to be higher than in the LCFS. Two likely caveats to the data probably create discrepancies relative to the LCFS data. First, as discussed in Section 2.2, I drop unclassified transactions from the data, which biases both income and spending downwards (i.e. spending would probably be even higher compared to the LCFS). Second, it is likely that it is more challenging to automatically classify income transactions than spend transactions, which might create an additional downward bias on the income distribution.[14] Panels C, D, and E show that, as discussed in Section 2.1, the sample is skewed towards users that are younger, male, and live in the South East. Finally, Panel D shows that most users use transact with one or two accounts per month.

Figure 1: Sample characteristics



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.

[13] I accessed the LCFS data via the UK Data Service at the following url: https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8686.

[14] Bourquin et al. (2020) present an alternative algorithm to identify income transactions and find that this leads to higher estimated incomes. I do not use their algorithm because I only use disposable income as a covariate to capture relative income differences between users.

Table 2 provides additional summary statistics. It shows, for instance, that the average number of monthly transactions is slightly above 100 (with a mean of 112 and a median of 101), that half of all user-month incomes lie between £1,407 and £3,596, and that inflows into savings accounts closely mirror outflows. It also suggests that highly discretionary spend accounts for about 30 percent of total monthly spend

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

3 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 papers in econometric research demonstrate that while this approach is extremely common in applied research, two-way fixed effects models do not produce valid estimates of dynamic treatment effects in most settings (Roth et al. 2022). 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 my own, this is a very strong assumption.[15]

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

[15]The problem is exacerbated in static two-way fixed effects models, which require homogeneity in treatment effects across treatment adoption cohorts and in time since treatment. Goodman-Bacon (2021) shows that the static two-way fixed effect DiD estimator is a weighted average of all possible two-units and two-time-periods difference-in-differences in which one unit changes its treatment status and the other does not. Because this includes "forbidden comparisons", where already treated units are used as control units, this can lead to a scenario where the treatment effect of some units has a negative weight, unless one makes the additional assumption of time-invariant treatment effects (in which case, the "forbidden" comparisons have an effect size of zero). These negative weights, in turn, can lead to situations where the treatment effect estimate is negative even though all unit-level treatment effects are positive. The dynamic specification suffers from very similar problems, as shown in Sun and Abraham (2021). For two excellent reviews of this new literature, see Roth et al. (2022) and A. C. Baker et al. (2022).

the causal parameter of interest, and then discussing identification, estimation, and inference, while using the simplified notation used in Roth et al. (2022).

Causal parameter of interest: 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:

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

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.

Identification: 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.[16] 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.[17]

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:

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

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

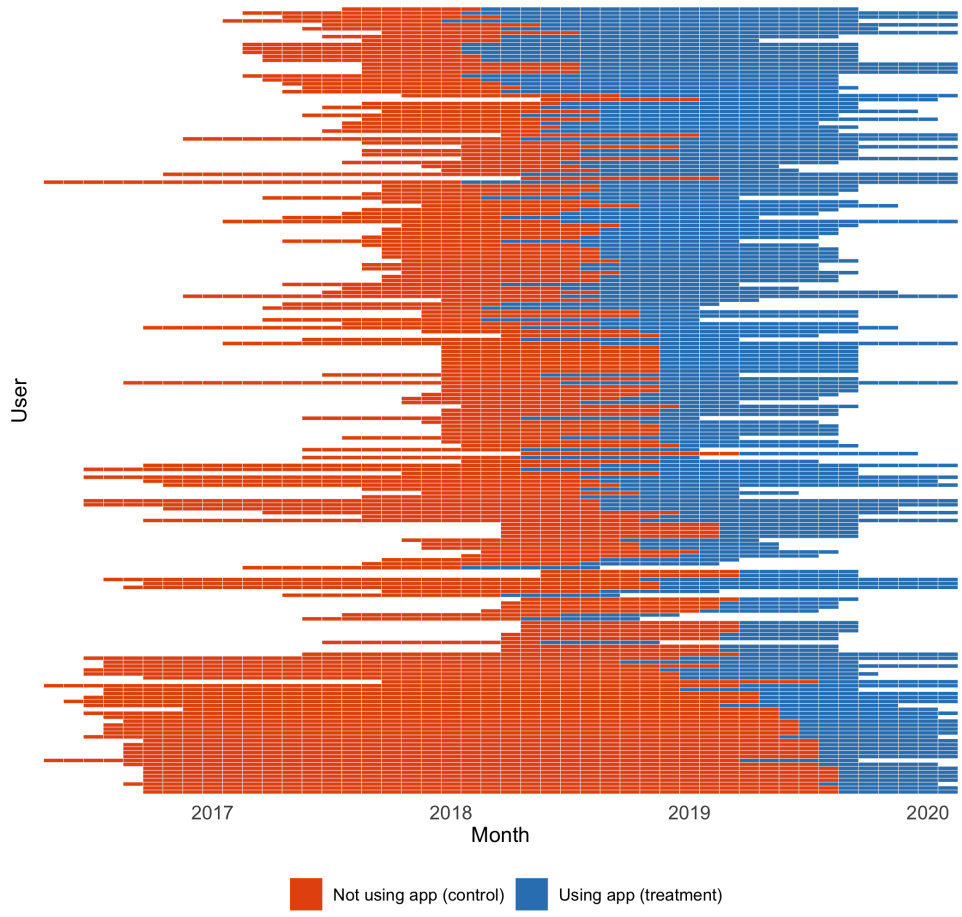
$$ATT(g,t) = \mathbb{E}[Y_{i,t} - Y_{i,g-1}|G_i = g] - \mathbb{E}[Y_{i,t} - Y_{i,g-1}|G_i \in \mathcal{G}]. \tag{3}$$

[16] Additional 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.

[17] Intuitively, 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. I use the first step regression estimator throughout for all results shown in this paper. See Callaway and Sant'Anna (2021) and Roth et al. (2022) for more details.

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.[18] 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.

Figure 2: Pre and post-signup data availability



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.

Figure 2 shows how we can leverage pre-signup data provided in the MDB data to construct valid control groups. Each row of cells shows data for one of 200 randomly selected users, with red cells indicating data from pre-signup months and blue cells indicating data from post-signup months. Signup thus occured sometime during the first blue coloured month, and we can use users who have not yet signed up (whose cells are still red) as a comparison group.

[18] If 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.

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

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

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 calculating

$$ATT_l = \sum_g w_g ATT(g, g+l), \tag{5}$$

where $w_g = P(G = g | g \in \mathcal{G})$ 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.[19]

Inference: Inference is based on a bootstrap proceedure that can account for clustering and produces simultaneous (or uniform) confidence bands that account for multiple hypothesis testing.[20]

4 Results

The main outcome variables in my analysis are "discretionary spend", defined as spend transactions over which users likely have a lot of control, and "net-inflows into savings-accounts", defined as the difference between inflows into and outflows from all of a user's savings accounts.[21]

4.1 Main results

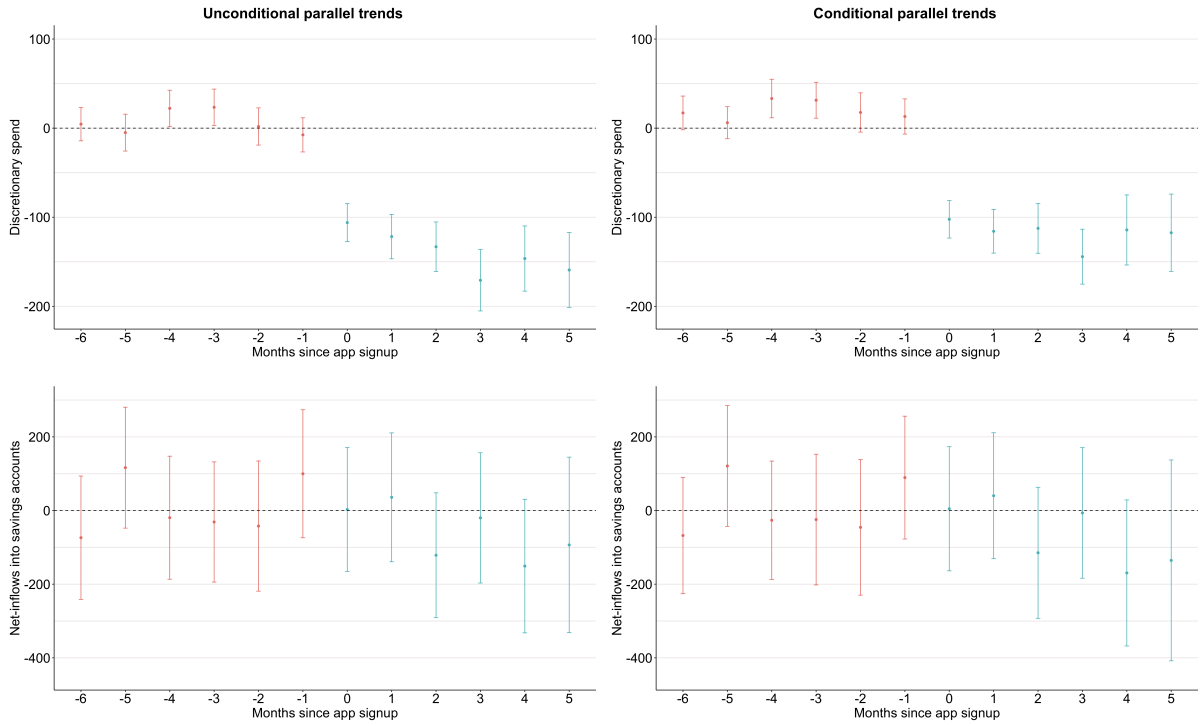
Figure 3 shows the effect of app use on monthly discretionary spend (top row) and monthly net-inflows into savings accounts (bottom row) under the unconditional (left column) and conditional (right column) parallel trends assumptions. As discussed in Section 3, the conditional parallel trends assumption states that the outcome variable of treatment and control units with the same set of covariates would have evolved in parallel fashion. Throughout my analysis, the set of covariates I use are month income, month spend, the number of active accounts, and age.

[19] See Section 3.1.1 in Callaway and Sant'Anna (2021) for a more detailed discussion.

[20] For more details on the inference proceedure, see Section 4.1 in Callaway and Sant'Anna (2021). The interpretation of uniform and pointwise confidence intervals differ in the following way: a 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.

[21] I do not include gasoline expenses in discretionary spend because many people use their car to commute to work so that it is unclear what proportion of these expenses are discretionary. A list of transaction tags used to identify discretionary spend transactions and the code used create the variable are available on Github. To capture only user-generated savings-account flows and not interest payments and similar automated transactions, I include only transaction with an absolute value of £5 or higher. The code used to calculate savings account flows is also available on Github.

Figure 3: 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 3. 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.

Estimates are group-time average treatment effects aggregated by time since treatment exposure, as defined in Equation 5. All results are presented with a uniform 95% confidence band, based on bootstrapped standard errors that account for autocorrelation in the data and are clustered at the user level, as discussed in Section 3.

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. These results are in line with those found in Levi and Benartzi (2020), which find a 11.6 percent reduction in discretionary spend following the use of an aggregator app. Interestingly, we can also see that at least for the first six months of app use, the effect is persistent.

That the results based on the unconditional and the conditional parallel trends assumptions are very similar but not identical is not surprising. 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 treatment and control groups. A classic example of the latter would be comparing individuals wo did and did not sign up for a job-training program, where characteristics such as age and employment history are often quite different (Heckman et al. 1997). In our context, however, where all individuals eventually sign up to Money Dashboard and our comparison group is a set of "not-yet-treated" rather than "never-treated" individuals,

we would not expect covariate values between individuals in the treatment and control groups to vary as much. At the same time, we would expect the results to differ somewhat. If, as is plausible, discretionary spend is a constant fraction of income and total spend, then we would expect parallel trends only for individuals with similar incomes and total spend. Similarly, if we think that discretionary spend increases in the number of accounts we observe per user, then parallel trends in spend only holds for users for whom we observe the same number of accounts.[22]

We can also see that in two periods, the null-hypothesis of identical pre-signup trends is being rejected (both under conditional and unconditional trends). The estimated differences from zero are not large, but a sign that the results should be interpreted with caution. In future work, I plan to use the approach introduced by Rambachan and Roth (2022) to test how sensitive my results are to deviations from parallel trends.

Finally, we can see that, in contrast to discretionary spending, net-inflows into savings accounts do not change once users start using the app. This is somewhat surprising, since we might have expected that users transfer at least some of their saved funds into their savings accounts to either build up a "rainy-day" savings cushion or as a contribution towards specific savings goals. The absence of such transfers naturally begs the question "where does the money go?", to which I turn in the next section.

4.2 Where do additional funds go?

Above, we found that while users reduce their discretionary spend after using MDB by between £100 and £150, they do not transfer additional funds into their savings accounts.

One possible explanation is that users do save more but transfer money either into longer-term savings vehicles such as investment funds or into savings accounts they have not linked to MDB. The first three panels (from the top left) in Figure 4 show flows into investment of pension funds, current account outflows that users manually labelled as "savings", and – as an alternative but cruder measure of transfers into savings accounts – the total amount of transfers from current accounts into other accounts held by the user.[23] None of these change significantly during users' pre and post signup periods, rejecting the idea that users systematically move additional funds into either investment vehicles or other savings accounts. An exception is the spikes in user-labelled savings in the month just before and the month of signup. Rather than suggesting an additional inflow of funds, however, this is is more likely to be an artifact of users only manully labelling transactions during active app use rather than also labelling historical transctions.[24]

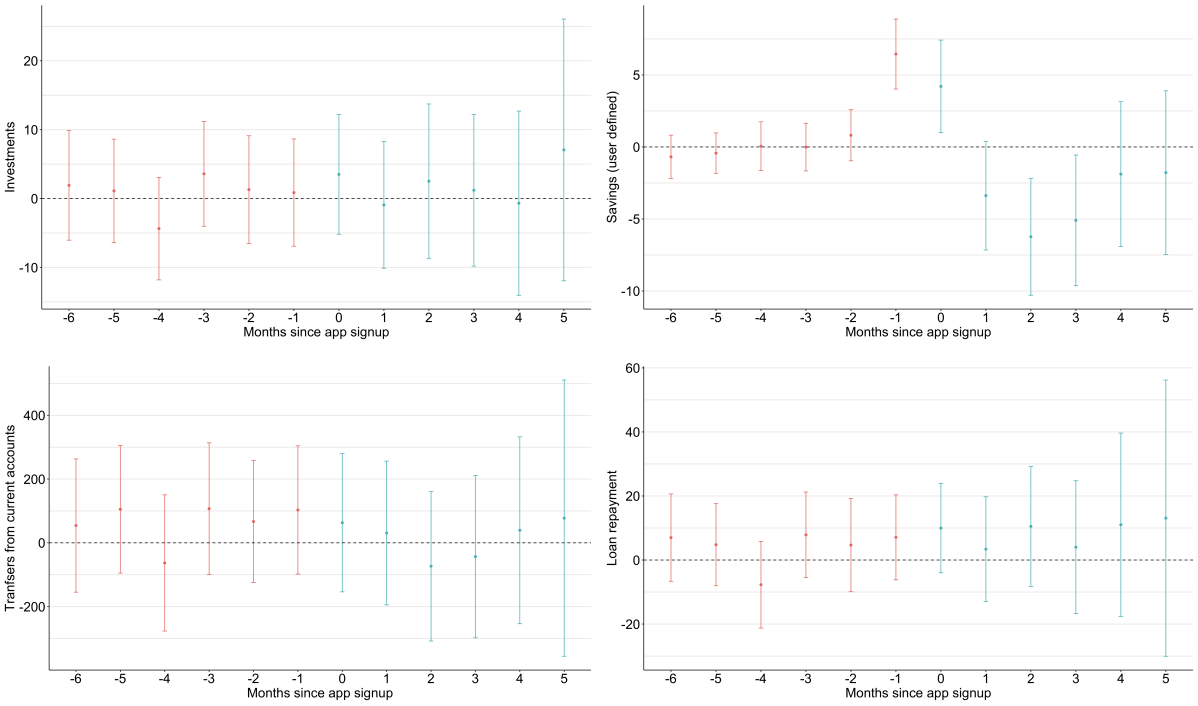
Another possibility is that users use the additional funds to pay down debt. The bottom right panel in Figure 4 shows loan repayments, which do not change significantly during the 12-month period around app signup either.

[22] As discussed in Section 2, users choose which accounts to link to Money Dashboard, and while I select for users that appear to have linked all their accounts and for whom we could observe the complete account history even if they added some of their accounts after joining, I cannot rule out that some users do add account after signup and MDB is unable to capture the complete history.

[23] The code where these variables are defined is available on Github.

[24] The pattern could be explained in two ways: either users can see the entire account historey available to MDB after linking an account but only label transactions during periods of active app use and the period immediately before signup, or MDB only shows users data from the month before signup onwards, so that users are unable to manually label transactions in earlier periods.

Figure 4: Possible destinations of additional savings



Notes: Point estimates represent group-time average treatment effects aggregated to periods since treatment exposure, as defined in Section 3. 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.

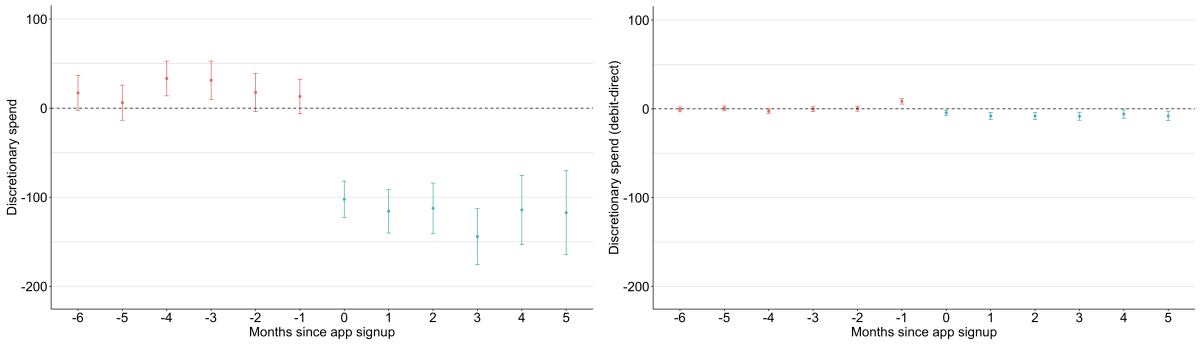
The finding that the saved funds do not flow into any of the plausible channels suggests that they might be channelled towards multiple uses either across or between individuals.

4.3 Disaggregated discretionary spend

Another question of interest based on the main findings above is along what dimensions users reduce their discretionary spend. There are three dimensions of interest. First, users can achieve the persistent reduction in discretionary spend seen in Figure 3 by either maintaining a change in behaviour consistently month-to-month by, for instance, making fewer purchases or spending less on certain items, or, alternatively, make a decision during the month of signup that automates the reductin in spend by cancelling debit-direct contracts. Figure 5 shows in the left panel the reduction of discretionary spending shows in Figure 3 above for reference, and in the right panel discretionary spend paid via debit-direct contracts. It shows that debit-direct discretionary spend falls only marginally after app use, indicating that it is a change in behaviour sustained month-to-month that drives the reduction in discretionary spend. A caveat to this analysis is that the MDB data allows for only imperfect identification of debit-direct transactions for two reasons: first, only some banks label debit-direct transactions as such in the transaction descriptions provided to MDB and it is commonly applied labels that I use to identify the transactions. Second, MDB redacts some transaction descriptions entirely or in part to ensure the anonymity of its users, and its possible that some of these transactions are debit-directs.[25]

[25]The code used to calculate debit-direct discretionary spend is available on Github.

Figure 5: Change in direct-debit discretionary spend



Notes: Effect of app use on total discretionary spend (for reference) and discretionary spend paid via debit-direct contracts. Point estimates represent group-time average treatment effects aggregated to periods since treatment exposure, as defined in Section **??**. 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.

The second interesting way to disaggregate discretionary spend is into different component groups based on the type of spend; for this purpose, I group transactions into "groceries", "clothes", "entertainment", "food", which captures spending on restaurant meals and take-away, and "other".[26] Figure 6 again reproduces the plot from the main results in the top-left panel for reference and then plots results for each of the five subgroups. While we can see that users mainly reduce spending on groceries and the residual category "other", followed by food and clothes, the overall picture that emergest is that these differences are quite small, and that the overall reduction in discretionary spend results from changes along all these margins.

Finally, it is interesting to consider whether users reduce discretionary spend along the extensive or the intensive margin – whether they make fewer transactions or reduce the average spend per transaction. Figure 7 shows the number of transactions (the extensive margin) in the left panel and the average spend per transaction (the intensive margin) in the right panel. We can see that the reduction in spend is clearly the result of changes along the extensive margin – on average, users make about five fewer discretionary spend purchases per month once they sign up to MDB. The average transaction value of a discretionary spend purchase in the data is about £25, adding up to the total effect we find in Figure 3.

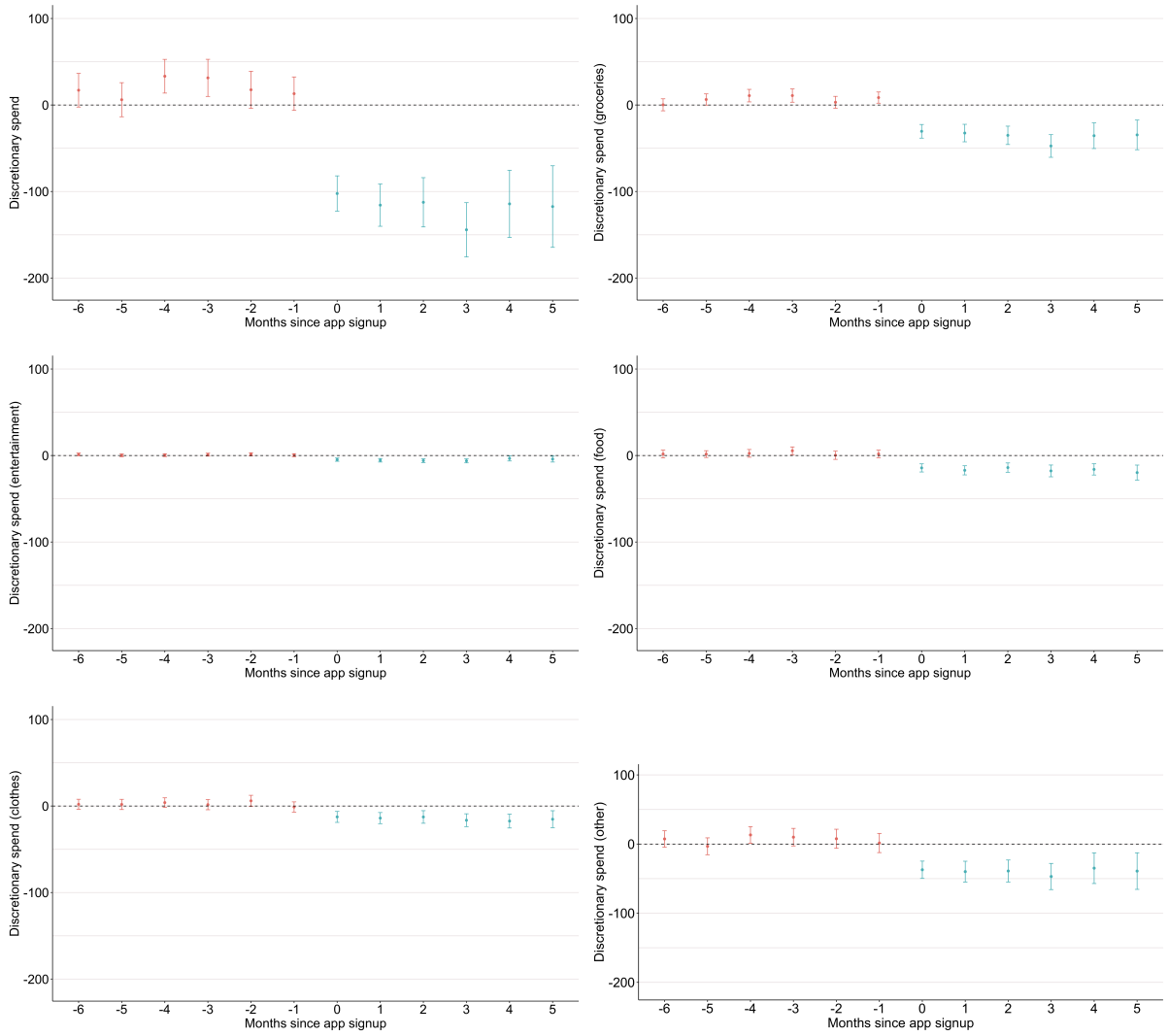
5 Conclusion

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,

[26]The list used to classify transactions is available on Github. I classify transactions tagged by MDB as "entertainment" as "other discretionary spend", since many such transactions are purchses with Amazon, which we cannot precisely classify.

Figure 6: Change in discretionary spend by subgroup



Notes: Effect of app use on total discretionary spend (for reference) and its subgroups. Point estimates represent group-time average treatment effects aggregated to periods since treatment exposure, as defined in Section **??**. 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.

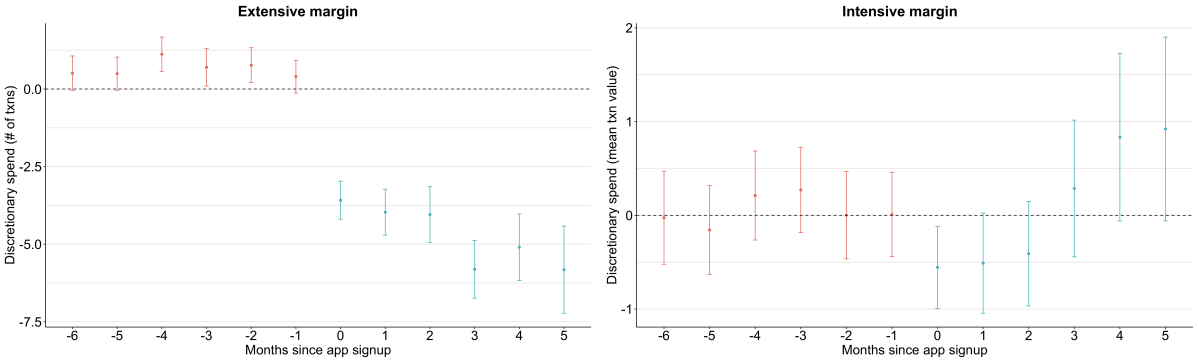
signup reflects something like openness to newness, then it's not necessarily correlated with financial savyness. Either way, we might ignore it for now. We could test whether behaviour differs between early or late adopters, but that doesn't seem important enough.

- We have good reason to believe that the conditional parallel trend assumption doesn't hold. For now, I ignore this. But in future work, I want to explore this further using approach by Rambachan and Roth (2022).

- We cannot say whether effect is driven by information or goal-setting.

- I ignore cash spending and focus on card spending only because cash spending cannot be grouped into expense categories and because it represents no more than 15 percent of total

Figure 7: Intensive and extensive margin



Notes: The effect of app use on monthly discretionary spend disaggregated into the effect on the extensive (left column) and intensive (right column) margins. The extensive margin is the number of discretionary transactions per month, and the intensive margin is the average value of a discretionary spend transaction. Point estimates represent group-time average treatment effects aggregated to periods since treatment exposure, as defined in Section **??**. 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.

spend.[27].

- It's plausible that app has negative effect as it crowds out use of superior means of budgeting.

Notes:

- Maybe not more effective because use of phone makes good decision making more difficult due to size of screen and layout and "distracted mindset" Levi and Benartzi (2020).

[27]This is an upper bound, since cash spending is the sum of all observed ATM withdrawals and some of that money will have been used for non-spend transactions

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.

Baker, Scott R (2018). "Debt and the response to household income shocks: Validation and application of linked financial account data". In: *Journal of Political Economy* 126.4, pp. 1504–1557.

Baugh, Brian, Itzhak Ben-David, and Hoonsuk Park (2014). "Disentangling financial constraints, precautionary savings, and myopia: household behavior surrounding federal tax returns". Tech. rep. National Bureau of Economic Research.

Benhabib, Jess and Alberto Bisin (2005). "Modeling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption–saving decisions". In: *Games and economic Behavior* 52.2, pp. 460–492.

Berk, Sarah Holmes, John Beshears, James J Choi, and David Laibson (2022). "Automating Short-Term Payroll Savings: Initial Evidence from a Large UK Experiment". In:

Beshears, John, James J Choi, J Mark Iwry, David C John, David Laibson, and Brigitte C Madrian (2020). "Building emergency savings through employer-sponsored rainy-day savings accounts". In: *Tax Policy and the Economy* 34.1, pp. 43–90.

Beshears, John, Katherine L Milkman, and Joshua Schwartzstein (2016). "Beyond beta-delta: The emerging economics of personal plans". In: *American Economic Review* 106.5, pp. 430–34.

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*.

Brunnermeier, Markus K and Stefan Nagel (2008). "Do wealth fluctuations generate time-varying risk aversion? Micro-evidence on individuals". In: *American Economic Review* 98.3, pp. 713–36.

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.

— (2022). "Mobile Apps and Financial Decision Making". In: *Review of Finance*.

Chater, Nick and George Loewenstein (2022). "The i-frame and the s-frame: How focusing on the individual-level solutions has led behavioral public policy astray". In: *Available at SSRN 4046264*.

Choi, James J, David Laibson, Brigitte C Madrian, and Andrew Metrick (2004). "For better or for worse: Default effects and 401 (k) savings behavior". In: *Perspectives on the Economics of Aging*. University of Chicago Press, pp. 81–126.

— (2009). "Reinforcement learning and savings behavior". In: *The Journal of finance* 64.6, pp. 2515–2534.

Choukhmane, Taha (2019). "Default options and retirement saving dynamics". In: *Working Paper*.

D'Acunto, F, AG Rossi, and M Weber (2020). "Crowdsourcing peer information to change spending behavior". In: *Chicago Booth Research Paper* 19-09.

DellaVigna, Stefano (2009). "Psychology and economics: Evidence from the field". In: *Journal of Economic literature* 47.2, pp. 315–72.

Fudenberg, Drew and David K Levine (2006). "A dual-self model of impulse control". In: *American economic review* 96.5, pp. 1449–1476.

Ganong, Peter and Pascal Noel (2019). "Consumer spending during unemployment: Positive and normative implications". In: *American Economic Review* 109.7, pp. 2383–2424.

Gargano, Antonio and Alberto G Rossi (2021). "Goal Setting and Saving in the FinTech Era". In:

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.

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". In: *Journal of Econometrics* 225.2, pp. 254–277.

Governors of the Federal Reserve System, Board of (2022). "Economic Well-Being of U.S. Households in 2021". Tech. rep.

Gul, Faruk and Wolfgang Pesendorfer (2001). "Temptation and self-control". In: *Econometrica* 69.6, pp. 1403–1435.

Hacıoğlu-Hoke, Sinem, Diego R Känzig, and Paolo Surico (2021). "The distributional impact of the pandemic". In: *European Economic Review* 134, p. 103680.

Heckman, James J, Hidehiko Ichimura, and Petra E Todd (1997). "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme". In: *The review of economic studies* 64.4, pp. 605–654.

Hsiaw, Alice (2013). "Goal-setting and self-control". In: *Journal of Economic Theory* 148.2, pp. 601–626.

Kuchler, Theresa and Michaela Pagel (2020). "Sticking To Your Plan: The Role of Present Bias for Credit Card Debt Paydown". In: *Journal of Financial Economics, forthcoming.*

Laibson, David (1997). "Golden eggs and hyperbolic discounting". In: *The Quarterly Journal of Economics* 112.2, pp. 443–478.

Laibson, David and Keith Marzilli-Ericson (2019). "Intertemporal choice". In: *Handbook of Behavioral Economics* 2.

Levi, Yaron and Shlomo Benartzi (2020). "Mind the app: Mobile access to financial information and consumer behavior". In:

Loewenstein, George and Ted O'Donoghue (2004). "Animal spirits: Affective and deliberative processes in economic behavior". In: *Available at SSRN 539843.*

Lukas, Marcel F. and Ray Charles Howard (2022). "The influence of budgets on consumer spending". In: *Journal of Consumer Research.*

Madrian, Brigitte C and Dennis F Shea (2001). "The power of suggestion: Inertia in 401 (k) participation and savings behavior". In: *The Quarterly journal of economics* 116.4, pp. 1149–1187.

Meyer, Steffen and Michaela Pagel (2018). "Fully closed: Individual responses to realized capital gains and losses". Tech. rep. Working paper.

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. URL: psyarxiv.com/qabgm.

O'Donoghue, Ted and Matthew Rabin (1999). "Doing it now or later". In: *American economic review* 89.1, pp. 103–124.

Olafsson, Arna and Michaela Pagel (2018). "The liquid hand-to-mouth: Evidence from personal finance management software". In: *The Review of Financial Studies* 31.11, pp. 4398–4446.

Philipps, Jo, Annick Kuipers, and Will Sandbrook (2021). "Supporting emergency savings: early learnings of the employee experience of workplace sidecar savings". Tech. rep.

Rambachan, Ashesh and Jonathan Roth (2022). "A More Credible Approach to Parallel Trends". Tech. rep. Working Paper.

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*.

Sims, Christopher A (2003). "Implications of rational inattention". In: *Journal of monetary Economics* 50.3, pp. 665–690.

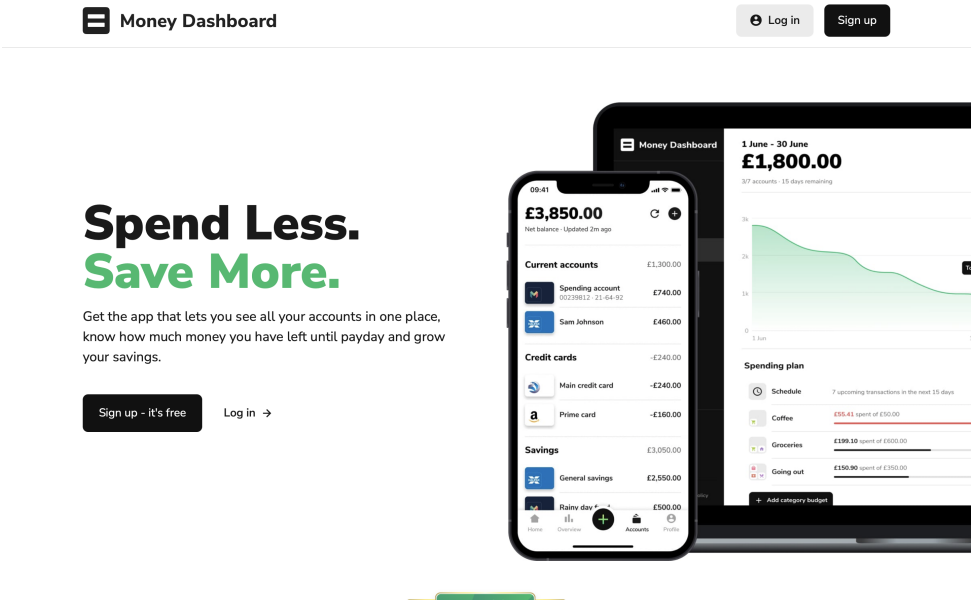
Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". In: *Journal of Econometrics* 225.2, pp. 175–199.

Thaler, Richard H and Shlomo Benartzi (2004). "Save more tomorrow^{TM}: Using behavioral economics to increase employee saving". In: *Journal of political Economy* 112.S1, S164–S187.

Thaler, Richard H and Hersh M Shefrin (1981). "An economic theory of self-control". In: *Journal of political Economy* 89.2, pp. 392–406.

A Money Dashboard application

Figure 8: Money Dashboard website screenshot



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

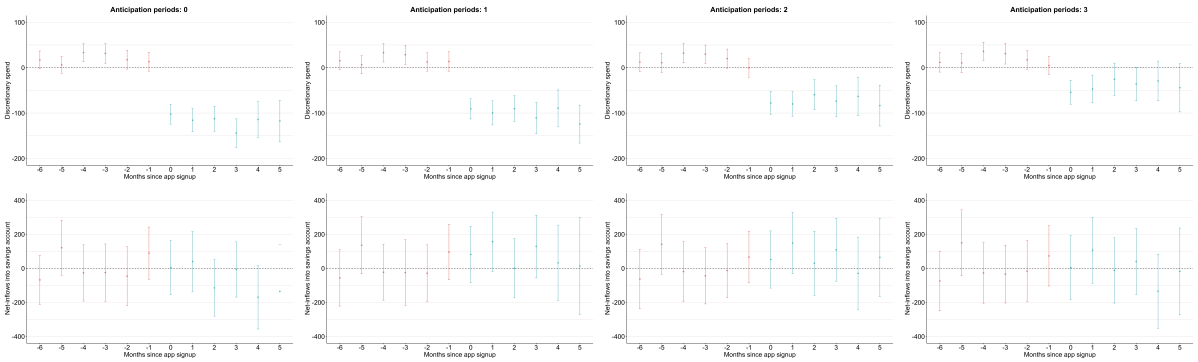
B Robustness checks

B.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 3 would underestimate
- Three reasonable scenarios: users just deciding to sign up for reasons unrelated to trajectory of dspend, decidign to cut back a few months earlier and wanting additional help, looking for tool to help them curb increase in dspend.
- 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 dspend 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 dspend before signup, which might be what causes them to sign up.

Figure 9: Anticipation ...



Notes: ...

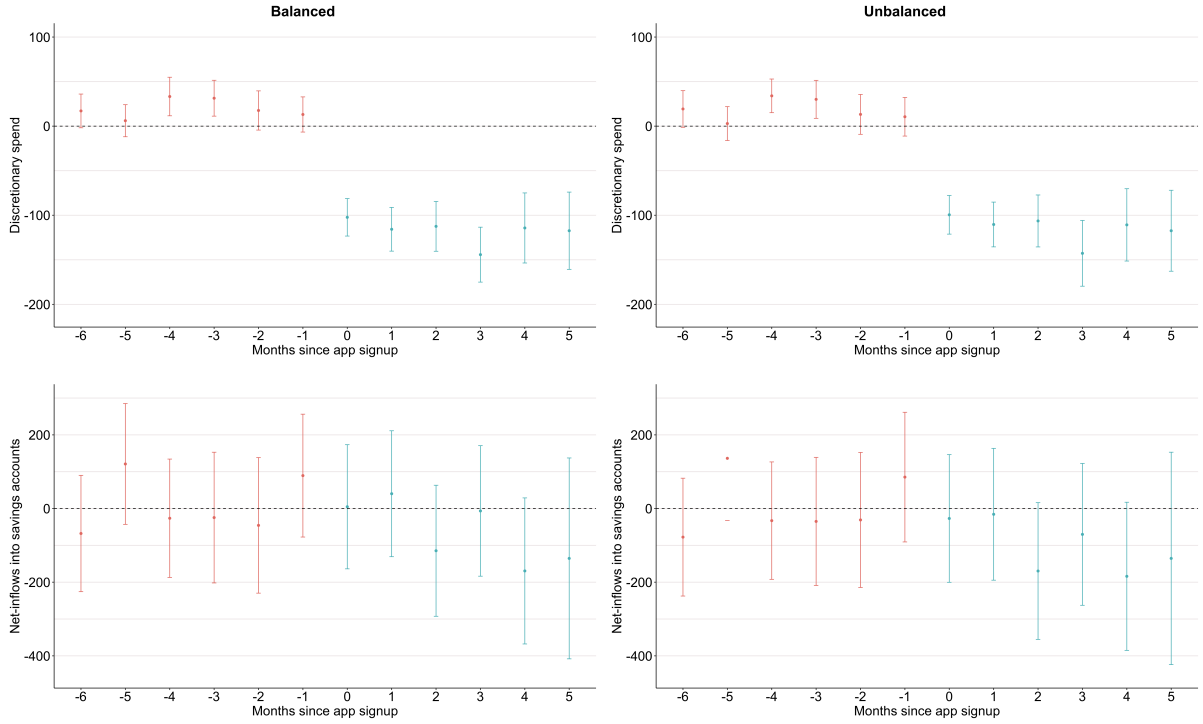
B.2 Unbalanced aggregation

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 10 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.

Figure 10: Balanced and unbalanced aggregation

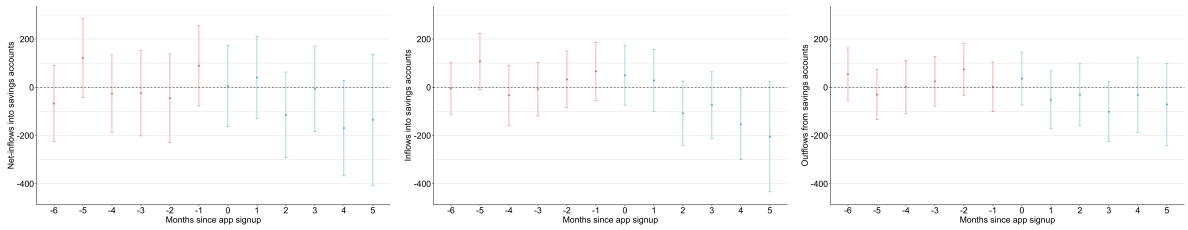


Notes: ...

B.3 Inflows and outflows

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

Figure 11: Inflows and outflows



Notes: ...