Evaluation*

Fabian Gunzinger
Warwick Business School

Neil Stewart
Warwick Business School

May 9, 2022

Contents

*This research was supported by Economic and Social Research Council grant number ES/V004867/1. WBS ethics code: E-414-01-20.

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 Dashoard 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

2.2 Sample selection

To assess the impact of MDB use on users' financial behaviour we need to observe their relevant financial history for a sufficiently long period of time prior to and after they started using the app. 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 X 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.

2.3 Treatment

Provide information about signup patterns

2.4 Outcomes

Primary outcome Our main outcome variable is savings as a proportion of monthly income, where we measure savings as the sum of all savings account inflows.

Secondary outcomes 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. We think of these additional 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

[1] In Appendix C.2 we show results with different window lengths. The results are unchanged.

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

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 1. 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]

2.5 Covariates

Description of covariates.

2.6 Difference-in-difference

Control group design:

- We only have data for a self-selected group of people who choose to use Money Dashboard. 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 use of Money Dashboard helps the average person in the population as a whole save more.[4] Instead, we are answering the question whether Money Dashboard succeeds in helping its *users* save more.
- 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.
- 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 tretment user.
- To perform the matching, we proceed as follows:
- Selection of covariates: all variables that simultaneously affect treatment and outcomes. No need to control for fixed effects: these capture unobserved time-invariant factors that make an individual sign up to MDB and affect its spending habits. Given that these are time invariant, and that all users eventually sign up, there is no difference between control and treatment units in these factors. Month of year: should probably include, as can affect p of signup and spending behaviour.

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

[4] One 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 is 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 first of these.

- For treatment units, we select data for six months before and after signup, where the month of the signup is treated as the first of the six after-signup months. For each user, we then calculate the mean value of each covariate for the pre-signup period.
- We construct potential control units as all 12-month data windows we observe before signup, and for each potential control unit calculate the mean value of each covariate for the first six months.
- Calculate propensity scores and eliminate obs outside common support?
- We use matching proceedure introduced in Ho et al. (2007) and implemented in the *MatchIt* R-package (Stuart et al. 2011) to find most similar comparison unit for each treatment unit.
- Choice of exact matching proceedure:
  - Matching with or without replacement? I'd think with, but read papers and check vignettes for trade-offs. Ho et al. (2007): if many more good (with common support) control than treatment units available, use many to one matching, else do matching with replacement. Check for common support using convex hull analysis from King and Zeng (2006).
  - Match more exactly on variables that are more correlated with outcome. How to determine? Correlation? Cutoff for "high" correlation?
  - We start with exact matching (matching all possible control units that exactly match the treatment unit). For continuous variables, we use 6 buckets (default in Stuart et al. (2011)).
  - If we match more than X percent of treatment observations, we stop. If not, we move to another approach. The goal of proceedure is to avoid biased estimates due to the deletion of too many treatment units (Rosenbaum and Rubin 1985).
  - We use nearest-neighbour matching based on propensity score. If we succeed in balancing covariates, we stop. If we don't we use more sophisticated specifications to estimate the propensity score. See Ho et al. (2007) section 6.4.
  - We asses balance following vignette.
  - Check for reduction in moden dependence producing equivalent of Fig. 2 in Ho et al. (2007).

Open questions:

- Do we include fixed effects after matching? Reason not to: we use the same units for treatment and control, so time-invariant unobserved differences should be equally distributed across treatment and control.
- How does event studies in Sun and Abraham (2021) relate to all this? Is key difference that event studies use periods since treatment rather than time?

Estimating treatment effects

- Our estimate is the ATT, not the ATE.
- First, we present pre and post signup comparisons without matching (i.e. control group is user pre-signup).
- Second, we present (static) pre-post using matched comparisons.
- Third, we present dynamic pre-post using matched comparisons. Need to think about how this relates to Sun and Abraham (2021), who propose an unbiased estimator for dynamic two-way FE event study designs. I think our analysis nests theirs, since we might still want to include fixed and time effects, though I need to think about this. (Reason to do so: we won't be able to match perfectly, so including user and time fixed effects to control for unit and time invariant variation still seems useful).
- As alternative to matching, consider synthetic controls for disaggregated data (Abadie and L'Hour 2021).

Is estimate causal?

- King and Zeng (2006) show that there are four sources of bias (ommitted 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).

3 Results

Main results: Figure and associated table akin to sakaguchi2022default Figure 2 and Table 6.

3.1 Descriptive analysis

3.2 Difference-in-difference analysis

3.3 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.[5]

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).

4 Discussion

[5]Based on age ranges provides by Beresford Research.

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.

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart (2007). "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference". In: *Political analysis* 15.3, pp. 199–236.

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.

Rosenbaum, Paul R and Donald B Rubin (1985). "The bias due to incomplete matching". In: *Biometrics*, pp. 103–116.

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.

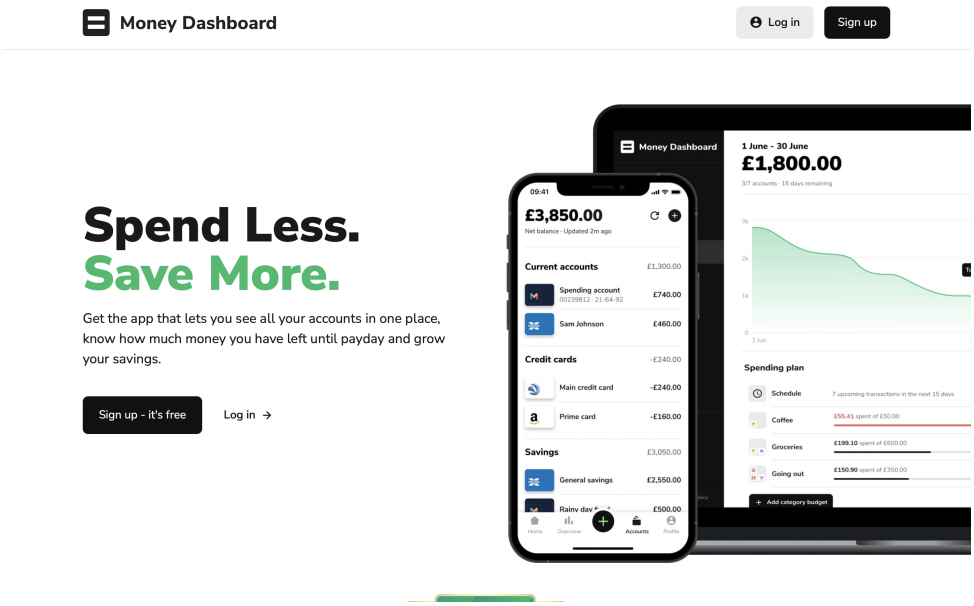
Stuart, Elizabeth A, Gary King, Kosuke Imai, and Daniel Ho (2011). "MatchIt: nonparametric preprocessing for parametric causal inference". In: *Journal of statistical software*.

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.

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 1: Money Dashboard website screenshot



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

B Variable definitions

Complete steps from raw data to variables used in analysis, with links to code on Github.

C Alternative control group designs

C.1 Alternative window lengths

Show results for 12 months on either end.

C.2 Alternative matching method

C.3 Post treatment periods as control

C.4 Synthetic controls

C.5 Classic two-way fixed-effects model

As discussed in Imai et al. (2021).

C.6 Alternative two-way fixed-effects model

Use fixest implementation of Sun and Abraham (2021).

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