The random filter identification strategy

Pierce Donovan[†]

May 2024

Abstract

The random filter identification strategy estimates a treatment effect by removing variation in treatment that is driven by unobservable confounding factors. This filtering is feasible when the treatment effect mechanism is exogenous and a mediator variable reflecting this mechanism is observable. The approach permits identification of an average treatement effect (ATE) or average treatment effect on the treated (ATT) in settings where the assumptions underlying other strategies are not met or yield a local average treatment effect (LATE). In this paper, I derive the identification assumptions for the random filter and demonstrate its applicability to economics research. My empirical example uses the random filter to measure the impact of the launch of a free-to-use mobile banking platform in El Salvador and explores the mechanisms that caused a push for financial inclusion to be unsuccessful.

Keywords: random filter, identification strategies, front door criterion, financial inclusion, digital banking, mobile money, Chivo Wallet (JEL: C13, D14, G21)

[†]Assistant Professor, Department of Economics, University of Nevada
email: pierce.donovan@unr.edu
web: piercedonovan.github.io

[††]I want to thank seminar participants at the University of Nevada, Reno, and the University of Nevada, Las Vegas, for their thoughtful comments that contributed to the framing and development of this paper. I also want to thank Max Edelstein for his help in facilitating fieldwork in El Salvador. Lastly, I am grateful for the financial support from the Lampert Institute at Colgate University that enabled that fieldwork.

1 Introduction

In observational studies, an individual's treatment status is rarely randomly assigned. This concern motivates the development of identification strategies that can yield treatment effects without selection bias. To justify a causal claim, economists frequently employ variations on matching, instrumental variables, regression discontinuity, and differences in differences in order to exploit variation in treatment that is as good as randomly assigned. This paper develops an additional avenue for causal effect identification that may succeed when the assumptions underlying these popular strategies are not met.

The random filter identification strategy leverages knowledge of a mechanism through which a treatment effect is facilitated. Even if treatment is partially driven by unobservable factors that also influence the outcome of interest, the treatment effect mechanism may be unaffected by these factors. In these cases, if the mechanism can be observed and a mediator variable can encode its status, then the relation between treatment and outcome can be "filtered" through the estimation of the two causal effects involving the mediator. These estimates, when chained together, yield a causally-interpretable treatment effect.

The ideal case for the random filter can provide some intuition regarding its operation. Imagine a treatment effect that is facilitated by an intermediate lottery, where the lottery result directly impacts an outcome of interest. Buying a lottery ticket (our treatment) changes the odds of winning—i.e. increases $P(win)$—but is otherwise unrelated to the [randomly-generated] lottery result. Thus the same odds of winning apply to any treated or untreated individual if they were to buy a ticket, and the groups provide credible counterfactuals for each other. Similarly, the lottery ensures that among ticket buyers, the losers provide credible counterfactual outcomes for the winners. Since neither causal link is confounded, multiplying estimates of the probability of winning and the effect of the win will yield the average treatment effect on the treated. If everyone is entered into the lottery so that $P(win) > 0$ for both treated and untreated, but this probability still increases with some voluntary treatment, then the win effect can also be identified for the untreated and the average treatment effect for the full population is estimable.

Like the instrumental variables approach, the random filter functions without explicit data on the drivers of selection into treatment. However, effect estimates produced by an instrumental variables strategy are weighted by an individual's susceptibility to a particular instrument. No such concession is needed for the random filter, because the approach exploits—i.e. the mediator intercepts—all of the available exogenous variation in treatment. The random filter is a thus valuable contribution that can become a regular feature in applied econometrics and expand the range of causal inference in empirical research.

This approach was first brought to light as the "Front-Door Criterion" (FDC) (Pearl, 1995, 2000). Despite its age, the FDC has not seen employment in social science research for three reasons. First, the typical presentation relies on prior knowledge of causal modeling techniques not frequently used by economists, namely Directed Acyclic Graphs (DAGs) and *do*-calculus (Imbens, 2020; Heckman and Pinto, 2022; Donovan, 2024).[1] Second, applications are difficult to imagine without an existing collection of examples (Huntington-Klein, 2022b; Bellemare et al., 2024). But perhaps most importantly, the FDC identification assumptions and the consequent treatment effect estimated had not been rigorously classified, and this information is critical for widespread adoption.[2] A proof of the external validity claim that I have made above is needed to complete an explanation of the method.

I provide a new motivation for the random filter. My approach employs the counterfactual language necessary to facilitate a treatment effect proof. The objectives of this paper are to provide a more rigorous description of the random filter using standard econometric theory, classify the treatment effect that is estimable when the random filter identification assumptions are met, discuss the novel intuition that develops from this new exposition, illustrate the identification of a treatment effect in an empirical setting, and suggest archetypal settings for which the random filter may be most applicable.

In my empirical example, I employ the random filter to determine whether the most promising facet of El Salvador's gamble on cryptocurrency and quasi-decentralized banking had a positive impact on financial well-being. As part of that agenda, the El Salvadoran government released a free-to-use financial application for savings, transactions, peer-to-peer transfers, and remittances. The civilian response was strong and immediate—a majority of the adult population voluntarily accessed the Chivo Wallet banking network within three months of its launch. The level of attention that this intervention received in its early stages suggested that the potential benefits of the program would be very high.

Since engaging with Chivo Wallet was a voluntary decision, it is likely that some of the drivers of adoption would also have some impact on financial well-being. Further, some of these factors, such as prudence or financial savvy, are unobservable. Thus an acceptable identification strategy would need to remove the influence of these confounding factors, rather than control for them directly. In the Chivo Wallet setting, the customer-facing application, ATMs, and back-end of the network were all laden with coding errors, making the network randomly inaccessible for reasons outside of users' control. These er-

[1] To ameliorate the FDC's inaccessibility and reduce dependence on graph-theoretic jargon, I am proposing an alternative nomenclature—the random filter—for use in econometrics. My term distinctly characterizes the identification strategy and provides a convenient mnemonic that aids in its understanding.

[2] Bellemare et al. (2024) confirmed the ability of the FDC to measure certain treatment effects through the use of simulation. However, they do not prove the necessity or sufficiency of their proposed assumptions.

rors introduced exogenous mediation that facilitated any potential treatment effect from downloading the application, enabling the use of the random filter. With this approach, I find little evidence that the push for financial inclusion made a positive impact—as might be expected given the lack of consistent functionality—while other methods would have produced positively-biased results.

The single contemporary use of the random filter is by Bellemare et al. (2024), who present the first credible application to observational data. Their core application estimates the effect of authorizing ride-sharing within ride-hailing applications on tipping behavior. In this setting, unobserved rider characteristics like frugality will partially influence both tipping and ride-sharing decisions, as frugal riders will choose the cheaper option and are less likely to tip. However, an exogenous mediator facilitates the treatment effect and can remove this selection bias. When someone authorizes ride-sharing, they will not necessarily share a ride with another passenger. Many potential ride-sharers go unmatched due to incompatible trips. Because the only plausible mechanism through which authorization would impact tipping behavior is through the matching process, and because the matching algorithm is known, the authors use this conditionally-exogenous matching event to filter any confounding variation between authorization and tipping. They find no effect of the former on the latter once selection bias is mitigated.

With this paper, I will demonstrate how the random filter can be broadly applied to economics research. Section 2 provides a rigorous introduction to the random filter identification strategy and classifies the identified treatment effect. Section 3 details the Chivo Wallet setting and estimates the treatment effect of signing up for an account on near-term financial outcomes. Section 4 concludes with suggestions for finding new empirical research opportunities that can make use of the random filter.

2 The random filter identification strategy

In this section, I introduce the identification assumptions for the random filter. This exposition is self-contained and does not require any understanding of DAGs or *do*-calculus.[3] I first use the potential outcomes framework to motivate the random filter (Rubin, 1974, 1977; Holland, 1986; Imbens and Rubin, 2015).[4] I then prove that the approach can identify the average treatment effect (ATE) for the population of interest or the average treatment effect on the treated (ATT) even amid the influence of unobserved confounding factors.

[3]Explanations that use these two approaches are provided by Donovan (2024) and Bellemare et al. (2024).

[4]While I take advantage of a Rubinesque treatment-control framework in this paper, the scope of the random filter is not limited to experimental settings or data with binary treatment or mediator designations.

2.1 First principles thinking for the random filter

In the typical empirical setting, we are interested in determining the impact of some treatment (T) on an outcome of interest (Y) for a group of treated individuals. The key obstacle to overcome is that an individual i, once treated, only reveals their outcome under treatment $\left(Y_i^1\right)$, and the untreated counterfactual $\left(Y_i^0\right)$ remains unobserved (Imbens and Rubin, 2015). The individual treatment effect is of course determined by the difference of these two outcomes, and the missing data problem is evident.

Since our objective is to infer what a particular treatment might do for a random individual, the missing counterfactual information presents a barrier to estimating the ATT:

$$ATT = E\left[Y_i^1 - Y_i^0 \mid T_i = 1\right]. \tag{1}$$

A separate untreated group often provides this missing data. If we focus on the difference in average group outcomes, we can relax the need for matching each treated individual and rely on group-level similarity to justify our comparison. But outside of an experimental setting, it is very likely that individuals will self-select into groups. This can drive a difference in outcomes that is not due to treatment, but other systematic differences between these groups. The identification assumption that fails is $Y_i^0 \perp\!\!\!\perp T_i$, i.e. the distribution of Y_i^0 now depends on the realized value of T_i, which implies that some factor beyond the treatment effect of interest will drive an observed difference in outcomes.

When this selection bias cannot be disentangled from the treatment effect, the untreated will not produce a credible counterfactual outcome for the treated (Duflo et al., 2006; Angrist and Pischke, 2009). This can be represented mathematically by Equation 2,

$$E\left[Y_i^0 \mid T_i = 1\right] \neq E\left[Y_i^0 \mid T_i = 0\right]. \tag{2}$$

The resulting bias prevents a comparison of outcomes $Y_i = Y_i^0 + \left(Y_i^1 - Y_i^0\right) \cdot T_i$ from having a causal interpretation:

$$\begin{aligned} &E[Y_i \mid T_i = 1] - E[Y_i \mid T_i = 0] \\ &= E\left[Y_i^1 \mid T_i = 1\right] - E\left[Y_i^0 \mid T_i = 0\right] \\ &= ATT + \left\{E\left[Y_i^0 \mid T_i = 1\right] - E\left[Y_i^0 \mid T_i = 0\right]\right\}. \end{aligned} \tag{3}$$

This difference in outcomes only yields the ATT in absence of selection bias (in braces). When $Y_i^0 \perp\!\!\!\perp T_i$ holds, the conditional statements in the latter two terms can be dropped, and the terms cancel. If we can make a stronger assumption that the untreated group would

also respond similarly to the treated, i.e. $Y_i^1, Y_i^0 \perp\!\!\!\perp T_i$, then the difference in outcomes would not only yield the ATT, but the ATE for the full population under study:

$$ATE = E\left[Y_i^1 - Y_i^0\right]. \tag{4}$$

If the confounding factor driving selection (U) is observed, a matching strategy can generate an otherwise-similar untreated group. This works by making comparisons of the treated and untreated groups conditional on each value of U where treated and untreated individuals are both observed, then averaging over these conditional average treatment effects (Heckman et al., 1997, 1998). Equation 5 is thus estimable on the common support of U, $S_T(U) = \text{supp}(U_i \mid T_i = 1) \cap \text{supp}(U_i \mid T_i = 0)$:

$$\begin{aligned}
&E_{S_T(U)}\left[E[Y_i \mid T_i = 1, U_i = u] - E[Y_i \mid T_i = 0, U_i = u]\right] \\
&= E_{S_T(U)}\left[E\left[Y_i^1 \mid T_i = 1, U_i = u\right] - E\left[Y_i^0 \mid T_i = 0, U_i = u\right]\right] \\
&= E_{S_T(U)}\left[E\left[Y_i^1 - Y_i^0 \mid T_i = 1, U_i = u\right]\right] \\
&= E_{S_T(U)}\left[ATT_u\right] = ATT_{S_T(U)}.
\end{aligned} \tag{5}$$

The third line above makes use of the fact that conditional on U, the untreated group provides a credible counterfactual outcome for the treated group, i.e. $Y_i^0 \perp\!\!\!\perp T_i \mid U_i$. The outer expectation is taken over the support of U common to both treated and untreated groups because values of U without variation in treatment cannot generate a conditional treatment effect.[5] A stronger form of the common support idea is $P(T_i = 1 \mid U_i = u) < 1$, which states that for every value of U where treatment occurs, untreated observations are also available. If this does not hold, we are only estimating the ATT for a subset of the treated (Heckman et al., 1998).

However, in many cases, U is unobserved, and this strategy is unworkable. Apart from settings involving randomized trials, selection into (or out of) treatment should be expected, even in cases where some of the variation in treatment status is purported to be exogenously-driven. In response to the selection threat, a common approach is to find an instrumental variable (Z) that drives some of this exogenous variation in T. Z will then be unresponsive to the variation in the confounding factor U and enable effect identification.

In the simplest case of a binary instrument, the independence assumption is succinctly stated as $T_i^0, T_i^1, Y_i^0, Y_i^1 \perp\!\!\!\perp Z_i$, where the superscripts on T and Y denote the status of Z and T, respectively. This notation encapsulates two additional assumptions. First, Z must have a causal effect on T, as a spurious correlation is not admissible from the potential outcome

[5] In practice, many U are continuous, so regression is used to automate weighting and matching.

independence. Second, an exclusion restriction asserts that Z impacts Y only through its impact on T, as no additional potential outcomes of Y are created and indexed by Z.

In absence of treatment, the two groups separated by an instrument meeting the above assumptions will have similar average outcomes, and their comparison will provide a causally-interpretable estimate of the intent to treat effect (ITT) (Duflo et al., 2006). A brief proof in the $Z, T \in \{0, 1\}$ case can demonstrate how the ITT can be decomposed into the impact of Z on the probability of treatment, times the treatment effect for those treated because of Z (Imbens and Angrist, 1994):

$$\begin{aligned} ITT &= E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0] \\ &= E\left[Y_i^0 + \left(Y_i^1 - Y_i^0\right) \cdot T_i^1 \mid Z_i = 1\right] - E\left[Y_i^0 + \left(Y_i^1 - Y_i^0\right) \cdot T_i^0 \mid Z_i = 0\right] \\ &= E\left[\left(Y_i^1 - Y_i^0\right) \cdot \left(T_i^1 - T_i^0\right)\right] \\ &= P\left(T_i^1 > T_i^0\right) \cdot E\left[Y_i^1 - Y_i^0 \mid T_i^1 > T_i^0\right]. \end{aligned} \tag{6}$$

The second line above utilizes the exclusion restriction assumption, which ensures that no other effects of Z on Y exist to be misattributed to T. The third line utilizes the independence of Z and any confounding factors that could pollute the $Z-T$ or $Z-Y$ relationships, which allows us to drop the conditional statements in the expectations from the second line. The final line utilizes a monotonicity assumption, which asserts that the instrument weakly influences take-up of treatment for all individuals, i.e. $T_i^1 \geq T_i^0$ (Angrist and Pischke, 2009).[6] This eliminates the $\left(T_i^1 < T_i^0\right)$ case and allows us to identify the resulting conditional effect. The ITT assigns zero weighting to the treatment effects for those whose treatment status is unresponsive to the instrument, as their $\left(T_i^1 - T_i^0\right)$ term equals zero.

To isolate the conditional effect in Equation 6, the ITT can be reduced by dividing by the impact of Z on T, which due to the independence and monotonicity assumptions above, is the output of the expected difference in T across the groups delineated by Z. The reduction provides the local average treatment effect (LATE)—an average treatment effect weighted by individual susceptibility to the instrument. In the case of binary Z and T, this is the average treatment effect on "compliers" (ATC)—those whose treatment status is set by the value of Z—as those not induced to take up treatment by the instrument receive zero weighting, and those who are will receive full weighting (Huntington-Klein, 2022a):

$$LATE = E\left[Y_i^1 - Y_i^0 \mid T_i^1 > T_i^0\right]. \tag{7}$$

[6]This assumption is commonly met in many research designs. For example, it is typical to observe noncompliance with treatment assignment in randomized trials, however, it is difficult to imagine "defiers" who would find a way to be treated if not assigned treatment, but eschew treatment if assigned to it.

This reduction in external validity is the main setback of the instrumental variables strategy, a concession made in order to ensure the causal interpretability of a difference in outcomes.[7] The intuitive cause of this issue is that the approach only exploits some—not all—of the exogenous variation in treatment, rather than filtering the endogenous variation directly. However, in some circumstances it is desirable to determine the treatment effect for those who were uninfluenced by the instrument and received zero weighting.

This direct filtering is possible in settings with the appropriate data generating process. If a variable M mediates the causal effect of T on Y and is exogenous with respect to U, then we can estimate the effect T has on Y that is facilitated by M. This is done in two [unbiased] stages. The first stage estimates the effect of T on M, and the second stage estimates the effect of M on Y, conditional on T; we then use the latter effect to scale up the former one. With this approach, the exogeneity of the mediator "filters" the endogenous variation in T generated by U before the influence of T reaches Y. Crucially, this approach does not reduce the external validity of the identified effect, as it does not introduce zero weighting on the treatment effects of any subgroups of the treated or untreated.

Below, I provide the identification assumptions for the random filter approach and prove that these are sufficient and necessary for the estimation of the ATE or ATT in settings where the data generating process admits the prerequisite mediator.

2.2 The random filter assumptions

To facilitate a proof of the Random Filter Theorem in Section 2.3, I first formalize the identification assumptions underlying the random filter identification strategy and provide some intuition for their necessity. Without loss of generality, I will work with the simplest case concerning binary treatment and mediator variables while occasionally referencing the more general case.

Assumption 1 (Mediation)**.** $Y_{Mi}^{T} = Y_{Mi}$.

We first assume that treatment can only impact the outcome via the mediator, and (A1) demonstrates that the status of T is immaterial to Y once M is fixed. We can now define the mediator and outcome variables in terms of their potential outcomes.

Definition 1 (Mediator). $M_i = M_{0i} + (M_{1i} - M_{0i}) \cdot T_i$.

Definition 2 (Outcome). $Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) \cdot M_i$.

[7] Interestingly, the instrumental variables approach doesn't typically measure a local average treatment effect on the treated (LATT) due to the fact that the most credible Z candidates are often randomly-assigned to observations. Thus the LATE indeed captures the effect on both treated and untreated compliers.

In the above definitions, the subscripts on M and Y denote the potential mediator and outcome under $T \in \{0,1\}$ and $M \in \{0,1\}$, respectively. The subscripted Y is needed to differentiate between Y_i^T and Y_{Mi} in the following proof and maintains compatibility with the previous section, while the subscript on M is for stylistic consistency with Y.

The mediation assumption is crucial if we are interested in measuring the full effect of T on Y, although this can be relaxed in the case of multiple observable causal mechanisms. This is discussed further in Section 4 and by Bellemare et al. (2024). If there is an additional causal channel between T and Y that M does not intercept, that partial effect will not be identified in the random filter. This is ultimately due to (A2) below, which requires us to control for T when estimating the impact of M on Y. This removes any correlation between M and another mechanism stemming from T, and automatically avoids the misattribution of other partial effects of T on Y to the causal channel facilitated by M.

Assumption 2 (Exogeneity)**.** $M_{1i}, M_{0i} \perp\!\!\!\perp T_i$ *and* $Y_{1i}, Y_{0i} \perp\!\!\!\perp M_i \mid T_i$.

The exogeneity assumption ensures that any observed differences in M and Y across groups delineated by T and M, respectively, are not driven by any dissimilarities between the groups themselves.[8] (A2) allows us to separate the estimation of the treatment effect into two causally-interpretable stages: the effect of T on M, and the effect of M on Y conditional on T. Clearly, the second independence claim doesn't hold unconditionally, since an unobserved factor confounding the relationship between T and Y will impact M through its effect on T. However, this assumption is met after conditioning on T, since variation in U cannot impact M after T is fixed.

In the case that there is some structural relationship between U and M, the random filter approach may still be admissible. If, for instance, there is another variable responsible for this link, and this variable is observable, we can use a matching approach to control for this factor in both stages of the estimation. Donovan (2024) discuses this idea in the context of analyzing the effectiveness of crop insurance programs, and Bellemare et al. (2024) use this approach in the identification of the ride-share effect on tipping behavior.

Condition 1 (Full Support). $P(M_i = m \mid T_i = t) \neq 1 \ \forall t, m$.

The random filter requires a data support condition to qualify which treatment effects can be measured. Pearl (2000) proposed $P(T_i = t \mid M_i = m) > 0 \ \forall t, m$, and Bellemare et al. (2024) support this condition via simulation. This alternative is equivalent to $P(M_i = m \mid T_i = t) > 0 \ \forall t, m$ after applying Bayes' Theorem, and is therefore equivalent to (C1) in the binary T, M case. In the more general case, however, (C1) is a slightly weaker condition that simply requires observed variation in M for each T group.

[8] If the strong form is implausible, this can be relaxed to mean-independence for linear model estimation.

To establish the consequence of a violation of this data requirement, imagine that the region of common support only includes $T = 1$, i.e. variation in M is not observed for the untreated. Then the mediator effect is only estimable for the treated, and the random filter estimator could only ever identify the ATT. To illustrate, the modified support condition below weakens (C1) while leaving the ATT estimable.[9]

Condition 2 (Partial Support). $P(M_i = m \mid T_i = 1) \neq 1 \ \forall m; \exists m$ s.t. $P(M_i = m \mid T_i = 0) = 1$.

In the binary case, the simplest way to encode (C2) in a data structure is to have $M = 0$ whenever $T = 0$, and let $T = 1$ signify access to a variable mediator.[10]

The support requirements for the observed data are not restrictions on the underlying data generating process, like (A1) and (A2). It is worth making an explicit statement about the potential support of the population, as an assumption concerning the potential outcomes of M aids in the decomposition of the random filter estimand in the next section.

Assumption 3 (Effect Support)**.** $P(M_{1i} > M_{0i}) > 0$ *and* $P(M_{1i} < M_{0i}) > 0$.

For there to be a treatment effect, exposure to treatment must change the value of the mediator and a change in the mediator must impact the outcome of interest. In the most general case, neither stage of the random filter has to have a monotonic effect. (A3) imposes no restriction on the first stage. This may not be immediately valuable, but in some common cases a data generating process will create a monotonic relation between T and M, and this will allow us to relax the requirements of (A2) while leaving the ATT estimable. I first prove the Random Filter Theorem given (A3) in the next section, then explore the consequences of restricting the data generating process in Section 2.4.

2.3 The random filter theorem

Random Filter Theorem. *Given (A1), (A2), (A3), and (C1), then the random filter estimand,*

$$\begin{aligned}\beta_{RF} = &\{E[M_i \mid T_i = 1] - E[M_i \mid T_i = 0]\} \\ &\cdot E_{S_{M(T)}}\left[E[Y_i \mid M_i = 1, T_i = t] - E[Y_i \mid M_i = 0, T_i = t]\right] \ , \end{aligned} \tag{8}$$

is equivalent to the ATE. If (C2) replaces (C1), then it is equivalent to the ATT.

The $E_{S_{M(T)}}[\cdot]$ notation clarifies that the outer expectation is taken over the common support of T. I first prove three lemmas (strictly in the binary T, M case), which aid the

[9] In a less policy-relevant case, the random filter can also potentially identify the average treatment on the untreated (ATU) if the partial support condition holds for $T = 0$ instead of $T = 1$.

[10] (C2) then allows us to relax the assumption that $M_{1i} \perp\!\!\!\perp T_i$, since only the ATT is estimable.

proof of the above theorem.[11] I reference the use of the above assumptions when applied.

Lemma 1 (First Stage)**.** *Given (A2) and (A3), the first stage estimand of the random filter is*

$$E[M_i \mid T_i = 1] - E[M_i \mid T_i = 0] = P(M_{1i} > M_{0i}) - P(M_{1i} < M_{0i}) \,. \tag{9}$$

Proof of Lemma 1.

$$\begin{aligned}
&E[M_i \mid T_i = 1] - E[M_i \mid T_i = 0] \\
&= E[M_{1i} \mid T_i = 1] - E[M_{0i} \mid T_i = 0] && (D1) \\
&= E[M_{1i} - M_{0i}] && (A2) \\
&= P(M_{1i} > M_{0i}) - P(M_{1i} < M_{0i}) && (A3)
\end{aligned}$$ □

Lemma 1 provides the rather intuitive result that the treatment must have some impact on the mediator if it is to impact the outcome of interest. In the final line, the difference $M_{1i} - M_{0i}$ may only take on values of $\{-1, 0, 1\}$. Archetypes whose M isn't driven by T pull the first stage estimate towards zero, which reflects that their Y will also be unmotivated by T. Note how these individuals do not receive zero weighting, in contrast to the LATE estimator for instrumental variables. That method assigns zero weight to those not affected by Z, which is what removes the ability to estimate the ATE or ATT.

Lemma 2 (Second Stage)**.** *Given (A1), (A2), and (C1), the second stage of the random filter,*

$$\begin{aligned}
&E_{S_{M(T)}}\left[E[Y_i \mid M_i = 1, T_i = t] - E[Y_i \mid M_i = 0, T_i = t]\right] \\
&= P(T_i = 1) \cdot E[Y_{1i} - Y_{0i} \mid T_i = 1] + P(T_i = 0) \cdot E[Y_{1i} - Y_{0i} \mid T_i = 0] \,.
\end{aligned} \tag{10}$$

Proof of Lemma 2.

$$\begin{aligned}
&E_{S_{M(T)}}\left[E[Y_i \mid M_i = 1, T_i = t] - E[Y_i \mid M_i = 0, T_i = t]\right] \\
&= P(T_i = 1) \cdot \{E[Y_i \mid M_i = 1, T_i = 1] - E[Y_i \mid M_i = 0, T_i = 1]\} \\
&+ P(T_i = 0) \cdot \{E[Y_i \mid M_i = 1, T_i = 0] - E[Y_i \mid M_i = 0, T_i = 0]\} && (C1) \\
&= P(T_i = 1) \cdot \{E[Y_{1i} \mid M_i = 1, T_i = 1] - E[Y_{0i} \mid M_i = 0, T_i = 1]\} \\
&+ P(T_i = 0) \cdot \{E[Y_{1i} \mid M_i = 1, T_i = 0] - E[Y_{0i} \mid M_i = 0, T_i = 0]\} && (A1), (D2) \\
&= P(T_i = 1) \cdot E[Y_{1i} - Y_{0i} \mid T_i = 1] + P(T_i = 0) \cdot E[Y_{1i} - Y_{0i} \mid T_i = 0] && (A2)
\end{aligned}$$ □

[11] The usual stable unit treatment value assumption (SUTVA) is implicit in this derivation. In the random filter setting, this assumption states that there are no treatment and mediator externalities generated between observations (otherwise recorded $T_i = 0$ or $M_i = 0$ may be false) and that the treatment/mediator doseage is identical for those with $T_i = 1$ or $M_i = 1$ (Duflo et al., 2006; Cunningham, 2021).

Corollary (Lemma 2)**.** *If (C1) does not hold, but (C2) does, then the estimand becomes*

$$E_{S_M(T)}\left[E[Y_i \mid M_i = 1, T_i = t] - E[Y_i \mid M_i = 0, T_i = t]\right] = E[Y_{1i} - Y_{0i} \mid T_i = 1]\,. \tag{11}$$

Lemma 2 provides the average mediator effect, which is [conceivably] a weighted average of the average mediator effect on the treated and the average mediator effect on the untreated. If part of the support of T does not include observations where $M_i = 1$ and $M_i = 0$, all of the density in $P(T_i = t \mid S_M(T))$ shifts to the region of common support and trivializes the $P(T_i = t)$ distribution. For example, if there is no variation in M for the untreated observations, $P(T_i = 1) = 1$, hence the result of the corollary.

Lemma 3 (Decomposition)**.** *(a) Given (A1), (A2), and (A3), the ATT estimand can be written*

$$E\left[Y_i^1 - Y_i^0 \mid T_i = 1\right] = \{P(M_{1i} > M_{0i}) - P(M_{1i} < M_{0i})\} \cdot E\left[Y_{1i} - Y_{0i} \mid T_i = 1\right], \tag{12}$$

and (b) under the same assumptions and (C1), the ATE can be decomposed into

$$\begin{aligned} E\left[Y_i^1 - Y_i^0\right] = {} & \{P(M_{1i} > M_{0i}) - P(M_{1i} < M_{0i})\} \\ & \cdot \{P(T_i = 1) \cdot E\left[Y_{1i} - Y_{0i} \mid T_i = 1\right] + P(T_i = 0) \cdot E\left[Y_{1i} - Y_{0i} \mid T_i = 0\right]\}\,. \end{aligned} \tag{13}$$

Proof of Lemma 3(a).

$$\begin{aligned} & E\left[Y_i^1 - Y_i^0 \mid T_i = 1\right] \\ & = P(M_{1i} > M_{0i} \mid T_i = 1) \cdot E\left[Y_i^1 - Y_i^0 \mid M_{1i} > M_{0i}, T_i = 1\right] \\ & + P(M_{1i} < M_{0i} \mid T_i = 1) \cdot E\left[Y_i^1 - Y_i^0 \mid M_{1i} < M_{0i}, T_i = 1\right] && (A3) \\ & = P(M_{1i} > M_{0i} \mid T_i = 1) \cdot E\left[Y_{1i} - Y_{0i} \mid M_{1i} > M_{0i}, T_i = 1\right] \\ & + P(M_{1i} < M_{0i} \mid T_i = 1) \cdot E\left[Y_{0i} - Y_{1i} \mid M_{1i} < M_{0i}, T_i = 1\right] && (A1), (D2) \\ & = P(M_{1i} > M_{0i} \mid T_i = 1) \cdot E\left[Y_{1i} - Y_{0i} \mid M_i = 1, T_i = 1\right] \\ & + P(M_{1i} < M_{0i} \mid T_i = 1) \cdot E\left[Y_{0i} - Y_{1i} \mid M_i = 0, T_i = 1\right] && (D1) \\ & = \{P(M_{1i} > M_{0i}) - P(M_{1i} < M_{0i})\} \cdot E\left[Y_{1i} - Y_{0i} \mid T_i = 1\right] && (A2) \end{aligned}$$ □

The following corollary is evident given the preceding proof.

Corollary (Lemma 3(b))**.** *Under the same assumptions, the ATU can be decomposed into*

$$E\left[Y_i^1 - Y_i^0 \mid T_i = 0\right] = \{P(M_{1i} > M_{0i}) - P(M_{1i} < M_{0i})\} \cdot E\left[Y_{1i} - Y_{0i} \mid T_i = 0\right], \tag{14}$$

and by (C1), the ATE,

$$E\left[Y_i^1 - Y_i^0\right] = P(T_i = 1) \cdot E\left[Y_i^1 - Y_i^0 \mid T_i = 1\right] + P(T_i = 0) \cdot E\left[Y_i^1 - Y_i^0 \mid T_i = 0\right],$$

matches the form in 3(b).

In the second line of the proof of Lemma 3(a), I list the archetypes in the data that could conceivably have a non-zero treatment effect. In the third line, I insert the potential outcomes of Y consistent with the conditional M statements in each expectation (for example, if treatment moves an individual's M from 0 to 1, then their treated outcome must have $M = 1$ and their untreated outcome must have $M = 0$). In the fourth line, I determine the value of M that must be observed by each archetype using the definition of M. In the final line, (A2) eliminates the conditional statements, allowing me to combine like terms.

The central result of this section is determined by combining the relevant conclusions from the three lemmas:

Proof of the Random Filter Theorem. Provided (A1), (A2), (A3), and (C1) hold, multiplying the resulting estimands of Equations 9 and 10 yields the estimand in Equation 13, thus the ATE and ATT are estimable by the random filter. If (C2) replaces (C1), then multiplying the resulting estimands of Equations 9 and 11 yields the estimand in Equation 12, thus only the ATT is estimable by the random filter. □

2.4 The random filter under a common treatment effect restriction

This section considers two common cases where (A2) can be relaxed. Many data generating processes admit a monotonicity restriction that limits the potential variability in M. For example, treatment is what often determines access to variation in the mediator, which is the case in both my empirical example in Section 3 and in Bellemare et al. (2024). Monotonicity can be written in a strong (a) or weak (b) form, seen in (A4).

Assumption 4 (Monotonicity)**.** *(a)* $M_{0i} = 0$*; (b)* $M_{1i} \geq M_{0i}$ *.*

(A4)(a) states that variation in the mediator is not possible for the untreated, which ensures only (C2) could be met and eliminates the ability to estimate the ATU (and thus the ATE). (A4)(b) is a weaker form of this assumption which is implied by (A4)(a).[12] The converse is not true, as (A4)(b) does not rule out the possibility of observing variation in M for those with $T = 0$. Thus (A4)(b) permits estimation of the ATE given (C1) holds.

[12] As written, both forms only allow for a positive first stage effect, but this can be reversed if desired.

When either the strong or weak form of (A4) holds, one can still estimate the ATT while relaxing the assumptions $M_{1i} \perp\!\!\!\perp T_i$ and $Y_{1i} \perp\!\!\!\perp M_i \mid T_i$.

Assumption 5 (Relaxed Exogeneity)**.** *$M_{0i} \perp\!\!\!\perp T_i$ and $Y_{0i} \perp\!\!\!\perp M_i \mid T_i$.*

Theorem (ATT Under Monotonicity and Relaxed Exogeneity)**.** *Given (A1), (A4), and (A5), the random filter estimand is equivalent to the ATT.*

This is shown by proving the Lemma below, which itself is evident following the steps taken in Lemmas 1-3.

Lemma 4 (ATT)**.** *Given (A1), (A4), and (A5), the first stage of the random filter estimates*

$$E[M_i \mid T_i = 1] - E[M_i \mid T_i = 0] = P(M_{1i} > M_{0i} \mid T_i = 1) \ ,$$

and the second stage (using (C2)) permits the estimation of

$$E_{S_{M(T)}}\left[E[Y_i \mid M_i = 1, T_i = t] - E[Y_i \mid M_i = 0, T_i = t]\right] = E[Y_{1i} - Y_{0i} \mid M_i = 1, T_i = 1] \ .$$

Under the same assumptions, the ATT can be decomposed into

$$E\left[Y_i^1 - Y_i^0 \mid T_i = 1\right] = P(M_{1i} > M_{0i} \mid T_i = 1) \cdot E\left[Y_{1i} - Y_{0i} \mid M_i = 1, T_i = 1\right] \ .$$

Multiplying the first two results provides the final one. The inclusion of the $M_i = 1$ in the conditional expectations arises due to the weaker exogeneity assumption. This does not reduce external validity, because those with $T_i = 1$ and $M_i = 0$ must have a treatment effect of zero due to the monotonicity assumption, as treatment failed to change the value of their mediators. Thus their impact is baked into the full ATT.

This last result greatly reduces the assumption burden for using the random filter while still yielding an estimate with more external validity than a LATE. I will now apply the random filter to my empirical example, which exhibits a data generating process that takes advantage of this new result.

3 Chivo Wallet

In this section, I describe the key events and motivations of the Chivo Wallet rollout, develop a rationale for expecting benefits from the intervention, and illustrate why it failed with the aid of a new dataset and the random filter identification strategy.[13]

[13]My exposition builds on findings by Alvarez et al. (2024) and interviews with individuals at PADE-COSMS, Credicampo, SPTF, and ASEI—four organizations providing financial services to disadvantaged

3.1 The boom and bust of Chivo Wallet

In September 2021, the El Salvadoran government deployed "Chivo Wallet," a free-to-use mobile banking application that allowed citizens to pay businesses, save money securely, and send money to others. In-network transactions such as deposits and withdrawals, transfers, and currency conversions carried no fees.[14] Those who enrolled received $30—8.5% of the median monthly salary in El Salvador—for creating an account.[15] 53% of the adult population attempted to create an account by the end of 2021, with 40% of downloads occurring within a month of the initial release (Alvarez et al., 2024).

The launch of Chivo Wallet was sudden, and users faced many issues while interacting with the network. Numerous programming errors in the phone application, ATM software, and server code prevented users from claiming the $30 sign-up bonus, interacting with ATMs, sending funds to other users, or paying for goods with Chivo Wallet. Additionally, new users occasionally discovered that an account had already been created using their government identification number. These programming and security issues were eventually fixed in March 2022, but by this time the majority of users had stopped using the app entirely. Interest in Chivo Wallet had declined due to the chaotic implementation, unease generated by political opposition and the volatility of bitcoin, and a lack of incentive beyond the initial sign-up bonus (*ASEI*; *Credicampo*; *PADECOSMS*; *SPTF*).

A second issue stemming from the hastiness of the launch was that many individuals and businesses had little to no instruction on the benefits of using Chivo Wallet unless they purposefully sought out this information.[16] Many individuals reasonably conflated Chivo Wallet with cryptocurrency and did not realize that an account was able to hold and transact in both U.S. dollars and bitcoin (*Credicampo*). Firms and would-be remittance-receivers largely shunned Chivo Wallet because they did not want to be exposed to the volatility of bitcoin, even though received bitcoin could be—and in practice, was—immediately converted to dollars by users (*Credicampo*). Had the government focused on delineating the functionality of Chivo Wallet and the network's [rather tenuous] relation to cryptocurrency, much of this confusion could have been avoided. A relaunch may be more effective with sufficient financial education, but it is unlikely that the government will attempt this due to the cost of the initial intervention (*ASEI*; *SPTF*).

communities in El Salvador. References to these interviews are noted in parentheses.

[14]Chivo Wallet is compatible with the El Salvadoran tax authority, bank accounts, and certain decentralized finance applications, but transfers out of the Chivo network incur small transaction fees.

[15]Users also qualified for an 8% discount on gas bought using Chivo Wallet.

[16]This is not for a lack of support infrastructure, however. Government officials distributed materials to help users with the Chivo Wallet app and employees were stationed around Chivo ATMs to help users navigate the machines for at least a year after the launch. Additionally, the President of El Salvador, Nayib Bukele, was found providing technical support via Twitter during the initial launch of Chivo Wallet.

3.2 Chivo Wallet's wasted potential

The intervention's emphasis on cryptocurrency adoption hampered a push for financial inclusion with significant potential. The El Salvadoran government had created Chivo Wallet to facilitate and promote bitcoin usage under the 2021 "Bitcoin Law"—a play to attract foreign private investment during a period of exceptionally-low creditworthiness. The law established bitcoin as an alternative to the U.S. dollar for paying taxes and required businesses to accept bitcoin as legal tender (Alvarez et al., 2024).

This conflation of cryptocurrency propaganda and increased access to affordable banking made any real progress towards financial inclusion unlikely. The benefits from having a bank account would not come from cryptocurrency adoption because the structure of decentralized finance actively disenfranchises smaller players (Cong et al., 2023).[17] Access to decentralized finance has had no positive impact for users except in niche cases where individuals aim to escape hyperinflation—a problem which El Salvador has not faced since its adoption of the U.S. dollar in 2001—or perpetrate scams and fraud.[18]

Nevertheless, Chivo Wallet brought a high amount of attention to virtual currencies and banking in a short period of time (*Credicampo*). In a country where 64% of adults had access to a mobile phone but 70% of adults were unbanked prior to Chivo Wallet (Alvarez et al., 2024), the potential for improving financial inclusion was great. Access would reduce the risk of carrying cash and remove the need for rural customers to travel long distances for physical cash transfers or micro-financing (*PADECOSMS*). The reduction in remittance fees should have also posed a massive benefit. Remittances constitute nearly a quarter of El Salvador's GDP, and El Salvadorans outside of the country were able to access the network and send [subsidized] funds home (Alvarez et al., 2024).

Chivo Wallet bears a resemblance to mobile money applications that have increased financial inclusion through access to peer-to-peer transfers via mobile phone accounts (Batista and Vicente, 2020). M-Pesa—the most recognizable mobile money system—had a similarly explosive start in 2007, with over 1.1 million Kenyans enrolling within eight months of its launch (Mbiti and Weil, 2011). M-Pesa was particularly successful in developing a resilient informal credit system (Jack et al., 2013; Jack and Suri, 2014), and digital traces of economic behavior allowed formal banking institutions to assess creditworthi-

[17] In a study of the Ethereum platform, Cong et al. (2023) show that transactions, mining, and wealth are concentrated among a few large players and that the bidding structure that determines transaction costs forces disproportionate fees on smaller players. High percentage fees, congestion-induced fluctuations in transaction costs, misunderstandings regarding reserve prices, and volatility in the Ether token all contribute to diminished consumer surplus. These issues are present with all other popular cryptocurrencies as well.

[18] The general erosion of societal welfare due to decentralized finance makes the whole movement disreputable. For examples of the failed, but persistent decentralized finance experiment, see the many investigations by Stephen Findeisen, Dan Olsen, or Molly White.

ness (Björkegren and Grissen, 2018). These gains in access to credit could be expected of Chivo Wallet as well, given that transaction costs were even lower than that of M-Pesa and all activity could be observed by the government.

There are many other mechanisms through which Chivo Wallet could have offered a viable pathway out of poverty. Expanding access to formal financial products has had a remarkable impact on asset accumulation, the ability to protect against income shocks, and the relaxation of credit constraints—relative to informal mechanisms such as storing cash at home or buying durable, but illiquid assets (Demirgüç-Kunt et al., 2015; Demirgüç-Kunt and Singer, 2017). The introduction of high-value savings products in particular has led to improvements in financial well-being (Prina, 2015), higher income through enabled entrepreneurship (Dupas and Robinson, 2013; Schaner, 2018), increased workforce participation in response to higher returns on capital (Callen et al., 2019), and greater trust and engagement with financial institutions (Bachas et al., 2021). Interpersonal financial relationships can also generate additional positive spillover effects for those connected to someone who is formally banked (Dupas et al., 2019). However, low-cost savings accounts are not necessarily sufficient for improving financial well-being, and positive results are very much context and mechanism-dependent (Dupas et al., 2018; de Mel et al., 2022). Chivo Wallet provides another case study to support this last point.

3.3 Personal finance survey and data generating process

The Chivo Wallet story provides an excellent opportunity to demonstrate the utility of the random filter. The general threat to identification is that downloading the Chivo Wallet application was a voluntary choice and ostensibly driven by several factors that would also have some effect on an outcome related to financial well-being, thus it is likely that a naïve comparison of those who engaged with Chivo Wallet and those who did not would reveal a large positive treatment effect. But the impact of access to an institution that did not function effectively in the first six months of use should intuitively be much more muted than this. This study demonstrates that the true impact is very small—and likely zero—by filtering out the variation in treatment that contributes no causal interpretation.

In September 2022, on the anniversary of the Chivo Wallet launch, I collected data on Chivo Wallet enrollment, the barriers faced while using the application, and coarse measures of financial well-being with the help of an enthusiastic undergraduate assistant and CID Gallup, a prominent Latin American enumerator and research company previously utilized by Alvarez et al. (2024). The survey generated a nationally-representative sample of 700 El Salvadoran residents that were eligible to download Chivo Wallet (i.e. had a

phone and were an adult) through randomized dialing of a large set of active phone numbers. The interviews occurred throughout a single week, with calls attempted throughout the day from 8:30 AM to 5:30 PM. Each number was tried up to three times, at different times of day. Respondents were asked to engage in an anonymous, three-to-five minute survey on "personal finances" that later pivoted towards questions on Chivo Wallet if an individual had attempted to create an account.

The partner enumerators first conducted a 100-person pilot sample and provided feedback on questions before the final survey. The most exigent finding in the pilot with respect to survey design was that most individuals could not provide a reliable dollar amount for a year-over-year change in savings or income since the initial launch, so we opted for Likert-scale questions to capture changing financial well-being in order to fulfil the primary empirical goal of demonstrating the mitigation of bias by the random filter. This would muddy the interpretation of the measured effect size, but in the present setting, the purpose is to show the data are consistent with a null hypothesis of no effect.

Table 1 shows that those who download Chivo Wallet are systematically different from those who do not. This is not surprising and in-agreement with the previous demographic survey by Alvarez et al. (2024). To illustrate, individuals who downloaded Chivo Wallet were more likely to already interact with formal banking institutions, use digital forms of payment, be young and male, and complete high school but not college.

The majority of those who did not download Chivo Wallet prefer to use cash. Alvarez et al. (2024) determine that this preference is due to privacy and security concerns. Most transactions in El Salvador are made with cash—an anonymous form of payment—and half of El Salvadorans use cash exclusively according to both surveys. In contrast, Chivo Wallet account transactions are not private or anonymous, as the accounts are linked to El Salvadoran identification and phone numbers. I find that these concerns, as well as a lack of trust in the application and perceived difficulty of using it were the three main reasons (cited by 75% of individuals) that someone chose not to download Chivo Wallet.

These systematic differences between treatment and control groups generate selection biases when considering differences in financial outcomes. Experience with formal financial institutions and interest in cryptocurrency are likely positively correlated with financial well-being (and improvements in well-being), as the former signals higher wealth and the latter signals higher discretionary income. Thus a naïve regression of the change in well-being on downloading Chivo Wallet will provide a positively biased result.

The hastiness of the Chivo Wallet launch provides a unique way to mitigate the selection bias detailed above. As mentioned previously, many users found out that due to a poor validation procedure within the application, their identity had already been used

Table 1: Demographic imbalance across treatment groups, yet balance across mediator groups.

	Downloaded Chivo			Experienced no Chivo issue		
Variable	$T = 0$	$T = 1$	Difference	$M = 0$	$M = 1$	Difference
no savings	0.347	0.321	-0.026	0.301	0.330	0.029
	(0.477)	(0.467)	(0.041)	(0.460)	(0.471)	(0.044)
money in bank	0.318	0.392	0.075*	0.380	0.398	0.017
	(0.467)	(0.489)	(0.043)	(0.487)	(0.490)	(0.046)
money in house	0.324	0.245	-0.078**	0.276	0.232	-0.044
	(0.469)	(0.431)	(0.039)	(0.448)	(0.422)	(0.041)
unbanked	0.512	0.389	-0.123***	0.429	0.371	-0.059
	(0.501)	(0.488)	(0.043)	(0.497)	(0.484)	(0.046)
bank account	0.424	0.511	0.088**	0.442	0.542	0.101**
	(0.496)	(0.500)	(0.044)	(0.498)	(0.499)	(0.047)
bank trips/month	1.244	1.949	0.705***	1.908	1.967	0.059
	(1.728)	(2.144)	(0.181)	(2.390)	(2.029)	(0.202)
bank travel time	43.556	36.954	-6.602	39.443	35.941	-3.502
	(39.343)	(42.611)	(4.223)	(48.681)	(39.917)	(4.418)
credit card	0.112	0.145	0.034	0.172	0.134	-0.038
	(0.316)	(0.353)	(0.030)	(0.378)	(0.341)	(0.033)
remittance change	2.435	2.527	0.092	2.506	2.535	0.029
	(0.866)	(1.107)	(0.142)	(1.193)	(1.075)	(0.143)
cash vs digital	1.854	2.517	0.663***	2.426	2.557	0.131
	(1.235)	(1.476)	(0.132)	(1.419)	(1.501)	(0.142)
only uses cash	0.609	0.404	-0.205***	0.419	0.397	-0.022
	(0.490)	(0.491)	(0.046)	(0.495)	(0.490)	(0.047)
dollars vs bitcoin	1.066	1.391	0.325***	1.283	1.438	0.155**
	(0.275)	(0.769)	(0.064)	(0.665)	(0.806)	(0.074)
female	0.524	0.368	-0.156***	0.423	0.343	-0.080*
	(0.501)	(0.483)	(0.043)	(0.496)	(0.475)	(0.045)
$18 \leq$ age ≤ 39	0.382	0.564	0.182***	0.528	0.580	0.053
	(0.487)	(0.496)	(0.044)	(0.501)	(0.494)	(0.047)
$40 \leq$ age ≤ 62	0.482	0.355	-0.128***	0.337	0.362	0.025
	(0.501)	(0.479)	(0.043)	(0.474)	(0.481)	(0.045)
age ≥ 63	0.135	0.081	-0.054**	0.135	0.057	-0.078***
	(0.343)	(0.273)	(0.026)	(0.343)	(0.233)	(0.026)
primary school	0.447	0.325	-0.123***	0.288	0.341	0.052
	(0.499)	(0.469)	(0.042)	(0.454)	(0.475)	(0.044)
high school	0.241	0.360	0.119***	0.387	0.349	-0.038
	(0.429)	(0.481)	(0.041)	(0.488)	(0.477)	(0.045)
college	0.312	0.315	0.003	0.325	0.311	-0.015
	(0.465)	(0.465)	(0.041)	(0.470)	(0.463)	(0.044)
Observations	170	530	700	163	367	530

Notes: The mediator panel makes its comparison conditional on being treated. The remittance and currency variables have a Likert scale from one to five. A remittance value of three implies remittances received did not change in the year since the launch, and lower/higher numbers imply a decrease/increase. A cash vs digital score of three implies cash is used as frequently as digital payments, with a score of one meaning that the individual only uses cash; this logic applies to the dollars vs bitcoin score as well. Bank trips per month is a count, and travel time to a bank/ATM is in minutes. All other variables are binary. Not shown: downloads and errors are geographically-distributed in proportion with population.

to collect the $30 incentive. Those who did receive the incentive could then be impacted by the inoperable programming errors that plagued the application. These complications were widespread, with a third of surveyed users revealing difficulties with withdrawing money from ATMs, making purchases, sending money to others, receiving remittances, and other technical glitches. Any potential impact of downloading Chivo Wallet on financial well-being is therefore mediated by surviving these early issues, satisfying (A1).

Table 1 shows that while there is a lack of balance in demographic and financial characteristics across treatment vs control groups, the mediator is as good as randomly assigned with respect to these factors. This is because the vast majority of reported issues with Chivo Wallet had nothing to do with user error.[19] The exogeneity of these barriers is therefore evident, given that problems with Chivo Wallet cannot be confounded with the download decision or a change in finances (conditional on downloading). Further, due to the survey's generic framing, volunteering to participate in the survey was unrelated to whether a subject faced a barrier, thus the sample selection process will not introduce any additional bias.[20] Therefore, (A2) is also satisfied.

Since those who chose not to download Chivo Wallet never faced a problem with its use, the stronger behavioral restriction (A4)(a) holds, and thus the data generating process created by this simple single-wave phone survey facilitates the estimation of the ATT.[21]

3.4 Random filter estimates of the impact of Chivo Wallet

The null hypothesis tested in this paper is that downloading Chivo Wallet did not increase financial well-being one year after its launch.[22] From Section 2, the causal relationship between T (downloading Chivo Wallet) and Y (a change in a financial well-being score) can be decomposed into two separately identifiable estimands representing the effect of T on M (access to a "functional" Chivo Wallet application) and the effect of M on Y

[19] Those who faced issues were slightly less likely to have banking or bitcoin familiarity and more likely to be older, but the odds of 3/19 rejections of group-level similarity given similar groups is 62%. Nonetheless, it could be expected that some issues were due to user error. Thus exogeneity may only be met conditional on demographic factors. A robustness test reveals no difference in the main results.

[20] Regarding external validity, the phone survey could still yield a different treatment effect relative to that of the general adult population because everyone surveyed had some free time when they were called and higher-than-average interest in answering a survey about personal finance. However, this doesn't seem likely given that the demographic results here are consistent with the in-person Alvarez et al. (2024) sample.

[21] Following this reasoning, Bellemare et al. (2024) therefore identify the ATT in their setting as well.

[22] Theft or loss of deposited funds beyond the sign-up bonus was not reported, hence the one-sided test.

(conditional on T). A linear model makes their combination straightforward:[23]

$$M_i = \gamma + \delta \cdot T_i + \varepsilon_i \tag{15}$$

$$Y_i = \theta + \lambda \cdot M_i + \phi \cdot T_i + \nu_i \,. \tag{16}$$

From (A1), (A2), and (A4), $\hat{\beta}_{RF} = \hat{\delta} \cdot \hat{\lambda}$ provides an estimate of the effect of T on Y. Following Bellemare et al. (2024), I use seemingly unrelated regression (SUR) to recover and combine estimates of each component of this effect, although the joint estimation is only used to improve the precision of the estimates within each regression and has no bearing on the estimation of the standard error of the random filter estimate.[24,25]

In small samples, bootstrapping allows us to recover the distribution of this estimator, and with larger samples, the delta-method approximation is permissible. I use the former approach here. Table 2 provides the random filter estimate of the Chivo Wallet effect alongside the [biased] estimate from an ordinary least squares (OLS) estimation.

Table 2: Random filter (via SUR) and naïve (OLS) Chivo Wallet effect estimates.

	OLS		SUR	Random Filter
	finance score	functional Chivo	finance score	finance score
intercept	3.073***	0.000	3.073***	.
	(0.089)	(0.031)	(0.088)	
download Chivo	**0.301*** **	*0.691***	0.316***	**-0.015**
	(0.102)	*(0.036)*	(0.126)	**(0.075)**
functional Chivo	.	.	*-0.022*	.
			(0.107)	
Observations	690	690	530	

Notes: "finance score" refers to the Likert-scale question asking about improvements in financial well-being at the time of the survey, one year after the launch of Chivo Wallet. Bolded estimates are the naïve and random filter estimates of the effect of Chivo Wallet on financial well-being. Italicized estimates represent the relevant first and second stage SUR estimates that are multiplied to create the random filter estimate. The random filter estimate utilizes bootstrapped standard errors.

Table 2 presents little evidence that downloading Chivo Wallet made a lasting positive impact on financial well-being. A naïve observation (the OLS estimate) would suggest a

[23] Non-linear estimators can be used here as well if one wants to make the stronger independence assumptions in Section 2 rather than weaker mean-independence claims. However, in the present ATT estimation, something like logit or probit cannot estimate the first stage since $M = 0$ whenever $T = 0$. Ultimately, the linear probability model specification will not make any predictions outside of the M and Y domains since T and M are binary, and the sample size is large enough that each regression coefficient is approximately t-distributed, so neither of the downsides typically associated with the linear model apply here.

[24] Using ordinary least squares for each stage will provide identical effect and standard error estimates, provided that one uses a bootstrap approach to measure the standard error of the random filter estimate.

[25] The results of a matching estimator, which are suppressed for clarity, provide very similar estimates of the ATT and the reported standard error of this effect.

significant positive impact, but this measurement is driven by selection bias. The random filter instead measures a precise zero for the treatment effect, owing to the fact that even a functional Chivo Wallet application failed to have any meaningful impact on the financial well-being score.

The random filter decomposition allows us to determine the size of the selection bias because the spurious correlations driven by unobserved factors are captured by the second stage regression. Since the mediator intercepts all of the exogenous variation in treatment, the second stage regression only has the residual endogenous variation to assign to the "effect" of treatment. The non-causal relationship between treatment and outcome (controlling for the mediator) is unsurprisingly close to the naïve treatment effect estimate.

The financial score variable ranges from "significant decrease" (one) to "significant increase" (five) in perceived financial well-being between September 2021 and September 2022, with a score of three signalling no meaningful change. To put the magnitude of the two point estimates in context, an effect size of 0.015 on the Likert scale is akin to claiming that ten individuals in the sample experienced a one-point improvement in this scale due to Chivo Wallet, while the naïve estimate of 0.301 equates to 210 people reporting a one-point improvement—an estimate that is off by a factor of 20. Applying the random filter estimate to the general population, we would reject any positive effect size greater than "one-point for 11% of the population" with a one-sided hypothesis test and 5% false positive rate. While this scale is somewhat coarse, we can rule out effects with the scope needed to call Chivo Wallet a financial inclusion success because those most likely to use the application are not those that stood to gain the most from adopting it.[26]

According to Alvarez et al. (2024), encountering an issue with Chivo Wallet did not discourage an individual's continued use, thus Chivo Wallet's impact was not reduced due to error-induced attrition. Instead, the treatment group's collective behavioral response to the launch explains the mechanism behind this null result. It was widely reported that a large portion of the population was having difficulty using Chivo Wallet. Public perception of the value of Chivo Wallet diminished quickly, which modified the underlying mechanism of the treatment effect of the functional application (the effect of M on Y). Positive network externalities should have been a major driver of adoption and generator of value, but the general loss of public trust following the launch eliminated this potential through less frequent or discontinued use—relative to a scenario with a successful launch

[26] Additionally, this study is able to detect an effect size of "one-point for 19% of the population" with 80% power against the null hypothesis of "no positive effect." If we apply the quasi-Bayesian approach of Lang (2023), the likelihood that the null is true (instead of the alternative 19%) given a test statistic of -0.2 is 81%, even with a prior of 10% (representing an initial strong belief in the effectiveness of Chivo Wallet). This dramatic change in belief demonstrates that the random filter's null result is a not due to a lack of precision.

and thus higher user engagement.[27]

Because the existence of our mediator shaped the effective treatment, any random variation in usability cannot be used to measure the effect of the intended treatment. But the observed intervention is shown to have zero effect, and it is easy to understand why this may be expected. The greatest stated motivator for downloading the application was the $30 incentive, and many downloads were claimed to have been solely to acquire this bonus—with no planned future use. The program was therefore equivalent to a small unconditional cash transfer, which wasn't likely to have a lasting impact on its own.[28]

The Chivo Wallet launch was not solely a push for financial inclusion—it was also a push for cryptocurrency adoption, and facilitating this second aim complicated the programming of the application, hastened the rollout, and eroded any potential long-term benefits. My data show that 76% of the eligible population downloaded Chivo Wallet and 31% of those who downloaded the application encountered some sort of issue with it. An actionable policy implication of my findings suggests that, had El Salvadorans been subject to a traditional finance institution with the sole purpose of education and financial inclusion, there could have been significant successes with respect to continued use, and eventually, financial well-being.[29]

4 Discussion

In this paper, I provide a rigorous introduction to the random filter identification strategy and demonstrate its usefulness in settings where the identification assumptions of other popular approaches are not met. Like the instrumental variables approach, the random filter functions without explicit data on the drivers of selection into treatment, however, the random filter can uncover the ATE or the ATT, rather than a LATE. The random filter is therefore a valuable addition to any empiricist's toolkit and expands the set of data generating processes that can be leveraged for causal inference.

A data generating process permits the random filter strategy when an exogenous mechanism facilitates a treatment effect on an outcome of interest. At first, this may not seem like a common pattern, but with time, researchers may learn to spot candidate mediator variables as easily as new instruments. Below, I discuss three general settings where

[27] The lack of network effects ensures that there were no spillovers to those who did not download Chivo Wallet, which would have biased the random filter estimate toward zero.

[28] The size of the transfer was also unlikely to have generated any spillovers that would have reduced the gap between treatment and control group outcomes.

[29] For example, if the government had simply applied subsidies to remittances via traditional bank transfers (a much simpler and cheaper policy), El Salvadorans could have experienced a significant and repeatable positive wealth shock in response to creating a savings account.

the random filter may be applied most effectively, and conclude with some comments on causal modeling and identification strategy discovery.

4.1 Partial treatment effects

It is likely that mediators will occasionally fail to fully intercept the impact of some intervention on an outcome of interest. Bellemare et al. (2024) have shown that data on multiple parallel mediators can be combined to estimate the full effect, if every mediator is observed.[30] If a researcher's goal is to identify the full treatment effect, imposing a greater data requirement allows us to satisfy a weaker form of (A1). However, this brings up an interesting opportunity for those interested in partial effect identification. Crucially, even if another causal effect exists, the random filter approach will remove this variation rather than misattributing it to the effect facilitated by the mediator. This allows for an unbiased estimate of partial effects—with no loss in external validity—that may be of broad interest.

In the present empirical example, had the Chivo Wallet sign-up bonus been significantly larger and reliably distributed without error (while errors continued elsewhere in the application), the random filter would allow for the estimation of the impact of Chivo Wallet after removing the immediate [likely positive] effect of the larger unconditional cash transfer. The random filter would therefore continue to measure a null result, since the direct effect of downloading the application would not be intercepted by the mediator.

4.2 Natural experiments

The random filter creates new opportunities for conducting natural experiments with observational data.[31] Random filter designs may therefore improve the financial feasibility of empirical research. In the Chivo Wallet setting, there was no way to influence who had downloaded the application, however, its usability was determined by a random process outside of individual control. If researchers have relatively small budgets and an inability to randomize treatment, observational studies with data that fit the mediation motif can still meet the prerequisite random filter assumptions for estimating treatment effects.

To illustrate, Donovan (2024) provides a potential application of the random filter to measure the effect of crop insurance programs on seasonal revenues. If crop insurance is a voluntary choice, one might guess that the farmers most interested in crop insurance may enroll because of a general business savvy that already leads to increased seasonal

[30] In contrast, only one mediator in a series of mediators facilitating a single causal effect needs to be observed for identification (Bellemare et al., 2024).

[31] In their discussion, Bellemare et al. (2024) name several theoretical examples applied to technology, labor, agriculture, and health settings.

revenues through other causal channels. Thus the crop insurance decision and seasonal revenue observation are confounded by an unobservable factor. However, crop insurance will only have an impact on seasonal revenues if a farmer is able to make an insurance claim due to some deleterious weather event. While there is selection into treatment, this insurance claim mediator is plausibly exogenous, conditional on the perceived risk of such an event (stratifying on risk blocks any potential correlation between savvy and the likelihood of a claim). Thus the random filter can recover the impact of crop insurance without an expensive research design.

The random filter may sometimes identify the effect of a precursor to the treatment of interest. With Chivo Wallet, for example, it is perhaps the effect of the mediator that is of greater interest, rather than the effect of the download itself.[32] Conveniently, the random filter identification assumptions provide guidance for measuring the causal effect of the mediator that may not be immediately obvious. The mediator in the present example is only as good as randomly assigned conditional on the treatment, and any selection bias that impacts the treatment would affect the mediator as well without this control.

4.3 Ambition effects

Another promising opportunity for the random filter is the estimation of "ambition effects." For example, in over-subscription designs for randomized controlled trials, it is ultimately a selected group of individuals upon which randomization is done. In this case, everyone in the treated and control groups have some level of ambition to receive the treatment that is different from the rest of the population. This would imply that the sample ATE measured by this design would be higher than the ATE for the general population. This is typically acceptable, as it often isn't desirable from a policy perspective to provide treatment to those who do not want it.

Using the random filter approach, the over-subscription design can yield an estimate of the impact of wanting to receive the treatment, which is distinct from the sample selection effect described above. If data can be collected for those who did not desire to be treated, we can compare outcomes for those who enrolled to those who did not, and use the random assignment to the actual treatment—an exogenous mediator between enrollment and outcome—to remove the bias associated with selection. This ambition effect—which applies to everyone who enrolled for treatment regardless if they were treated—provides additional behavioral insight complementary to the ATE generated by the intended intervention. Summing the ATE and ambition effects—since they are independent—reveals

[32] For this reason, the more policy-relevant effect in Bellemare et al. (2024) is also their mediator effect—the effect of ride-sharing, as opposed to authorizing ride-sharing.

the full impact of the treatment if one were to change an individual's mind about enrolling in the program of interest.

4.4 Causal models and new identification strategies

The random filter identification strategy has not yet been integrated into econometrics research or instruction. This is partly due to a dependence on knowledge of a niche literature in computer science and a lack of empirical examples. But another, more subtle setback was that the previous theory supporting the identification strategy was not sufficiently developed for use in economics. DAGs and *do*-calculus are tools designed to determine the causal interpretability of an estimand given an assumed data generating process. This provides a claim of internal validity, but does not establish the population to which the causal effect applies. To address this lacuna, this paper determines the identification assumptions for the random filter and classifies the treatment effect identified when these assumptions are met. These are the key prerequisites for widespread dissemination and employment in applied economics.

In empirical studies, causal inference typically considers variations on one of five established identification strategies—randomization, matching, instrumental variables, regression discontinuity, and differences in differences (Angrist and Pischke, 2015). This paper introduces a sixth strategy—the *random filter*. But the random filter may be one of many methodological discoveries made possible by embracing a transdisciplinary approach to causality (Donovan, 2024). Model paradigms like DAGs and *do*-calculus make causal discovery—the determination of the validity of an identification strategy given assumptions about a data generating process—fairly straightforward. The potential outcomes framework is not proficient here, but it can reveal crucial complementary information about the estimand of interest that DAGs and *do*-calculus are not capable of explaining. Using both of these modeling techniques together may empower researchers to answer new questions previously thought to be inaccessible.

References

Alvarez, F., Argente, D., and Van Patten, D. (2024). Are Cryptocurrencies Currencies? Bitcoin as Legal Tender in El Salvador. *Science*, Forthcoming.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press, Princeton.

Angrist, J. D. and Pischke, J.-S. (2015). *Mastering 'metrics: the path from cause to effect*. Princeton University Press, Princeton.

Bachas, P., Gertler, P., Higgins, S., and Seira, E. (2021). How Debit Cards Enable the Poor to Save More. *The Journal of Finance*, 76(4):1913–1957.

Batista, C. and Vicente, P. C. (2020). Adopting Mobile Money: Evidence from an Experiment in Rural Africa. *AEA Papers and Proceedings*, 110:594–598.

Bellemare, M. F., Bloem, J. R., and Wexler, N. (2024). The Paper of How: Estimating Treatment Effects Using the Front-Door Criterion. *Oxford Bulletin of Economics and Statistics*, Forthcoming.

Björkegren, D. and Grissen, D. (2018). The Potential of Digital Credit to Bank the Poor. *AEA Papers and Proceedings*, 108:68–71.

Callen, M., de Mel, S., McIntosh, C., and Woodruff, C. (2019). What Are the Headwaters of Formal Savings? Experimental Evidence from Sri Lanka. *The Review of Economic Studies*, 86(6):2491–2529.

Cong, L. W., Tang, K., Wang, Y., and Zhao, X. (2023). Inclusion and Democratization Through Web3 and DeFi? Initial Evidence from the Ethereum Ecosystem.

Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press.

de Mel, S., McIntosh, C., Sheth, K., and Woodruff, C. (2022). Can Mobile-Linked Bank Accounts Bolster Savings? Evidence from a Randomized Controlled Trial in Sri Lanka. *The Review of Economics and Statistics*, 104(2):306–320.

Demirgüç-Kunt, A., Klapper, L. F., Singer, D., and van Oudheusden, P. (2015). The Global Findex Database 2014: Measuring Financial Inclusion Around the World.

Demirgüç-Kunt, A. and Singer, D. (2017). Financial Inclusion and Inclusive Growth: A Review of Recent Empirical Evidence.

Donovan, P. (2024). Visualizing Causal Hypotheses in Environmental Economics. *Review of Environmental Economics and Policy*, Forthcoming.

Duflo, E., Glennerster, R., and Kremer, M. (2006). Using Randomization in Development Economics Research: A Toolkit.

Dupas, P., Karlan, D., Robinson, J., and Ubfal, D. (2018). Banking the Unbanked? Evidence from Three Countries. *American Economic Journal: Applied Economics*, 10(2):257–297.

Dupas, P., Keats, A., and Robinson, J. (2019). The Effect of Savings Accounts on Interpersonal Financial Relationships: Evidence from a Field Experiment in Rural Kenya. *The Economic Journal*, 129(617):273–310.

Dupas, P. and Robinson, J. (2013). Savings Constraints and Microenterprise Development: Evidence from a Field Experiment in Kenya. *American Economic Journal: Applied Economics*, 5(1):163–192.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an Econometric Evaluation Estimator. *The Review of Economic Studies*, 65(2):261–294.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, 64(4):605–654.

Heckman, J. J. and Pinto, R. (2022). The Econometric Model for Causal Policy Analysis. *Annual Review of Economics*, 14(1):893–923.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Huntington-Klein, N. (2022a). *The effect: an introduction to research design and causality*. CRC Press, Taylor & Francis Group, Boca Raton.

Huntington-Klein, N. (2022b). Pearl before economists: the book of why and empirical economics. *Journal of Economic Methodology*, 29(4):326–334.

Imbens, G. and Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: an introduction*. Cambridge University Press, New York.

Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature*, 58(4):1129–1179.

Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475.

Jack, W., Ray, A., and Suri, T. (2013). Transaction Networks: Evidence from Mobile Money in Kenya. *American Economic Review*, 103(3):356–361.

Jack, W. and Suri, T. (2014). Risk Sharing and Transactions Costs: Evidence from Kenya's Mobile Money Revolution. *American Economic Review*, 104(1):183–223.

Lang, K. (2023). How Credible is the Credibility Revolution?

Mbiti, I. and Weil, D. N. (2011). Mobile Banking: The Impact of M-Pesa in Kenya.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge, U.K. ; New York.

Prina, S. (2015). Banking the poor via savings accounts: Evidence from a field experiment. *Journal of Development Economics*, 115:16–31.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*, 2(1):1–26.

Schaner, S. (2018). The Persistent Power of Behavioral Change: Long-Run Impacts of Temporary Savings Subsidies for the Poor. *American Economic Journal: Applied Economics*, 10(3):67–100.