OESTERREICHISCHE NATIONALBANK
EUROSYSTEM

WORKING PAPER 230

Financial Innovation, Payment Choice and Cash Demand – Causal Evidence from the Staggered Introduction of Contactless Debit Cards

Martin Brown, Nicole Hentschel, Hannes Mettler, Helmut Stix

The *Working Paper series of the Oesterreichische Nationalbank* is designed to disseminate and to provide a platform for discussion of either work of the staff of the OeNB economists or outside contributors on topics which are of special interest to the OeNB. To ensure the high quality of their content, the contributions are subjected to an international refereeing process. The opinions are strictly those of the authors and do in no way commit the OeNB.

The Working Papers are also available on our website (http://www.oenb.at) and they are indexed in RePEc (http://repec.org/).

Publisher and editor *Oesterreichische Nationalbank*
Otto-Wagner-Platz 3, 1090 Vienna, Austria
PO Box 61, 1011 Vienna, Austria
www.oenb.at
oenb.info@oenb.at
Phone (+43-1) 40420-6666
Fax (+43-1) 40420-046698

Coordinating editor Martin Summer

Design *Communications and Publications Division*

DVR 0031577

ISSN 2310-5321 (Print)
ISSN 2310-533X (Online)

Financial Innovation, Payment Choice and Cash Demand – Causal Evidence from the Staggered Introduction of Contactless Debit Cards[*]

Martin Brown Nicole Hentschel Hannes Mettler Helmut Stix

This draft: April 20, 2020

Abstract

We examine how an innovation in payment technology impacts on consumer payment choice and cash demand. We study the staggered introduction of contactless debit cards between 2016-2018. The timing of access to the contactless technology is quasi-random across clients, depending only on the expiry date of the existing debit card. Our analysis is based on administrative data for over 21’000 bank clients and follows a pre-analysis plan. Average treatment effects show that the receipt of a contactless card increases the use of debit cards especially for small-value payments. However, we find only a moderate average reduction in the cash share of payments and no reduction of average cash demand. Treatment effects on payment choice are strongest among consumers with an intermediate pre-treatment use of cash. Explorative analyses reveal that effects are largely driven by young consumers in urban locations.

Keywords: Financial innovation, cash, money demand, payment choice, pre-analysis plan.
JEL Codes: E41, G20, O33, D14

[*] Corresponding author: Martin Brown: University of St. Gallen, Unterer Graben 21, CH-9000 St. Gallen, Email: martin.brown@unisg.ch. Affiliations: Hentschel and Mettler, University of St.Gallen, Stix: Oesterreichische Nationalbank.

The findings, interpretations and conclusions presented in this article are entirely those of the authors and should not be attributed in any manner to the Oesterreichische Nationalbank or the Eurosystem.

Non-technical summary

Cash still accounts for a significant share of payment transactions in most advanced economies. However, it is a widely held presumption that the recent innovations of contactless, mobile and instant payments will accelerate the move to a cashless society. This would pose challenges to central banks who have a mandate to guarantee a safe, efficient and broadly accessible payment system. To counterbalance ongoing payment innovations and an expected strong decline in cash demand – as has been observed e.g. in Sweden – many central banks are now contemplating the introduction of electronic cash substitutes, i.e. central bank digital currencies.

But are recent digital payment innovations really accelerating the move to a cashless society? To answer that question, this paper studies the introduction of contactless debit cards by a Swiss retail bank. We provide a clean identification of the causal effect of contactless cards on cash use and demand – disentangling the direct impact of a recent payment innovation from the broader trend in cash use that is unrelated to this innovation.

The analysis is based on strictly anonymized administrative data for over 21’000 randomly selected bank clients from one bank in Switzerland over the period 2015-2018. For each client we observe the number and value of annual point of sale payments by debit card as well as the number and value of cash withdrawals from ATMs and bank branches. The bank rolled out contactless debit cards to clients at year-end in a staggered manner depending on the expiry date of existing cards. Specifically, clients can be divided into three groups: *Early Adopters* received a new contactless-enabled debit card at the end of 2016, *Late Adopters* at the end of 2017 and *Non Adopters* at the end of 2018. Because card expiry dates are random, the three groups are very similar with regard to socio-demographic characteristics such as gender, age or income and their payment habits and money demand. Thus, any change in behavior after receiving the new debit card can be attributed to the contactless function. This setting constitutes a "natural experiment" that allows us to isolate the causal effect of the new payment technology on customers' payment behavior. In addition, we aim to strengthen the credibility of results by following a pre-analysis plan.

Our results show that the introduction of contactless debit cards causes a strong increase in the use of debit cards. After receiving a new card, on average 7 additional purchases were paid cashless (+8.6% relative to the sample mean of 79 transactions per year). The increased card use is most pronounced for small, PIN-exempt payments under CHF 20, which rise by 21%. Contactless cards do further reduce the cash share of purchases, but only by 0.6 percentage

points per year. The impact on the cash-share of payments is relatively weak as the level of debit card payments is initially low and most additional debit cards payments are of small value. Furthermore, we find no causal effect of contactless cards on cash demand as measured by the frequency and average size of cash withdrawals.

Overall, our results document statistically significant effects of payment innovation on payment choice, but the economic magnitude of these effects are small. By comparison, our data reveal a strong trend decline in the use of cash of about 2 percentage points per year, which in descriptive analyses may be confused for a causal impact of recent payment innovations. This highlights the importance of disentangling causal effects of payment innovations from overarching trends in payment behavior.

While the average treatment effect of contactless cards is small, our subsample analyses reveal substantial and informative heterogeneities across households: the impact of contactless cards is strongest among consumers with an intermediate cash share of payments. By contrast, the impact is negligible among “cash lovers”. Explorative analyses reveal that the impact of contactless cards on payment choice is largely driven by young consumers, but only those in urban locations. The latter finding suggests that recent payment innovations may accelerate the trend towards cashless transactions among technology affine consumers in locations with dense networks for cashless payments. Our results also document significant and persistent heterogeneities in payment choice across consumers which points towards the importance of habit and / or behavioral motives for payment choice and cash demand.

"The only thing useful banks have invented in 20 years is the ATM" — Paul Volcker, 2009

1. Introduction

Over the past decades, the introduction of ATMs, debit and credit cards or online-banking have revolutionized the way consumers pay for goods and services. Understanding how these significant innovations in retail payment technology affect money demand has been of first-order interest to monetary policy makers. First, changes in the structure of money demand impacts on the welfare costs of inflation (Attanasio et al. 2002, Alvarez and Lippi 2009). Second, the stability of money demand impacts on the optimal choice of a nominal anchor, i.e. the targeting of inflation as opposed to monetary aggregates (Mishkin 1999).

While previous innovations in payment technology may have altered the structure of money demand, they did not question the existence of physical central bank issued money, i.e. cash. Today, cash still accounts for a significant share of payment transactions in most advanced economies (Bagnall et al. 2016). However, this may be about to change. Recent innovations of contactless, mobile and instant payments are widely believed to be "game changing" with a higher potential of making cash obsolete.[2] A marked decline in cash demand – as has been observed e.g. in Sweden or Norway – poses two novel and important challenges to central banks[3]: First, most central banks are mandated to guarantee a safe and accessible payment system to consumers and firms. General accessibility to the payment system may be undermined if cash is no longer a universal means of payment. In addition, the overall stability of the payment system may be undermined in the event of a systemic shock to the electronic payment system. Second, in a cashless society, consumers no longer have access to an alternative safe and liquid asset in times of distress to the banking sector. For these reasons, many central banks are today contemplating

[2] The development of private digital currencies is also challenging the role of central bank issued money. A significant decline in money demand due to the use of private digital currencies have major consequences for the conduct of monetary policy, the provision of credit and liquidity to the private sector, financial stability (see e.g. Brunnermeier et al. 2019, Friedman 2000, Schilling and Uhlig 2019, Woodford 2000).

[3] The value share of cash transactions in Sweden declined from about 60% in the year 2000 to about 10% recently. Cash in circulation in percent of nominal GDP has steadily trended downwards from 3% in the year 2000 to less than 2% in 2018 (Engert et al. 2019). Two-thirds of Swedish consumers say that they can manage without cash (Sveriges Riksbank 2017). In many other countries, e.g. Canada, the U.K., Denmark, cash use declined but cash demand remained stable or even increased.

the introduction of electronic cash substitutes, i.e. central bank digital currencies (Bindseil 2020, Brunnermeier and Niepelt 2020).

Are recent digital payment innovations accelerating the move to a cashless society? We provide causal evidence on how an innovation in payment technology impacts on payment choice and cash demand. We study the staggered introduction of contactless debit cards in Switzerland. The timing of access to the contactless technology is quasi-random across clients, depending only on the expiry date of the existing debit card. Our analysis is based on administrative data for over 21’000 bank clients. For these clients we observe account-level information including point of sale (PoS) payments by debit card as well as cash withdrawals from ATMs and bank branches, over the period 2015-2018. We group the sampled clients by the timing of receipt of a contactless debit card: *Early adopters* are clients who received a contactless card at the end of 2016, *Late adopters* are clients who received the card at the end of 2017, and *Non adopters* are clients who did not receive a contactless card until end 2018. These three groups are similar with respect to pre-treatment socioeconomic characteristics as well as their pre-treatment payment and cash withdrawal behavior. Therefore, we can assign post-treatment differences in payment behavior and cash demand to the receipt of a contactless card.

Our focus on the contactless payment technology is well warranted: First, such payments are fast and convenient, especially for small value payments which typically have been the exclusive domain of cash.[4] Second, contactless payments have been growing strongly in almost all developed economies and empirical evidence indicates a concurrent decline in the use of cash (Doyle et al. 2017, Henry et al. 2018).[5] Third, the study of contactless payments is conceptually interesting because this technology lowers consumers’ costs of card vis-à-vis cash payments while leaving cash withdrawal costs unchanged. Alvarez and Lippi (2017) suggest that cash may have been resilient to earlier financial innovations, like debit cards, because these innovations have often made both the card and the withdrawal technology more efficient such that relative costs of cash and cards may not have changed much.

[4] We will henceforth refer to Near-Field-Communication debit card payments as contactless payments or as NFC payments, neglecting that such payments are also possible by credit cards or mobile devices as these payments are of low quantitative significance in Switzerland.

[5] In Canada, the share of cash in terms of the number of transactions has decreased from 54% in 2009 to 33% in 2017 (Henry et al. 2018). In Australia, the respective case share has decreased from 69% to 37% within 10 years (Doyle et al. 2017). In both economies, contactless card payments have strongly increased.

Our analysis follows a pre-analysis plan (PAP) which has been registered and time-stamped at https://osf.io/scvbq/ before data delivery. In this plan we have pre-specified the hypotheses, the data cleaning and sample selection, the definition of outcome and explanatory variables, the econometric specification and statistical inference (Olken 2015).[6]

Our hypotheses are derived from Alvarez and Lippi (2017). Their model provides an ideal conceptual framework for our research as it integrates payment instrument choice (cash vs. cards) into an inventory model of cash-management. Within this framework, the introduction of contactless cards can be seen as a reduction in the relative costs of card versus cash payments. As a consequence, contactless cards should on average reduce (i) the cash share of payments, (ii) the frequency of cash withdrawals, and (iii) the average cash withdrawal amount.

We test these hypotheses by estimating a difference-in-difference model with staggered adoption (Athey and Imbens 2018). Our estimates control for client-level and location*year-level fixed effects. They thus account for differences in unobserved transaction costs and payment preferences across consumers as well as time-varying differences in the local payment infrastructure. Our estimates of average treatment effects offer three main findings. First, the receipt of a contactless debit card causes a sizeable increase in the use of debit cards (+8.6%, relative to the sample mean of 79 debit card transactions per year). Second, the contactless payment technology reduces the cash share of payments. However, given that contactless cards mainly increase small value debit card transactions, the impact on overall payment volume is modest (0.6 percentage points (pp) relative to the average cash share of 68%). Our data reveal a downward trend of 2 pp per year in the cash share of payments that is unrelated to the contactless technology. Contactless cards thus add about 30% to this downward trend. This result signifies the importance of causal inference as the decline in the use of cash could be misinterpreted as being mainly caused by concurrent contactless cards. Third, we find no measurable effect of the contactless payment technology on cash demand, i.e. the frequency of cash withdrawals, or the average cash withdrawal amount.

In a (pre-registered) test of heterogenous treatment effects we study the impact of contactless cards across consumers with varying pre-treatment payment behavior. Pre-treatment payment behavior varies strongly in our sample: One-quarter of the sample pays almost exclusively by cash, while

[6] The use of a PAP intends to eliminate biases arising from model selection as well as from the non-reporting of insignificant findings and should thus strengthen the credibility of results, in particular for proprietary data (Casey et al. 2012). While PAPs are common in randomized control trial studies, they are much less frequent in studies using observational data (Burlig 2018). We are unaware of other papers in the monetary economics and finance literature which are based on a PAP.

another quarter pays more by card than by cash. This variation in initial behavior partly reflects differences in local payment infrastructure as well as individual cash preferences related to e.g. budget monitoring, anonymity concerns, or habit. Our results show that the impact of contactless debit cards is particularly strong among consumers with an intermediate initial cash-share of payments.

In an exploratory analysis we study the impact of contactless cards on payment behavior by consumer age and rural vs. urban location. Our findings confirm previous evidence suggesting that younger consumers are more likely to adopt financial technology (see e.g. Yang and Ching 2013). However, we show that contactless cards only exert a strong causal effect on payment behavior among those younger consumers who reside in urban locations. This suggests that technology affinity per se does not drive the adoption of the contactless payment technology. Rather it is likely that local developments in the (contactless) payment infrastructure and / or salience of the new technology among young consumers are responsible for the observed effects on payment choice.

Our paper contributes to the literature on the transaction demand for money (e.g. Baumol 1952, Tobin 1956), as well as to the literature on payment choice (e.g. Whitesell 1989). Recent theoretical approaches account for the interrelatedness of both the transaction demand for money and payment choice (e.g. Alvarez and Lippi 2017). In these models, withdrawal costs, the cost of foregone interest and differences in the costs of using cash or cards jointly determine payment choice and cash demand. The empirical literature on payment choice and cash demand has established significant and persistent heterogeneities in the use of payment instruments across households which cannot be accounted for by observed differences in transaction costs (Schuh and Stavins 2010, Arango et al. 2015, Wang and Wolman 2016, Brancatelli 2019, Stavins 2017). Further models thus emphasize behavioral determinants of payment choice and cash demand, e.g. the role of payment choice for budget control (von Kalckreuth et al. 2014, Ching and Hayashi 2010).

We contribute to this literature in three important ways:

First, in line with the recent theory (Alvarez and Lippi 2017) we empirically test the implications of financial innovation in an inventory model which jointly analyzes payment choice *and* cash demand. By contrast the previous empirical literature mostly analyzes these aspects separately. Here, our analysis complements recent work by Briglevics and Schuh (2014). While those authors examine the dynamic (short-run) sequence of payments our analysis examines the reaction of payment choice and money demand to a change in payment technology.

Second, our research design allows us to provide causal estimates of the impact of payment innovation on payment choice and cash demand. Here, our study builds on previous analyses of payment innovations and money demand. Attanasio et al. (2002), Lippi and Secchi (2009) as well as Alvarez and Lippi (2009) examine how the diffusion of cash withdrawal points (ATMs) impacts on the cash demand of Italian households. More recently, Chen et al. (2017) and Trütsch (2016) use survey data to examine the impact of contactless cards and mobile payments on payment choice and cash demand in Canada and the U.S., respectively.[7] Compared to these papers, our research design allows to better disentangle the causal effect of payment innovation from (unobserved) variation in payment behavior across households and concurrent time trends in overall payment behavior.

Third, the administrative data at hand as well as our pre-analysis plan offer two methodological novelties to the empirical literature on money demand. The bank-account-level data allow us to measure both payment choice and cash demand using precise and reliable indicators at the consumer-level over a significant period of time.[8] The existing empirical literature is based either on survey data (e.g. Borzekowski and Kiser 2008; Koulayev et al. 2016; Schuh and Stavins 2009), payment diary data (e.g. Bagnall et al. 2016; Wakamori and Welte 2017) or grocery store scanner data (Klee 2008, Wang and Wolman, 2016; Brancatelli 2019). None of these sources provide precise measures of the use of cash and cards for payments and on cash demand by the same consumers over a long period of time. Moreover, our pre-analysis plan lends credibility to the empirical results based on this data, as our reported analysis adheres to a pre-specified choice of outcome variables, econometric specifications, and subsample splits.

[7] Bounie and Camara (2019) provide evidence on the real effects of payment innovation by estimating the effects of contactless card acceptance on the profits of French merchants.

[8] Magnac (2017) uses account data to study the effects of ATM withdrawal fees.

2. Research Design, Institutional Background and Hypotheses

2.1. Research Design

We study the staggered introduction of contactless debit cards (Maestro PayPass) by one medium sized bank ("the Bank") in Switzerland over the period 2016-2018.[9] Debit cards at the Bank are valid for three calendar years, expire in December and are automatically replaced two months earlier by new cards. Starting in late 2016 (for calendar year 2017), the Bank replaced conventional debit cards with new debit cards featuring the contactless NFC function. Our research design exploits the fact that the timing of access to this new payment technology depends solely on the expiry date of the previous card, and thus is arguably exogenous from the perspective of an individual bank client.

We observe payment behavior and cash withdrawal behavior from 2015 to 2018 for a random sample of clients who all hold a transaction account and a debit card with the Bank. Our treatment variable captures the timing of receipt of a contactless debit card. The structure of our data is that of panel data with staggered adoption as discussed in Athey and Imbens (2018). As illustrated by Figure 1, clients can be separated into three groups based on the expiry date of their existing debit card. Existing debit cards of *Early adopters* expire at the end of 2016 so that their new contactless card is valid from 2017. *Late adopters* have an expiry date of end 2017 so that their new contactless card is valid from 2018. The existing debit cards of *Non adopters* expire only at the end of 2018, the end of our observation period. We use data from 2015 to conduct balancing tests of outcome variables and covariates as well as to split the sample according to pre-treatment behavior.

--- *Insert Figure 1 here* ---

2.2. Institutional Background

In Switzerland, as in many other European countries, the payment card system is dominated by debit cards which can be used to withdraw cash from ATMs of any bank as well as to make PoS

[9] Our agreement with the Bank includes its anonymity. The account-level data which we receive were strictly anonymized.

payments.[10] When opening a transaction account, bank clients receive a debit card by default. In addition to a debit card, bank clients can further request a credit card subject to an annual fee.

The 2017 survey on payment methods confirms that the overwhelming majority of PoS payments by Swiss consumers are conducted in cash or by debit card (SNB 2018). By contrast, credit cards[11] are mostly used for online purchases or for specific transactions (e.g. travel expenses, durables). According to this survey, 45% of the value and 70% of the volume of consumer transactions in 2017 were paid in cash. This widespread use of cash is similar to that observed in Germany, Italy, Austria and other Euro area economies (see Bagnall et al. 2016, Esselink and Hernández 2017), and significantly above that in Australia, Canada or the UK, for example.[12] It is important to note that the use of cash seems to be governed by a strong cash preference and not by an underdeveloped card infrastructure network. In 2018, Switzerland had 40 PoS terminals per 1,000 inhabitants, which compares with 39 in Australia, 38 in Canada and 41 in the United Kingdom.[13]

The period we study marks the widespread introduction of contactless debit cards in Switzerland. The share of debit cards featuring the NFC technology was 10% at the end of 2015, 28% in 2016, 51% in 2017 and 71% at the end of 2018.[14] While the density of PoS terminals changed little over our sample period, the share of PoS terminals which accepted contactless cards increased from 25% in 2015 to 62% in 2018.[15] In our analysis we control for time-varying heterogeneities in local payment infrastructure by employing location*year fixed effects.

2.3. Hypotheses

We derive our empirical predictions from the theoretical model of Alvarez and Lippi (2017). This model integrates payment instrument choice into an inventory model of money demand. The model thus allows us to make predictions about how the introduction of contactless cards impacts

[10] Bank clients in our sample do not have to pay fees for ATM withdrawals, regardless of whether the withdrawal occurs at an ATM from a different bank.

[11] The vast majority of credit cards are "delayed debit cards", i.e. card balances have to be paid off in full at the end of the billing period.

[12] The volume share of cash was 37% in Australia 2016 (Doyle et al., 2017) and 33% in Canada in 2017 (Henry et al. 2018).

[13] BIS (CT14B: Number of terminals per inhabitant, https://stats.bis.org/statx/srs/table/CT14b).

[14] Section 2 in the PAP summarizes the dissemination of NFC debit and credit cards and presents evidence on the share of payment instruments. A significant share of credit cards already featured a contactless payment function prior to the beginning of our observation period. However, as mentioned above, credit cards are hardly used for PoS payments in Switzerland (SNB 2018).

[15] Comparable data on contactless terminals are not available for Australia, Canada or the UK.

both on payment choice and cash demand. In the model, consumers can either make payments in cash or with cards. Cash is obtained by ATM withdrawals which can be free or costly, e.g. due to transaction fees or shoe-leather costs. Card payments always involve costs, which can either be transaction fees or the time-cost of transactions relative to cash.[16] In equilibrium, consumers either (i) use cash only or (ii) they act as *cash burners*; i.e. they use cards only when they run out of cash.[17] The model assumes a representative agent, and thus does not explore heterogeneities in payment behavior and cash demand across households. However, it is straightforward to assume that the relative cost of cash versus card payments varies across consumers depending on individual behavioral traits (budget monitoring) or the local payment infrastructure.

Within the Alvarez and Lippi (2017) framework, the introduction of contactless cards can be interpreted as a reduction in the relative costs of card payments, with cash withdrawal costs remaining constant. This implies that for all consumers who initially use cash and cards (i) the cash share of payments should decline, (ii) the average withdrawal amount should decline, (iii) the frequency of (free) ATM withdrawals should remain unaffected,[18] and (iv) the average demand for cash should therefore decline. The model further predicts that some cash-only consumers start using card payments after the introduction of contactless cards.[19] These consumers should hence reduce their number of (costly) cash withdrawals such that their overall number of withdrawals should decline.

Based on the above predictions we establish two main hypotheses for the average treatment effect of the introduction of contactless debit cards:

[16] Studies which measure the time to conduct transactions show that contactless card payments are 10 to 20 seconds below those of PIN-based card payments (Kosse et al. 2017, Polasik et al. 2010). Cash is slightly faster than contactless card payments.

[17] Consumers continue to use cash despite their ownership of cards because they have a certain number of "free" withdrawals whereas card transactions always involve "costs". The model predicts that consumers only use cards when they run out of cash which they previously withdrew at no cost. This prediction is not entirely borne out by empirical evidence. One possible reason for consumers using cards despite the availability of cash is that they want to retain cash for future purchases (c.f. Briglevics and Schuh, 2014 or Huynh et al., 2014).

[18] In this model cash-burning consumers (who use both cash and cards) do not make costly ATM withdrawals as such withdrawals are strictly dominated by cashless payments which are always possible. In the data, we presume that costly ATM withdrawals may exist also for cash-burning consumers as cards are not always accepted which could trigger a costly withdrawal. A reduction in the costs of card payments would not affect the frequency of costly withdrawals if they arise from the non-acceptance of cards.

[19] The threshold costs of withdrawals $(\underline{b})$ decreases. Thus, some consumers should move from cash-only use to cash-card use.

H1: Contactless debit cards and payment choice: The contactless payment technology reduces the use of cash as a means of payment.

H2: Contactless debit cards and cash demand: The contactless payment technology reduces the demand for cash, i.e. the frequency and the average size of cash withdrawals.[20]

For some consumers, the shift in relative costs may not be large enough and hence one might not observe a change in payment choice and cash demand. Such a prediction would be supported by behavioral models which suggest persistent heterogeneities in cash preference, e.g. due to the valuation of anonymity, budget monitoring or habit (e.g. Kahn et al. 2005, von Kalckreuth et al. 2014). Thus, we expect significant heterogeneity in the effect of the introduction of contactless cards on payment choice and cash demand across consumers which is systematically related to consumers' pre-treatment behavior: Consumers who previously only used cash are least likely to react to the payment innovation.

In our test of heterogenous treatment effects we thus predict that the magnitude of the casual effect of contactless cards is systematically related to past payment behavior:

H3: The role of past payment behavior: The impact of the contactless payment technology on cash usage and cash demand differs according to the pre-treatment use of cash. The impact should be stronger for consumers with a low pre-treatment use of cash than for consumers with a high pre-treatment use of cash.

The Alvarez and Lippi (2017) framework suggests that the demand for cash is affected by local payment infrastructure: localities with weak PoS terminal infrastructure and high density of withdrawal opportunities should feature more cash-only consumers.[21] This suggests that a reduction in the relative costs of debit card payments will have heterogenous treatment effects on payment choice and cash demand depending on the locally available payment infrastructure. In our pre-analysis plan we established a hypothesis (H4) that the effect of contactless cards on payment choice and cash demand should be stronger in locations with more PoS terminals and

[20] We focus on the frequency of withdrawals and on the average withdrawal amount as we do not observe average cash balances.

[21] See also Hyunh et al. (2014) or Arango et al. (2015) who find that payment choice decisions and cash holding decisions are affected by the availability of payment terminals.

fewer ATMs. Due to the unavailability of data on the location of PoS terminals we cannot test this hypothesis. [22]

Our conjecture is that access to the contactless payment technology reduces cash demand as consumers increasingly use debit cards for small-value, contactless-eligible payments. [23] In order to shed light on the mechanism behind the effect of the contactless payment technology on payment choice and cash demand we will explore the following auxiliary hypotheses:

H5: The contactless payment technology increases the number of small-value PoS payments (0-20 CHF) relative to all card-based PoS payments.[24]

H6: The contactless payment technology increases the number of medium sized cashless PoS payments which are eligible for the contactless technology (20-40 CHF) relative to medium sized cashless PoS payments which are not eligible for the contactless technology (40-60 CHF).

3. Data and Methodology

3.1. Sample

Our data is based a random sample of retail clients (private individuals only) of the Bank with a transaction account and at least one debit card in 2015.[25] We obtained data on 30,000 randomly drawn clients holding 30,330 accounts and 33,165 debit cards. We apply a series of restrictions to this raw sample (see Appendix A1). First, we restrict our main analysis to the overwhelming majority of clients with one account and one card only (90%=26,934 clients).[26] Second, we

[22] We collect publicly available data on the number of ATMs, population size and settlement area (km^2) for each municipality relevant to our sample. We hand collected information on ATM locations from an ATM locator webpage: https://www.mastercard.ch/de-ch/privatkunden/services-wissenswertes/services/bankomaten-suche.html as per March 2020. As discussed in detail below we define 22 locations of residence for our sample based on the local economic region (MS-region) and municipality size the consumer lives in. The data reveals that the density of the ATM-network varies from 0,29 to 1,02 per 1’000 inhabitants across our 22 locations. This compares well to the national average of 0.84 per 1’000 inhabitants (see section 2.2). Unfortunately, comparable public information on the location of PoS terminals is not available.

[23] In Switzerland contactless payments (without the typing of a PIN code) are possible for amounts up to 40 CHF.

[24] Payment diary survey data suggests that in Switzerland roughly 20% (40%) of all payments feature a value in the range of 0-5 CHF (5-20 CHF) and that more than 90% (80%) of these payments are conducted in cash (SNB 2018).

[25] The PAP details the sampling, e.g. the sample was drawn only among *active* accounts, i.e. accounts with at least 1200 CHF of incoming payments in 2015 and accounts with at least 1200 CHF of cash withdrawals or debit and credit card payments in 2015.

[26] In the PAP, we planned to include accounts with multiple cards in our sample and we described how we will handle the case of accounts with multiple debit cards (and possibly, different expiry dates). In the sample, we found out that

exclude all debit cards which experience irregular changes in the expiry date during our observation period. Irregular changes in expiry dates may occur because a card is lost or stolen or if a client demands a change of his/her card, e.g. because he/she wishes (earlier) access to the contactless technology. This results in 24,021 clients of which 22,504 have complete information on covariates. Finally, we exclude clients whose incoming or outcoming account flows are less than 1,200 CHF or more than 500,000 CHF in any year. The final sample comprises 21,122 clients, of which 8,487 are *Early adopters*, 6,150 are *Late adopters* and 6,485 are *Non adopters*.[27]

We aggregate the account-level data from a monthly to an annual frequency to account for seasonalities in payment behavior and cash demand, e.g. due to festivities or holidays. We thus obtain a balanced panel of client*year data with four observations per client *i* for periods *t= 2015, 2016, 2017, 2018* for a total of 84,488 client*year observations. As illustrated by Figure 1, our main analysis is based on a sample of 63’366 observations for the period 2016-2018. Table A2 presents the definition of all variables used in our analysis. Tables 1 and 2 present pre-treatment summary statistics and balancing tests based on the 2015 data.

3.2. Outcome Variables

As specified in our pre-analysis plan, we study three primary outcome variables which are each measured at the client*year level**.**

Our first outcome variable *Cash ratio* measures the share of annual payments (in CHF value) paid in cash. The value of total payments made in cash is hereby proxied by the total value of cash withdrawals. The total value of non-cash payments is proxied by the sum of PoS debit card payments and total credit card payments from the account.

$$Cash\ ratio\ (\%) = \frac{Value\ in\ CHF\ of\ Cash\ Withdrawals}{Value\ in\ CHF\ of\ [Cash\ withdrawals + Debit\ PoS\ payments + Credit\ card\ payments]}$$

26,923 out of 30,000 accounts (90%) have just one card (see Table A1). Therefore, we focus our analysis on accounts with one card and present robustness checks for accounts with multiple cards.

[27] The separation of clients into the three groups is not fully balanced as there was an irregular renewal of cards by the Bank in 2010 so that some cards were replaced even though they did not expire in that year. As a result, a disproportionate share of clients belongs to the early adopter group (i.e. they received a new card in 2010, in 2013 and in 2016). Importantly, this does not affect the exogeneity of the timing of access to contactless cards. However, it does explain why some covariates (e.g. age) do not fully balance across the groups of *Early, Late, and Non adopters* (see Table 2, Panel B).

We study two measures of cash demand which are central to inventory models. First, we measure the *Cash withdrawal frequency* which captures the total annual number of cash withdrawals from ATMs or from bank branches. Second, we measure the average *Cash withdrawal amount* (in CHF) as:

$$Cash\ withdrawal\ amount = \frac{Value\ in\ CHF\ of\ Cash\ withdrawals\ from\ ATMs\ or\ bank\ branches}{Number\ of\ Cash\ withdrawals\ from\ ATMs\ or\ bank\ branches}$$

The variable $Cash\ ratio$ proxies the value share of PoS payments which are made in cash. *Cash ratio* has the important advantage to be based on a precise measure of cash withdrawals from both ATMs and bank counters, which is difficult to obtain in survey data due to people's limited recall. However, the variable is also subject to measurement error arising from several sources: First, consumers may use other payment methods for PoS payments that are not covered in the denominator of $Cash\ ratio$ (e.g. mobile payments or gift cards). Evidence from payment survey data (SNB 2018) suggests, however, that this is rarely the case for PoS transactions. Second, credit card payments might include non-PoS transactions (e.g. online purchases). Again, payment diary data (SNB, 2018) suggest that this source of measurement error is small relative to the sum of cash, debit and credit transactions. Third, consumers may withdraw cash to conduct non-PoS payments (payment of recurring bills) or to hoard cash. According to SNB (2018) less than 20% of Swiss households report that they withdraw cash to pay bills or to store it. Although this might seem non-negligible, we note that the separation between cash withdrawn for transaction or for hoarding purposes is not straightforward conceptually and practically (i.e. for survey participants) as cash might be stored for ensuing purchases. Our annual aggregation of data alleviates this problem to a large degree.[28] More importantly, our panel data allows us to control for idiosyncratic – time invariant – patterns in the use of credit cards or cash for non-PoS transactions. Finally, we provide robustness tests with several alternative definition of *Cash ratio* (excluding credit cards, including e-banking payments, focusing only in domestic transactions, see Appendix A4).

The variables $Cash\ withdrawal\ frequency$ and $Cash\ withdrawal\ amount$ both proxy for the transaction demand for cash. Both variables are also subject to measurement error if consumers make withdrawals to hoard cash. SNB (2018) report that the vast majority of surveyed households

[28] The fact that cash withdrawals might also contain hoarding can also be seen as an advantage as central banks are interested in the overall demand for cash (transaction balances, precautionary balances, hoarding, etc.).

withdraw cash to make PoS payments. And, our panel data allows us to control for idiosyncratic, time invariant, patterns in cash hoarding with client-level fixed-effects.

Note that all three of our outcome variables might additionally be subject to measurement error as they may not capture all cash, debit card and credit card transactions of the households in question. In particular, this could arise if households use other current accounts (of the Bank or another bank) to conduct cash withdrawals and PoS payments we will not observe their entire payment behavior and cash demand. Survey data suggests that less than half of all Swiss households hold transaction accounts at multiple banks (Brown et al. 2020). Our account-level fixed effects also allow us to control for time-invariant variation in the use of accounts in our sample for transaction purposes.

To examine the mechanism by which the contactless payment technology affects cash use and cash demand we study six auxiliary outcome variables. These measure the frequency of *Debit PoS transactions* in total as well as by transaction size (0-20 CHF: 20-40 CHF; 40-60 CHF; 60-100 CHF; more than 100 CHF). While we do observe debit card transactions by size, we do not observe whether a debit card payment employed the contactless (NFC) technology. However, the use of the contactless feature can be inferred indirectly by separately analyzing debit card payments according to their eligibility for no PIN contactless payments (up to 40 CHF).

--- Insert Table 1 about here ---

Panel A of Table 1 presents descriptive statistics for all outcome variables based on pre-treatment (2015) observations. The table documents the importance of cash as a means of payment in our sample. The median *Cash ratio* is 78%, while the interquartile range spans 52%-96%. Thus, only one quarter of the consumers in our sample pay more with cards than they do with cash, while another quarter pay almost exclusively in cash.[29] The median of *Cash withdrawal frequency* is 39 while that of *Cash withdrawal amount* is 344 CHF, implying that the average consumer in our sample makes less than 1 cash withdrawal per week and withdraws an amount equal to roughly 258 CHF per week. A closer look at the data reveals that median number of withdrawals from ATMs (36) by far outweighs that from bank branches withdrawals (1). By contrast the median size

[29] The ratio is higher than in SNB (2018), because the latter study includes payments via bank transfer in the denominator. If we include bank transfer payments that are conducted via e-banking, we obtain a cash share of 51% (see the robustness tests in Appendix A4).

of withdrawals from ATMs (270 CHF) is significantly lower than that from bank branches (1625 CHF). The median number of *Debit PoS transactions* is 36 in 2015, while the interquartile range spans from 6 to 95. Thus, the average consumer in our sample uses the debit card only 3 times per month, while one quarter of our sample use the debit card at most every second month. The average consumer in our sample rarely uses the debit card for small-value transactions: The median number of debit transactions below 20 CHF is only 2 (!) per year in 2015. These descriptive statistics confirm the presence of pronounced heterogeneities in payment behavior that have also been noted in other studies (e.g. Attanasio et al 2002, Bagnall et al. 2016, Koulayev et al. 2016).

3.3. Methodology

The structure of our data is that of *panel data with staggered adoption* as discussed in Athey and Imbens (2018). Defining $t \in \{2016, 2017, 2018\}$ as our observation periods and $a \in \{2017, 2018\}$ as the possible adoption dates during this observation period we can identify three relevant groups of clients in our sample (see Figure 1): *Early adopters* are those clients who have a debit card which expired at end 2016 and thus adopt the contactless payment technology as per the beginning of 2017. For these clients we have adoption date $a_i = 2017$. *Late adopters* are those clients who have a debit card which expired at end 2017 and thus adopt the contactless payment technology at the beginning of 2018. For these clients we have $a_i = 2018$. *Non adopters* are those clients who have a debit card which expires at end 2018 and thus do not adopt the contactless payment technology during our observation period. In line with the notation of Athey and Imbens (2018) these clients have $a_i = \infty$.

We define $Y_{i,t}(a)$ as the potential outcome (cash use or cash demand) of client i in period t conditional on the adoption date a. We can define $\tau_{t;a,a'} = E[Y_{i,t}(a)] - E[Y_{i,t}(a')]$ as the treatment effect of adopting the technology in period a instead of period a' on outcome in period t. In this framework, the treatment effect of adoption may depend on (i) which pair of adoption dates we are comparing (a, a') and (ii) the period for which we are measuring outcomes (t).

Given our empirical setting, there are three separate treatment effects of particular interest:

- Early adoption vs. Non adoption on outcomes in 2017: $\tau_{t=2017;a=2017,a'=\infty}$
- Early adoption vs. Non adoption on outcomes in 2018: $\tau_{t=2018;a=2017,a'=\infty}$

- Late adoption vs. Non adoption on outcomes in 2018: $\tau_{t=2018;a=2018,a'=\infty}$

One may also be interested in the effect of early adoption vs. later adoption on outcomes in 2018: $\tau_{t=2018;a=2017,a'=2018}$. This can be calculated from $\tau_{t=2018;a=2017,a'=\infty}$ - $\tau_{t=2018;a=2018,a'=\infty}$.

Following Athey and Imbens (2018) we will consider a difference-in-difference (DiD) estimand τ estimated by the following regression:

[1] $$Y_{i,t} = \beta_i + \beta_t + \tau \cdot A_{i,t} + \varepsilon_{i,t}$$

where $Y_{i,t} \in \{Cash\ ratio_{i,t}, Cash\ withdrawal\ frequency_{i,t}, Cash\ withdrawal\ amount_{i,t}\}$ and $t \in \{2016, 2017, 2018\}$. In this regression β_i , β_t are client and year fixed effects respectively. $A_{i,t}$ is set to 1 for all accounts *i* in period *t* which have already adopted the technology, i.e. $a_i \leq t$ (and 0 otherwise). Athey and Imbens (2018) show that under the assumption of random assignment of adoption and no anticipation effects the DiD estimator $\hat{\tau}$ is a weighted average of the three causal treatment effects of interest listed above ($\tau_{t=2017;a=2017,a'=\infty}$; $\tau_{t=2018;a=2017,a'=\infty}$; $\tau_{t=2018;a=2018,a'=\infty}$).

Our observation of pre-adoption realizations ($t < a_i$) of the outcome variables allow us to verify the assumption of no anticipation. In particular we can compare the $Y_{i,t} \in \{Cash\ ratio_{i,t}, Cash\ withdrawal\ frequency_{i,t}, Cash\ withdrawal\ amount_{i,t}\}$ by adoption date $a_i \in \{2017, 2018, \infty\}$ for the period $t \in \{2015\}$. Panel B of Table 1 presents summary statistics for all outcome variables by treatment groups. The table displays similar pre-treatment payment behavior and cash demand across the three groups.

Our administrative data provides us with a broad set of socioeconomic and account-level covariates measured as per December 2015 (see Appendix A2, Panel B for details). Table 2 (Panel B) presents balancing tests for all covariates which allow us to verify the assumption of randomized adoption. While t-tests indicate statistically significant differences for some covariates across the treatment groups, the magnitude of these differences is negligible for most variables. We thus argue that our data largely meet the assumptions of randomized adoption as well as no anticipation.

--- Insert Table 2 about here ---

Our DiD estimator $\hat{\tau}$ provides us with a measure of the "average" effect of contactless debit cards on subsequent payment and cash holding behavior during our observation period. However, as discussed above this estimator is a weighted average of three separate treatment effects: $\tau_{t=2017;a=2017,a'=\infty}$, $\tau_{t=2018;a=2017,a'=\infty}$ and $\tau_{t=2018;a=2018,a'=\infty}$.[30]

To better understand the dynamics of this treatment effect we will explore the heterogeneity of the three individual treatment effects by running the following regression:

[2] $$Y_{i,t} = \beta_i + \beta_t + \tau_{2017,2017} \cdot A_{2017,2017} + \tau_{2017,2018} \cdot A_{2017,2018} + \tau_{2018,2018} \cdot A_{2018,2018} + \varepsilon_{i,t}$$

where $Y_{i,t} \in \{Cash\ ratio_{i,t}, Casg\ withdrawal\ frequency_{i,t}, Cash\ withdrawal\ amount_{i,t}\}$ and $t \in \{2016, 2017, 2018\}$. In this regression β_i , β_t are again individual and time fixed effects respectively. $A_{2017,2017}$ is set to 1 for all observations in period $t \in \{2017\}$ of clients who adopted the technology in 2017 (and 0 otherwise). $A_{2017,2018}$ is set to 1 for all observations in period $t \in \{2018\}$ of clients who adopted the technology in 2017 (and 0 otherwise). $A_{2018,2018}$ is set to 1 for all observations in period $t \in \{2018\}$ of clients who adopted the technology in 2018 (and 0 otherwise).

3.4. Inference

Our null-hypotheses suggest no effect of the contactless payment technology on the outcome variables *Cash ratio, Cash withdrawal frequency* and *Cash withdrawal amount.* Our statistical inference is therefore based on two-sided tests of the DiD estimators $\hat{\tau}$ in regression equations [1] and [2]. The DiD estimation of the treatment variable τ is based on data at the client*year level which includes multiple pre-treatment and post-treatment observations per account. We therefore account for potential serial correlation in the outcome variable and its effect on the standard error of our estimate for the treatment variable $\hat{\tau}$ (see Bertrand et al. 2004). We do so by adjusting standard errors for clustering at the client-level.

[30] Athey and Imbens (2018) show that two key assumptions are required for these treatment effects to be homogenous $\left(\tau = \tau_{t;\,a,a'} \forall t, a, a'\right)$. The first assumption is history invariance, i.e. the treatment effect for period *t* is independent of adoption period a, i.e. $Y_{i,t}(1) = Y_{i,t}(a)\ \forall\ \alpha \leq t$. The second assumption is constant treatment effect over time, i.e. the treatment effect of adoption period α is identical for all subsequent periods, i.e. $Y_{i,t}(a) - Y_{i,t}(\infty) = Y_{i,t'}(a) - Y_{i,t'}(\infty)\ \forall\ t, t' \geq \alpha$. In our setting neither of these assumptions are likely to hold as it is very likely that the treatment effect of contactless debit cards on payment behavior and cash demand is dynamic within subject.

We account for multiple hypothesis testing (three primary outcome variables) by adjusting our inference tests according to the Bonferroni method (see Olken, 2015). Thus, to reject either of our null-hypotheses at the 5% level we require the estimated coefficient of our treatment variables $\hat{\tau}$ in equations [1] and [2] to be significant at a level of $p<0.0167$.

4. Average Treatment Effects

4.1. Debit Card PoS Transactions

Panel A of Figure 2 depicts the average number of debit card, PoS transactions by treatment group over the period 2015 - 2018. The figure documents an increase in the number of debit card transactions for all groups during our period of interest. The increase for the group of *Non adopters* documents that even without access to the contactless payment technology there is a strong upward trend in the use of debit cards for PoS transactions. The average number of transactions per year increases for this group by 7.5% in 2016, 6.3% in 2017 and 8.4% in 2018. By comparison, however, the growth rate for debit card PoS transactions of *Early adopters* increases after they receive a contactless card (at the end of 2016) from 10.5% in 2016 to 14.2% in 2017 and 14.8% in 2018. Similarly, the growth rate for debit card PoS transactions of *Late adopters* increases after they receive a contactless card (at the end of 2017) from 9.1% in 2016 and 8.2% in 2017 to 17.9% in 2018. Panel B of Figure 2 shows that these effects are even more pronounced for transactions with a value below 20 CHF (see Appendix A3 for larger transaction amounts).

--- Insert Figure 2 about here ---

Our visual inspection in Figure 2 suggests a strong causal effect of the contactless payment technology on the use of debit cards for PoS payments. This finding is confirmed by the regression estimates presented in Table 3. The column 1 results show that the use of debit cards increases by 6.8 transactions on average per year after the receipt of a contactless card. [31] This average treatment

[31] Note that we apply standard critical values for parameter tests in Table 3, because the dependent variable does not belong to the group of primary outcome variables.

effect amounts to an 8.6% increase relative to the sample mean of 79 transactions. The bulk of this increase occurs for small transaction values: 4.9 transactions per year for amounts below 20 CHF (column 2) and 1.1 transactions per year for amounts between 20 and 40 CHF (column 3). In relation to the baseline sample mean, the increase declines from 21% for transactions up to 20 CHF to 6.1% for transactions between 20 and 40 CHF.

As we observe debit card transactions by amounts, we can test whether contactless cards trigger increases in (contactless) debit card payments also for amounts above 40 CHF still requiring the introduction of the PIN. Such effects would arise if consumers start to more frequently use their debit card through comfort-with technology effects or learning, for example. The results of Table 3, columns (4-6) suggest that these spillover effects are present for payment amounts beyond 40 CHF, although they are considerably weaker than for smaller payment amounts. For example, the relative increase in card use is just 1.8% for transactions larger than 100 CHF (relative to the sample mean).[32] Overall, the Table 3 results confirm our auxiliary hypotheses: The receipt of a contactless debit card increases the number of small-value debit card transactions relative to all such transactions (Hypothesis 5). Also, the receipt of a contactless card increases the number of medium-sized debit card transactions which are eligible for the contactless technology relative to medium transactions for which a PIN has to be entered (Hypothesis 6)

--- Insert Table 3 about here ---

4.2. Payment choice and cash demand

Access to the contactless payment technology increases the use of debit cards for PoS payments. But to what extent does this payment innovation decrease the cash share of payments and cash demand? Figure 3 illustrates the impact of the contactless payment technology on our primary outcome variables; the *Cash ratio*, the *Cash withdrawal frequency*, and the *Cash withdrawal amount*. The figure provides two key insights. First, we observe a significant trend decline in the cash ratio and the number of cash withdrawals from 2015 to 2018, while there is no change in the

[32] The quantitative impact on the number of payments should not be mistaken with the impact on cash use as a small increase of higher value payments may have a bigger effect on cash use than a larger increase of small value payments. In fact, a back-of-the-envelope calculation shows that the increase in debit card payments up to 40 CHF exerts a similar decrease in cash use as the increase in debit card payments of more than 40 CHF.

average size of cash withdrawals. Second, while there does appear to be a steeper decline of the *Cash ratio* for *Early adopters* and *Late adopters* than for *Non adopters*, the effect seems less substantial than observed in Figure 2 for debit card transactions.

--- Insert Figure 3 about here ---

Table 4 presents our estimates of the average treatment effect of the contactless payment technology on cash use and cash demand. The column 1 results indicate that contactless cards cause a decline in the *Cash ratio* by -0.6 pp per year. This amounts to an average annual treatment effect of -0.9% relative to the mean cash ratio of 68.1% in our sample for the period 2016-2018. This modest decrease fits well to the Table 3 results on debit card payments. Although the causal increase in debit card transactions is substantial, the overall number and value of such transactions is low. This implies that even a significant increase in the number of debit card transactions leads only to a small decline in the cash share of payments. The column (1) regression results also reveal a trend decrease in the cash ratio of -1.5 pp from 2016 to 2017 and -2 pp from 2017 to 2018. Thus, the causal effect of contactless cards per year is less than one-third of the annual trend. Columns (3) and (5) of Table 4 summarize the findings regarding cash demand. We find no significant effect of contactless cards on the *Cash withdrawal frequency* or *Cash withdrawal amount*.

Our main estimates in columns (1, 3, 5) of Table 4 are based on the regression specification in equation [1] including client and year fixed effects. This specification accounts for any time-invariant heterogeneity in the access to local payment infrastructure across households. As the timing of access to contactless cards is largely orthogonal to household characteristics, including the place of residence (see Table 2), it is very unlikely that our estimates are biased by unobserved heterogeneity in the development of local payment infrastructure. This is confirmed by our estimates in columns (2, 4, 6) of Table 4. There we additionally include location*year fixed effects to account for time-varying heterogeneity in local payment infrastructure.[33] Our estimates of the causal effect of contactless cards on the *Cash ratio*, *Cash withdrawal frequency*, and *Cash withdrawal amount* are unaffected.

[33] For reasons of data-protection we do not observe the exact zip-code / municipality of clients. See section 5 for a detailed discussion of how we define location based on available information on region of residence and municipality size.

--- Insert Table 4 about here ---

As discussed in section 3.3 on methodology, the average treatment effect estimates presented in Table 4 are a weighted average of three distinct treatment effects; the treatment effect on *Early adopters* in 2017, the treatment effect on *Early adopters* in 2018, and the treatment effect on *Late adopters* in 2018. In Table 5 we present separate estimates of these three treatment effects based on regression equation [2]. The results confirm our main findings from Table 4: While contactless cards impact on the *Cash ratio* we find no treatment effect at all on *Cash withdrawal frequency* or *Cash withdrawal amount*. Interestingly, Table 5 shows that the average treatment effect of contactless cards on the *Cash ratio* is largely driven by the impact on *Early adopters* and *Late adopters* in 2018. By contrast the impact on *Early adopters* in 2017 is small and statistically insignificant. It appears that the initial impact of contactless debit cards on *Early adopters* was muted - either due to lack of salience of the new payment technology or a lack in access to corresponding payment infrastructure. To sum up, the average treatment effects confirm Hypothesis 1 as a negative impact of contactless debit cards on *Cash ratio* indicates a reduced use of cash as a means of payment. By contrast, we do not find evidence for our second hypothesis, that the contactless payment technology reduces the demand for cash as *Cash withdrawal frequency* and *Cash withdrawal amount* remain unaffected.

--- Insert Table 5 about here ---

5. Heterogenous Treatment Effects

Given that Table 4 and 5 document an average treatment effect for *Cash ratio* only, we focus our analysis of heterogenous treatment effects on this outcome variable. Theory suggests that cross-sectional differences in payment behavior across households may result due to transaction costs (Alvarez and Lippi 2017) as well as persistent differences in cash preferences due to budget monitoring (von Kalckreuth et. al. 2014), habit (van der Cruijsen et al. 2017) or preferences towards anonymity (Kahn et al. 2005). As a consequence, we hypothesize that the impact of contactless payment technology on cash use and cash demand will be related to pre-treatment payment behavior. In particular, Hypothesis 3 suggests a stronger effect of contactless cards among those consumers who already frequently use non-cash payment technologies.[34]

We split our sample into four groups which correspond to four quartiles of the pre-treatment *Cash ratio,* as measured in 2015. Note from Table 1 (Panel B) that this pre-treatment level of cash use is all but identical across our three treatment groups (*Early adopters*, *Late adopters, Non adopters*). We expect that the treatment effect of contactless cards on the *Cash ratio* should be smaller for consumers with a higher pre-treatment cash use. As predicted, the groups of consumers with the highest pre-treatment cash use (columns 3 and 4 in Table 6) reveal the lowest (relative) treatment effect. In these groups contactless cards lead to a statistically insignificant reduction of the *Cash ratio* by 0.35 pp, compared to a pre-treatment level of more than 78%. Interestingly, the group of consumers who used cards most intensively before treatment (column 1) also reveal a low and insignificant treatment effect. In this group, contactless cards lead to a reduction of the *Cash ratio* by only 0.17 pp, compared to a pre-treatment level of 35%. This insignificant treatment effect may indicate either demand-side saturation effects or supply side constraints.

Table 6 (column 2) documents a sizeable and significant treatment effect of contactless cards for the group with an intermediate pre-treatment cash ratio. In this group, contactless cards reduce the cash ratio by 1.3 pp per year compared to an average pre-treatment cash ratio of 60%. This finding suggests contactless cards may have the largest impact on card vs. cash payments among those clients who initially make regular, but few card payments. A closer look at the frequency of debit card payments for this group of clients supports this conjecture. In unreported regressions we replicate our Table 3 analysis only for this group of clients. In this group the average number of debit transactions increases from 94 in 2016 to 119 in 2018. The average treatment effect of contactless cards is estimated to be 9 transactions per year in this subsample. In line with the Table 3 findings, this treatment effect is mainly driven by debit card payments for small value transactions (below 20 CHF), where contactless cards lead to an increase by 6.3 transactions per year.

--- Insert Table 6 about here ---

[34] As noted in section 2.3. we cannot test Hypothesis 4 from our pre-analysis plan due to a lack of data on locations of PoS terminals.

In Table 7 we present an explorative (not pre-registered) subsample analysis. Here we examine whether the treatment effect of contactless debit cards on *Cash ratio* differs by location (urban vs. rural) and age of consumers. Survey evidence shows that the payment behavior of consumers within Switzerland varies cross-sectionally both by age and location (SNB 2018). There are many reasons why this may be the case: Local payment infrastructure (PoS terminals vs. ATMs) and thus relative transaction costs of cards vs. cash for the same type of purchases may differ between urban and rural areas. Individual consumption behavior (types of goods and services purchased, timing of purchases) may differ by age group, so that differences in payment infrastructure across types of purchases would lead to differences in observed payment behavior. Differences in behavioral traits (budget monitoring), habits as well as network effects may also affect payment behavior across locations and age groups. If payment behavior differs cross-sectionally by location and age-group it is also plausible that we could see a heterogenous impact of a change in payment technology on this behavior. Young and urban consumers may be more likely to adopt the contactless payment technology than older consumers in rural areas.

Based on our administrative data we split our sample by three, similarly sized age groups: less than 35 years old, 35-55 years and above 55 years. We also split our sample, by whether the client resides in an urban or rural area. For reasons of data-protection we do not observe the zip-code of clients. We do, however, observe the local economic region (MS-region) as well as the size (number of inhabitants) of the municipality in which the client resides (0-5’000; 5’001-10’000; 10’001-20’000; 20’001-50’000; more than 50’000). Crossing this information, we can distinguish 22 locations based on a combination of the local economic region and the size of the municipality within that region that the client resides in. We collect publicly available data on population size and settlement area (km^2) for each municipality relevant to our sample. Aggregating this information for each location we obtain a measure of population density per region.[35] We categorize locations with a population density of more (less) than 3’000 inhabitants per km^2 as urban (rural).

Table 7 presents our subsample estimates for the impact of contactless cards on *Cash ratio* by age and location. The results are striking. First, we observe that the cash share of payments depends strongly on client age, but hardly on client location. In urban locations the mean *Cash ratio* varies

[35] The data reveals that the population density varies from just under 1’500 inhabitants per km^2 to just over 4’500 inhabitants per km^2. The median population density is just under 3’000 inhabitants per km^2.

from 58% for consumers below 35 years to 66% for 35-55 year olds and 78% for clients above 55 years. The mean cash share of payments is almost identical by age group for clients in rural areas. Second, younger consumers exhibit a stronger trend decline in the cash share of payments than older consumers. And again the time trend per age-group is independent of urban vs. rural location. Consumers aged below 35 years display a decline in the *Cash ratio* by 3-4 pp per year in 2017 and 2018 compared to 2016. The trend decline for 35-55 year olds is 1-2 pp per year while it is roughly half a percentage point per year for clients above 55 years. Third, the causal impact of contactless cards on the *Cash ratio* is large and statistically significant only for young consumers in urban areas (column 1). In this subsample, the receipt of a contactless card reduces the *Cash ratio* by 1.25 pp per year. This effect is sizeable as it amounts to 2% of the subsample mean and more than one-third of the annual trend decline. By comparison, the estimate of the causal effect of contactless cards is smaller and statistically insignificant for young consumers in rural areas (column 4) as well as for older consumers (columns 2-3, 5-6).

What could explain that a substantial causal effect of contactless cards on payment choice is limited to young urban consumers? Previous studies suggest that young consumers are more likely to adopt new (financial) technologies due to lower resistance and greater ability to learn new technologies and a longer time horizon (see e.g. Yang and Ching, 2013). However, if affinity to new technology were the driving force in our case, we should observe a similar effect for all young consumers. After all, young consumers in rural areas display not only an identical level for the *Cash ratio* but also an identical time-trend as young consumers in urban areas. For the same reason, it seems unlikely that general changes in local payment infrastructure (e.g. self-checkouts in grocery stores) are the driver of our results. One potential driver may, however, be changes in payment infrastructure which are specific to contactless cards, i.e. the faster dissemination of NFC enabled terminals in urban areas. A further potential driver is a heightened awareness of the new payment technology and potential network effects among young urban consumers.

--- Insert Table 7 about here ---

6. Robustness tests

In accordance with our pre-analysis plan, we conduct a series of robustness tests. First, we replicate our main analysis from Table 4 applying alternative definitions of our primary outcome variables. The definitions and summary statistics of these alternative outcome variables as well as the corresponding regression results are provided in Appendix A4. We first alter our definition of *Cash ratio* to (i) omit credit card payments, (ii) include e-banking payments and (iii) focus only on domestic card transactions. These adjustments have no effect on the causal effect of contactless cards on cash use, qualitatively (Panel A, columns 1-3). We further alter our measures of *Cash withdrawal frequency* and *Cash withdrawal amount* to focus on ATM withdrawals only (columns 4-5) and on domestic transactions only (columns 6-7). Again, our baseline results of Table 4 are confirmed.

Second, we replicate regression equation [1] measuring the outcome variables not by calendar year, but from the month of November to the following month of October. This robustness test accounts for the fact that replacement debit cards are sent to clients 2 months prior to the expiry of their old card and can be used immediately after receipt. Appendix A5 presents regression estimates which confirm our baseline results from Table 4.

Third, we replicate our subsample analysis of Table 6 employing an alternative definition of pre-treatment payment behavior. Specifically, we separate clients according to their pre-treatment number of debit card transactions below 20 CHF. Again, our results are confirmed (see Appendix A6).

Next, we report on a placebo test to disentangle the effect of a new payment card per se from the effect of receiving a payment card with a contactless function. To this end we exploit the fact that our control group (*Non adopters*) receive a new payment card at the end of 2015 (valid from beginning 2016) but this card does not yet feature the contactless technology (see Figure 1). Our placebo test therefore compares the payment behavior of *Non adopters* to early and *Late adopters* over the period 2015:01 to 2016:12.

$$[4] \qquad Y_{i,t} = \beta_i + \beta_t + \tau_{placebo} \cdot New\ card_{i,2016} + \varepsilon_{i,t}$$

where

$$Y_{i,t} \begin{Bmatrix} Debit\ card\ use_{i,t}, Cash\ ratio_{i,t}, Cash\ withdrawal\ frequency_{i,t}, \\ Cash\ withdrawal\ amount_{i,t} \end{Bmatrix} \text{and } t \in \{2015, 2016\}.$$

,

In this regression β_i , β_t are individual and time fixed effects respectively. $New\ card_{i,2016}$ is set to 1 for all individuals *i* of *Non adopters* in year 2016 (and 0 otherwise). Table A7 summarize the respective findings for the number of debit card transactions and the results suggest that *Non adopters* decrease rather than increase their use of debit cards after receipt of a new card. The respective results for our primary outcome variables are shown in Table A8. Reassuringly, the estimate of *New card* is insignificant in all specifications.

Finally, we replicate our analysis with a sample of clients which hold multiple debit cards. In this sample, we define treatment at the card level and not at the account level because expiry dates of cards might differ. Therefore, we can only conduct the analysis for the number of debit card transactions but cannot compute *Cash ratio* or withdrawal variables, which would require aggregation at the account level. Moreover, the number of observations (cards) in this sample is just 1,412 which limits the statistical power of our analysis. The respective results in Appendix A9 confirm, nevertheless, that small value card transactions strongly increase after the receipt of a contactless card.

7. Discussion

We study the causal effect of a payment technology innovation on payment choice and cash demand. We examine the staggered introduction of contactless debit cards in Switzerland over the period 2016-2018. We thus focus on an economy with a high level of financial development and a well-established payment infrastructure. Yet, like in many other European economies, Swiss consumers are strikingly cash intensive in their payment behavior. Studying how financial innovation affects payment behavior and money demand in cash intensive, advanced economies is important. The future use of cash as opposed to electronic private money, and hence the future design of the monetary system, will arguably be strongly influenced by these economies.[36]

Our analysis is based on account-level, administrative data for over 21,000 retail bank clients. The date at which these clients receive a contactless debit card the first time depends only on the expiry

[36] As a case in point, the Euro area, Japan and Switzerland account for roughly 40% of world currency in circulation. The card-intensive economies Australia, Canada, the UK, Sweden and Norway account for about 4% (own calculations). Even if we abstract from currency which is circulating abroad, the quantitative difference is large.

date of their previous card. Our results show that the introduction of contactless debit cards causes a strong increase in the use of debit cards at PoS. The impact on the cash-share of payments is weaker as the level of debit card payments is initially low and most additional debit cards payments are of small value. We find no effect of contactless cards on cash demand as measured by the frequency and average size of cash withdrawals.

Overall, our results document statistically significant effects of payment innovation on payment choice, but the economic magnitude of these effects are small. By comparison, our data reveal a strong decline in the use of cash which in descriptive analyses may be confused for a causal impact of recent payment innovations. This highlights the importance of disentangling causal effects of payment innovations from overarching trends in payment behavior

While the average treatment effect of contactless cards is underwhelming, our subsample analyses reveal substantial and informative heterogeneities across households: the impact of contactless cards is strongest among consumers with an intermediate cash share of payments. By contrast, the impact is negligible among extensive margin "cash lovers". Explorative analyses reveal that the impact of contactless cards on payment choice is largely driven by young consumers, but only those in urban locations. The latter finding suggests that recent payment innovations may accelerate the trend towards cashless transactions among technology affine consumers in locations with dense networks for cashless payments. By contrast, digital payment innovations may not trigger a widespread jump to a cashless society – at least in presently cash-intensive advanced economies.

Our findings speak to – and qualify – recent inventory theories of money demand which jointly model payment choice and cash demand (Alvarez and Lippi, 2017). First, our data reveal that a financial innovation may impact differently on payment choice and cash demand. While payment choice reacts to payment innovations, the frequency and average amount of cash withdrawals does not. In cash-intensive economies, even a strong increase in cashless payments - especially for small value transactions – has a limited impact on aggregate cash demand. We suspect that this low sensitivity of cash demand is related to the exceptionally low interest rates. Second, our results reveal significant and persistent heterogeneities in payment choice across consumers which can hardly be explained by variation in local payment infrastructure and corresponding transaction costs. Thus, it appears that habit and / or behavioral motives may exert a stronger impact on payment choice and cash demand than is typically assumed in inventory models.

References

Alvarez, F. and Lippi, F. (2009). Financial innovation and the transactions demand for cash. *Econometrica*, 77(2), 363-402.

Alvarez, F. and Lippi, F. (2017). Cash Burns: An Inventory Model with a Cash-Credit Choice. *Journal of Monetary Economics*, 90 (October), Pages 99-112.

Alvarez, F., Lippi, F. and Robatto, R. (2019). Cost of Inflation in Inventory Theoretical Models. *Review of Economic Dynamics*, 32, 206-226.

Arango, C., Huynh, K. P. and Sabetti, L. (2015). Consumer Payment Choice: Merchant Card Acceptance versus Pricing Incentives, *Journal of Banking and Finance*, 55, 130–141.

Athey, S. and Imbens, G. W. (2018) Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption. NBER Working Paper No. 24963.

Attanasio, O. R., Guiso, L. and Jappelli, T. (2002) The demand for money, financial innovation, and the welfare cost of inflation: An analysis with household data. *Journal of Political Economy*, 110, 317–351.

Bagnall, J., Bounie, D., Huynh, K. P., Kosse, A., Schmidt, T., Schuh, S. and Stix, H. (2016). Consumer Cash Usage: A Cross-Country Comparison with Payment Diary Survey Data. *International Journal of Central Banking*, 12(4), 1-61.

Baumol, W. J. (1952). The transactions demand for cash: An inventory theoretic approach. *The Quarterly Journal of Economics*, 66(4), 545-556.

Bech, M. L., Faruqui, U., Ougaard, F. and Picillo, C. (2018). Payments are a-changin' but cash still rules. *BIS Quarterly Review*, March 2018. Available at: https://www.bis.org/publ/qtrpdf/r_qt1803g.htm.

Bertrand, M., E. Duflo, and Mullainathan, S. (2014). How Much Should We Trust Differences-In-Differences Estimates?, *The Quarterly Journal of Economics*, 119(1), 249–275.

Bindseil, Ulrich (2020). Tiered CBDC and the Financial System, ECB Working Paper No. 2351.

Bounie, D. and Camara, Y. (2019). Card-Sales Response to Merchant Contactless Payment Acceptance: Causal Evidence. Available at SSRN: http://dx.doi.org/10.2139/ssrn.3459419.

Borzekowski, R. and Kiser, E.K. (2008). The choice at the checkout: quantifying demand across payment instruments. *International Journal of Industrial Organization*, 26 (4), 889-902.

Brancatelli, C. (2019). Preferences for Cash vs. Card Payments: An Analysis using German Household Scanner Data. Mimeo.

Briglevics, T. and Schuh, S. (2014). This Is What's in Your Wallet…and Here's How You Use It, Federal Reserve Bank of Boston No 14-05.

Brown, M., Guin, B. and Morkoetter, S. (2020). Deposit Withdrawals from Distressed Banks: Client relationships matter. *Journal of Financial Stability* 46, Article 100707.

Brunnermeier, M. K., H. James and Landau, J.-P. (2019). The Digitalization of Money. Working Paper.

Brunnermeier, M. K. and Niepelt, D. (2020). On the Equivalence of Private and Public Money. Forthcoming *Journal of Monetary Economics*.

Burlig, F. (2018). Improving transparency in observational social science research: A pre-analysis plan approach. *Economics Letters*, 168, 56–60.

Casey, K., Glennerster, R. and Miguel, E. (2012). Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan, *The Quarterly Journal of Economics*, 127(4), pages 1755-1812.

Chen, H., Felt, M-H. and Hunyh, K.P. (2017). Retail payment innovations and cash usage: accounting for attrition by using refreshment samples. *Journal of the Royal Statistical Society,* 180, 503-530.

Ching, A. T. and Hayashi, F. (2010). Payment card rewards programs and consumer payment choice. *Journal of Banking & Finance*, 34(8), 1773-1787.

Cohen, M., Rysman, M. and Wozniak, K. (2018). Payment choice with consumer panel data. Mimeo.

Doyle, M.-A., Fisher, C., Telez, E. and Yaday, A. (2017). How Australians Pay: Evidence from the 2016 Consumer Payments Survey. Reserve Bank of Australia Research Discussion Paper 2017-04.

Engert, W., Fung, B. S. C. and Segendorf, B. (2019). A Tale of Two Countries: Cash Demand in Canada and Sweden. Bank of Canada Staff Discussion Paper 2019-7.

Esselink, H. and Hernández, L. (2017). The use of cash by households in the euro area. ECB. Occasional Paper Series 201.

Finlay, R., Staib A. and Wakefield, M. (2018). Where's the Money? An Investigation into the Wereabouts and Uses of Australian Banknotes. Reserve Bank of Australia. Research Discussion Paper 2018-12.

Friedman, B. M. (2000). Decoupling at the Margin: The Threat to Monetary Policy from the Electronic Revolution in Banking, *International Finance,* 3(2), Juli, 261-72.

Henry, C. S., Huynh, K. P. and Welte, A. (2018). 2017 Methods-of-Payment Survey Report. Bank of Canada Staff Discussion Paper 2018-17.

Huynh, K. P., Schmidt-Dengler, P. and Stix, H. (2014). The Role of Card Acceptance in the Transaction Demand for Money. Oesterreichische Nationalbank Working Paper No. 196.

Kahn, C. McAndrews, J. and Roberds, W. (2005). Money Is Privacy. *International Economic Review*, *46* (2), 377-399.

Klee, E. (2008) How People Pay: Evidence from Grocery Store Data. *Journal of Monetary Economics*, 55, 526–41.

Kosse, A., Chen H., Felt, M.-H., Jiongo, V. D., Nield, K. and Welte, A. (2017). The Costs of Point-of-Sale Payments in Canada. Bank of Canada Staff Discussion Papers 2017-4.

Koulayev, S., Rysman, M., Schuh, S. and Stavins, J. (2016). Explaining adoption and use of payment instruments by US consumers. *The RAND Journal of Economics*, 47(2), 293-325.

Lippi, F. and Secchi, A. (2009). Technological change and the households' demand for currency. *Journal of Monetary Economics*, 56(2), 222-230.

Magnac, T. (2017). ATM foreign fees and cash withdrawals, *Journal of Banking & Finance*, 78(C), 117-129.

Mishkin, F. (1999). International Experiences International Experiences with Different Monetary Policy Regimes, NBER Working Paper No. w6965.

Olken, B. A. (2015). Promises and Perils of Pre-analysis Plans. *Journal of Economic Perspectives* 29 (3), 61-80.

Polasik, M. M., Górka, J., Wilczewski, G., Kunkowski, J. and Przenajkowska, K. (2010). Time Efficiency Of Point-Of-Sale Payment Methods: Preliminary Results, *Journal of Internet Banking and Commerce*, 15(3), 277-287.

Schilling, L. and Uhlig, H. (2019). Some Simple Bitcoin Economics, *Journal of Monetary Economics*, 106, 16-26.

Schuh, S. and Stavins, J. (2010). Why are (some) consumers (finally) writing fewer checks? The role of payment characteristics, *Journal of Banking & Finance*, 34(8), 1745-1758.

SNB (2018). Survey on payment methods 2017. Swiss National Bank.

Stavins, J. (2017). How do consumers make their payment choices? Federal Reserve Bank of Boston Research Data Report 17-1.

Sveriges Riksbank (2017). The Riksbank's e-krona project. Report 1. Available at: https://www.riksbank.se/globalassets/media/rapporter/e-krona/2017/rapport_ekrona_uppdaterad_170920_eng.pdf

Tobin, J. (1956). The interest-elasticity of transactions demand for cash. *The Review of Economics and Statistics*, 38(3), 241-247.

Trütsch, T. (2016).The impact of mobile payment on payment choice. *Financial Markets and Portfolio Management*, 30. 299-336.

van der Cruijsen, C., Hernandez, L., and Jonker, N. (2017). In love with the debit card but still married to cash. *Applied Economics*, 49(30), 2989-3004.

von Kalckreuth, U., Schmidt, T. and Stix, H. (2014). Using cash to monitor liquidity: implications for payments, currency demand, and withdrawal behavior. *Journal of Money, Credit and Banking*, 46(8), 1753-1786.

Wakamori, N. and Welte, A. (2017). Why do shoppers use cash? Evidence from shopping diary data. *Journal of Money, Credit and Banking*, 49(1), 115-169.

Wang, Z. and Wolman, A. (2016), Payment choice and currency use: Insights from two billion retail transactions, *Journal of Monetary Economics*, 84(issue C), 94-115.

Whitesell, W. 1989. The demand for currency versus debitable account. *Journal of Money Credit & Banking*, 21(2), 246–251.

Woodford, M. (2000). Monetary Policy in a World Without Money. *International Finance*, 3(2), 229-260.

Yang, B. and Ching, A. T. (2013). Dynamics of consumer adoption of financial innovation: The case of ATM cards. *Management Science*, 60(4), 903-922.

Figure 1. Research Design

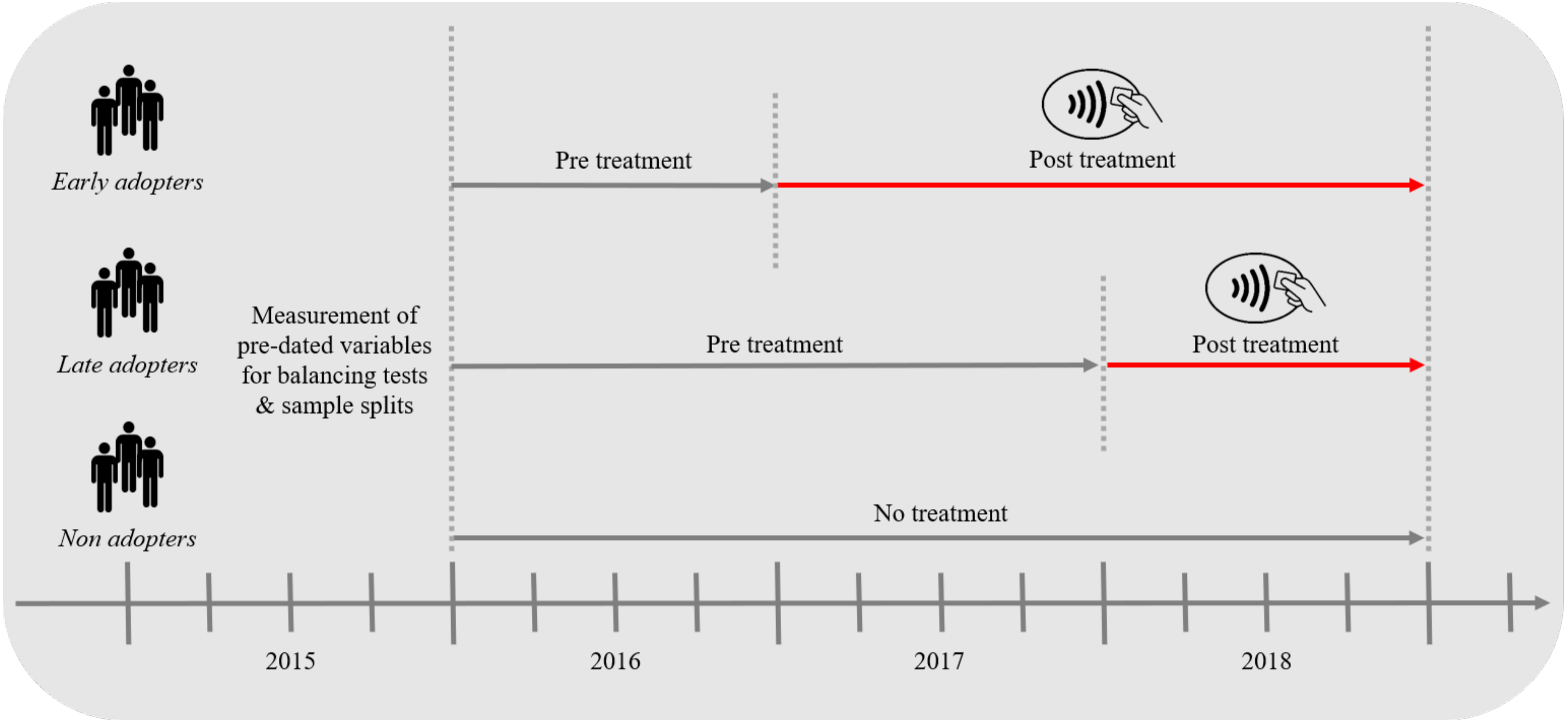
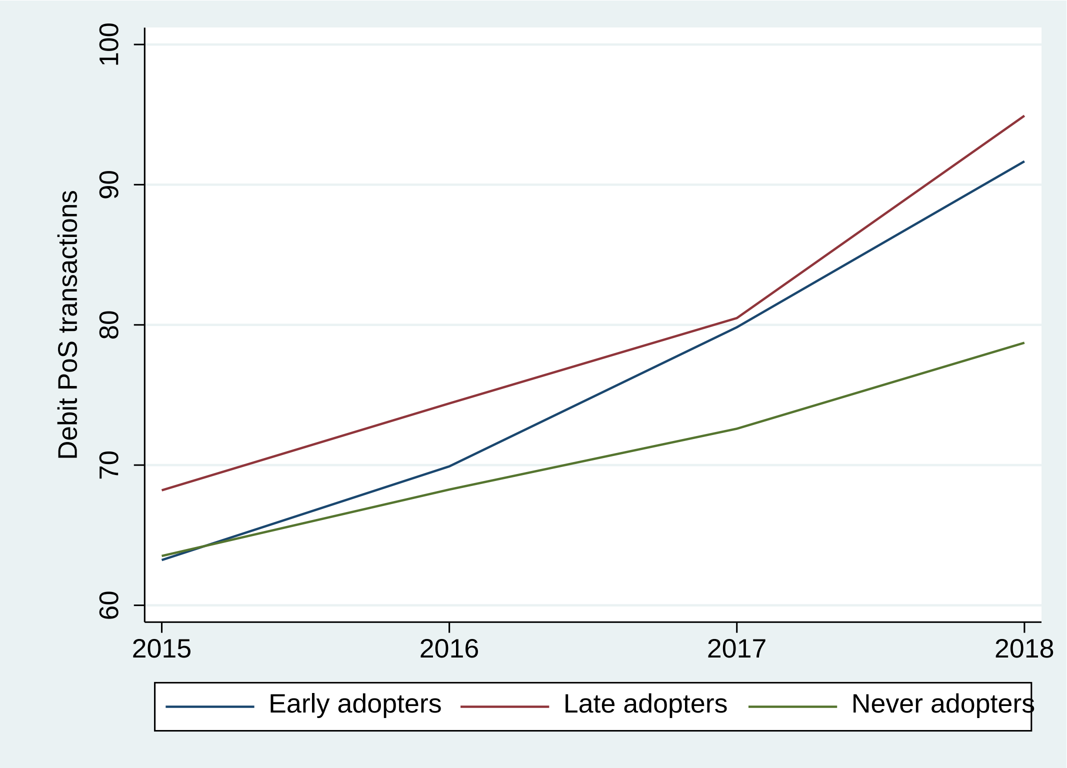


Figure 2. Debit Card PoS Transactions

This figure displays the average number of Point of Sale (PoS) transactions conducted by debit card per client and year by treatment group. Panel A displays the total number of PoS debit card transactions. Panel B displays the number of transactions with a value of at most 20 CHF. Appendix A2 presents definitions of all variables. Table 1 presents pre-treatment (2015) summary statistics.

Panel A- Total number of transactions



Panel B. Transactions below 20 CHF only

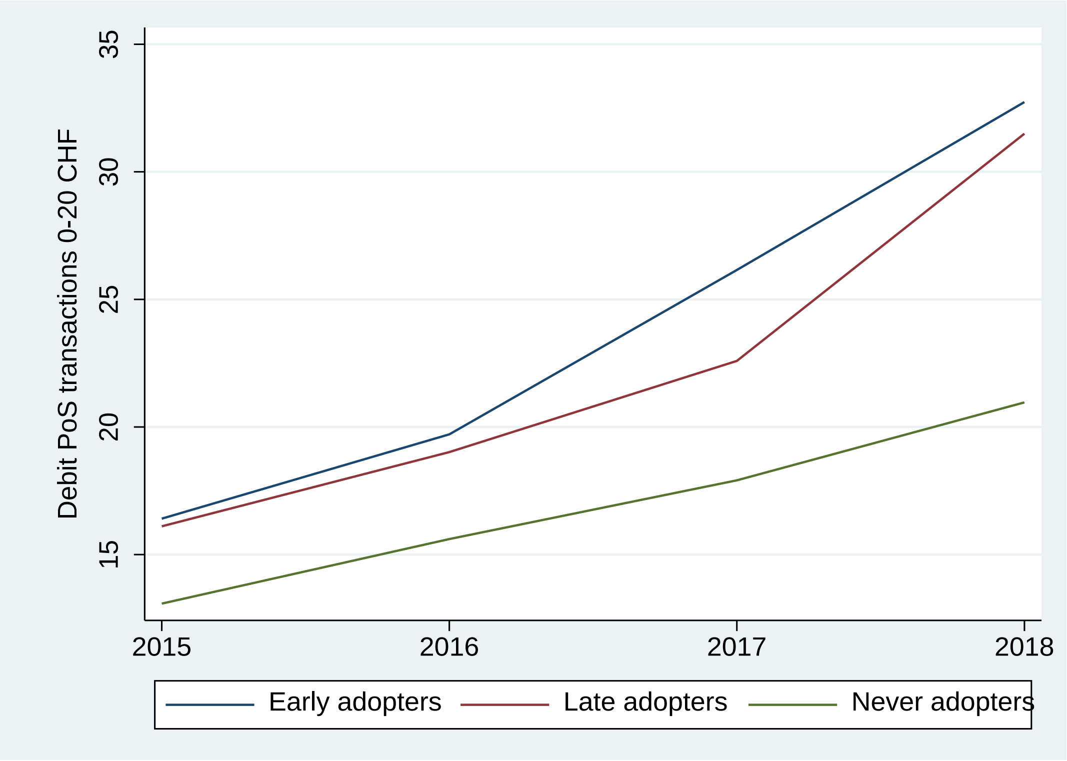
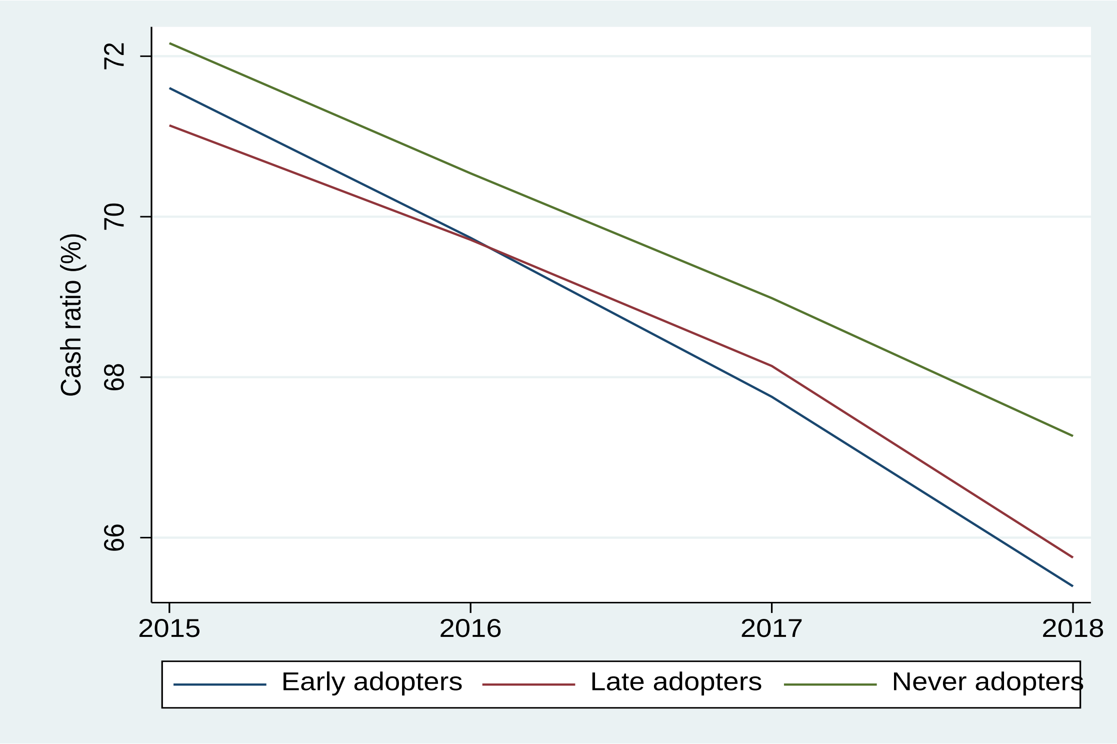


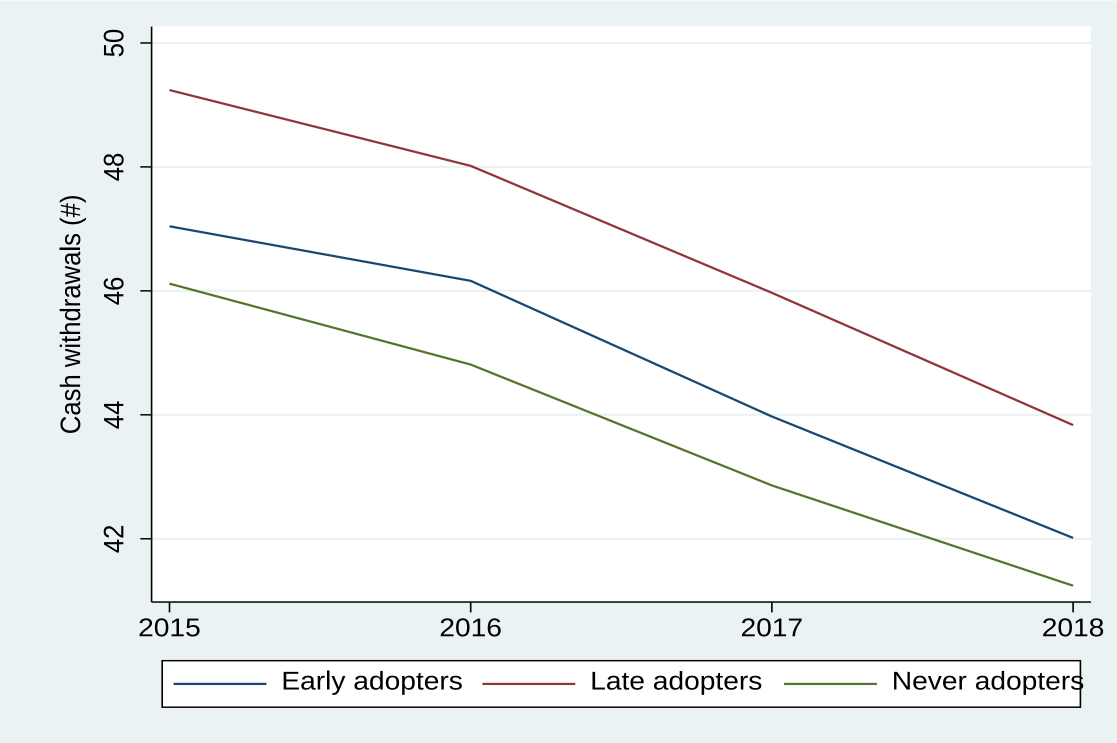
Figure 3. Payment choice and Cash demand

This figure displays the payment choice and cash demand per client and year by treatment group. Panel A displays the cash ratio of payments in %. Panel B displays the number of cash withdrawals. Panel C displays the average size of cash withdrawals in CHF. Appendix A2 presents definitions of all variables. Table 1 presents pre-treatment (2015) summary statistics.

Panel A- Cash ratio (%)



Panel B. Cash withdrawal frequency



Panel C. Cash withdrawal amount

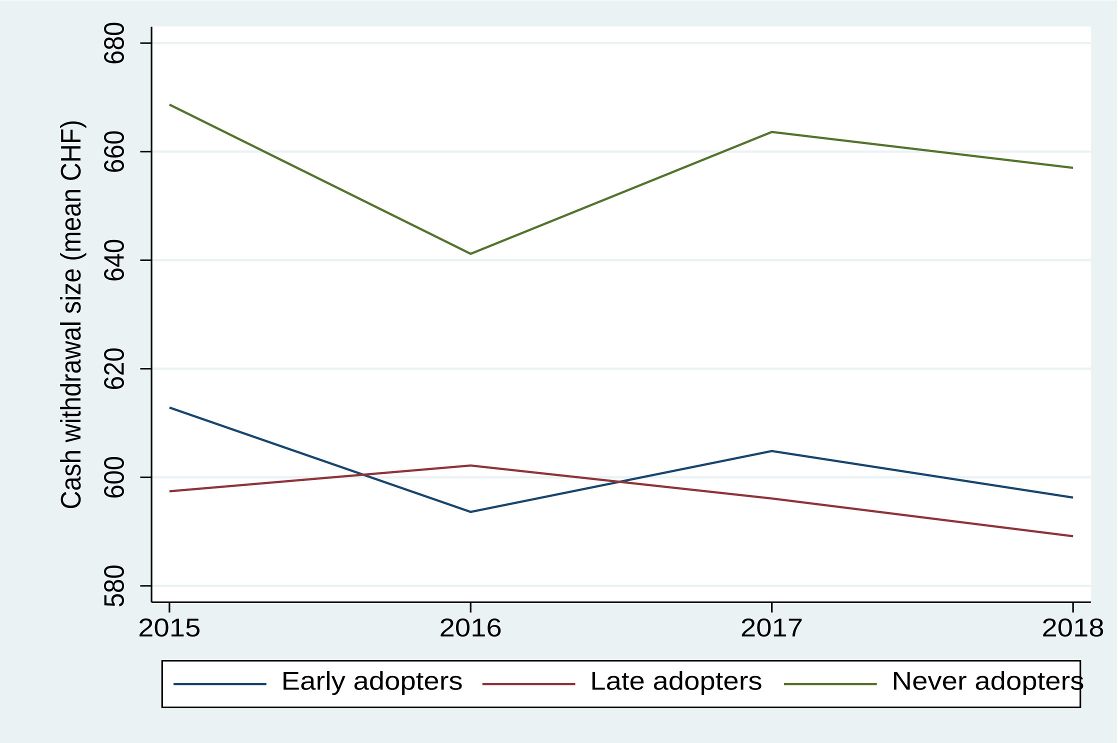


Table 1. Outcome Variables

The table presents descriptive statistics for our main and auxiliary outcome variables as measured in 2015 (pre-treatment). Panel A displays detailed summary statistics for all variables. Panel B displays comparisons of sample means by treatment group. Variable definitions are presented in Appendix Table A2.

Panel A. Summary Statistics (Pre-treatment = 2015)

	mean	min	p25	p50	p75	max	n
Main Outcome Variables							
Cash ratio (%)	71.6	0	52	78	96	100	21'122
Cash withdrawal frequency	47.4	0	20	39	64	594	21'122
Cash withdrawal amount	625	20	189	344	677	25'000	20'992
Auxillary Outcome Variables							
Debit PoS transactions	64.8	0	6	36	95	909	21'122
Debit PoS transactions (0-20 CHF)	15.3	0	0	2	15	633	21'122
Debit PoS transactions (20-40 CHF)	14.4	0	0	5	19	288	21'122
Debit PoS transactions (40-60 CHF)	10.7	0	0	5	15	178	21'122
Debit PoS transactions (60-100 CHF)	12.3	0	1	6	18	278	21'122
Debit PoS transactions (>100 CHF)	12.1	0	1	6	16	195	21'122

Panel B. Sample Means by Treatment Group (Pre-treatment = 2015)

* (**) indicate significance levels of T-tests at the 5%-level (1%-level), respectively.

	Early adopters	Late Adopters	Non adopters		T-tests	
	[1]	[2]	[3]	[1 vs. 2]	[1 vs. 3]	[2 vs. 3]
Main Outcome Variables						
Cash ratio (%)	71.6	71.1	72.2			*
Cash withdrawal frequency	47.0	49.2	46.1	**	**	
Cash withdrawal amount	613	597	669		**	**
Auxillary Outcome Variables						
Debit PoS transactions	63.2	68.2	63.5	**		**
Debit PoS transactions (0-20 CHF)	16.4	16.1	13.1		**	**
Debit PoS transactions (20-40 CHF)	13.9	15.1	14.4	**		
Debit PoS transactions (40-60 CHF)	10.1	11.3	10.9	**	**	
Debit PoS transactions (60-100 CHF)	11.5	13.0	12.7	**	**	
Debit PoS transactions (>100 CHF)	11.3	12.8	12.5	**	**	

Table 2. Covariate Variables

The table presents descriptive statistics for our client-level and account-level covariates as measured in 2015 (pre-treatment). Panel A displays detailed summary statistics for all variables. Panel B displays comparisons of sample means by treatment group. Variable definitions are presented in Appendix Table A2.

Panel A. Summary Statistics (Pre-treatment = 2015)

	mean	min	p25	p50	p75	max	n
Client-level Variables							
Age	3.52	1	2	4	5	6	21'122
Male	0.51	0	0	1	1	1	21'122
Nationality Swiss	0.71	0	0	1	1	1	21'122
Size municipality	2.63	1	2	2	3	5	21'122
Income	2.62	1	1	2	4	6	21'122
Wealth	2.02	1	1	2	3	6	21'122
Retirement account	0.53	0	0	1	1	1	21'122
Savings account	0.22	0	0	0	0	1	21'122
Custody account	0.19	0	0	0	0	1	21'122
Mortgage	0.07	0	0	0	0	1	21'122
Ebanking	0.54	0	0	1	1	1	21'122
Account-level Variables							
Account opening year	1998	1972	1990	2000	2008	2014	21'122
Direct debiting	0.55	0	0	1	1	1	21'122
Standing order Ebanking	0.15	0	0	0	0	1	21'122
Standing order paper	0.36	0	0	0	1	1	21'122
Ebanking payments	19'335	0	0	0	30'227	435'745	21'122
Transfers	3'938	0	0	0	400	420'000	21'122
Incoming payments	58'663	1'200	28'413	53'169	76'518	471'408	21'122
Outgoing payments	64'466	1'206	30'862	56'377	82'371	499'429	21'122
Account balance	3.4	1	1	3	6	6	21'122

Panel B. Sample Means by Treatment Group (Pre-treatment = 2015)

* (**) indicate significance levels of T-tests at the 5%-level (1%-level), respectively.

	Early adopters	Late Adopters	Non adopters	T-tests		
	[1]	[2]	[3]	[1 vs. 2]	[1 vs. 3]	[2 vs. 3]
Client-level Variables						
Age	3.41	3.49	3.68	**	**	**
Male	0.51	0.53	0.50			**
Nationality Swiss	0.72	0.70	0.71	**	*	
Size municipality	2.64	2.64	2.61			
Income	2.53	2.71	2.64	**	**	*
Wealth	2.03	1.98	2.05	*		**
Retirement account	0.54	0.53	0.52		*	
Savings account	0.21	0.23	0.23	**	**	
Custody account	0.19	0.18	0.21		**	**
Mortgage	0.07	0.07	0.08			
Ebanking	0.54	0.55	0.52		**	**
Account-level Variables						
Account opening year	1998	1999	1997	**	**	**
Direct debiting	0.54	0.56	0.55	**		
Standing order Ebanking	0.15	0.17	0.15	**		**
Standing order paper	0.35	0.36	0.38		**	
Ebanking payments	18'493	20'428	19'401	**		
Transfers	3'632	4'293	4'000	**		
Incoming payments	56'351	60'366	60'073	**	**	
Outgoing payments	61'858	66'614	65'842	**	**	
Account balance	3.42	3.34	3.42	*		*

Table 3. Debit PoS transactions

The table shows the results of an OLS regression. The dependent variables measure the number of debit PoS transactions per client and year. In column (1) the dependent variable covers all transactions, in columns (2-6) the dependent variable covers transactions of specific values only (0-20 CHF, 20-40 CHF, 40-60 CHF, 60-100 CHF, 100+ CHF). Each regression includes 3 annual observations (2016, 2017, 2018) for 21'122 clients. The explanatory variable *Contacless* is 1 for early adopters in years 2017 and 2018 and for late adopters in year 2018. All regressions include client fixed effects. Robust standard errors are reported in parentheses. *, **,*** denote significance at the 0.05, 0.01, and 0.001-level.

	(1)	(2)	(3)	(4)	(5)	(6)
		Debit card PoS transactions by transaction value				
Outcome variable	All	below 20 CHF	20-40 CHF	40-60 CHF	60 - 100 CHF	above 100 CHF
Contactless	6.786***	4.888***	1.092***	0.322***	0.242**	0.241**
	(0.506)	(0.316)	(0.140)	(0.087)	(0.091)	(0.087)
Year = 2017	4.365***	2.371***	1.061***	0.221***	0.638***	0.074
	(0.323)	(0.186)	(0.096)	(0.063)	(0.065)	(0.064)
Year = 2018	13.227***	7.122***	3.033***	0.797***	1.593***	0.681***
	(0.493)	(0.288)	(0.142)	(0.090)	(0.094)	(0.090)
Client fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Clients	21'122	21'122	21'122	21'122	21'122	21'122
Client * Year observations	63'366	63'366	63'366	63'366	63'366	63'366
Mean of dependent variable	79.05	23.20	17.85	11.97	13.30	12.73
Method	OLS	OLS	OLS	OLS	OLS	OLS

Table 4. Payment choice and cash demand: Average treatment effect

The table shows the results of an OLS regression. The dependent variables measure payment choice and cash demand per client and year. In columns (1-2) the dependent variable is *Cash ratio* , in columns (3-4) *Cash withdrawals frequency* , in columns (5-6) *Cash withdrawal amount* . Appendix A2 presents definitions of each variable. Each regression includes 3 annual observations (2016, 2017, 2018) per client. The explanatory variable *Contactless* is 1 for early adopters in years 2017 and 2018 and for late adopters in year 2018. All regressions include client fixed effects. Columns (1,3,5) include year fixed effects. Columns (2,4,6) include year*location fixed effects. We distinguish 22 locations based on a combination of the local economic region (MS-region) and the size of the municipality within that region that the client resides in. Robust standard errors are reported in parentheses. *, **,*** denote significance at the 0.017, 0.01, and 0.001-level.

	(1)	(2)	(3)	(4)	(5)	(6)
Outcome variable	Cash ratio (%)		Cash withdrawal frequency		Cash withdrawal amount	
Contactless	-0.581***	-0.574***	-0.362	-0.346	-1.138	-0.944
	(0.144)	(0.144)	(0.169)	(0.169)	(7.602)	(7.685)
Year = 2017	-1.496***		-1.929***		8.943	
	(0.104)		(0.122)		(5.611)	
Year = 2018	-3.518***		-3.729***		4.262	
	(0.143)		(0.168)		(7.081)	
Client fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year*Location fixed effects	No	Yes	No	Yes	No	Yes
Clients	21'112	21'112	21'122	21'122	21'047	21'047
Client * Year observations	63'169	63'169	63'366	63'366	62'544	62'544
Mean of dependent variable	68.10	68.10	44.27	44.27	614.62	614.62
Method	OLS	OLS	OLS	OLS	OLS	OLS

Table 5. Payment choice and cash demand: Dynamic treatment effect

The table shows the results of an OLS regression. The dependent variables measure payment choice and cash demand per client and year. In columns (1-2) the dependent variable is *Cash ratio* , in columns (3-4) *Cash withdrawal frequency* , in columns (5-6) *Cash withdrawal amount* . Appendix A2 presents definitions of each variable. Each regression includes 3 annual observations (2016, 2017, 2018) per client. The explanatory variables are Early adopter in 2017, Early adopter in 2018 and Late adopter in 2018. In addition we report the estimate for Late adopter in 2017 as an anticipation / placebo effect. All regressions include client fixed effects. Columns (1,3,5) include year fixed effects, columns (2,4,6) include year*location fixed effects. We distinguish 22 locations based on a combination of the local economic region (MS-region) and the size of the municipality within that region that the client resides in. Robust standard errors are reported in parentheses. *, **, *** denote significance at the 0.017, 0.01, and 0.001-level.

	(1)	(2)	(3)	(4)	(5)	(6)
Outcome variable	Cash ratio (%)		Cash withdrawal frequency		Cash withdrawals amount	
Early adopter, 2017	-0.46	-0.451	-0.239	-0.221	-9.911	-10.44
	(0.214)	-0.215	(0.241)	(0.241)	(12.553)	-12.645
Early adopter, 2018	-1.128***	-1.124***	-0.58	-0.552	-8.63	-9.204
	-0.257	(0.257)	(0.306)	(0.306)	(12.073)	-12.198
Late adopter, 2017	0.006	0.001	-0.098	-0.079	-27.064	-28.113
	(0.225)	(0.225)	(0.265)	(0.265)	(12.111)	-12.116
Late adopter, 2018	-0.709**	-0.710**	-0.614	-0.58	-19.981	-20.045
	(0.271)	(0.271)	(0.339)	(0.339)	(12.545)	-12.536
Client fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	No	Yes	No	Yes	No
Year*Location fixed effects	No	Yes	No	Yes	No	Yes
Clients	21'112	21'112	21'122	21'122	21'047	21'047
Client * Year observations	63'169	63'169	63'366	63'366	62'544	62'544
Mean of dependent variable	68.10	68.10	44.27	44.27	614.62	614.62
Method	OLS	OLS	OLS	OLS	OLS	OLS

Table 6. Payment choice: By pre-treatment payment behavior

The table shows the results of an OLS regression for subsamples of clients based on their pre-treatment payment behavior. We split clients by quartile of *Cash ratio* (%) in 2015. The dependent variable is *Cash ratio* (%) in all columns. Appendix A2 presents definitions of each variable. Each regression includes 3 annual observations (2016, 2017, 2018) per client. The explanatory variable *Contactless* is 1 for early adopters in years 2017 and 2018 and for late adopters in year 2018. All regressions include client fixed effects. Robust standard errors are reported in parentheses. *, **, *** denote significance at the 0.017, 0.01 and 0.001-level.

	(1)	(2)	(3)	(4)
Outcome variable		Cash ratio (%)		
Cash ratio (%) in 2015 (subsample):	[0-52%]	(52%-78%]	(78%-96%]	(96%-100%]
Contactless	-0.172	-1.292***	-0.347	-0.343
	(0.333)	(0.326)	(0.276)	(0.191)
Year = 2017	-0.620*	-1.973***	-2.296***	-1.144***
	(0.244)	(0.240)	(0.202)	(0.129)
Year = 2018	-2.226***	-4.775***	-5.061***	-2.102***
	(0.329)	(0.325)	(0.289)	(0.183)
Client fixed effects	Yes	Yes	Yes	Yes
Year*Location fixed effects	No	No	No	No
Clients	5'278	5'278	5'280	5'276
Client * Year observations	15'801	15'805	15'820	15'743
Mean of dependent variable	35.56	59.60	81.07	96.24
Method	OLS	OLS	OLS	OLS

Table 7. Payment choice: By client location and age-group

This table shows the results of an OLS regression for subsamples of clients based on the population-density of their residential location and the clients age. We distinguish urban locations (columns 1-3) from rural locations, whereby locations of residence are categorized as urban (rural) if they have above (below) 3'000 inhabitants per km2 settlement area. The dependent variable is *Cash ratio* in all columns. Appendix A2 presents definitions of each variable. Each regression includes 3 annual observations (2016, 2017, 2018) per client. The explanatory variable *Contactless* is 1 for early adopters in years 2017 and 2018 and for late adopters in year 2018. All regressions include client fixed effects. Robust standard errors are reported in parentheses. *, **, *** denote significance at the 0.017, 0.01, and 0.001-level.

	(1)	(2)	-3	(4)	-5	(6)
Outcome variable	Cash ratio (%)					
Location	Urban			Rural		
Client age (years)	below 35	35-55	above 55	below 35	35-55	above 55
Contactless	-1.246**	-0.717	0.092	-0.390	-0.333	0.365**
	(0.411)	(0.303)	(0.348)	(0.396)	(0.307)	(0.364)
Year = 2017	-3.085***	-0.858***	-0.643*	-3.244***	-1.308***	-0.549***
	(0.301)	(0.217)	(0.259)	(0.300)	(0.217)	(0.265)
Year = 2018	-7.139***	-2.720***	-1.198***	-7.137***	-2.961***	-1.164***
	(0.428)	(0.294)	(0.346)	(0.411)	(0.305)	(0.347)
Client fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year*Location fixed effects	No	No	No	No	No	No
Clients	3'041	4'033	3'262	3'323	4'417	3'036
Client * Year observations	9'105	12'085	9'738	9'958	13'214	9'069
Mean of dependent variable	58.44	66.22	77.73	61.86	66.97	78.45
Method	OLS	OLS	OLS	OLS	OLS	OLS

Apppendix A1. Sample composition

Raw data sample	
Clients:	30'000
Accounts:	30'330
Debit cards:	33'165

Main sample (1 account, 1 card)		**Robustness sample (1 account, multiple cards)****		
Single account /single card:	26'934	Single account / multiple cards:	2'735	(5470 accounts)
with regular expiry date	24'021	with regular expiry date	2'582	(5164 accounts)
with account opened before 2015:	23'957	with account opened before 2015	2'576	(5152 accounts)
non-missing covariates:	22'504	non-missing covariates:	1'485	(2970 accounts)
No outlier turnover*:	21'122	No outlier turnover*:	1'396	(2792 accounts)
		Multiple expiry dates	706	(1412 accounts)
Final sample, # clients:	**21'122**	**Final sample, # clients:**	**706**	**(1412 accounts)**

* Outlier turnovers are defined as incoming /outgoing account flows below 1200 CHF or exceeding 500'000 CHF in any year.

** Our robustness sample includes only clients with 1 account and 2 debit cards. We drop 6 clients with 1 account and 3 debit cards.

Appendix A2. Definition of Variables

Panel A. Outcome variables

Variable	Definition	Unit	Range
	Main outcome variables		
Cash ratio	Cash withdrawals (ATM & Branch in CHF) / [Cash withdrawals (ATM & Branch in CHF) + Debit PoS transactions (in CHF) + Credit Card transactions (in CHF)], annual	%	[0,100]
Cash withdrawal frequency	Number of cash withdrawals (ATM & Branch), annual	number	>=0
Cash withdrawal amount	Cash withdrawals (ATM & Branch) in CHF / Cash withdrawals frequency	CHF	>0
	Auxillary outcome variables		
Variable	Definition	Unit	Range
Debit PoS transactions	Number of PoS transactions by debit card, annual		
Debit PoS transactions (0-20 CHF)	Number of PoS transactions with volume of (0,20] CHF by debit card, annual	number	>=0
Debit PoS transactions (20-40 CHF)	Number of PoS transactions with volume of (20,40] CHF by debit card, annual	number	>=0
Debit PoS transactions (40-60 CHF)	Number of PoS transactions with volume of (40,60] CHF by debit card, annual	number	>=0
Debit PoS transactions (60-100 CHF)	Number of PoS transactions with volume of (60,100] CHF by debit card, annual	number	>=0
Debit PoS transactions (>100 CHF)	Number of PoS transactions with volume of >100 CHF by debit card, annual	number	>=0

Panel B. Covariates

Client-level variables

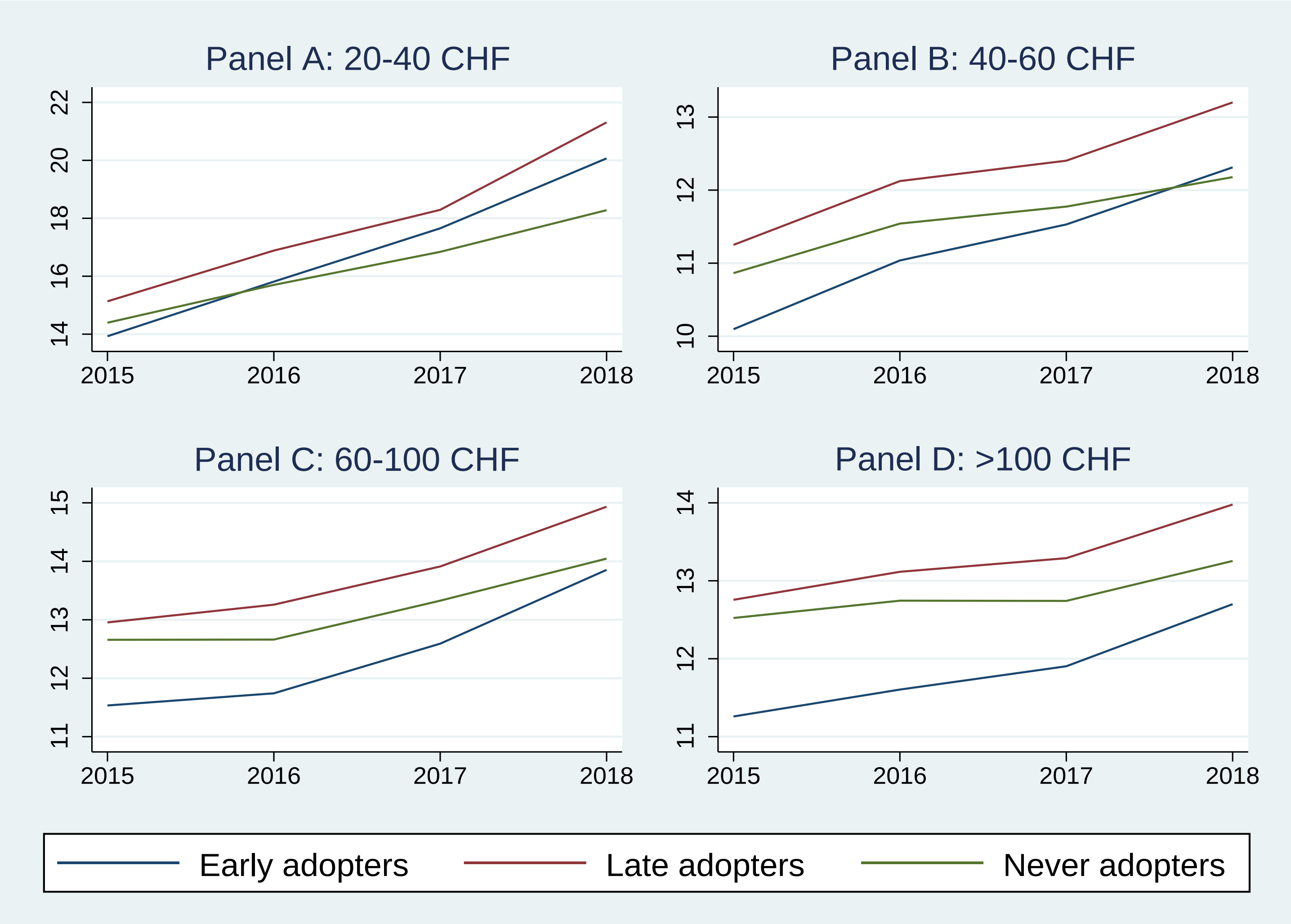
Variable	Definition	Unit
Age	Age of client in years: 1=25 or younger; 2=26-35; 3=36-45; 4=46-55; 5=56-65; 6= 66 and older	[1;..;6]
Male	Gender of client: 1=male; 0=female.	[0;1]
Nationality Swiss	Nationality of client 1=Swiss; 0=other nationality	[0;1]
Size municipality	Population of municipality in which client resides. 1= (0,5'000] ; 2=(5'000-10'000]; 3=(10'000-20'000]; 4=(20'000-50'000]; 5= more than 50'000	[1;..;5]
Income	Monthly income of client in CHF as estimated by the Bank in December 2015. 1 = [0,3'000]; 2= (1'000, 2'500]; 3= (2'500, 5'000]; 4= (5'000, 7'500]; 5= (7'500. 10'000]; 6= >10'000	[1;..;6]
Wealth	Total financial assets under management of the client with the Bank in December 2015 in CHF. 1 = [0,10'000]; 2= (10'000, 50'000]; 3= (50'000, 100'000]; 4= (100'000, 250'000]; 5= (250'000, 1'000'000]; 6=more than 1'000'000.	[1;..;6]
Retirement account	Dummy variable = 1 if client has a voluntary retirement savings account with the Bank, 0=otherwise	%
Savings account	Dummy variable = 1 if client has an ordinary savings account with the Bank, 0=otherwise	number
Custody account	Dummy variable = 1 if client has a custody account for securities with the Bank, 0=otherwise	
Mortgage	Dummy variable = 1 if client has a mortgage with the Bank, 0=otherwise	
E-banking	Dummy variable = 1 if client has an Ebanking contract with the Bank, 0=otherwise	

Account-level variables (measured in 2015)

Variable	Definition	Unit
Account opening year	Year in which account was opened	Year
Direct debiting	Dummy variable = 1 if client uses direct debiting with this account, 0=otherwise	[0;1]
Standing order Ebanking	Dummy variable = 1 if client uses Ebanking standing orders with this account, 0=otherwise	[0;1]
Standing order paper	Dummy variable = 1 if client uses ordinary standing orders with this account, 0=otherwise	[0;1]
Ebanking payments	Volume of outgoing Ebanking transactions in CHF, 2015	CHF
Transfers	Volume of outgoing account transfers in CHF, 2015	CHF
Incoming payments	Total volume of incoming payments in CHF, 2015	CHF
Outgoing payments	Total volume of outgoing payments in CHF, 2015	CHF
Account balance	Account balance in CHF as per end December 2015. 1 = [0,1'000]; 2= (1'000, 2'500]; 3= (2'500, 5'000]; 4= (5'000, 7'500]; 5= (7'500. 10'000]; 6=more than 10'000	[1;..;6]

Appendix A3 Debit Card PoS Transactions

The figure displays the average number of Point of Sale (PoS) transactions conducted by debit card per client and year by treatment group. Panel A displays the number of transactions with a value of (20-40] CHF. Panel B displays the number of transactions with a value of (40-60] CHF. Panel C displays the number of transactions with a value of (60-100] CHF. Panel A displays the number of transactions with a value of >100 CHF.



Appendix A4. Payment choice and cash demand: Alternative outcome variables

Panel A. Average treatment effect estimates

Panel A shows the results of OLS regressions, where the dependent variables are alternative indicators of payment choice and cash demand per client. Panel B presents definitions of each variable. Panel C presents (pre-treatment) summary statistics for each variable. Each regression includes 3 annual observations (2016, 2017, 2018) per client. The explanatory variable *Contactless* is 1 for early adopters in years 2017 and 2018 and for late adopters in year 2018. All regressions include client fixed effects. Robust standard errors are reported in parentheses. *, **, *** denote significance at the 0.017, 0.01 and 0.001-level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Outcome variable	Cash ratio without credit (%)	Cash ratio with Ebanking (%)	Cash ratio - domestic (%)	Cash withdrawal frequency - ATM	Cash withdrawals amount - ATM	Cash withdrawal number - domestic	Cash withdrawal amount - domestic
Contactless	-0.514***	-0.427**	-0.550***	-0.420*	-0.188	-0.367	-1.908
	(0.143)	(0.143)	(0.154)	(0.168)	(1.906)	(0.155)	(8.043)
Year = 2017	-1.192***	-2.251***	-1.322***	-1.736***	3.927**	-1.871***	7.537
	(0.104)	(0.103)	(0.112)	(0.121)	(1.369)	(0.112)	(6.149)
Year = 2018	-2.944***	-4.778***	-3.234***	-3.354***	8.072***	-3.532***	3.281
	(0.142)	(0.148)	(0.153)	(0.167)	(1.970)	(0.156)	(7.466)
Client fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Region*Year fixed effects	No	No	No	No	No	No	No
Clients	21'096	21'118	21'079	21'122	20'341	21'122	21'000
Client * Year observations	63'036	63'289	62'911	63'366	59'810	63'366	62'172
Mean of dependent variable	71.60	50.70	70.30	41.17	360.96	38.80	638.00
Method	OLS	OLS	OLS	OLS	OLS	OLS	OLS

Panel B. Alternative outcome variables - Definitions

Variable	Definition	Unit	Range
Cash ratio without credit	Cash withdrawals (ATM & Branch, value) / [Cash withdrawals (ATM & Branch, value) + Debit PoS transactions (value) + Credit Card transactions (value)], annual	%	[0,100]
Cash ratio with ebanking	Cash withdrawals (ATM & Branch, value) / [Cash withdrawals (ATM & Branch, value) + Debit PoS transactions (value) + Credit Card transactions (value)+ Ebanking transactions (value)], annual	%	[0,100]
Cash ratio - domestic	*Cash ratio without credit* , calculated based on transactions in CHF in Switzerland only	%	[0,100]
Cash withdrawal frequency - ATM	Number of ATM withdrawals, annual	number	>=0
Cash withdrawal amount - ATM	ATM withdrawals (value) / Cash withdrawals - number	CHF	>0
Cash withdrawal frequency - domestic	*Cash withdrawals CHF - number* , calculated based on transactions in CHF in Switzerland only	number	>=0
Cash withdrawal amount - domestic	*Cash withdrawals - average size* , calculated based on transactions in CHF in Switzerland only	CHF	>0

Panel C. Summary Statistics (Pre-treatment = 2015)

	mean	min	p25	p50	p75	max	n
Cash ratio without credit	74.5	0	57	81	97	100	21'094
Cash ratio with Ebanking	55.5	0	20	56	93	100	21'122
Cash ratio - domestic	73.3	0	55	81	97	100	21'076
Cash withdrawal frequency - ATM	44.0	0	17	36	62	592	21'122
Cash withdrawal amount - ATM	358.5	20	163	270	438	5'000	20'031
Cash withdrawal frequency - domestic	42.0	0	17	34	57	594	21'122
Cash withdrawal amount - domestic	645.5	20	181	339	700	25'000	20'907

Appendix A5. Payment choice and cash demand: Alternative treatment period definition

The table shows the results of robustness tests with an alternative definition of treatment periods. In our main analysis we define treatment periods by calendar year (January - December). In this robustness test we define treatment periods from November to the following year October. This accounts for the fact that new debit cards are typically issued at end of October of the previous year. The dependent variables measure payment choice and cash demand per client and year. Appendix A2 presents definitions of each variable. Each regression includes 3 annual observations (2016, 2017, 2018) per client. The explanatory variable *Contactless* is 1 for early adopters in years 2017 and 2018 and for late adopters in year 2018. All regressions include client and year fixed effects. Robust standard errors are reported in parentheses. *, ** denote significance at the 0.017, and 0.001-level.

	(1)	(3)	(5)
Outcome variable	Cash ratio (%)	Cash withdrawal frequency	Cash withdrawal amount (CHF)
Contactless	-0.471***	-0.279	-13.932
	(0.142)	(0.170)	(13.151)
Year = 2017	-1.533***	-1.815***	29.663
	(0.105)	(0.124)	(17.312)
Year = 2018	-3.370***	-3.705***	24.213
	(0.143)	(0.171)	(15.391)
Client fixed effects	Yes	Yes	Yes
Region*Year fixed effects	No	No	No
Clients	20'928	20'934	20'861
Client * Year observations	62'634	62'802	62'058
Mean of dependent variable	68.70	44.90	621.60
Method	OLS	OLS	OLS

Table A6. Payment choice: By pre-treatment payment behavior

The table shows the results of an OLS regression for subsamples of clients based on their pre-treatment payment behavior. We split clients by quartile of PoS debit transactions (below 20 CHF) in 2015. The dependent variable is *Cash ratio* (%) in all columns. Appendix A2 presents definitions of each variable. Each regression includes 3 annual observations (2016, 2017, 2018) per client. The explanatory variable *Contactless* is 1 for early adopters in years 2017 and 2018 and for late adopters in year 2018. All regressions include client fixed effects. Robust standard errors are reported in parentheses. *, **, *** denote significance at the 0.017, 0.01 and 0.001-level.

	(1)	(2)	(3)	(4)
Outcome variable	Cash ratio (%)			
PoS debit transactions (below 20 CHF) in 2015 (subsample):	[0]	(1-2]	(3-15]	(16-633]
Contactless	-0.398	-0.286	-0.374	-0.931**
	(0.226)	(0.410)	(0.294)	(0.292)
Year = 2017	-0.885***	-1.717***	-1.505***	-2.395***
	(0.166)	(0.300)	(0.206)	(0.215)
Year = 2018	-1.851***	-3.954***	-4.109***	-5.368***
	(0.220)	(0.409)	(0.290)	(0.301)
Client fixed effects	Yes	Yes	Yes	Yes
Region*Year fixed effects	No	No	No	No
Clients	5278	3022	5287	5068
Client * Year observations	7'735	9'048	15'849	15'193
Mean of dependent variable	1.5	5.1	16.3	74.4
Method	OLS	OLS	OLS	OLS

Appendix A7. Debit PoS transactions - Placebo Test

The table shows the results of a placebo test with observations from year 2015 and 2016 only. The dependent variables measure the number of debit PoS transactions per client and year. Appendix A2 presents definitions of each variable. The explanatory variable *New card* is 1 for all cards which expire at end 2015 and thus receive a new card (albeit one without a contactless function) for 2016. All regressions include client and year fixed effects. Robust standard errors are reported in parentheses. *, **,*** denote significance at the 0.05, 0.01, and 0.001-level.

	(1)	(2)	(3)	(4)	(5)	(6)
			Debit card transactions by transaction value			
Outcome variable	All	below 20 CHF	20-40CHF	40-60 CHF	60 - 100 CHF	above 100 CHF
New card	-1.740***	-0.608*	-0.524***	-0.235*	-0.245*	-0.129
	(0.508)	(0.271)	(0.157)	(0.110)	(0.112)	(0.113)
Year = 2016	6.478***	3.138***	1.829***	0.913***	0.248***	0.351***
	(0.307)	(0.173)	(0.091)	(0.060)	(0.063)	(0.062)
Client fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Clients	21'122	21'122	21'122	21'122	21'122	21'122
Card * Year observations	42'244	42'244	42'244	42'244	42'244	42'244
Mean of dependent variable	67.70	16.80	15.30	11.10	12.40	12.20
Method	OLS	OLS	OLS	OLS	OLS	OLS

Appendix A8. Payment choice and cash demand: Placebo test

The table shows the results of a placebo test with observations from year 2015 and 2016 only. The dependent variables measure payment choice and cash demand per client and year. Appendix A2 presents definitions of each variable.The explanatory variable *New card* is 1 for all cards which expire at end 2015 and thus are replaced with a new card (albeit one without a contactless function) for 2016. All regressions include client and year fixed effects. Robust standard errors are reported in parentheses. *, ** denote significance at the 0.017, and 0.001-level.

	(1)	(3)	(5)
Outcome variable	Cash ratio (%)	Cash withdrawals frequency (#)	Cash withdrawal amount (CHF)
New card	0.06	-0.282	-16.65
	(0.188)	(0.226)	(9.809)
Year = 2016	-1.673***	-1.024***	-3.877
	(0.105)	(0.134)	(5.079)
Client fixed effects	Yes	Yes	Yes
Region*Year fixed effects	No	No	No
Clients	21'122	21'122	21'052
Client * Year observations	42'193	42'244	41'896
Mean of dependent variable	70.80	46.80	618.10
Method	OLS	OLS	OLS

Appendix A9. Debit PoS transactions - Multiple card holders

This table shows the results of OLS regressions for the sample of clients with one account and two debit cards in 2015. The dependent variables measure the number of debit PoS transactions per card and year. In column (1) the dependent variable covers all transactions, in columns (2-6) the dependent variable covers transactions of specific values only (0-20 CHF, 20-40 CHF, 40-60 CHF, 60-100 CHF, 100+ CHF). Each regression includes 3 annual observations (2016, 2017, 2018) per card. The explanatory variable *Contactless* is defined at the card level. It is 1 for cards replaced in years 2017 and 2018. All regressions include card fixed effects. Robust standard errors are reported in parentheses. *, **,*** denote significance at the 0.05, 0.01, and 0.001-level respectively

	(1)	(2)	(3)	(4)	(5)	(6)
		Debit card transactions by transaction value				
Outcome variable	All	below 20 CHF	20-40CHF	40-60 CHF	60 - 100 CHF	above 100 CHF
Contactless	2.291	1.376*	-0.112	0.028	0.412	0.588
	(1.516)	(0.696)	(0.457)	(0.339)	(0.383)	(0.361)
Year = 2017	1.355	1.591***	0.850*	-0.319	-0.271	-0.497
	(1.066)	(0.417)	(0.337)	(0.258)	(0.279)	(0.264)
Year = 2018	5.820***	4.469***	2.187***	-0.388	0.065	-0.513
	(1.624)	(0.705)	(0.495)	(0.367)	(0.399)	(0.385)
Client fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Clients	706	706	706	706	706	706
Cards	1'412	1'412	1'412	1'412	1'412	1'412
Card * Year observations	4236	4236	4236	4236	4236	4236
Mean of dependent variable	81.50	13.90	18.20	13.90	17.00	18.40
Method	OLS	OLS	OLS	OLS	OLS	OLS

Index of Working Papers:

June 15, 2015	Anil Ari	202	Sovereign Risk and Bank Risk-Taking
June 15, 2015	Matteo Crosignani	203	Why Are Banks Not Recapitalized During Crises?
February 19, 2016	Burkhard Raunig	204	Background Indicators
February 22, 2016	Jesús Crespo Cuaresma, Gernot Doppelhofer, Martin Feldkircher, Florian Huber	205	US Monetary Policy in a Globalized World
March 4, 2016	Helmut Elsinger, Philipp Schmidt-Dengler, Christine Zulehner	206	Competition in Treasury Auctions
May 14, 2016	Apostolos Thomadakis	207	Determinants of Credit Constrained Firms: Evidence from Central and Eastern Europe Region
July 1, 2016	Martin Feldkircher, Florian Huber	208	Unconventional US Monetary Policy: New Tools Same Channels?
November 24, 2016	François de Soyres	209	Value Added and Productivity Linkages Across Countries
November 25, 2016	Maria Coelho	210	Fiscal Stimulus in a Monetary Union: Evidence from Eurozone Regions
January 9, 2017	Markus Knell, Helmut Stix	211	Inequality, Perception Biases and Trust
January 31, 2017	Steve Ambler, Fabio Rumler	212	The Effectiveness of Unconventional Monetary Policy Announcements in the Euro Area: An Event and Econometric Study
May 29, 2017	Filippo De Marco	213	Bank Lending and the European Sovereign Debt Crisis
June 1, 2017	Jean-Marie Meier	214	Regulatory Integration of International Capital Markets

October 13, 2017	Markus Knell	215	Actuarial Deductions for Early Retirement
October 16, 2017	Markus Knell, Helmut Stix	216	Perceptions of Inequality
November 17, 2017	Engelbert J. Dockner, Manuel Mayer, Josef Zechner	217	Sovereign Bond Risk Premiums
December 1, 2017	Stefan Niemann, Paul Pichler	218	Optimal fiscal policy and sovereign debt crises
January 17, 2018	Burkhard Raunig	219	Economic Policy Uncertainty and the Volatility of Sovereign CDS Spreads
February 21, 2018	Andrej Cupak, Pirmin Fessler, Maria Silgoner, Elisabeth Ulbrich	220	Exploring differences in financial literacy across countries: the role of individual characteristics and institutions
May 15, 2018	Peter Lindner, Axel Loeffler, Esther Segalla, Guzel Valitova, Ursula Vogel	221	International monetary policy spillovers through the bank funding channel
May 23, 2018	Christian A. Belabed, Mariya Hake	222	Income inequality and trust in national governments in Central, Eastern and Southeastern Europe
October 16, 2018	Pirmin Fessler, Martin Schürz	223	The functions of wealth: renters, owners and capitalists across Europe and the United States
October 24, 2018	Philipp Poyntner, Thomas Reininger	224	Bail-in and Legacy Assets: Harmonized rules for targeted partial compensation to strengthen the bail-in regime
Dezember 14, 2018	Thomas Breuer, Martin Summer	225	Systematic Systemic Stress Tests
May 20, 2019	Helmut Stix	226	Ownership and purchase intention of crypto-assets – survey results
October 17, 2019	Markus Knell Helmut Stix	227	How Peer Groups Influence Economic Perceptions

February 26, 2020	Helmut Elsinger	228	Serial Correlation in Contingency Tables
March 2, 2020	Mariarosaria Comunale, Markus Eller, Mathias Lahnsteiner	229	Assessing Credit Gaps in CESEE Based on Levels Justified by Fundamentals –A Comparison Across Different Estimation Approaches
April 30, 2020	Martin Brown Nicole Hentschel Hannes Mettler Helmut Stix	230	Financial Innovation, Payment Choice and Cash Demand – Causal Evidence from the Staggered Introduction of Contactless Debit Cards

This Is “What’s in Your Wallet”...and Here’s How You Use It*

Tamás Briglevics[†]
Magyar Nemzeti Bank

Scott Schuh[‡]
West Virginia University

April 2020

Abstract

Consumer wallets have more means of payment yet cash still is used most. We develop a dynamic structural model blending cash inventory management and payment instrument choice. For each expenditure, consumers endogenously pay with cash, debit card, or credit card with an option to withdraw cash beforehand. The model is estimated with transaction-level data from a daily consumer payment diary and reveals that utility from payment services exceed cash management costs. For payment card owners, optimal cash holdings are about $50 and jointly determined with the share of cash payments. Eliminating either cash or payment cards reduces consumer welfare significantly.

Keywords: Money demand; cash inventory management; payment demand; debit cards; credit cards; structural estimation; discrete-continuous choice; Diary of Consumer Payment Choice

JEL Classification: E41, E42, D12, D14

[*] The views expressed in this paper are those of the authors and do not necessarily represent the views of West Virginia University or Magyar Nemzeti Bank. For helpful comments and suggestions, we thank: Victor Aguirregabiria, Susanto Basu, Santiago Carbo-Valverde, Kim Huynh, Peter Ireland, Marc Rysman, Philipp Schmidt-Dengler, Zsolt Sándor; Judit Temesváry, and Randall Wright; participants at the ECB-BdF Retail Payments at a Crossroads Conference, Hungarian Academy of Sciences Summer Workshop, Econometric Society European and North American Meetings and World Congress, International Association for Applied Econometrics Annual Conference, International Industrial Organization Conference, the Bundesbank Cash Conference, Jornadas Economia Industrial; the Bank of Canada Retail Payments Conference; the WEAI Annual Conference; and seminar participants at the Bank of Canada, Boston College, and Federal Reserve Banks of Boston and Atlanta.

[†] *E-mail*: tbriglevics@gmail.com

[‡] *E-mail*: scott.schuh@mail.wvu.edu

1 Introduction

A popular advertising campaign for a U.S. bank asks, "What's in *your* wallet?" For years the answer was "cash and checks," plus maybe one credit card for high-income consumers. Today, U.S. consumer wallets are thick and diverse following a quarter-century transformation of payments from paper to cards and electronic means of payment.[1] Most consumers have five or six types of payment instruments; the average wallet holds nearly a dozen (two per type). Now, three-fourths of consumers have at least one credit card and the average consumer has 3-1/2. The average (median) wallet still has $70 ($30) of cash despite ardent efforts to eliminate it. For reasons not fully understood, consumers have adopted new instruments without discarding older ones.[2] And there is no representative wallet—more than 100 unique portfolios of instruments exist. Only one in seven consumers holds the most popular combination of cash, check, debit card, credit card, and two types of electronic bank payments.

One possible reason for thicker wallets is heterogeneous utility from payment services and no instrument emerging as "one size fits all." U.S. consumers make about three-quarters of their payments (volume, not value) with cash, debit cards, and credit cards, mainly for retail and other low-value payments; consumers often turn to electronic instruments for bills and other higher-value payments (see Greene and Schuh 2017). Some consumers rely heavily on one type of payment card (debit, credit, or prepaid) for their card payments, a practice called "single-homing" by Rysman (2007) and Shy (2013). But scant few consumers single-home for all payments, and even less report never using cash (see Briglevics, Schuh, and Zhang 2016). Klee (2008) found that instrument choices are correlated with the dollar values of payments—cash for low values and debit or credit cards for higher values. Non-acceptance of payment instruments occurs, but it is too rare to explain the U.S. diversity choices. However, using new data from the Diary of Consumer Payment Choice (DCPC), we find the probability of cash use is roughly con-

[1] This transformation is being measured by the Federal Reserve Payment Study and the Survey and Diary of Consumer Payment Choice from the Federal Reserve Bank of Atlanta. Unless noted otherwise, statistics cited in this paper are from Greene, Schuh, and Stavins (2016) and Greene and Schuh (2016).

[2] The exception is checks, which most consumers still have but are using less often. See Gerdes and Walton (2002), Benton et al. (2007), Schuh and Stavins (2010), and Gerdes et al. (2019).

stant around 50 percent for most payments (i.e., less than \$100) when consumers have sufficient cash in their wallets at the point of sale. Hence, the negative correlation between the probability of choosing cash and payment values depends on consumers' cash management policies. Thus, analyzing payment choices independently of cash holdings may lead to incorrect inferences about consumers' preferences for payment services.

Theoretical models generally have not kept pace with the remarkable scope of transformation in money and payments because two strands of literature have not been fully connected. One strand is the demand for money, where prototypical models of cash inventory management are Alvarez and Lippi (2009, 2017).[3] This research includes a few means of payment—cash, debit cards, and credit cards—but the adoption, characteristics, and suitability for expenditure of payment instruments are not central to the problem. Instead, these models impose *a priori* temporal orderings on the use of assets and liabilities, which are not consistent with transactions-level data. The other strand is the demand for payment instruments, where a protoypical model is Koulayev et al. (2016).[4] This research examines a wide range of payment instruments, modeling their adoption and use based on a rich array of instrument characteristics and payment conditions, including dollar value, that yield utility and influence endogenous choices at the point of sale. However, these models tend to be static, ignore cash inventory management, and abstract from consumption-saving and portfolio allocation decisions that are central to monetary models.

To better understand simultaneous demand for money and payments, we propose a dynamic optimizing model of consumers making daily cash management and payment choices that blends the theoretical approaches in the two literatures. As in monetary models, consumers manage cash inventories to fund current and future payments.[5] As in payments models, agents endogenously choose an instrument for each transaction to

[3] Other research examining money demand with an option for credit payments includes Telyukova (2013), Briglevics and Schuh (2013), Fulford and Schuh (2017), and Alvarez and Argente (2019).

[4] Other research examining payment choice includes Schuh and Stavins (2010), Wakamori and Welte (2017) and Hunyh, Nicholls, and Shcherbakov (2019).

[5] Limited data availability prevents the inclusion of similar management tasks for other liquid assets and liabilities, such as checking accounts and credit card accounts. The potential benefits of doing so are illustrated in Samphantharak, Schuh, and Townsend (2018).

maximize utility from payment services. This way the model can replicate empirically observed orderings and substitution patterns among instruments across transaction values, and provide a framework for evaluating the relative importance of cash management costs and utility from payment services for consumer welfare.

The model is estimated with transactions-level longitudinal micro data that tracks each consumer payment and cash management decision. The data are from the DCPC, the U.S. version of daily diary surveys developed by central banks and other researchers to record consumer cash management and payment activity in industrial countries documented in Bagnall et al. (2016). In addition to capturing the richness of cash management and payment choices, diary surveys have less error than recall-based survey data used in previous research, and diaries provide relatively accurate estimates of aggregate consumer expenditures (see Schuh (2018)). Although the theoretical model does not yield closed-form solutions, its structural parameters can be estimated using the method described in Bajari, Benkard, and Levin (2007).

The estimated model reveals important insights that extend the cash demand and payment choice literatures. Two key conclusions emerge. First, the estimated model provides statistically and economically significant evidence that consumers jointly determine cash demand and payment choice, so models that focus on just one of these decisions are incomplete. Second, the estimated model reveals that utility gains from payment choices are about an order of magnitude larger than losses from cash management costs. In retrospect, the latter finding should not be surprising. The average U.S. consumer only makes five cash withdrawals per month but 59 payments, so opportunities to reap utility from optimal payment choices exceed the incidence of costs in managing cash.

Cash management in the estimated model is qualitatively similar to existing models with fixed or exogenous cash payments but now exhibits fluctuations in the share of cash payments due substitution among instruments. This feature leads to changes in the utility derived from payment services that are of comparable magnitude to changes in cash withdrawal or holding costs. Thus, the monetary literature's focus on cash managment costs misses an important source of consumer welfare derived from the functioning of the payment system.

Likewise, payments in the estimated model are qualitatively similar to existing models without cash management but instrument choice probabilities now depend on cash holdings and the random costs of withdrawals. The probability of cash use declines much faster with payment value when cash holdings are smaller because consumers try to postpone withdrawals until a favorable opportunity is available. Conversely, consumers with very large amounts of cash in their wallets are much *more likely* to use cash. We estimate the optimal cash holdings to be around \$50, so consumers with larger cash stocks will want to spend cash. Alvarez and Lippi (2017) describe this phenomenon as “cash burns” in a model where cash is assumed to be used first; our model exhibits this behavior when consumers are not constrained to order their use of assets and liabilities and consumers make optimal dynamic choices.

Finally, the structural model enables us to run counterfactual simulations of restrictions on payment choices at the point of sale. Most notably, decreases in utility stemming from eliminating (or not accepting) a single payment instrument are notably larger than changes in utility associated with changes in cash management costs. As a practical matter, cash still contributes significantly to consumer welfare despite criticisms and calls for its removal by Rogoff (2016) and others. However, eliminating both debit and credit cards would reduce utility by almost an order of magnitude more than any single instrument, reflecting the large value of technological innovations embodied in electronic card networks. These findings likely have implications for the operation of monetary and payment systems, and the public policies governing them.

2 Literature Review

This section provides a brief but overview of two literatures, monetary and payments, that are inherently related but remain largely disconnected. This paper is part of an emerging research program that is attempting to more fully integrate them.

2.1 Demand for Money and credit

Modeling money demand as the optimal solution of an inventory management problem has a long tradition in monetary economics starting with Allais (1947) and popularized by Baumol (1952) and Tobin (1956). The core objective of this problem, the minimization of opportunity and transactions costs, remains central to the current literature. Changes in transactions costs are most often specified as improvements in withdrawal technologies such as ATMs (for examples, see Lippi and Secchi 2009; Alvarez and Lippi 2009; Amromin and Chakravorti 2009). Opportunity costs arise from interest-differentials between liquid assets serving as a medium of exchange without bearing interest, like currency, and interest-bearing assets that cannot be used for payment.

The opportunity cost distinction has been evolving as the number of assets serving as a medium of exchange and the number bearing interest both have increased over time. Whitesell (1989) extended the Baumol-Tobin model to allow payments from currency and debitable (checkable) demand deposits that do not pay interest but have a fee differential. The elimination of Regulation Q in the early 1980s permitted interest payments on demand deposits, but still only about half of consumers have an interest-bearing checking account. Mulligan and Sala-i-Martin (2000) show that failure to adopt interest-bearing transaction accounts affects the interest-elasticity of money demand. Subsequent financial innovations increased the variety of interest-bearing liquid assets available to settle payments. For example, Ball (2012) and Lucas and Nicolini (2015) argue that money market deposit accounts (MMDA), which now are used as a medium of exchange, can be added to transactions balances to mitigate the historical destabilization of M1 velocity.[6]

Other theoretical approaches to modeling the demand for money go beyond the framework proposed in this paper. One approach is the shopping-time model in which money balances produce utility by saving time or energy in the shopping process (see McCallum and Goodfriend 1987),which is similar to a money-in-utility function specification. A related, but deeper, approach is search-theoretic models in the New Monetarist Economics (NME) tradition, which motivate demand for cash balances because they facilitate

[6] Also, Hester (1972) accurately predicted that money velocity would be affected by the introduction of electronic funds transfers (Automated Clearing House network).

exchange (see Lagos, Rocheteau, and Wright 2017).

Demand for transactions balances to fund consumer expenditures also includes short-term (revolving) credit. Sastry (1970), Bar-Ilan (1990), and Alvarez and Lippi (2017) offer models that allow consumers to pay with credit *after* they run out of cash. Microeconometric studies similar to this paper estimate more stable money demand by controlling for adoption of credit cards (Reynard 2004; Briglevics and Schuh 2013). Alternatively, studies like Townsend (1989), Telyukova and Wright (2008) and Telyukova (2013) offer NME style models in which consumers hold cash balances because they are unable to buy certain goods using credit. From this line of research, Chiu and Molico (2010) is closest to our work; their calibrated general equilibrium model features cash withdrawal decisions resulting from a stochastic dynamic optimization problem.

Models of demand for money and credit often assume a temporal ordering of use based on *a priori* beliefs about the relative costs and benefits—lowest net cost funds are used first—rather than allowing transaction-specific variation in net benefits. Strict temporal orderings of settlement funds are inconsistent with empirical evidence found in daily payment diaries where the choice of money or credit varies by transaction.[7] NME models that allow non-acceptance of money or credit by sellers can generate alternating use of funds in environments where exchange opportunities and outcomes are random. But payment choices become more systematic when acceptance is universal or agents have foreknowledge of acceptance and preferences for household financial decisions, especially cash management.

Recent research has begun to address the need for transaction-specific endogenous demand for money and credit that may vary across types of consumers. For example, (Nosal and Rocheteau 2011, chapter 8) presents a tractable model in which consumers endogenously choose between credit and cash and can reset their cash holdings at a fixed cost. Fulford and Schuh (2017) build a model with endogenous payment choices that embodies the relative net benefits of money and credit and links them to consumption expenditures and debt accumulation. Following the model of Duca and Whitesell (1995),

[7] Table 1 in Huynh, Schmidt-Dengler, and Stix (2014) details the predictions of some models that are not borne out in Canadian and Austrian data.

Briglevics and Schuh (2013) find microeconomic evidence that demand for currency is less interest sensitive for credit card revolvers with high-interest debt than for convenience users who pay no interest on their credit card use.

In general, the monetary literature has abstracted from details about the choice of instrument used to authorize payment. Tobin (2008) defined payment instruments as "derivative media" linked to monetary assets (currency, demand deposits, etc.) and to liabilities (such as credit card limits). For currency, the instrument and asset are the same, but multiple instruments can be used to access demand deposits (checks, debit cards, prepaid cards, and online banking payments). Prescott and Weinberg (2003) show that non-pecuniary characteristics of payment instruments, such as communication and commitment, also can be important determinants of their use. This decision has become more complex as payment instruments once limited to demand deposits now can be used to make payments directly from more favorable liquid assets, like MMDAs, or from liquid liabilities, like a home equity line of credit (HELOC). And, of course, not all credit cards are alike in terms of their fees, rewards and rates paid to revolve balances—prompting a bank to ask which card is in our wallets. Thus, studying payment choices jointly with demand for money and credit may expand our ability to understand and explain the payments transformation and financial innovations in assets and liabilities.[8]

2.2 Demand for Payments

A key segment of the payments literature is modeling consumer demand for instruments to authorize retail payments.[9] An early innovation is Stavins (2001), which investigated slow *adoption* of electronic payments methods by heterogeneous consumers using the

[8] The advent of new technologies such as e-money and mobile payments also may have similar implications. Recent technology has even altered the concept of "money" itself, with Bitcoin and M-PESA (Jack, Suri, and Townsend 2010) serving jointly as an electronic payment network and private money in the form of "virtual currency." For extended definitions and discussions of "e-money," see ECB (2012, 2015) and Committee on Payments and Market Infrastructure and Markets Committee (2018).

[9] Research on supply of payment services—provision of payment networks and the acceptance of payment instruments by merchants—also is important in general equilibrium. Humphrey, Kim, and Vale (2001) argue that adoption of electronic methods lowers the social costs of payment systems. See Hunyh, Nicholls, and Shcherbakov (2019) for an estimated model of merchant acceptance. We exclude this part of the literature because it goes beyond the scope of our partial equilibrium consumer model, and because acceptance is not measured well in the DCPC.

limited data on payments in the Survey of Consumer Finances. Subsequent research by Borzekowski, Kiser, and Ahmed (2008) and Schuh and Stavins (2010), as well as references therein, also modeled the *use* of payment instruments (number of payments) as a function of technology and instrument-specific characteristics like cost, convenience, security, and record-keeping using better-suited recall-based survey data. This research relies on two-step discrete-continuous models of adoption and use of individual payment instruments. Koulayev et al. (2016) extended this approach by simultaneously modeling adoption of a bundle of instruments (the wallet), and including random utility from the use of payment instruments in various payment contexts. This model focuses primarily on costs and benefits of instruments used to make heterogeneous payments by a cross-section of consumers, but abstracts from consumer demand for money and credit needed to settle payments.

An alternative approach is to model consumer demand for payments at the point of sale (POS) over time. Starting with Klee (2008), and followed by Cohen and Rysman (2013) and Wang and Wolman (2016), researchers used scanner data from retail stores to document instrument choices at checkout to estimate multinomial logit models. These studies found notable correlation between the dollar values of individual transactions and the choice of payment instruments, with cash being far more likely to be used for payments of small dollar values.[10] This result added a new perspective unavailable from survey data, which generally do not contain individual payments or dollar values. However, except for Cohen and Rysman (2013), scanner data do not provide information about the demographics of each consumer, their options at the time of payment (cash in their wallet or instrument adoption), or the longitudinal behavior of individual consumers. In particular, scanner data do not reveal single-homing behavior by individual consumers (see Rysman 2007; Shy 2013), which (Briglevics, Schuh, and Zhang 2016) show is obscured by the aggregate correlation between payment values and instrument choices across all consumers.

[10] Arango, Hogg, and Lee (2015), Eschelbach and Schmidt (2013), Briglevics and Schuh (2014), and Huynh, Schmidt-Dengler, and Stix (2014) also provide evidence that cash holdings are correlated with payment instrument choices.

Shortcomings of recall-based surveys and scanner data motivated development of daily consumer payment diaries used in the cross-country study by Bagnall et al. (2016). In real time, payment diaries track the dollar value of each transaction, the payment instrument used, and information about the consumer and merchant involved in each payment.[11] Recent research uses payment diary data to estimate POS choice probability models for various countries and non-retail transactions.[12] Wakamori and Welte (2017) extended this research using the Canadian data to estimate a random coefficients model where not all respondents switch from cash to a debit or credit card at the same transaction value. They found the dominance of cash for low-value transactions is primarily driven by consumer preferences for cash. A limitation of econometric models applied to diary data thus far is they are not derived from a dynamic optimizing framework for consumers' joint payment and cash management choices that provides cash-flow accounting of money holdings (stock) and payments, withdrawals, and deposits (flows).

2.3 Joint demand for money, credit, and payments

The unique role of payment instruments offers the potential to better connect demand for money and credit, on one had, with the demand for specific consumer expenditures. An early example is Prescott (1987), which enhances cash-in-advance constraints by jointly modeling the choice of payment instruments (currency and interest-bearing bank drafts). Fulford and Schuh (2017) jointly models credit card spending, revolving debt, and payments settled with money over the life-cycle. Alvarez and Argente (2019) models the cash-credit card tradeoff for consumers paying for Uber rides. And Stokey (2019) develops an extensive general equilibrium model that includes banks and a monetary authority to assess the macroeconomic impact of payment choices. In each case, however, the models only determine the aggregate shares of expenditures and funding paid for with each

[11] Cohen and Rysman (2013) resolved the scanner data anonymity problem by surveying participating consumers and asking them to re-scan their products. This strategy produces data similar to a payment diary but requires *ex post* recall of real-time POS conditions.

[12] See Fung, Huynh, and Sabetti (2012) and Arango, Hogg, and Lee (2015) for Canada; van der Cruijsen, Hernandez, and Jonker (2015) for The Netherlands; Bounie and Bouhdaoui (2012) for France; von Kalckreuth, Schmidt, and Stix (2009) and Eschelbach and Schmidt (2013) for Germany, and Briglevics and Schuh (2014) for the United States.

instrument type during a period of time, not the choice of payment instrument and settlement funds for individual payment opportunities.

The model proposed in this paper models each sequential payment choice for individual consumer expenditures while tracking consumer cash management and the corresponding cash-flow for currency. To our knowledge, this is the first attempt to use longitudinal panel data with individual transactions from payment diaries to estimate a dynamic optimizing model that jointly explains consumer payment instrument use and cash management linked by the accounting cash-flow identity at the transaction level. Samphantharak, Schuh, and Townsend (2018) illustrate the empirical potential of this approach using the 2012 DCPC data to demonstrate how household financial statements can track exact cash-flows connecting the payment instrument used to authorize a specific consumer expenditure directly to the monetary asset or credit liability (balance sheet) used to settle the exchange.

3 Data

This section provides a brief overview of the primary data sources for this paper, the 2012 Diary of Consumer Payment Choice (DCPC) and corresponding 2012 Survey of Consumer Payment Choice. More details can be found in Schuh (2018) and Appendix B.

The SCPC and DCPC are complementary surveys that measure detailed payment choices and cash management of U.S. consumers. SCPC respondents complete an online survey and *recall* from memory their adoption of financial acccounts and payment instruments, cash management, and (not used in this paper) frequency of use of payment instruments. DCPC respondents *record* their payment transactions and cash management for three consecutive days. We use SCPC consumer data on adoption of accounts and payment instruments plus DCPC transactions data on: 1) payment values, instrument used, location, and type; 2) cash holdings, deposits, and withdrawals by location; and 3) time of day.

DCPC data are a balanced longitudinal panel of a representative sample of about

2,500 U.S. consumers during October 1-31, 2012. Respondents were selected from the RAND *American Life Panel* to match the population of U.S. adults (ages 18 years and older). After completing their SCPC, respondents were assigned to complete their DCPC on randomly selected days throughout the month so panel entry and exit is deterministic and fixed. This diary design produces representative samples for each day of the month as well as for the entire month.

The DCPC panel data mimic the transaction records of monthly statements for checking and credit card accounts. Thus, they are essentially the same as transactions data from financial institutions provided by the kinds of personal financial management (PFM) services and applications used by Baker (2018), Pagel and Olafsson (2018), and Gelman et al. (2018). Data from financial institutions may have less measurement and reporting error than consumer diary data, but the DCPC data are superior in other respects. For example, the DCPC tracks what consumers do with cash withdrawn from banks, not just how much they withdrew. The DCPC also collects additional relevant information at the time of transaction, such as cash held in wallet. And, importantly, the DCPC data are based on sampling and implementation methods that are designed to produce representative samples of U.S. consumers whereas PFM data are not.

We restrict the sample for model estimation to *in-person* POS transactions, including person-to-person (P2P) payments, by consumers who had *both* a debit card (hence checking account) and credit card. The restricted sample represents the bulk of cash use because online payments don't accept cash and few bill payments are made with cash. Wallet restrictions are made to sidestep the theoretical complication of modeling adoption; in practice, respondents are unlikely to adopt or discard payment cards during the three diary days. The restricted sample accounts for 62 percent of POS transactions and 57 percent of respondents, who are not quite representative of the U.S. population. However, payment card adopters rely on cash relatively less than other consumers, so our results likely serve as a lower bound on the usefulness of cash.

4 Empirical Evidence

This section provides evidence on consumer payment choices and cash management to motivate the model and enhance understanding of the estimation results.[13]

4.1 Payment Adoption and Use

The first two panels of Table 1 report statistics on consumer adoption and use of payment instruments for the DCPC ("full sample") and sub-sample used in estimation ("estimation sample"). In the full sample, all respondents adopted cash, 78 percent had a debit card, 69 percent had a credit card, 57 percent had both payment cards, and only 10 percent had neither card.[14] In the estimation sample, respondents have all three payment instruments by construction. Despite thicker consumer wallets, cash is still king at the point of sale. In the full sample, cash accounted for half (51 percent) of POS payments by volume (number of transactions). Even in the estimation sample, where respondents have both payment cards, cash accounted for a higher share (44 percent) than either debit cards (31 percent) or credit cards (24 percent). Thus, the estimation subsample understates the full use and value of cash.

Ching and Hayashi (2010) showed that consumer use of payment cards can be influenced by monetary incentives, such as cash back or airline mileage, that entice consumers to use payment cards more often. "Convenience users" who pay off their credit card charges in full each month receive the full benefit of rewards, but "revolvers" who carry high-interest unpaid balances on their cards have an offsetting cost. Table 2 shows consumer payment choices broken down by credit card use (convenience or revolving) and type (with rewards or not) in the estimation sample. Not surprisingly, consumers with a rewards card are more likely to pay with a credit card—convenience users are nearly twice as likely (40.0 versus 23.1 percent), and revolvers more than three times (19.6 versus 5.8).

[13] Reported sample moments are unweighted because the structural model is estimated without weights. The DCPC data are collected using stratified random sampling, so weighted sample means are required to estimate population moments for all U.S. consumers, which can be found in Schuh and Stavins (2014) and Greene, Schuh, and Stavins (2018).

[14] The weighted population estimates are quite similar: 100 percent for cash, 79 for debit card, and 72 percent credit card. Cash "adoption" actually is measured in the SCPC and DCPC questionnaires

Variable	DCPC Sample Full	DCPC Sample Estimation
Adoption rates (share of respondents)		
Cash	1.00	1.00
Debit card	.78	1.00
Credit card	.69	1.00
Debit and credit card	.57	1.00
Neither debit nor credit card	.10	0.00
Payment use (share of transactions)		
Cash	.51	.44
Debit	.28	.31
Credit	.21	.24
Transactions at POS with cash, debit, credit (#)		
Total	10,822	6,707
When CIA binds	2,803	2,044
When $m < \$2$	1,206	850
Values at POS with cash, debit, credit ($)		
Median	12.60	13.41
Average	27.99	29.66
Standard deviation	66.66	73.89

NOTE: The number of respondents is 2,468 in the full DCPC sample and 1,272 in the estimation sample.

Table 1: Payment instruments and transactions, 2012

Credit card type	Number of transactions	Percentage of transactions (%) Cash	Debit	Credit	Preceded by withdrawals
Convenience users					
Rewards	1,661	42.6	17.5	40.0	7.5
No rewards	2,582	42.6	34.3	23.1	9.3
Revolvers					
Rewards	1,860	46.0	34.4	19.6	8.3
No rewards	604	47.9	46.4	5.8	9.1
All types	6,707	44.0	31.2	24.8	8.5

Table 2: Payment choices by credit card type, 2012

However, adoption of a rewards card has little effect on cash activity because higher credit card use is largely offset by lower debit card use. Table 2 shows that revolvers use cash 3-5 percentage points more often than convenience users, but cash shares are essentially the same for consumers with and without rewards. Although rewards card holders are less likely to withdraw cash before a transaction, the differences are less than 2 percentage points.[15] These results are fortuitous because the DCPC data do not track whether specific card payments were made with a rewards card or not. Therefore, the model and estimation can focus on cash management without specifying separate decision rules for different types of debit and credit card adopters and users.

4.2 Transactions

The remaining two panels of Table 1 report statistics on the volume and values of transactions for which consumers made payments. Nearly 11,000 POS transactions are recorded in the diary. The estimation sample includes 57 percent of all DCPC respondents who account for a slightly disproportionate amount of payments at 62 percent ($\sim 6,707/10,822$). For close to one-third of transactions ($\sim 2,044/6,707$), cash is not an option because the consumer does not have enough in their wallet to fund the payment and hence the cash-in-advance (CIA) constraint is binding. For almost one in eight transactions ($\sim 850/6,707$), consumers have essentially no cash in their wallet ($<$ \$2).

Table 1 also reveals that most POS transactions are relatively low-value. The median consumer payment was \$13, so half of all recorded POS transaction values do not require consumers to hold large amounts of cash. Some merchants impose minimum values (typically \$10) for credit card transactions, which also helps cash to compare favorably. Even the average transaction value was only slightly more than double the median (about \$30) despite large variation (standard deviations). However, the left panel of Figure 1 shows that the full distribution of POS transaction values is skewed to the right by much larger amounts, even after excluding bill payments.

rather than assumed. It is defined as having or using cash at some point during the year.

[15] Using SCPC data, Briglevics and Schuh (2013) found no effect of credit card rewards or debt on average cash holdings but showed that cash demand of revolvers is less interest sensitive than cash demand of convenience users.

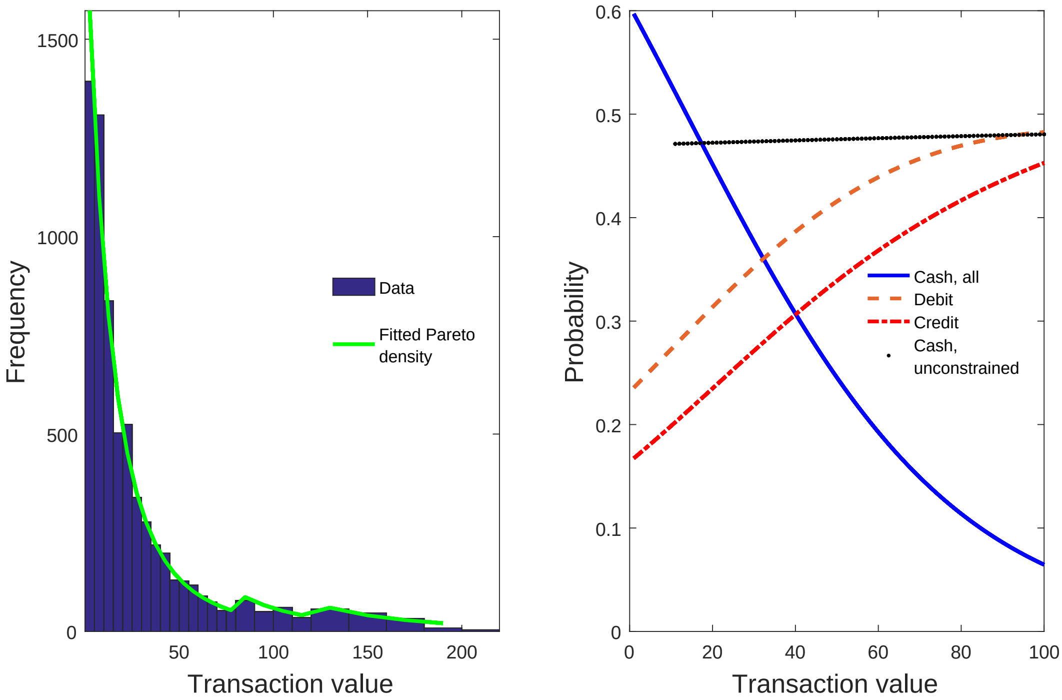


Figure 1: Distribution of POS transaction values (left) and payment probabilities (right)

As noted in Section 2, transaction values are good predictors of the payment instruments consumers choose. Following the literature, we estimated a multinomial logit model of payment choice and plot the unconditional probabilities of each instrument as a function of transaction value in the right panel of Figure 1. Like the scanner data, DCPC data reflect a negative correlation between cash use and transaction values. Payment cards are used more often for larger values, with debit cards slightly higher than credit.[16] These payment choice probabilities are central to estimation of the structural model, which adds controls for consumer-level cash management.

To preview later results showing the sensitivity of cash use to cash holdings, the right panel of Figure 1 also includes the estimated probability of cash use for the subset of transactions that were unconstrained by the amount of cash in their wallets (dotted black line).[17] When consumers had enough cash to pay for their next transaction in full

[16] The modest dominance of debit differs from prior estimates using retail-store scanner data that showed credit more common than debit. The reason is that scanner data combines signature debit and credit card payments, which run on the same networks, and could not be identified separately due to technical limitations. Instead, the DCPC measures signature and PIN debit card payments separately, so debit and credit use are identified accurately.

[17] The multinomial logit of payment choice simply adds an indicator variable for a binding CIA constraint to the variables in the utility functions (a constant, an indicator variable for transaction values under \$10, and a linear term in the transaction value).

with cash, the probability of using cash was remarkably stable at just under 50 percent for transaction values up to \$100. Thus, the overall negative correlation between cash use and transaction values, observed in the data and noted in the literature, appears to be explained by cash holding behavior. Indirectly, however, the occurrence of payment values that exceed the amount of cash held in wallet reflects consumers' endogenous decision to forego a cash withdrawals that would have removed their cash-in-advance constraint. Our main contribution is to build and estimate a model that can assess whether reluctance to withdraw cash primarily reflects the costs of cash management or consumers' inherent preferences for using cash to pay for transactions, especially those with low value.

4.3 Cash Management

Table 3 reports statistics on cash holdings and withdrawals. In addition to providing context for model estimation, these statistics suggest how well cash demand models in prior research could explain the DCPC data.

4.3.1 Cash holdings

Most consumers hold low amounts of cash, but some hold relatively large amounts (first two panels of Table 3). The median consumer in the estimation sample only has \$20 stored at home (first panel) compared with \$36 in the median consumer's wallet before a transaction (second panel). However, average cash held at home is \$202, whereas the average held in a wallet is only \$76. Thus, while most consumers would require a cash withdrawal to pay for a large-value transaction, some have a large stash of cash they can tap to replenish their cash-in-wallet holdings.[18] The average cash in a wallet can fund 2-1/2 average-sized transactions ($\sim$ 75.57/29.66) and 6 median-sized transactions ($\sim$ 75.57/12.60), but median cash in wallet can fund less than 2 median transactions ($\sim$ 20/13).

[18] As explained in Appendix B, these cash-at-home stocks are used to handle cases where the cash-flow identity does not hold. We construct an artificial withdrawal category (not reported in the diary) called "beginning-of-day adjustment" that accounts for about one-fifth of all withdrawals.

Variable	DCPC Sample Full	 Estimation
Cash held at home ($)*		
Median	20.00	20.00
Average	234.23	202.02
Standard deviation	583.15	466.62
Cash in wallet		
Before POS transaction ($)		
Median	40.00	36.00
Average	80.98	75.57
Standard deviation	145.40	130.58
*Before card transactions (ratio)****		
Median debit card	.61	.61
Median credit card	1.37	1.10
Average debit card	3.68	3.62
Average credit card	6.02	4.77
Before withdrawal ($)		
Median	10.00	11.00
Average	41.32	43.09
Standard deviation	107.63	114.10
*Cash withdrawals***		
Number (#)	1,024	573
Median amount ($)	40.00	40.00
Average amount ($)	81.30	77.27

NOTES: *Excludes observations above $5,000. **Excludes observations above $1,100. Outliers are excluded because they significantly influence estimated moments. ***Value of cash in wallet relative to value of the current card transaction.

Table 3: Cash holdings and withdrawals, 2012

4.3.2 Payments and cash holdings

Although most consumers have non-trivial amounts of cash in their wallets, many pay with a debit or credit card instead of using their available cash. The third panel of Table 3 quantifies this fact by reporting the ratios of cash in wallet to the value of the next card payment; ratios of 1.0 or greater indicate transactions where the CIA constraint was not binding and vice versa for ratios below 1.0. For most credit card payments, the CIA constraint was not binding (median ratio $>$ 1.0), but for most debit card payments it was (ratio of .61). The average ratios of cash to debit and cash to credit payment values are much higher (3.62 and 4.77, respectively), which indicates that even consumers with very large amounts of cash in their wallets still make card payments for some reason.

The relationship between cash-in-wallet and POS transaction values (including card payments) appears in their joint distribution depicted in Figure 2. Both axes are in logs and the transaction value axis is inverted; the heat map denotes the number of transactions. The diagonal between the northwest corner (low transaction values and cash holdings) and southeast corner (high transaction values and cash holdings) demarcates the feasible region for cash payments. Above the diagonal, consumers held sufficient cash to pay for the transaction; below the diagonal, consumers faced a CIA constraint and paid with a card. The key fact in Figure 2 is that most transactions occurred when the CIA constraint was *not* binding. A non-trivial mass of transactions also exists where consumers had very low cash balances (orange-yellow region along the left vertical axis) and thus had to use a payment card.

Narrowing the focus to cash payments only, Figure 3 displays the shares of cash payments for combinations of transaction values and cash on hand. The flat portion of the floor is the infeasible region where the CIA constraint binds. Two important facts are evident. First, cash shares generally decline as transaction values increase for essentially all levels of cash on hand but bottom out at around 0.4, even for large transactions by consumers with enough cash in their wallet (see also right side of Figure 1). Second, the cash share for each transaction value increases slightly with the level of cash on hand. This finding is consistent with consumers worrying about running out of cash and trying

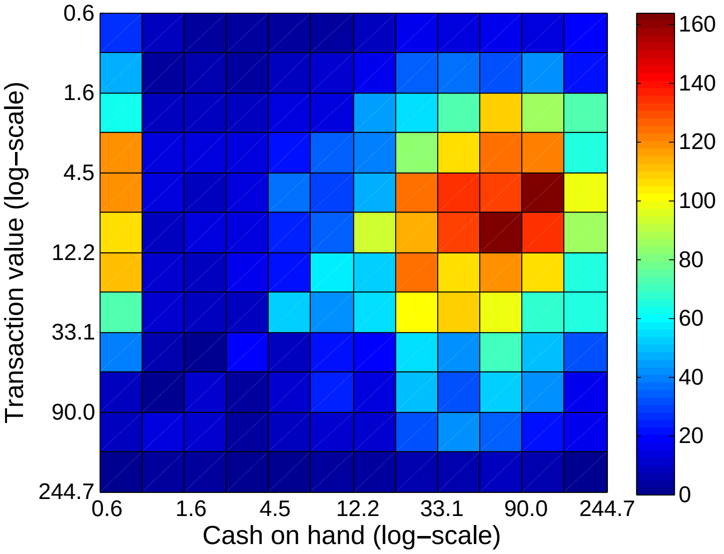


Figure 2: Joint distribution of POS transaction values and cash holdings

to conserve their holdings.

Overall, this subsection provides evidence against the hypothesis that consumers follow a lexicographic ordering of payment instrument choices across their sequential transactions. Consumers make card payments under a variety of cash holding conditions, and vice versa, so models that assume ordering of assets and liabilities (hence payment instrument choices) miss a salient feature of the data. To fit the data, models of cash demand must introduce structure to motivate different payment choices for each transaction and amount of cash holding. The model in the next section does this by introducing instrument-specific random utility that varies across payment opportunities and transaction values.

4.3.3 Withdrawals

The last two panels of Table 3 report cash withdrawals and their relation to cash holdings. Unlike transactions, consumer withdrawals are relatively rare. The estimation sample contains only 573 withdrawals for October 2012, an average of less than one per month (.45) per consumer. In the estimation sample, the median cash withdrawal was $40 and the average withdrawal amount was almost twice as much ($77). Figure 4 shows that the full distribution of withdrawal amounts is not smooth. The global mode is $20 and local modes occur at $40, $60, $100, and $200—all multiples of the two largest denominations.

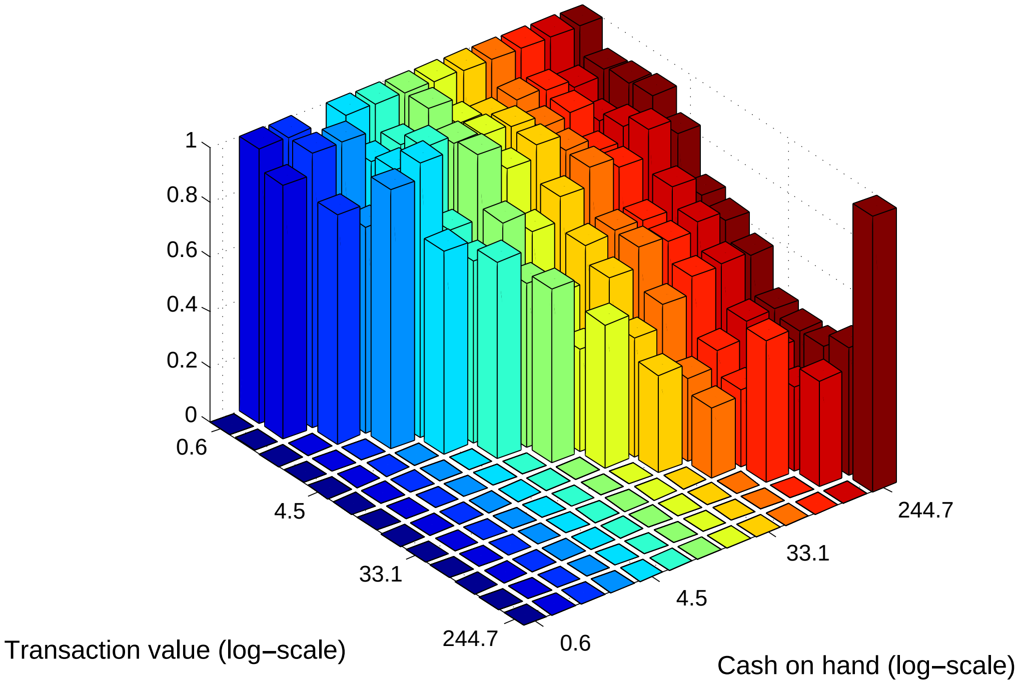


Figure 3: Shares of POS cash transactions

More than one in five withdrawals is less than $20.

An important feature of these withdrawal data is the heterogeneity of locations shown, in Table 4. ATMs are most common, but obtaining cash from family and friends or from the beginning-of-day adjustment are tied for the second most frequent. These three locations account for nearly two-thirds of all withdrawals, while the remaining third represent a diverse range locations. The average withdrawal amount varies by more than $100 across locations, which may reflect heterogeneity in the cost of withdrawals at each location. Little evidence is available on the cost of withdrawals by location, but some (bank teller, check cashing store) may be higher cost than others (ATM or cash back). Because there are not enough observations to identify withdrawal costs for each location, our model incorporates this feature with an unobserved random cost.

The penultimate panel of Table 3 shows that most consumers held some cash when making a withdrawal (median of $11), while some had considerably more (average of $43 compared to average transaction of $30). This finding contrasts with the basic Baumol-Tobin framework in which withdrawals only occur when cash holdings reach $0, but it is consistent with the models in Lippi and Secchi (2009) and Alvarez and Lippi (2009) that

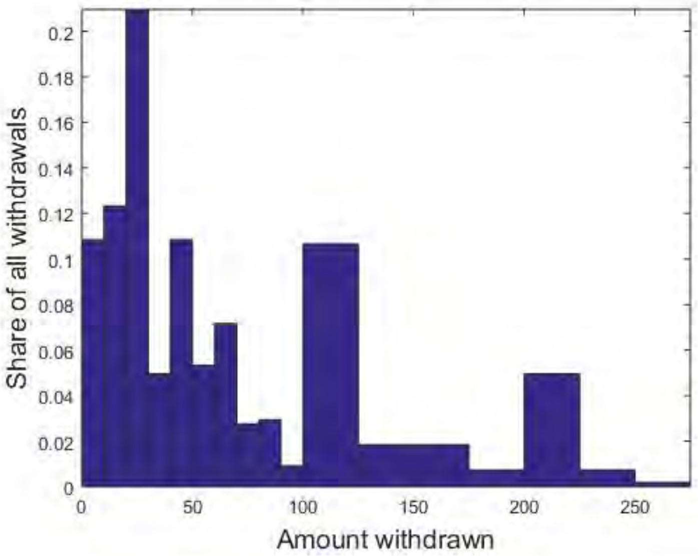


Figure 4: Distribution of withdrawal amounts, 2012

		Withdrawal amount ($)		
Location	Number	Average	Median	90th percentile
Bank teller	64	156	80	400
ATM	147	103	60	200
Cash back (retail store)	48	31	20	50
Cash refund (retail store)	7	30	21	75
Employer	25	104	70	200
Check cashing store	3	88	68	149
Family or friend	112	44	20	100
Other location	55	53	25	112
Beginning-of-day adjustment	112	60	26	167
Total	573	77	40	200

Table 4: Withdrawals by location, 2012

account for non-zero cash holdings at withdrawal by assuming random free withdrawals. However, the ratio of cash held before withdrawal ($41-43) to average cash in wallet ($76-81) is 0.5-0.6, notably higher in the 2012 DCPC than in Alvarez and Lippi (2009) for Italy (0.4) and the United States (0.3) in the 1980s. Lower interest rates and technological changes through 2012 may explain at least part of these differences.

Figure 5 depicts the relationship between withdrawals and transactions by the amount of cash holdings. Symbols (+ and o) indicate the shares of POS transactions (left scale) preceded by a withdrawal when the CIA constraint was binding (+) or slack (o). Stacked bars represent the number of transactions (right scale) used to calculate these shares

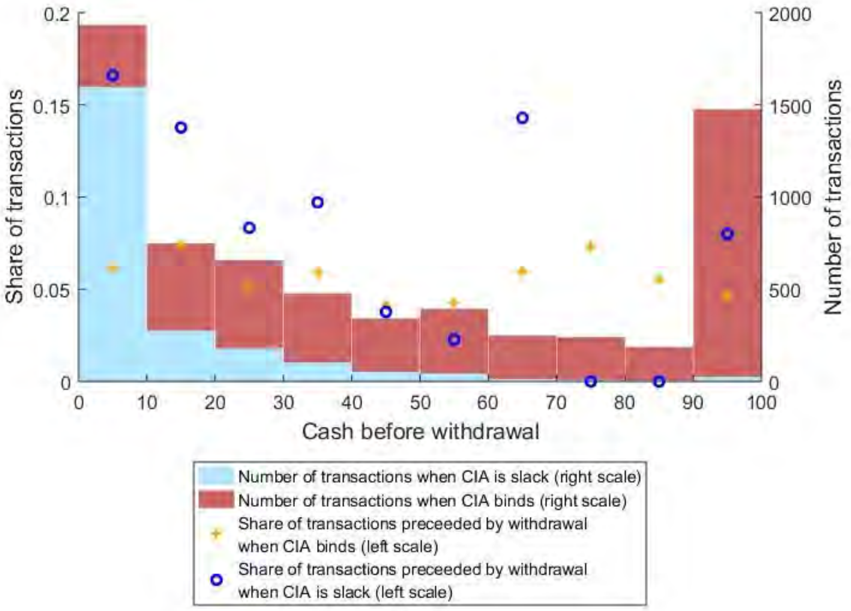


Figure 5: Share of withdrawals by amount of cash holdings

when the CIA constraint was binding or slack. Not surprisingly, consumers are more likely to make a withdrawal when the CIA constraint is binding. For example, when cash holdings are \$10 or less, cash-constrained consumers make a withdrawal for every six transactions whereas unconstrained consumers make one for every 16. When cash holdings reach \$40, the effect of the CIA constraint on withdrawals disappears. Very few consumers with more than \$50 face a binding CIA constraint, so the estimates of pre-transaction withdrawals are erratic in these small samples.

The evidence in this subsection, combined with the evidence in Figure 3 showing cash is used primarily for small transactions, suggests that short-term cash needs are an important driver of withdrawals. On the other hand, payment card holders can keep making purchases long after they run out of cash. These findings illustrate the simultaneity of cash management and payment choice, underscoring the importance of jointly modeling of these consumer decisions.

5 Model

This section describes our model of cash management and payment instrument choice, which blends and builds on Alvarez and Lippi (2009, 2017) and Koulayev et al. (2016).

Consumers finance a stream of transactions that have a stochastic value (p). Before each payment, consumers may withdraw cash; if so, they pay a stochastic withdrawal cost (b) and fixed holding (opportunity) cost of cash between transactions (R). Then, at the point of sale, consumers choose cash, debit card, or credit card to make each payment based on transaction-specific random utility derived from the payment services provided by the payment instrument chosen.

As noted in Section 2, existing models tend to impose a temporal ordering of cash use based on *a priori* assumptions about its cost relative to other means of payment. However, the evidence in Section 4 shows that consumers do not follow lexicographic ordering of payment instrument use, suggesting that the utility of payment services varies across transactions and time. Instead of imposing *a priori* restrictions on instrument value and timing, we parameterize the utility functions and estimate them.

Using a random utility framework to model payment instrument choice means that, unlike traditional inventory management models of cash demand, the withdrawal and holding costs become parameters of a utility function and are not measured in units of money or interest rates. While this feature is important when interpreting the econometric estimates later, it nevertheless fits into the literature that usually interprets these costs broadly. For example, withdrawal costs are usually thought of as including shoe-leather costs of finding an ATM; holding costs capture the inconvenience associated with keeping a certain amount of cash in one's wallet, not just foregone interest.[19]

Currency payments are subject to a CIA constraint. If cash balances are insufficient to settle a transaction, consumers cannot take advantage of high utility opportunities associated with cash transactions.[20]. As a result, their expected utility from future transactions falls as they run out of cash. This change in expected utility is balanced against the costs

[19] Given that consumers in the estimation sample make 2.3 (2.0) transactions per day on average (median), the opportunity cost or risk of theft should be small and we interpret holding costs primarily as the “inconvenience" of carrying cash. A generous 2 percent annual rate for checking accounts interest translates into a 0.00002 ($\sim 1.02^{\frac{1}{2.3*365}} - 1$) percent interest rate over the average holding period.

[20] In reality, debit and credit card payments are subject to funding constraints as well (checking account balances have a zero minimum and credit card borrowing has an upper limit). Ideally, the model would incorporate these constraints too, but the DCPC does not provide data on them. However, the CIA constraint on currency is likely to bind most frequently at the point of sale because some consumers have overdraft protection on debit cards and some consumers can exceed their credit card limits.

of acquiring and holding cash associated with cash inventory management. Since the costs and benefits of holding cash accrue over multiple transactions, consumers take into account current and future costs and utility when making withdrawal and payment decisions. Importantly, in our blended model consumers can adjust their inflows and outflows of cash holdings *continually*, and thus have an extra margin on which to change cash holdings compared to other models of cash demand in the literature.

5.1 The dynamic problem

The formal consumer's problem involves finding the optimal withdrawal and payment choices of a consumer who settles an infinite sequence of transactions with stochastic transaction values, p. Each transaction involves two sequential decisions: (1) a decision whether to withdraw cash before that transaction, followed by (2) a choice of payment instrument for that transaction.

Consider first the problem of choosing a payment instrument for a consumer who already made her withdrawal decision and holds m dollars of cash in her wallet. She can choose **c**redit, **d**ebit, or cas**h** (provided she has enough) to pay for the current transaction. Following Koulayev et al. (2016), the model contains a random utility framework where each payment method yields an indirect utility flow, $u^i(p) + \epsilon(i)$, associated with each instrument $i = \{c, d, h\}$. The stochastic part of utility, $\epsilon(i)$, is revealed to the consumer just before she chooses the payment instrument and captures the random value of each transaction that depends on payment choice but is unobservable to the econometrician.[21] The deterministic part of utility, $u^i(p)$, depends only on the current transaction value, p, which is assumed to be known by the consumer. However, the consumer does not know future realizations of p or $\epsilon(i)$, only the distributions from which they will be drawn.

At each point-of-sale, the consumer solves the Bellman equation

$$V(m;p) = \max_{i \in \{c,d,h\}} u^i(p) + \epsilon(i) + \beta E\left[W(m';p',b')\right] \tag{1}$$

[21] Examples of the random value may include non-acceptance of cash or card payments; discounts or surcharges associated with a payment instrument; unsafe environments where risk of theft is high for cash or where consumers prefer not to share their card information; and store clerks that are slow at dealing with cash.

where $V(m;p)$ denotes the value of having m dollars of cash before making the current p-dollar transaction, and $E[W(m';p',b')]$ denotes the expected continuation value of reaching the withdrawal decision before the next withdrawal decision with m' dollars of cash. $E[.]$ is the mathematical expectation operator taken over the realizations of all stochastic variables related to the next transaction. The $\epsilon(i)$'s are assumed to be independently and identically distributed Type I extreme value. The law of motion for m is given by $m' = m - p \cdot \mathcal{I}(i = h)$, where $\mathcal{I}$ is an indicator function taking the value of 1 if cash is chosen $(i = h)$ and 0 otherwise. β is used to discount the utility from future transactions.

Prior to each transaction, the consumer decides whether to withdraw cash by solving another Bellman equation,

$$W(m;p,b) = \max_{m^* \geq m} -b \cdot \mathcal{I}(m^* \neq m) - R \cdot m^* + E\left[V(m^*;p)\right], \tag{2}$$

where $W(m;p,b)$ denotes the value of having m dollars of cash before making a withdrawal decision knowing that the next transaction to be financed is p dollars. The withdrawal cost, b, is drawn from a uniform distribution on the interval $[b_L, b_U]$ before each withdrawal decision, while the holding cost of each dollar of cash between transactions is fixed at R. The consumer will increase cash holdings from m to m^* by making a withdrawal $(m^* - m)$ if the expected value of having more cash at the next payment choice, $E[V(m^*;p)]$, exceeds the transaction and opportunity costs of withdrawal. In this case the indicator function $\mathcal{I}(m^* \neq m)$ will equal 1, otherwise it is 0. A unique feature of this model is that *the endogenous withdrawal decision and amount are time-varying* because they depend on the consumer's upcoming transaction value p and on the expected utility of using cash for that transaction.

Assuming consumers know the *exact* value of their next transaction when making withdrawal decisions is convenient and tractable but admittedly strong. It would be preferable to introduce uncertainty about transaction values, but there is no feasible way to infer the magnitude and variation of this uncertainty from the available data. Most of the time, consumers probably know where they plan to shop, what they will buy, and how much they will spend before making a transaction. In reality, consumers may plan spending for multiple future transactions. In any case, the expected transaction value probably is not the unconditional mean of p in reality. The conditional expected transaction value is important because Figure 5 shows that the actual transaction value explains variation in the likelihood of observing a withdrawal for low cash balances reasonably well.

Our specification of withdrawal costs extends the models of Alvarez and Lippi (2009, 2017) where consumers are randomly offered an opportunity to make free withdrawals, which would appear as a Bernoulli distributed b. Table 4 showed numerous methods to obtain cash, which consumers use to varying degrees. Specifying a continuous distribution for withdrawal costs, b, captures this variation in the data simply. The withdrawal cost only has first-order effects on whether consumers make a withdrawal, not how much they withdraw. Withdrawal amounts would vary even more if holding costs, R, also had a stochastic component, which would improve the fit of our estimated model. Unfortunately, the estimation method cannot handle errors in both b and R.[22]

5.2 Timing

Following is a summary of the timing structure of the model.

1. Before each transaction, a consumer with m dollars of cash in her wallet has the option to withdraw cash:

   (i) Random transaction value, p, and random withdrawal cost, b, are realized and observed by the consumer

   (ii) Consumer decides how much cash (if any) to withdraw

   - If withdrawing, consumer adjusts her holdings to m^* and incurs fixed withdrawal cost b and cash holding costs $R \cdot m^*$
   - If not withdrawing, she incurs cash holding costs $R \cdot m$

[22] With an additional shock to R, the one-to-one mapping between the probability of making a withdrawal (observed in the data) and the percentiles of b (the unobserved structural shock) is broken. However, this mapping is crucial, as it allows us to link the observed behavior to the unobserved states of the model when forward-simulating the value functions. See Section 6 and (Ackerberg et al. 2007, , page 103) for more details.

2. After withdrawal decision, the consumer proceeds to the transaction:

    (i) Random components of utility for the current transaction, $\epsilon(i)$, are realized

    (ii) Payment instrument is chosen, $i = \{c, d, h\}$

    (iii) Cash on hand decreases by p, if consumer pays with cash

3. Return to step #1.

6 Estimation

To estimate the model, the deterministic part of the utility function for each payment instrument, $u^i(p)$, is parameterized as

$$u^i(p) = \gamma_0^i + \gamma_{p\leq 10}^i \mathcal{I}(p \leq 10) + \gamma_p^i p \qquad i \in \{c, d, h\},$$

which includes a constant, γ_0, an indicator variable for low-value transactions, $\mathcal{I}(p \leq 10)$, and a linear term in p. The dummy variable for transactions less than \$10 controls for the effects of potential supply-side constraints where vendors do not accept cards due to fees or other costs.[23] If the cash in advance constraint binds, $u^h(p) = -\infty$. The evidence in Section 4 suggests that $\gamma_p^h < 0$ and $\gamma_{p\leq 10}^h > 0$. These utility functions introduce channels for the transaction value to influence payment choice beyond the effects of cash management costs (b and R).

In addition to computational ease, this parsimonious specification of utility is warranted for several reasons. First, Cohen and Rysman (2013) provide evidence from a large U.S. scanner data set that the effect of transaction values on payment instrument choice are not correlated with demographic variables or even individual fixed-effects. Second, although most prior studies use demographic variables as regressors, demographics tend to matter more for adoption of payment instruments than for use conditional on adoption, and our estimation is conditional on adoption of payment cards. Finally, we did not control for card rewards because Section 4.1 showed they had little effect on cash use.

[23] We chose \$10 as the cutoff based on U.S. anecdotal evidence and the discrete drop in the probability of cash use at that transaction value seen in Figures 1 and 3.

The model is estimated using the methods described in Bajari, Benkard, and Levin (2007), or BBL, which is an extension of the Hotz and Miller (1993) conditional choice probability (CCP) estimator used in the empirical IO literature to estimate dynamic structural models with discrete and continuous variables. This approach differs from the methodology used in prior studies of cash management or payment instrument choice. In the monetary literature, dynamic models typically are constructed to yield closed-form solutions for withdrawal policies that can be matched to data using GMM estimators. In the payments literature, static models typically are constructed for discrete choices where the likelihood functions have a closed-form that can be estimated or simulated as in Koulayev et al. (2016).

Like CCP estimators, the BBL procedure has two steps. The first-step involves estimating reduced-form models for state transitions, which are used to characterize the expected value function $E[W(m;p,b)]$. As shown in BBL, the linearity of the utility functions (in structural parameters) and the error specifications imply that $E[W(m;p,b)]$ will be a product of the vector of structural parameters and some basis functions that are derived from the observed choices and state variables. The basis functions can be recovered with forward simulations. In our model, this means: 1) a Pareto-distribution is estimated for transaction amounts; 2) a nonparametric estimate describes payment instrument choice; and 3) the observed nonparametric distribution is used to describe withdrawals. In accordance with Figure 5, separate withdrawal functions are used for when the CIA constraint is binding and non-binding. These reduced-form policy functions are used to construct estimates of the basis functions of $E[W(m;p,b)]$ at a number of grid points in the state space. At each grid-point, we drew 10,000 paths of the stochastic variables with 7,200 transactions for each.[24]

In the second stage of estimation, the structural parameters, $\theta = \{b_L, b_U, R, \gamma_0^h, \gamma_{p\leq 10}^h, \gamma_p^h, \gamma_0^d, \gamma_{p\leq 10}^d, \gamma_p^d, \gamma_0^c, \gamma_{p\leq 10}^c, \gamma_p^c\}$, are recovered using a simulated method of moments estimation as in Pakes, Ostrovsky, and Berry (2007), or POB. β is assumed to be fixed at .995. Cash management costs are restricted to be positive $(b, b_L, b_U, R > 0)$ because

[24] After about 7,200 transactions, the discount factor falls below machine precision so the present value of additional transactions is zero.

they enter equation (2) with negative signs. Using the basis functions from the first-stage simulations and a vector of structural parameters $\hat{\theta}$, the model's prediction is computed for each observation in the sample. As noted in POB, the maximum-likelihood (ML) estimator is not asymptotically efficient because the second stage uses the simulated value function (a function of the basis functions from the first-stage simulations) and not the true value function. Moreover, the ML estimate of the structural parameters can be very sensitive to this error if only a few withdrawals are observed in parts of the state space, resulting in poor small-sample performance. Figure 5 shows this is a realistic concern in the DCPC data.

In the estimation routine, six moments are simulated and matched to their data counterparts: the probabilities of withdrawal for low-value ($m \leq \$25$) and high-value ($m > \$25$) cash holdings; the probabilities of cash use for low-value ($p \leq 10$) and high-value ($p > \$10$) transactions; the average amount of cash purchases; and the average amount of cash withdrawn. Separating withdrawal probabilities for low and high values of cash holdings and transactions is important, as Figure 5 shows these could be quite different. Careful inspection of equation (1) reveals that when the CIA is binding the continuation value of the two remaining options (debit and credit) is the same since $m' = m$ regardless of which payment card is chosen. Therefore, a simple multinomial logit estimation will identify γ_0^d, $\gamma_{p\leq 10}^d$ and γ_p^d. Because the model only identifies utility differences and not the absolute level, we normalize utility from choosing a credit card to zero ($\gamma_0^c = \gamma_{p\leq 10}^c = \gamma_p^c = 0$). The six moment conditions are used to estimate the six remaining structural parameters $\{b_L, b_U, R, \gamma_0^h, \gamma_{p\leq 10}^h, \gamma_p^h\}$.

7 Results

The estimated coefficients are supportive of the theoretical model, as shown in Table 5. All estimates are statistically significant at the 5-percent level or better except the lower bound on cash withdrawal costs (b_L), which is not significantly different from zero. The cash holding and opportunity cost parameters (b_L, b_U, and R) are restricted to plausible

b_L	b_U	R	γ_0^h	$\gamma_{p\leq 10}^h$	γ_p^h	γ_0^d	$\gamma_{p\leq 10}^d$	γ_p^d
0.0003	7.99	0.0049	2.20	0.79	-0.12	.57	.51	-.0037
(0.08)	(1.57)	(0.001)	(0.43)	(0.37)	(0.03)	(0.13)	(0.22)	(0.0016)

Table 5: Structural parameter estimates (standard errors)

ranges, but the remaining unrestricted parameters have expected signs and plausible magnitudes. Relative utility declines with the transaction price for cash (γ_p^h) and debit card (γ_p^d) payments, although the latter is close to zero. Even after controlling for the costs of managing cash, consumers prefer cards for larger transaction values. Cash and debit card payments less than \$10 offer additional relative utility, suggesting that credit cards have lower acceptance or convenience for small-value payments.

The estimates are parameters of a utility function that do not have natural units and thus can be hard to interpret beyond signs. For examples, b_U, b_L and R do not represent a dollar value or rate of interest, respectively, although R represents units of utility *per dollar* by virtue of multiplying cash holdings (m). Thus, the parameter estimates merit additional interpretation.

7.1 Parameter interpretation

A key result is the distribution of cash withdrawal costs $[b_L, b_U]$. Despite the relatively wide estimated range, in our simulations consumers never withdraw cash if withdrawal costs are greater than 4. That is, withdrawals only happen in the most favorable lower half of the estimated distribution; the average withdrawal cost estimate, $\bar{\hat{b}} = -0.75$, reveals that consumers time most of their withdrawals strategically. One way to evaluate the economic magnitude of this relative utility estimate is to compare it with another estimated parameter of the inventory problem, such as the holding cost ($\hat{R}$). In that case, the fixed cost of withdrawals is roughly equal to the utility loss, or “inconvenience," of carrying \$153 $\left(= \bar{\hat{b}}/\hat{R}\right)$ between two transactions.

Another way to gauge the size of the withdrawal cost is to compare it with the benefit of a cash withdrawal that gives a consumer the option to pay with cash, which is particularly valuable for small-value transactions. We measure this benefit as the difference between expected instantaneous utility flow for a consumer who makes a transaction of size p with and without sufficient cash in her wallet. Formally, we calculate

$$\Delta E[u(p)] = \log\left[\sum_{i=\{c,d,\mathbf{h}\}} \exp(u^i(p))\right] - \log\left[\sum_{i=\{c,d\}} \exp(u^i(p))\right],$$

where the log-sum formula computes the expected utility derived from the payment choice. This formula abstracts from continuation values and thus reduces the problem to a multinomial choice model. Comparing this benefit to the fixed cost of withdrawals, it takes about two median-sized transactions to recoup the fixed cost of a withdrawal:

$$\frac{\bar{\hat{b}}}{\Delta E[u(p = 13.41)]} = 1.82$$

About 43 percent of POS payments were \$10 or less (see Figure 1), which explains the popularity of cash even though consumers receive relatively low payment-service utility from large-value cash transactions.

7.2 Cash holdings and use

Using the estimated model and data on cash holdings, Figure 6 illustrates the effects of CIA constraints on the probability of cash use by consumers. The four colored line types in Figure 6 plot the estimated probabilities of cash use for amounts of cash held in wallet ranging from \$25-250. When the CIA binds at the wallet amount, cash probabilities reach zero for larger transactions. Even with a roughly average amount of cash (\$75), consumers are reluctant to use cash for larger transactions; less than 20 percent of purchases of \$30 or more are made with cash. The tradeoff changes rapidly with cash holdings; consumers with \$25 make only about one-third of their very small-value transactions with cash and less than 5 percent of \$20 transactions. In contrast, for large cash holdings (e.g., \$250), the probability of cash use is nearly 80 percent and stable up to \$80.

The results in Figure 6 relate to other recent research. Eschelbach and Schmidt (2013) found that cash in wallets *after* transactions is strongly negatively correlated with the

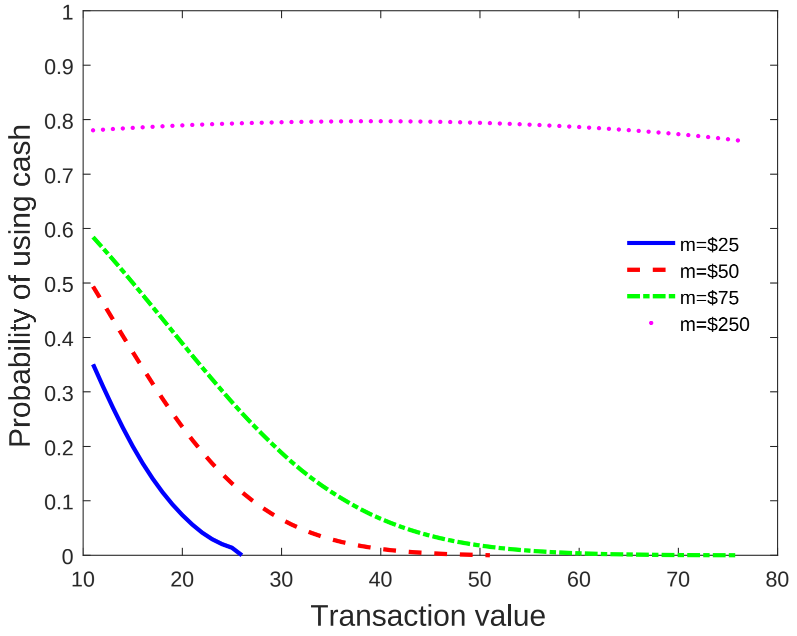


Figure 6: Probability of cash use by transaction value and cash holding

probability of cash use. However, cash holding and withdrawals are jointly determined (see Figure 5), so it is inappropriate to include cash holdings as an explanatory variable in a multinomial logit model without controlling for the endogeneity. Alvarez and Lippi (2017) assume credit card payments are more costly than cash payments on the margin so consumers spend cash as long as they have enough of it—a behavior they call "cash burns." Figure 6 shows this behavior arises even in a model where the relative value of cash payments fluctuates across transactions and consumers can substitute payment cards for cash at each transaction. Thus, consumers with \$75 of cash and above are *more* likely (greater than 50 percent) to use cash for transactions under \$20 than consumers without a binding CIA constraint (see right panel of Figure 1, black dotted line).

The cash-burn result also is illustrated with the estimated model in Figure 7. To minimize withdrawal costs, consumers defer withdrawals and run down cash inventories until a favorable withdrawal opportunity arises, represented by low value of random cost $b \in [b_L, b_U]$. The intuition underlying this behavior appears in the continuation value, $E\left[W\left(m'; p'\right)\right]$, plotted in the left panel of Figure 7 for each amount of cash held *after* a point of sale was made (and *before* the next holding cost shock and transaction value are realized). The continuation value is hump-shaped with a maximum just below \$50.

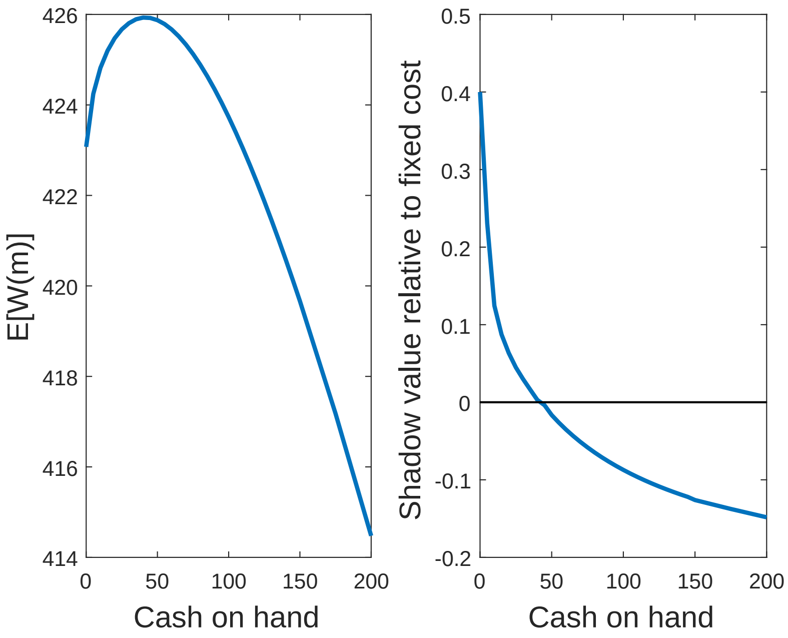


Figure 7: Expected continuation values before holding cost shocks and transaction values are drawn (left); shadow value of an additional dollar in cash (right)

Consumers gladly make cash payments that decrease their holdings to around \$50 but tend to avoid cash purchases that reduce their holdings below \$50.

The shadow value of cash, shown in the right panel of Figure 7, is the marginal utility an extra unit of cash provides by relaxing the CIA constraint for current or future transactions. We compute the shadow value as the difference between the expected continuation values (before p and b are known) of holding $m+1$ and m dollars of cash,

$$\lambda(m) = E[W(m+1;p,b)] - E[W(m;p,b)],$$

where the expectation is taken over the realizations of p and b. The plotted shadow value (right panel) is the derivative of the continuation value (left panel) measured relative to the average cost of withdrawals ($\bar{\tilde{b}} = .75$) for different values of m. The shadow value rises rapidly as cash falls below \$50, reaching about 40 percent of the average withdrawal cost when cash is depleted. But when cash rises above \$50 the shadow value turns negative and declines steadily because consumers are made worse off with more cash. Although having more cash relaxes the likelihood of a binding CIA constraint, consumers with more than \$50 in their wallet are not particularly worried about the constraint because most transactions are low value.

7.3 Consumer welfare

The welfare cost of inflation is a central concern in the monetary literature. Bailey (1956) measured the welfare cost of inflation in a static model with zero-interest money as the area under the interest-elastic money demand curve. More recently, Alvarez and Lippi (2009) computed welfare cost estimates in a dynamic stochastic model with a CIA constraint and inventory management, and Alvarez, Lippi, and Robatto (2019) showed the Baily approach still is appropriate in a wide range of modern inventory theoretic models. However, few studies of money demand consider the effects of payment choice on welfare, so this subsection explores these effects in detail.

7.3.1 Holding costs with instrument choice

Another key result is the magnitude of the estimated cost of holding cash ($\hat{R} = .0049$), which includes the interest elasticity of cash demand among other factors. As holding costs increase, consumers should hold lower cash balances and make more withdrawals, thereby incurring more costs that are pure deadweight loss. However, in a model with non-cash means of payment consumers have an additional margin of response to changes in holding costs—substituting card payments for cash—that may have welfare implications. To gauge the importance of substitution among payment instruments, we simulated the estimated model for different values of the cash holding cost. Because R is a utility parameter, not the interest rate on an alternative asset, we do not know how much R would change if inflation rose one percentage point. Thus, we varied R by about half the estimated value and calculated implied elasticities.

The simulation results in Table 6 reveal the sensitivity of cash management to changes in the holding cost of cash.[25] A 50-percent decrease in the holding cost (.0049 to .0025) would raise cash holdings before a transaction about 44 percent (\$25.49 to \$36.59). This

[25] The reported figures are averages from simulating the choices of 2,000 consumers, who each start with zero cash, for 7,200 periods.

result implies a holding-cost elasticity of demand for cash of −.85, larger in absolute value than the prediction of −0.5 in the basic Baumol-Tobin model. Analogous elasticities for cash holdings before withdrawals and for withdrawal amounts are roughly similar. Table 6 also reveals a non-trivial asymmetry. A roughly 50-percent *increase* in holdings costs (.0049 to .0075) causes cash holdings before a transaction to decline about 24 percent ($25.49 to $19.47), an elasticity of −.44. The probability of making a withdrawal only falls about one-half of 1 percentage point.

	Cash holdings before		Withdrawal		Cash use	Cash	Payment
R	transaction	withdrawal	amount	probability	share	costs	utility
.0025	36.59	15.57	43.94	.049	.35	26.5	465.5
.0030	33.36	14.01	40.48	.051	.34	28.7	464.1
.0035	30.76	13.21	37.25	.053	.33	30.4	462.7
.0040	28.31	11.28	36.22	.052	.33	31.8	461.1
.0045	26.50	11.03	33.23	.055	.32	33.2	459.9
.0049	25.49	10.68	31.90	.056	.32	34.6	459.0
.0055	23.58	9.69	29.71	.058	.31	35.9	457.4
.0060	22.71	9.43	28.77	.058	.31	37.2	456.5
.0065	21.33	8.65	27.68	.058	.30	37.6	454.5
.0070	20.04	8.23	26.14	.059	.30	38.2	453.0
.0075	19.47	7.79	25.77	.059	.30	39.5	452.4

Table 6: Cash management with different cash holding costs

The estimated model exhibits a novel sensitivity of payment choices to holding costs that differs from inventory theoretic models that assume no change in the cash share of payments. The decrease in holding costs induces a modest increase in the share of transactions made with cash from .32 to .35, or about 9 percent, an elasticity of −.2. Given the results in Figure 6, the magnitude of changes in cash holdings and cash share recorded in Table 6 would lead to non-trivial changes in the probabilities of choosing cash. These results reveal that cash holdings are more responsive to R than what standard inventory-theoretic models would predict. Table 6 shows that unless one can directly control for cash spending, estimates of the interest elasticity of cash demand will confound two effects: 1) a change in cash spending, and 2) a change in cash holdings to finance a constant stream of cash spending. Because there is little reason to believe that cash spending remains constant over time when alternative payment methods emerge, there is no reason to believe that the estimated interest elasticity of cash demand should stay constant over time either.

A reduction in holding costs ambiguously improves consumer welfare, defined as payment utility net of cash management costs, for two reasons. Total cash management costs decline (8.1 units of utility), naturally, in part due to a slight decline in the probability of withdrawal. At the same time, payment utility rises by almost the same amount in absolute terms as the reduction in costs (6.5 units of utility) as consumers take advantage of more cash payments. Cash costs fall much more in percentage terms (23.4 percent) than utility rises (1.4 percent), but the absolute changes in utility are similar and the change in net utility is small. In any case, these additional changes in consumer welfare due to changes in payment choices has been missing from previous research on the demand for money.

7.3.2 Withdrawal costs and technological change

As noted in Section 2, the literature widely acknowledges that considerable improvements in technology such as ATM networks and cash back withdrawals from retail stores have reduced the costs of cash management significantly. To measure the effects of technological change in our model, we ran counter-factual simulations with variation in the lower bound of the cash withdrawal cost from the estimated value ($\hat{b}_L = .0003$) to the midpoint of the estimated range ($b_L = 4$) and compared the models' predicted changes in cash management.

Reducing the lower bound of withdrawal costs affects withdrawals notably more than cash holdings or use, as shown in Table 7. The probability of a withdrawal more than doubles (.023 to .056) and the withdrawal amount nearly falls by half ($61 to $32). But cash holdings before a transaction decline less than 20 percent and the cash share only rises 4 percentage points (.28 to .32). As with holding costs, a reduction in cash withdrawal costs make consumers unequivocally better off. These changes primarily impact cash management costs, which fall by one-third (53.2 to 34.6), whereas payment utility rises by just over 1 percent. Collectively, these economically significant changes provide a quantitative guide to the potential effects of recent technological changes.

	Cash holdings before		Withdrawal		Cash use	Cash	Payment
b_L	transaction	withdrawal	amount	probability	share	costs	utility
.0003	25.49	10.68	31.90	.056	.32	34.6	459.0
1	26.49	6.49	43.56	.038	.31	41.3	457.2
2	27.73	5.12	50.66	.031	.30	46.3	456.0
4	29.04	3.56	60.71	.023	.28	53.2	453.1

Table 7: Cash management with different withdrawal costs

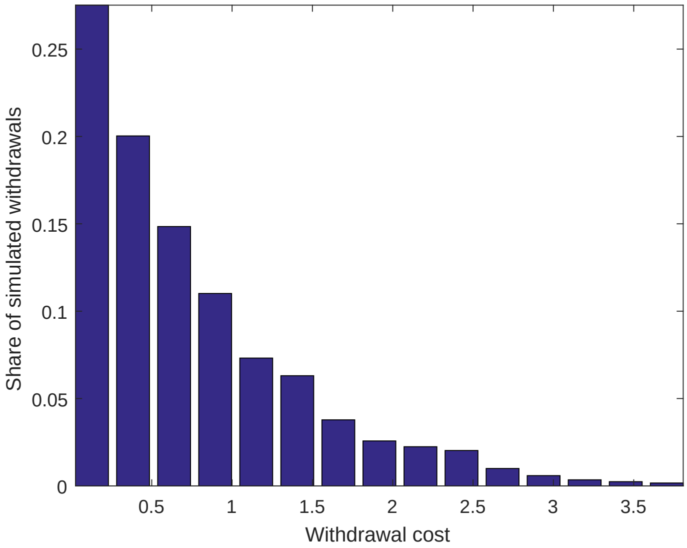


Figure 8: Distribution of simulated withdrawal costs

The estimated costs of withdrawal suggest that the scope for additional cost-saving technology in cash withdrawals going forward may be modest. The full distribution of simulated costs reveals that most are close to zero, as shown in Figure 8, with the median $\hat{b} = .58 < \bar{\hat{b}} = .75$. Some withdrawals are made at high cost, and these might benefit from further technological changes. But the distribution of withdrawal costs decays rapidly from the lower bound because consumers already strategically make most of their withdrawals at the plentiful number of relatively favorable (low-cost) opportunities available to them.

7.3.3 Value of payment instruments

The emergence of electronic means of payment, including credit and debit cards, has coincided with growing anti-cash sentiment. A leading opponent is Rogoff (2016), who describes cash as a “curse” because it aids crime and tax evasion, and constrains monetary policy by inhibiting negative interest rates. Evidence on the consumer welfare of cash relative to other payment instruments is limited and varied, however. Alvarez and Lippi (2017) estimated that eliminating cash altogether and forcing consumers to pay with credit would cost a mere $2 per year, but Alvarez and Argente (2019) find that Uber customers who prefer cash (disproportionately lower income) suffer an average loss of 50 percent of the ride value when they have to use payment cards. Fulford and Schuh (2017) estimated the value of credit card payments is 0.3 percent of annual consumption for convenience users (no high-interest debt). Koulayev et al. (2016) estimated that consumer welfare declines 1-3 percent in response either to a per-transaction fee of 3.6 cents for debit cards or to surcharging credit card payments that offset the merchant discount fee. And consumers lose utility when they prefer cash but it is not accepted for payment, of course.[26]

To measure consumer welfare associated with payment instruments, we simulated the estimated model under different counter-factual scenarios with exclusion of instruments (equivalently, non-acceptance). Table 8 reports simulation results for cash management decisions and consumer utility in each scenario. For reference, the first row repeats the estimation results of the full model with all instruments. See Appendix A for details of modifications made to the model for the counterfactual simulations.

Eliminating any single payment instrument would entail much larger welfare declines than previous simulations. Elimination of debit cards is the most welfare-reducing, as payment utility would be 22 percent lower and cash management costs would more than triple. Eliminating cash would entail an even larger reduction in payment utility (27 percent), but cash management and related costs would disappear so consumer welfare would be slightly higher than without debit cards. Eliminating credit cards is the least welfare-reducing counterfactual, as payment utility falls less than eliminating cash or debit cards, but cash costs increase less than eliminating debit cards. In every case, welfare declines by about an order of magnitude more than in the counterfactual simulations of

[26] None of these studies provides a comprehensive general equilibrium analysis of social welfare, which requires incorporating a market for revolving credit, details of bank and non-bank payment services, and the fee structure of the two-sided credit card markets.

	Cash holdings before		Withdrawal		Cash use	Cash	Payment
Model	transaction	withdrawal	amount	probability	share	costs	utility
Full	25.49	10.68	31.9	.056	.32	16.6	459.0
No cash	0	0	0	0	0	0	336.1
No debit	36.52	15.42	45.3	.072	.47	52.0	357.8
No credit	29.60	12.66	36.8	.063	.37	40.8	401.3
No cards	123.95	55.42	162.1	.177	1.00	219.4	-76.7

Table 8: Cash management with counterfactual payment instruments

changes in cash costs. Note that eliminating just one of the payment cards would not alter dramatically the cash landscape, however. Withdrawal probabilities and cash holdings would be modestly higher, and the cash share would be 5 to 15 percentage points higher; these effects are slightly greater for debit cards.

Eliminating *both* payment cards would make consumers markedly worse off and entail much larger increases in cash activity. Payment utility would decline 117 percent and the cost of cash managment would rise more than 1,300 percent. The probability of cash withdrawals would more than triple to nearly one in five payments being preceded by a withdrawal instead of one in 26. Cash holdings before a transaction would increase roughly five-fold to $124. For perspective on the last outcome, note that Briglevics and Schuh (2013) reports consumers holding $110 (inflation-adjusted to 2010 dollars) in the mid-1980s.[27] At that time, debit cards had not fully diffused yet and credit cards were not used as widely for smaller value payments, so the counterfactual simulation provides a reasonable comparison with actual cash holdings between the two periods.

8 Conclusions

This paper demonstrates that daily transactions-level data on cash demand and payment use from diary surveys can be used successfully to estimate a dynamic optimizing model blending modern elements of cash inventory managment and payment choice. The estimated model shows cash demand and payment use are jointly determined, influencing each other in economically meaningful ways. Two important insights for consumer wel-

[27] See their Table 1 based on the Survey of Currency and Transactions Account Usage conducted by the Federal Reserve Board in 1984 and 1986.

fare are: 1) the level of utility from optimal payment choices is much larger than utility lost from cash management costs; and 2) changes in economic conditions affecting cash management or payment opportunities produce roughly similar magnitudes of change in utility from payment choices and cash costs. Together, the results motivate the need for future research that builds on the blended model.

Relaxing the model's theoretical restrictions on consumers' payment planning is an important direction. Endogenizing the number and value of payments (expenditures), planning more than one payment into the future, allowing for bill payments, and introducing shopping time and trips with multiple payments all could lead to broader and deeper insights. Exploring heterogeneity in cash withdrawal opportunities and management of new payment technologies would enhance understanding as well. Introducing merchant acceptance of payments (as in Hunyh, Nicholls, and Shcherbakov 2019, for example) is essential for capturing demand and supply effects in general equilibrium. More generally, integration of the process of search, exchange, and settlement of transactions that is central to New Monetarist models (as in Chiu and Molico 2010, for example) is a natural direction to extend our framework.

Although impressive and valuable, the new payments diary data merit further development that would enable vital enhancements to the theoretical model. Over time, simply having more data will eventually make it feasible to incorporate variation in the precise costs of withdrawals across locations. But extensions and improvements to the data also are needed. Perhaps most importantly, the balances of non-cash assets and liabilities—especially money in checking or other payment accounts plus credit limits and revolving debt from credit card accounts—are essential for completely characterizing CIA—more generally, liquidity in advance (LIA)—constraints that affect the linkage between portfolio management and settlement of payment for consumer expenditures envisioned by Samphantharak, Schuh, and Townsend (2018). More details about the nature of asset and liability accounts, such as the costs and benefits of specific credit cards, and tracking of the exact payment card or instrument used (instead of a simple category like "credit card") would allow useful enhancements of the theoretical specification of payment utility. Accurately measuring merchant acceptance for each payment opportunity also is essential to relaxing the assumption that sellers accept every payment instrument.

The estimated model's characterization of consumer welfare effects from completely restricting payment instrument use (or acceptance) provides a step toward the evaluation of social welfare and optimal public policies related to currency and other payment systems. However, it is not yet sufficient for comprehensive assessments of the many important policy issues of the day. For example, the future of physical currency in an electronic world that has spawned the re-emergence of private currencies like Bitcoin remains uncertain. And neither regulation of payment card interchange fees, such as Federal Reserve Regulation II, nor provision of payment services with faster or real-time settlement, such as the Federal Reserve's FedNow$^{\text{SM}}$ Service, have been evaluated with an economically adequate specification of consumer demand for money and payments.[28]

Bibliography

Ackerberg, Daniel, C. Lanier Benkard, Steven Berry, and Ariel Pakes. 2007. "Econometric Tools for Analyzing Market Outcomes." In *Handbook of Econometrics*, eds. James Heckman and Edward Leamer, vol. 6A, chap. 63, 4171–4276. Elsevier.

Allais, Maurice. 1947. *Économie & Intérêt.* Paris: Librairie des Publications Officielles.

Alvarez, Fernando, and David Argente. 2019. "Consumer Surplus of Alternative Payment Methods: Paying Uber with Cash." Unpublished working paper.

Alvarez, Fernando, and Francesco Lippi. 2009. "Financial Innovation and the Transactions Demand for Cash." *Econometrica* 77(2): 363–402.

Alvarez, Fernando, and Francesco Lippi. 2013. "The Demand of Liquid Assets with Uncertain Lumpy Expenditures." *Journal of Monetary Economics* 60(7): 753–770.

[28] For more details, see `https://www.federalreserve.gov/paymentsystems/regii-about.htm` for Reg II and `https://www.frbservices.org/financial-services/fednow/index.html` for FedNow$^{\text{SM}}$.

Alvarez, Fernando, and Francesco Lippi. 2017. "Cash burns: An Inventory Model with a Cash-Credit Choice." *Journal of Monetary Economics* 90: 99–112.

Alvarez, Fernando, Francesco Lippi, and Roberto Robatto. 2019. "Cost of Inflation in Inventory Theoretical Models." *Review of Economic Dynamics* 32: 206–226.

Amromin, Gene, and Sujit Chakravorti. 2009. "Whither Loose Change? The Diminishing Demand for Small-Denomination Currency." *Journal of Money, Credit, and Banking* 41(2-3): 315–335.

Angrisani, Marco, Kevin Foster, and Marcin Hitczenko. 2014. "The 2011 and 2012 Surveys of Consumer Payment Choice: Technical Appendix." Research Data Reports 14-2. Federal Reserve Bank of Boston.

Arango, Carlos A., Dylan Hogg, and Alyssa Lee. 2015. "Why Is Cash (Still) So Entrenched? Insights from Canadian Shopping Diaries." *Contemporary Economic Policies* 33(1): 141–158.

Bagnall, John, David Bounie, Kim P. Huynh, Anneke Kosse, Tobias Schmidt, Scott Schuh, and Helmut Stix. 2016. "Consumer Cash Usage: A Cross-Country Comparison with Payment Diary Survey Data." *International Journal of Central Banking* 12(4): 1–61.

Bailey, Martin J. 1956. "The Welfare Cost of Inflationary Finance." *Journal of Political Economy* 64(2): 93–110.

Bajari, Patrick, C. Lanier Benkard, and Jonathan Levin. 2007. "Estimating Dynamic Models of Imperfect Competition." *Econometrica* 75(5): 1331–1370.

Baker, Scott R. 2018. "Debt and the Response to Household Income Shocks: Validation and Application of Linked Financial Account Data." *Journal of Political Economy* 126(4): 1504–1557.

Ball, Laurence. 2012. "Short-run Money Demand." *Journal of Monetary Economics* 59(7): 622–633.

Bar-Ilan, Avner. 1990. "Overdrafts and the Demand for Money." *American Economic Review* 80(5): 1201–16.

Baumol, William J. 1952. "The Transactions Demand for Cash: An Inventory Theoretic Approach." *The Quarterly Journal of Economics* 66(4): 545–556.

Benton, Marques, Krista Blair, Marianne Crowe, and Scott Schuh. 2007. "The Boston Fed Study of Consumer Behavior and Payment Choice: A Survey of Federal Reserve System Employees." Public Policy Discussion Paper 07-1. Federal Reserve Bank of Boston.

Borzekowski, Ron, Elizabeth K. Kiser, and Shaista Ahmed. 2008. "Consumers' Use of Debit Cards: Patterns, Preferences, and Price Response." *Journal of Money, Credit, and Banking* 40(1): 149–172.

Bounie, David, and Yassine Bouhdaoui. 2012. "Modeling the Share of Cash Payments in the Economy: An Application to France." *International Journal of Central Banking* 8(4): 175–195.

Briglevics, Tamás, and Scott Schuh. 2013. "U.S. consumer demand for cash in the era of low interest rates and electronic payments." Working Papers 13-23. Federal Reserve Bank of Boston.

Briglevics, Tamás, and Scott Schuh. 2014. "An Initial Look at How the Electronic Payments Transformation is Changing Consumer Payments." In *The Usage, Costs and Benefits of Cash - Revisited.* Deutsche Bundesbank.

Briglevics, Tamás, Scott Schuh, and David Zhang. 2016. "Homing in Payment Instrument Choice." Unpublished working paper.

Ching, Andrew T., and Fumiko Hayashi. 2010. "Payment card rewards programs and consumer payment choice." *Journal of Banking & Finance* 34(8): 1773–1787.

Chiu, Jonathan, and Miguel Molico. 2010. "Liquidity, Redistribution, and the Welfare Cost of Inflation." *Journal of Monetary Economics* 57(4): 428–438.

Cohen, Michael, and Marc Rysman. 2013. "Payment Choice with Consumer Panel Data." Working Paper 13-6. Federal Reserve Bank of Boston.

Committee on Payments and Market Infrastructure and Markets Committee. 2018. "Central Bank Digital Currencies." Tech. rep. Bank for International Settlements, Basel, Switzerland.

Duca, John V, and William C Whitesell. 1995. "Credit Cards and Money Demand: A Cross-sectional Study." *Journal of Money, Credit and Banking* 27(2): 604–23.

ECB. 2012. "Virtual Currency Schemes." Tech. rep. European Central Bank.

ECB. 2015. "Virtual Currency Schemes – a Further Analysis." Tech. rep. European Central Bank.

Eschelbach, Martina, and Tobias Schmidt. 2013. "Precautionary Motives in Short-term Cash Management - Evidence from German POS Transactions." Discussion Paper Series 2013,38. Deutsche Bundesbank, Research Centre.

Fulford, Scott L., and Scott Schuh. 2017. "Credit card utilization and consumption over the life cycle and business cycle." Working Papers 17-14. Federal Reserve Bank of Boston.

Fung, Ben, Kim Huynh, and Leonard Sabetti. 2012. "The Impact of Retail Payment Innovations on Cash Usage." Working Papers 12-14. Bank of Canada.

Gelman, Michael, Shachar Kariv, Matthew D. Shapiro, Dan Silverman, and Steven Tadelis. 2018. "How Individuals Respond to a Liquidity Shock: Evidence from the 2013 Government Shutdown." *Journal of Public Economics* Online.

Gerdes, Geoffrey, Claire Greene, Xuemei (May) Liu, and Emily Massaro. 2019. "The 2019 Federal Reserve Payments Study." Brief. Federal Reserve System.

Gerdes, Geoffrey R., and Jack K. Walton. 2002. "The Use of Checks and Other Noncash Payment Instruments in the United States." *Federal Reserve Bulletin* 88(8): 360–74.

Greene, Claire, and Scott Schuh. 2016. "The 2016 Diary of Consumer Payment Choice." Research Data Report 17-7. Federal Reserve Bank of Boston.

Greene, Claire, and Scott Schuh. 2017. "The 2016 Diary of Consumer Payment Choice." Research Data Report 17-7. Federal Reserve Bank of Boston.

Greene, Claire, Scott Schuh, and Joanna Stavins. 2016. "The 2014 Survey of Consumer Payment Choice: Summary Results." Research Data Report 16-3. Federal Reserve Bank of Boston.

Greene, Claire, Scott Schuh, and Joanna Stavins. 2018. "The 2012 Diary of Consumer Payment Choice." Research Data Report 18-1. Federal Reserve Bank of Boston.

Hester, Donald D. 1972. "Monetary Policy in the "Checkless" Economy." *The Journal of Finance* 27(2): 279–293.

Hitczenko, Marcin. 2015. "Estimating Population Means in the 2012 Survey of Consumer Payment Choice." Research Data Report 15-2. Federal Reserve Bank of Boston.

Hotz, Joseph V., and Robert A. Miller. 1993. "Conditional Choice Probabilities and the Estimation of Dynamic Models." *Review of Economic Studies* 60(3): 497–529.

Humphrey, David B., Moshe Kim, and Bent Vale. 2001. "Realizing the Gains from Electronic Payments: Costs, Pricing, and Payment Choice." *Journal of Money, Credit, and Banking* 33(2): 216–234.

Hunyh, Kim, Gradon Nicholls, and Oleksandr Shcherbakov. 2019. "Explaining the Interplay between Merchant Acceptance and Consumer Adoption in Two-Sided Markets for Payment Methods." Staff Working Paper 2019-32. Bank of Canada.

Huynh, Kim, Philipp Schmidt-Dengler, and Helmut Stix. 2014. "The Role of Card Acceptance in the Transaction Demand for Money." Discussion Paper 10183. CEPR.

Jack, William, Tavneet Suri, and Robert M. Townsend. 2010. "Monetary theory and electronic money : reflections on the Kenyan experience." *Economic Quarterly* (1Q): 83–122.

Klee, Elizabeth. 2008. "How People Pay: Evidence from Grocery Store Data." *Journal of Monetary Economics* 55(3): 526–541.

Koulayev, Sergei, Marc Rysman, Scott Schuh, and Joanna Stavins. 2016. "Explaining Adoption and Use of Payment Instruments by U.S. Consumers." *RAND Journal of Economics* 47(2): 293–325.

Lagos, Ricardo, Guillaume Rocheteau, and Randall Wright. 2017. "Liquidity: A New Monetarist Perspective." *Journal of Economic Literature* 55(2): 371–440.

Lippi, Francesco, and Alessandro Secchi. 2009. "Technological Change and the Households' Demand for Currency." *Journal of Monetary Economics* 56(2): 222–230.

Lucas, Robert E., and Juan Pablo Nicolini. 2015. "On the stability of money demand." *Journal of Monetary Economics* 73(C): 48–65.

McCallum, Bennett T., and Marvin S. Goodfriend. 1987. "Money: Theoretical Analysis of the Demand for Money." In *The New Palgrave Dictionary of Economics*, eds. John Eatwell, Murray Milgate, and Peter Newman. London: Palgrave Macmillan UK.

Mulligan, Casey B., and Xavier Sala-i-Martin. 2000. "Extensive Margins and the Demand for Money at Low Interest Rates." *Journal of Political Economy* 108(5): 961–991.

Nosal, Ed, and Guillaume Rocheteau. 2011. *Money, Payments, and Liquidity.* MIT Press Books. The MIT Press.

Pagel, Michaela, and Arna Olafsson. 2018. "The Liquid Hand-to-Mouth: Evidence from Personal Finance Management Software." *The Review of Financial Studies* 31(11): 4398–4446.

Pakes, Ariel, Michael Ostrovsky, and Steven Berry. 2007. "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)." *RAND Journal of Economics* 38(2): 373–399.

Prescott, Edward C. 1987. "A Multiple Means-of-Payment Model." In *New Approaches to Monetary Economics*, eds. William A. Barnett and Kenneth J. Singleton, Proceedings of the Second International Symposium in Economic Theory and Econometrics.

Prescott, Edward Simpson, and John A. Weinberg. 2003. "Incentives, Communication, and Payment Instruments." *Journal of Monetary Economics* 50: 433–454.

Reynard, Samuel. 2004. "Financial Market Participation and the Apparent Instability of Money Demand." *Journal of Monetary Economics* 51(6): 1297–1317.

Rogoff, Kenneth S. 2016. *The Curse of Cash.* Princeton, NJ: Princeton University Press.

Rysman, Marc. 2007. "An Empirical Analysis of Payment Card Usage." *Journal of Industrial Economics* 55(1): 1–36.

Samphantharak, Krislert, Scott Schuh, and Robert M. Townsend. 2018. "Integrated Household Surveys: An Assessment of U.S. methods and an innovation." *Economic Inquiry* 56(1): 50–80.

Samphantharak, Krislert, and Robert M. Townsend. 2009. *Households as Corporate Firms.* Cambridge: Cambridge University Press.

Sastry, A. S. Rama. 1970. "The Effect of Credit on Transactions Demand for Cash." *The Journal of Finance* 25(4): 777–781.

Schuh, Scott. 2018. "Measuring Consumer Expenditures with Payment Diaries." *Economic Inquiry* 56(1): 13–49.

Schuh, Scott, and Joanna Stavins. 2010. "Why Are (Some) Consumers (Finally) Writing Fewer Checks? The Role of Payment Characteristics." *Journal of Banking and Finance* 34: 1745–1758.

Schuh, Scott, and Joanna Stavins. 2014. "The 2011 and 2012 Surveys of Consumer Payment Choice." Research Data Reports 14-1. Federal Reserve Bank of Boston.

Sexton, Steven. 2015. “Automatic Bill Payment and Salience Effects: Evidence from Electricity Consumption.” *The Review of Economics and Statistics* 97(2): 229–241.

Shy, Oz. 2013. “How Many Cards Do You Use?” Working Paper 13-13. Federal Reserve Bank of Boston.

Stavins, Joanna. 2001. “Effect of Consumer Characteristics on the Use of Payment Instruments.” *New England Economic Review* (3): 19–31.

Stokey, Nancy L. 2019. “Means of Payment.” Unpublished working paper.

Telyukova, Irina A. 2013. “Household Need for Liquidity and the Credit Card Debt Puzzle.” *Review of Economic Studies* 80(3): 1148–1177.

Telyukova, Irina A., and Randall Wright. 2008. “A Model of Money and Credit, with Application to the Credit Card Debt Puzzle.” *Review of Economic Studies* 75(2): 629–647.

Tobin, James. 1956. “The Interest-Elasticity of Transactions Demand For Cash.” *The Review of Economics and Statistics* 38(3): 241–247.

Tobin, James. 2008. “Money.” In *The New Palgrave Dictionary of Economics*, eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2nd ed.

Townsend, Robert M. 1989. “Currency and Credit in a Private Information Economy.” *Journal of Political Economy* 97(6): 1323–1344.

van der Cruijsen, Carin, Lola Hernandez, and Nicole Jonker. 2015. “In love with the debit card but still married to cash.” DNB Working Papers 461. Netherlands Central Bank, Research Department.

von Kalckreuth, Ulf, Tobias Schmidt, and Helmut Stix. 2009. “Choosing and Using Payment Instruments: Evidence from German Microdata.” Working Paper Series 1144. European Central Bank.

Wakamori, Naoki, and Angelika Welte. 2017. "Why Do Shoppers Use Cash? Evidence from Shopping Diary Data." *Journal of Money Credit and Banking* 49(1): 115–169.

Wang, Zhu, and Alexander L. Wolman. 2016. "Payment Choice and the Future of Currency: Insights from Two Billion Retail Transactions." *Journal of Monetary Economics* 84: 94–115.

Whitesell, William C. 1989. "The Demand for Currency versus Debitable Accounts." *Journal of Money, Credit, and Banking* 21(2): 246–251.

Appendix A Counterfactual Models

For clarity, we briefly spell out the models used in the counterfactual simulations. The simplest cases are the models with cash and one type of payment card. These models retain the structure of the benchmark model (described by equations (1) and (2)), but the payment instrument choice equation (1) only includes either debit or credit cards. Formally, either $i \in \{h, c\}$ or $i \in \{h, d\}$.

A.1 No cash

In these simulations consumers choose between credit and debit cards at the point of sale. The model collapses to a sequence of logit models, with a value function of

$$V(p) = \max_{i \in \{d,h\}} u^i(p) + \epsilon(i) + \beta E\left[V(p')\right]. \tag{3}$$

Since the only endogenous state variable in the benchmark model was cash holdings, decisions made in the current choice situation have no effect on subsequent transactions.

A.2 No cards

The counterfactual model is an extension of the Baumol–Tobin model with stochastic transaction values and withdrawal costs. Consumers choose withdrawal policies to solve

$$\begin{aligned} W(m; p, b) = \max -b \cdot \mathcal{I}(m^* \neq m) - R \cdot m^* + \beta E\left[W(m^* - p; p', b')\right] \\ m^* \geq m, \quad m^* \geq p. \end{aligned}$$

After observing the value of their next transaction, p, and the withdrawal cost, b, consumers decide whether to adjust their cash holdings. Then they make a cash payment (only choice) and move on to another withdrawal decision before their next transaction. Without payment cards, consumers must always have enough cash to pay for the current transaction, p.

The counter-factual model uses the same withdrawal and holding costs as in Table 5, but no utility from card payments. Timing in the counter-factual model also is the same. Thus, consumers know with certainty the amount of their next transaction and are not forced to hold precautionary balances to accommodate the low-probability occurrence of very large-value transactions as in Alvarez and Lippi (2013), which are much less likely for retail payments.

Appendix B Data Appendix

This appendix provides additional details about the Survey (SCPC) and Diary (DCPC) of Consumer Payment Choice and their data. Originally, the SCPC and DCPC were produced by the Federal Reserve Bank of Boston but these data programs are now managed by the Federal Reserve Bank of Atlanta. Data, questionnaires, and associated data reports for each year and survey can be obtained from the Atlanta Fed's consumer payment website.[29] For specific details about the 2012 SCPC and DCPC, see Schuh and Stavins (2014), Angrisani, Foster, and Hitczenko (2014), Hitczenko (2015), and Greene, Schuh, and Stavins (2018).

[29] https://www.frbatlanta.org/banking-and-payments/consumer-payments/.

B.1 Survey Instruments

The SCPC is a 30-minute online questionnaire based on respondent recall that is administered annually each fall beginning in 2008. In most cases, respondents completed the 2012 SCPC at least one day before the DCPC, although the lag may be up to several weeks. SCPC respondents received $20 incentive compensation for completing the survey. The SCPC is taken first and responses are used to tailor the design of the DCPC for each respondent's adoption patterns.

The DCPC is a 20-minute mixed-mode diary survey that was administered for the first time in October 2012. For three consecutive days, respondents were asked to record all payment and cash management transactions in a physical memory aid. Each night, respondents also completed an online survey to report their cash holdings (including denominations) and the transactions recorded in their memory aid, and to answer follow-up questions about the transactions. If they completed the SCPC, DCPC respondents also received additional incentive compensation of $60 for completing all three diary days.

The survey instruments primarily are designed to track payment and cash management activity for nine common instruments: cash, checks (personal, certified, or cashier's), money orders, traveler's checks, debit cards (also ATM cards), credit cards, prepaid cards, online banking bill payment and bank account number payment.[30] The SCPC also measures consumer adoption of bank accounts that are associated with the payment instruments: checking, saving, credit card, and prepaid card (some of which may be managed by non-banks).

Performance of the survey instruments was relatively good in all dimensions. Item response rates for most survey questions were well above 90 percent. Both survey instruments included real-time error checking methods, and respondents had access to RAND staff for technical and conceptual assistance. The vast majority of respondents rated their interest in both surveys as 4 or 5 on a five-point Likert scale (5 being most interesting).

[30] Newer payment instruments such as text/SMS (Venmo and Zelle) and cryptocurrencies (bitcoin) are not included. Applications like PayPal or ApplePay are not payment instruments *per se* but use them to process payments in ways that compete with traditional banking services.

B.2 Sampling Methodology

Respondents in the 2012 SCPC and DCPC were selected from the RAND Corporation's *American Life Panel* (ALP).[31] Currently, the ALP "is a nationally representative, probability-based panel of more than 6,000 participants who are regularly interviewed over the internet." In 2012, however, the ALP was in the process of transitioning from a convenience sample to nationally representative over multiple years. Consequently, the 2012 SCPC and DCPC subsamples of the ALP were randomly re-selected using standard methods to match the U.S. population characterized by the Current Population Survey. The matched 2012 SCPC-DCPC sample included 2,468 respondents who completed all three days of the DCPC. The participation rate of respondents selected for the survey and diary participation was nearly 100 percent. Hitczenko (2015) and Angrisani, Foster, and Hitczenko (2014) provide details of the joint sampling methodology for the 2012 survey instruments.

The primary reporting unit in the ALP is a consumer rather than household. Sampling consumers is easier and less expensive than surveying all members of a household. Consumer-based sampling also is likely to produce better estimates of individual payment choices, especially for currency where the head of household may not track all activity. Sampling consumers could lead to mismeasurement of other aspects of payments, like joint bank accounts and shared household bills like utilities. However, proper random selection of consumers should yield a sample that is representative of U.S. households and produces unbiased aggregate U.S. estimates.[32] A separate quarterly survey provides a wide array of time series demographic characteristics for each ALP consumer that can be merged with the SCPC and DCPC.

[31] See `https://www.rand.org/research/data/alp.html`.

[32] In 2012, the convenience sample nature of the ALP produced around 100 households with two cohabitating adults. This household subsample does not exhibit any large differences from the single-adult sample.

B.3 Survey Design

The SCPC and DCPC were jointly implemented with a common sample of respondents. Starting in September, the SCPC was implemented first and completed prior to the DCPC. In most cases, respondents completed their SCPC at least one day prior to their DCPC. In some cases, the delay may have been a month or so, which could have had minor effects on the synchronization of responses between survey instruments related to adoption of accounts or payment instruments.

Respondents who completed their SCPC were randomly assigned to start their consectutive three-day diaries from September 29 through October 31, with the last diaries being completed on November 2. Each wave of more than 200 DCPC respondents also was randomly selected to be representative of U.S. consumers and staggered across the month so that each day had (in expectation) an equal share of respondents who were completing days one, two, and three of the diary. This procedure is designed to smooth any possible effects of diary fatigue that might lead to incomplete diaries or reduced response quality during a diary period and requires "burn in" (September 29-30) and "cool down" (November 1-2) periods from which the data are not used.

The resulting DCPC data form a balanced longitudinal panel for October 1-31 with fixed entry and exit predetermined by the sampling design and diary methodology. Together, the sampling methodology and survey design make the DCPC sample representative of U.S. consumers for each day of the month and for the entire month. However, the data for individual consumers only extend three days and may not be representative of the individual consumer's monthly payment and cash management behavior. Thus, individual consumer data cannot be projected to the full month.

B.4 Data Measurement

The primary input for this paper is the DCPC transactions-level data on payments and cash management. For payments, the DCPC measures the following seven items: 1) exact time of day (hour, minute, and a.m. or p.m.); 2) the payment value (dollars and cents); 3) the payment instrument; 4) the location (in-person or not); 5) the device used (computer, mobile phone, etc. or none); 6) payment type (retail, person-to-person, or bill); and 7) the merchant type (payee). The SCPC measures payment use as the number of payments per month made (volume), which is measured implicitly in the DCPC as the recorded number of payments per day. However, we do not use the SCPC payment volume data because they rely on respondent recall, hence more susceptible to potential measurement error, and do not include dollar values.[33]

For cash management, the DCPC measures cash holdings (stock) and other cash-related activities (flows). Every night, respondents record the total dollar values of currency held in their "pocket, purse, or wallet" by denomination (the number and value of $1 bills, $5 bills, etc.) but excluding coins. Every day, respondents record the number and dollar values of cash withdrawals by location, cash deposits, and other aspects of cash-related transactions such as conversion of coins to notes.

The 2012 DCPC did not collect stock and flow data on other assets or liabilities, such as bank checking and credit card accounts. The 2012 DCPC collected data on re-loadings of prepaid cards, which are quite similar to cash, but did not collect the balances and withdrawals of specific prepaid cards. Subsequent DCPC's have collected data on balances in *primary* checking accounts only. However, these data are insufficient to track the cash flow of demand deposits if there are multiple accounts, joint account holders, or other complexities in household management of checking account stocks and flows.

B.5 Data Cleaning

For every consumer and every day, the DCPC data should measure exactly the following cash-flow identity:

$$\text{cash tonight} = \text{cash last night} + \text{withdrawals} - (\text{deposits} + \text{cash payments}).$$

In practice, however, there is potential error in this measurement. To minimize the potential measurement error, the online diary survey uses this exact accounting cash-flow

[33] Despite relying on recall, the SCPC data on payment use are surprisingly close to the DCPC estimates except for cash, where the DCPC estimates are significantly higher perhaps due to better tracking.

identity and other techniques for real-time error checking and data correction to ensure that the daily cash-flow identity holds. More than 70 percent of daily consumer-level cash-flow identies held within a rounding error ($1 per transaction allowing for coins).

When individual consumer-day cash-flow identities did not hold, we cleaned the micro data following methods used in other consumer or household surveys that collect dynamic cash data, such as the Townsend-Thai Monthly Survey (see Samphantharak and Townsend 2009). When cash-flow errors were negative, suggesting that respondents spent more cash (or made more deposits) during the day than they recorded, we increased their end-of-day cash holdings sufficiently to eliminate negative cash-flow entries. One explanation for these negative errors is that respondents used cash stored in their home or elsewhere, which was not collected in the 2012 DCPC but is estimated in the SCPC to be much larger than cash in wallet. Measurement errors also may have occurred in reporting of the cash stocks or withdrawals but positive cash-flow errors are smaller and less common. In any case, we trusted respondent reporting of cash management and adjusted end-of-day cash holdings whenever the cash-flow identity was violated.

In the few cases where cash was used to pay bills (which were excluded from the sample), we adjusted the respondent's cash holdings by subtracting the amount of the bill so our measure of cash holdings reflects only cash balances held for making POS transactions. This procedure is not entirely innocuous. For example, consumers who make a large bill payment with cash may make a withdrawal beforehand, in which case they might withdraw cash to cover POS expenses as well. However, our estimation sample has only five instances where a cash bill payment is preceded by a withdrawal that is larger than the amount of the bill payment, so this restriction is unlikely to influence our results. In any case, bill payments often involve different means of payment (online banking, bank account number payment) that are unavailable at the point of sale and likely entail different decision making than POS payments such as planning and budgeting at monthly or annual frequencies. Sexton (2015) also argues that bill payments involve aspects of behavioral economics.

Income, Liquidity, and the Consumption Response to the 2020 Economic Stimulus Payments[*]

Scott R. Baker[†] R.A. Farrokhnia[‡] Steffen Meyer[§]
Michaela Pagel[¶] Constantine Yannelis[‖]

June 3, 2020

Abstract

The 2020 CARES Act directed large cash payments to households. We analyze households' spending responses using high-frequency transaction data from a FinTech nonprofit, exploring heterogeneity by income levels, recent income declines, and liquidity. Households respond rapidly to the receipt of stimulus payments, with spending increasing by \$0.25-\$0.30 per dollar of stimulus during the first weeks. Households with lower incomes, greater income drops, and lower levels of liquidity display stronger responses highlighting the importance of targeting. Liquidity plays the most important role, with no observed spending response for households with high levels of bank account balances. Relative to the effects of previous economic stimulus programs in 2001 and 2008, we see faster effects, smaller increases in durables spending, and larger increases in spending on food, likely reflecting the impact of shelter-in-place orders and supply disruptions. Additionally, we see substantial increases in payments like rents, mortgages, and credit cards reflecting a short-term debt overhang. We formally show that these differences can make direct payments less effective in stimulating aggregate consumption.

JEL Classification: D14, E21, G51
Keywords: Household Finance, CARES, Consumption, COVID-19, Stimulus, MPC, Transaction Data

[*] The most up-to-date version of this paper is available at Columbia Fintech. The authors wish to thank Sylvain Catherine, Arpit Gupta, Jonathan Parker and Joe Vavra for helpful discussion and comments as well as seminar participants at the MoFiR Banking Workshop, the Toulouse School of Economics, Columbia Graduate School of Business, the University of Chicago Booth School of Business and the Federal Reserve Bank of Philadelphia. Constantine Yannelis is grateful to the Fama Miller Center for generous financial support. R.A. Farrokhnia is grateful to Advanced Projects and Applied Research in Fintech at Columbia Business School for support. We would like to thank Suwen Ge, Spyros Kypraios, Rebecca Liu, and Sharada Sridhar for excellent research assistance. We are grateful to SaverLife for providing data access.

[†] Northwestern University, Kellogg, NBER; s-baker@kellogg.northwestern.edu

[‡] Columbia Business School, Columbia Engineering School; farrokhnia@gsb.columbia.edu

[§] University of Southern Denmark (SDU) and Danish Finance Institute (DFI); stme@sam.sdu.dk

[¶] Columbia Business School, NBER, and CEPR; mpagel@columbia.edu

[‖] University of Chicago Booth School of Business, NBER; constantine.yannelis@chicagobooth.edu

1 Introduction

One of the tools used by governments in response to recessions is direct cash payments to households. These payments are generally meant to alleviate the effects of a recession and stimulate the economy through a multiplier effect, i.e., by increasing households' consumption which then translates to more production and employment. The effectiveness of these payments relies on households' marginal propensities to consume, or MPCs, out of these stimulus payments.

In this paper, we estimate households' marginal propensity to consume in response to the 2020 CARES Act stimulus payments using data from a non-profit FinTech. We also look at how these MPCs vary with household characteristics, such as income, income declines, and cash on hand. Finally, we describe how household MPCs vary across categories of consumption and how these categorical responses differ from those seen in previous recessions. Understanding these MPCs is key to targeting policies to households where effects will be largest, as well as testing between different models of household consumption behavior.

MPCs are particularly important to both policy and economic theory as they determine fiscal multipliers in a wide class of models. More specifically, heterogeneity in MPCs impacts which households are most responsive to stimulus payments. In turn, targeting can have large impacts on the effectiveness of stimulus payments on consumption and the aggregate economy. This paper shows that liquidity is a key determinant of MPC heterogeneity during the 2020 contraction, with highly liquid households showing no response to stimulus payments. Even among households with higher levels of income, low levels of liquidity are associated with high MPCs.

We explore responses to stimulus payments and individual heterogeneity in MPCs by using high frequency transaction data from SaverLife, a non-profit that helps families to develop long-term savings habits and meet financial goals. Individuals can link their accounts to the service, and we have access to de-identified bank account transactions and balances data from August 2016 to May 2020 for these users. The fact that we observe inflows and outflows from individual accounts as well as balances in this dataset allows us to explore heterogeneity in levels of income, changes in income, and liquidity.

We use this detailed data to look at CARES Act stimulus payments distributed in April and May 2020. The first stimulus payments were made in mid April via direct deposit from the IRS, and we can observe the user-specific stimulus amounts as well as spending daily before and after stimulus payments are made. We see sharp and immediate responses to the stimulus payments, and continued elevated spending even ten days after payments were received. Within ten days, users spend 29 cents of every dollar received in stimulus payments. The largest increases in spending are on food, non-durables, and payments like rent, mortgages, and student loans. In contrast to the 2008 stimulus payments (**?**), there is relatively little increase in spending on durables.

We exploit the fact that we observe paychecks and account balances to explore heterogeneity across important financial characteristics. Greater income and less liquidity are associated with larger MPCs while recent drops in income seem to have only small effects. Individuals with less than $500 in their accounts spend over one third of their stimulus payments within ten days – 36 cents out of every dollar – while we observe no response for individuals with more than $3,000 in their accounts.

These heterogeneity results are important in terms of targeting stimulus policies towards groups most impacted by them. The theory behind stimulus payments rests on multipliers, which are determined by MPCs in most models. The results of this study suggest that targeting stimulus payments to households with low levels of liquidity in a type of recession where large sectors of the economy are shut down will have the largest effects on MPCs, and hence on fiscal multipliers.

We then show in a macroeconomic model with multiple sectors that untargeted fiscal stimulus payments in environments like the 2020 COVID-19 epidemic may be less effective than the payments in response to the 2001 and 2008 economic downturns. Reflecting the current situation, we map out a three sector model in which one sector employing lower wage agents is shut down while a second low-wage essential sector remains operational alongside a higher-wage sector that can largely work from home.

Due to the shut down of one low wage sector, those poorer and higher MPC agents are largely excluded from benefiting from additional spending induced by stimulus payments, thereby reducing the fiscal multiplier effect. We also see that agents in the lower wage sectors tend to accumulate more debt by borrowing from the higher wage sector. Agents end up using the stimulus payments to repay debt to high wage individuals who have the lowest MPCs out of income. In short, workers will spend their stimulus payment on mortgages and loan repayments as well as non-durable essentials which implies that the cash flows immediately to agents with lower MPCs. This tends to make fiscal stimulus less effective overall.

There is extensive literature on households' responses to tax rebates and previous stimulus payments. The existing studies exploit the differences in timing of the arrival of the payment to infer causal effects. Our results are generally comparable. However, the three main differences are: 1) during the 2020 stimulus, households spend much of their stimulus checks in a shorter period of time, 2) they spend more on food and non-durables than on durable consumption like furniture, electronics, or cars, and 3) they repay credit cards, rent, mortgages, and other overdue bills.

Using spending data from the Consumer Expenditure Survey, **?** and **?** look at the tax rebates granted in 2001 and the economic stimulus payments in 2008. For the 2001 rebates, **?** find that households spend 20-40% on non-durable goods during the quarter in which they received the rebate - the effect also carried over to the next quarter. **?** focus on the stimulus payment in 2008 and find large and positive effects on spending in the same range. The authors document positive effects on spending in both non-durable and durable goods. **?** use high-frequency scanner data and find large positive effects on spending. In Section **??**, we discuss some of the differences between our estimates and the previous literature that analyze previous stimulus programs.

Besides looking at aggregate effects, studies have also found heterogeneous effects across agents. **?** work with credit card accounts and found that customers initially saved the tax rebates in 2001, but then increased spending later on. In their setting, customers with low liquidity were most responsive. **?** use a quantile framework to look at the 2001 tax rebates and the 2008 economic stimulus payments on the distribution of changes in consumption.

? focus on the 2001 tax rebates and use a structural model to document that responsiveness to rebates is driven by liquid wealth. Households with sizable quantities of illiquid assets but low liquidity are an important driver of the magnitude of the response. To our knowledge, our study is the first to look at stimulus payments using high-frequency transaction data, as such data did not exist in 2008.[1] The use of transaction data allows us to explore very-short term responses across categories, minimize measurement error, and explore individual daily heterogeneity in income declines and available cash on hand.

In this paper, we focus on a very different type of contraction relative to those faced during previous stimulus programs: one stemming from an infectious disease outbreak that caused

[1] A number of papers use transaction-level data to look at spending responses to other income, such as **?**, **?**, **?**, **?**, **?**, and **?**. **?** explore some higher frequency weekly responses using Nielsen Homescan data.

widespread business and government shutdowns. In comparison to the 2001 and 2008 economic downturns, the downturn due to COVID-19 was inflicted on households at a much faster pace, causing large job losses much more quickly. In addition, the pandemic has the potential to have large initial effects on income and liquidity, but potentially comparatively less on future income and wealth.

While previous studies have pointed out that stimulus payments have positive but heterogeneous effects on spending, analyzing the 2020 stimulus program will help us learn more about effects on spending in different economic circumstances. In particular, this crisis was so fast moving that households had little ability to increase precautionary savings. Additionally, many sectors of the economy were shut down due to state and local orders, which can impact the effectiveness of fiscal stimulus, as discussed above. Some policymakers argued that shutdowns make conventional fiscal stimulus obsolete.[2]

Our results are also important for the ongoing discussion of Representative Agent Neo-Keynesian (RANK) and Heterogeneous Agent Neo-Keynesian (HANK) models. RANK and HANK models often offer starkly different predictions, and the observed MPC heterogeneity highlights the importance of the HANK framework. In a recent attempt to study pandemics in a HANK framework, **?** show that for income declines up to 70%, consumption declines by 10%, and GDP per capita by 6% in a lockdown scenario coupled with economic policy responses. In another recent working paper, **?** calibrate a HANK model to study the impact of the quarantine shock on the US economy in the case of a successful suppression of the pandemic. In their model, the stimulus payment help stabilize consumption and results in an output decline of less than 3.5%. Additionally, **?** study multipliers in a HANK framework, whose size can depend on market completeness and the targeting of the stimulus.

This paper also joins a fast-growing literature on the effects of the COVID-19 pandemic on the economy, and policy responses. Several papers develop macroeconomic frameworks of epidemics, e.g. **?**, **?**, **?**, and **?**. **?** use stock prices and dividend futures to back out growth expectations. **?** study short-term employment effects and **?** analyze risk expectations. **?** study the targeting and impact of the Paycheck Protection Program (PPP) on employment. **?** and **?** show that political affilia-

[2] For example, Joshua Rauh the former chair of the President's Council of Economic advisers noted that: *"A contraction cannot be addressed via conventional fiscal stimulus since no increase in consumer demand will cause restaurants closed on government orders to re-open."*

tions impact the social distancing response to the pandemic, and **?** study disparities in COVID-19 infections and responses.

Our related paper, **?**, studies household consumption during this onset of the pandemic in the United States using the same data source. **?**, **?**, **?**, **?** perform similar analyses as the one in this paper using transaction-level data from the Spain, Denmark, France, and China. **?** uses transaction-level data from the US provided by merchants rather than individual-level data and find similar results to **?**. We join this emerging and rapidly-growing literature by providing early evidence on how households responded to the crisis and on the details of the impacts of federal stimulus policy. The results suggesting that MPCs are much higher for low liquidity households are important in designing future rounds of stimulus, if the effects of the epidemic persist over the next months.

The remainder of this paper is organized as follows. Section **??** provides background information regarding the 2020 stimulus and our empirical strategy. Section **??** describes the main transaction data used in the paper. Section **??** presents the main results and Section **??** discusses heterogeneity by income, income drops, and liquidity. Section **??** compares around findings to similar stimulus programs, discusses results for mortgage, credit card and other payments and presents a simple model to explain how fiscal multiplier effects may differ from prior stimulus programs. Section **??** concludes and suggests directions for future research.

2 Institutional Background and Empirical Strategy

2.1 2020 Household Stimulus

COVID-19, a novel coronavirus, was first identified in Wuhan, China and subsequently spread worldwide in early 2020. By some estimates, the new virus had a mortality rate which is ten times higher than the seasonal flu and has at least twice the rate of infection. The first case in the United States was identified in late January in Washington state and spread within the country in February. By mid-March, the virus was spreading rapidly, with significant clusters in New York, San Francisco, and Seattle. Federal, state, and local governments responded to the COVID-19 pandemic in a number of ways: by issuing travel restrictions, shelter-in-place orders, and closures of many non-essential businesses.

The federal government soon passed legislation aimed at ameliorating economic damage stemming from the spreading virus and shelter-in-place policies. The CARES Act was passed on March 25, 2020 as a response to the economic damage of the new virus. The Act deployed nearly $2 trillion across a range of programs for households and businesses. This study focuses on the portion of the Act that directed cash transfers to the vast majority of American households. These one-time payments consist of $1,200 per adult and an additional $500 per child under the age of 17. For an overview of amounts by household, see Appendix Table **??**. These amounts are substantially larger than the 2001 and 2008 stimulus programs. In 2020, a married couple with two children would be sent $3,400, a significant amount, particularly for liquidity-constrained households.

Most American households qualified for these payments. All independent adults who have a social security number, filed their tax returns, and earn below certain income thresholds qualified for the direct payments. Payments begin phasing out at $75,000 per individual, $112,500 for heads of households (single parents with children), and $150,000 for married couples. No payments were made to individuals earning more than $99,000 or married couples earning more than $198,000.[3]

Payments are made by direct deposit whenever available, or by paper check when direct deposit information was unavailable. Funds are disbursed by the IRS, and the first payments by direct deposit were made on April 9th. The IRS expected that direct deposits would largely be completed by April 15th. In practice, the timing varied across banks and financial institutions, with some making payments available earlier than others, and direct deposits being spread out across more than one week. Amounts and accounts for direct deposits were determined using 2019 tax returns, or 2018 tax returns if the former were unavailable.

For individuals without direct deposit information, paper checks were scheduled to be mailed starting on April 24th. Approximately 70-80% of taxpayers use direct deposit to receive their tax refunds, though given changes in banking information or addresses, many individuals were unable to receive their payments through direct deposit even when they had received prior tax refunds via direct deposit. In the case of paper checks, the order of payments across households is not random. The IRS directed to send individuals with the lowest adjusted gross income checks first in late April, and additional paper checks will be sent throughout May. Appendix **??** provides further details regarding the timing of payments and the stimulus.

[3] Due to data limitations, in identifying stimulus payments, we are unable to identify these partial payments from these higher-income households. However, these individuals are a very small fraction of total households, both overall and particularly among our sample which is skewed towards lower income households.

2.2 Empirical Strategy

Our empirical strategy exploits our high-frequency data and the timing of stimulus payments to capture spending responses. We first show estimates of β_i from the following specification:

$$c_{it} = \alpha_i + \alpha_t + \sum_{t=-7}^{23} \beta_i \mathbb{1}[t = i]_{it} + \varepsilon_{it} \tag{1}$$

c_{it} denotes spending by individual i aggregated to the daily level t. α_i are individual fixed effects, while α_t are date fixed effects. Individual fixed effects α_i absorb time invariant user-specific factors, such as some individuals having greater average income or wealth. The date fixed effects α_t absorb time-varying shocks that affect all users, such as the overall state of the economy and economic sentiment. $\mathbb{1}[t = i]$ is an indicator of a time period i days after receipt of the stimulus payment.

In some specifications, we interact individual fixed effects with day of the week or day of the month fixed effects to capture consistent time-varying spending patterns over the week and month. For example, some individuals may spend more on weekends, or on their paydays. We run regressions at an individual-day level to examine more precisely the high frequency changes in behavior brought about by the receipts of the stimulus payments. Standard errors are clustered at the individual level. The coefficient β_i captures the excess spending on a given day before and after stimulus payments are made. In our graphs, the solid lines show point estimates of β_i, while the dashed lines show 95$\%$ confidence intervals.

We identify daily MPCs using the following specification:

$$c_{it} = \alpha_i + \alpha_t + \sum_{t=-7}^{23} \gamma_i P_i \times \mathbb{1}[t = i]_{it} + \varepsilon_{it} \tag{2}$$

where P_i are stimulus payments for individual i at time t. To identify cumulative MPCs since the first payment, we scale indicators of a time period being after a stimulus payment by the amount of the payment over the number of days since the payment. That is, our estimate of a cumulative MPC ζ comes from the following specification:

$$c_{it} = \alpha_i + \alpha_t + \zeta \left(\frac{Post_{it} \times P_i}{D_{it}} \right) + \varepsilon_{it} \tag{3}$$

where P_i is the stimulus payment an individual i is paid, and D_{it} is the total number of days over which we estimate the MPC and $Post_{it}$ is an indicator of the time period t being after individual i receives a stimulus payment. The coefficient ζ thus captures the aggregate effect of of the stimulus in the time period in question, by scaling the average effect per day by the number of days since receipt. The resulting coefficients can be interpreted as the fraction of stimulus money spent during that period: a coefficient of 0.05 corresponds to the user spending 5% of their stimulus check during their observed post-stimulus period.[4]

3 Data

3.1 Transaction Data

In this paper, we utilize de-identified transaction-level data from SaverLife, a non-profit helping working families develop long-term savings habits and meet financial goals. As with a number of other personal financial apps, SaverLife allows users to link their main bank accounts to their service. Users can link their checking, savings, as well as their credit card accounts. SaverLife offers users the ability to aggregate financial data and observe trends and statistics about their own spending.

Figure **??** shows two screenshots of the online interface in the app. The first is a screenshot of the linked main account while the second is a screenshot of the savings and financial advice resources that the website provides. This data is described in more detail in **?**.

Overall, we have been granted access to de-identified bank account transactions and balances data from August 2016 to May 2020. We observe 44,660 users in total who live across the United States. In addition, for a large number of users, we are able to link financial transactions to self-reported demographic and spatial information such as age, education, ZIP code, family size, and the number of children they have.

We also observe a category that classifies each transaction. Spending transactions are cate-

[4] As an example to illustrate this, imagine that a \$1 transfer leads to \$1 dollar of additional spending in the day immediately after receipt. Thus if we estimated the effect over one day, we would scale by 1 and $\zeta = 1$. If we estimate the effect over 10 days, the average effect each day is 0.1, which would be the coefficient on a regression of $Post_{it} \times P_i$ and we scale by 10 so again $\zeta = 1$. If we estimate the effect over 100 days, the average effect per day is 0.01, again we would scale by 100 and so on.

gorized into a large number of categories and subcategories. For the purposes of this paper, we mostly analyze and report spending responses into the following aggregated categories: food, household goods and personal care, durables like auto-related spending, furniture, and electronics, non-durables and services, and payments including check spending, loans, mortgages, and rent. Across all specifications, we exclude transactions that represent transfers between accounts like transfers to savings or investment accounts.

Looking only at the sample of users who have updated their accounts reliably up until May 2020, we have complete data for 6,033 users to analyze in this paper. We require these users to have several transactions per month in 2020 and have transacted at least $1,000 in total during these three months of the year. Requiring regular prior account usage is frequently used as a completeness-of-record check when using bank-account data (**?**).

In Table **??** we report descriptive statistics for users' spending in a number of selected categories as well as their incomes at the monthly level. We note that income is relatively low for many SaverLife users, with an average level of post-tax income being approximately $25,000 per year. In addition, we show the distribution of balances across users' accounts during the week before most stimulus checks arrived. Consistent with the low levels of income, we see that most users maintain a fairly low balance in their linked financial account, with the median balance being only $141.02.

We identify stimulus payments using payment amounts stipulated by the CARES Act, identifying all payments at the specific amounts (eg. $1,200, $1,700, $2,400) paid after April 9th in the categories 'Refund', 'Deposit', 'Government Income', and 'Credit.' Figure **??** shows the identified number of payments of this type, relaxing the time restrictions in 2019 and 2020. While there are a small number of payments in these categories at the exact stimulus amounts prior to the beginning of payments, there is a clear massive increase in frequency after April 9th. This suggests that there are relatively few false positives, and that the observed payments are due to the stimulus program and not other payments of the same amount.

As of May 16th, approximately 53% of users have received a stimulus payment into their linked account. The remainder of the sample may be still waiting for a stimulus check or may be ineligible for one. Some banks and credit unions had issues processing stimulus deposits and these deposits were still pending for a number of Americans. In addition, users may not have had direct deposit information on file with the IRS and would then need to wait for a check to be mailed. Finally, users may be ineligible for stimulus checks due to their status as a dependent, because they did not file their taxes in previous years, or because they made more than the eligible income thresholds for receipt. Of those who receive payments, two-thirds received them by April 15, with 40% of all payments occurring on April 15. 92% of those who received payments in our sample did so in April.

While most American households were due to receive a stimulus check, the amount varied according to the number of tax filers and numbers of children. Table **??** gives an accounting of amounts due to a range of household types. While we cannot observe the exact household composition for each user, we are able to observe a self-reported measure of household size. Our measure matches up reasonably well with the received stimulus payments.

Appendix **??** provides further details regarding payments in our sample. Payments line up closely with self-reported household size. Because of our strategy for picking out stimulus checks, being within the 'phase-out' region of income would mean that we would falsely classify an individual as having not received a stimulus check, since his or her check would be for a non-even number. This would likely attenuate our empirical estimates slightly. We conduct a placebo exercise in the appendix, and look at spending around April for households that do not receive a check. We do not see any sharp breaks in spending beyond day of the week effects, suggesting that the impact of mis-categorization is small.

4 Effects of Stimulus Payments

Looking at the raw levels of spending for users receiving stimulus payments, Figure **??** shows mean daily spending before and after the receipt of a stimulus payment without any other controls or comparison group. In this figure, we only show spending data for users who receive a stimulus check in our sample period. Prior to receiving a check, the typical individual in the sample is spending under \$100 a day. There is a sharp and immediate increase in spending following the receipt of a stimulus deposit. Mean daily spending rises on the day of receipt to approximately \$150 and continues to increase, to over \$200, for the two days after the receipt of the stimulus payment.

Observed spending declines substantially in the third and fourth days, though most of this is driven by the fact that a plurality of 'treated' users in our sample received the stimulus check on Wednesday, April 15th and spending tends to decline on weekends. After the weekend period, observed spending rises to $250 before beginning to decline.[5]

While Figure **??** provides some evidence that spending was affected by the stimulus payments, we want to directly compare users receiving stimulus payments to those that did not receive one on that day. Figure **??** shows estimates of β_i from the equation: $c_{it} = \alpha_i + \alpha_t + \sum_{t=-7}^{23} \beta_i \mathbb{1}[t = i]_{it} + \varepsilon_{it}$. 'Time to Payment' is equal to zero for a user on the day of receiving the stimulus check. Here, we see that users who receive stimulus checks tend to not behave differently than those that do not in the days before they receive the checks. Upon receiving the stimulus check, users dramatically increase spending relative to users who do not receive the checks.

Similar to what we saw looking at the raw spending data, users show large increases in spending in the first days following the stimulus check receipt and keep spending significantly more than those who have not received checks for the entirety of the post-check period that we observe. The relative difference in spending declines during weekends, mostly driven by the fact that observed levels of spending tend to be depressed during these days for reasons described above.

In Figure **??**, we break down users' spending responses by categories of spending. We map our categories to roughly correspond to those reported in **?** from the CEX: food, household goods and personal care, durables like auto-related spending, furniture, and electronics, non-durables and services, and payments including check spending, loans, mortgages, and rent.

Across all categories, we find statistically significant increases in spending following the receipt of a stimulus check. These responses are widely distributed across categories, with cumulative spending on food, household, non-durables, and payments each increasing by approximately $50-$75 in the three days following receipt of a check. Durables spending sees a significant increase, but it is much smaller in economic terms with only a $20 relative increase in spending during the first three days.

Table **??** presents similar information, presenting coefficients from the regression $c_{it} = \alpha_i +$

[5] Observed spending tends to decline dramatically on weekends throughout our sample. This is likely driven by two factors. The first is that actual transactions and spending declines during these days. The second is that transactions that occurred during the weekend may process only on the Monday that follows. We are unable to distinguish between these cases using our data.

$\alpha_t + \sum_{t=-7}^{23} \beta_i \mathbb{1}[t = i]_{it} \times P_i + \varepsilon_{it}$. That is, we examine the excess spending among users who received stimulus payments on each day following the receipt of their stimulus checks, scaled by the size of their payment. A value of 0.03 can be interpreted as the user spending, on day t, 3% of their stimulus check (eg. $36 out of a $1,200 stimulus check) more than a user who did not receive a check.

Columns 1-3 test how total user spending responds with three different sets of fixed effects. Column 1 presents results using individual and day of the month fixed effects. Column 2 also includes individual-by-day-of-month fixed effects, and Column 3 includes individual, calendar date, and individual-by-day-of-week fixed effects. We find similar effects across all specifications, with spending among those who received a stimulus check tending to increase substantially in the first week after stimulus receipt.

Spending on days during this period is economically and statistically significantly higher for those receiving stimulus checks and there are no days with significant reversals – days with stimulus check recipients having lower spending than those who did not. Overall, for each dollar of stimulus received, households spent approximately $0.25-0.3 more in the month following the stimulus.

The remainder of the columns in Table **??** decompose the effect that we see in overall spending according to the category of spending. We find significant increases in spending in all of these categories, with the largest increases coming from non-durables and payments. We find muted effects of the stimulus payments on durables spending. In previous recessions, noted by **?**, spending on durables (mainly auto-related spending), was a large component of the household response to stimulus checks. At least in the short-term, we find significantly different results, with durables spending contributing negligibly to the overall household response. We discuss some of these differences relative to past stimulus programs in Section **??**

5 Income, Liquidity, and Drops in Income

The 2020 CARES Act stimulus payments were sent to taxpayers with minimal regard for current income, wealth, and employment status. While there was an income threshold above which no stimulus would be received, this threshold was fairly high relative to average individual income and most Americans were eligible for payments. During debates about the size and scope of the stimulus, a common question was whether Americans with higher incomes, unaffected jobs, and higher levels of wealth needed additional financial support. With data on both the income and bank balances of SaverLife users, we are able to test whether the consumption and spending responses differed markedly between users who belonged to these different groups.

In Figures **??**-**??**, we show the cumulative estimated MPCs from regressions of spending on an indicator of a time period being after a stimulus payment is received. Each figure contains the results of multiple regressions, with users broken down into subsamples according to a number of financial characteristics that we can observe. That is, the graphs represent the sum of daily coefficients seen in a regression as in Table **??**, by group. In these figures, we divide the samples of users by their level of income, the drop in income we observed over the course of 2020, and their levels of liquidity prior to the receipt of stimulus payments.

Figure **??** splits users by their average income in January and February 2020 (prior to the major impacts of the pandemic). We see clear evidence that users with lower levels of income tended to respond much more strongly to the receipt of a stimulus payment than those with higher levels of income. Users who had earned under $1,000 per month saw an MPC more than twice as large as users who earned $5,000 a month or more.

We also split our sample of users according to their accounts' balances at the beginning of April, before any stimulus payments were made. We separate users into four groups: those with balances under $500, between $500 and $1,000, between $1,000 and $3,000, and over $3,000. Figure **??** displays results from these four regressions. We see dramatic differences across groups of users. Users with the highest balances in their bank accounts tend not to respond to the receipt of stimulus payments, while those who had under $500 respond the most. The low balance group has an MPC out of the stimulus payment of about 0.36 across the following weeks.

In Figure **??**, we examine whether a similar pattern can be seen among users who have had declines in income following the COVID outbreak. For each user, we measure the change in income received in March 2020 relative to how much was received, on average, in January and February 2020. We split users into those who had a decline in monthly income and those who saw no decline in income (or had an increase). In contrast with heterogeneity across levels of income and levels of liquidity, we find only a weak difference between these two groups. This may be driven by the fact that the federal government had also made generous unemployment insurance available to nearly all workers, mitigating the potential loss of income from job loss for many lower income households.

Table **??**, Table **??**, and Table **??** display some of these results in regression form. In general, we find that users with lower incomes, larger drops in income, and lower pre-stimulus balances tend to respond more strongly than other users. Again, across all subsamples of our users based on financial characteristics, we see that low liquidity tends to be the strongest predictor of a high MPC and high liquidity tends to be the strongest predictor of low MPCs.

Of particular note, in Table **??**, is the fact that we find that MPCs among low-liquidity individuals tend to be high even for those with relatively high levels of income. High income and low balance individuals have MPCs that are significantly different than high income and high balance individuals, but are indistinguishable from the MPCs among low balance individuals, in general. That is, when splitting the sample by both levels of income and liquidity, a household's liquidity tends to drive the observed MPC from stimulus payments to a much larger extent than a household's income.

6 The 2020 Stimulus and Previous Economic Stimulus Programs

6.1 Comparison to Previous Economic Stimulus Programs

? and **?** examine the response of households to economic stimulus programs during the previous two recessions (2001 and 2008). These programs were similar in nature to the stimulus program in 2020 but were smaller in magnitude. In 2001, individuals generally received $300 rebates, with married couples generally eligible for $600. In 2008, couples could receive $1,200 and $300 for each dependent child. In the 2020 stimulus program, couples could receive $2,400 and each dependent child would be eligible for $500.

In these previous stimulus programs, households also tended to respond strongly to the receipt of their checks. For instance, in 2008, **?** estimated that households spent approximately 12-30% of their stimulus payments on non-durables and services and a total of 50-90% of their checks on total additional spending (including durables) in the six months following receipt. In 2001, approximately 20-40% of stimulus checks were spent on non-durables and services in the six months following receipt.

In one paper examining the high-frequency responses (**?**), the authors are able to use Nielsen Homescan data to examine weekly spending responses to the 2008 stimulus payments. They find that a household's spending on covered goods increased by approximately ten percent in the week that it received a payment, with an MPC of approximately 0.5 by the month after stimulus check receipt. Spending remained elevated for approximately three months following stimulus payment receipt, although more than 70% of the excess spending is in the month that the check is received.

While they were not always able to examine the timing of all types of spending in more detail due to data limitations in previous recessions, we demonstrate that households respond extremely quickly to receiving stimulus checks. Rather than taking weeks or months to spend appreciable portions of their stimulus checks, we show that households react extremely rapidly, with household spending increasing by approximately one third of the stimulus check within the first 10 days. Given that previous stimulus programs saw sustained increases in spending lasting six months or more, we would expect that the long-run impact of the 2020 stimulus program would be much larger than the already sizable short-run effect that we have seen so far.

Another notable difference from the stimulus program during the 2008 recession is the variation in magnitudes and spending responses across categories. We find smaller estimates of MPCs, which is driven by low durable spending. In non-durable categories, we find similar estimates relative to previous work. Previous research has found strong responses of durables spending to large tax rebates and stimulus programs, especially on automobiles (about 90% of the estimated impact on durables spending in the 2008 stimulus program was driven by auto spending). In contrast, despite a sizable response in non-durables and service spending, we see little immediate impact on durables. Even if we attribute the entirety of our observed response in the 'Payments' category to spending on durables, the magnitude is much smaller than the combined response in food and non-durables categories. Moreover, the payments category also includes rent and bill payments which compose a portion of the 'Payments' category increase.

This difference becomes even starker if we consider the fact that some prior literature has shown that larger payments often result in spending responses that skew more towards durables.

Given the size of the 2020 stimulus checks, we might have expected large impacts on categories like automobile spending, electronics, appliances, and home furnishings. Instead it seems that individuals are catching up with rent and bill payments as well as engaging in spending on food, personal care, and non-durables.

In part, this discrepancy with past recessions may be driven by the fact that automobile use and spending is highly depressed, with many cities and states being under shelter-in-place orders and car use being restricted. Similarly, as these orders hinder home purchases, professional installment, and moves, spending on home furnishings and other related durables may be lower as well (the stimulative effects of home purchases on home durables are demonstrated in **?**).

While increases in durables spending were limited in the 2020 stimulus setting, we find substantial increases in spending on food. This again stands in contrast to some of the effects seen in earlier stimulus programs. Again, it may reflect the unique economic setting in which the 2020 economic stimulus took place. While many outlets for consumer spending were closed by government order, restaurants remained open; we find that household spending on food delivery was one category in particular that increased following the receipt of a stimulus check.

Finally, across both 2001 and 2008, **?** note that lower income households tend to respond more, and that households with either larger declines in net worth or households with lower levels of assets also tend to respond more strongly to stimulus checks. These results are largely consistent with the patterns we observe in 2020. We find that households with low levels of income and lower levels of wealth tend to respond much more strongly. In addition, our measure of available liquidity from actual account balances arguably suffers from much less measurement error than the measures used in previous research on stimulus checks, giving additional confidence in our estimates.

6.2 Payments

In Figure **??** and Table **??**, we report the impact of the stimulus check on financial payments. In particular, we examine the impact on total financial payments as well as payments on several subsets of financial payments such as credit card payments as well as rent and mortgage payments. Rent payments are not always able to be accurately identified due to the number of users who utilize checks or online transfer tools like Chase QuickPay, Zelle, or Venmo to pay their rent. Such payments will still be accurately captured by the 'Total Financial Payments' category.

We find that financial payments surge substantially upon receipt of the 2020 stimulus payments. Marginal spending on total financial payments totals about one third of total MPC out of the stimulus payments. In the following subsection, we argue that our empirical findings imply that the fiscal stimulus payments may be less effective in stimulating aggregate consumption in the 2020 environment relative to previous downturns.

6.3 Modeling the Effectiveness of Fiscal Stimulus Payments

We now present a simple model that outlines two reasons, consistent with our empirical findings, that the fiscal stimulus in 2020 may be less effective in actually stimulating the economy than the 2001 or 2008 payments. The basic reason for this lack of effectiveness is that sectors of the economy employing workers with the lowest levels of liquidity are shut down, leading to lower fiscal multipliers.

Suppose that we have three types of sectors and workers employed by those sectors. First, we have a sector that we call groceries and necessities. Here, we refer to large firms that sell groceries and basic household supplies that are both essential and non-durable (moderate depreciation). For instance, large supermarkets or stores such as Target, Walmart, and CVS. At the same time, the grocery and necessity sector is moderately labor intensive. This sector is not shut down in response to an epidemic.

In turn, we have a second sector, called restaurants and hospitality, that produces non-durable consumption which depreciates immediately and is more labor intensive than the first sector. Being less essential to households, the second sector is shut down in response to the crisis.

Finally, we have a third sector of the economy. This sector is broader, and encompasses durables production as well as many white-collar services like banking and tech. This sector can avoid being locked down through employing safety measures in production or by working remotely. This sector pays higher wages than in sectors 1 and 2. Consequently, the corporations in sectors 1 and 2 are owned by the workers in sector 3. We assume that workers in sector 1 and 2 borrow (for example, rent, mortgages, or financial lending) from workers in sector 3.

The effectiveness of fiscal stimulus rests on the idea that stimulus checks induce extra spending by recipients. For example, workers in sector 3 spend in sector 2 and generate income for workers in that sector that is then spent again. Thus, if the MPC out of a stimulus payment is 0.8, then out of a \$100 payment, \$80 is consumed, generating \$80 of income for another worker. That worker then again consumes \$64 which generates income for another worker, and so on. In the classic Keynesican framework the equation for the fiscal multiplier is given by $1/(1 - MPC)$. The more cash arrives with agents that have high MPCs, the higher the fiscal multiplier.

In our framework, there are two reasons why fiscal stimulus is less effective in this environment relative to the 2001 and 2008 recessions. First, in a lockdown induced by an epidemic, neither group of workers can spend in sector 2. At the same time, workers in sector 2 are the poorest and have the highest MPCs. Second, workers in sectors 1 and 2 (that are poorer) use the stimulus payment to pay down debt held by sector 3 workers. Therefore, the excess spending from the stimulus flows to workers that have a lower MPC.

More formally, we have a three-period model inspired by **?** and consider an economy with three sectors. All sector s agents' preferences are represented by the utility function:

$$\sum_{t=0}^{3} \beta^t U(c_t^s) \tag{4}$$

where c_t^s is consumption and $U(c) = c^{1-\sigma}/(1-\sigma)$ is a standard power utility function. Each agent is endowed with $\bar{n}_t^s > 0$ units of labor which are supplied inelastically but they can only work in their own sector. Competitive firms in each sector s produce the final good from labor using the linear technology:

$$Y_t^s = \bar{n}_t^s. \tag{5}$$

Each agent maximizes utility subject to:

$$c_t^s + a_t^s \leq w_t^s \bar{n}_t^s + (1 + r_{t-1}) a_{t-1}^s. \tag{6}$$

As the initial condition, we assume that agents in sectors 1 and 2 borrow from agents in sector 3, such that $a_1^1 < 0$, $a_1^2 < 0$, and $a_1^1 + a_1^2 = -a_1^3$. Given the economy is frictionless, agents choose their consumption to satisfy their Euler equation:

$$U'(c_t^s) = \beta(1 + r_t) U'(c_{t+1}^s). \tag{7}$$

Because preferences are homothetic, we can think of all agents in each sector as just being represented by one agent. In turn, each agent can consume consumption goods from any sector, denoted by c_t^{ss}. The consumption composite, c_t^s, over the three sectors' consumption goods equals $f_c(c_t^{s1}, c_t^{s2}, c_t^{s3})$ and relative goods prices meeting the composite constraint $p_t c_t^s = p_t^1 c_t^{s1} + p_t^2 c_t^{s2} + p_t^3 c_t^{s3}$ adjust to ensure full employment in each sector. Additionally, we assume that $\frac{\partial f_c}{c_t^{s1}}|_{c_t^{s1} \to 0} = \infty$ whereas $\frac{\partial f_c}{c_t^{s2}}|_{c_t^{s2} \to 0}$ and $\frac{\partial f_c}{c_t^{s2}}|_{c_t^{s2} \to 0}$ approach finite numbers, which implies that consumption purchased in sector 1 is necessary, whereas it is not necessary when it comes from sectors 2 and 3. Finally, the goods market clearing condition has to hold in each period:

$$c_t^{1s} + c_t^{2s} + c_t^{3s} = \bar{n}_t^s. \tag{8}$$

Suppose the central bank implements a fixed rate $1 + r_0 = 1/\beta$ and the economy starts from a state in which each agent consumes his or her labor income in composite consumption $c_1^s = w_1^s \bar{n}_1^s$ and does not accumulate or decumlate their debt or savings. In turn, in period 2, an unexpected shock hits that restricts agents working in sector 2 in periods 2 and 3, i.e., $w_2^2 = w_3^2 = 0$, and the government promises a stimulus payment S in period 3. Then agents in sector 2 allocate consumption in periods 2 and all the following periods according to their Euler equation and budget constraints.

$$U'(c_2^2) = U'(c_3^2),\ c_2^2 + a_2^2 \leq 1/\beta a_1^2, \text{ and } c_3^2 = S + 1/\beta a_2^2. \tag{9}$$

In turn, we obtain:

$$c_2^2 = c_3^2 = \frac{S + 1/\beta^2 a_1^2}{1 + 1/\beta} \text{ and } a_2^2 = 1/\beta a_1^2 - \frac{S + 1/\beta^2 a_1^2}{1 + 1/\beta}. \tag{10}$$

Agents in sector 1 allocate consumption in periods 2 and 3 according to their Euler equation and budget constraints in the same manner and we obtain:

$$c_2^1 = c_3^2 = \frac{S + w_3^1 \bar{n}_3^1 + 1/\beta(w_2^1 \bar{n}_2^1 + 1/\beta a_1^1)}{1 + 1/\beta} \text{ and} \tag{11}$$

$$a_2^1 = w_2^1 \bar{n}_2^1 + 1/\beta a_1^1 - \frac{S + w_3^1 \bar{n}_3^1 + 1/\beta(w_2^1 \bar{n}_2^1 + 1/\beta a_1^1)}{1 + 1/\beta}. \tag{12}$$

Consumption for agents in sector 3 follows the above straightforwardly.

Proposition 1. *The MPC out of income (or fiscal stimulus payments) is larger for agents in sector 2 than for agents in sectors 1 or 3.*

Proof. Compare MPCs, i.e., how much out of income (or fiscal stimulus payments) are consumed:

$$\frac{\partial c_2^2}{\partial S} = \frac{\partial c_3^2}{\partial S} = \frac{\partial(\frac{S+1/\beta^2 a_1^2}{(1+1/\beta)})}{\partial S} = \frac{1}{1+1/\beta} > \frac{\partial c_2^1}{\partial(S + w_2^1 \bar{n}_2^1 + w_3^1 \bar{n}_3^1)}$$

as $\dfrac{\partial(\frac{S+w_2^1\bar{n}_2^1+w_3^1\bar{n}_3^1+(1/\beta-1)w_2^1\bar{n}_2^1+1/\beta^2 a_1^1}{1+1/\beta})}{\partial(S + w_2^1 \bar{n}_2^1 + w_3^1 \bar{n}_3^1)} = \dfrac{1}{1+1/\beta} + \underbrace{\dfrac{\partial(\frac{(1/\beta-1)}{1+1/\beta} w_2^1 \bar{n}_2^1)}{\partial(S + w_2^1 \bar{n}_2^1 + w_3^1 \bar{n}_3^1)}}_{<0}$ and $\dfrac{(1/\beta - 1)}{1+1/\beta} < 0.$

This argument extends straightforwardly to the comparison of agents in sectors 2 and 3. □

Proposition 2. *The marginal propensity to repay debt out of income (or fiscal stimulus payments) is larger for agents in sector 2 than for agents in sector 1.*

Proof. Compare the propensity to repay mortgages, i.e., how much out of income (or fiscal stimulus payments) are used to repay debt:

$$\frac{\partial(-a_2^2)}{\partial S} = \frac{\partial(-1/\beta a_1^2 + \frac{S+1/\beta^2 a_1^2}{(1+1/\beta)})}{\partial S} = \frac{1}{1+1/\beta} > \frac{\partial(-a_2^2)}{\partial(S + w_2^1 \bar{n}_2^1 + w_3^1 \bar{n}_3^1)}$$

as $\dfrac{\partial(-w_2^1\bar{n}_2^1 - 1/\beta a_1^1 + \frac{S+w_2^1\bar{n}_2^1+w_3^1\bar{n}_3^1+(1/\beta-1)w_2^1\bar{n}_2^1}{1+1/\beta})}{\partial(S + w_2^1 \bar{n}_2^1 + w_3^1 \bar{n}_3^1)} = \dfrac{1}{1+1/\beta} + \underbrace{\dfrac{\partial(-w_2^1\bar{n}_2^1 + \frac{(1/\beta-1)}{1+1/\beta} w_2^1 \bar{n}_2^1)}{\partial(S + w_2^1 \bar{n}_2^1 + w_3^1 \bar{n}_3^1)}}_{<0}.$

□

If we now compare this economy's to one in which sector 2 would not shut down, there are three differences that each diminish the amount of consumption induced by the stimulus payment S. First, agents in all sectors cannot consume in sector 2, thereby foregoing increases in employment and income in that sector. Secondly, sector 2 agents are the poorest agents with the highest MPC out of their income, so declines in their income disproportionately decrease the fiscal multiplier. Finally, agents in sector 2 choose to accumulate more debt in period 2 planning to repay it with their stimulus payment. In turn, the stimulus payment goes to agents in sector 3 that have lower MPCs out of the stimulus payment.

In summary, in this economy, workers in sectors 1 and 2 will spend their stimulus payment on mortgages and loan repayments as well as non-durable necessary consumption (sector 1). As shown above, this means that the fiscal stimulus payments flows to households with less high MPCs and directly decreases the fiscal multiplier, i.e., $1/(1 - MPC)$, making fiscal stimulus less effective.

7 Conclusion

This paper studies the impact of the 2020 CARES Act stimulus payments on household spending using detailed high-frequency transaction data from SaverLife, a non-profit helping working families develop long-term savings habits and meet financial goals. We utilize this dataset to explore heterogeneity of MPCs in response to the stimulus payments, an important parameter both in determining multipliers and in testing between representative and heterogeneous agent models. We hope that our results inform the ongoing debate about appropriate policy measures and next steps in the face of the COVID-19 pandemic.

We find large consumption responses to fiscal stimulus payments and significant heterogeneity across individuals. Income levels and liquidity play important roles in determining MPCs, with liquidity being the strongest predictor of MPC heterogeneity. We find substantial responses for households with low levels of liquidity and no response to stimulus payments for households with high levels of account balances or cash on hand. The results will potentially be important for policy-makers in terms of designing future rounds of stimulus if the 2020 crisis persists. Our results suggest that the effects of stimulus are much larger when targeted to households with low levels of liquidity.

More work should be done to study how targeting can be designed to have large impacts on consumption without generating significant behavioral effects. Just as unemployment benefits may increase unemployment durations (**?**), policies targeting stimulus payments towards households with low levels of liquidity could discourage liquid savings.

References

Agarwal, Sumit, Chunlin Liu, and Nicholas S Souleles, "The Reaction of Consumer Spending and Debt to Tax Rebates-Evidence from Consumer Credit Data," *Journal of Political Economy*, dec 2007, *115* (6), 986–1019.

Allcott, Hunt, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Y Yang, "Polarization and Public Health: Partisan Differences in Social Distancing During the Coronavirus Pandemic," *NBER Working Paper*, 2020.

Andersen, Asger Lau, Emil Toft Hansen, Niels Johannesen, and Adam Sheridan, "Consumer Reponses to the COVID-19 Crisis: Evidence from Bank Account Transaction Data," Technical Report, Working Paper 2020.

Baker, Scott R, "Debt and the Response to Household Income Shocks: Validation and Application of Linked Financial Account Data," *Journal of Political Economy*, 2018, *126* (4), 1504–1557.

Baker, Scott R. and Constantine Yannelis, "Income Changes and Consumption: Evidence from the 2013 Federal Government Shutdown," *Review of Economic Dynamics*, 2017, *23*, 99–124.

Baker, Scott R, Nicholas Bloom, Steven J Davis, and Stephen J Terry, "Covid-Induced Economic Uncertainty," *National Bureau of Economic Research Working Paper*, 2020.

_ , RA Farrokhnia, Steffen Meyer, Michaela Pagel, and Constantine Yannelis, "How Does Household Spending Respond to an Epidemic? Consumption During the 2020 COVID-19 Pandemic," *National Bureau of Economic Research Working Paper*, 2020.

Barrios, John and Yael Hochberg, "Risk Perception Through the Lens Of Politics in the Time of the COVID-19 Pandemic," *Working Paper*, 2020.

Barro, Robert J, José F Ursua, and Joanna Weng, "The Coronavirus and the Great Influenza Epidemic," *Working Paper*, 2020.

Baugh, Brian, Itzhak Ben-David, Hoonsuk Park, and Jonathan A Parker, "Asymmetric Consumption Response of Households to Positive and Negative Anticipated Cash Flows," *NBER Working Paper*, 2018.

Bayer, Christian, Benjamin Born, Ralph Luetticke, and Gernot J. Müller, "The Coronavirus Stimulus Package: How Large is the Transfer Multiplier?," *Working Paper*, 2020.

Benmelech, Efraim, Adam Guren, and Brian Melzer, "Making the House a Home: The Stimulative Effects of Home Purchases on Consumption and Investment," *Working Paper*, 2019.

Bounie, David, Youssouf Camara, and John W Galbraith, "Consumers' Mobility, Expenditure and Online-Offline Substitution Response to COVID-19: Evidence from French Transaction Data," *Available at SSRN 3588373*, 2020.

Broda, C and J Parker, "The Economic Stimulus Payments of 2008 and the Aggregate Demand for Consumption," *Journal of Monetary Economics*, 2014.

Carvalho, Vasco M., Juan R. Garcia, Stephen Hansen, Alvaro Ortiz, Tomasa Rodrigo, Jose V. Rodriguez Mora, and Jose Ruiz, "Tracking the COVID-19 Crisis with High-Resolution Transaction Data," Technical Report, Working Paper 2020.

Chen, Haiqiang, Wenlan Qian, and Qiang Wen, "The Impact of the COVID-19 Pandemic on Consumption: Learning from High Frequency Transaction Data," *Available at SSRN 3568574*, 2020.

Coibion, Olivier, Yuriy Gorodnichenko, and Michael Weber, "Labor Markets During the COVID-19 Crisis: A Preliminary View," *Fama-Miller Working Paper*, 2020.

Coven, Joshua and Arpit Gupta, "Disparities in Mobility Responses to COVID-19," Technical Report 2020.

Dunn, Abe, Kyle Hood, and Alexander Driessen, "Measuring the Effects of the COVID-19 Pandemic on Consumer Spending Using Card Transaction Data," *BEA Working Paper Series WP2020-5*, 2020.

Eichenbaum, Martin S, Sergio Rebelo, and Mathias Trabandt, "The Macroeconomics of Epidemics," *NBER Working Paper*, 2020.

Gormsen, Niels Joachim and Ralph SJ Koijen, "Coronavirus: Impact on Stock Prices and Growth Expectations," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, 2020, pp. 2020–22.

Granja, Joao, Christos Makridis, Constantine Yannelis, and Eric Zwick, "Did the Paycheck Protection Program Hit the Target?," *NBER Working Paper*, 2020.

Guerrieri, Veronica, Guido Lorenzoni, Ludwig Straub, and Iván Werning, "Macroeconomic Implications of COVID-19: Can Negative Supply Shocks Cause Demand Shortages?," Technical Report, National Bureau of Economic Research 2020.

Hagedorn, Marcus, Iourii Manovskii, and Kurt Mitman, "The Fiscal Multiplier," Technical Report 2019.

Johnson, David S., Jonathan A. Parker, and Nicholas Souleles, "Household Expenditure and the Income Tax Rebates of 2001," *American Economic Review*, 2006, *96* (5), 1589–1610.

Jones, Callum, Thomas Philippon, and Venky Venkateswaran, "Optimal Mitigation Policies in a Pandemic," *Working Paper*, 2020.

Kaplan, Greg and Gianluca Violante, "A Model of the Consumption Response to Fiscal Stimulus Payments," *Econometrica*, 2014, *82*, 1199–1239.

_ , Ben Moll, and Gianluca Violante, "Pandemics According to HANK," Technical Report, Working Paper 2020.

Kuchler, Theresa and Michaela Pagel, "Sticking to Your Plan: Hyperbolic Discounting and Credit Card Debt Paydown," *Journal of Financial Economics*, 2020.

Kueng, Lorenz, "Excess Sensitivity of High-Income Consumers," *The Quarterly Journal of Economics*, 2018, *133* (4), 1693–1751.

Meyer, Bruce D, "Unemployment Insurance and Unemployment Spells," *Econometrica (1986-1998)*, 1990, *58* (4), 757.

Misra, Kanishka and Paolo Surico, "Consumption, Income Changes, and Heterogeneity: Evidence from Two Fiscal Stimulus Programs," *American Economic Journal: Macroeconomics*, 2014, *6* (4), 84–106.

Olafsson, Arna and Michaela Pagel, "The Liquid Hand-to-Mouth: Evidence from Personal Finance Management Software," *Review of Financial Studies*, 2018, *31* (11), 4398–4446.

Parker, Jonathan A, Nicholas S Souleles, David S Johnson, and Robert McClelland, "Consumer Spending and the Economic Stimulus Payments of 2008," *American Economic Review*, 2013, *103* (6), 2530–53.

Figure 1: Example of Platform

Notes: The figures show screenshots of the SaverLife website. The upper part of the screenshot shows the app's landing page and the lower part illustrates the offered financial advice pages. Source: SaverLife.

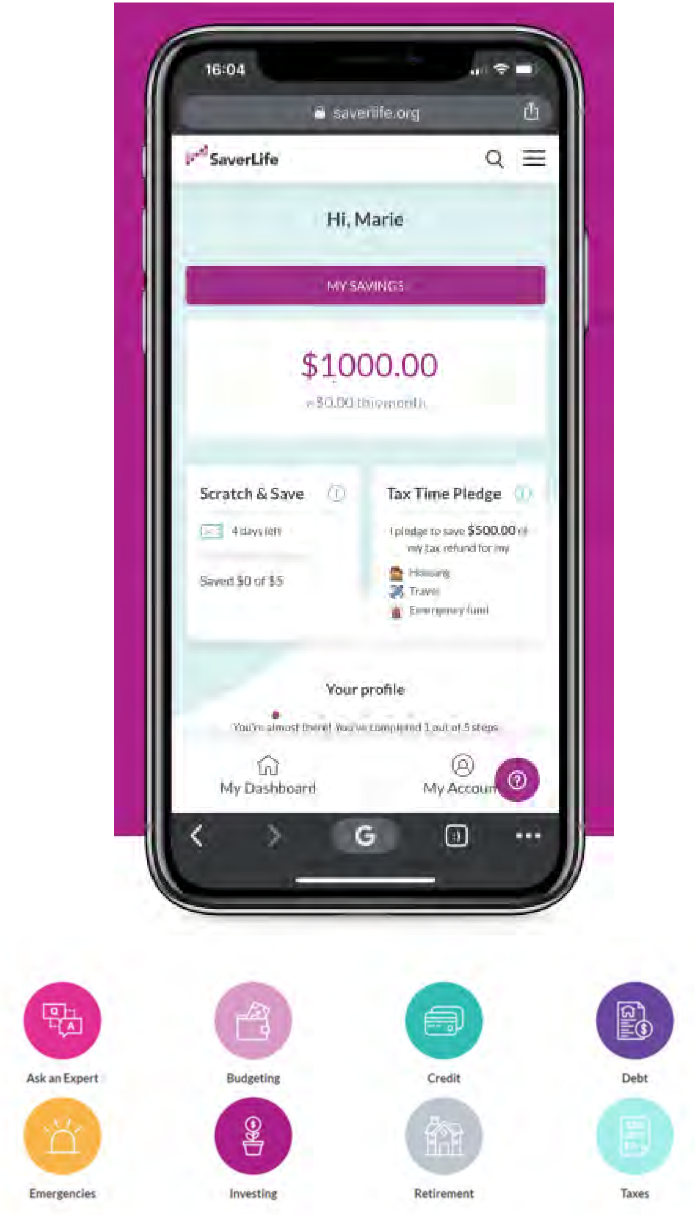


Figure 2: Daily Number of Government Payments at Stimulus Amounts

Notes: The top panel shows the number of payments users receive that match the amounts of the 2020 government stimulus payment by day in 2019 and 2020. Potential payments are classified by the specified amounts of the stimulus checks and need to appear as being tax refunds, credit or direct deposits. The bottom panel restricts the time period to February through April in 2020. Source: SaverLife.

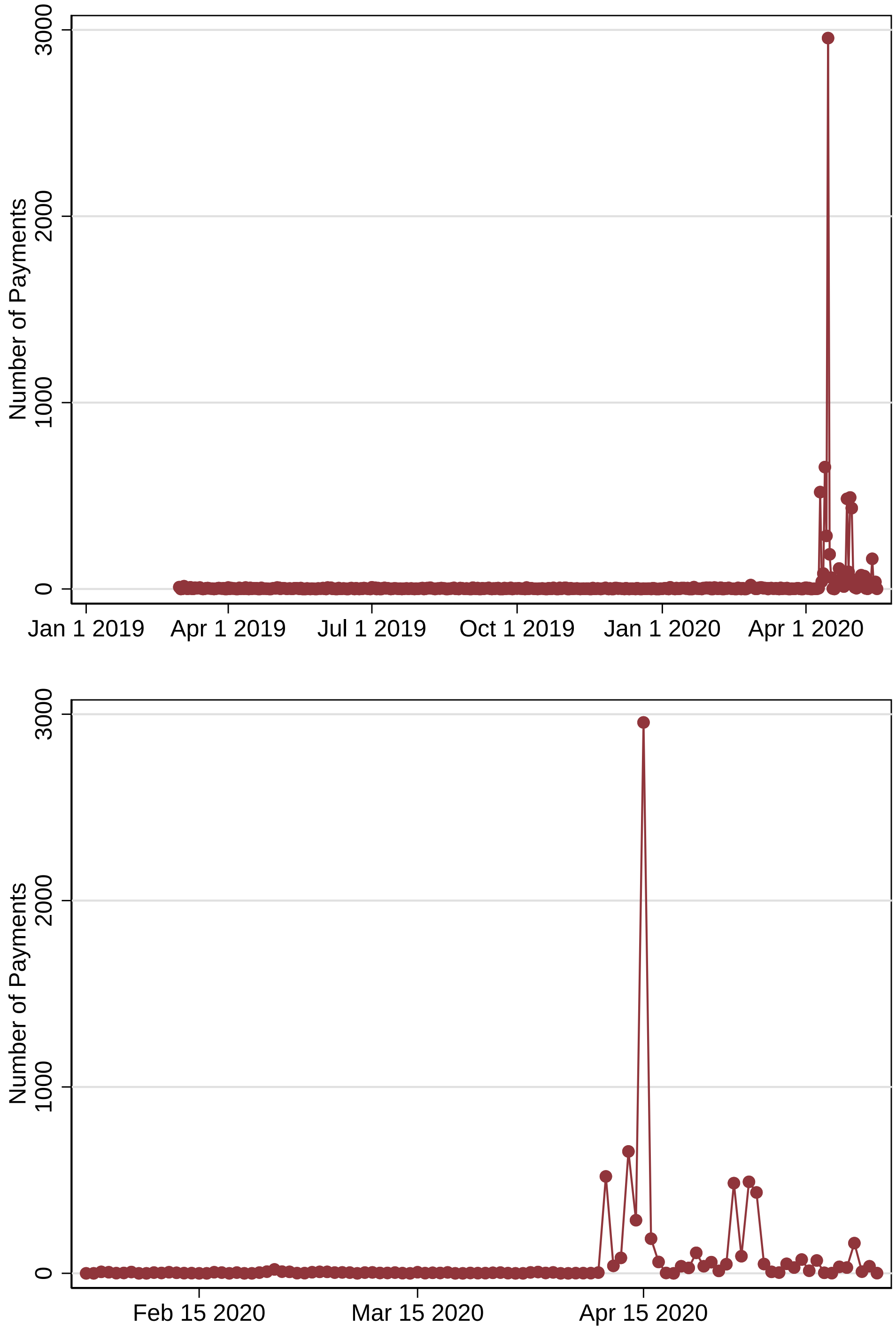


Figure 3: Mean Spending Around Receiving the Stimulus Payments - Raw Spending

Notes: This figure shows mean spending around the receipt of stimulus payments. The sample includes only users who receive a stimulus payment during our sample period. The vertical axis measures spending in dollars, and the horizontal axis shows time in days from receiving the stimulus check which is defined as zero (0). Shaded days represent weekends for the majority of stimulus-recipients who receive their payment on Wednesday April 15th. Source: SaverLife.

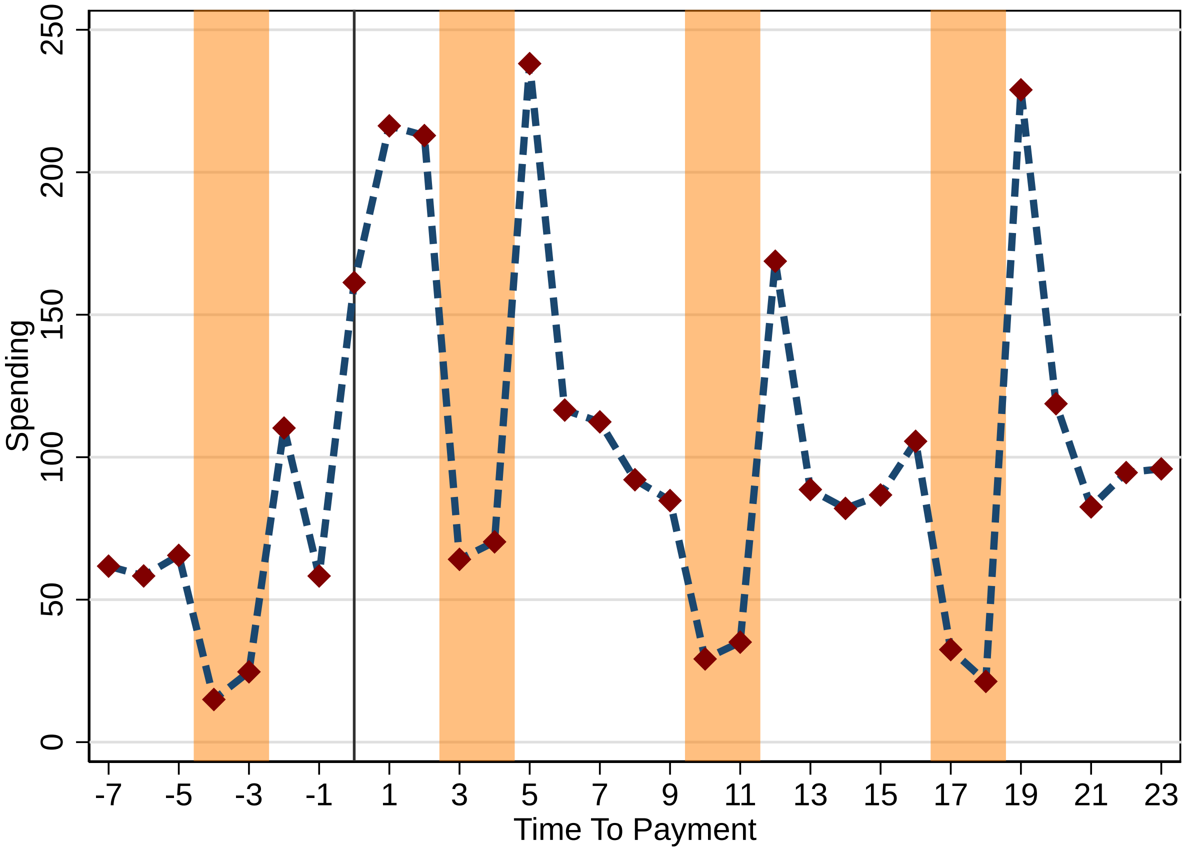


Figure 4: Spending Around Stimulus Payments - Regression Estimates

Notes: This figure shows estimates of β_i from $c_{it} = \alpha_i + \alpha_t + \sum_{t=-7}^{23} \beta_i \mathbb{1}[t=i]_{it} + \varepsilon_{it}$. The sample includes all users in our sample period (both those who do and do not receive stimulus payments). The solid line shows point estimates of β_i, while the dashed lines show 95% confidence interval. Time to payment is equal to zero on the day of receiving the stimulus check. Source: SaverLife.

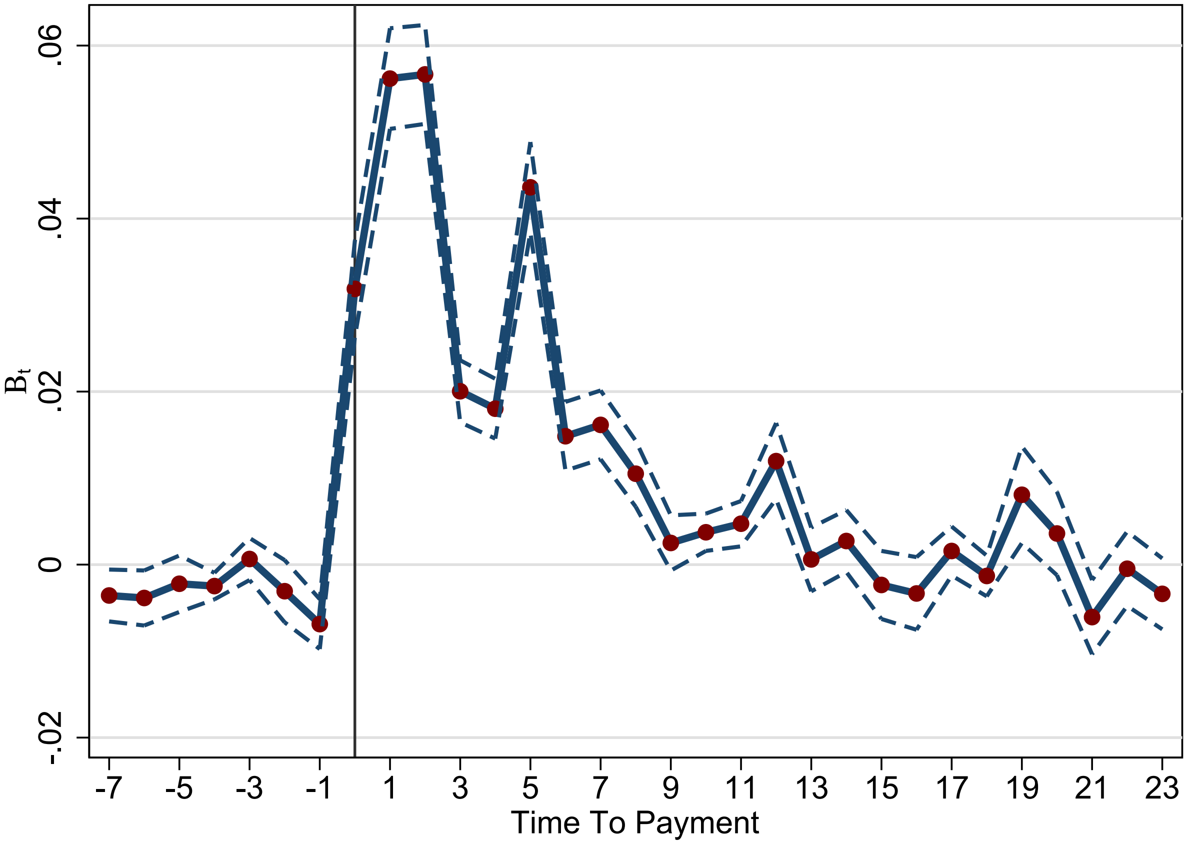
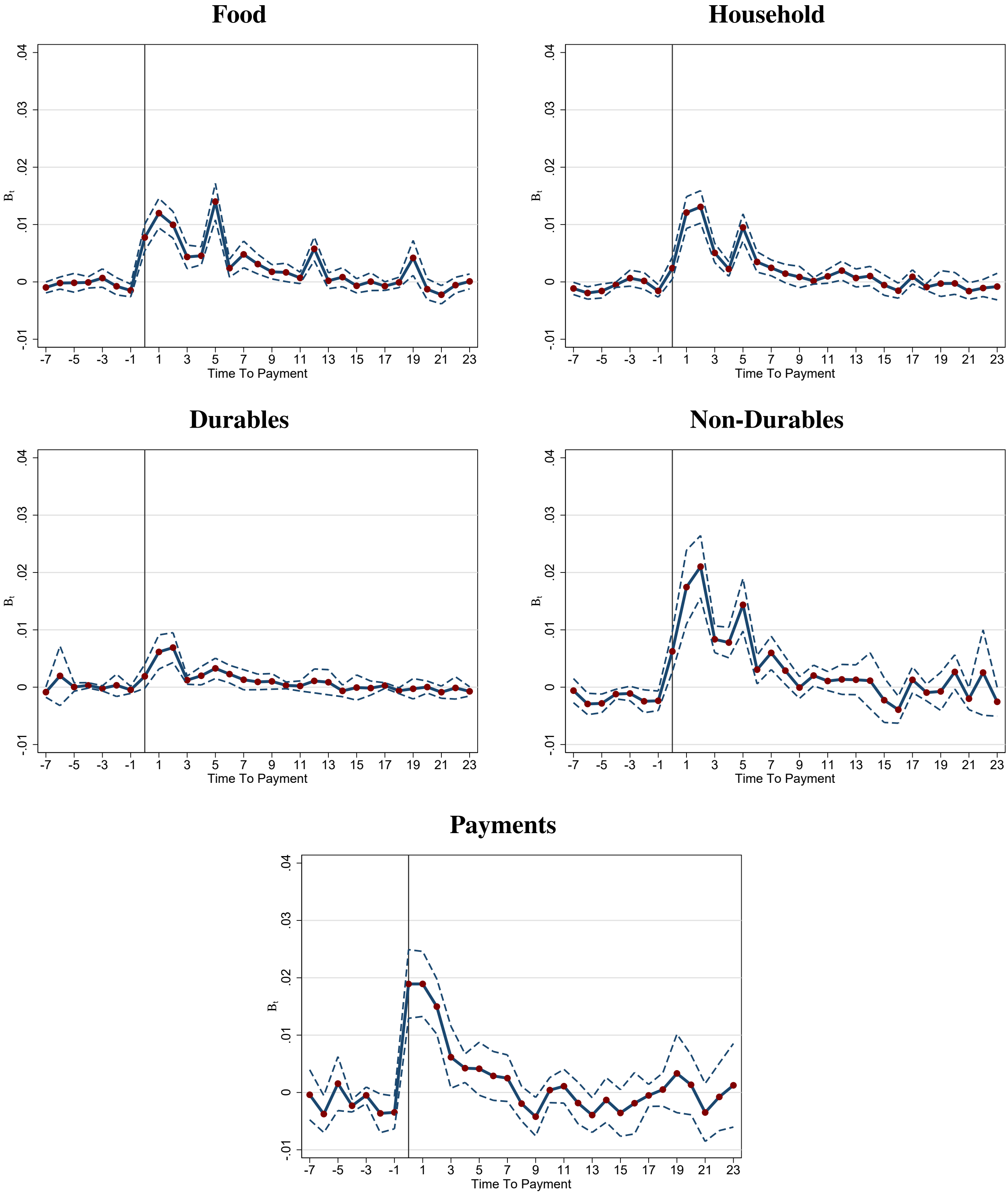


Figure 5: Spending Around Stimulus Payments by Categories

Notes: This figure shows estimates of β_i from $c_{it} = \alpha_i + \alpha_t + \sum_{t=-7}^{23} \beta_i \mathbb{1}[t = i]_{it} + \varepsilon_{it}$, broken down by spending categories. The solid line shows point estimates of β_i, while the dashed lines show the 95% confidence interval. Time to payment is equal to zero on the day of receiving the stimulus check. Source: SaverLife.



Food

Household

Durables

Non-Durables

Payments

Figure 6: MPC by Income Groups

Notes: This figure shows cumulative MPCs estimated from coefficients from regressions of spending on an indicator of a time period being after a stimulus payment, scaled by the amount of the payment over the number of days since the payment. That is, of ζ from $c_{it} = \alpha_i + \alpha_t + \zeta \frac{Post_{it} \times P_i}{Days_{it}} + \varepsilon_{it}$, broken down by monthly income groups. Year and week by individual fixed effects are included. Standard errors are clustered at the user level. The bar shows point estimates, while the thin lines show the 95% confidence interval. Source: SaverLife.

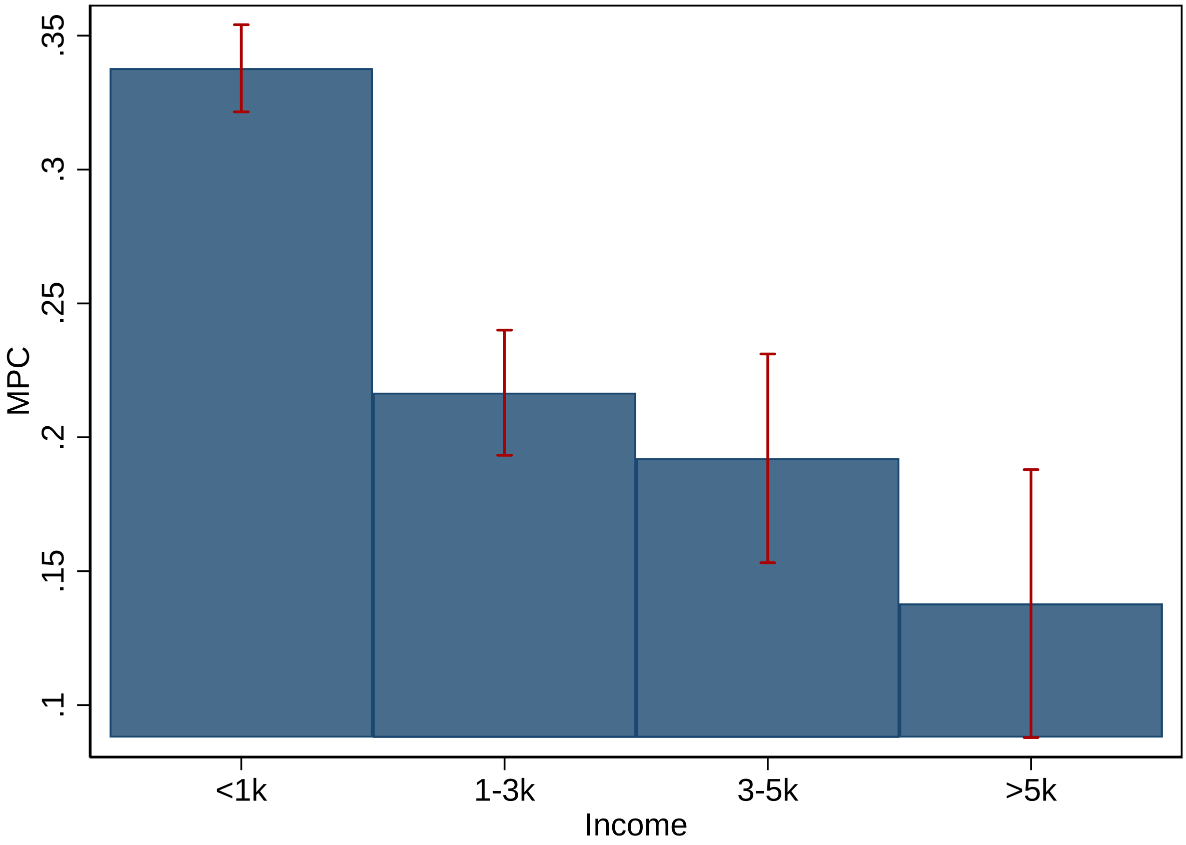


Figure 7: MPC by Liquidity

Notes: This figure shows cumulative MPCs estimated from coefficients from regressions of spending on an indicator of a time period being after a stimulus payment, scaled by the amount of the payment over the number of days since the payment. That is, of ζ from $c_{it} = \alpha_i + \alpha_t + \zeta \frac{Post_{it} \times P_i}{Days_{it}} + \varepsilon_{it}$, broken down by account balances. Year and week by individual fixed effects are included. Standard errors are clustered at the user level. The bar shows point estimates, while the thin lines show 95% confidence interval. Source: SaverLife.

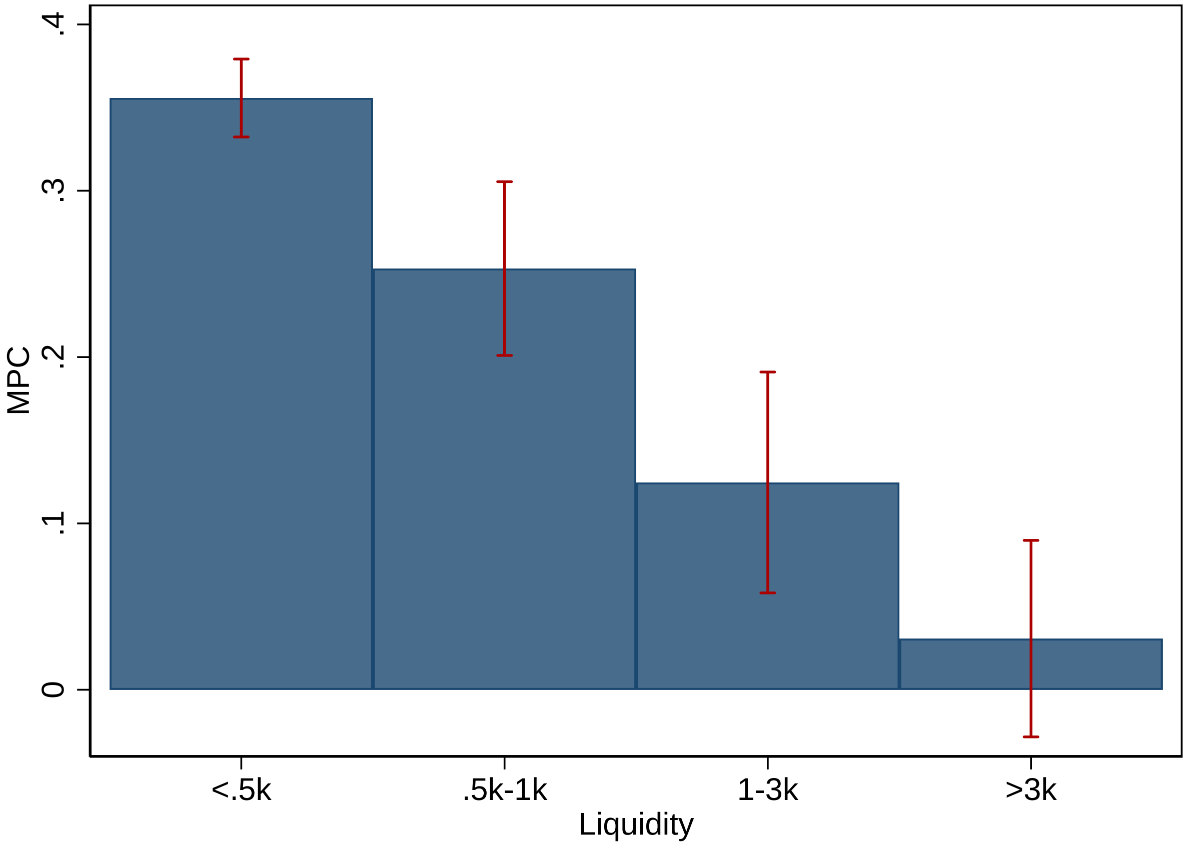


Figure 8: MPC by Drop in Income

Notes: This figure shows cumulative MPCs estimated from coefficients from regressions of spending on an indicator of a time period being after a stimulus payment, scaled by the amount of the payment over the number of days since the payment. That is, of ζ from $c_{it} = \alpha_i + \alpha_t + \zeta \frac{Post_{it} \times P_i}{Days_{it}} + \varepsilon_{it}$, broken down by the drop in income between January/February 2020 and March 2020. Year and week by individual fixed effects are included. Standard errors are clustered at the user level. The bar shows point estimates, while the thin lines show 95% confidence interval. Source: SaverLife.

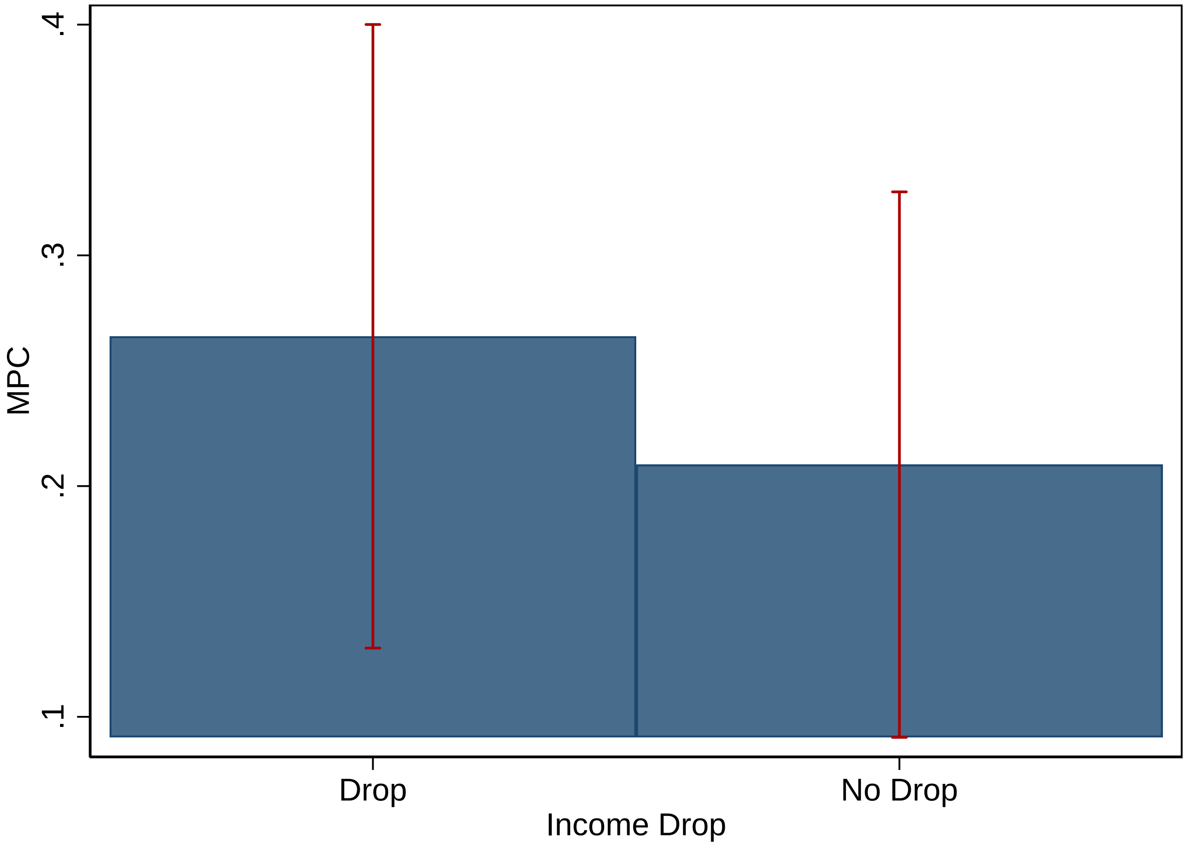
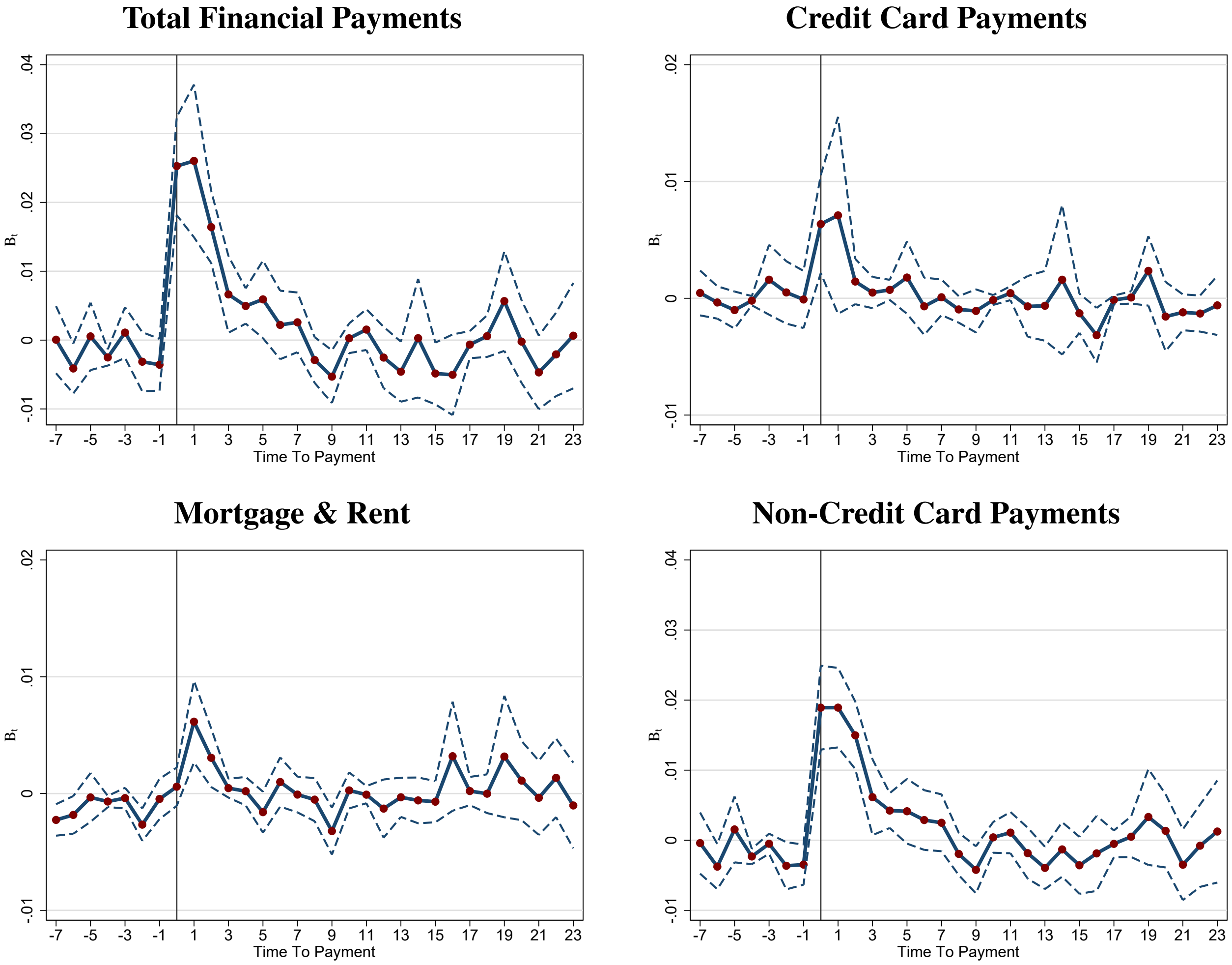


Figure 9: Payment Spending Around Stimulus

Notes: This figure shows estimates of β_i from $c_{it} = \alpha_i + \alpha_t + \sum_{t=-7}^{23} \beta_i \mathbb{1}[t = i]_{it} + \varepsilon_{it}$, broken down by payment categories. The solid line shows point estimates of β_i, while the dashed lines show the 95% confidence interval. Time to payment is equal to zero on the day of receiving the stimulus check. Source: SaverLife.



Total Financial Payments

Credit Card Payments

Mortgage & Rent

Non-Credit Card Payments

Table 1: Summary Statistics

Notes: Summary statistics for spending and income represent user-month observations. Statistics regarding user characteristics are given at an individual level. Stimulus Income (Cond) refers to the distribution of stimulus income conditional on receiving a stimulus payment.

Variable	# Obs.	Mean	10th	25th	Median	75th	90th
User-Month							
Income	22,826	1,913.87	100	600.96	1,562.76	2,982.05	4,662.17
Balance	22,826	650.42	3.91	38.48	148.30	825.73	2,451.86
Durables	22,826	48.41	0	0	0	30	159.90
Food	22,826	242.07	0	21.185	151.315	371.25	651.02
Household	22,826	225.11	0	30	151.22	350.82	598.39
Non-Durables	22,826	320.37	0	50.32	213.53	478.39	850.16
Payments	22,826	402.93	0	0	132.61	659.48	1,278.29
Transfers	22,826	515.74	0	31.98	256.39	801.44	1,521.58
User							
Stimulus Income	6,033	840.85	0	0	0	1,200	2,400
Stimulus Income (Cond)	2,665	1,903.52	1,200	1,200	1,700	2,200	3,200

Table 2: Stimulus Payments and Spending

The table shows regressions of overall spending and categories of spending on the one-day lag of the stimulus payment. We run separate regressions for overall spending, food, non-durables, household items, durables and payments. For total spending, we run three specifications with varying fixed effects. We use individual by day of the month fixed-effects, individual and calendar date and individual times day of month fixed-effects, or individual and day of the month and individual times day of week fixed-effects. Standard errors are clustered at the user level. *$p < .1$, ** $p < .05$, *** $p < .01$. Source: SaverLife.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Total	Total	Total	Food	Durables	Household	Durables	Payments
Stimulus Payment	0.0310***	0.0345***	0.0396***	0.00731***	0.00527*	0.00259**	0.00216**	0.0195***
	(0.00552)	(0.00391)	(0.00469)	(0.00126)	(0.00284)	(0.00105)	(0.000966)	(0.00528)
Stimulus Payment$_{t+1}$	0.0599***	0.0590***	0.0589***	0.0125***	0.0188***	0.0128***	0.00713***	0.0219***
	(0.00604)	(0.00563)	(0.00768)	(0.00225)	(0.00298)	(0.00263)	(0.00143)	(0.00408)
Stimulus Payment$_{t+2}$	0.0603***	0.0578***	0.0626***	0.00993***	0.0218**	0.0141*	0.00705***	0.0168***
	(0.0163)	(0.0161)	(0.0137)	(0.00309)	(0.00822)	(0.00779)	(0.00174)	(0.00456)
Stimulus Payment$_{t+3}$	0.00835	0.0205	0.00484	0.00185	0.00520	0.00262	0.000123	0.00203
	(0.0166)	(0.0140)	(0.0179)	(0.00646)	(0.00714)	(0.00612)	(0.000743)	(0.00756)
Stimulus Payment$_{t+4}$	0.00808	0.0185	0.00875	0.00226	0.00577	0.000525	0.00180	0.000347
	(0.0172)	(0.0131)	(0.0144)	(0.00658)	(0.00836)	(0.00365)	(0.00285)	(0.00573)
Stimulus Payment$_{t+5}$	0.0609***	0.0440***	0.0644***	0.0192**	0.0203***	0.0125**	0.00463**	0.0112***
	(0.0114)	(0.00807)	(0.0101)	(0.00904)	(0.00432)	(0.00486)	(0.00186)	(0.00309)
Stimulus Payment$_{t+6}$	0.0216***	0.0162***	0.0196***	0.00390**	0.00591**	0.00526***	0.00241***	0.00502***
	(0.00440)	(0.00488)	(0.00496)	(0.00158)	(0.00244)	(0.00124)	(0.000571)	(0.00139)
Stimulus Payment$_{t+7}$	0.0140***	0.0172***	0.0213***	0.00372**	0.00444***	0.00281***	0.00150**	0.00243
	(0.00346)	(0.00398)	(0.00388)	(0.00144)	(0.00143)	(0.000926)	(0.000586)	(0.00194)
Stimulus Payment$_{t+8}$	0.0130***	0.0118***	0.0138**	0.00370***	0.00396***	0.00255***	0.00119***	0.00103
	(0.00326)	(0.00359)	(0.00540)	(0.000920)	(0.00130)	(0.000861)	(0.000402)	(0.00140)
Stimulus Payment$_{t+9}$	0.00603**	0.00327	0.00155	0.00160	0.00131	0.00171**	0.000963**	-0.00112
	(0.00282)	(0.00250)	(0.00508)	(0.000970)	(0.00140)	(0.000798)	(0.000392)	(0.00157)
Date FE	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
User FE	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
User*Day of Month FE		**X**						
User*Day of Week FE			**X**					
Observations	560,711	560,711	560,711	560,711	560,711	560,711	560,711	560,711
R^2	0.178	0.291	0.409	0.080	0.062	0.077	0.024	0.051

Table 3: Stimulus Payments, Spending and Income

This figure shows cumulative MPCs estimated from coefficients from regressions of spending on an indicator of a time period being after a stimulus payment, scaled by the amount of the payment over the number of days since the payment. That is, of ζ and ξ from $c_{it} = \alpha_i + \alpha_t + \zeta \frac{Post_{it} \times P_i}{Days_{it}} + \xi \frac{Post_{it} \times P_i}{Days_{it}} \times I_i + \phi Post_{it} \times I_i + \varepsilon_{it}$. Average monthly income is approximately $2,000, yielding a logged income value of 7.6. Columns (4) and (5) drop the interaction, and split the sample by January and February monthly income above and below $2,000. The inclusion of fixed effects is denoted beneath each specification. Standard errors are clustered at the user level. $^{*}p < .1$, $^{**}\ p < .05$, $^{***}\ p < .01$. Source: SaverLife.

	(1)	(2)	(3)	(4)	(5)
	Total	Total	Total	Low Inc	High Inc
Post-Stimulus*Stimulus	0.703***	0.732***	0.684***	0.337***	0.180**
	(0.129)	(0.0993)	(0.132)	(0.0534)	(0.0803)
Post-Stimulus*Stimulus*ln(Inc)	-0.0629***	-0.0656***	-0.0593**		
	(0.0213)	(0.0152)	(0.0228)		
Date FE	**X**	**X**	**X**	**X**	**X**
Individual FE	**X**	**X**	**X**	**X**	**X**
Individual X Day of Month FE		**X**			
Individual X Day of Week FE			**X**	**X**	**X**
Observations	560,711	560,711	560,711	350,177	210,534
R^2	0.172	0.287	0.404	0.130	0.196

Table 4: Stimulus Payments, Spending and Liquidity

This figure shows cumulative MPCs estimated from coefficients from regressions of spending on an indicator of a time period being after a stimulus payment, scaled by the amount of the payment over the number of days since the payment. That is, of ζ and ξ from $c_{it} = \alpha_i + \alpha_t + \zeta \frac{Post_{it} \times P_i}{Days_{it}} + \xi \frac{Post_{it} \times P_i}{Days_{it}} \times L_i + \phi Post_{it} \times L_i + \varepsilon_{it}$. The second row of columns (1) through (3) interacts with the individual's bank account balance prior to the arrival of the stimulus payment, in thousands of dollars. Columns (4) and (5) drop the interaction, and split the sample by having more or less than $500 in a bank account. Columns (6) and (7) does the same split as in columns (4) and (5), restricting to individuals who make more than $4,000 a month. The inclusion of fixed effects is denoted beneath each specification. Standard errors are clustered at the user level. *$p < .1$, ** $p < .05$, *** $p < .01$. Source: SaverLife.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Total	Total	Total	Low Bal	High Bal	High Inc/Low Bal	High Inc/High Bal
Post-Stimulus*Stimulus	0.293***	0.300***	0.297***	0.327***	0.130***	0.353***	0.0944***
	(0.0391)	(0.0332)	(0.0424)	(0.0474)	(0.0381)	(0.0749)	(0.0305)
Post-Stimulus*Stimulus*Balance	-0.0665***	-0.0690***	-0.0562**				
	(0.0236)	(0.0199)	(0.0240)				
Date FE	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Individual FE	**X**	**X**	**X**	**X**	**X**	**X**	**X**
Individual X Day of Month FE		**X**					
Individual X Day of Week FE			**X**	**X**	**X**	**X**	**X**
Observations	560,711	560,711	560,711	374,975	185,736	32,833	55,063
R^2	0.172	0.287	0.404	0.150	0.212	0.214	0.243

Table 5: Stimulus Payments, Spending and Income Declines

This figure shows cumulative MPCs estimated from coefficients from regressions of spending on an indicator of a time period being after a stimulus payment, scaled by the amount of the payment over the number of days since the payment. That is, of ζ and ξ from $c_{it} = \alpha_i + \alpha_t + \zeta \frac{Post_{it} \times P_i}{Days_{it}} + \xi \frac{Post_{it} \times P_i}{Days_{it}} \times D_i + \phi Post_{it} \times D_i + \varepsilon_{it}$. The second row of columns (1) through (3) interacts with the fraction of January and February income that an individual earned in March (ie. a lower value means a larger decline in income). Columns (4) and (5) drop the interaction, and split the sample by whether a household had an income drop in March relative to January and February. The inclusion of fixed effects is denoted beneath each specification. Standard errors are clustered at the user level. *$p < .1$, ** $p < .05$, *** $p < .01$. Source: SaverLife.

	(1)	(2)	(3)	(4)	(5)
	Total	Total	Total	Income Decline	No Decline
Post-Stimulus*Stimulus	0.231***	0.238***	0.233***	0.265***	0.209***
	(0.0620)	(0.0468)	(0.0583)	(0.0689)	(0.0603)
Post-Stimulus*Stimulus*Inc Drop	-0.0341**	-0.0374**	-0.0265		
	(0.0164)	(0.0164)	(0.0165)		
Date FE	**X**	**X**	**X**	**X**	**X**
Individual FE	**X**	**X**	**X**	**X**	**X**
Individual X Day of Month FE		**X**			
Individual X Day of Week FE			**X**	**X**	**X**
Observations	560,711	560,711	560,711	301,137	259,574
R^2	0.172	0.287	0.404	0.179	0.169

Table 6: Payments

This figure shows cumulative MPCs estimated from coefficients from regressions of spending on an indicator of a time period being after a stimulus payment, scaled by the amount of the payment over the number of days since the payment. That is, of ζ from $c_{it} = \alpha_i + \alpha_t + \zeta \frac{Post_{it} \times P_i}{Days_{it}} + \varepsilon_{it}$. Each column shows a different payment category. The inclusion of fixed effects is denoted beneath each specification. Standard errors are clustered at the user level. $*p < .1$, $**\ p < .05$, $***\ p < .01$. Source: SaverLife.

	(1)	(2)	(3)	(4)	(5)
	Total Spending	Total Financial Payments	Non-CC Payments	CC Payment	Rent and Mortgage
Post-Stimulus*Stimulus	0.243***	0.0812***	0.0670***	0.0143	0.0122**
	(0.0423)	(0.0261)	(0.0173)	(0.0123)	(0.00585)
Date FE	**X**	**X**	**X**	**X**	**X**
Individual FE	**X**	**X**	**X**	**X**	**X**
Observations	560,711	560,711	560,711	560,711	560,711
R^2	0.287	0.141	0.123	0.127	0.113

A Details on the CARES Act

The COVID-19 pandemic and the following policy responses had a large impact on the US economy. To combat the adverse consequences, Congress passed the Coronavirus Aid, Relief and Economic Security Act (CARES Act) which was passed on March 25, 2020 and signed into law on March 27, 2020. The CARES Act is the third act in a sequence of responses to the outbreak of the coronavirus by Congress. The first act was focused on spurring coronavirus vaccine research and development (Coronavirus Preparedness and Response Supplemental Appropriations Act, March 6, 2020) with a volume of $8.2 billion. The second act was a package of approximately $104 billion in paid sick leave and unemployment benefits for workers and families (the Families First Coronavirus Response Act, March 18, 2020).

The CARES Act was a $2.2 trillion economic stimulus package and is by far the largest part in this sequence of responses to the pandemic up to that point. The act splits up into $500 billion support for companies in distress, $350 billion in loans for small businesses, and over $300 billion in stimulus payments for most American workers. The rebate provides a direct payment, which is treated as a refundable tax credit against 2020 personal income taxes. Thus, the rebates would not be counted as taxable income for recipients, as the rebate is a credit against tax liability and is refundable for taxpayers with no tax liability to offset. Figure **??** shows an example of a letter sent out announcing stimulus payments.

All individuals were eligible for the stimulus if they had a valid social security number and if they were not depending on someone else. Individuals must have filed tax returns in 2018 or 2019. Individuals who did not need to file tax returns because their income was below $12,200 ($24,400 for married couples) were eligible but needed to register through a website at the Internal Revenue Service. Recipients of social security benefits did not need to register but were also eligible.

Single individuals received up to $1,200, while those who filed jointly received $2,400. Those with children under 17 received an add-on of $500 per child. The tax rebate phased out for higher levels of income. The payment was declined by 5 percent of the amount of adjusted gross income. The phase-out started at $75,000 for singles or at $150,000 for married couples. For households heads with dependents (e.g. one person with a child) the phase-out began at an income of $112,500. For details see Figure **??**.

Due to the phase-out provisions, singles (couples) above $99,000 ($198,000) did not qualify for a rebate. In Figure **??**, we plot the average size of the identified stimulus by users who report living in a household of a given size. In general, we see a clear upward trend in stimulus check size received as households get larger, again reinforcing the likelihood that we are truly picking up stimulus check receipt by users.

The House Ways & Means Committee, using information from the IRS, estimates that 171 million people were eligible for receiving rebate payments under the CARES Act. The 171 million people split up into 145-150 million taxpayers who file returns and are were eligible for the stimulus, 20-30 million Social Security beneficiaries and SSI recipients who do not file returns, 15 million non-filers below the filing threshold, 6 million veterans, and 500-600,000 from the Railroad Retirement Board.

In comparison to previous stimulus payments in 2001 or 2008, the IRS did not communicate an exact schedule for sending out the stimulus payments. An approximate schedule for the payments can still made based on the information available (see table **??** and figure **??**). Taxpayers received the first payments, using direct deposit information from the tax filings from 2018 or 2019, during the week of April 13. The House Ways & Means Committee estimates that during this first week, over 80 million Americans received payments in their bank accounts. During the following weeks the IRS continued weekly rounds of direct deposits to those who provided direct deposit information through the website of the IRS. All taxpayers who had not registered their bank account information by May 13 received their stimulus payment as paper checks. The issuing and mailing of paper checks started in the week of April 20. The checks were sent out at a rate of 5 million checks per week.

During the end of April and beginning of May, Social Security retirement, survivor and disability insurance (SSDI) beneficiaries who did not file tax returns in 2018 or 2019 received their payments via direct transfer (nearly 100% of Social Security beneficiaries). Adult Supplemental Security Income (SSI) recipients received their payments by early May, in the same way, they received their normal benefits (see AARP).

Banks like e.g. the Bank of America and Wells Fargo allowed customers to deposit their checks using mobile solutions to make the stimulus available during the physical lockdown period and to reduce delays. Wells Fargo also allowed non-customers to cash checks with no fees charged. As of May 8, 2020 CNN reported that more than 130 million eligible households had already received their stimulus payment. This lines up closely with the fraction receiving payments in our sample.

In addition to the economic stimulus package, the CARES Act made two additional provisions that are relevant. People who filed for unemployment or were partly unemployed due to the coronavirus received an additional $600 per week on top of their state benefits, until July 31. Whether a person is entitled to the extra money depends on whether an individual qualifies for state or other federal unemployment benefits. The extra $600 also applies to self-employed, part-time workers and gig-workers. Individuals receive their extra unemployment benefits with their state or federal benefits.

The CARES Act suspends minimum distributions from Individual Retirement Accounts (IRAs), 401(k)s, 403(b)s, 457(b)s, and inherited retirement accounts for 2020. It also waives the 10% tax penalty for early distributions of up to $100,000 retroactively by January 1, 2020 if an individual, their spouse, or dependent others is hit by negative consequences of the COVID-19 pandemic.

Figure **??** presents a placebo exercise. We show spending for individuals in April who did not observe receiving a stimulus check. There is no sharp uptick in spending beyond day of the week effects, consistent with there not being significant measurement error in our sample.

Table A.1: Household Composition and Stimulus Payments Under the CARES Act

Notes: This table shows statutory payment amounts for household stimulus payments under the CARES Act (for households not subject to an income-based means test).

Household Composition	Expected Stimulus Payment
Single	\$1,200
Single with one child	\$1,700
Single with two children	\$2,200
Single with three children	\$2,700
Single with four children	\$3,200
Couple	\$2,400
Couple with one child	\$2,900
Couple with two children	\$3,400
Couple with three children	\$3,900
Couple with four children	\$4,400

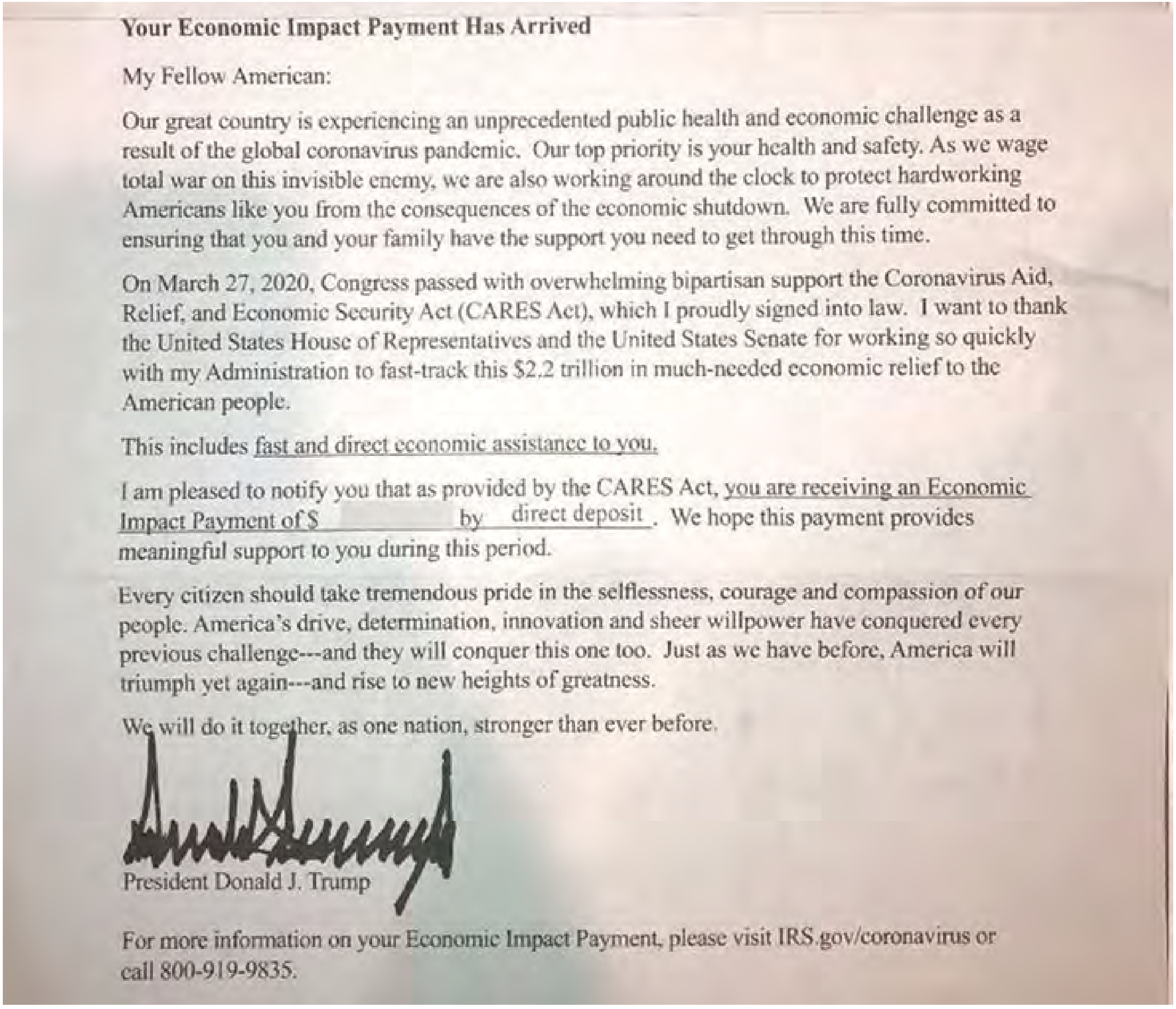
Table A.2: The Timing of the CARES Act Stimulus Payments of 2020

Notes: The table is based on information from The House Ways & Means Committee. The table displays payments disbursed by end of week dates (Fridays). Payments received counts the number of individuals.

Payments by electronic funds transfer		*Payments by check*		*Payments received*
Taxpayer group	Date funds transferred by	Taxpayer group (if no bank account information available)	Date checks received by	Direct deposit and check (cumul.)
Direct deposit information on file	Apr 17			80 mil.
Registered direct deposit information with IRS until Apr 17	Apr 24	< 10k gross income	Apr 24	
Registered direct deposit information with IRS until Apr 24	May 1	10k - 20k gross income	May 1	
Registered direct deposit information with IRS until May 1	May 8	20k - 30k gross income	May 8	130 mil.
Registered direct deposit information with IRS until May 13	May 15	30k - 40k gross income	May 15	
Website for registering direct deposit information closed on May 13		40k - 50k gross income	May 22	152 mil.
		50k - 60k gross income	May 29	
		60k - 70k gross income	Jun 05	
		Further increments of 10k (= 5 mil. checks)	Weekly until August 28	171 mil. (expected)

Figure A.1: Example of Notification Letter for Direct Deposit Transfer

Notes: This figure shows an example of a notification letter for stimulus payments.



Your Economic Impact Payment Has Arrived

My Fellow American:

Our great country is experiencing an unprecedented public health and economic challenge as a result of the global coronavirus pandemic. Our top priority is your health and safety. As we wage total war on this invisible enemy, we are also working around the clock to protect hardworking Americans like you from the consequences of the economic shutdown. We are fully committed to ensuring that you and your family have the support you need to get through this time.

On March 27, 2020, Congress passed with overwhelming bipartisan support the Coronavirus Aid, Relief, and Economic Security Act (CARES Act), which I proudly signed into law. I want to thank the United States House of Representatives and the United States Senate for working so quickly with my Administration to fast-track this $2.2 trillion in much-needed economic relief to the American people.

This includes fast and direct economic assistance to you.

I am pleased to notify you that as provided by the CARES Act, you are receiving an Economic Impact Payment of $ __________ by direct deposit. We hope this payment provides meaningful support to you during this period.

Every citizen should take tremendous pride in the selflessness, courage and compassion of our people. America's drive, determination, innovation and sheer willpower have conquered every previous challenge---and they will conquer this one too. Just as we have before, America will triumph yet again---and rise to new heights of greatness.

We will do it together, as one nation, stronger than ever before.

President Donald J. Trump

For more information on your Economic Impact Payment, please visit IRS.gov/coronavirus or call 800-919-9835.

Figure A.2: CARES Act Economic Relief

Notes: This figure shows the expected stimulus payment for different household compositions and income levels. Source: Coronavirus Aid, Relief and Economic Security Act.

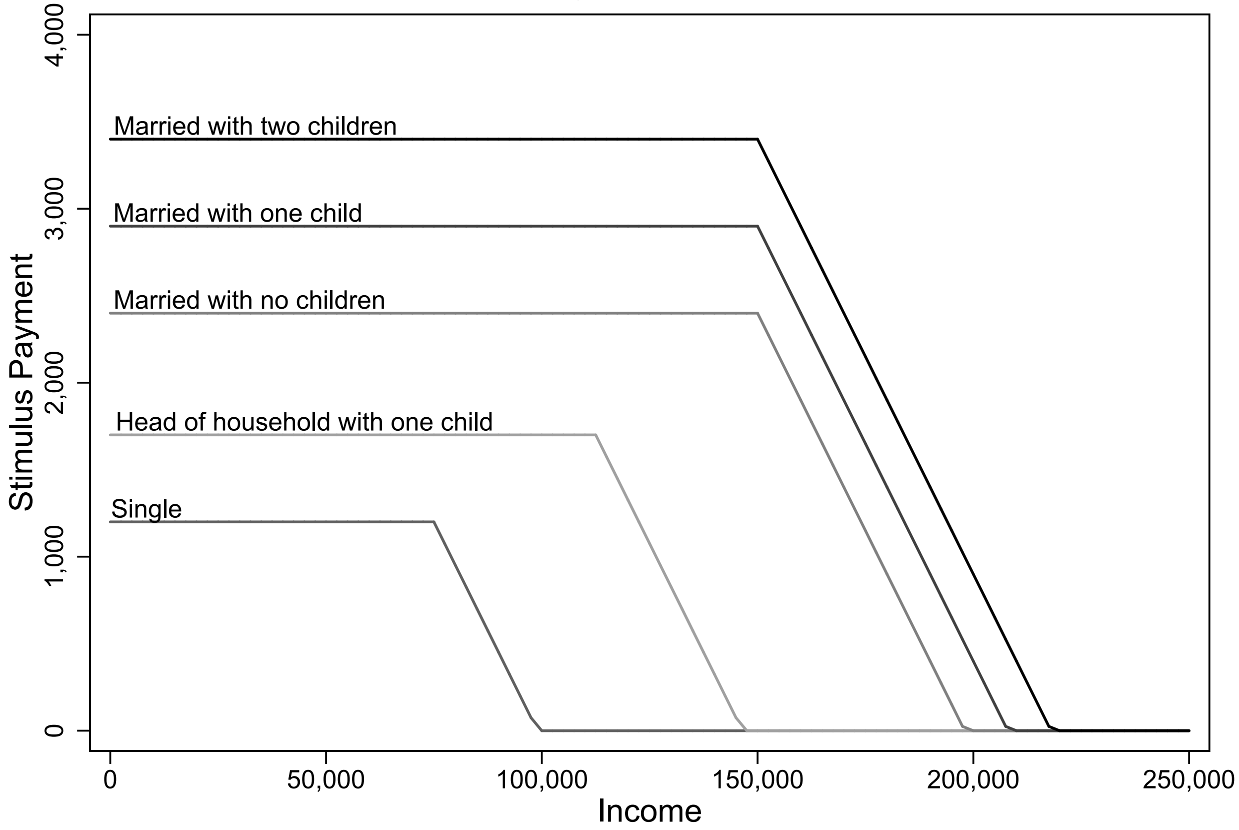


Figure A.3: Stimulus Amount Received by Household Size

Notes: This figure shows the average stimulus amount for users receiving stimulus checks, by self-reported household size. Source: SaverLife.

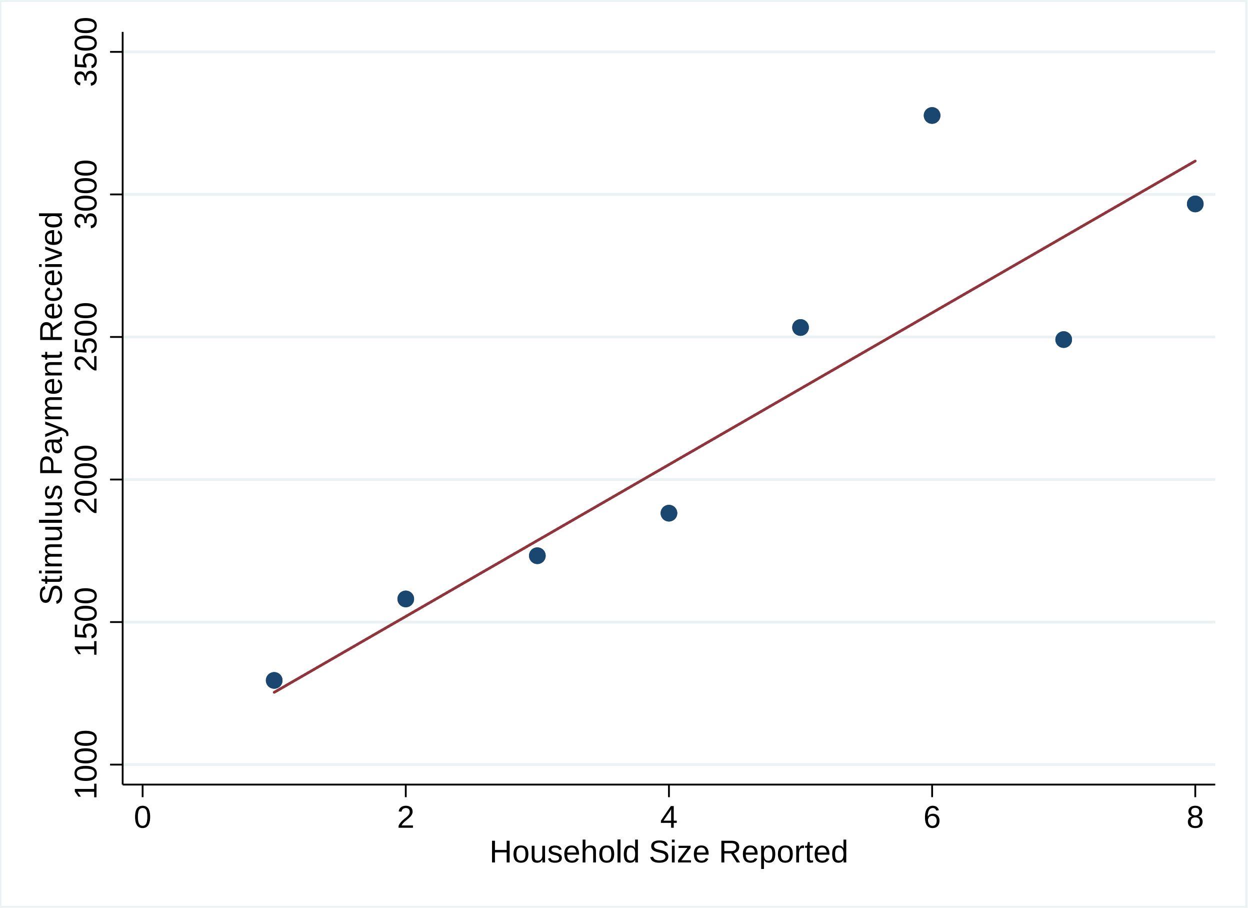
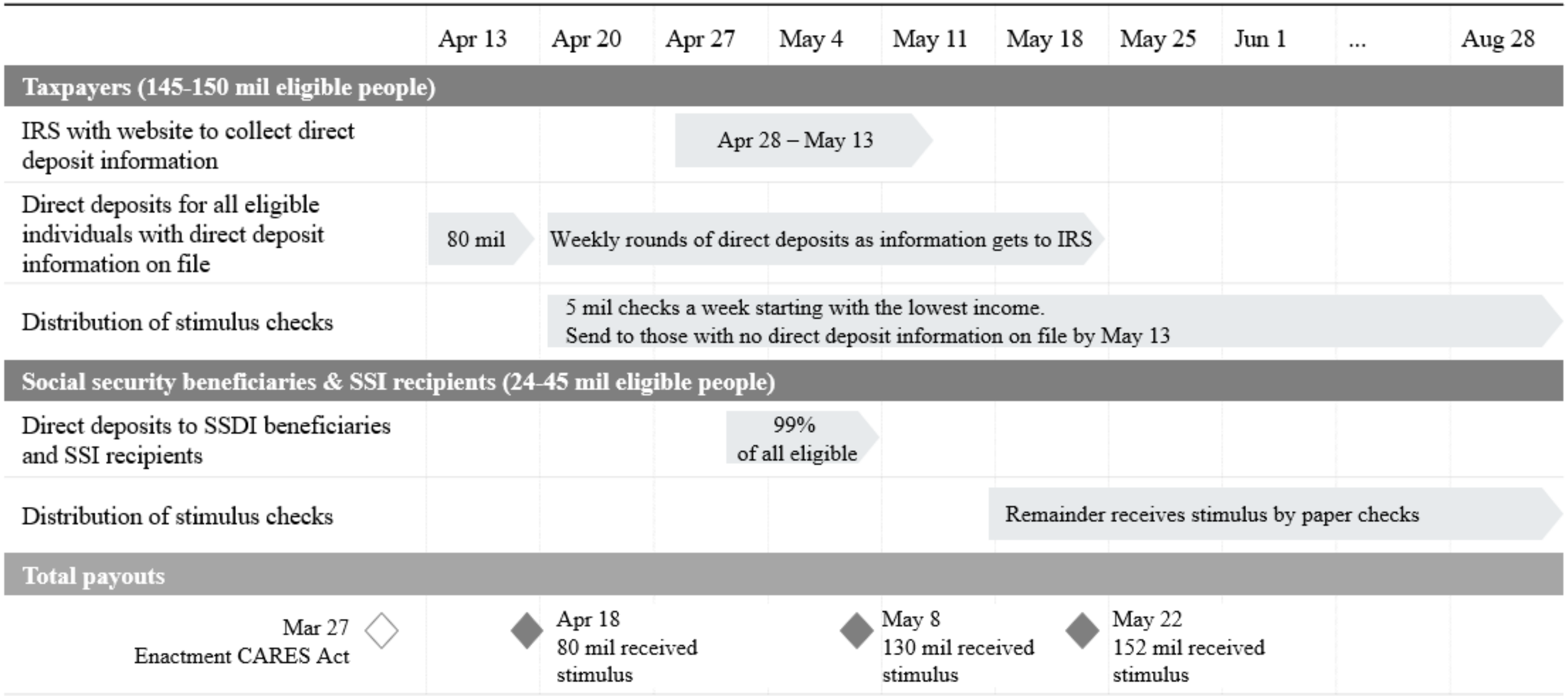


Figure A.4: Timeline of stimulus payouts

Notes: The figure presents a timeline of stimulus payments to different households. Source: House Ways & Means Committee.



Apr 13 | Apr 20 | Apr 27 | May 4 | May 11 | May 18 | May 25 | Jun 1 | ... | Aug 28

Taxpayers (145-150 mil eligible people)

IRS with website to collect direct deposit information: Apr 28 – May 13

Direct deposits for all eligible individuals with direct deposit information on file: 80 mil; Weekly rounds of direct deposits as information gets to IRS

Distribution of stimulus checks: 5 mil checks a week starting with the lowest income. Send to those with no direct deposit information on file by May 13

Social security beneficiaries & SSI recipients (24-45 mil eligible people)

Direct deposits to SSDI beneficiaries and SSI recipients: 99% of all eligible

Distribution of stimulus checks: Remainder receives stimulus by paper checks

Total payouts

Mar 27 Enactment CARES Act

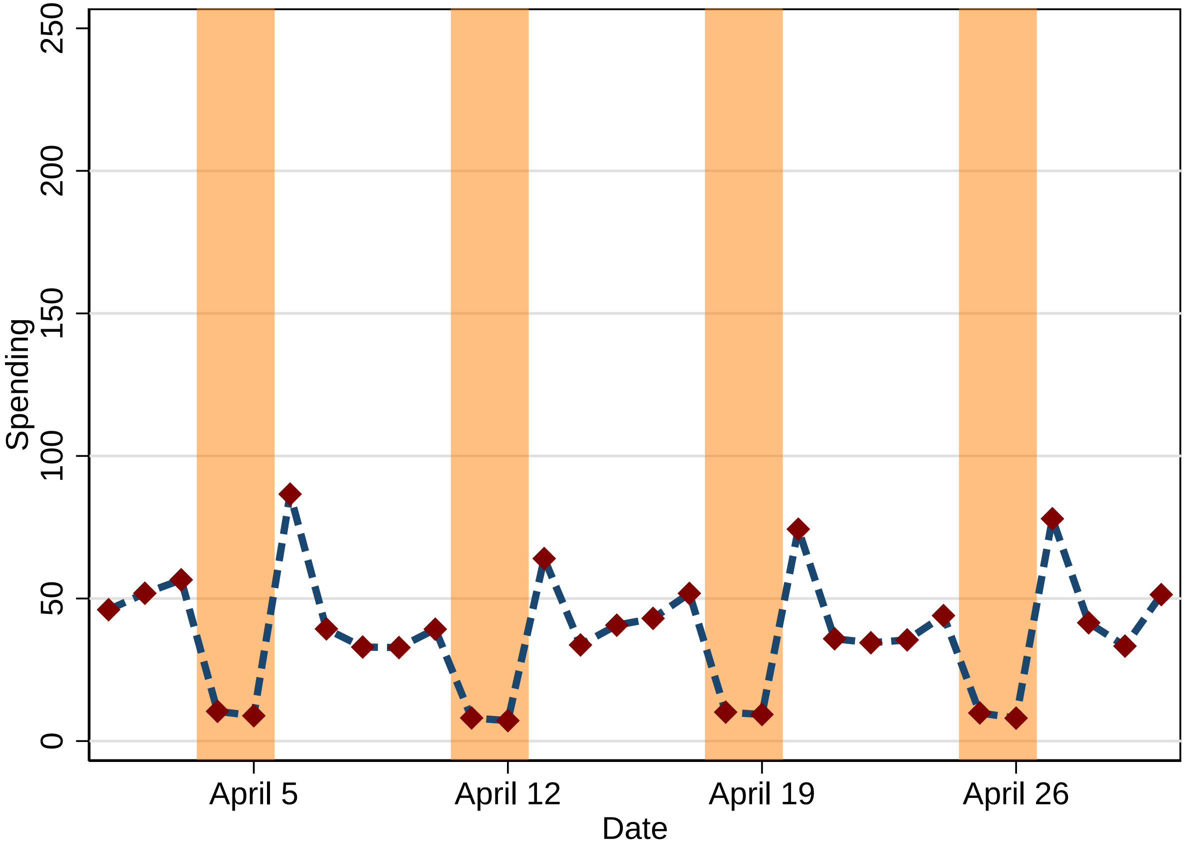
Apr 18 80 mil received stimulus

May 8 130 mil received stimulus

May 22 152 mil received stimulus

Figure A.5: Mean Spending in April for Individuals Not Receiving Payment- Raw Spending

Notes: This figure shows mean daily spending in April for individuals who did not receive payments in that month. Sample includes only users who do not receive a stimulus payment during our sample period. The vertical axis measures spending in dollars, and the horizontal axis shows the date. Shaded days represent weekends for the majority of stimulus-recipients who receive their payment on Wednesday April 15th. The graph is based on data from SaverLife.



The Initial Household Spending Response to COVID-19: Evidence from Credit Card Transactions

Diana Farrell: JPMorgan Chase Institute, President & CEO; Fiona Greig: JPMorgan Chase Institute, Director of Consumer Research; Natalie Cox: Assistant Professor of Economics, Princeton University; Peter Ganong: Assistant Professor at the University of Chicago Harris School of Public Policy; Pascal Noel: Neubauer Family Assistant Professor of Finance at the University of Chicago Booth School of Business

Introduction

COVID-19 has rapidly transformed our nation. Following the declaration of a national emergency on March 13, 2020, the U.S. caseload exceeded 100,000 on March 29, and by April 6, 90 percent of the U.S. population was subject to "stay-at-home" orders. Within a matter of weeks, vacations and special events were cancelled, and routine trips to the store, workplace, and restaurants became hindered by both the virus and the policies designed to prevent its spread.

These almost universal disruptions to normal activity have already had unprecedented consequences for the economy. The pandemic has shut down large sectors of the economy deemed "non-essential," leaving millions of workers jobless. Social distancing restrictions have all but prohibited the consumption of certain goods and services. The government has responded with a massive recovery act to bolster income by funding stimulus checks, Unemployment Insurance (UI) supplements, and the Payroll Protection Program.

In this report, we provide preliminary high-frequency evidence of the reaction of consumer spending to these events. We ask two main questions. First, how much has individual spending fallen, and how does this drop vary across households?[1] Second, can heterogeneity across households provide suggestive evidence about the spending decline caused by the nearly ubiquitous pandemic and policies intended to contain it versus the initial round of income losses during that period? With consumer spending accounting for roughly 70 percent of GDP, understanding the magnitude and causes of changes in consumption is critical to identifying policy interventions that could aid in accelerating an economic recovery. This will be increasingly important as the pandemic and policy impacts interact with increasing job loss and additional policies to ameliorate the job loss, such as stimulus payments and UI.

To answer these questions, we use a dataset based on the universe of transactions made on Chase credit cards through April 11, 2020. We focus on a sample of 8 million families across all fifty states who have been active users of their credit cards since January 2018.[2] For a subset of our analyses we pair these credit card data with checking account data through February 2020, which allow us to segment our population by income levels and industry of employment before the COVID shock.[3]

The key strengths of these data are a large sample size and the ability to track the spending patterns of specific households across time. We are thus able to provide detailed estimates of the spending drop and to analyze heterogeneity in this drop across household characteristics and across categories of spending. We decompose the spending drop into non-essential and essential spending, speaking to the pandemic-induced closure of many non-essential businesses. We also look at changes in spending across the pre-COVID income distribution. Finally, we stratify the sample by individuals' industry of employment to test whether those employed in sectors with higher expected rates of job loss cut spending by a larger amount.

Despite these strengths, our findings come with several important caveats that stem from the fact that, at the time of writing, we only observe the subset of spending that occurs on a household's Chase credit cards through April 11. We do not observe spending using debit cards, cash, electronic payments, and non-Chase credit cards. Our estimates could be biased to the extent that there is substitution between these alternative channels and Chase credit cards coincident with our analysis period. We may be particularly concerned about this type of substitution at a time of acute economic disruption when households might turn to credit cards to smooth their consumption. In addition, the pandemic might accelerate the growth in card transactions to as people avoid the risks associated with exchanging physical cash and because of the growth in online spend. This might cause us to *understate* the drop in spending.

Second, while our data include households across a wide cross-section of income levels and geographies, Chase credit card holders tend to be more affluent than the average U.S. household. As we show below, if higher-income families cut their spending to a greater extent, the sample frame could cause us to *overstate* the drop in spending. Thus, the net effect of these biases on our spend estimates is ambiguous.

Third, at the time of this release our preliminary data only cover the initial phase of the pandemic. Spending changes, and how these vary with household characteristics, may evolve over time particularly as income disruptions become more widespread.

In the future, we will be able to partially address these three limitations by looking at a longer time-period of data and examining checking account transactions to provide an integrated view of income and spending.

We have four main findings. First, we find that average weekly household credit card spending fell by 40 percent year-over-year by the end of March 2020, coinciding with a dramatic increase in COVID-19 cases, social distancing policies, and job losses. The magnitude of the spending drop is enormous; it is eight times larger than the spending drop typically observed among UI recipients in the first month after job loss. Second, spending cuts on non-essential goods and services account for nearly all of the total spending decline. Spending on essentials initially spiked 20 percent before falling back, while spending on non-essentials declined by 50 percent. Third, spending dropped substantially for households across the entire income distribution, with slightly larger drops for higher-income households driven by cuts in non-essential goods and services. Fourth, spending dropped dramatically for workers in all industries of employment. Similar drops occurred in industries with high and low rates of job loss as of April 2020.

In summary, we provide evidence suggesting that, as of the second week of April, the 40 percent drop in consumer spending appears to be driven to a greater extent by the pandemic and social distancing policies implemented across the country to prevent its spread and to a lesser extent by the initial round of income losses. However, as the pandemic unfolds, the balance of factors contributing to spending behavior could change dramatically. We will continue to track and disentangle these dynamics over time using administrative banking data.

Finding 1: Average household credit card spending had fallen by 40 percent year-over-year by the end of March 2020.

Figure 1 plots the year-over-year percentage change in weekly credit card spending in 2020 and in 2019, and Figure 2 shows levels of average weekly credit card spending in 2020 and in 2019.

Changes in spending follow a distinctive pattern — spend is stable through the beginning of March, then declines precipitously by 40 percent relative to 2019 from the second through fourth week of March. It then appears to stabilize at this lower level in the first two weeks of April. The size of the spending drop is largely consistent with other estimates from similar administrative data sources during the same time frame.

Figure 1: Average weekly household credit card spending had fallen by 40 percent year-over-year by the end of March 2020.

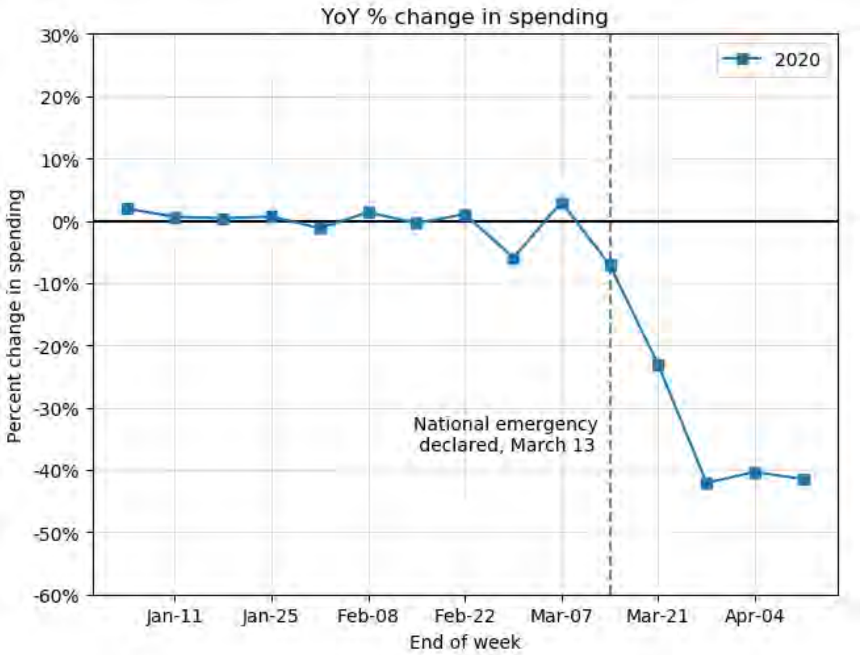
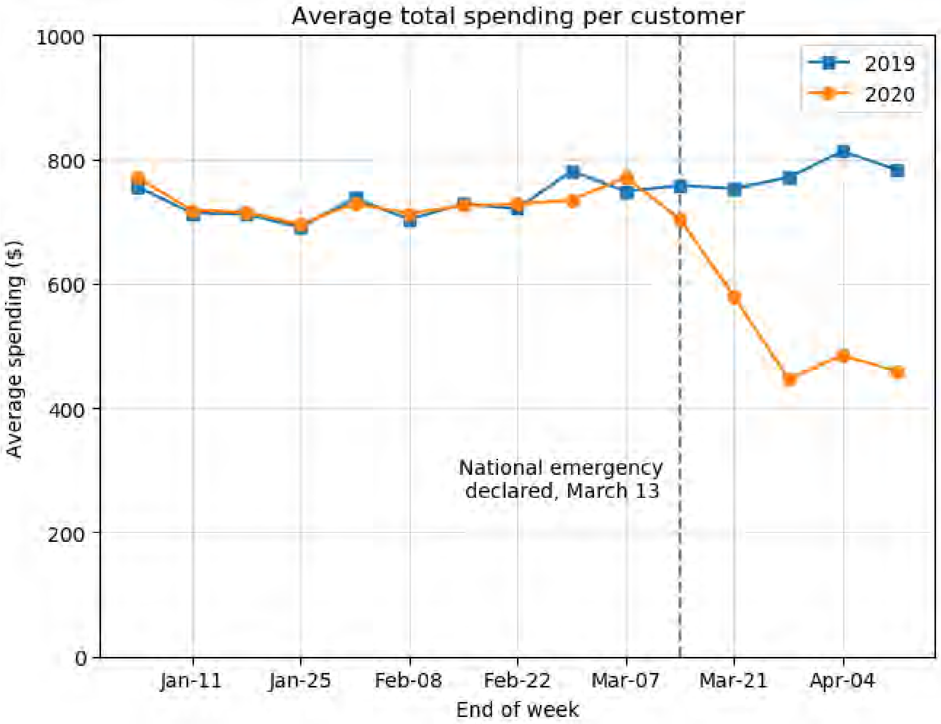


Figure 2: Average weekly credit card spending per household was more than $300 lower in April 2020 compared to April 2019.



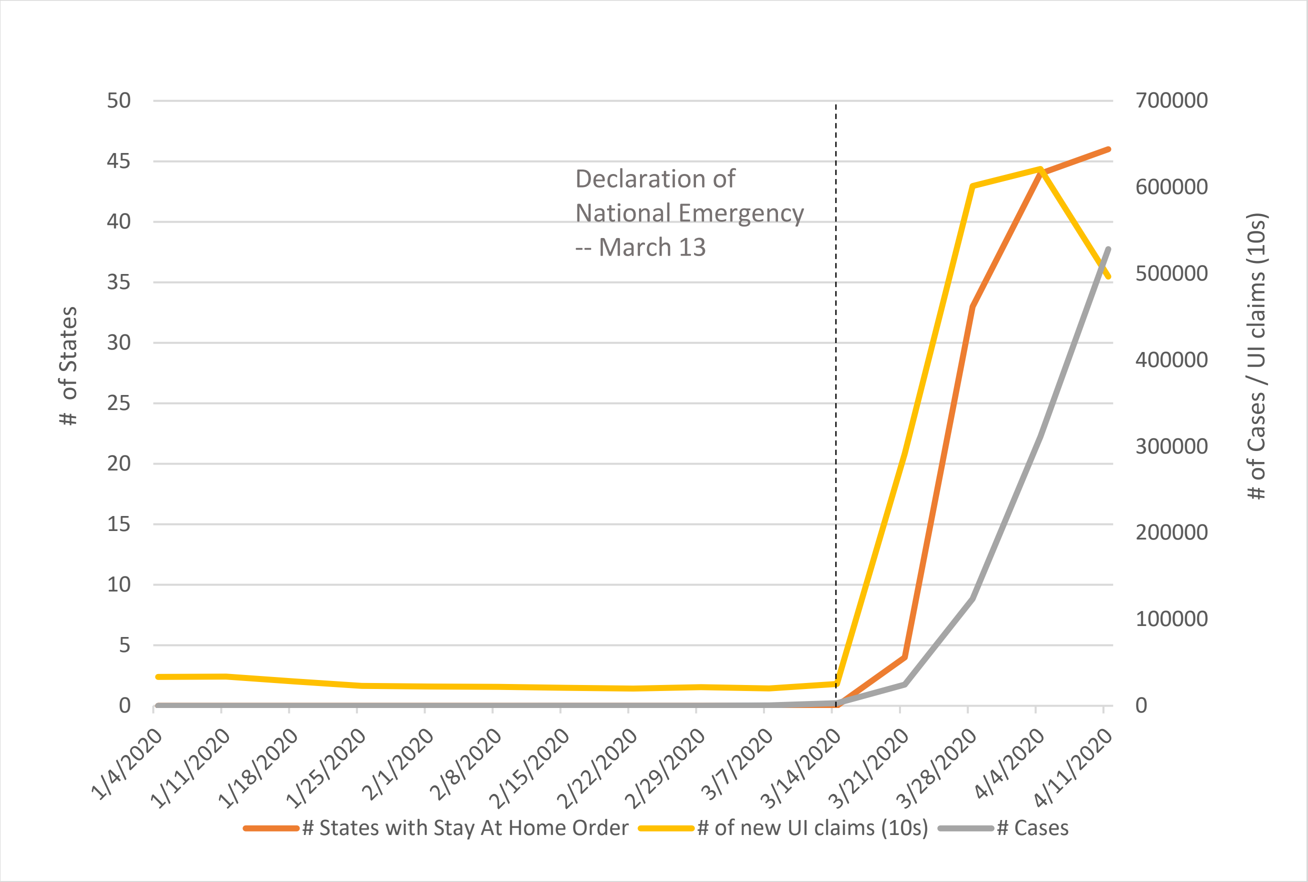
The timing of the spending drop mirrors the spread of the virus and staggered national implementation of government social distancing orders. A national emergency was declared on March 13, 2020. Over the following three weeks, the number of states with stay-at-home orders increased from zero to forty-five, and then also remained stable (see Figure 3). The prevalence of COVID-19 also increased dramatically with over 300,000 cases and 5,000 COVID-related deaths in the U.S. by the month's end.

At the same time, the drop in spending also closely tracks the pattern of initial job losses. UI claims began spiking in the third week of March, with more than 20 million UI claims filed by April 11. This raises the question as to how much of the 40 percent drop in credit card spending is due to the pandemic itself, the social distancing policies, or income losses.

Importantly, while we know from UI claims that jobs have been lost, it is unlikely that the income supports extended by the government in response to COVID-19 would have been received by the end of our time series—the second week of April. The median time between job loss and the first UI benefit receipt is roughly five weeks, which would mean that many of the first 3 million people to file for UI during the third week of March—the initial surge in UI claims—may not have started receiving their UI benefits until late April. In addition, families did not start receiving stimulus checks until the third week of April. Thus, to the extent that income losses are playing a role, they would likely not yet have been offset by policy interventions to mitigate those losses.

Nonetheless, it is still useful to calibrate the size of the spending drop relative to what we have observed among those who lose a job involuntarily during normal times. We have previously used these data to measure the spending drop around job loss among UI recipients, and observed an initial credit card spending drop of roughly 5 percent (Ganong and Noel 2019). In other words, the spending drop in March 2020 is roughly *eight times larger* than the average household credit card spending drop in the first month of unemployment for UI recipients in normal times. This puts into perspective how dramatic the spending drop is and suggests that the pandemic and policies aimed at preventing its spread are contributing substantially to the drop in spending. We explore this possibility further in Findings 3 and 4.

Figure 3: UI claims, social distancing policies, and COVID-19 cases all increased dramatically during late March and early April.

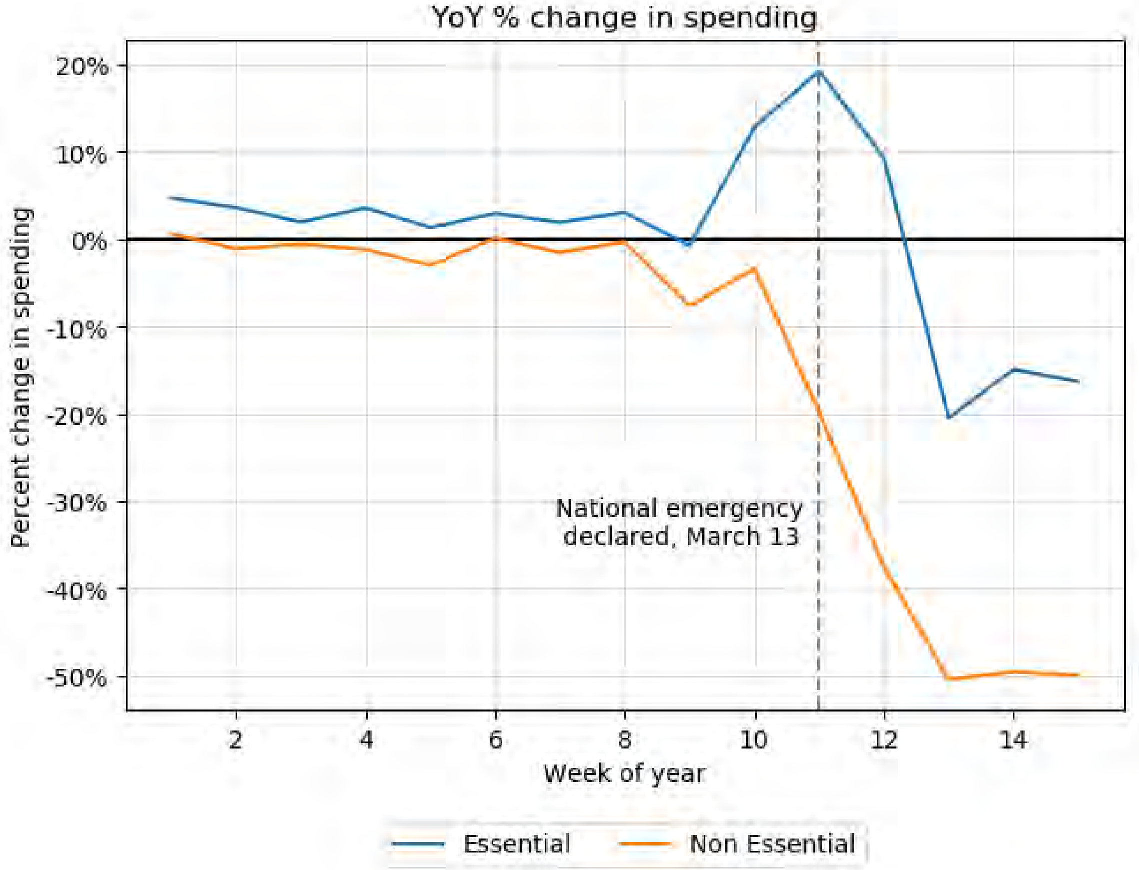


Finding 2: Spending on essentials initially spiked 20 percent before falling to below pre-pandemic levels, while spending on non-essentials declined by 50 percent and accounted for nearly all of the total spending decline.

While Finding 1 shows a sharp drop in aggregate spending, there is reason to think that specific spending categories would be differentially impacted. Many non-essential businesses, like bars and salons, were closed by state and local governments. Similarly, stay-at-home orders limited the ability of individuals to travel. Beyond the mechanical effect of social distancing regulations, individuals may also have independently curtailed spend in certain categories to avoid risk of infection or as a response to income loss.

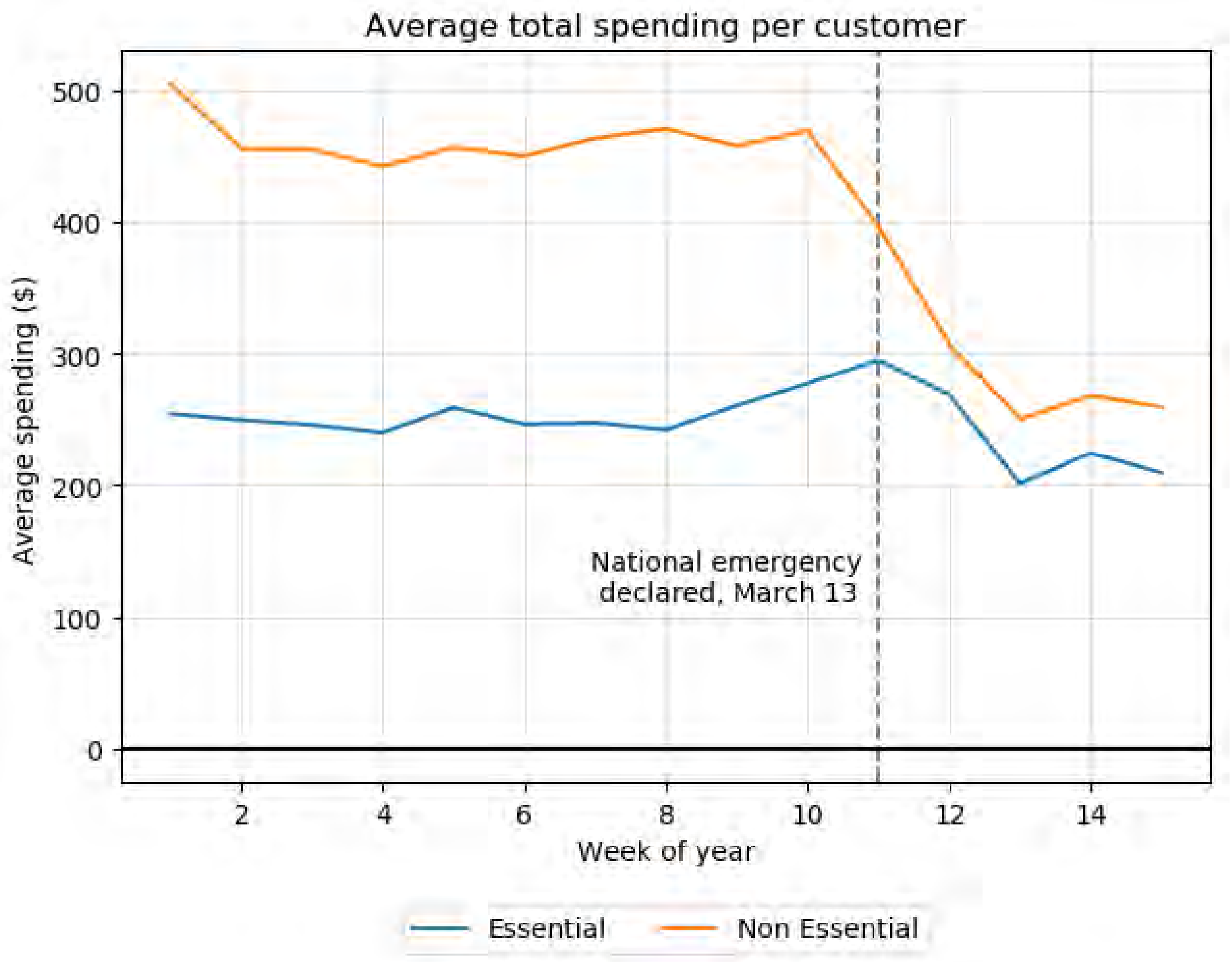
We begin by disaggregating total spending into essential and non-essential categories, as commonly defined in state “stay-at-home” orders. Figures 4 and 5 show a dramatic difference in the path of essential and non-essential spending. Essential spending spiked in early March, up almost 20 percent by the second week. It then fell back down, stabilizing at a decline of around 20 percent by early April. In contrast, spending on non-essential categories fell sharply throughout March before stabilizing down 51 percent in early April.[4]

Figure 4: Spending in non-essential categories dropped by roughly 50 percent year-over-year compared to 20 percent for essential categories.



Note: We use state social distancing orders that restricted non-essential goods and services to categorize spend. “Essential” categories include fuel, transit, cash, drug stores, discount stores, auto repair, groceries, telecom, utilities, insurance, and healthcare. “Non-essential” includes department stores, other retail, restaurants, entertainment, retail durables, home improvement, professional and personal services, and miscellaneous. Although flights, hotels, and rental cars are sometimes categorized as “essential” and not technically closed, we include them in the “non-essential” group because they are affected by stay-at-home restrictions on non-essential travel.

Figure 5: Average weekly household spending on non-essential categories dropped by roughly $200



Given the fact that households were ordered to stay at home except to make essential trips in most states, one might ask why households were still spending roughly $250 a week on non-essential categories in April. First, there is variation in the degree of closures across geographies and in what is deemed non-essential in each place. Second, our spending categories do not map perfectly to each specific non-essential category. Third, households may be able to switch some non-essential services from in-person to remote — for example from movie theatre entertainment to online streaming or from in-restaurant dining to take-out.

Figure 4 shows the percentage change in spending *within* each category, but how much did each category then contribute to the *aggregate* drop in spend? This requires understanding what share of aggregate spending went towards essential and non-essential categories at baseline.[5] These shares are shown, before and during the pandemic, in Table 1. Multiplying the baseline shares by their relative percentage drops, we find that non-essential spending accounts for 84 percent of the aggregate decline, and essential spending accounts for 16 percent.

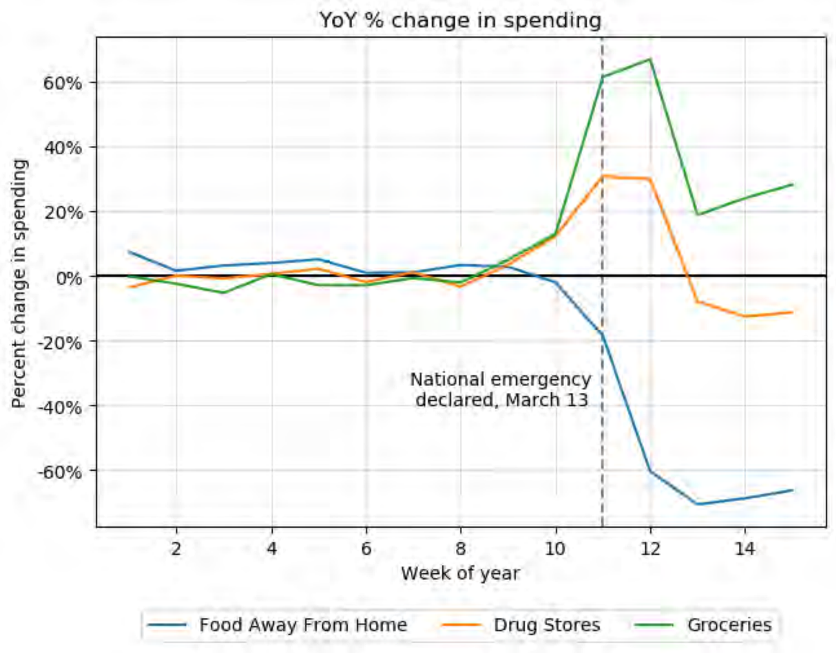
Table 1: The drop in non-essential spending accounted for 84 percent of the aggregate drop in spend.

	Essential		Non-Essential	
	Share of spending	Year-over-year percent change	Share of spending	Year-over-year percent change
April 2019	33%		67%	
April 2020	45%	-20%	55%	-51%
Contribution to Aggregate Drop in Spend*	16%		84%	

* Percent contribution to aggregate drop in spend is calculate as: (% Drop in Category A)*(Baseline Share of Category A)/(% Drop in Aggregate).

To further illustrate the divergence in spending patterns across essential and non-essential categories, we show the year-over-year change in spending at grocery stores, drug stores, and restaurants. Figure 6 shows that spending spiked dramatically on groceries and remained elevated relative to baseline. Spending on drugstores also increased initially, before declining slightly by the end of March and early April. In contrast, spending on restaurants fell by about 70 percent.

Figure 6: Year-over-year percent change in spending at grocery stores and drug stores surged initially, while spending at restaurants dropped by 70 percent.



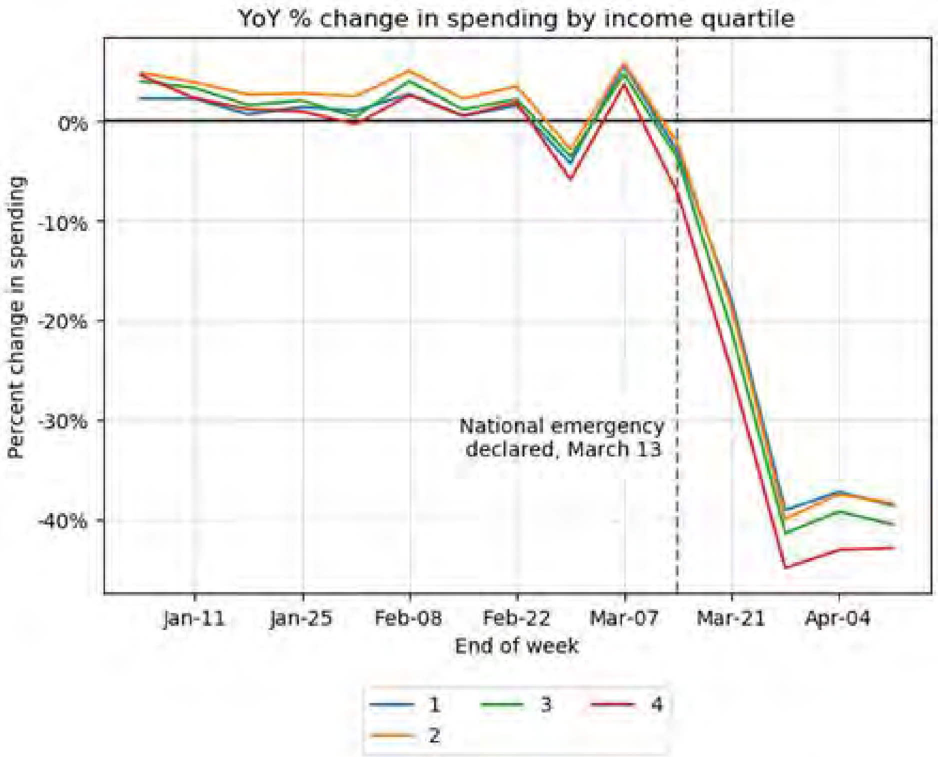
Finding 3: Spending dropped substantially for households across the entire income distribution, with slightly larger drops for higher-income households

We next explore whether spending reductions (both in aggregate and by category) varied across the pre-pandemic income distribution. We stratify our sample into income quartiles based on total labor inflows in 2019. For context, those in the bottom quartile make less than $39,000 in take-home labor income per year, while those in the top quartile earn more than $92,000.[6]

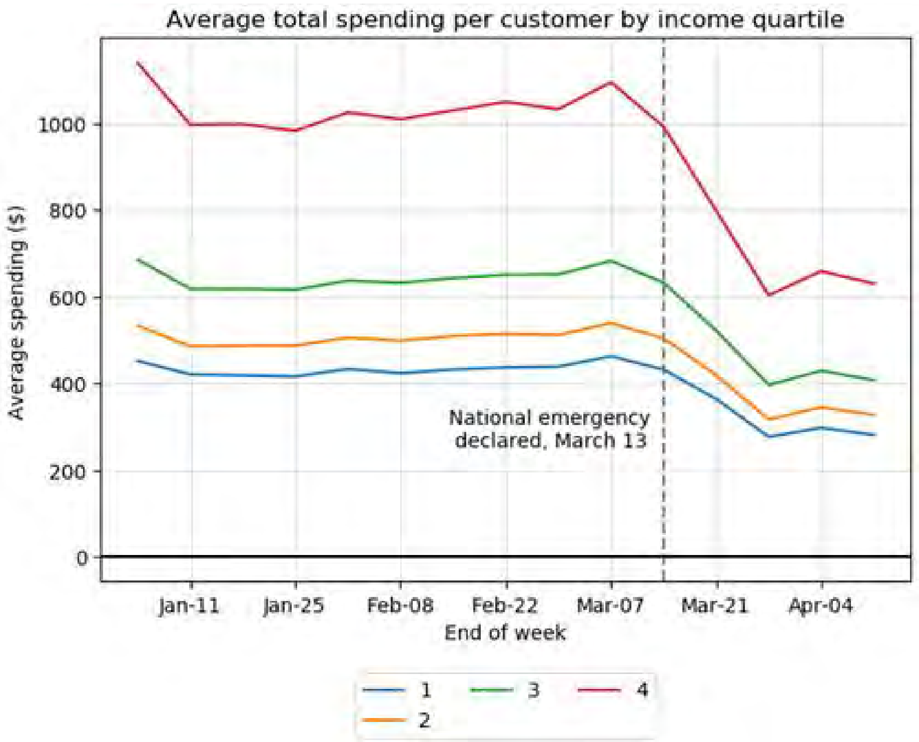
Figure 7 plots the year-over-year change in spending for each quartile, both in percentage and dollar terms. The top income quartile reduces spend by about 46 percent, or $400, by the second week of April, while the bottom quartile reduces spend by 38 percent, or $150. The difference in the spending drop between income quartiles is starker in dollar terms than percentages, since high-income households have a higher baseline level of spending.[7]

The sharp decline in spending across the entire income distribution may be surprising. Recent research suggests that lower-income households work in jobs that are harder to perform at home, require higher physical proximity, and therefore may be more impacted by distancing restrictions (Mongey, Pilossoph, and Weinberg 2020). Perhaps as a result, recent evidence from administrative ADP data shows that job losses were four times higher for workers in the bottom income quintile than in the top income quintile, with a staggering 35 percent employment decline for the lowest-income workers (Cajner et al 2020). In response to greater income losses, we might have expected lower-income workers to have cut their spending by more. If anything, we find the reverse—higher-income households cut their spending by slightly more.

Figure 7: Year-over-year reductions in aggregate spending are slightly larger for households in the upper portion of the income distribution.



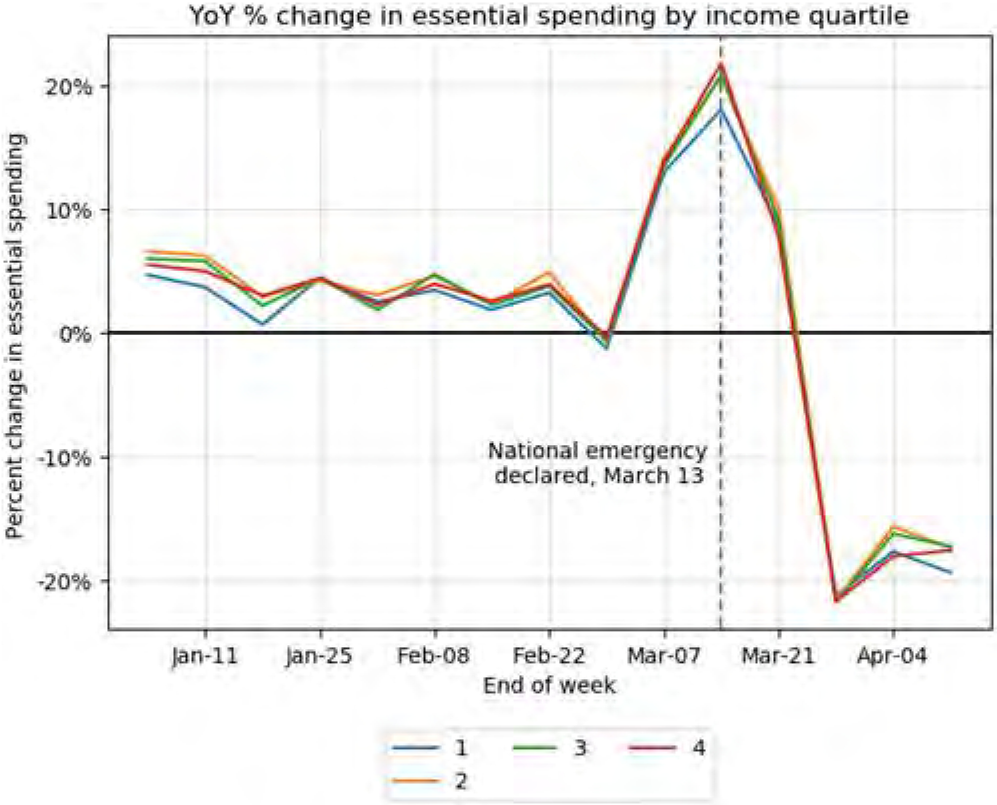
Note: Income quartiles are defined as follows: Quartile 1: less than $39,200; Quartile 2: $39,200 - $58,900; Quartile 3: $58,900- $91,800; Quartile 4: greater than $91,800.



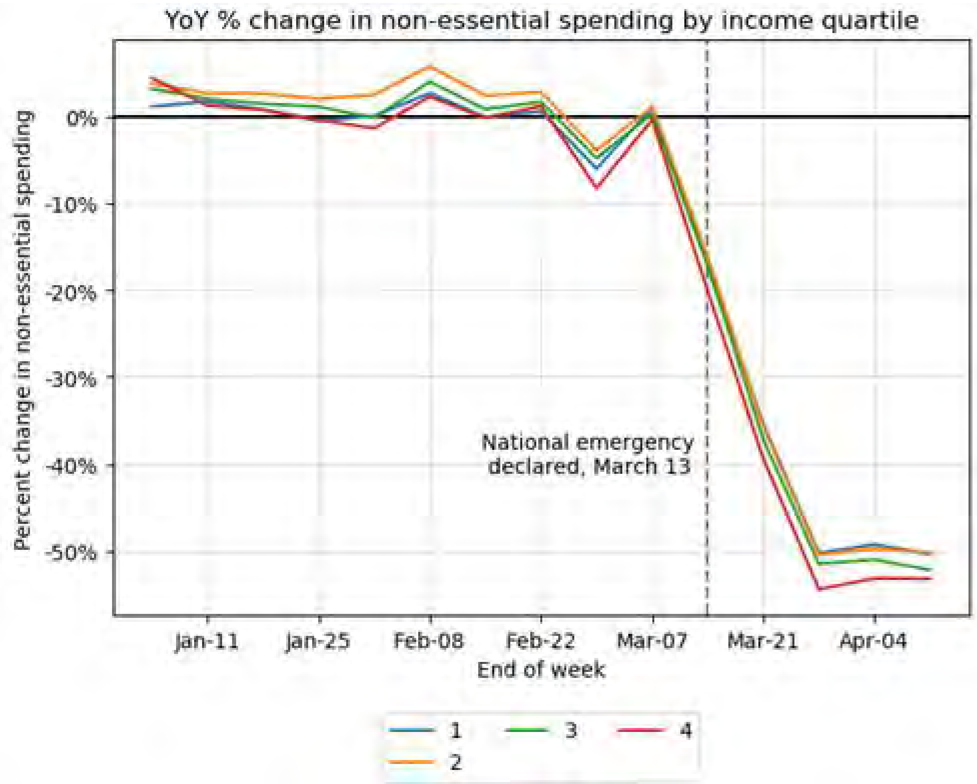
Note: Income quartiles are defined as follows: Quartile 1: less than $39,200; Quartile 2: $39,200 - $58,900; Quartile 3: $58,900- $91,800; Quartile 4: greater than $91,800.

One potential reason that high-income households cut total spending slightly more than low-income households could be that non-essential categories represent a larger share of spending for high-income households — 70 percent of spending in April 2019 for households in the top income quartile compared to 61 percent for those in the bottom income quartile. Additionally, higher-income families exhibited a slightly larger drop in non-essential spending, while we see little divergence across the income distribution in essential spending (Figure 8). Thus, the reduction in non-essential spending accounted for a slightly larger share of the total spending decline for high- versus low-income households (88 percent compared to 81 percent, Figure 9).

Figure 8: Year-over-year changes in essential spending were consistent across the income spectrum, while higher-income households cut non-essential spending slightly more than lower-income households.

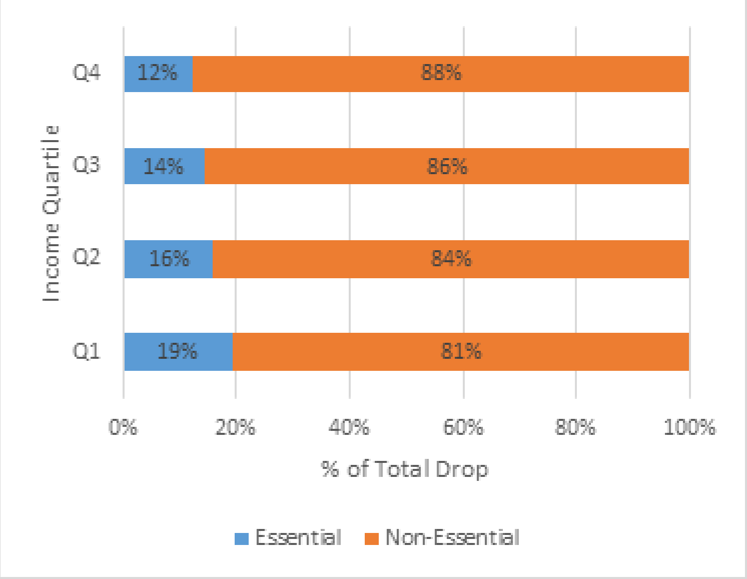


Note: Income quartiles are defined as follows: Quartile 1: less than $39,200; Quartile 2: $39,200 - $58,900; Quartile 3: $58,900- $91,800; Quartile 4: greater than $91,800.



Note: Income quartiles are defined as follows: Quartile 1: less than $39,200; Quartile 2: $39,200 - $58,900; Quartile 3: $58,900- $91,800; Quartile 4: greater than $91,800.

Figure 9: The drop in non-essential spending accounted for a slightly larger share of the drop in total spending among higher-income households



Note: Income quartiles are defined as follows: Quartile 1: less than $39,200; Quartile 2: $39,200 - $58,900; Quartile 3: $58,900- $91,800; Quartile 4: greater than $91,800.

Finding 4: Spending dropped dramatically for workers in all industries of employment.

Findings 1 and 2 show that drops in spend are especially pronounced in non-essential categories, and mirror the timing of emergency declarations, the implementation of social distancing policies, and the prevalence of the disease. Finding 3 shows that the spending drops were dramatic across the income distribution, even though income losses may have been more concentrated among lower wage earners unable to perform their duties from home. This suggests that the pandemic largely contributed to the reduction in spending.

Here we further examine whether income losses could also be playing a central role: as individuals working in affected sectors lose their jobs or see hours reduced, they may additionally cut down on spending. Indeed, Figure 3 shows that UI claims started spiking in the same week that commerce would have been restricted by state stay-at-home orders.

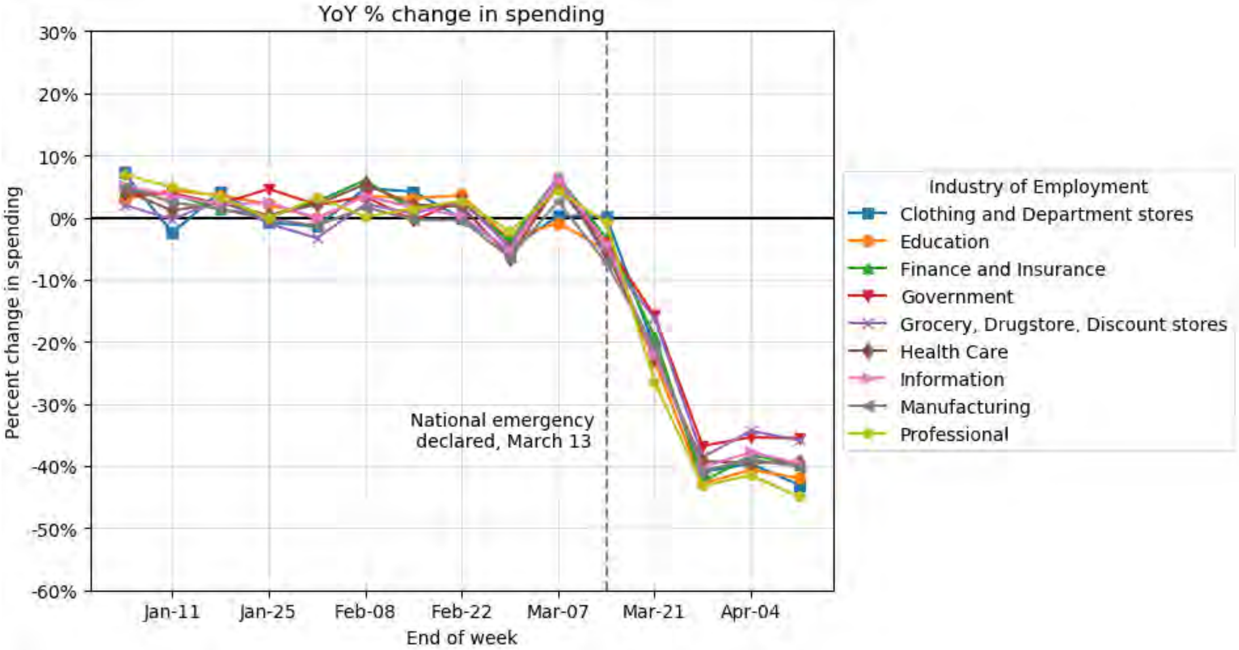
We speak to this hypothesis by splitting the sample by industry of employment and comparing spending across industries that may have been differentially impacted by earnings losses. We examine a subset of our credit card sample who also have a Chase checking account and infer their industry of employment based on the payer associated with their payroll income received in February 2020. However, we observe the payer associated with their payroll income for only 24 percent of households, and most of these payers tend to be large employers.

Figure 10 plots spending changes by industry of employment for each industry where we have significant sample size. We aggregate to industries at the two-digit NAICS code. The one exception is retail, which we break out into grocery stores, drug stores, and discount stores—generally considered essential businesses and kept open under social distancing policies—and clothing and department stores, which were generally deemed non-essential businesses and where layoffs have been greater (Cajner et al 2020).

We find that spending declined dramatically across all industries of employment. Workers in professional services, manufacturing, healthcare, education, and finance all cut spending similarly. For the most part, differences in spending declines between industries are not statistically significant. This may be surprising given initial evidence of large differences across industries in hours reductions (Bartik et al. 2020) and employment (Cajner et al 2020). Even government workers, who have experienced some of the lowest employment losses since the beginning of the pandemic, cut spending by about 35 percent. This is only a few percentage points lower than the 40 percent spending cut for all other workers.

Perhaps the most direct test of the income channel is to compare retail workers employed by different types of retail stores. Workers employed in grocery, drugstore, and discount stores cut spending by 35 percent, only a few percentage points less than the 41 percent cut in spending observed among workers employed by clothing and department stores, who might have experienced larger drops in earnings. Similarly, when we disaggregate the spending behavior into essential and non-essential spending, we see comparable spending drops across households with individuals who work in government and retail sectors (Figure 11).

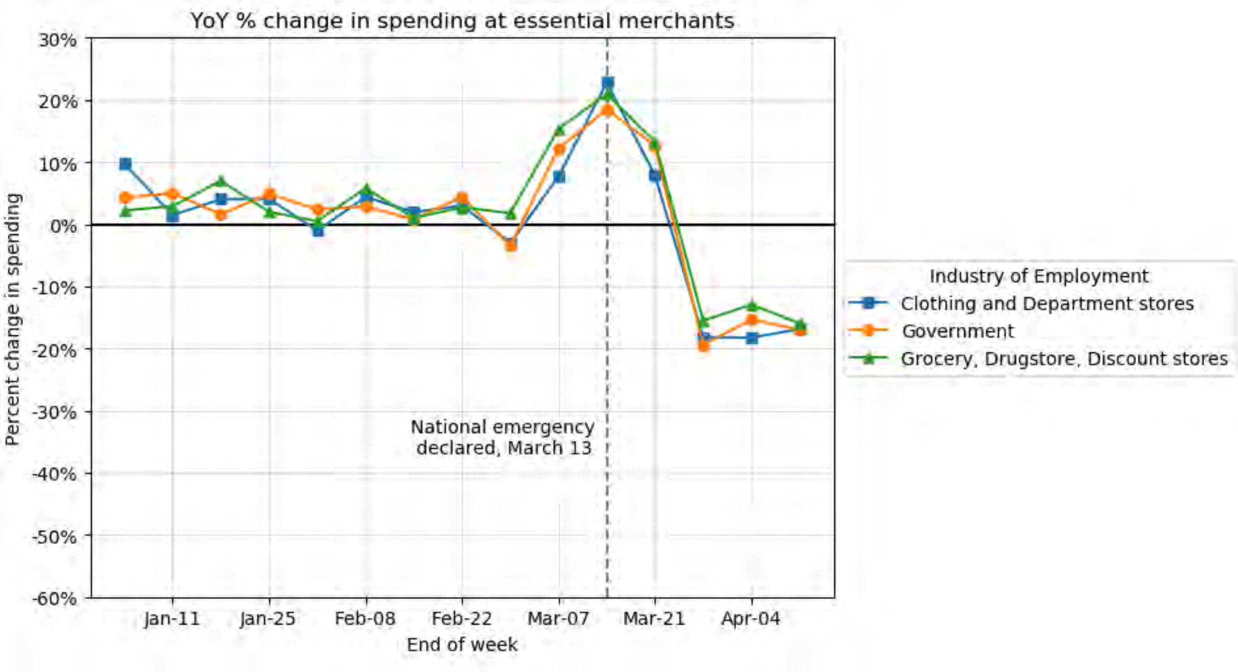
Figure 10: Spending dropped dramatically for workers in all industries of employment.



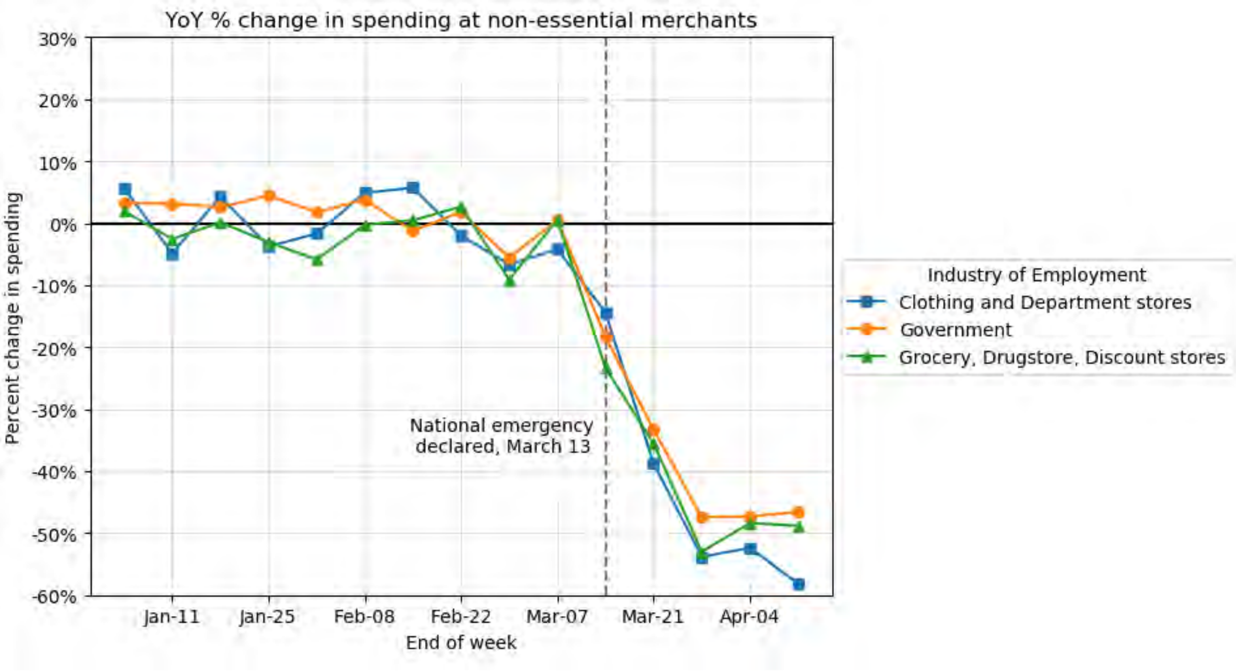
Note: Industry of employment defined as of February 2020.

Figure 11. Similar drops in spending on essentials and non-essentials occurred in industries with high and low rates of job loss.

Essential spending by industry of employment



Non-essential spending by industry of employment



Assuming income losses do vary systematically across sectors, one potential interpretation is that the income channel accounts for only a small share of the initial spending decline through mid-April. This may not be surprising given the magnitude of the spending decline. As mentioned previously, we document that average household spending fell almost 40 percent, while the typical unemployed worker receiving UI only cuts credit card spending around 5 percent in normal times (Ganong and Noel 2019).

However, there are at least four reasons for caution in concluding that the income channel is playing a small role even in this early phase of the pandemic. First, industry of employment may be a poor proxy for job loss in our sample. To the extent that we can ascertain industry of employment primarily for employees of large firms, we may not be capturing the income losses for employees of small businesses. Second, job losses may not yet have translated into income losses within our time frame. The peak of the UI claims occurred the penultimate week of our sample. Households who have lost jobs may still be receiving their last paychecks. We may expect to see larger income-related spending declines over time due to both past and future job losses.[8]

Third, current conditions of the pandemic make comparing the magnitude of the spending response in April 2020 to that of UI recipients during normal times highly uncertain. In normal times only one in four unemployed households receives UI, and spending declines may be greater for those who do not receive such benefits. Presently, due to the CARES act, UI benefits are much more generous in level and duration, available to many more workers, and may coincide with stimulus checks. Thus, current income supports might buffer against income-related spending declines to a greater extent. On the other hand, the economic situation is highly uncertain, and the labor market is rapidly weakening, which could cause the unemployed to cut spending to a greater extent.

Finally, we analyze spending solely on the universe of Chase credit cards, which may not fully capture the spending response to income loss due to both sample selection and measurement error. Since Chase credit card holders tend to be more affluent than the average U.S. household, we may be missing those households who might cut spending the most due to income declines. In addition, impacted individuals may turn to credit cards to finance their spending and to avoid the risk of infection posed by other means of transacting.[9] In the future we can test the limitations of the credit card sample by studying checking account transactions and spending on debit cards.

Conclusion

In summary, we provide two pieces of evidence suggesting that, as of the second week of April, the 40 percent drop in consumer spending appears to be driven to a greater extent by the pandemic and social distancing policies implemented across the country to prevent its spread and to a lesser extent by the initial round of income losses. First, the 40 percent drop in spending was observed across the income distribution and regardless of industry of employment. Second, the drop in spending was most dramatic at merchants which provide non-essential goods and services.

However, we analyze only the initial, short-run reaction of spending to the pandemic. The balance of factors contributing to spending behavior could change dramatically as the pandemic unfolds. If the virus and economic disruptions remain widespread even after social distancing restrictions are lifted, or if income supports, such as UI and stimulus payments, provide only temporary relief, consumer spending may not return to baseline levels. In future work we will continue to track the path of consumer spending and evaluate the extent and impact of income disruptions by extending and complementing our current view of credit card spending with checking account transactions.

[1] For the purposes of this report we have aggregated account activity to the primary account holder of the Chase credit card account, which could have one or more authorized users. For shorthand, we refer to this unit as a household, recognizing that members of households do not always link financial accounts.

[2] We included only customers who had a Chase credit card and at least three transactions in every month from January 2018 through March 2019. We do not apply this sample screen to April since we do not observe a full month of spending, though we report credit card spending through April 11. Ninety-seven percent of the sample has at least one credit card swipe the April 2020 period we observe, and our results are unchanged if we alternatively drop all households with no card swipes in this period.

[3] Specifically, in Finding 3, we take the subset of our 8 million Chase credit card customers who also have a Chase checking account and at least 5 ACH checking account transactions in every month between January of 2018 and December of 2019 and at least $12,000 in labor income inflows in 2018 and 2019, respectively. We use this sample of roughly 1 million account holders to segment our population by take-home income in 2019. In Finding 4 we take a different subset of credit card customers who also have a Chase checking account and who received labor income via direct deposit in February of 2020. For roughly 24 percent of these 1.3 million customers we observe the payer associated with their payroll income, which we then categorized by industry. These payers tend to be large employers with a median of 870 employees per employer observed in the Chase data. We additionally focus on the 97 percent of households in this sample with only one identified industry.

[4] Other data sources reported in the NYTimes, Baker et al. (2020), and Opportunity Insights show similar divergence in trends in essential versus non-essential spending categories.

[5] For example, the table shows that 33 percent of the average pre-COVID consumption basket was composed of non-essential items. A back-of-the-envelope decomposition suggests that if non-essential spending was completely disallowed during the pandemic, while essential spending stayed constant, we would see a 33 percent drop in overall spend. The drop in aggregate spending would be 100 percent attributable to a non-essential spending decline.

[6] Specifically the income quartile cutoffs are as follows: Quartile 1: less than $39,200; Quartile 2: $39,200 - $58,900; Quartile 3: $58,900- $91,800; Quartile 4: greater than $91,800.

[7] Nonetheless, the drop in spending is on the same order of magnitude across the income spectrum, a finding consistent with Baker et al. (2020), who show using a sample of relatively low-income users from a non-profit Fintech company that there is little difference in the spending pattern between households with incomes above and below $40,000.

[8] We are unable to address these concerns now because our income data are only available through February 2020. However, we will be able to directly measure income losses by industry in future work once we observe income during the pandemic.

[9] Although Ganong and Noel (2019) found that total nondurable spending fell similarly to total credit card spending at the onset of unemployment, households may behave differently during this crisis.

A Survey of Machine Learning in Credit Risk

Joseph L Breeden
Prescient Models LLC
breeden@prescientmodels.com

May 30, 2020

Abstract

Machine learning algorithms have come to dominate some industries. After decades of resistance from examiners and auditors, machine learning is now moving from the research desk to the application stack for credit scoring and a range of other applications in credit risk. This migration is not without novel risks and challenges. Much of the research is now shifting from how best to make the models to how best to use the models in a regulatory-compliant business context.

This article seeks to survey the impressively broad range of machine learning methods and application areas for credit risk. In the process of that survey, we create a taxonomy to think about how different machine learning components are matched to create specific algorithms. The reasons for where machine learning succeeds over simple linear methods is explored through a specific lending example. Throughout, we highlight open questions, ideas for improvements, and a framework for thinking about how to choose the best machine learning method for a specific problem.

Keywords: Machine learning, artificial intelligence, credit risk, credit scoring, stress testing

1 Introduction

The greatest difficulty in writing a survey of machine learning (ML) in credit risk is the extraordinary volume of published work. Just in the area of comparative analyses of machine learning applied to credit scoring, dozens of articles can be found. The goal of this survey cannot be to index all work on machine learning in credit risk. Even listing all of the worthy articles is beyond the attainable scope.

Rather, this survey seeks to identify the major methods being used and developed in credit risk and to document the breadth of application areas. Most importantly, this article seeks to provide some intuitive insights on why certain methods work in specific areas,. When does machine learning work better than linear methods only because it was a quicker path to an answer versus discovering something about the problem that was undiscoverable with traditional methods? Further, as a result of this research, we hope to identify some areas of investigation that could be fruitful but have not yet been fully explored.

In attempting to provide a balanced view of the state of machine learning, some passages herein may take a tone that machine learning is ”much ado about nothing”. In other discussions, we are clearly singing the virtues of deep learning with discussions of ensemble methods for robustness, deep learning to analyze alternate data, and techniques for modeling the smallest data sets. Machine learning can be seen to be clearly successful in some cases and disturbingly overblown in others, bringing new innovations in important areas and painfully rediscovering old methods in some cases, and overall has made significant strides toward mainstream application while still having significant challenges to overcome.

The article begins with a definition of machine learning intended in part to limit the scope of this survey to a manageable breadth. The next section offers a modeling taxonomy based upon defining data structure, architectures, estimators, optimizers, and ensembles. From this perspective, much machine learning research is a human-based search of the meta-design space of what happens when you mix and match among those categories. Then, Section 4 provides a discussion of the many application areas within credit risk and some of the model approaches found within each. Section 5 reviews a specific example of testing many machine learning algorithms to illustrate the differences relative to traditional methods. Section 6 follows with a discussion of significant challenges in creating machine learning models and using them in business contexts. The conclusion pulls these thoughts together to highlight areas where future comparative studies could provide significant value to practitioners.

2 What is Machine Learning?

We tend to think of statistical models and linear methods as something other than machine learning, and yet simple linear regression can take on unbounded complexity through factor variables, spline approximations, interaction terms, and input massive numbers of descriptive variables through dimension reduction methods such as singular value decomposition. The heart of many machine learning algorithms is a search or optimization method that was pioneered decades or centuries ago in other contexts. Bagging, boosting, and random forests harken back to earlier work on ensemble methods [66, 213].

Harrell [126] proposes a distinction between statistical modeling and machine learning:

- Uncertainty: Statistical models explicitly take uncertainty into account by specifying a probabilistic model for the data.
- Structural: Statistical models typically start by assuming additivity of predictor effects when specifying the model.
- Empirical: Machine learning is more empirical including allowance for

high-order interactions that are not pre-specified, whereas statistical models have identified parameters of special interest.

The above items carry other implications. For example, search-based methods such as Monte Carlo simulation, genetic algorithms, and various forms of gradient descent usually do not provide confidence intervals for the parameters, and correspondingly are usually considered as machine learning. Ensemble methods where multiple models are combined are generally considered to be machine learning, even when the constituent models are statistical. One might also say that traditional statistical methods rely on analyst selection of input features and interaction terms whereas machine learning methods emphasize algorithmic selection of features, discovery of interaction terms, and even creation of features from raw data.

Drawing the line between machine learning and traditional modeling is challenging for the best scientific linguist. Practically speaking, machine learning seems like it should include models that emphasize nonlinearity, interactions, and data-driven structures and exclude simple additive linear methods with moderate numbers of inputs. The distinction may be more in the specific application than the method used. For example, an artificial neural network could be dumbed down to a nearly-linear adder, and common logistic regression can incorporate almost all the learnings from a sophisticated machine learning algorithm through artful use of binning, interaction terms, and segmentation.

Some methods might be viewed as intermediates, like transitional species in evolution. (The author recognizes that "transitional species" is a misnomer in evolutionary taxonomy, but the perspective is not inappropriate here.) Forward stepwise regression or backward stepwise regression automate feature selection while being statistically grounded. Principal Components Analysis (PCA) is an inherently linear, statistical method of dimensionality reduction via eigenvalue estimation, whereas other dimensionality reduction methods lean much more to machine learning. One of the greatest strengths of neural networks is as a nonlinear dimensionality reduction algorithm.

Within this attempted dichotomy, many machine learning techniques are rapidly taking on statistical rigor. This maturing process is what we see in any field where rapid advances are followed by a team of scientists filling in theoretical and technical details.

Many of the most public successes of machine learning are coupled with big data, massive data sets that allow equally massive parameterizations of the problem so that the optimal transformations of the inputs and dimensionality reduction are learned from the data rather than via human effort. However, machine learning should not be viewed as synonymous with big data. Some machine learning methods appear well suited to very thin data sets where even linear regression struggles. Eventually, as we truly move into human-style AI, the ability to learn from a single event in the context of a 'physical' model of the world would show the power of machine learning with the smallest of data.

In credit risk, we are often stuck with small data. This was observed in the credit scoring survey by Lessmann, et. al. [175] where only five of the

48 papers surveyed had 10,000 accounts or more to test, quite small samples compared to the big data headlines, but this is often the reality of credit risk modeling. For many actual portfolios, number of accounts * loss rate = very few training events. Even in subprime consumer lending where loss rates are higher, only the largest lenders have had the data sets needed to apply the most data-hungry techniques like deep learning, or so it seems. However, machine learning is succeeding in credit risk modeling even on smaller data sets, apparently by emphasizing robustness and simpler interactions as opposed to the extreme nonlinearities in big data contexts such as image processing [162], voice recognition [120], and natural language processing [69].

Machine learning is generating successes in credit risk, although less dramatically in well-worn domains like prime mortgages. The biggest wins appear to be in niche products, alternate channels, serving the underbanked [5], and alternate data sources. A well-trained machine learning algorithm may be preprocessing deposit histories [2], corporate financial statements, twitter posts [199], social media [97, 24, 10] or mobile phone use [38, 232] to create input factors that eventually feed into deceptively simple methods like logistic regression models.

Also, in looking at applications of machine learning to credit risk, we must look beyond predicting probability of default (PD). One of the great early success stories of ML was in fraud detection [110]. Anti-fraud [297], anti-money laundering [254, 273, 20, 217, 178], and target marketing applications [180, 22] make heavy use of machine learning, but are outside the boundaries we will draw here around credit risk applications. Still we must consider applications to predicting exposure at default, recovery modeling, collections queuing, and asset valuation, to name a few.

The following sections aim to provide an introduction to the literature on machine learning methods, applications in credit risk, what makes machine learning work, and what are the challenges with employing machine learning in credit risk.

3 Machine Learning Methods

Providing an exhaustive list of machine learning methods would not be possible, particularly when we look beyond credit scoring to the broader applications of machine learning across credit risk modeling. One of the greatest challenges of creating any list of models is the difficulty in defining a model. The name given to a model typically represents a combination of data structure, architecture, estimator, selection or ensemble process, and more. Authors may swap out one estimator for another or add ensembles on top and describe it as a new model. This abundant hybridization leads to exponential growth in the literature and model names. Finding the right combination is, of course, very valuable, but the human search through this model component space with publications as measurement points is more than can be cataloged here.

In this section we will identify key sets of available components behind the models and then categorize some of the most studied models according to the components used. Of course, each of these lists can never be complete. They are intended only to be representative.

3.1 Data Structures

Choosing a data structure is the first step in either statistical modeling or machine learning. That model must be chosen to align with the data being modeled. A range of target variables are possible in credit risk and those variables can be observed with different frequencies and aggregation, depending upon the business application.

Table 1 lists some of the outputs one might wish to model in the domain of credit risk. Items like PPNR [183] and Prepayment [240] might not seem like credit risk tasks, but when they are modeled divorced from credit risk modeling, the result can be conflicting predictions leading to nonsensical financial projections. Taking a consistent, coordinated perspective of all account outcomes and performance as in competing risk architectures and models [168, 92] is the best hope of predicting pricing and profitability.

Even deposit modeling can leverage very similar methods and works best when a total customer view is taken. Deposit balances are a potentially valuable input to credit risk models, but are not always categorized as credit risk targets. Anti-fraud, anti-money laundering, and target marketing were considered as separate from credit risk, because they are not part of the analysis of an active customer relationship, although even here the boundaries are weak.

Target Variables
Loss Balance
PD, EAD, LGD [and PA (Probability of Attrition)]
Prepayment
Pre-provision net revenue (PPNR)
Asset Values
Deposit Balance*
Time Deposit Renewal Rate*

Table 1: List of target variables that can be modeled in credit risk applications. The items marked with * are likely candidates for using the same methods as the other items, but not strictly considered credit risk issues.

For any target to be modeled, a decision must be made on the aggregation level and performance to be predicted. Table 2 lists the most common answers. Each type of data usually has a corresponding literature. Econometric models [85, 277] focus on time series data, either for a portfolio or segments therein; Age-Period-Cohort models [113, 195, 106] are applied to vintage performance time series; survival models [256, 152] and panel data models [283, 140] are applied to account performance time series; and the large literature on credit scoring [257, 14] focuses mostly on account outcomes, using a single binary performance indicator for each account.

Data Types
Segment Time Series
Vintage Performance Time Series
Account Performance Time Series
Account Outcomes

Table 2: List of data types that can be modeled in credit risk applications.

By starting the discussion with target variables, what follows is immediately focused on supervised learning. The assumption is that unsupervised learning techniques might be used to create input factors. Many forms of dimensionality reduction and factor creation can be conducted using unsupervised methods. PCA and most segmentation methods can be considered unsupervised learning. However, a credit risk model will ultimately always finish with a supervised learning technique.

3.2 Architectures

Once the problem is stated as a target variable to be predicted and its data structure as in Tables 1 and 2, an architecture must be chosen for the problem, Table 3. This is the point where the distinction between traditional methods and machine learning can appear.

Additive effects refers to regression approaches [153, 131]. Additive fixed effects includes the use of fixed effects (dummy variables), again in a regression approach, panel model, etc.

State transition models, [206, 26] (also known as grade, rating, or score migration models depending upon whether they are applied to delinquency states, risk grades, agency ratings, or credit scores) are all variations on Markov chains [207]. Roll rate models [89] capture the net forward transition of a state transition model and are used throughout credit risk modeling. Generally, this architecture involves identifying a set of key intermediate states and modeling the transitions between those states and to the target state. Usually the target is a terminal state like charge-off or pay-off.

Going beyond the above architectures leads more into the realm of machine learning, although there are again few fixed boundaries. Convolutional networks [162], feed-forward networks [253, 15], and recurrent neural networks [182] are all kinds of artificial neural networks and are just a few of the many structures being tested.

Whenever the nonlinearity of a problem exceeds the flexibility of the underlying model, segmenting the analysis is a common solution. The more nonlinear the base model, the less segmentation is required. Traditional logistic regression models may actually be a collection of many separate regression models applied to different segments, whereas a neural network or decision tree may use a single model.

Some models are themselves segmentation engines. Methods such as support vector machines (SVM) [266] use hyperplanes or other structures to segment the parameter space. Decision trees [226] can also be viewed as a high-dimensional segmentation technique and are employed in a variety of machine learning approaches. Nearest neighbor methods [71, 130] are difficult to classify in this architectural taxonomy, but seem closer to these than the rest.

Architecture
Additive Effects
Additive Fixed Effects
Convolutional Network
Clustering
Feed-forward Network
Fuzzy Rules and Rough Sets
Nearest Neighbors
Recurrent Neural Network
Segmentation
State Transitions
Trees

Table 3: List of internal architectures used in modeling.

Fuzzy rules are used to capture uncertainty directly in the forecasting process [202] and are often combined with other methods [220]. Rough sets [219] can be seen as having a similar objective of considering the vagueness and imprecision of available information, but using a different theoretical framework.

Recurrent neural networks (RNN) are used primarily to model time series data. By making the forecast from one period and input to the network for the next period, they are effectively a nonlinear version of vector ARMA models ((Multivariate Box-Jenkins) [197, 177]. Long short term memory (LSTM) networks apply a specific architecture to the recurrent neural network framework in order to scale and refine the use of memory in the forecasting.

Overall, many architectures can be used in time series forecasting. The same lagged inputs used in linear distributed lag models [11, 291] can be used as inputs to machine learning methods. To reduce the dimensionality of the problem and aid visualization, optimal state space reconstruction can be used, also known as the method of delays [216, 234, 52, 165].

Convolutional neural networks (CNN) being applied to consumer transaction data [167] seems far from the leading applications in image processing, but many more applications of CNNs are likely, particularly with recent advances incorporating rotational [78, 65] and other symmetry transformation to increase the generalization power of CNNs.

Not shown is the list of possible inputs, because this would be too extensive.

3.3 Estimators and Optimizers

The primary purpose of this modeling taxonomy is to illustrate that, for example, a genetic algorithm is not a model. Practitioners, both experienced and novice, often use sloppy terminology confusing data structures, architectures, and optimizers. Here we illustrate that many different estimators and optimizers can be applied in an almost mix-and-match fashion across the range of architectures. By clearly identifying the components of a model, researchers can find opportunities for creating useful hybrids.

The literature also attempts to carefully distinguish between estimators and optimizers. In simple terms, estimators all rely on a statistical principle to estimate values for the model's parameters, usually with corresponding confidence intervals and statistical tests in the traditional statistical framework. Optimizers generally follow an approach of specifying a fitness criteria to be optimized. As parameter values are changed, the fitness landscape can be mapped. Each optimizer follows a specific search strategy across that fitness landscape. Of course, here again it can be difficult to draw bright lines between these categories as estimators and optimizers can take on properties of each other.

Table 4 lists some of the many methods used to estimate parameters or even meta-parameters (architectures) of a model. Items such as back propagation are specific to a certain architecture, e.g. back prop as a way to revise the weights of a feed-forward neural network. Most, however, can be applied creatively across many architectures for a variety of problems.

Estimators
Least Squares
Maximum Likelihood [218]
Partial Likelihood [72]
Bayes Estimator [35]
Method of Moments [196, 121, 98]

Table 4: List of statistical estimators used in modeling.

Maximum likelihood estimation is the dominant statistical estimator, which is, for example, behind the logistic regression estimation that is ubiquitous in scoring and many other contexts. Least squares estimation predated maximum likelihood but can be derived from it. Partial likelihood estimation was a clever efficiency developed for estimating proportional hazards models without estimating the hazard function parameters needed in the full likelihood function.

Aside from some deep philosophical issues, Bayesian methods are particularly favored when a prior is available to guide the solution. Markov chain Monte Carlo (MCMC) starts with a Bayesian prior distribution for the parameters and uses a Markov chain to step toward the posterior distribution given the data, somewhat like a correlated random walk.

In data-poor settings, Bayesian methods provide a powerful mechanism for combining expert knowledge from the analyst with available observations to obtain a more robust answer. Computing a batting average in baseball is an easy way to illustrate this. Someone who has never swung could be assumed to have a 50/50 chance of hitting the ball, a .500 average. After their first swing, a miss would take his batting average to .333 and a hit would take it to .667. With a maximum likelihood estimation, the best fit to the data would be .000 for a miss and 1.000 for a hit, which seems less helpful until more observations are acquired. This is Laplace's Rule of Succession. Not coincidentally, Laplace also formulated Bayes' Theorem.

With method of moments, the moments of the distribution are expressed in terms of the model parameters. These parameters are then solved by setting the population moments equal to the sample moments.

Linear programming and quadratic programming are methods for incorporating constraints. Many other constrained optimization methods exist, such as Lagrange multipliers which provide a mechanism for adjusting the fitness function to incorporate penalty terms.

Optimizers
Gradient Descent
Simulated annealing [159] Back Propagation [128]
Reinforcement Learning
Genetic Algorithms [115]
Evolutionary Computation [151]
Genetic Programming [161]
Markov Chain Monte Carlo [108]
Kalman Filter [111]
Linear Programming [265]
Quadratic Programming [36]

Table 5: List of optimizers used in modeling.

Gradient descent can be accomplished via several specific algorithms, but it generally refers to computing the local gradient of the fitness landscape at a test point and stepping in the direction with the steepest slope, hopefully toward the desired minimum. Back propagation is gradient descent in the context of a neural network where the gradient is computed for each node's parameters. Reinforcement learning is the more general concept of adjusting parameters, usually in a neural network context, based upon new experiences. Kalman filters are an optimal update procedure for linear, normally distributed models, which could be thought of as a subset of reinforcement learning.

Genetic algorithms, evolutionary computation, and genetic programming are all modeled on evolutionary principles. In an optimization setting, mutation operations with survivor selection are equivalent to stochastic gradient descent. Including cross-over between candidates works if symmetries exist in the fitness landscape such that sets of parameters form a useful sub-solution within the model.

Also not shown are the many estimation methods developed to handle correlated input factors such as Lasso [258] and ridge regression [133].

Of course, many of these concepts can be combined. Stochastic back propagation and stochastic gradient descent [41] are widely used. Simulated annealing can be thought of as combining the stochastic gradient descent concept with the multiple candidate solutions approach of evolutionary methods. Bayesian methods can be combined with many other optimization approaches, such as Bayesian back propagation [56] or MCMC as described above.

3.4 Heterogenous Ensembles

Ensemble modeling is actually a general technique that can combine forecasts from different model types. "Triangulation" has been a common technique over several decades for portfolio managers to create loss forecasts by comparing the outputs of several different models, each with different confidence intervals and known strengths and weaknesses. Voting is largely a formalization of what managers have been doing intuitively, with several interesting variations [263, 166].

Ensemble modeling [73, 66, 213, 79, 222] has been in use well before the burst of activity in machine learning, but has quickly proven itself to be a valuable addition to most any machine learning technique, particularly in credit risk [271]. Most research into ensemble modeling can be split between homogenous methods, where multiple models of the same type are combined to create better overall forecasts and heterogenous methods where any types of models can be combined. We also consider a third category of hybrid ensembles where two complimentary model types are integrated via mechanisms more specific to the methods than in the generic heterogenous ensemble approaches.

For an ensemble to be more effective than the individual contributors, Hansen & Salamon [123] showed that the individual models must be more accurate than random and the models must not be perfectly correlated. In other words, we cannot create useful forecasts from a collection of random models, and the best ensembles have constituents that have complimentary strengths.

Ensemble modeling seems particularly well suited to credit risk, because of the typically limited data sets available. Although the underlying dynamics can be quite complex and explainable with a rich variety of observed and unobserved factors, the actual data available may support models of only limited complexity. Even though many factors can be important, issues of multicolinearity [205] can limit the modeler's ability to include more than a few factors and is often a deeper problem than is generally recognized [118, 49]. Dimensionality reduction methods such as singular value decomposition, principle components analysis, [150] and projection pursuit [103, 102, 146] are methods to address multicolinearity, but they do not address the sensitivity to outliers and overfitting questions as well as the full nonlinearity treatment available in machine learning.

The basic principle behind ensemble modeling is that different models can capture different aspects of the data. This can provide robustness to outliers and anomalies [281] as well as which factors are included in the modeling. Both theoretical [123, 163, 141] and empirical studies have shown that this diversity when obtained for individually accurate predictors has significant out-of-sample advantages.

Heterogenous Ensemble Methods

Binary	Categorical	Continuous
Plurality Voting	Plurality Voting	Average
Sum rule	Majority voting	Median
Product rule	Sum rule	Confidence weighted
Stacking	Product rule	Stacking
	Amendment vote	
	Runoff vote	
	Condorcet count	
	Pandemonium	
	Borda count	
	Single transferable vote	
	Stacking	

Table 6: List of methods used to combine forecasts in heterogenous ensemble modeling.

Table 6 lists some of the methods used for combining forecasts in ensemble modeling of potentially heterogenous models. Many of these methods were developed from the perspective of choosing from several possible categories [263]. In a broader credit risk context, we can have situations with binary outcomes, e.g. default or not; multiple (categorical) outcomes, e.g. transition to different states; or continuous outcomes, e.g. forecasting a default rate.

Combining forecasts for binary events can be performed with several methods. Voting methods are the most common, where each constituent model gets one vote. Plurality voting is the simplest of these, where the outcome with the most votes is chosen. If the constituent models produce probabilities or some kind of fractional forecast, then each constituent model can divide its vote proportionally between the two outcomes, which are then summed. Classification methods can be modified to produce probabilities to facilitate their use more broadly [221, 164]. In the product rule, these fractional votes are multiplied, which means extremely confident models can dominate an outcome.

When predicting multiple possible outcomes (categorical outputs), the above methods can be generalized easily. If addition, majority voting is different from plurality voting, where one outcome must have a majority of the votes. In no outcomes have a majority, the least favored outcome is removed and a majority is sought among the remaining outcomes. A run-off vote is a simple extension of the majority voting process until a single outcome remains.

Amendment voting starts with a majority vote between the first two candidate outcomes. The most favored is tested against the next candidate until one outcome remains. However, this procedure can be biased depending upon the sequence of comparisons.

The Condorcet count performs pairwise comparisons of all outcomes. The favored outcome from each comparison receives one point and the outcome with the most points is chosen. Although complex, this has many favorable properties.

In Selfridges Pandemonium [241] method each model would choose one outcome, but that vote is stated with a confidence. Those weighted votes are summed to choose a winner, meaning that model confidence intervals become important.

If the constituent models cannot assign a probability to all possible outcomes, as needed for sum rule and product rule, but the models can rank the outcomes, then ranked voting can be used. The outcome can be chosen by mean rank [40], median rank, or a trimmed mean or median rank.

Single transferable vote also works from ranks, although not every model must rank every outcome. If one outcome has a majority of the top ranks, it is chosen. If not, the least preferred outcome is eliminated and the top ranks are re-aggregated. The procedure continues until one outcome receives a majority.

Beyond voting, one could imagine creating a model of models. In a linear regression context, this does not introduce any new information beyond the initial estimate. However, with stacking [282] the initial models are trained on a subset of the total data. Then a secondary, often linear regression, model is trained on the hold-out sample, considering model accuracies and correlations. Machine learning methods can also be used to create models of models [259].

One advantage of ensembles is the ability to create confidence measures for classification models, although direct, single-model approaches are also available [224].

For continuous-valued predictions, averages, medians, trimmed values, and stacking all apply. Continuous forecasts are often or in best practice should be accompanied by confidence intervals. Therefore, weighted averages or some method that incorporates those confidences would be preferable.

3.5 Homogeneous Ensembles

Any method for combining heterogenous model predictions can of course be applied to homogenous models, where multiple models of the same type are built to be combined. However, some methods have been specifically designed to work with homogenous ensembles.

3.5.1 Bagging

Bootstrap aggregation (bagging) [53, 173, 179] is a simple process of subsampling the available training data with replacement. Considering the typically limited size of the training samples in credit risk, the subsets can be 75% of the available data and upward. Bagging can be used with any model type and the resulting forecasts combined as described for heterogenous models, although the sum rule is used most often [160].

For random subspace modeling [132], a random sample of the available input factors is drawn for each model. This could also be done sequentially deterministic fashion, where the strongest explanatory variable from the first model is excluded from the next model in order to find structure among other variables, and so forth. The first application was for creating decision trees, leading to the literature on random forecasts, but the technique is generic to any model type.

Rotation forests [230] follow the random forests idea, but all of the data is used each time. Instead, a rotation of the axes in the data space for a subset of input factors is performed prior to building each model. This has the effect of testing many different projections for predictive ability.

Similar to the bagging concept is to use all of the training data each time, but different initial conditions for the parameter estimates. For model types such as neural networks [66] or decision trees [9] that employ some form of learning or gradient descent, this can also create a robust ensemble.

3.5.2 Boosting

Conceptually, one could say that boosting is a process of building subsequent models on the residuals of previous models, though for model types that have no explicit measure of residuals [236, 235]. AdaBoost [99] reweights the training data with each iteration to emphasize the points that were not predicted as well in the previous iterations. Gradient boosting [100] computes the gradient of a fitness function in order to provide weights to each model trained. Stochastic gradient boosting [101] combines bagging with gradient boosting, building an ensemble of ensembles where different gradient boosted ensembles are built for each data sample. These methods can also be applied to any model type. The popular XGBoost package (eXtreme Gradient Boosting) [63] is a highly optimized version of gradient boosting.

Many studies have been performed to compare ensemble methods [204, 271], but the winning approach probably depends upon the specific problem and data set. For example, gradient boosting has been reported to be more susceptible to outliers.

3.6 Hybrid Ensembles

A very large area of research involves creating hybrid models, where specific model types are chosen that are intended to be integrated in non-trivial ways, usually via an algorithm specifically tailored to the models chosen and to the application area. This is different from heterogenous ensembles where the forecasts are combined via one of the voting schemes in Table 6. Instead, hybrid ensembles create an architecture that leverages the specific traits of the models. The criterion for success is not about choosing which models are most orthogonal and accurate [123]. Rather, it involves combining models that may (1) use different data sources, (2) predict over different forecast horizons, or (3) identify different problem structures. So the models are inherently complimentary, often making measures like orthogonality or comparative accuracy undefined.

A classic example in credit risk is the use of roll rate models [89] for portfolio forecasting for the first six months combined with vintage models [45] for the longer horizon forecasts. In this case, the analyst would usually switch from one model to the other at a certain forecast point or use a weighting between the models that is a function of forecast horizon. Some version of this approach has been in use for decades, because roll rates are known to be accurate for the short term and vintage models for the long term.

The list of hybrid ensembles (or hybrid models) in the literature is far too great, but these provide a few examples: decision trees and neural networks [171], support vector machines and neural networks [70, 6], naive Bayes and support vector machines [201], a classifier ensemble with genetic algorithms [294], and genetic algorithm and artificial neural networks [214]. Some authors provide surveys of collections of hybrid ensembles generally [17] or for specific application areas such as bankruptcy prediction [269]. Hybrids combining age-period-cohort (APC) models [195, 113, 136, 106] with origination scores, behavior scores, neural nets, or gradient boosted trees were created specifically to better solve the short economic cycle data described above [46, 50, 51]

4 Applications in Credit Risk

Machine learning methods received early attention from researchers, but adoption into operational contexts has been understandably cautious for reasons to be discussed in Section 6. The earliest experiments were primarily in fraud detection, credit scoring [188, 77, 129, 279, 288], corporate bankruptcy and default forecasting [209]. As machine learning methods have matured along the lines described above, parallel efforts occurred in the application of those techniques to areas of credit risk, resulting in a wide range of new applications.

4.1 Credit Scoring

Credit scores were created to predict the relative risk of default among borrowers [68, 176]. Their success as compared to human judgment was so great that they became part of the standard credit bureau offering and an essential part of the lending ecosystem. These bureau scores have been developed and refined over decades and are essentially the result of an optimization process where the disparate and complex consumer performance history has been linearized into factors what fit well into a logistic-regression model. This would seem to be the same kind of work done automatically by machine learning, but historically done through human intuition and experimentation.

Anecdotally, developers of modern bureau scores are said to use machine learning methods to search for additional interaction terms and nonlinearities. Those lessons are taken back to the original logistic regression-based model to create small improvements, but the advances available from machine learning appear to be small compared to the decades of human optimization already performed. However, Hand & Henley [129] showed that even small enhancements to credit score performance can have significant returns.

In principle any institution can purchase data from the bureaus similar to what goes into creating the bureau scores and do a head-to-head test of in-house machine learning model to bureau score. In any such test, the in-house model has a great advantage in that the target is known. When developing a bureau score, the model is attempting to predict default without knowing what product the consumer will be offered, or if default will come in the absence of new loans and based purely on existing loans. An in-house model is typically built to predict the outcome of offering a new loan of a specific type and perhaps even incorporating the terms of that loan. Fair comparisons are difficult, but perhaps unimportant. A developer creating an in-house model can jump straight to sophisticated modern methods, either taking the bureau score as an input or starting fresh, in each case bypassing the decades of labor put into the original bureau scores.

Machine learning in credit scoring is not new. Comparative surveys can be found as far back as 1994 [228]. New comparative analyses continue to appear as new methods are developed and more data becomes available. One of the most complete surveys was conducted by Lessmann, et. al. (2015) [175] in which they noted the irony that most published work on machine learning in credit scoring leveraged only very small data sets for comparing "big data" machine learning methods. Lessmann, et. al. sought to resolve that shortcoming by testing multiple methods on multiple, larger data sets. These surveys are useful both in bringing the readers up to date on the latest methods and in suggesting which methods could be best, but no single method wins in all studies [21]. The obvious conclusion is that not all data sets have similar structures, and the analyst can still expect to test several approaches to find which is most effective on a specific data set. Similarly, researchers need to be careful to avoid publishing conclusions that one method is better than another based only upon one data set over one time period.

4.1.1 Neural Networks

Neural networks are one of the most extensively tested methods for credit scoring and one of the first machine learning methods employed [149, 77, 268, 279, 190]. They can function like a nonlinear version of dimension reduction algorithms such as principal components analysis or as factor discovery methods in deep learning contexts. They offer additive and comparative interaction terms between variables. On the most basic level, neural networks provide a nonlinear response function between input and output. With enough training data, these attributes can be a powerful combination.

The first challenge with applying neural networks is in choosing an architecture. In theory, with enough data, a fully connected, feed-forward neural network should be able to learn its own architecture, but reality is more challenging. Some of the biggest success stories in using deep learning neural networks required vast amounts of training to determine the meta-parameters for the networks: number of inputs, number of hidden layers, number of nodes in each layers, activation functions for the nodes, etc.

Therefore, much of the work around neural networks is in how to choose or learn an optimal architecture. Genetic algorithms have been used to select the optimal set of inputs [288, 23]. Classic genetic algorithms performed cross-over and mutation on a binary encoding of the parameter space [116]. That binary encoding is rarely optimal for applications in credit risk [43]. A more general evolutionary approach [151] could operate on the full architecture of the neural network in order to share optimal subnets across candidate networks within a population.

Feed-forward networks are the most commonly used, largely because they are the easiest to train and comprehend. However, recurrent neural networks have been used to create memory within the network rather than have the analyst provide lagged inputs of dynamic variables in behavior scoring contexts [142]. When applied to massive amounts of input data, such as transactional information, convolutional neural networks have been applied [167].

Even with an optimal architecture, limiting overfitting [255, 172, 249] is a significant problem. Much work has been done in this area with some surprising findings that the number of parameters in deep learning networks may not be as much of a problem as we think [30]. One explanation may be that the initial random assignment of many small parameters might actually create robustness to input noise rather than the multicolinearity nightmare we would otherwise expect.

Even worse can be transient structures that are actually present in the data, but only for a short period of time. When we know that a certain structure will not persist in the future, such as an old account management policy of an expiring government program, how does one get the neural network to forget? One answer could be the ‘given knowledge’ approach suggested by Breeden and Leonova (2019) [51] where we could train a subnet on just the transient structure, embed this as a fixed component of a network trained to solve the larger problem on the full data set, and then remove the subnet when creating forecasts out of sample.

Neural networks are data hungry and time intensive to train, but can be successfully used. Many authors have studied these effects, comparing different neural network designs and comparing them to other methods [21, 190, 233, 170, 284, 3, 93]. When the available data is wide in the number of inputs but short in the number of observations, ensembles of small networks can also be effective [280].

4.1.2 Support Vector Machines

Support vector machines (SVMs) excel at creating segmentations of the input vector space for classification. The ability to segment the observation space with arbitrary hyperplanes provides an effective classification technique for an arbitrary number of end states and without assumptions about the distributions of the input factors or target categories. They are less well suited to continuous prediction problems, although techniques mentioned earlier can be applied to product continuous outputs. SVMs have been applied to credit scoring by multiple authors and found to be an effective approach in many cases [264, 285, 21, 237].

One of the biggest advantages in SVMs is the ability to use kernels to create optimally separating hyperplanes (OSHs). The "kernel trick" refers to the chosen maps the data to a higher-dimensional space, which can in some cases dramatically simplify the process of finding OSHs. The placement of the hyperplanes is a nonlinear problem requiring an optimizer.

As with neural networks, the challenge is optimizing the architecture. With SVMs, the input features and the kernel parameters must be optimized. The choice of whether to use a linear, polynomial, radial basis function, or other kernel is a matter of experimentation given a specific data set. No universal best answer exists, but the best advice is to start simple (linear) and move toward complex as required.

These choices across meta-parameters are interdependent. To optimize these meta-parameters, GAs have again been applied [104] and other hybrid approaches [143]. The lesson from studies into neural networks and SVMs is that optimizing the meta-parameters is essential to success.

4.1.3 Decision Trees

Decision trees are a simple concept that can be used to create sophisticated models. The concept is a recursive partitioning of the input space until enough confidence is achieved to make a prediction. They have been used for decades in credit risk [188, 75, 105] where the earliest decision trees were heuristically created. Modern algorithms can use a variety of partitioning criteria: misclassification error, Gini index, information gain, gain ratio, ANOVA, and others. The final forecast can be the state with the greatest representation in the final leaf, a probability based upon representation, or a small model as in regression trees [54, 91]. The meta-parameters are how to optimize the partitioning, the input factors, and when to stop partitioning. As usual, these need to be optimized.

A single tree can have the same overfitting concerns as previous methods, but the explosion in the use of decision trees has come with the introduction of ensembles. Bagged decision trees [293], boosted decision trees [27], random forests [164, 189, 109], rotation forests [204, 193], and stochastic gradient boosted trees [261, 60] are some of the most popular. Most authors agree that this list represents a steady improvement in methodology, currently with stochastic gradient boosted trees as the usual winner. Although ensemble methods are most popular in scoring when applied to decision trees, these methods are found combined with all credit scoring techniques [7].

One advantage of decision trees is the mapping between trees and rules. Trees can be compared to known rules and rules can be learned from trees [64].

In general, trees have an advantage in handling sparse data or data with outliers. Binning is a simple method to limit outlier sensitivity that is lacking in continuous methods like neural networks. In situations where the data is abundant, of good quality, and with clear nonlinearities, neural networks are often the reported winners.

4.1.4 Nearest-neighbors and Case-based Reasoning

One category of models could be defined as those that learn from past examples. Case-Based Reasoning (CBR) [57] searches through past lending experiences to find a comparable loan. In commercial lending where examples are few and nearly unique, this can be an effective approach. Where most data is available, as with consumer lending, a kNN (k-nearest neighbors) [129, 191] approach is conceptually equivalent.

The challenge with both CBR and kNN lies with identifying comparables. This is not unlike the challenge for home appraisals. If the closest comparable home is at a distance, in a different kind of neighborhood, is it really comparable? This concept applies to both methods here. Any data set will be non-uniformly distributed along the explanatory factors. When optimizing the metric for identifying comparable loans or choosing ”k” in kNN, the definition of a near neighbor that works well in one region of the space may be a poor choice in another.

Using geography as an example, finding 20 neighbors in an urban setting might provide a roughly homogenous set, whereas finding the same 20 neighbors in a sparse geography could span counties or even states. Of course, using CBR or kNN geographically could create a redlining risk, but the same concept applies, if more abstractly, to any set of explanatory factors. Therefore, the success of these methods appears to be tied to the uniformity of the distribution of the data set.

Where CBR and kNN may excel are in extremely sparse data situations. When tens of events or less are available, especially when the events are very heterogeneous in their properties, matching to prior experience without attempting to interpolate or extrapolate as in estimation-based approaches may be more effective.

4.1.5 Kernel Methods, Fuzzy Methods, and Rough Sets

Kernel methods, fuzzy methods, and rough sets are best viewed as a method to augment other modeling approaches. Decision trees, support vector machines, or any method that performs classification by drawing hard boundaries among the input factors will inevitably have uncertainty in the location of those boundaries. In general, one would assume that the greatest forecast errors should occur near the boundaries. Incorporating estimation kernels [287, 59] into these methods or treating the boundaries as fuzzy [134] can capture this uncertainty and potentially improve accuracy by reporting appropriate probabilities. Estimation kernels or fuzzy logic have been incorporated into many credit scoring methods [275, 289, 295, 119]. This may be particularly valuable in sparse data settings where the boundaries can only be approximations.

Rough sets have also seen application to credit scoring [198]. With an objective similar to kernel and fuzzy methods, rough sets have been combined with other base modeling techniques to incorporate the imprecision of the available information. Along these lines, rough sets have been combined with decision trees [296] and with SVMs [62].

4.1.6 Genetic Programming

Genetic programming (GP) employs trees to perform computation. The leaves are input values or numerical constants. The branching nodes contain numerical operators or functions. In this way, nested algebraic operations can be performed to create predictions for credit scoring [212, 144, 4].

The genetic aspect refers to how the tree structure, constants, and input factors are chosen. As with genetic algorithms, concepts of mutation and crossover are employed. Mutation is a simple change in a constant, swapping an input factor, or swapping an operator or function. Crossover is the more interesting process of swapping subtrees between two trees. In genetic algorithms applied to binary representations, crossover rarely produces viable offspring because the fitness landscape lacks useful symmetries. In GP applications to credit scoring, such symmetries exist if subtrees can capture conceptual subsets of the problem, such as swapping the proper transformation of an input factor between candidate trees..

For credit scoring, the fitness function will be one or several measures of forecast accuracy in predicting the target variable. The optimization naturally occurs on an ensemble of candidate trees. The best tree at the end of the optimization process can be used as the model, but following the ensemble concept, one could also apply a voting algorithm across all qualifying trees. However, one challenge with genetically learned ensembles is that they tend to cluster around a single peak in the fitness landscape. A similarity penalty could be added to the fitness function to encourage diversity in the population, both to reduce the risk of being stuck in a local optimum and to increase the usefulness of the ensemble.

GP appears to be useful as a highly nonlinear method. To justify the slow search speed of genetic methods, one needs a problem that is equally complex. Simple credit scoring problems may not qualify, but the use of alternate data sources might make GP more interesting.

4.1.7 Alternate data sources

Some machine learning methods for credit scoring are specifically focused on how to incorporate new data sources into the scores. Cash flow analysis using data scraped from demand deposit accounts is a successful area of business application, particularly during the COVID-19 pandemic during which so much traditional scoring data is in doubt. Although the data source is new to scoring, the methods for analysis are more traditional. One seeks to determine the frequency and reliability of income by different sources. During the COVID-19 pandemic, someone with periodic, steady income could be a good credit risk regardless of credit score, industry of employment, or many other underwriting criteria.

Mobile phone data is potentially an important data source in emerging markets and specifically for underbanked consumers. Research on credit risk for Chinese consumers using mobile phone calling records and billing information has been found to be effective for credit risk assessment [274]. Research in underdeveloped markets using smartphone metadata such as types of apps installed, text message history, etc. [274]. Both studies used well known credit scoring and machine learning methods, just with emphasis on sourcing and regularizing new data types.

Some novel data sources can require corresponding innovations in analysis. Social network data has proven to be quite interesting [278, 97], but incorporating data from networks into a credit score can be a challenge. Low-dimensional embeddings of network graphs [122] are the standard approach to creating a usable input factor for modeling. However, recent research [242] suggests that low dimensional embeddings lose much of the information in the network. The best approach for incorporating social network data will continue to be a topic of research for some time, as will the ethics and legality of incorporating such data within the underwriting process.

4.2 Corporate Defaults

Discussions of credit scoring usually carry an implication of consumer loans and large volumes of training data. Modeling corporate defaults and bankruptcies is a similar problem, but with fewer events in the training data and less standardized inputs. A panelist at a conference on machine learning in finance explained humbly that they used machine learning just to read the corporate financial statements. The scoring models were trivial. In fact, standardizing diverse and heterogenous inputs may be one of the best uses of machine learning in lending applications.

Even so, some large data sets on corporate defaults do exist, and a variety of papers have been published to apply ML to the problem [262, 247, 13] . Bankruptcy and default are not exactly the same thing, but bankruptcy filings are public so many works have focused there [200, 248, 198, 209, 67, 19, 267]. Across both applications, the methods tested cover the full range of machine learning techniques.

4.3 Other Scoring Applications

Published work often lags what is being done in-house at lenders around the world. For example, the author knows that prepayment and attrition models have been created using machine learning, with the short study in Section 5 as one such previously unpublished example. At this point, one can assume that machine learning is being tested everywhere models can be employed in lending.

4.3.1 Loss Given Default

The natural companion to credit risk forecasting is modeling loss severity, loss given default (LGD). LGD, or conversely, recovery modeling, has always been a challenging problem because of the inherent tri-modal distribution [239, 28]. Some percentage of borrowers will show no net loss in event of default because of the collateral value. Another significant percentage can be expected to have 100% LGD because of failure to recover the collateral (such as a totaled vehicle), and a distribution can exist between the two extremes. LGD has been modeled as a multi-stage problem where the first step is to predict 0, 1, or Intermediate and the second stage attempts to predict the specific value for the intermediates.

In addition to economic sensitivity [33], LGD can also depend on the age of the loan and the time since default. One approach is to use survival or age-period-cohort methods to predict monthly recoveries from the date of default with vintage defined by month of default.

Naturally, given the importance of LGD (the 2009 US mortgage crisis was as much an LGD crisis as a PD crisis because of the collapse in property values), work has also been done to apply machine learning to LGD [252, 246] or recovery rates [32]. Given the complexity of the problem, research into which approach is best for different asset classes could continue for some time.

4.3.2 Automated Valuation Models

The valuation of property in collateralized loans is part of the underwriting process providing a preview of what LGD could be in the event of default, much the way en primeur wine ratings are an early estimate of what the quality of a finished wine will be [8]. Automated valuation models (AVMs) replace the human property appraiser with a data driven model to speed the approval process and lower the origination costs. Machine learning methods are also being applied to AVMs [260, 31] where regression-based approaches have been previously deployed [86, 107].

4.4 Portfolio Forecasting and Stress testing

Time series applications of machine learning provide an interesting contrast to machine learning applications in credit scoring. The challenges and best methods are almost completely opposite. In credit scoring, large data sets are obtained by observing many accounts, transactions, or behaviors over a short period of time. Success comes largely through identifying nonlinearities and interactions. For time series modeling, the available data sets are very short relative to macroeconomic cycles [51] and credit cycles [48]. Some studies have reached questionable conclusions, because what looks like a linear response over a short time period may in fact be a cyclical response to a completely different factor when observed over a longer period.

The longest data sets in lending usually extend only as far back as the mid-1990s. For the US, that translates to only two clear recessions (2001 and 2009), although some subcycles can also be observed [47]. At the time of this writing, the COVID-19 Global Recession is just beginning. This event may add more clarity about tail risk in our models and the kind of government responses we can expect in the event of extreme events.

When faced with short (in time) and wide (in variables) data sets, different approaches will be preferable to what was seen in credit scoring. In fact, most lenders struggle to obtain data back to 2006, which would qualify as one full economic cycle in the US. Modeling portfolio responses through a single economic cycle is akin to having four data points in a credit score: one good, one bad, one mediocre trending worse, and one mediocre trending better. The usual validation of testing on the last 12 months of data is no more than a continuity test, not a true out of sample test.

Regardless of the technique employed, creating time series forecasts and stress test models is about creating robustness much more than worrying about subtle interaction effects or subtle nonlinearities. As such, an ensemble of small regression models is more likely to succeed in the next recession than a deep learning neural network. Although decision trees are very successful in scoring, any binned method is less suitable to forecasting rates if it truncates the tail of the distribution. Continuous models only extrapolate to the tails of the distribution by making assumptions, but those can be explicitly expressed.

Therefore, forecasting and stress testing are applications where machine learning must be combined with intuition. Business and analyst experience serve as a human-powered smoothing and regularization technique for models that do not truly have enough examples to train upon. Some commentators have suggested that state-level or MSA-level modeling can solve the problem of not having enough economic cycles for training, but US states are highly correlated. Some lead-lag effects are present, but no state missed the 2009 recession. Oil shocks can create recessions in energy states like Alaska, North Dakota, West Virginia, and West Texas, but we are far from having 50 separate macroeconomic responses to model with 50 states.

This discussion should not be taken to imply that nonlinearities are unimportant. In fact, transformations of macroeconomic inputs must be carefully chosen. Percentage change in gross domestic product (GDP) is not a good fit to a linear regression model, because increases and decreases are not symmetric. Taking the logarithm of the ratio of the values would produce a roughly normally distributed distribution that is symmetric in changes, and thus better suited to use in a linear regression model. In short, either variables need to be transformed to scale linearly with the target, or the model needs to be flexible enough to learn the nonlinearities. However, in a limited data environment, there may not be enough information to learn the nonlinearities from the data, so human assistance through choosing the transforms is one path to success.

Using interest rates in loan default models provides an effective example. Over the last decade, analysts have commonly taken the natural log of interest rates or the log of the ratio of interest rates in order to control for the fact that a change from 4% to 3% is much more important than a change from 15% to 14%. Unfortunately, we have entered a realm where interest rates can go negative and logarithm-based transformations are not suitable. Therefore, we need to find transformations that are more linear through zero but less sensitive for large values. Transformations such as $y = tanh(x)$ and sigmoid functions in general, as well as $y = sign(x)\sqrt{|x|}$ are reasonable candidates.

Conveniently, this observation about suitable transformations fits well with neural networks, leading to the thought that ensembles of short, wide neural networks could be an effective approach for portfolio forecasting and stress testing. Short because sufficient data is only available to support one to a few hidden layers. Wide because many macroeconomic factors can be taken as inputs and the neural network provides a nonlinear equivalent to dimensionality reduction like PCA. Ensemble because many such models trained on randomly selected data subsets when combined provide robustness relative to the limited economic cycles available.

Genetic methods like GP can and probably have been used to swap transformed macroeconomic factors between models, much as has been demonstrated in credit scoring [288]. However, the data sets in time series models are small enough that exhaustive search among input factors is often possible. Forward stepwise regression and backward stepwise regression are also common approaches.

Although the author has observed that ensembles of small time series models can be quite effective, they pose model risk management challenges under current practices, which is discussed in Section 6.

The target variables for time series modeling can be delinquency rates, default rates, charge-off rates, prepayment rates, and recovery rates. All of this can be combined to create time series forecasts of expected losses, payments, and revenue ultimately leading to cash flow modeling needed for estimating yield or loss reserves under CCAR, CECL or IFRS 9. No technical obstacle exists to creating the outputs with machine learning enhancements such as ensembles of nonlinear models, but in the author's experience, auditors are not yet ready to use them to produce numbers in financial statements.

4.4.1 Recurrent Neural Networks

Recurrent neural networks and LSTM were designed specifically to learn lag structures from data in time series problems, so one would expect that they be tested for loan loss stress testing. Although quite successful in speech recognition, challenges exist in application to credit risk time series modeling.

As mentioned in the previous section, the primary issue is with the number of events in the data. In a training data set for speech recognition, every vowel is another cycle in the data, unlike the extreme data sparsity in stress testing. However, they may yet find a niche.

We already know that each recession has unique aspects. Models of loan defaults need to focus on the direct drivers of borrower cash flows. However, when we build models across multiple recessions, in cases where we have data on multiple recessions, the lags and cross-correlations between economic factors change. In 2008, a collapse in house prices preceded a decline in GDP and a subsequent drop in unemployment. In the 2020 COVID-19 recession, declines in GDP and unemployment rate are the leading effects and house price and commercial real estate declines follow. A simpler model will simply average across these structures, producing an unfocused model structure.

One alternative could be to use some form of regime switching [225] that detects the nature of the recession and switches between models corresponding to different types of crisis [26, 186]. Although plausible, the data is limited. The question not yet answered is whether some form of recurrent neural network could perform the equivalent of regime switching in a smoother, more continuous way and thereby adapt better to differing macroeconomic structures. This would only be conceivable with the longest data sets, possibly between 1995 and 2021, for example, to capture three or four recessions with at least three clearly different types of economic crises. That time is not so far in the future and it will be interesting to see what can be done to improve the state of the art in stress testing.

4.4.2 Survival and Vintage Models

Survival and vintage models occupy a middle ground between scoring methods and top-down time series models. Vintage models, such as Age-Period-Cohort (APC) models, operate simultaneously on multiple time series segmented by vintage (origination date) cohort so that dynamics versus age of the loan, credit risk by vintage, and environmental impacts may be quantified and used in forecasting [113, 286, 106]. Survival models operate on individual account performance data, but with the important addition of when an event occurred, rather than simply if something occurred, as with traditional credit scores [95]. Both methods can produce the periodic forecasts required for forecasting, stress testing, cash flow modeling, and pricing.

Both survival and vintage models include a function of risk by age of the account known as the hazard function or lifecycle function, respectively. The estimation of this function is inherently nonlinear, so the now-standard methods developed decades ago should fairly be considered machine learning methods. Nonparameteric estimation [82, 157], parametric and spline estimation [231] as in APC, and Bayesian methods [238] are all standard approaches. Neural networks or even decisions trees will probably be tested for estimating hazard functions, although the necessity is not clear given available nonparametric methods.

Account-level Cox proportional hazards models [256, 250, 80] and other survival scoring techniques [50] importantly include a scoring aspect that can be performed with a version of regression or more generally with machine learning techniques as well [87, 37]. Wang, Li, and Reddy (2019) [272] provide a thorough survey of machine learning survival methods to date.

The environmental or econometric modeling aspect of survival and vintage models can be addressed via the time series methods discussed in the previous section. Therefore, although survival and vintage models were machine learning methods from the start, they are being aggressively hybridized with the latest techniques [51]. One of the great advantages of these methods is the proven separability of nonlinear effects in age of the account, vintage, and environment by calendar date [135]. That separability creates a semi-structured approach where each of those pieces can be estimated by the methods and data set most suitable to the problem while the underlying mathematical structure guarantees a consistent framework for combining the pieces.

Naturally, ensemble methods for survival and vintage models have also appeared. Random survival forecasts and other ensembles [147, 139, 138] have been reported to be quite successful for credit risk modeling.

4.5 Portfolio Optimization

Modern portfolio theory [192, 84] is based upon stable linear expected returns and covariances for a set of possible instruments. Experience has shown that expected returns and covariances are rarely stable or linear, so this area is also being explored for enhancement with machine learning. In a lending context, optimization is constrained to the limits of how much certain asset types can be grown and whether or how much holdings can be reduced.

Optimization under modern portfolio theory can be viewed as optimizing the Sharpe ratio [245], defined as the ratio of expected return to expected volatility. Even without machine learning, many enhancements have been offered to this view, such as the Sortino ratio [88] which looks only at the volatility arising from negative returns.

In the context of optimizing a lending portfolio or any investment portfolio that includes loans, one needs to consider unique aspects of loan losses. As seen more dramatically for retail portfolio, increases in losses can occur because of lifecycle (loss timing) effects or intended changes in credit quality. Similar to the way the Sortino ratio computes volatlity without penalizing for positive increases, loan loss volatility and correlations should not include structure that is a feature of the product or intended management. Simulation-based methods exist for recreating historic loss time series to remove such expected variances [44].

Once we understand the true covariance structure with loan products, portfolio optimization methods based upon machine learning should apply equally well here as with general investment portfolios where they were originally developed. Early work focused on capturing nonlinearities in the covariance structure, tail risk, and boundaries. Copulas have seen significant application in this area [42, 154], but neural networks [58], genetic methods [243, 169], fuzzy optimization [184] and others have also been used. Ban (2018) [25] provides a review of available methods.

5 What Makes ML Work

In a 2018, Casey Foltz at Oregon Community Credit Union (OCCU) used 23 different machine learning algorithms readily available in R to compare checking account attrition models. More than just a comparison of AUC values, the project's goal was also to understand the reasons for the winners and losers. Experiments were conducted on how to modify the inputs and model meta-parameters in order to explore what made one method work better than another and how to improve the weaker performers.

The explanatory factors included a number of measures of fees incurred, transaction errors by the financial institution, complaints and denied credit applications. As would become important later, most of these variables were not normally or even lognormally distributed. The outcome variable was binary, attrite or not attrite during the two year observation period.

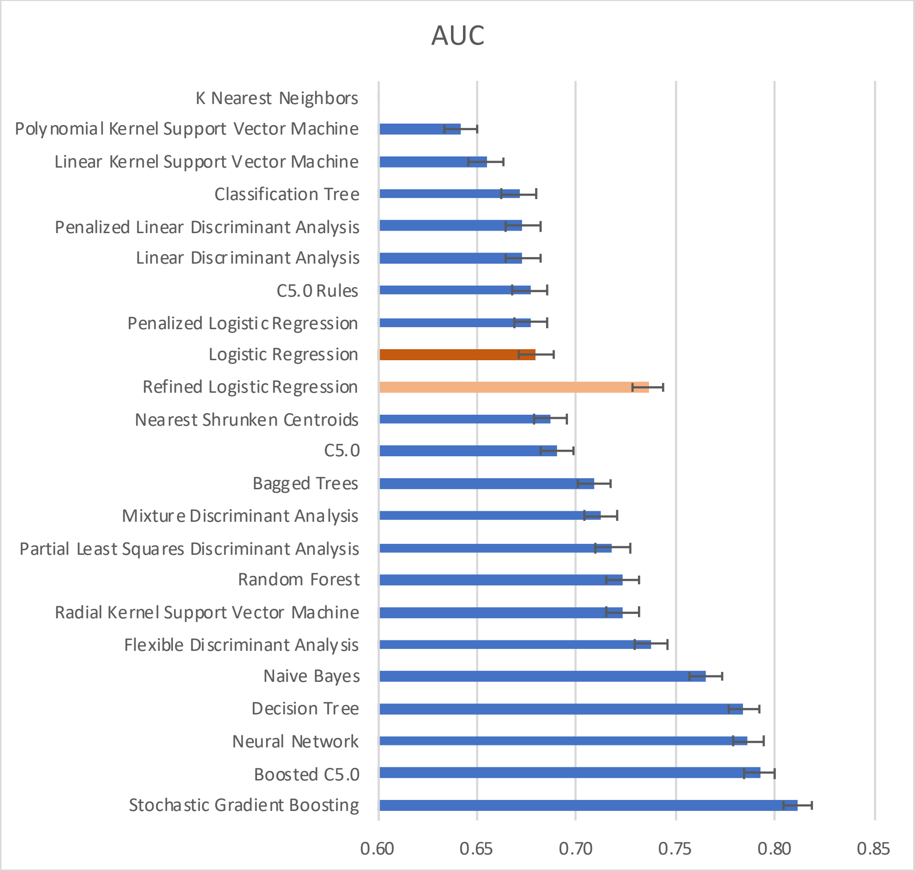


Figure 1: A comparison of AUC values for models of checking account attrition probability.

The explanatory factors and target variable were fed unmodified into each of the available algorithms with default parameters. Figure 1 shows the initial comparison between the methods. In reviewing these results, the first step was to review the meta-parameters of each method. For example, the neural network was implemented with `nnet` in R, which only allows for a single hidden layer. Packages like `tensorflow` allow for significantly more flexibility, but that would involve quite a bit more exploration. However, even unoptimized, the neural net was the third best approach.

Other methods could also benefit from optimization. For example, the linear discriminant analysis performs best with normally distributed variables. Applying logarithmic transformations to the variables with roughly lognormal distributions added 5% to the AUC value.

Stochastic gradient boosted trees was the winning method in this study of checking account attrition. However, the deeper question was why. How far could we push the logistic regression model toward the stochastic gradient boosted tree's performance?

To investigate this question, three simple things were done. Lognormally distributed variables were transformed with a logarithmic function. All other variables were binned so that graphs of test factor versus probability of attrition were created. For a large number of variables, those graphs showed the data to exhibit two regimes with linear relationships to attrition on either side of a break point. Therefore, those variables were split with an interaction term allowing for two different linear responses. Finally, the odd variables were just manually binned. With the couple dozen available input variables, this exercise took about an hour of manual work. The result is also shown in Figure 1. Confidence intervals for the AUC values were computed according to DeLong, et. al. [76].

The logistic regression model moved from the bottom third of the methods to the upper third. Decision trees and neural networks still performed better than the refined logistic regression, implying that more could be done to linearize the input factors and identify needed interaction terms. Capturing nonlinearities has previously been shown to be important for credit risk modeling [185], so this result is not surprising.

The improvements come from the boosted methods, employing multiple models rather than a single model. Figure 2 shows the ROC curves for the original logistic regression model, the refined logistic regression model, and the stochastic gradient boosted trees.

This is just one of numerous examples in the literature, but it illustrates the progression of predictability gained through adapting to nonlinear responses, interaction terms, and ensemble models.

6 Challenges of Employing Machine Learning

For all its promise, machine learning presents some unique challenges to application in credit risk. Unlike applications in speech recognition or image processing, accuracy alone is not sufficient in lending. FCRA guidelines require that lenders not discriminate against protected classes and that consumers are offered explanations for denial of credit. Such concerns have dramatically slowed the

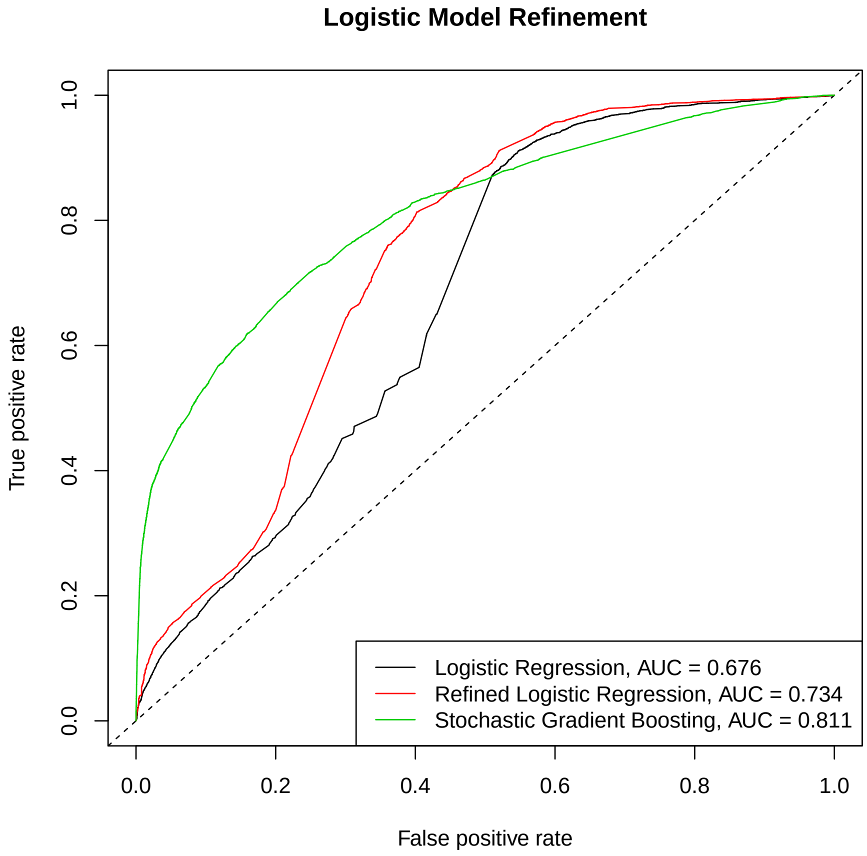


Figure 2: ROC curve comparison for the original logistic regression model, a refined logistic regression, and stochastic gradient boosting.

adoption of machine learning, and with good reason. These and other valid concerns in model risk management must be addressed before the models can be widely adopted.

Note that Sections 6.4 Unintended Bias, 6.5 Adverse Action Notices, 6.6 Predicting without Understanding, and 6.7 Adapting to Sudden Behavioral Shifts are all facets of explainability. Collectively, model explainability is the most critical challenge to widespread adoption of machine learning in credit risk. Financial institutions and regulatory bodies [39] cannot rely on models that they do not understand, for all of the reasons listed in these sections.

6.1 Large data needs

The promise of machine learning comes largely from the ability to incorporate nonlinearities in the input variables and interdependence between variables. That promise is fulfilled only in the presence of very large data sets both for identifying the structure and for testing to make sure the structure is not spurious. While those data sets are appearing in some contexts, some machine learning models are being built without the requisite data.

Conversely, the many studies that compare machine learning methods in hopes of identifying which is best often fail to note that the answer is strongly tied to how much data is available. In sparse data environments, k-nearest neighbor models may beat neural networks. With large, complex data sets, deep learning neural networks are likely to win. With intermediate data sets that produce many spurious to transient correlations, boosted trees might come out ahead. The simple answer is that we are unlikely ever to crown a single winning method, because the data sets and output requirements vary so widely, even in specific contexts like credit risk modeling.

Interestingly, ensembles of models can be used to identify where more data is needed [163]. This has been raised as an aid in the reject inference problem. Testing the model in regions where unlabeled data (rejection applications) predominates can highlight where the model is most in need of additional data.

6.2 Imbalanced data sets

Another problem that is more prevalent in credit risk than generic machine learning applications is the extreme imbalance between outcomes [148, 29]. For example, in a commercial loan portfolio, defaults might occur for only 0.1% of accounts. This imbalance means that many machine learning algorithms will be happy to classify the non-defaults while largely ignoring the defaults, resulting in ever poorer performance where it is needed most [270, 158].

Two main approaches have been explored to address the data imbalance problem. Brown and Mues (2012) [55] tested a range of machine learning methods across data sets with varying levels of imbalance in defaults to identify those methods best suited to modeling data sets with different default rates. One notable result was that traditional methods like logistic regression and linear discriminant analysis are robust to the degree of imbalance in the data, so this is largely a machine learning question.

Others have pursued various strategies of modifying the training data to create more balance [145, 194]: over-sample the smaller class, under-sample the larger class, apply weights to the training data, or generate synthetic data to augment the lesser class, as with SMOTE [61].

Overall, the results appear to show that adding to the under-represented class is most effective, with SMOTE being a commonly used approach. SMOTE is basically a random sampling along hyperplanes connecting pairs of points in the smaller class, a linear interpolation. Since this is using a simple model to generate data to feed into a more sophisticated model, it is no surprise that other methods have been proposed.

With any data manipulation approach, the analyst must remember that the underlying probabilities are being modified. The resulting model may be used for scoring, but will require work in order to reintroduce predictions of probabilities. The simplistic approach of introducing a scalar to adjust for the over-sampling is risky, as the sampling will not be perfectly uniform across the feature space, so the probabilities likely will not be accurately recreated locally.

6.3 Overfitting the overfitting tests

Many machine learning methods use performance on an out-of-sample data set to determine when to stop training the model. This approach to preventing overfitting is generally effective, but it carries a caveat. As a rule, the more often a data point is tested the less it can be considered out-of-sample. This was recognized several decades ago. In the case of hypothesis testing, the significance of the result should be adjusted based upon the number of tests conducted, as in the HolmBonferroni method [137].

When repeatedly testing scoring metrics or goodness-of-fit measures on an out-of-sample data set rather than hypothesis testing as above, the author is not aware of an equivalent adjustment, but the same principles apply. Simply stated, a good result with fewer out-of-sample tests is better than a slightly better result after many more tests. This needs to be considered when creating machine learning models and when reviewing work done.

These principles apply to both scoring models tested across hold out samples and time series models tested on an out-of-time sample. Rerunning an out-of-time test repeatedly can result in ”look-ahead bias” where the meta-parameter decisions are based upon the analyst’s judgement of accuracy on data that was supposed to be out-of-sample. This problem is particularly acute when modeling a short time series relative to the cycle being studied.

6.4 Unintended bias

Machine learning has been in production for fraud detection longer than any other application in lending. Conversations with those involved at the beginning suggest that the earliest efforts did not have zip code as an input, but were essentially zip code detection tools. Using or inferring zip codes in loan underwriting or pricing is called redlining and is prohibited [1]. In fraud detection, no such prohibition exists, and one wonders why they didn't just give it zip code to start with.

This story is useful only in the notion that given many other inputs, a sophisticated machine learning algorithm recreated the data that it needed most. That is the greatest danger for using machine learning in credit risk. With linear methods, we generally feel safe in saying that no information on protected class status was given to the model, so the results are unbiased. The same cannot be said of machine learning [90, 223], especially when given alternate inputs. Big Data and sophisticated modeling approaches create significant unobserved risks of inequality and unfair treatment [211].

Consider the case of Amazon's AI-based attempt to find the best job applicants [74]. It was apparently shut down because it was identifying female applicants based upon association with women's groups, and Amazon didn't hire many female engineers, so following the pattern meant that women were rejected. That tale could easily be replayed in credit risk, where a machine learning algorithm infers protected class status using social media data, credit card transactions, branch transactions, etc. One such example showed that the digital footprint of an online borrower was as predictive as FICO score, yet all of those digital footprint data elements probably correlate to protected class status [34]. Excluding protected data is insufficient to assert that the final model's forecasts do not correlate to protected status. Simple linear correlation is the standard for discrimination.

A significant amount of research is being conducted on how to identify and mitigate disparate impacts from machine learning. Current methods can largely be grouped into two approaches. One group is modifying the input data to prevent models from finding biases [155, 298, 292, 90, 124]. The second group modifies the learning algorithm to add constraints that would enforce fairness conditions. [229, 94, 114, 156, 290].

The challenge with both approaches is the need to tag the data with information about protected class status. If we knew the demographic data for each account in the training data, one could trivially run correlations to prove that no bias exists after applying one of the above methods or others. Unfortunately, a linear mindset underlies the regulations. US lenders are not allowed to save data about race, gender, and such for anything except mortgages, so they lack the data necessary to prove that the models are performing fairly. Something will need to change here.

The risk of unintended bias is one of the greatest obstacles to widespread adoption of machine learning models. The solutions will be legal as much as statistical [174].

6.5 Adverse Action Notices

The Equal Credit Opportunity Act (ECOA) [12], as implemented by Regulation B, and the Fair Credit Reporting Act (FCRA), require lenders to provide

Adverse Action Notices when a consumer is denied credit. These notices are specifically intended to be both understandable by the consumer and actionable in the sense that the consumer can make improvements in their financial position in order to qualify in the future. Machine learning has many applications in credit risk, but when it is the primary underwriting tool, it must have good answers for consumers.

Unlike the previous discussion about global interpretability, providing reasons for specific decisions is an inherently local problem. Several methods exist for this, but it remains an important area of research, referred to as the quest for explainable AI (XAI) [81, 203, 181, 112].

The first widely adopted method was Local Interpretable Model-agnostic Explanations (LIME) [227]. LIME samples the space around the decision point to generate a small data set. These points are weighted by distance from the original point, and a small local linear model is built. In fact, the original idea was that any model could work, but the standard implementation is linear. So, it's making a local linear model of a potentially highly nonlinear model overall and using the smaller model to explain the decision just as one would with a linear model.

Shapley values [244] use game theory to allocate significance across input factors. The focus here is local rate of change of the forecast relative to a specific input The approach leverages the original model rather than a locally created simplified model as in LIME. This concept has been enhanced for application to XAI by several authors [251, 187], including integrating elements of LIME [18].

Unfortunately, LIME can be unstable, and both Shapley values and LIME can suffer when a forecast point is at an inflection point in the input variables. In such cases, important dependencies will be missed. Significant research into XAI is currently happening in image processing. Recent work there has developed an approach of explaining an answer relative to a reference image [81]. Work in credit risk has shown that the same reference approach can be effective. Moreover, using a distribution of reference points can provide both explainability and robustness [125].

Certainly more methods will follow. For linear methods, explainability is inherent. Hopefully in the near future, XAI will be an integral part of all machine learning methods.

6.6 Predicting without Understanding

Henley and Hand's work is often cited [129] showing that even small gain in a credit score adds business value. This is taken as proof that prediction is important above all, presumably including explanation. However, those in business know that understanding gained from the modeling process can be used in intangible ways during the underwriting process to add value. One of the greatest risks with machine learning is that analysts can create effective models without learning about the problem they are modeling. For both the analyst and the business, learning matters.

One of our deepest insights from the checking attrition project in Section 5 was the realization that readily available machine learning packages allow analysts to create highly predictive models without understanding what is driving those models. Even when used with default settings, many of these algorithms performed quite well, but is it a good thing to be able to create such models without seizing the opportunity to learn more about the business? In our attempt to understand the relative performance, we actually did learn more about the underlying dynamic between customer and lender, but this was not necessary for the model's success. The importance of explanatory methods for machine learning is not just about educating customers and regulators, but also so that analysts learn about the business.

Some of the understanding gained from a detailed explanation of the model can be more about the data itself. Several machine learning methods are robust to outliers, but if those outliers are data errors, this robustness can lead to a false extrapolation. Robust machine learning models put a greater burden on the analyst to validate the data to assure the model does not just learn an entry error.

Part of the solution also comes back to the disparate impact analysis. We need to recognize that explainable AI is valuable and necessary not just for consumers but also for analysts. Model risk managers need to start asking for a deeper inspection of what makes a machine learning model work, what are the key structures being leveraged, and what can we do with this knowledge to improve the input data and model development process.

Some have gone so far as to say that a bad model that can be understood is better than a good model that cannot be. Let's be clear. That is also a bad answer. The correct answer is to work harder to explain the good models.

6.7 Adapting to Sudden Behavioral Shifts

This article is being written during the depths of the COVID-19 recession. As soon as shelter-in-place orders were issued in the US, we knew that the models would have a problem. All of the algorithms discussed here are data-driven pattern recognition engines. When past patterns are not predictive of future behavior, the models will fail.

Asking for forbearance on a mortgage was no longer a risk indicator, just sensible cash flow management. Job loss and filing for unemployment might be a joint strategy of employer and employee to maximize government benefits until the business reopens. "Strategic delinquency" will probably appear in the research literature in a year or two, exploring the behavioral dynamics leading consumers to go delinquent even when they have money just in order to hoard cash. Sudden increases in deposits, drops in spending, and increases in forbearance reinforce this perspective. Early monitoring of machine learning models in the crisis suggest that exactly these kinds of failures are occurring [127].

In a model driven world, we cannot just wait months or years for new data to arrive to allow us to retrain the models. Model triage becomes an immediate top priority. Human judgment is required to create intuitive models of how behavior is shifting and what adjustments or overlays should be deployed to compensate. In such crises, linear methods or models with separable pieces have the advantage, because then their human masters can understand more easily where model weaknesses might lie, the presumed sensitivities that are no longer true, and what adjustments might compensate for the new situation.

A complex machine learning model could essentially be picking up on the same structures as a linear model, yet lack of interpretability will be a major obstacle to use. The best way to make the machine learning methods robust through such behavior shifts is to make them explainable globally [203] so that the managers can understand enough to compensate.

That is precisely the objective with global interpretability methods in machine learning. Permutation tests [210, 96] randomize either the outcome labels or the input values to measure the significance of the model and specific inputs. Partial dependence plots [100] create a graph of the average forecast versus values for a given input , where the test value is substituted into each input data element. This idea spawned several others including accumulated local effects (ALE) [16], where the change in the model forecast for small changes about a test value is averaged across all corresponding values of the other inputs. Individual Conditional Expectation is essentially a disaggregation of the partial dependence plots, showing the forecasts for each input at the test values rather than simply aggregating to an average value. That disaggregated visualization can provide additional insights into what drives the model [117]. These are a few of the methods available for visualizing the dynamics of machine learning models.

This section could have been called “Run global interpretability tests”. However, in the rush to finish a model and start the next task, analysts usually leave them for “later”. In a crisis, the tools that get used are the ones that are already in place and are understandable. Therefore, measures to provide insight into a model must be part of model development and validation. Such insights are obvious with linear models. Machine learning must adopt such measures as standard practice before models are deployed.

6.8 p-Value arbitrage

In comparing machine learning with traditional methods, the worst reason to choose a winner would be if they were being judged by different standards. For the most part, machine learning models are considered acceptable if they test well out-of-sample, provide a reasonable disparate impact analysis, and do not appear to be biased. For logistic regression, the list is a bit longer.

The most notable difference is the use of p-values to screen for insignificant factors in logistic regression models. Standard practice among model validators and auditors is to make sure that all coefficients in the model are statistically significant according to the p-value, given a reasonably chosen threshold. The p-value is essentially measuring the distance from zero considering the estimation uncertainty. For binned variables where each bin has a corresponding coefficient, the appropriate interpretation is that ”some” of the bins should have statistically significant coefficients. For example, if month-of-year were an input with one coefficient for each month, you would not delete June from the model if its coefficient were zero so long as other months were significantly non-zero. Given that, let's focus on coefficients for continuous variables.

Consider Figure 3. The figure compares coefficients estimated for three different input variables. The first coefficient fails the p-value test, because with it's 95% confidence interval touching zero, it would not meet the 5% p-value threshold, i.e. its coefficient is not provably non-zero. The second coefficient also fails the p-value test for the same reason. However, assuming the inputs are standardized, the first variable is potentially much more important than the second, just equally uncertain. The third coefficient passes and would be allowed into the model, even though it is only weakly useful.

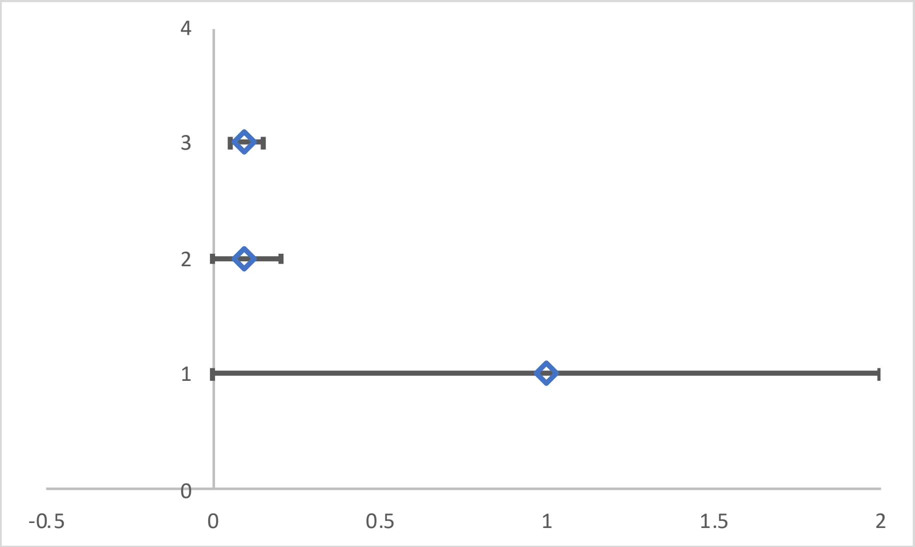


Figure 3: Coefficients with confidence intervals are shown for three hypothetical input variables. The x-axis shows the estimated value and confidence interval. The y-axis just lists three different events.

The American Statistical Society says this is not a correct use or interpretation of p-values [276, 208], and yet it is standard practice in credit risk modeling. By using a p-value criterion for screening variables in regression models but not in machine learning models, we are creating a p-value arbitrage situation. In one of the model comparison studies, we should test the significance of the input factors to see if machine learning models are including factors deleted from the regression models.

It is important that we avoid creating a situation where analysts inadvertently choose machine learning methods over regression methods just because of inconsistent evaluation standards by those in model risk management.

7 Conclusions

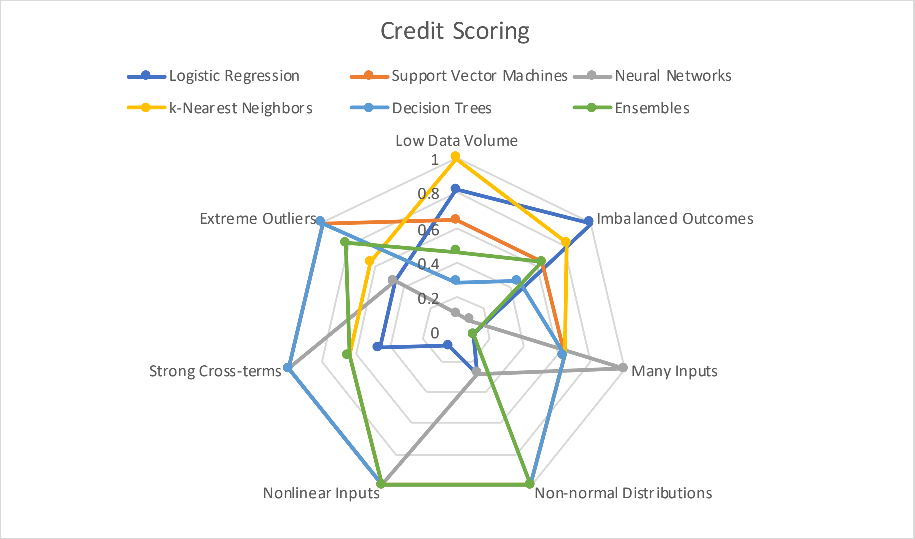


Figure 4: An intuitive comparison of potential strengths and weaknesses of various models for credit scoring. 0 is the weakest and 1 is the strongest under a given challenge.

In reviewing the many machine learning methods available and the equally numerous applications, it becomes clear that declaring a single best method is impossible. Methods have specific strengths and weaknesses that align to different applications. In a specific application, the best method often involves the combination of elements of several methods, both statistical and machine learning.

For academics and researchers, the goal should be to develop a problem space map showing the optimal domains for the methods. Figures 4 and 5 give the author's rough intuitive assessment relative to common modeling challenges for credit scoring and credit risk time series modeling. Each vertex gives a modeling challenge and each modeling technique is rated from 0 (worst) to 1 (best) at addressing that challenge. In the course of creating this survey, the author did not find any method that would be best against all challenges.

If these model rankings could truly be quantified, we could create a recommendation engine that would assess a modeling task and recommend a subset of methods that are likely candidates. Of course, these maps are only guesses, do not include all variants of all methods, and do not consider all modeling challenges. Adding those details would be worthy additions to the literature.

As far as where the field goes from here, several trends are apparent. Research continues into how best to model image-like data sets. We noted that some researchers used image processing techniques to analyze credit card trans-

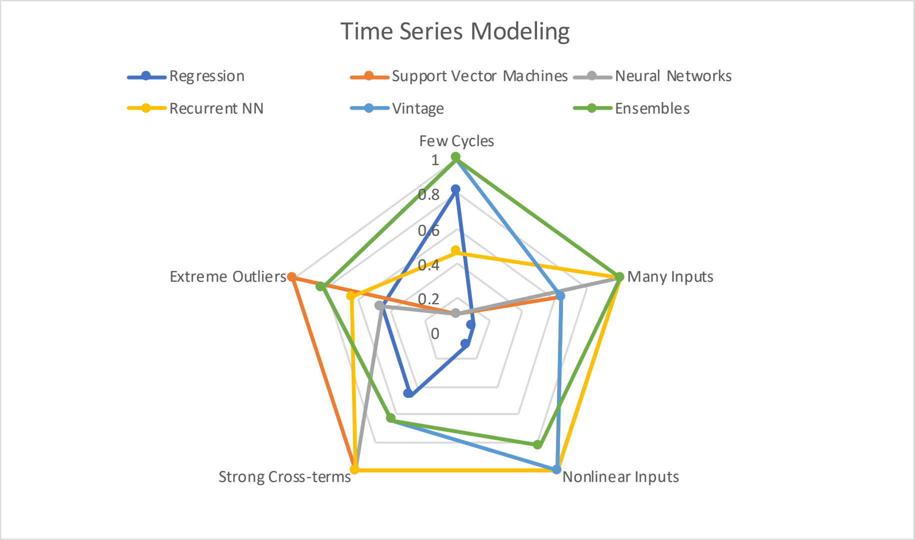


Figure 5: An intuitive comparison of potential strengths and weaknesses of various models for time series modeling in credit risk. 0 is the weakest and 1 is the strongest under a given challenge.

action data, so those could find application in credit risk. Memory-dependent methods such as time series modeling is still seeing rapid development. For all methods reviewed, methods for selecting meta-parameters could be dramatically improved.

Overall, however, we must note that good machine learning methods exist across a range of scoring and time series modeling applications. The greatest advances from here are likely to be more in addressing the challenges of Section 6. All of those challenges involve looking past model accuracy to issues of how to make the models function productively in the real world.

One thing missing from all of these methods is that they produce expectation values. Distributions of possible outcomes are obtained only by running multiple input scenarios, as in stress testing, or looking across many models, as with ensemble distributions. Could we move beyond these forecasts of expectation values to performing calculations upon entire distributions so that the final output of any model is immediately a distribution?

Perhaps, this is where quantum computing [215, 83] could revolutionize credit risk modeling (and many other industries as well). With quantum calculations could we incorporate the full uncertainty of the non-normal distributions of our problems through each step to the final answer? Clearly the greatest failing in using credit risk models is the infrequent generation of confidence intervals and the even rarer use of those in decision making. If all forecasts had accurate measures of uncertainty attached expressing their full non-normal distributions, we would find a great deal of false precision being employed.

Machine learning in some form is clearly the future, but that does not mean it is the present. The challenges listed are not insignificant. The institutional knowledge required for the proper development, validation, monitoring, and overriding of machine learning models currently exists only in pockets. Recognizing those challenges is the best way to speed wider adoption with the fewest possible number of newsworthy blow-ups.

References

[1] Federal Trade Commision, September 2012. 15 U.S.C. S 1681.

[2] New credit score unveiled drawing on bank account data. ABA Banking Journal, October 2018. Newsbytes, Retail and Marketing, Technology.

[3] Hussein Abdou, John Pointon, and Ahmed El-Masry. Neural nets versus conventional techniques in credit scoring in egyptian banking. *Expert Systems with Applications*, 35(3):1275–1292, 2008.

[4] Hussein A Abdou. Genetic programming for credit scoring: The case of egyptian public sector banks. *Expert systems with applications*, 36(9):11402–11417, 2009.

[5] Umar Farouk Ibn Abdulrahman, Joseph Kobina Panford, and James Ben Hayfron-acquah. Fuzzy logic approach to credit scoring for micro finance in Ghana: a case study of kwiqplus money lending. *International Journal of Computer Applications*, 94(8), 2014.

[6] Mohammad Zoynul Abedin, Guotai Chi, Sisira Colombage, and Fahmida-E Moula. Credit default prediction using a support vector machine and a probabilistic neural network. *Journal of Credit Risk*, 14(2).

[7] Joaquín Abellán and Carlos J Mantas. Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, 41(8):3825–3830, 2014.

[8] Héla Hadj Ali, Sébastien Lecocq, and Michael Visser. The impact of gurus: Parker grades and en primeur wine prices. *The Economic Journal*, 118(529):F158–F173, 2008.

[9] K. Ali and M. Pazzani. Error reduction through learning multiple descriptions. *Machine Learning*, 24:172–202, 1996.

[10] Linda Allen, Lin Peng, and Yu Shan. Social networks and credit allocation on fintech lending platforms. Technical report.

[11] Shirley Almon. The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society*, pages 178–196, 1965.

[12] Sarah Ammermann. Adverse action notice requirements under the ECOA and the FCRA. Consumer Compliance Outlook, 2013.

[13] Ioannis Anagnostou, Javier Sánchez Rivero, Sumit Sourabh, and Drona Kandhai. Contagious defaults in a credit portfolio: A bayesian network approach. *Journal of Credit Risk*, 16(1):1–26, 2019.

[14] Raymond Anderson. *Credit Intelligence & Modelling: Many Paths through the Forest*. Independently Published, 2019.

[15] Eliana Angelini, Giacomo di Tollo, and Andrea Roli. A neural network approach for credit risk evaluation. *The quarterly review of economics and finance*, 48(4):733–755, 2008.

[16] Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*, 2016.

[17] Sina Ardabili, Amir Mosavi, and Annamária R Várkonyi-Kóczy. Advances in machine learning modeling reviewing hybrid and ensemble methods. In *International Conference on Global Research and Education*, pages 215–227. Springer, 2019.

[18] Miller Janny Ariza-Garzón, Javier Arroyo, Antonio Caparrini, and Maria-Jesus Segovia-Vargas. Explainability of a machine learning granting scoring model in peer-to-peer lending. *IEEE Access*, 8:64873–64890, 2020.

[19] Amir F Atiya. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on neural networks*, 12(4):929–935, 2001.

[20] Abhishek Awasthi. Clustering algorithms for anti-money laundering using graph theory and social network analysis. 2012.

[21] Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635, 2003.

[22] Bart Baesens, Stijn Viaene, Dirk Van den Poel, Jan Vanthienen, and Guido Dedene. Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, 138(1):191–211, 2002.

[23] A. C. Bahnsen and A. M. Gonzalez. Evolutionary algorithms for selecting the architecture of a MLP neural network: A credit scoring case. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 725–732, 2011.

[24] Michael Bailey, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3):259–80, 2018.

[25] Gah-Yi Ban, Noureddine El Karoui, and Andrew EB Lim. Machine learning and portfolio optimization. *Management Science*, 64(3):1136–1154, 2018.

[26] Anil Bangia, Francis X Diebold, André Kronimus, Christian Schagen, and Til Schuermann. Ratings migration and the business cycle, with application to credit portfolio stress testing. *Journal of Banking & Finance*, 26(2-3):445–474, 2002.

[27] Joao Bastos. Credit scoring with boosted decision trees. Technical Report MPRA Paper No. 8034, CEMAPRE, School of Economics and Management (ISEG), Technical University of Lisbon, April 2007.

[28] João A Bastos. Forecasting bank loans loss-given-default. *Journal of Banking & Finance*, 34(10):2510–2517, 2010.

[29] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[30] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

[31] Anthony Bellotti. Reliable region predictions for automated valuation models. *Annals of Mathematics and Artificial Intelligence*, 81(1-2):71–84, 2017.

[32] Anthony Bellotti, Damiano Brigo, Paolo Gambetti, and Frédéric D Vrins. Forecasting recovery rates on non-performing loans with machine learning. In *Credit Scoring and Credit Control XVI Conference*, Edinburgh, August 2019.

[33] Tony Bellotti and Jonathan Crook. Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, 28(1):171–182, 2012.

[34] Tobias Berg, Valentin Burg, Ana Gombović, and Manju Puri. On the rise of fintechs–credit scoring using digital footprints. Technical report, National Bureau of Economic Research, 2018.

[35] James O Berger. *Statistical decision theory and Bayesian analysis.* Springer Science & Business Media, 2013.

[36] Michael J Best. *Quadratic programming with computer programs.* Chapman and Hall/CRC, 2017.

[37] Elia Biganzoli, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, 17(10):1169–1186, 1998.

[38] Daniel Björkegren and Darrell Grissen. Behavior revealed in mobile phone usage predicts loan repayment. *arXiv preprint arXiv:1712.05840*, 2017.

[39] Financial Stability Board. Artificial intelligence and machine learning in financial services. *Market developments and financial stability implications*, 1, 2017.

[40] Jean-Charles de Borda. Mémoire sur les élections au scrutin: Histoire de lacadémie royale des sciences. *Paris, France*, 12, 1781.

[41] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[42] Heni Boubaker and Nadia Sghaier. Portfolio optimization in the presence of dependent financial returns with long memory: A copula based approach. *Journal of Banking & Finance*, 37(2):361–377, 2013.

[43] Joseph L Breeden. GA-optimal fitness functions. In *International Conference on Evolutionary Programming*, pages 95–102. Springer, 1998.

[44] Joseph L. Breeden. Portfolio optimisation. In *Reinventing Retail Lending Analytics: Forecasting, Stress Testing, Capital and Scoring for a World of Crises*, pages 299–321. Risk Books, London, 2010.

[45] Joseph L. Breeden. *Reinventing Retail Lending Analytics: Forecasting, Stress Testing, Capital and Scoring for a World of Crises, 2nd Impression.* Risk Books, London, 2014.

[46] Joseph L. Breeden. Incorporating lifecycle and environment in loan-level forecasts and stress tests. *European Journal of Operational Research*, 255(2):649 – 658, 2016.

[47] Joseph L. Breeden. Measuring economic cycles in data. *Journal of Risk Model Validation*, 14(1):1–17, March 2020.

[48] Joseph L. Breeden and Jose J. Canals-Cerdá. Consumer risk appetite, the credit cycle, and the housing bubble. *Journal of Credit Risk*, 14(2):1–30, 2018.

[49] Joseph L Breeden, Anthony Bellotti, E Leonova, and A Yablonski. Instabilities using cox ph for forecasting or stress testing loan portfolios. In *Credit Scoring and Credit Control Conference XIV*, 2015.

[50] Joseph L. Breeden and Jonathan Crook. Multihorizon survival models. In *Credit Scoring and Credit Control XVI Conference*, Edinburgh, August 2019.

[51] Joseph L Breeden and Eugenia Leonova. When big data isnt enough: Solving the long-range forecasting problem in supervised learning. In *2019 International Conference on Modeling, Simulation, Optimization and Numerical Techniques (SMONT 2019)*. Atlantis Press, 2019.

[52] Joseph L. Breeden and Norman Packard. A learning algorithm for optimal representation of experimental data. *International Journal of Bifurcation and Chaos*, 4(2):311–326, April 1994.

[53] Leo Breiman. Bagging predictors. *Machine Learning*, pages 123–140, 1996.

[54] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[55] Iain Brown and Christophe Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3):3446–3453, 2012.

[56] Wray L Buntine and Andreas S Weigend. Bayesian back-propagation. *Complex systems*, 5(6):603–643, 1991.

[57] Paul Buta. Mining for financial knowledge with cbr. *Ai Expert*, 9(2):34–41, 1994.

[58] C Casas. Reducing portfolio volatility with artificial neural networks. In *Artificial Intelligence and Applications: IASTED International Conference Proceedings*, 2005.

[59] Stephan K Chalup and Andreas Mitschele. Kernel methods in finance. In *Handbook on information technology in finance*, pages 655–687. Springer, 2008.

[60] Yung-Chia Chang, Kuei-Hu Chang, and Guan-Jhih Wu. Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, 73:914–920, 2018.

[61] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

[62] Fei-Long Chen and Feng-Chia Li. Combination of feature selection approaches with svm in credit scoring. *Expert systems with applications*, 37(7):4902–4909, 2010.

[63] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[64] Yen-Liang Chen and Lucas Tzu-Hsuan Hung. Using decision trees to summarize associative classification rules. *Expert Systems with Applications*, 36(2):2338–2351, 2009.

[65] Gong Cheng, Peicheng Zhou, and Junwei Han. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016.

[66] Robert T. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5:559–583, 1989.

[67] Pamela K Coats and L Franklin Fant. Recognizing financial distress patterns using a neural network tool. *Financial management*, pages 142–155, 1993.

[68] John Y. Coffman and Gary G. Chandler. Applications of performance scoring to accounts receivable management in consumer credit. Technical report, Credit Research center, Krannert Graduate School of Management, Purdue University, 1983.

[69] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML 08, page 160167, New York, NY, USA, 2008. Association for Computing Machinery.

[70] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[71] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[72] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[73] Belur V Dasarathy and Belur V Sheela. A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, 67(5):708–713, 1979.

[74] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Technology News*, 2018.

[75] Rober Hunter DAVIS, DB Edelman, and AJ Gammerman. Machine-learning algorithms for credit-card applications. *IMA Journal of Management Mathematics*, 4(1):43–51, 1992.

[76] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[77] Vijay S Desai, Jonathan N Crook, and George A Overstreet Jr. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, 95(1):24–37, 1996.

[78] Sander Dieleman, Kyle W Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly notices of the royal astronomical society*, 450(2):1441–1459, 2015.

[79] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[80] Viani Biatat Djeundje and Jonathan Crook. Dynamic survival models with varying coefficients for credit risks. *European Journal of Operational Research*, 275(1):319–333, 2019.

[81] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, pages 13567–13578, 2019.

[82] Bradley Efron. The two-way proportional hazards model. *Journal of the Royal Statistical Society B*, 64:899 − 909, 2002.

[83] Daniel J Egger, Ricardo Gacía Gutiérrez, Jordi Cahué Mestre, and Stefan Woerner. Credit risk analysis using quantum computers. *arXiv preprint arXiv:1907.03044*, 2019.

[84] Edwin J Elton, Martin J Gruber, Stephen J Brown, and William N Goetzmann. *Modern portfolio theory and investment analysis.* John Wiley & Sons, 2009.

[85] Walter Enders. *Applied Econometric Time Series, Fourth Edition.* John Wiley & Sons, 2014.

[86] Muhammad Faishal Ibrahim, Fook Jam Cheng, and Kheng How Eng. Automated valuation model: an application to the public housing resale market in singapore. *Property Management*, 23(5):357–373, 2005.

[87] David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

[88] Simone Farinelli, Manuel Ferreira, Damiano Rossello, Markus Thoeny, and Luisa Tibiletti. Beyond sharpe ratio: Optimal asset allocation using different performance ratios. *Journal of Banking & Finance*, 32(10):2057–2063, 2008.

[89] FDIC. Credit card activities manual. https://www.fdic.gov/regulations/examinations/credit_card/ch12.html, 2007.

[90] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[91] Holmes Finch and Mercedes K Schneider. Classification accuracy of neural networks vs. discriminant analysis, logistic regression, and classification and regression trees. *Methodology*, 3(2):47–57, 2007.

[92] Jason P. Fine and Robert J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

[93] Steven Finlay. Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, 210(2):368–378, 2011.

[94] Benjamin Fish, Jeremy Kun, and Adám D Lelkes. Fair boosting: a case study. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*. Citeseer, 2015.

[95] T.R. Fleming and D.P. Harrington. *Counting Processes and Survival Analysis*. Wiley, New York, 1991.

[96] Eibe Frank and Ian H Witten. Using a permutation test for attribute selection in decision trees. 1998.

[97] Seth Freedman and Ginger Zhe Jin. The information value of online social networks: lessons from peer-to-peer lending. *International Journal of Industrial Organization*, 51:185–222, 2017.

[98] Christoph Frei and Marcus Wunsch. Moment estimators for autocorrelated time series and their application to default correlations. *Journal of Credit Risk*, 14(1).

[99] Yoav Freund and Robert E Schapire. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

[100] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[101] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[102] Jerome H Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.

[103] Jerome H Friedman and John W Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, 100(9):881–890, 1974.

[104] Holger Frohlich, Olivier Chapelle, and Bernhard Scholkopf. Feature selection for support vector machines by means of genetic algorithm. In *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*, pages 142–148. IEEE, 2003.

[105] Halina Frydman, Edward I Altman, and Duen-Li Kao. Introducing recursive partitioning for financial classification: the case of financial distress. *The Journal of Finance*, 40(1):269–291, 1985.

[106] Wenjiang Fu. *A Practical Guide to Age-Period-Cohort Analysis: The Identification Problem and Beyond*. Chapman and Hall/CRC, 2018.

[107] Joshua Gallin, Raven Molloy, Eric Reed Nielsen, Paul A Smith, and Kamila Sommer. Measuring aggregate housing wealth: New insights from an automated valuation model. 2018.

[108] Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

[109] Nazeeh Ghatasheh. Business analytics using random forest trees for credit risk prediction: A comparison study. *International Journal of Advanced Science and Technology*, 72(2014):19–30, 2014.

[110] Sushmito Ghosh and Douglas L Reilly. Credit card fraud detection with a neural-network. In *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on*, volume 3, pages 621–630. IEEE, 1994.

[111] Bruce P Gibbs. *Advanced Kalman filtering, least-squares and modeling: a practical handbook*. John Wiley & Sons, 2011.

[112] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[113] Norval D. Glenn. *Cohort Analysis, 2nd Edition*. Sage, London, 2005.

[114] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

[115] David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, 1989.

[116] DE Goldberg. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, Reading, MA, 1989.

[117] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[118] Dale L Goodhue, William Lewis, and Ronald L Thompson. A dangerous blind spot in is research: False positives due to multicollinearity combined with measurement error. In *AMCIS*, 2011.

[119] Asogbon Mojisola Grace and Samuel Oluwarotimi Williams. Comparative analysis of neural network and fuzzy logic techniques in credit risk evaluation. *International Journal of Intelligent Information Technologies (IJIIT)*, 12(1):47–62, 2016.

[120] Alex Graves, Santiago Fernndez, Faustino Gomez, and Jrgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks. volume 2006, pages 369–376, 01 2006.

[121] Alastair R Hall. *Generalized method of moments*. Oxford university press, 2005.

[122] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.

[123] L. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:993–1001, 1990.

[124] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[125] Ian Hardy. Robust explainability in AI models, 2020.

[126] Frank Harrell. Road map for choosing between statistical modeling and machine learning. https://www.fharrell.com/post/stat-ml, 2019.

[127] Will Douglas Heaven. Our weird behavior during the pandemic is messing with AI models. MIT Technology Review, May 2020.

[128] Robert Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural networks for perception*, pages 65–93. Elsevier, 1992.

[129] WE Henley and D Hand. Construction of a k-nearest-neighbour credit-scoring system. *IMA Journal of Management Mathematics*, 8(4):305–321, 1997.

[130] WE Henley and David J Hand. Ak-nearest-neighbour classifier for assessing consumer credit risk. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45(1):77–95, 1996.

[131] Joseph M. Hilbe. *Logistic Regression Models.* Taylor & Francis, 5 2009.

[132] T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[133] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[134] F Hoffmann, Bart Baesens, Christophe Mues, Tony Van Gestel, and Jan Vanthienen. Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. *European journal of operational research*, 177(1):540–555, 2007.

[135] T R Holford. The estimation of age, period and cohort effects for vital rates. *Biometrics*, 39(2):311–324, 1983.

[136] Theodore Holford. *Encyclopedia of Statistics in Behavioral Science*, chapter Age-Period-Cohort Analysis. Wiley, 2005.

[137] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[138] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

[139] Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. Bagging survival trees. *Statistics in medicine*, 23(1):77–91, 2004.

[140] Cheng Hsiao. *Analysis of Panel Data.* Cambridge University Press, 2014.

[141] Kuo-Wei Hsu. A theoretical analysis of why hybrid ensembles work. *Computational intelligence and neuroscience*, 2017, 2017.

[142] T. Hsu, S. Liou, Y. Wang, Y. Huang, and Che-Lin. Enhanced recurrent neural network for combining static and dynamic features for credit card default prediction. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1572–1576, 2019.

[143] Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4):847–856, 2007.

[144] Jih-Jeng Huang, Gwo-Hshiung Tzeng, and Chorng-Shyong Ong. Two-stage genetic programming (2sgp) for the credit scoring model. *Applied Mathematics and Computation*, 174(2):1039–1053, 2006.

[145] Yueh-Min Huang, Chun-Min Hung, and Hewijin Christine Jiau. Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications*, 7(4):720–747, 2006.

[146] Peter J Huber. Projection pursuit. *The annals of Statistics*, pages 435–475, 1985.

[147] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

[148] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[149] Herbert L Jensen. Using neural networks for credit scoring. *Managerial finance*, 1992.

[150] I.T. Jolliffe. *Principal Component Analysis, second edition.* Springer, 2002.

[151] Kenneth A. De Jong. *Evolutionary Computation: A Unified Approach.* The MIT Press, 2016.

[152] Jr., David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis: Regression Modeling of Time to Event Data, Second Edition.* Wiley Series in Probability and Statistics, New York, 2008.

[153] Frank E. Harrell Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* Springer, 2015.

[154] Iakovos Kakouris and Berç Rustem. Robust portfolio optimization with copulas. *European Journal of Operational Research*, 235(1):28–37, 2014.

[155] Faisal Kamiran and Toon Calders. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, pages 1–6, 2010.

[156] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[157] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[158] Kenneth Kennedy, Brian Mac Namee, and Sarah Jane Delany. Learning without default: A study of one-class classification and the low-default portfolio problem. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pages 174–187. Springer, 2009.

[159] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[160] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[161] John R Koza et al. *Genetic programming.* MIT press Cambridge, 1994.

[162] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS12, page 10971105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[163] A. Krogh and J. Vedelsby. *Neural network ensembles, cross validation, and active learning*, pages 231–238. MIT Press, Cambridge, 1995.

[164] Jochen Kruppa, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13):5125–5131, 2013.

[165] Dimitris Kugiumtzis. State space reconstruction parameters in the analysis of chaotic time seriesthe role of the time window length. *Physica D: Nonlinear Phenomena*, 95(1):13–28, 1996.

[166] Ludmila I Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Transactions on pattern analysis and machine intelligence*, 24(2):281–286, 2002.

[167] Hvard Kvamme, Nikolai Sellereite, Kjersti Aas, and Steffen Sjursen. Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, 102:207 – 217, 2018.

[168] Prentice R. L. the analysis of failure times in the presence of competing risks. *Biometrics*, 34:541, 1978.

[169] Kin Keung Lai, Lean Yu, Shouyang Wang, and Chengxiong Zhou. A double-stage genetic optimization algorithm for portfolio selection. In *International Conference on Neural Information Processing*, pages 928–937. Springer, 2006.

[170] Kin Keung Lai, Lean Yu, Shouyang Wang, and Ligang Zhou. Neural network metalearning for credit scoring. In *International Conference on Intelligent Computing*, pages 403–408. Springer, 2006.

[171] William B Langdon, SJ Barrett, and Bernard F Buxton. Combining decision trees and neural networks for drug discovery. In *European Conference on Genetic Programming*, pages 60–70. Springer, 2002.

[172] Steve Lawrence, C Lee Giles, and Ah Chung Tsoi. Lessons in neural network training: Overfitting may be harder than expected. In *AAAI/IAAI*, pages 540–545. Citeseer, 1997.

[173] Tae-Hwy Lee and Yang Yang. Bagging binary and quantile predictors for time series. *Journal of econometrics*, 135(1-2):465–497, 2006.

[174] David Lehr and Paul Ohm. Playing with the data: what legal scholars should learn about machine learning. *UCDL Rev.*, 51:653, 2017.

[175] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124 – 136, 2015.

[176] Edward M. Lewis. *An Introduction to Credit Scoring.* The Athena Press, San Rafael, California, 1994.

[177] WK Li and Al McLeod. Distribution of the residual autocorrelations in multivariate arma time series models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(2):231–239, 1981.

[178] Xiangfeng Li, Shenghua Liu, Zifeng Li, Xiaotian Han, Chuan Shi, Bryan Hooi, He Huang, and Xueqi Cheng. Flowscope: Spotting money laundering based on graphs.

[179] Guohua Liang, Xingquan Zhu, and Chengqi Zhang. An empirical study of bagging predictors for different learning algorithms. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[180] Charles X Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *Kdd*, volume 98, pages 73–79, 1998.

[181] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[182] Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[183] Laura Liu, Hyungsik Roger Moon, and Frank Schorfheide. Forecasting with dynamic panel data models. Working Paper 25102, National Bureau of Economic Research, September 2018.

[184] Yan-Chun Liu, Tie Wang, Li-Qun Gao, Ping Ren, and Bao-Zheng Liu. Fuzzy portfolio optimization model based on worst-case var. In *2005 International Conference on Machine Learning and Cybernetics*, volume 6, pages 3512–3516. IEEE, 2005.

[185] Christian Lohmann and Thorsten Ohliger. Nonlinear relationships in a logistic model of default for a high-default installment portfolio. *Journal of Credit Risk*, 14(1).

[186] André Lucas and Pieter Klaassen. Discrete versus continuous state switching models for portfolio credit risk. *Journal of Banking & Finance*, 30(1):23–35, 2006.

[187] Scott Lundberg and Su-In Lee. An unexpected unity among methods for interpreting model predictions. *arXiv preprint arXiv:1611.07478*, 2016.

[188] Paul Makowski. Credit scoring branches out. *Credit World*, 75(1):30–37, 1985.

[189] Milad Malekipirbazari and Vural Aksakalli. Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10):4621–4631, 2015.

[190] Rashmi Malhotra and Davinder K Malhotra. Evaluating consumer loans using neural networks. *Omega*, 31(2):83–96, 2003.

[191] Yannis Marinakis, Magdalene Marinaki, Michael Doumpos, Nikolaos Matsatsinis, and Constantin Zopounidis. Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment. *Journal of Global Optimization*, 42(2):279–293, 2008.

[192] Harry M Markowitz and G Peter Todd. *Mean-variance analysis in portfolio choice and capital markets*, volume 66. John Wiley & Sons, 2000.

[193] Ana I Marqués, Vicente García, and José Salvador Sánchez. Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, 39(11):10244–10250, 2012.

[194] Ana Isabel Marqués, Vicente García, and José Salvador Sánchez. On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*, 64(7):1060–1070, 2013.

[195] W.M. Mason and S. Fienberg. *Cohort Analysis in Social Research: Beyond the Identification Problem.* Springer, 1985.

[196] László Mátyás, Christian Gourieroux, Peter CB Phillips, et al. *Generalized method of moments estimation*, volume 5. Cambridge University Press, 1999.

[197] Jose Alberto Mauricio. Exact maximum likelihood estimation of stationary vector arma models. *Journal of the American Statistical Association*, 90(429):282–291, 1995.

[198] Thomas E Mckee. Developing a bankruptcy prediction model via rough sets theory. *Intelligent Systems in Accounting, Finance & Management*, 9(3):159–173, 2000.

[199] Aaron Mengelkamp, Sebastian Hobert, and Matthias Schumann. Corporate credit risk analysis utilizing textual user generated content-a twitter based feasibility study. In *PACIS*, page 236, 2015.

[200] Jae H Min and Young-Chan Lee. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert systems with applications*, 28(4):603–614, 2005.

[201] Jun-Ki Min and Sung-Bae Cho. Activity recognition based on wearable sensors using selection/fusion hybrid ensemble. In *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1319–1324. IEEE, 2011.

[202] Asunción Mochón, David Quintana, Yago Sáez, and Pedro Isasi. Soft computing techniques applied to finance. *Applied Intelligence*, 29(2):111–115, 2008.

[203] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2019.

[204] Loris Nanni and Alessandra Lumini. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications*, 36(2):3028–3033, 2009.

[205] J Neeter, W Wasserman, and MH Kutner. Applied linear statistics. *Irwin, Homeland, IL*, 1985.

[206] Pamela Nickell, William Perraudin, and Simone Varotto. Stability of rating transitions. *Journal of Banking & Finance*, 24(1-2):203–227, 2000.

[207] J. R. Norris. *Markov Chains*. Cambridge University Press, 1998.

[208] Regina Nuzzo. Scientific method: statistical errors. *Nature News*, 506(7487):150, 2014.

[209] Marcus D Odom and Ramesh Sharda. A neural network model for bankruptcy prediction. In *1990 IJCNN International Joint Conference on neural networks*, pages 163–168. IEEE, 1990.

[210] Markus Ojala and Gemma C Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(Jun):1833–1863, 2010.

[211] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books, 2016.

[212] Chorng-Shyong Ong, Jih-Jeng Huang, and Gwo-Hshiung Tzeng. Building credit scoring models using genetic programming. *Expert Systems with Applications*, 29(1):41–47, 2005.

[213] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[214] Stjepan Oreski, Dijana Oreski, and Goran Oreski. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert systems with applications*, 39(16):12605–12617, 2012.

[215] Roman Orus, Samuel Mugel, and Enrique Lizaso. Quantum computing for finance: overview and prospects. *Reviews in Physics*, page 100028, 2019.

[216] Norman H Packard, James P Crutchfield, J Doyne Farmer, and Robert S Shaw. Geometry from a time series. *Physical review letters*, 45(9):712, 1980.

[217] Ebberth L Paula, Marcelo Ladeira, Rommel N Carvalho, and Thiago Marzagão. Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 954–960. IEEE, 2016.

[218] Yudi Pawitan. *In all likelihood: statistical modelling and inference using likelihood.* Oxford University Press, 2001.

[219] Zdzisław Pawlak. Rough sets. *International journal of computer & information sciences*, 11(5):341–356, 1982.

[220] Selwyn Piramuthu. Financial credit-risk evaluation with neural and neurofuzzy systems. *European Journal of Operational Research*, 112(2):310–321, 1999.

[221] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[222] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[223] Anya ER Prince and Daniel Schwarcz. Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev.*, 105:1257, 2019.

[224] Foster Provost and Pedro Domingos. Tree induction for probability-based ranking. *Machine learning*, 52(3):199–215, 2003.

[225] Richard E Quandt. A new approach to estimating switching regressions. *Journal of the American statistical association*, 67(338):306–310, 1972.

[226] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[227] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[228] Leslie Richeson, Raymond A Zimmermann, and Kevin Gregory Barnett. Predicting consumer credit performance: Can neural networks outperform traditional statistical methods? *International Journal of Applied Expert Systems*, 2(2):116–130, 1994.

[229] Goce Ristanoski, Wei Liu, and James Bailey. Discrimination aware classification for imbalanced datasets. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 1529–1532, 2013.

[230] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630, 2006.

[231] Philip S Rosenberg. Hazard function estimation using b-splines. *Biometrics*, pages 874–887, 1995.

[232] Jose San Pedro, Davide Proserpio, and Nuria Oliver. Mobiscore: towards universal credit scoring from mobile phone data. In *international conference on user modeling, adaptation, and personalization*, pages 195–207. Springer, 2015.

[233] Natasa Sarlija, Mirta Bensic, and Marijana Zekic-Susac. A neural network classification of credit applicants in consumer credit scoring. In *Artificial Intelligence and Applications*, pages 205–210, 2006.

[234] Tim Sauer, James A Yorke, and Martin Casdagli. Embedology. *Journal of statistical Physics*, 65(3-4):579–616, 1991.

[235] Robert E Schapire. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer, 2003.

[236] Robert E Schapire and Yoav Freund. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

[237] Klaus B Schebesch and Ralf Stecking. Support vector machines for classifying and describing credit applicants: detecting typical and critical regions. *Journal of the Operational Research Society*, 56(9):1082–1088, 2005.

[238] Volker Schmid and Leonhard Held. Bayesian age-period-cohort modeling and prediction - bamp. *Journal of Statistical Software, Articles*, 21(8):1–15, 2007.

[239] Til Schuermann. What do we know about loss given default? 2004.

[240] Eduardo S. Schwartz and Walter N. Torous. Prepayment and the valuation of mortgage-backed securities. *The Journal of Finance*, 44(2):375–392, 1989.

[241] OG Selfridge. Pandemonium: a paradigm for learning, mechanism of thought processes: Proceedings of a symposium held at the national physical laboratory, 1958.

[242] C. Seshadhri, Aneesh Sharma, Andrew Stolman, and Ashish Goel. The impossibility of low-rank representations for triangle-rich complex networks. *Proceedings of the National Academy of Sciences*, 117(11):5631–5637, 2020.

[243] John Shapcott. Index tracking: genetic algorithms for investment portfolio selection. *Edinburgh Parallel Computing Centre*, 1992.

[244] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[245] William F Sharpe. The sharpe ratio. *Journal of portfolio management*, 21(1):49–58, 1994.

[246] Han Sheng Sun and Zi Jin. Estimating credit risk parameters using ensemble learning methods: An empirical study on loss given default. *Journal of Credit Risk, Forthcoming*, 2016.

[247] Kyung-shik Shin and Ingoo Han. A case-based approach using inductive indexing for corporate bond rating. *Decision Support Systems*, 32(1):41–52, 2001.

[248] Kyung-Shik Shin and Yong-Joo Lee. A genetic algorithm application in bankruptcy prediction modeling. *Expert systems with applications*, 23(3):321–328, 2002.

[249] Nitish Srivastava. Improving neural networks with dropout. *University of Toronto*, 182(566):7, 2013.

[250] Maria Stepanova and Lyn Thomas. Survival analysis methods for personal loan data. *Operations Research*, 50(2):277–289, 2002.

[251] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

[252] Han Sheng Sun and Zi Jin. Estimating credit risk parameters using ensemble learning methods: an empirical study on loss given default. *Journal of Credit Risk*, 12(3):43–69, 2016.

[253] Kar Yan Tam and Melody Y Kiang. Managerial applications of neural networks: the case of bank failure predictions. *Management science*, 38(7):926–947, 1992.

[254] Jun Tang and Jian Yin. Developing an intelligent data discriminating system of anti-money laundering based on svm. In *2005 International conference on machine learning and cybernetics*, volume 6, pages 3453–3457. IEEE, 2005.

[255] Igor V Tetko, David J Livingstone, and Alexander I Luik. Neural network studies. 1. comparison of overfitting and overtraining. *Journal of chemical information and computer sciences*, 35(5):826–833, 1995.

[256] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2000.

[257] Lyn C. Thomas, Jonathan N. Crook, and David B. Edelman. *Credit Scoring and Its Applications, Second Edition*. Society for Industrial and Applied Mathematics, Philadelphia, 2017.

[258] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[259] Ljupčo Todorovski and Sašo Džeroski. Combining classifiers with meta decision trees. *Machine learning*, 50(3):223–249, 2003.

[260] Bogdan Trawiński, Zbigniew Telec, Jacek Krasnoborski, Mateusz Piwowarczyk, Michał Talaga, Tedeusz Lasota, and Edward Sawiłow. Comparison of expert algorithms with machine learning models for real estate appraisal. In *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 51–54. IEEE, 2017.

[261] Bhekisipho Twala. Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, 37(4):3326–3336, 2010.

[262] Parastoo Rafiee Vahid and Abbas Ahmadi. Modeling corporate customers credit risk considering the ensemble approaches in multiclass classification: evidence from iranian corporate credits. *Journal of Credit Risk*, 12(3):71–95, 2016.

[263] Merijn Van Erp, Louis Vuurpijl, and Lambert Schomaker. An overview and comparison of voting methods for pattern recognition. In *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 195–200. IEEE, 2002.

[264] Tony Van Gestel, Johan AK Suykens, Bart Baesens, Stijn Viaene, Jan Vanthienen, Guido Dedene, Bart De Moor, and Joos Vandewalle. Benchmarking least squares support vector machine classifiers. *Machine learning*, 54(1):5–32, 2004.

[265] Robert J Vanderbei. *Linear Programming: Foundations and Extensions.* International Series in Operations Research & Management Science, 2013.

[266] Valdimir N Vapnik. T he nature of statistical l earning t heory. *New York: Springer2V erlag*, 1995.

[267] P-CG Vassiliou. Fuzzy semi-markov migration process in credit risk. *Fuzzy Sets and Systems*, 223:39–58, 2013.

[268] Alfredo Vellido, Paulo JG Lisboa, and J Vaughan. Neural networks in business: a survey of applications (1992–1998). *Expert Systems with applications*, 17(1):51–70, 1999.

[269] Antanas Verikas, Zivile Kalsyte, Marija Bacauskiene, and Adas Gelzinis. Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. *Soft Computing*, 14(9):995–1010, 2010.

[270] Veronica Vinciotti and David J Hand. Scorecard construction with unbalanced class sizes. 2003.

[271] Gang Wang, Jinxing Hao, Jian Ma, and Hongbing Jiang. A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38:223–230, 2011.

[272] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

[273] Su-Nan Wang and Jian-Gang Yang. A money laundering risk evaluation method based on decision tree. In *2007 International Conference on Machine Learning and Cybernetics*, volume 1, pages 283–286. IEEE, 2007.

[274] Xiaofei Wang. Machine learning-driven credit risk modelling using smartphone metadata. In *Proceedings of the 2019 Credit Scoring and Credit Control Conference, Edinburgh, UK*, volume 31.

[275] Yongqiao Wang, Shouyang Wang, and Kin Keung Lai. A new fuzzy support vector machine to evaluate credit risk. *IEEE Transactions on Fuzzy Systems*, 13(6):820–831, 2005.

[276] Ronald L Wasserstein and Nicole A Lazar. The asa statement on p-values: context, process, and purpose, 2016.

[277] William W. S. Wei. *Time Series Analysis: Univariate and Multivariate Models, 2nd Edition.* Addison-Wesley Pub Co, 1990.

[278] Yanhao Wei, Pinar Yildirim, Christophe Van den Bulte, and Chrysanthos Dellarocas. Credit scoring with social network data. *Marketing Science*, 35(2):234–258, 2016.

[279] D West. Neural network credit scoring models. *Comput Opns Res*, 27:1131, 2000.

[280] David West, Scott Dellana, and Jingxia Qian. Neural network ensemble strategies for financial decision applications. *Computers & operations research*, 32(10):2543–2559, 2005.

[281] T. Windeatt and G. Ardeshir. Decision tree simplification for classifier ensembles. *International Journal of Pattern Recognition*, 18(5):749–776, 2004.

[282] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[283] Jeffrey Wooldridge. *Econometric Analysis of Cross Section and Panel Data, Second Edition.* The MIT Press, 2010.

[284] Wenbing Xiao, Qian Zhao, and Qi Fei. A comparative study of data mining methods in consumer loans credit scoring management. *Journal of Systems Science and Systems Engineering*, 15(4):419–435, 2006.

[285] Xiujuan Xu, Chunguang Zhou, and Zhe Wang. Credit scoring algorithm based on link analysis ranking with support vector machine. *Expert Systems with Applications*, 36(2):2625–2632, 2009.

[286] Y. Yang and K.C. Land. *Age-Period-Cohort Analysis.* Taylor and Francis, Boca Raton, 2014.

[287] Yingxu Yang. Adaptive credit scoring with kernel learning methods. *European Journal of Operational Research*, 183(3):1521–1536, 2007.

[288] Mumine B Yobas, Jonathan N Crook, and Peter Ross. Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics*, 11(2):111–125, 2000.

[289] Lean Yu, Shouyang Wang, and Kin Keung Lai. An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring. *European journal of operational research*, 195(3):942–959, 2009.

[290] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

[291] Antonella Zanobetti, MP Wand, J Schwartz, and LM Ryan. Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, 1(3):279–292, 2000.

[292] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[293] Defu Zhang, Xiyue Zhou, Stephen CH Leung, and Jiemin Zheng. Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, 37(12):7838–7843, 2010.

[294] Wenyu Zhang, Hongliang He, and Shuai Zhang. A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring. *Expert Systems with Applications*, 121:221–232, 2019.

[295] Xiaofei Zhou, Wenhan Jiang, Yong Shi, and Yingjie Tian. Credit risk evaluation with kernel-based affine subspace nearest points learning method. *Expert Systems with Applications*, 38(4):4272–4279, 2011.

[296] XiYue Zhou, DeFu Zhang, and Yi Jiang. A new credit scoring method based on rough sets and decision tree. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 1081–1089. Springer, 2008.

[297] Xun Zhou, Sicong Cheng, Meng Zhu, Chengkun Guo, Sida Zhou, Peng Xu, Zhenghua Xue, and Weishi Zhang. A state of the art survey of data mining-based fraud detection and credit scoring. In *MATEC Web of Conferences*, volume 189, page 03002. EDP Sciences, 2018.

[298] Indre Zliobaite, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14*, pages 992–1001, 2011.

Dear Organizers

We are attaching two pieces.

First is the Law Review (which is a completed manuscript).

We are presenting this in tandem with some new evidence to bring the punchline of the legal treatise to financial economists. In Particular, we gathered credit scoring data and are applying the algorithmic accountability arguments from the law review in a finance manuscript.

The results are in the powerpoint slides, which are also attached.

We have presented the tandem at Berkeley, Chicago-Booth, and UnivWashington. (In other words, we are well developed in the application paper, but are touring with the combo offering because it is important and rich enough for economists, but not too long.)

ALGORITHMIC ACCOUNTABILITY:
A LEGAL AND ECONOMIC FRAMEWORK

Robert P. Bartlett, III[*]
Adair Morse[‡]
Nancy Wallace[†]
Richard Stanton[°]

Abstract

Despite the potential for machine learning and artificial intelligence to reduce face-to-face bias in decision-making, a growing chorus of scholars and policymakers have recently voiced concerns that if left unchecked, algorithmic decision-making can also lead to unintentional discrimination against members of historically marginalized groups. These concerns are being expressed through Congressional subpoenas, regulatory investigations, and an increasing number of algorithmic accountability bills pending in both state legislatures and Congress. To date, however, prominent efforts to define policies whereby an algorithm can be considered accountable have tended to focus on output-oriented policies and interventions that either may facilitate illegitimate discrimination or involve fairness corrections unlikely to be legally valid.

We provide a workable definition of algorithmic accountability that is rooted in the caselaw addressing statistical discrimination in the context of Title VII of the Civil Rights Act of 1964. Using instruction from the burden-shifting framework, codified to implement Title VII, we formulate a simple statistical test to apply to the design and review of the inputs used in any algorithmic decision-making processes. Application of the test, which we label the *input accountability test*, constitutes a legally viable, deployable tool that can prevent an algorithmic model from systematically penalizing members of protected groups who are otherwise qualified in a target characteristic of interest.

[*] I. Michael Heyman Professor of Law & Faculty Co-Director of the Berkeley Center for Law and Business - UC Berkeley School of Law.

[‡] Associate Professor & Soloman P. Lee Chair in Business Ethics – UC Berkeley Haas School of Business.

[†] Professor & Lisle and Roslyn Payne Chair in Real Estate Capital Markets, Co-Chair, Fisher Center for Real Estate and Urban Economics – UC Berkeley Haas School of Business.

[°] Professor & Kingsford Capital Management Chair in Business Economics – UC Berkeley Haas School of Business.

ALGORITHMIC ACCOUNTABILITY:
A LEGAL AND ECONOMIC FRAMEWORK

TABLE OF CONTENTS

I. INTRODUCTION

In August 2019, Apple Inc. debuted its much-anticipated Apple Card, a no fee, cash-rewards credit card "designed to help customers lead a healthier financial life."[1] Within weeks of its release, Twitter was abuzz with headlines that the card's credit approval algorithm was systematically biased against

[1] Press Release, Apple Inc., *Introducing Apple Card, A New Kind of Credit Card Created by Apple* (March 25,2019), https://www.apple.com/newsroom/2019/03/introducing-apple-card-a-new-kind-of-credit-card-created-by-apple/.

women.[2] Even Apple co-founder Steve Wozniak weighed in, tweeting that the card gave him a credit limit that was ten times higher than what it gave his wife, despite the couple sharing all their assets.[3] In the days that followed, Goldman Sachs—Apple's partner in designing the Apple Card—steadfastly defended the algorithm, insisting that "we have not and will not make decisions based on factors on gender."[4] Yet doubts persisted. By November, the New York State Department of Financial Services had announced an investigation into the card's credit approval practices.[5]

Around that same time, buzz spread across the media about another algorithm, that of health insurer UnitedHealth.[6] The algorithm was used to inform hospitals about patients' level of sickness so that hospitals could more effectively allocate resources to the sickest patients. However, an article appearing in *Science* showed that because the company used cost of care as the metric for gauging sickness and because African-American patients historically incurred lower costs for the same illnesses and level of illness, the algorithm caused them to receive substandard care as compared to white patients.[7]

Despite the potential for algorithmic decision-making to eliminate face-to-face biases, these episodes provide vivid illustrations of the widespread concern that algorithms may nevertheless engage in objectionable discrimination.[8] Indeed, a host of regulatory reforms have emerged to contend with this challenge. For example, New York City has enacted an algorithm accountability law, which creates a task force to recommend procedures for determining whether automated decisions by city agencies disproportionately impact protected groups.[9] Likewise, the Washington State House of Representatives introduced an algorithm accountability bill, which would require the state's chief information officer assess whether any automated decision system used by a state agency "has a known bias, or is untested for

[2] *See* Sridhar Natarajan & Shahien Nasiripour, *Viral Tweet About Apple Card Leads to Goldman Sachs Probe*, BLOOMBERG (Nov. 19, 2019), https://www.bloomberg.com/news/articles/2019-11-09/viral-tweet-about-apple-card-leads-to-probe-into-goldman-sachs.

[3] *See* Isobel Asher Hamilton, *Apple Cofounder Steve Wozniak Says Apple Card Offered His Wife a Lower Credit Limit*, BUSINESSINSIDER (Nov. 11, 2019), https://www.businessinsider.com/apple-card-sexism-steve-wozniak-2019-11.

[4] *Id.*

[5] *See* Neil Vigdor, *Apple Card Investigated After Gender Discrimination Complaints*, NY TIMES (Nov. 10, 2019), https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.

[6] Melanie Evans & Anna Wilde Mathews, *New York Regulator Probes UnitedHealth Algorithm for Racial Bias*, WSJ (Oct. 26, 2019), https://www.wsj.com/articles/new-york-regulator-probes-unitedhealth-algorithm-for-racial-bias-11572087601

[7] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447 (2019).

[8] *See, e.g.*, Salon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL L. REV. 671, 673 (2016) ("If data miners are not careful, the process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination.").

[9] *See* Zoë Bernard, *The First Bill to Examine 'Algorithmic Bias' in Government Agencies Has Just Passed in New York City*, BUSINESSINSIDER (Dec. 19, 2017), http://www.businessinsider.com/algorithmic-bias-accountability-bill-passes-in-new-york-city-2017-12?IR=T.

bias."[10] Federally, the Algorithmic Accountability Act of 2019, which is currently pending in Congress, would require large companies to audit their algorithms for "risks that [they] may result in or contribute to inaccurate, unfair, biased, or discriminatory decisions impacting consumers."[11]

Yet, a notable absence in these legislative efforts is a formal standard for courts or regulators to deploy in evaluating algorithmic decision-making, raising the fundamental question: *What exactly does it mean for an algorithm to be accountable?* The urgency of this question follows from the meteoric growth in algorithmic decision-making, spawned by the availability of unprecedented data on individuals and the accompanying rise in techniques in machine learning and artificial intelligence.[12]

In this Article, we provide an answer to the pressing question of what accountability is, and we put forward a workable test that regulators, courts, and data scientists can apply in examining whether an algorithmic decision-making process complies with long-standing antidiscrimination statutes and caselaw. Central to our framework is the recognition that, despite the novelty of artificial intelligence and machine learning, existing U.S. antidiscrimination law has long provided a workable definition of accountability dating back to Title VII of the Civil Rights Act of 1964.[13]

Title VII and the caselaw interpreting it define what it means for any decision-making process—whether human or machine—to be accountable under U.S. antidiscrimination law. At the core of this caselaw is the burden-shifting framework initially articulated by the Supreme Court in *Griggs v. Duke Power Co.*[14] Under this framework, plaintiffs putting forth a claim of unintentional discrimination under Title VII must demonstrate that a particular decision-making practice (e.g., a hiring practice) lands disparately on members of a protected group.[15] If successful, the framework then demands that the burden shift to the defendant to show that the practice is "consistent with business necessity."[16] If the defendant satisfies this requirement, the burden returns to the plaintiff to show that an equally valid and less discriminatory practice was available that the employer refused to use.[17] The focus of Title VII is on discrimination in the workplace, but the analytical framework that emerged from the Title VII context now spans

[10] House Bill 1655, 66th Leg., Reg. Sess. (Wash. 2019), http://lawfilesext.leg.wa.gov/biennium/2019-20/Htm/Bills/House%20Bills/1655-S.htm.
[11] H.R. 2231, 116th Cong. (2019).
[12] *See* C. Scott Hemphill, Disruptive Incumbents: Platform Competition in an Age of Machine Learning, 119 COL. L. REV. 1973, 1975-1979 (2019) (surveying rapid deployment of machine learning technologies).
[13] 42 U.S.C. § 2000e (2012).
[14] Griggs v. Duke Power Co., 401 U.S. 424, 432 (1971).
[15] *See* Dothard v. Rawlinson, 433 U.S. 321, 329 (1977).
[16] 42 U.S.C. § 2000e–2(k); *see also Griggs*, 401 U.S. at 431 (in justifying employment practice that produces disparate impact, [t]he touchstone is business necessity").
[17] *See* Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975).

other domains and applies directly to the type of unintentional, statistical discrimination utilized in algorithmic decision-making.[18]

Despite the long tradition of applying this framework to cases of statistical discrimination, it is commonly violated in the context of evaluating the discriminatory impact of algorithmic decision-making. Instead, for many, the legality of any unintentional discrimination resulting from an algorithmic model is presumed to depend on simply the accuracy of the model—that is, the ability of the model to predict a characteristic of interest (e.g., productivity or credit risk) generally referred to as the model's "target."[19] An especially prominent example of this approach appears in the Department of Housing and Urban Development's 2019 proposed rule revising the application of the disparate impact framework under the Fair Housing Act (FHA) for algorithmic credit scoring.[20] The proposed rule provides that, after a lender shows that the proxy variables used in an algorithm do not substitute for membership in protected group, the lender may defeat a discrimination claim by showing that the model is "predictive of risk or other valid objective."[21] Yet this focus on predictive accuracy ignores how courts have applied the *Griggs* framework in the context of statistical discrimination.

To see why, consider the facts of the Supreme Court's 1977 decision in *Dothard v. Rawlinson.*[22] There, a prison system desired to hire job applicants who possessed a minimum level of strength to perform the job of a prison guard, but the prison could not directly observe which applicants satisfied this requirement.[23] Consequently, the prison imposed a minimum height and weight requirement on the assumption that these observable characteristics were correlated with the requisite strength required for the job.[24] In so doing, the prison was thus engaging in statistical discrimination: It was basing its hiring decision on the statistical correlation between observable proxies (an applicant's height and weight) and the unobservable variable of business necessity (an applicant's job-required strength).

[18] For example, this general burden-shifting framework has been extended to other domains where federal law acknowledges the possibility for claims of unintentional discrimination under a disparate impact theory. *See, e.g.,* Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507 (2015), (adopting burden-shifting framework for disparate impact claims under the Fair Housing Act); Ferguson v. City of Charleston, 186 F.3d 469, 480 (4th Cir. 1999) (discussing cases adopting the Title VII burden-shifting framework in Title VI disparate impact cases), *rev'd on other grounds*, 532 U.S. 67 (2001).

[19] *See infra* Part 2(C).

[20] *See* Department of Housing and Urban Development, *HUD's Implementation of the Fair Housing Act's Disparate Impact Standard*, 84 FR 42,854 (August 19, 2019) [hereinafter "2019 HUD Proposal"]. The rulemaking was intended to amend HUD's interpretation of the disparate impact standard "to better reflect" the Supreme Court's 2015 ruling in *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*, 135 S. Ct. 2507 (2015), which upheld the ability of plaintiffs to bring disparate impact cases of discrimination under the FHA.

[21] *Id*.

[22] 433 U.S. 321 (1977).

[23] *Id*. at 331-32.

[24] *Id.*

Because this procedure resulted in adverse hiring outcomes for female applicants, a class of female applicants brought suit under Title VII for gender discrimination.[25] Deploying the burden-shifting framework, the Supreme Court first concluded that the plaintiffs satisfied the disparate outcome step,[26] and it also concluded that the prison had effectively argued that hiring applicants with the requisite strength could constitute a business necessity.[27] However, the Court ultimately held that the practice used to discern strength—relying on the proxy variables of height and weight—did not meet the "consistent with business necessity" criteria.[28] Rather, absent evidence showing the precise relationship between the height and weight requirements to "the requisite amount of strength thought essential to good job performance,"[29] height and weight were noisy estimates of strength that risked penalizing females over-and-above these variables' relation to the prison's business necessity goal. In other words, height and weight were likely to be biased estimates of required strength whose use by the prison risked systematically penalizing female applicants who were, in fact, qualified.

The Court thus illustrated that in considering a case of statistical discrimination, the "consistent with business necessity" step requires the assessment of two distinct questions. First, is the unobservable "target" characteristic (e.g., requisite strength) one that can justify disparities in hiring outcomes across members of protected and unprotected groups? Second, even with a legitimate target variable, are the proxy "input" variables used to predict the target noisy estimates that are biased in a fashion that will systematically penalize members of a protected group who are otherwise qualified? In this regard, the Court's holding echoes the long-standing prohibition against redlining in credit markets. A lender who engages in red-lining refuses to lend to residents of a majority-minority neighborhood on the assumption that the average unobservable credit risk of its residents is higher than those of observably-similar but non-minority neighborhoods.[30] Yet while differences in creditworthiness can be a legitimate basis for racial or ethnic disparities to exist in lending under the FHA,[31] courts have consistently held that the mere fact that one's neighborhood is correlated with predicted credit risk is insufficient to justify red-lining.[32] By assuming that all residents

[25] *Id.* at 323.

[26] *Id*. at 330-31.

[27] *Id*. at 332.

[28] *Id*.

[29] *Id*.

[30] The term red-lining derives from the practice of loan officers evaluating home mortgage applications based on a residential map where integrated and minority neighborhoods are marked off in red as poor risk areas. Robert G. Schwemm, *Housing Discrimination* 13–42 (Release # 5, 1995).

[31] *See infra* note 170.

[32] *See* Laufman v. Oakley Building & Loan Company, 408 F. Supp. 489 (S.D. Ohio 1976)(redlining on the basis of race violates the "otherwise make unavailable or deny" provision of § 3604(a) of the FHA); Wai v. Allstate Ins. Co., 75 F. Supp. 2d 1, 7 (D.D.C. 1999)(interpreting identical language in § 3604(f)(2)

of minority neighborhoods have low credit, redlining systematically penalizes minority borrowers who actually have high credit worthiness.

These two insights from *Dothard*—that statistical discrimination must be grounded in the search for a legitimate target variable and that the input proxy variables for the target cannot systematically discriminate against members of a protected group who are qualified in the target—remain as relevant in today's world of algorithmic decision-making as they were in 1977. The primary task for courts, regulators, and data scientists is to adhere to them in the use of big data implementations of algorithmic decisions (e.g., in employment, performance assessment, credit, sentencing, insurance, medical treatment, college admissions, advertising, etc.).

Fortunately, the caselaw implementing the Title VII burden-shifting framework, viewed through basic principles of statistics, provides a way forward. This is our central contribution: We recast the logic that informs *Dothard* and courts' attitude towards redlining into a formal statistical test that can be widely deployed in the context of algorithmic decision-making. We label it the *Input Accountability Test (IAT)*.

As we show, the IAT provides a simple and direct diagnostic that a data scientist or regulator can apply to determine whether an algorithm is accountable under U.S. antidiscrimination principles. For instance, a statistician seeking to deploy the IAT could do so by turning to the same training data that she used to calibrate the predictive model of a target. In settings such as employment or lending where courts have explicitly articulated a legitimate business target (e.g., a job required skill or creditworthiness),[33] the first step would be determining that the target is, in fact, a business necessity variable. Second, taking a proxy variable (e.g., height) that her predictive model utilizes, she would next decompose the proxy's variation across individuals into that which correlates with the target variable and an error component. Finally, she would test whether that error component remains correlated with the protected category (e.g., gender). If a proxy used to predict a legitimate target variable is unbiased with respect to a protected group, it will pass the IAT, even if the use of the proxy disparately impacts members of protected groups. In this fashion, the test provides a concrete method to harness the benefits of statistical discrimination with regard to predictive accuracy while avoiding the risk that it systematically

of the FHA as prohibiting insurance redlining); Laufman, 408 F. Supp. at 496–497 (mortgage redlining); Nationwide Mut. Ins. Co. v. Cisneros, 52 F.3d 1351 (6th Cir. 1995)(insurance redlining); American Family Mut. Ins., 978 F.2d at 297 (insurance redlining); Lindsey v. Allstate Ins. Co., 34 F. Supp. 2d 636, 641–643 (W.D. Tenn. 1999)(same); Strange v. Nationwide Mut. Ins. Co., 867 F. Supp. 1209, 1213–1214 (E.D. Pa. 1994)(same). The regulatory agencies charged with interpreting and enforcing the lending provisions of the FHA have defined redlining to include "the illegal practice of refusing to make residential loans or imposing more onerous terms on any loans made because of the predominant race, national origin, etc. of the residents of the neighborhood in which the property is located. Redlining violates both the FHA and ECOA." Joint Policy Statement on Discrimination in Lending, 59 Fed. Reg. 18266 (1994).

[33] *See* Part 4(A).

penalizes members of a protected group who are, in fact, qualified in the target characteristic of interest.

We provide an illustration of the IAT in the *Dothard* setting, not only to provide a clear depiction of the power of the test, but also to introduce several challenges in implementing it and suggested solutions. These challenges include multiple incarnations of measurement error in the target, as exemplified by the UnitedHeath use of cost as a target, rather than the degree of illness, mentioned previously. These challenges also include understanding what "significantly correlated" means in our era of massive datasets. We offer an approach that may serve as a way forward. Beyond the illustration, we also provide a simulation of the test using a randomly constructed training dataset of 800 prison employees.

Finally, we illustrate how the IAT can be deployed by courts, regulators, and data scientists. In addition to employment, we list a number of other sectors – including credit, parole determination, home insurance, school and scholarship selection, and tenant selection – where either caselaw or statutes have provided explicit instructions regarding what can constitute a legitimate business necessity target.[34] We also discuss other domains such as automobile insurance and health care where claims of algorithmic discrimination have recently surfaced, but where existing discrimination laws are less clear whether liability can arise for unintentional discrimination. Businesses in these domains are thus left to self-regulating and have generally professed to adhering to non-discriminatory business necessity targets.[35] For firms with an express target delineation (whether court-formalized or self-imposed), our IAT provides a tool to pre-test their models.

We highlight, however, that firm profit margins and legitimate business necessity targets can easily be confounded in the design of machine learning algorithms, especially in the form of exploiting consumer demand elasticities (e.g., profiling consumer shopping behavior).[36] In lending, for instance, courts have repeatedly held that creditworthiness is the sole business necessity target that can justify outcomes that differ across protected and unprotected groups.[37] Yet, newly-advanced machine learning techniques make it possible to use alternative targets, such as a borrower's proclivity for comparing loan products, that focus on a lender's profit margins in addition to credit risk. In other work, we provide empirical evidence consistent with FinTech algorithms' engaging in such profiling, with the result that minority borrowers face higher priced loans, holding constant the price impact of borrowers' credit risk.[38] As such, these findings illustrate how the incentive

[34] *See Id*.
[35] *See* Part 4(B).
[36] *See* Part 4(C).
[37] *See infra* note 170.
[38] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace, Consumer Lending Discrimination in the FinTech Era, NBER Working Paper No. 25943, available at https://www.nber.org/papers/w25943.

of firms to use shopping behavior as a target can lead to discrimination in lending—a practice that could be detected by application of the IAT.[39] Profiling for shopping behavior is a subject applicable to many settings beyond the lending context and a leading topic for future research and discourse.

Our approach differs from other approaches to "algorithmic fairness" that focus solely on ensuring fair outcomes across protected and unprotected groups.[40] As we show, by failing to distinguish disparities that arise from a biased proxy from those disparities that arise from the distribution of a legitimate target variable, these approaches can themselves run afoul of U.S. antidiscrimination law. In particular, following the Supreme Court's 2009 decision in *Ricci v. DeStefano*,[41] efforts to calibrate a decision-making process to equalize outcomes across members of protected and unprotected groups—regardless of whether individuals are qualified in a legitimate target of interest—are likely to be deemed impermissible intentional discrimination.[42]

This Article proceeds as follows. In Part 2, we begin by articulating a definition for algorithmic accountability that is at the core of our input accountability test. As we demonstrate there, our definition of algorithmic accountability is effectively a test for "unbiasedness," which differs from various proposals for "algorithmic fairness" that are commonly found in the statistics and computer science literatures. Building on this definition of algorithmic accountability, Part 3 formally presents the IAT. The test is designed to provide a workable tool for data scientists and regulators to use to distinguish between legitimate and illegitimate discrimination. The test is directly responsive to the recent regulatory and legislative interest in understanding algorithmic accountability, while being consistent with long-standing U.S. antidiscrimination principles. Part 4 follows by exploring how the IAT can likewise be applied in other settings where algorithmic decision-making has come to play an increasingly important role. Part 5 concludes.

II. ACCOUNTABILITY UNDER U.S. ANTIDISCRIMINATION LAW

A. Accountability and the Burden-Shifting Framework of Title VII

We ground our definition of accountability in the antidiscrimination principles of Title VII of the Civil Rights Act of 1964.[43] Title VII, which focuses on the labor market, makes it "an unlawful employment practice for an employer (1) to ... discriminate against any individual with respect to his

[39] *See* Part 4(C).
[40] *See* Part 2(B).
[41] 557 U.S. 557 (2009).
[42] We discuss this challenge in more detail in Part 2(B).
[43] 42 U.S.C. § 2000e (2012).

compensation, terms, conditions, or privileges of employment, because of such individual's race, color, sex, or national origin; or (2) to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities ... because of such individual's race, color, religion, sex, or national origin."[44] Similar conceptualizations of antidiscrimination law were later written to apply to other settings, such as the prohibition of discrimination in mortgage lending under the FHA.[45]

In practice, Title VII has been interpreted as covering two forms of impermissible discrimination. The first and "the most easily understood type of discrimination"[46] falls under the *disparate-treatment* theory of discrimination and requires that a plaintiff alleging discrimination prove "that an employer had a discriminatory motive for taking a job-related action."[47] Additionally, Title VII also covers practices which "in some cases, … are not intended to discriminate but in fact have a disproportionately adverse effect on minorities."[48] These cases are usually brought forth under the *disparate-impact* theory of discrimination and allow for an employer to be liable for "facially neutral practices that, in fact, are 'discriminatory in operation,'" even if unintentional.[49]

Critically, in cases where discrimination lacks an intentional motive, an employer can be liable only for disparate outcomes that are unjustified. The process of how disparities across members of protected and unprotected groups might be justified is articulated in the *burden-shifting framework* initially formulated by the Supreme Court in *Griggs v. Duke Power Co.*[50] and subsequently codified by Congress in 1991.[51] This delineation is central to the definition of accountability in today's era of algorithms.

Under the burden-shifting framework, a plaintiff alleging discrimination under a claim without intentional motive bears the first burden. The plaintiff must identify a specific employment practice that causes "observed statistical disparities"[52] across members of protected and unprotected groups.[53] If the plaintiff succeeds in establishing this evidence, the burden shifts to the

[44] 42 U.S.C. § 2000e-2(a) (2012).

[45] 42 U.S.C. § 3605 (2012) ("It shall be unlawful for any person or other entity whose business includes engaging in residential real estate-related transactions to discriminate against any person in making available such a transaction, or in the terms or conditions of such a transaction, because of race, color, religion, sex, handicap, familial status, or national origin.")

[46] Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 335 n.15 (1977).

[47] Ernst v. City of Chi., 837 F.3d 788, 794 (7th Cir. 2016).

[48] Ricci v. DeStefano, 557 U.S. 557, 577 (2009).

[49] *Id*. at 577-78 (quoting Griggs, 401 U.S. at 431).

[50] Griggs, 401 U.S. at 432.

[51] Civil Rights Act of 1991, Pub. L. No. 102-66, 105 Stat. 1071 (1991).

[52] Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 979 (1988).

[53] *See also* Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975) (holding that the plaintiff has the burden of making out a prima facie case of discrimination).

defendant.[54] The defendant must then "demonstrate that the challenged practice is job related for the position in question and consistent with business necessity."[55] If the defendant satisfies this requirement, then "the burden shifts back to the plaintiff to show that an equally valid and less discriminatory practice was available that the employer refused to use."[56]

This overview highlights two core ideas that inform what it means for a decision-making process to be accountable under U.S. antidiscrimination law. First, in the case of unintentional discrimination, disparate outcomes must be justified by reference to a legitimate "business necessity."[57] In the context of employment hiring, for instance, this is typically understood to be a job-related skill that is required for the position.[58] Imagine, for instance, an employer who made all hiring decisions based on applicant's level of a direct measure of the job-related skill necessary for the job. Even if the outcome of these decision-making processes results in disparate outcomes across minority and non-minority applicants, these disparities would be justified as nondiscriminatory with respect to a protected characteristic.

Second, in invalidating a decision-making process, U.S. antidiscrimination law does so because of invalid "inputs" rather than invalid "outputs" or results. This feature of U.S. antidiscrimination law is most evident in the case of disparate treatment claims involving the use by a decision-maker of a protected category in making a job-related decision. For instance, Section (m) of the 1991 Civil Rights Act states that "an unlawful employment practice is established when the complaining party demonstrates that race, color, religion, sex, or national origin was a motivating factor for any employment practice, even though other factors also motivated the

[54] *See* Albemarle, 422 U.S. at 425 (noting that the burden of defendant to justify an employment practice "arises, of course, only after the complaining party or class has made out a prima facie case of discrimination.")

[55] 42 U.S.C. § 2000e-2(k)(1(A)(i); *see also* Griggs, 401 U.S. at 432("Congress has placed on the employer the burden of showing that any given requirement must have a manifest relationship to the employment in question.")

[56] Puffer v. Allstate Ins. Co., 675 F.3d 709, 717 (7th Cir. 2012); *see also* 42 U.S.C. § 2000e-2(k)(1(A)(ii), (C).

[57] 42 U.S.C. § 2000e-2(k)(1(A)(i). Likewise, even in the case of claims alleging disparate treatment, an employer may have an opportunity to justify the employment decision. In particular, absent direct evidence of discrimination, Title VII claims of intentional discrimination are subject to the burden-shifting framework established in *McDonnell Douglas Corp. v. Green*, 411 U.S. 792 (1973). Under the *McDonnell Douglas* framework, a plaintiff must first "show, by a preponderance of the evidence, that she is a member of a protected class, she suffered an adverse employment action, and the challenged action occurred under circumstances giving rise to an inference of discrimination." *Bennett v. Windstream Communications, Inc*., 792 F.3d 1261, 1266 (10th Cir.2015). If the plaintiff succeeds in establishing a prima facie case, the burden of production shifts to the defendant to rebut the presumption of discrimination by producing some evidence that it had legitimate, nondiscriminatory reasons for the decision. *Id*. at 1266.

[58] *See, e.g*., Griggs, 401 U.S. at 432 (holding that the employer's practice or policy in question must have a "manifest relationship" to the employee's job duties); *see also* Albermarle, 422 U.S. at 425 ("If an employer does then meet the burden of proving that its tests are 'job related,' it remains open to the complaining party to show that other tests or selection devices, without a similarly undesirable racial effect, would also serve the employer's legitimate interest in 'efficient and trustworthy workmanship.'")

practice."[59] However, this focus on inputs is also evident in cases alleging disparate impact, notwithstanding the doctrine's initial requirement that a plaintiff allege disparate outcomes across members of protected and unprotected groups. Recall that even with evidence of disparate outcomes, an employer that seeks to defend against a claim of disparate impact discrimination must demonstrate why these outcomes were the result of a decision-making process based on legitimate business necessity factors (i.e., the disparate outcomes were the result of legitimate decision-making inputs).[60] This focus on "inputs" underscores the broader policy objective of ensuring a decision-making process that is not discriminatory.

The practical challenge in implementing this antidiscrimination regime is that the critical decision-making input—an individual's possession of a job-related skill—cannot be perfectly observed at the moment of a decision, inducing the decision-maker to turn to proxies for it. However, the foregoing discussion highlights that the objective in evaluating these proxy variables should be the same: ensuring that qualified minority applicants are not being systematically passed over for the job or promotion. As summarized by the Supreme Court in *Ricci v. DeStefano*, "[t]he purpose of Title VII 'is to promote hiring on the basis of job qualifications, rather than on the basis of race or color.'"[61]

This objective, of course, is the basis for prohibiting the direct form of statistical discrimination famously examined by economists Kenneth Arrow[62] and Edmund Phelps.[63] In their models, an employer uses a job applicant's race as a proxy for the applicant's expected productivity because the employer assumes that the applicant possesses the average productivity of his or her race. If the employer also assumes the average productivity of minority applicants is lower than non-minorities (e.g., because of long-standing social and racial inequalities), this proxy will ensure that above-average productive minorities will systematically be passed over for the job despite being qualified for it. Because this practice creates a direct and obvious bias against minorities, this practice is typically policed under the disparate treatment theory of discrimination.[64]

Beyond this clearly unlawful form of statistical discrimination, a decision-maker can use statistical discrimination to incorporate not just the protected-class variable but also other proxy variables for the business-necessity unobservable attributes. For instance, an employer might seek to

[59] 42 U.S.C. § 2000e-2(m).

[60] *See, e.g.*, Dothard, 433 U.S. at 331 (holding that, to satisfy the business necessity defense, an employer must show that a pre-employment test measured a characteristic "essential to effective job performance" given that the test produced gender disparities in hiring).

[61] Ricci, 557 U.S. at 582 (citing Griggs , 401 U.S. at 424).

[62] Kenneth J. Arrow, *The Theory of Discrimination*, *in* DISCRIMINATION AND LABOR MARKETS 3 (Orley Ashenfelter & Albert Rees eds., 1973).

[63] Edmund S. Phelps, *The Statistical Theory of Racism and Sexism*, 62 AM. ECON. REV. 659 (1972).

[64] *See* text accompanying note 59.

predict a job applicant's productivity based on other observable characteristics that the employer believes are correlated with future productivity, such as an applicant's level of education or an applicant's performance on a personality or cognitive ability test.[65] Indeed, it is the possibility of using data mining to discern new and unintuitive correlations between an individual's observable characteristics and a target variable of interest (e.g., productivity or creditworthiness) that has contributed to the dramatic growth in algorithmic decision-making.[66] The advent of data mining has meant that thousands of such proxy variables are sometimes used.[67]

As the UnitedHealth algorithm revealed, however, the use of these proxy variables can result in members of a protected class experiencing disparate outcomes. The problem arises from what researchers call "redundant encodings"—the fact that a proxy variable can be predictive of a legitimate target variable *and* membership in a protected group.[68] Moreover, there are social and economic factors that make one's group membership correlated with virtually any observable proxy variable. As one proponent of predictive policy declared, "If you wanted to remove everything correlated with race, you couldn't use anything. That's the reality of life in America."[69] At the same time certain proxy variables may predict membership in a protected group over-and-above their ability to predict a legitimate target variable; relying on these proxy variables therefore risks penalizing members of the protected group who are otherwise qualified in the legitimate target variable.[70] In short, algorithmic accountability requires a method to limit the use of redundantly encoded proxy variables to those that are consistent with the anti-discrimination principles of Title VII of the Civil Rights Act and to prohibit the use of those that are not.[71]

[65] *See, e.g.,* Neal Schmitt, *Personality and Cognitive Ability as Predictors of Effective Performance at Work*, 1 ANNUAL REVIEW OF ORGANIZATIONAL PSYCHOLOGY AND ORGANIZATIONAL BEHAVIOR 45, 56 (2014) (describing web-based tests pre-employment tests of personality and cognitive ability).

[66] *See* Barocas & Selbst, supra note 8, at 677 ("By definition, data mining is always a form of statistical (and therefore seemingly rational) discrimination.")

[67] *See, e.g.*, Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data* 18 YALE. J. L. TECH. 148, 164 (2020)(describing how ZestFinance uses an "all data is credit data" approach to predict an individual's creditworthiness based on "thousands of data points collected from consumers' offline and online activities").

[68] *See* Barocas & Selbst, *supra* note 8,, at 691 (citing Cynthia Dwork et al., *Fairness Through Awareness*, 3 PROC. INNOVATIONS THEORETICAL COMPUTER SCI. CONF. 214 app. at 226 (2012)).

[69] Nadya Labi, Misfortune Teller, ATLANTIC (Jan.–Feb. 2012), http://www.theatlantic.com/magazine/archive/2012/01/misfortune-teller/308846 (quoting Ellen Kurtz, Director of Research for Philadelphia's Adult Probation and Parole Department).

[70] As noted in the Introduction, redlining represents a classic example: An individual's zip code may be somewhat predictive of one's creditworthiness, but given racialized housing patterns, it is almost certainly far more accurate in predicting one's race. Assuming that all residents in a minority-majority zip code have low creditworthiness will therefore result in systematically underestimating the creditworthiness of minorities whose actual creditworthiness is higher than the zip code average.

[71] In theory, there are statistical methods that would estimate the precise degree to which a redundantly encoded proxy variable predicts a legitimate target variable that is independent of the degree to which it predicts membership in a protected classification. We discuss these methods and their shortcomings *infra* at notes 144 to 146 and in the Appendix.

Our central contribution is in developing accountability input criteria that speak directly to the process demanded by Title VII. Specifically, we use these accountability input criteria to develop a statistical test for whether a proxy variable (or each proxy variable in a set of proxy variables) is being used in a way that causes illegitimate statistical discrimination and should therefore not be used in an algorithmic model. Fundamentally, it is a test for "unbiasedness" designed to ensure that the use of a proxy input variable does not systematically penalize members of a protected group who are otherwise qualified with respect to a legitimate-business-necessity objective. We refer to this test as the *input-accountability test*. We illustrate the test and its application with a simple pre-employment screening exam designed to infer whether a job applicant possesses sufficient strength to perform a particular job. Before doing so, however, we differentiate the input-accountability test from other approaches to algorithmic accountability.

B. The Input Accountability Test Versus Outcome-Oriented Approaches

Our input-based approach differs significantly from that of other scholars who have advanced outcome-oriented approaches to algorithmic accountability. For instance, Talia Gillis and Jann Spiess have argued that the conventional focus in fair lending on restricting invalid inputs (such as a borrower's race or ethnicity) is infeasible in the machine-learning context.[72] The reason, according to Gillis and Spiess, is because a predictive model of default that excludes a borrower's race or ethnicity can still penalize minority borrowers if one of the included variables (e.g., borrower education) is correlated with both default and race.[73] Gillis and Spiess acknowledge the possibility that one could seek to exclude from the model some of these correlated variables on this basis, but they find this approach infeasible given that "a major challenge of this approach is the required articulation of the conditions under which exclusion of data inputs is necessary."[74] They

[72] *See* Talia B. Gillis and Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHICAGO L. REV 459 (2019).

[73] *Id*. at 468-69.

[74] *Id*. at 469. Elsewhere in their article, Gillis and Spiess also suggest that input-based analysis may be infeasible because "in the context of machine-learning prediction algorithms, the contribution of individual variables is often hard to assess." *Id*. at 475. They illustrate this point by showing how in a simulation exercise, the variables selected by a logistic lasso regression in a predictive model of default differed each time the regression was run on a different randomly-drawn subsample of their data. However, this evidence does not speak to how an input-based approach to regulating algorithms would be deployed in practice. A lasso regression—like other models that seek to reduce model complexity and avoid over-fitting—seeks to reduce the number of predictors based on the underlying correlations among the full set of predictor variables. Thus, it can be used in training a model on a set of data with many proxy variables, and running a lasso regression multiple times on different subsamples of the data should be expected to select different variables with each run. However, once a model has been trained and the model's features are selected, the model must be deployed, allowing the features used in the final model to be evaluated and tested for bias. That is, regardless of the type of model fitting technique one uses in the training procedure (e.g., lasso regression, ridge regression, random forests, etc.), the model that is ultimately deployed will utilize a set of features that can be examined.

therefore follow the burgeoning literature within computer science on "algorithmic fairness"[75] and advocate evaluating the outcomes from an algorithm against some baseline criteria to determine whether the outcomes are fair.[76] As examples, they suggest a regulator might simply examine whether loan prices differ across members of protected or unprotected groups, or a regulator might look at whether "similarly situated" borrowers from the protected and nonprotected groups were treated differently.[77]

Gillis and Spiess are, of course, correct that simply prohibiting an algorithm from considering a borrower's race or ethnicity will not eliminate the risk that the algorithm will be biased against minority borrowers in a way that is unrelated to their creditworthiness (which is a legitimate-business-necessity variable).[78] Indeed, we share this concern about redundant encodings, and it motivates our empirical test. However, we part ways with these authors in that we do not view as insurmountable the challenge of articulating the conditions for excluding variables that are correlated with a protected classification, as we illustrate in Part 3.

Equally important, it is with an outcome-based approach rather than with an input-based approach where one encounters the greatest conceptual and practical challenges for algorithmic accountability. As Richard Berk and others have noted, efforts to make algorithmic outcomes "fair" pose the challenge that there are multiple definitions of fairness, and many of these

[75] For a summary, *see* Sam Corbett-Davies and Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* (arXiv.org, August 2018), available at https://arxiv.org/pdf/1808.00023.pdf. In particular, a common approach to algorithmic fairness within computer science is to evaluate the fairness of a predictive algorithm by use of a "confusion matrix." *Id.* at 4. A confusion matrix is a cross-tabulation of actual outcomes by the predicted outcome. For instance, the confusion matrix for an algorithm that classified individuals as likely to default on a loan would appear as follows:

	Default Predicted	**No Default Predicted**
Default Occurs	# Correctly Classified as Defaulting = N_{TP} (True Positives)	# Incorrectly Classified as Non-Defaulting = N_{FN} (False Negatives)
Default Does Not Occur	# Incorrectly Classified as Defaulting = N_{FP} (False Positives)	# Correctly Classified as Non-Defaulting = N_{TN} (True Negatives)

Using this table, one could then evaluate the fairness of the classifier by inquiring whether classification error is equal across members of protected and unprotected groups. *Id.* at 5. For example, one could use as a baseline fairness criterion a requirement that the classifier have the same false positive rate (i.e., $N_{FP} / (N_{FP} + N_{TN})$) for minority borrowers as for non-minority borrowers. Alternatively, one could use as a baseline a requirement of treatment equality (e.g., the ratio of False Positives to False Negatives) across members of protected and unprotected groups.

[76] *See* Gillis & Spiess, *supra* note 72, at 480 ("In the case of machine learning, we argue that outcome analysis becomes central to the application of antidiscrimination law.")

[77] *Id.* at 484-85.

[78] *See also* Jon Kleinberg, et al, *Algorithmic Fairness*, 108 AEA PAPERS AND PROCEEDINGS 22, 22–23 (2018) ("Our central argument is that across a wide range of estimation approaches, objective functions, and definitions of fairness, the strategy of blinding the algorithm to race inadvertently detracts from fairness.")

definitions are incompatible with one another.[79] The central challenge is that an outcome test will often result in *some* form of residual discrimination, raising the inevitable question: *how much* discrimination should be permissible in the outcomes?[80]

In a concrete illustration of this challenge, Richard Berk and a team of researchers at the University of Pennsylvania describe a criminal-risk-assessment tool they designed for a jurisdiction that was concerned about racial bias in the pre-trial release rates among criminal defendants who were awaiting trial.[81] In general, when a defendant was arraigned in this jurisdiction, a magistrate judge was required to decide whether the defendant should be released until the trial date, considering (among other things) the defendant's threat to public safety.[82] Berk and his team developed a forecasting algorithm of a defendant's risk, using a subsequent arrest for a violent crime within 21 months of release as a proxy for the defendant's threat to public safety.[83]

To reduce racial disparities, Berk and his team tuned the algorithm so that it was equally accurate at predicting release across racial categories; that is, the rate of re-arrest for a violent crime among minority and non-minority defendants was the same.[84] However, the base rate of re-arrest among minority defendants was higher than non-minority defendants, meaning that the chosen fairness objective could be accomplished only by making the algorithm biased. In particular, the algorithm had to classify more "violent" non-minority defendants as "nonviolent" (thus resulting in their release), and it had to classify more "nonviolent" minority defendants as "violent" (thus resulting in their detention).[85] The need to bias the algorithm in this fashion

[79] *See* Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art* 33 (arXiv.org, May 30, 2017), available at https://arxiv.org/pdf/1703.09207.pdf (arguing that "[t]here are different kinds of fairness that in practice are incompatible").

[80] *See, e.g.,* Gillis & Spiess, *supra* note 72, at 486 (advocating an outcome test in which a regulator evaluates whether lending outcomes differ by race among "similarly situated" borrowers "should include a degree of tolerance set by the regulator").

[81] *See* Berk et al., *supra* note 79, at 31-33.

[82] *Id*. at 31.

[83] *Id*. at 31-33.

[84] *Id.*

[85] As Sam Corbett-Davies and Sharad Goel note, it is the different underlying distribution of risk (or other unobservable characteristic of interest) among minority and non-minority populations that gives rise to the alternative and incompatible definitions of fairness based on classification errors. *See* Corbett-Davies & Goel, *supra* note 75, at 2 ("When the true underlying distribution of risk varies across groups, differences in group-level error rates are an expected consequence of algorithms that accurately capture each individual's risk."). Within the antidiscrimination literature, this statistical challenge is known as the problem of infra-marginality and has long plagued outcome tests of discrimination in human decisions. *See* Ian Ayres, *Outcome Tests of Racial Disparities in Police Practices*, 4 JUSTICE RESEARCH AND POLICY 131 (2002). The central problem is that an inquiry into whether a decision is unbiased is concerned with what happens at the margin (i.e., is the same standard being applied to everyone?). Error rates, however, are evaluated away from the margin as they rely on evaluating outcomes following the application of a cut-off standard to all individuals (both those who might be near the cut-off and those who might be far from it). If the risk distributions differ across minority and non-minority individuals, lumping together both marginal and infra-marginal individuals will produce error rates that differ by race.

arose from the fact that minority defendants had a higher baseline re-arrest rate.[86] As a result, the algorithm could achieve its particular definition of fairness—equality of accuracy, conditional on release—only by releasing more non-minority defendants that were more likely be arrested (to compensate for the overall lower rate of arrest for non-minority defendants) and by not releasing some minority defendants that were unlikely to be arrested (to compensate for their overall higher rate of arrest).[87] As they note, in sacrificing one form of fairness for another, the resulting "differences [in error rates] can support claims of racial injustice."[88]

To be sure, applying our test for "unbiasedness" does not solve the challenge of addressing concerns about fairness. A decision that passes our test might still be objectionable for other distributional reasons. In the case of lending, for instance, creditworthiness is a well-recognized target variable, but the determinants of creditworthiness (e.g., income, income growth, wealth) also reflect long-standing racial and economic inequalities, ensuring that creditworthiness will likewise reflect these racial and economic inequalities. Thus, even an unbiased lending rule would result in lending outcomes that reflect these structural inequalities, and rectifying them would require an additional intervention, such as through subsidized loan programs and other policies designed to encourage lending to low and moderate-income families. Indeed, this approach is reflected in existing U.S. housing programs such as the Federal Housing Administration mortgage program (which seeks to provide mortgages to low and moderate-income borrowers)[89] and the Community Reinvestment Act (which seeks to encourage lenders to provide loans to residents of low and moderate-income neighborhoods).[90]

This two-step approach—ensuring that decision-making processes are unbiased and then subsequently addressing distributional concerns directly through transfers and subsidies—is consistent with democratic principles. As we show, it is conceivable to design a decision-making process that is unbiased with respect to a legitimate business necessity. This is the objective of the IAT. But as Berk's study illustrates, it is not possible to design a decision-making process that satisfies every possible definition of "fairness." Evaluating an algorithm for whether it is "fair" rather than "unbiased" thus risks enforcing a particular vision of fairness and doing so in a way that lacks transparency. Indeed, Berk et al. themselves provide no explanation for why

[86] *Id*. at 32.

[87] *Id*.

[88] *Id*.

[89] *See* James H. Carr, Katrin B. Anacker, *The Complex History of the Federal Housing Administration: Building Wealth, Promoting Segregation, and Rescuing the U.S. Housing Market and The Economy*, 34 BANKING & FIN. SERVICES POL'Y REP. 10 (2015) (describing program).

[90] *See* Keith N. Hylton, *Banks and Inner Cities: Market and Regulatory Obstacles to Development Lending*, 17 YALE J. ON REG. 197 (2000) (describing Community Reinvestment Act).

they opted to implement their particular definition of fairness.[91] Likewise, an algorithm that seeks to "fix" disparate outcomes that arise from an unbiased decision-making process can risk diminishing the ability to identify the source of the underlying structural inequalities and/or measurement error in the decision-making process. In Berk's setting, for instance, a risk-assessment algorithm that results in equality of release rates across minority and non-minority defendants could hide the possibility that minority defendants have a higher re-arrest rate because of prejudice among police, which in turn would raise the question of whether re-arrests are truly a decent proxy for a defendant's level of violence. For all of these reasons, determination of distributional equity is accordingly best left to institutions that can evaluate the relevant trade-offs in a transparent fashion.

Regardless of these conceptual and practical challenges, outcome-based approaches to algorithmic fairness would almost certainly be deemed legally problematic following the Supreme Court's 2009 decision in *Ricci v. DeStefano.*[92] The facts giving rise to *Ricci* involved a decision by the city of New Haven to discard the results of an "objective examination" that sought to identify city firefighters who were the most qualified for promotion.[93] The city justified its decision to discard the results on the basis that there was a statistical racial disparity in the results, raising the risk of disparate impact liability under Title VII.[94] A group of white and Hispanic firefighters sued, alleging that the city's discarding of the test results constituted race-based disparate-treatment.[95] In upholding their claim, the Court emphasized the extensive efforts that the city took to ensure the test was job-related[96] and that there was "no genuine dispute that the examinations were job-related and consistent with business necessity."[97] Nor did the city offer "a strong basis in evidence of an equally valid, less-discriminatory testing alternative."[98] Prohibiting the city from discarding the test results was therefore required to prevent the city from discriminating against "qualified candidates on the basis of their race."[99]

The Court's assumption that the promotion test identified the most qualified firefighters makes it difficult to see a legal path forward for explicit race-based adjustments of algorithmic outcomes. Assuming the algorithm

[91] Berk et al, *supra* note 79, at 32 (describing their choice of error metric as "conditional use accuracy equality, which some assert is a necessary feature of fairness.")
[92] 557 U.S. 557 (2009)
[93] *Id*. at 562.
[94] *Id*.
[95] *Id*. at 562-63.
[96] *Id*. 586-588.
[97] *Id*. at 587; see also id at 589 ("The City, moreover, turned a blind eye to evidence that supported the exams' validity.")
[98] *Id*. at 589.
[99] *Id*. at 584 ("Restricting an employer's ability to discard test results (and thereby discriminate against qualified candidates on the basis of their race) also is in keeping with Title VII's express protection of bona fide promotional examinations.")

properly functions to identify individuals who are qualified in a specified target, such race-based adjustments would appear to be no different than what the city of New Haven attempted to do with the promotion test results. Rather, *Ricci* underscores the fundamental importance of ensuring that decision-making processes do not systematically discriminate against qualified individuals because of their race. And as noted previously, this is the objective of the IAT.

C. The Input Accountability Test Versus the "Least Discriminatory Alternative" Test

We differ also from scholars and practitioners who focus only on the final step in the disparate-impact burden-shifting framework. Recall that according to this burden-shifting framework, an employer who establishes that a business practice can be justified by a legitimate business necessity shifts the burden back to the plaintiff to show that an equally valid and less discriminatory practice was available that the employer refused to use.[100] Some commentators have mistakenly assumed that this test implies that the critical question for an algorithm that produces a disparate impact is whether the algorithm uses the least discriminatory predictive model for a given level of predictive accuracy. Of course, in using machine learning over thousands of variables, it is easy to run many models and decide which creates the least disparate impact for a given level of accuracy in prediction. But this approach will not address whether any of the variables used in the model are systematically penalizing members of a protected group that are otherwise qualified in the skill or characteristic the model is seeking to predict.

Nonetheless, a number of commentators have mistakenly argued that the central test for whether an algorithm poses any risk of illegitimate discrimination should be whether there are alternative models that can achieve the same level of predictive accuracy with lower levels of discrimination.[101] For instance, in an oft-cited discussion paper regarding fair lending risk of credit cards, David Skanderson and Dubravka Ritter advocate that lenders should focus on this step of the disparate-impact framework when evaluating the fair-lending risk of algorithmic credit-card models.[102] Specifically, Skanderson and Ritter note that "a model or a model's predictive

[100] *See* text accompanying note 56.

[101] *See, e.g.,* Nicholas Schmidt and Bryce Stephens, *An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination*, (arXiv.org, Nov. 2019), available at https://arxiv.org/pdf/1911.05755.pdf (advocating for using "a 'baseline model' that has been built without consideration of protected class status, but which shows disparate impact, and then search[ing] for alternative models that are less discriminatory than that baseline model, yet similarly predictive.")

[102] *See, e.g.,* David Skanderson & Dubravka Ritter, Discussion Paper, Fair Lending Analysis of Credit Cards, Federal Reserve Bank of Philadelphia (August 2014), available at https://www.philadelphiafed.org/-/media/consumer-credit-and-payments/payment-cards-center/publications/discussion-papers/2014/d-2014-fair-lending.pdf?la=en (last visited February 2, 2020).

variable with a disproportionate adverse impact on a prohibited basis may still be legally permissible if it has a demonstrable business justification and there are no alternative variables that are equally predictive and have less of an adverse impact."[103] For Skanderson and Ritter, the business necessity defense for an algorithmic decision-making process therefore boils down to whether it is the most accurate possible test in predicting a legitimate target variable of interest. As they summarize in the context of lending, "If a scoring system is, in fact, designed to use the most predictive combination of available credit factors, then it should be unlikely that someone could demonstrate that there is an equally effective alternative available, which the lender has failed to adopt."[104]

To see how validating an algorithm based entirely on the fact that it is the most predictive model available would validate algorithms that are clearly biased against members of a protected group, we offer an example. Consider an employer who needs employees that can regularly lift 40 pounds as part of their everyday jobs. Imagine this employer designs a one-time test of whether applicants can lift 70 pounds as a proxy for whether the applicant can repetitively lift 40 pounds. The employer can show that this test has 90% prediction accuracy. However, those applicants that fail the test who in fact could regularly lift 40 pounds are disproportionately female. Thus, the test, because it is not a perfect proxy, causes a disparate impact on female applicants. Now assume that it can be shown that a one-time test of whether applicants can lift 50 pounds produces no disparate impact on females but has an accuracy rate of just 85%. Under Skanderson and Ritter's approach, the employer would have no obligation to consider the latter test, despite the fact that a 70-pound test will systematically penalize female applicants that can in fact satisfy the job requirement.

Not surprisingly, this approach to pre-screening employment tests has been expressly rejected by courts. In *Lanning v. Southeastern Pennsylvania Transportation Authority*,[105] for instance, the Third Circuit considered a physical fitness test for applicants applying to be transit police officers. The fitness test involved a 1.5 mile run that an applicant was required to complete

[103] *Id.* at 38.

[104] *Id.* at 43. This line of reasoning also informs Barocas and Selbst's conclusion that Title VII provides a largely ineffective means to police unintentional discrimination arising from algorithms. *See* Barocas & Selbst, *supra* note 66, at 701-714. According to Barocas and Selbst, the business necessity defense requires that an algorithm is "predictive of future employment outcomes." *Id.* If this is correct, it would logically follow that an employer will have no disparate impact liability from using the most predictive algorithmic model for a legitimate job-related quality since an equally predictive, less discriminatory alternative would not be available. However, this conclusion relies on an assumption that predictive accuracy is a necessary and sufficient condition to justify a decision-making process that produces a disparate impact. As we show, this is an incorrect assumption as courts have been careful not to conflate the business necessity defense with predictive accuracy. A predictive model may be accurate in predicting whether an individual is likely to have a legitimate target characteristic but nevertheless be biased against members of a protected group who are otherwise qualified in the target characteristic.

[105] 181 F.3d 478 (3rd Cir. 1999), cert. denied, 528 U.S. 1131 (2000).

within 12 minutes; however, the 12 minute cut-off was shown to have a disparate impact on female applicants.[106] The transit authority acknowledged that officers would not actually be required to run 1.5 miles within 12 minutes in the course of their duties, but it nevertheless adopted the 12 minute cut-off because the transit authority's expert believed it would be a more "accurate measure of the aerobic capacity necessary to perform the job of a transit police officer."[107]

In considering the transit authority's business-necessity defense, the court agreed that aerobic capacity was related to the job of a transit officer.[108] It also agreed that by imposing a12 minute cut-off for the run, the transit authority would be increasing the probability that a job applicant would possess high aerobic capacity.[109] Nonetheless, the court rejected this "more is better" approach to setting the cutoff time:

> Under the District Court's understanding of business necessity, which requires only that a cutoff score be "readily justifiable," [the transit authority], as well as any other employer whose jobs entail any level of physical capability, could employ an unnecessarily high cutoff score on its physical abilities entrance exam in an effort to exclude virtually all women by justifying this facially neutral yet discriminatory practice on the theory that more is better.[110]

Accordingly, the court required "that a discriminatory cutoff score be shown to measure the minimum qualifications necessary for successful performance of the job in question in order to survive a disparate impact challenge."[111] In other words, in determining whether disparate outcomes are justified, the question to ask in evaluating a predictive model of a legitimate target variable is not simply whether the model is accurate in predicting the target variable. Rather, the inquiry should be both whether the model is accurate and whether the cutoff score used to classify individuals serves the employer's legitimate business goals.[112]

[106] *Id.* at 482

[107] *Id.*

[108] *Id.* at 492.

[109] *Id.* ("The general import of these studies is that the higher an officer's aerobic capacity, the better the officer is able to perform the job.")

[110] *Id.* at 493.

[111] *Id.*

[112] The Third Circuit was even more explicit that setting the cut-off was effectively about calibrating the predictive accuracy of the employment test. *See* Lanning, 308 F.3d at 292 ("It would clearly be unreasonable to require SEPTA applicants to score so highly on the run test that their predicted rate of success be 100%. It is perfectly reasonable, however, to demand a chance of success that is better than 5% to 20%."); *see also* E.E.O.C. vs. Simpson Timber Co., 1992 WL 420897 (finding that a pre-employment step test accurately measured strength and endurance, which were legitimate business goals, but an equally

An even stronger rejection of the "more is better" approach to predictive accuracy appeared in *Murphy v. Derwinski*.[113] There, the plaintiff, Mary Murphy, applied to become a Roman Catholic chaplain at hospitals operated by the United States Veterans Administration (VA).[114] The VA rejected Murphy's application on the ground that VA guidelines required that all applicants be ordained in the relevant religion and receive an ecclesiastical endorsement from their churches.[115] However, within the Roman Catholic religion, only men can be ordained as priests, making it impossible for Murphy to satisfy these requirements.[116] In her subsequent Title VII lawsuit, the district court determined that Murphy made out a prima facie case of discrimination based on this policy and that the defendant articulated a legitimate business justification for it.[117] In particular, the court agreed with the VA that the agency's interest in providing a full range of ritual services to its Catholic patients creates a legitimate purpose for requiring ordination for VA chaplains.[118] The court nevertheless rejected the VA's argument that if the ordination requirement were eliminated, the VA would be unable to accommodate the needs of its patients by providing the full range of religious services.[119] Rather, the court held that by removing the ordination requirement and requiring only ecclesiastical endorsement, the VA could still ensure that its patients received the religious services that the Catholic Church deemed sufficient.[120]

On appeal, the Tenth Circuit affirmed and elaborated on why removing the ordination requirement would not impair the VA's legitimate interests.[121] Citing the VA's own administrative materials, the court noted that the chaplain service's primary objective was to "provide for the spiritual welfare"[122] of patients such as through establishing relationships with patients and providing patients and family members with ministry in crisis situations, and "[p]riests are not needed to administer these functions."[123] The court acknowledged that only priests could administer sacraments to patients subscribing to the Roman Catholic faith,[124] but it concluded that the VA would still be able to accommodate the religious needs of its Roman Catholic patients:

effective, less discriminatory alternative existed in the form of using a lower cut-off score to determine if an applicant passed the test).

[113] 990 F.2d 540 (10th Cir. 1993).

[114] *Id.* at 542.

[115] *Id.*

[116] *Id.* at 542 n. 5.

[117] Murphy v. Derwinski, 776 F. Supp. 1466, 1470 (D. Colo. 1991).

[118] *Id.*

[119] *Id.* at 1472-73

[120] *Id.*

[121] *Murphy* 990 F.2d at 545-547

[122] *Id.* at 546

[123] *Id.*

[124] *Id.* at 545.

> The experience of the VA hospital in Denver where Murphy sought to work suggests that removal of the ordination requirement will not disrupt services only priests may perform. Of the hospital's six chaplains at the time of this lawsuit, two were Catholic priests. Thus, four of the chaplains could not provide Roman Catholics with services unique to that religion. Similarly, none of the six could administer unique religious services to members of nonrepresented faiths. When a priest is needed but, for whatever reason, is unavailable, the VA Manual calls for supplementing its full-time chaplain services through contract help or other arrangements.[125]

Thus, the court held that requiring only the ecclesiastical endorsement was an alternative, nondiscriminatory requirement that could serve the VA's legitimate interest in providing the full range of religious services to its patients.[126] In so doing, note the inconsistency with the approach to the "less discriminatory alternative" inquiry as interpreted by Skanderson and Ritter. Like the transit authority in *Lanning*, the VA in *Murphy* was concerned about identifying job applicants who, at any given moment during their job performance, were likely to serve the VA's legitimate interest.[127] That is, the VA's hiring guidelines were designed to provide an answer to the question: When a Roman Catholic patient requires religious services, will this applicant be able to provide them? The two requirements—ordination and ecclesiastical endorsement—were clearly accurate in predicting whether an applicant could provide these services. And the requirement that applicants have both characteristics made it virtually certain that a VA chaplain could, in fact, provide any and all of these religious services, any time of the day (morning, noon or night). But like the court in *Lanning*, the *Murphy* court also concluded that setting the probability threshold so high—in this case, imposing an application requirement that made it close to 100% certain that a chaplain would be available to provide any and all Catholic religious services—was simply too high. As the court emphasized, most of the services required of chaplains did not require ordination. Thus, eliminating the ordination requirement might lessen the probability that a VA chaplain would actually be available to administer the sacraments if a patient happened to require them, but the probability would nonetheless remain high enough to

[125] *Id.* at 546.
[126] *Id.* at 545-546.
[127] *See Murphy*, 776 F. Supp. at 1472 ("The VA asserts that if ordination were not required, it would not be able to accommodate the needs of its patients by providing the full range of religious services. VA chaplains must be able to administer the various sacraments, and only ordained priests are qualified for these duties.")

satisfy the VA's legitimate interest in accommodating the religious needs of its Roman Catholic patients.

In short, in the era of algorithmic decision-making, we view the need to inquire into whether there exists a "less discriminatory alternative" to be fundamentally about the cut-off threshold applied to an algorithm that otherwise passes our test. Whether an algorithm is screening for acceptable job applicants or acceptable borrowers, the end result is both to estimate the probability that an individual has a legitimate target characteristic and then to apply a probability cut-off to make the ultimate accept/reject classification. In setting this cut-off, *Lanning* and *Murphy* are reminders of the need to consider whether the cut-off has been set at the minimum level required to advance a legitimate business interest, such as successful performance of the job in question.[128] As we show below, doing so can help ensure that a decision-making process that passes our test is not inappropriately biased against members of a protected group simply because of the unequal distribution of a legitimate target variable (e.g., strength or speed) across protected and unprotected groups.[129]

D. The Input Accountability Test Versus HUD's Mere Predictive Test

Finally, we consider the IAT against HUD's 2019 proposed rulemaking regarding the application of the disparate impact framework under the FHA.[130] Given the increasing role of algorithmic credit scoring, the proposed rule-making expressly provides for a new defense for disparate impact claims where "a plaintiff alleges that the cause of a discriminatory effect is a model used by the defendant, such as a risk-assessment algorithm…."[131] In particular, the proposed rule provides that in these cases, a lender may defeat the claim by "identifying the inputs used in the model and showing that these inputs are not substitutes for a protected characteristic and that the model is predictive of risk or other valid objective."[132] In other words, so long as a

[128] *See Lanning* F.3d. 481 ("[U]nder the Civil Rights Act of 1991, a discriminatory cutoff score on an entry level employment examination must be shown to measure the minimum qualifications necessary for successful performance of the job in question in order to survive a disparate impact challenge."); *see also* Association of Mexican-American Educators v. State of California, Nos. 96-17131 and 97-15422, 1999 WL 976720 (9th Cir. Oct. 28, 1999) (upholding, against a disparate-impact challenge under Title VII, a requirement that public-school teachers "demonstrate basic reading, writing and mathematics skills in the English language as measured by a basic skills proficiency test" and holding as not clearly erroneous the district court's finding that the cutoff scores "reflect[ed] reasonable judgments about the minimum levels of basic skills competence that should be required of teachers.").

[129] This interpretation of the third prong of the Title VII burden-shifting framework is also consistent with the common view that it is effectively a test for whether an articulated business necessity defense is a pretext for discrimination; that is, as noted in *Lanning*, one could purposefully set a threshold at a sufficiently high level to ensure that members of protected groups will fail the test. *See, e.g.,* Murphy 990 F.2d at 545 ("The focus on appeal is whether the VA's business justification for requiring an ordained clergy person constitutes a pretext for gender discrimination.")

[130] *See* 2019 HUD Proposal, *supra* note 20.

[131] *Id*. at 42,862.

[132] *Id*.

variable is not an undefined "substitute" for a protected characteristic, any model that predicts creditworthiness is sufficient to defeat a claim of disparate impact discrimination.

This approach to algorithmic accountability, however, suffers from the same defect noted previously with regard to those who have misapplied the "least discriminatory alternative" test. Specifically, by focusing solely on whether a model is "predictive of risk or other valid objective," HUD's test leaves open the possibility that a lender can adopt a model that systematically discriminates against borrowers who are, in fact, creditworthy. Recall that in our hypothetical strength test, the ability to lift 70 pounds was, in fact, predictive of whether an applicant could regularly lift 40 pounds; however, it systematically discriminated against women who were qualified for the job. Worse still, by not even requiring that a model have any particular level of accuracy, HUD's test would seemingly permit the use of any proxy so long as it has *some* correlation with credit risk. Indeed, this approach would even appear to permit the use of explicit redlining in a predictive model so long as a lender could show that the average credit risk of a majority-minority neighborhood is marginally higher than that of non-majority-minority neighborhoods.

In contrast, a central goal of the IAT is to ensure that in evaluating whether a model is consistent with a decision-maker's legitimate business necessity, it incorporates only those proxy variables that are not corelated with a protected characteristic beyond the proxy variables' correlation with a legitimate target variable.

III. THE INPUT ACCOUNTABILITY TEST

In this section, we formally present our *input accountability test* (IAT) for unintentional discrimination. We begin with some nomenclature. The design of a decision-making algorithm rests fundamentally on the relationships between a set of variables, referred to as "features," and an underlying latent skill or attribute of interest (creditworthiness, productivity, etc.), referred to as a "target." Today, the relationships between targets and features are increasingly analyzed and developed within artificial-intelligence and machine-learning processes, but it is just as likely that an organization uses an algorithmic decision process based on human-selected data or even on personal intuition. The IAT applies to both machine learning and human learning.

Our core contribution is a test that informs when a feature's (a proxy variable's) use has correlations with a target that produce statistical discrimination against a protected class that is unjustified according to the criteria developed in Part 2. That is, the IAT detects if the use of a feature results in systematically penalizing members of a protected group who are otherwise qualified in the target variable of interest. After illustrating the

IAT, we extend our analysis to consider the mis-assertion of a target cutoff that does not reflect the true level of the target that is required, reflecting the prior example we gave of requiring job applicants to lift 70 pounds as a mis-asserted target.

A. The Test

We illustrate our test throughout with the facts giving rise to the 1977 Supreme Court decision in *Dothard v. Rawlinson*.[133] As noted previously, in *Dothard*, female applicants for prison guard positions challenged a prison's minimum height and weight requirements as inconsistent with Title VII.[134] Because the average height and weight of females was less than that for males, the female applicants argued that the requirement created an impermissible disparate impact for females under Title VII.[135] In response, the prison argued that a height and weight requirement was a justified job requirement given that an individual's height and weight are predictive of strength, and strength was required for prison guards to perform their jobs safely.[136] In short, the prison took the position that the general correlation between one's height/weight and strength was sufficient to justify the disparate outcomes this requirement caused for women. The Supreme Court, however, rejected this defense.[137] Rather, to justify gender differences in hiring outcomes, the prison would need to show that it had tested for the *specific type of strength* required for effective job performance; [138] in other words the prison would have to be concerned with the aspects of strength that the proxy variables were and were not picking up that related to a prison guard's need for strength.

We use this setup and some hypothetical applicants to lay out the IAT. Imagine for example that twelve individuals apply for an open prison guard position, of which six applicants are male and six are female. In evaluating the applicants, the prison seeks to select those applicants who possess the actual strength required for successful job performance. For simplicity, assume that an individual's strength can be measured on a scale of zero to one hundred, and that a strength score of at least sixty is a true target for job effectiveness (in the Court's language a strength of sixty is a legitimate-business-necessity criterion). The challenge the prison faces in evaluating job applicants is that each applicant's actual strength is unobservable at the time of hiring, thus inducing the prison to rely on height as a proxy.

[133] 433 U.S. 321 (1977).
[134] *Id.* at 323-24.
[135] *Id.*
[136] *Id.* 331.
[137] *Id.* at 332.
[138] *Id.* at 332 ("If the job-related quality that the appellants identify is bona fide, their purpose could be achieved by adopting and validating a test for applicants that measures strength directly.")

Assume that the use of a minimum height requirement results in the following distribution of applicants according to their actual but unobservable strength (Figure 1).

Figure 1

	Results with Height Test		
Actual Strength	**Meets Height Requirement**	**Fails Height Requirement**	
100	x		
90	x		*Minimum Required Strength* ↓
80	x		
70	x	x	
60		x	
50	x		
40	x	x	
30		x	
20		x	
10		x	
0			

x = applicant

Consistent with the prison's argument, there is a clear correlation between an applicant's height and actual strength. However, when we examine the gender of the applicants, we discover that only the six male applicants satisfy the minimum height requirement (Figure 2).

Figure 2

	Results with Height Test		
Applicant's Strength	**Meets Height Requirement**	**Fails Height Requirement**	
100	□		
90	□		*Minimum Required Strength* ↓
80	□		
70	□	•	
60		•	
50	□		
40	□	•	
30		•	
20		•	
10		•	
0			

□ = male; • = female

In this situation, a basic correlation test between height and strength has produced exactly the same injury noted in Part 2: The imperfect relationship between height and strength results in penalizing otherwise qualified female applicants and benefiting unqualified male applicants. This can be seen from the fact that the only applicants who possessed sufficient strength but failed the height test were female. Likewise, the only applicants who met the height test but lacked sufficient strength were male. The screening test is thus systematically biased against female applicants for reasons unrelated to a legitimate business necessity.

This example points to the crux of the IAT. In general, the objective of the test is to ensure that a proxy variable is excluded from use if the imperfect relationship between the proxy variable and the target of interest results in systematically penalizing members of a protected group that are otherwise qualified in the target of interest. In other words, since the proxy variable (height) is not a perfect predictor of having the target strength, there is some residual or unexplained variation in height across applicants that is unrelated to whether one has the required strength. The question is whether that residual is correlated with gender. In Figure 2, it is.

To avoid this result in *Dothard*, the Supreme Court therefore required a better proxy for required strength. In particular, the prison would be required to "adopt[] and valida[te] a test for applicants that measures strength directly" in order to justify disparities in hiring outcomes.[139] For example, assume that the prison implemented as part of the job application a physical examination that accurately assessed required strength, which produced the following results (Figure 3).

[139] *Id*. at 332.

Figure 3

	Results with Perfect Strength Exam		
Actual Strength	**Passes Exam**	**Fails Exam**	
100	□		
90	□		
80	□		
70	□ ●		
60	●		*Minimum Required Strength* ↓
50		□	
40		□ ●	
30		●	
20		●	
10		●	
0			

□ = male; ● = female

The examination was perfect in classifying all individuals – male and female – as qualified if they in fact were so. Note that, even under this perfect exam, more males than females would be deemed eligible for the position. This disparity, however, arises solely through differences in actual strength (a legitimate business necessity).

Figure 3 is an ideal outcome in the sense that the prison was perfect in measuring each applicant's actual strength, but perfect proxy variables are rarely available. Imagine instead that the prison asks the applicants to complete a simple muscle-mass index assessment (Figure 4).[140]

[140] For instance, imagine the prison assesses each applicant's mid-arm muscle circumference (MAMC) and requires a minimum measure which the prison believes is associated with having a minimum strength of 60. The MAMC is one of several techniques to measure muscle mass. *See* Julie Mareschal et al., *Clinical Value of Muscle Mass Assessment in Clinical Conditions Associated with Malnutrition*, 8 J. CLIN. MED. 1040 (2019).

Figure 4

	Results with Muscle Mass Exam		
Actual Strength	**Meets Muscle Mass Requirement**	**Fails Muscle Mass Requirement**	
100	□		
90	□		*Minimum*
80	□		*Required*
70	●	□	*Strength*
60		●	↓
50	□		
40	●	□	
30		●	
20		●	
10		●	
0			

□ = male; ● = female

As can be seen, muscle mass proxies for required strength with a positive, significant correlation, but it does so with error. In particular, there are applicants who are sufficiently strong but fail the muscle mass requirement, and there are applicants who meet the muscle mass requirement but are not sufficiently strong. The difference from Figure 2, however, is that the proxy is unbiased: Neither male applicants nor female applicants are favored by the fact that the proxy does not perfectly measure required strength. This is illustrated by the fact that one male and one female fail the muscle mass requirement but possess sufficient strength for the job, and one male and one female meet the muscle mass requirement but lack sufficient strength. Because the proxy is unbiased with respect to gender, an employer should therefore be permitted to use muscle mass as a proxy for required strength.

B. The Test in Regression Form

Moving from concepts to practice, standard regression techniques provide a straightforward means to implement the IAT. In keeping with the foregoing example, we return to the modified facts of *Dothard,* in which a prison uses a job applicant's height as a proxy for whether they have the required strength to perform the job of a prison officer.[141] The prison does so based on the assumption that required strength is manifested in an

[141] Of course, there might be multiple proxies. For instance, imagine the job requirements were strength and IQ, in some combination. Such a specification could be handled by more complex formations on the right-hand side of the regression framework that we discuss in this subsection.

individual's height. However, height is also determined by a host of other causes that are unrelated to strength. If we represent this group of non-strength determinants of height for a particular individual i as ε_i, we can summarize the relationship between the height and strength as follows:

$$Height_i = \alpha \cdot Strength_i + \varepsilon_i,$$

where α is a transformation variable mapping the relationship of strength to expected height. If ε_i was zero for each individual i, the equation becomes $Height_i = \alpha \cdot Strength_i$. In such a setting, an individual's height would be precisely equal to the individual's strength, multiplied by the scalar α. Therefore, one could compare with perfect accuracy the relative strength of two individuals simply by comparing their heights.

Where ε is non-zero, using height as a proxy for strength will naturally be less accurate; however, using height in this fashion will pose no discrimination concerns if ε (the unexplained variation in height that is unrelated to strength) is uncorrelated with a protected classification. This was precisely the case in Figure 4: Strength was *somewhat* manifested through the muscle mass index. Thus, it would be a useful variable for predicting which job applicants had the required strength for the job. Moreover, while it was error-prone in measuring actual strength (i.e., $\varepsilon_i \neq 0$), using one's muscle mass index to infer strength would pass the IAT:

$$\varepsilon_i \perp gender;$$

the errors were not statistically correlated with gender, the protected category in our example. To implement this test empirically, the prison would use the historical data it holds concerning its existing employees' measured height and strength and regress employee height on employee strength to estimate α, which can be used to estimate ε_i for each employee.[142] Using these ε_i estimates, the prison would then examine whether they are correlated with employee gender.

How would the IAT be used in a setting where the proxy is not a continuous measure (such as one's height or muscle mass) but rather a binary outcome of whether an individual possesses a specified level of the measure? Recall that this was the case in our hiring example where the prison first assessed an applicant's height and then applied a cut-off score to eliminate from consideration those applicants who fell below it. As reflected in

[142] The regression will also estimate a constant term that is utilized in calculating the relationship between strength and height.

Dothard and *Lanning*, applying a minimum cut-off score to a proxy variable is a common decision-making practice, including within machine learning.[143]

The application of the IAT would use the same framework, but using as the left-hand-side variable an indicator variable for whether an individual i was above or below the cutoff—for our example, $Height_i = 1$ for applicants above the cutoff and $Height_i = 0$ for applicants below it. To estimate a discrete 0-1 variable (*Height*) as a function a target (e.g., *Strength*), the preferred model is a logistic regression (or a comparable model for use with a dichotomous outcome variable). Logistic regression is a transformation that takes a set of zeros and ones representing an indicator variable and specifies them in terms of the logarithm of the odds ratio of an outcome (in our example, the odds ratio is the probability of $Height_i$ being above the cut-off divided by the probability that it is below the cut-off). This formulation is then regressed on the target measure (*Strength*). To generate the residuals (ε_i) for the IAT test, one predicts the probability of a positive outcome and then generates the error as the true outcome minus the predicted probability. As above, to pass the test, the residuals should not be significantly correlated with gender.

Finally, we conclude this overview with a discussion of what happens when a proxy variable fails the input accountability test: exclusion from the model. If the residuals (ε_i) are correlated with a protected classification (e.g., gender), it may be possible to "de-bias" a model that predicts strength from height, most notably by adding an individual's membership (or lack of membership) in a protected class as an input in the predictive model.[144]

However, as shown in the Appendix, the fact that de-biasing requires us to include *Gender* in the predictive model impairs the utility of this approach. A predictive model that explicitly scores individuals differently according to gender constitutes disparate treatment, making it a legally impermissible means to evaluate individuals. To avoid this challenge, proponents of this approach have therefore advocated that, in making predictions, the model should assign all individuals to the mean of the protected classification;[145] in our example, one would do so by treating all individuals as if *Gender* = 0.5 (i.e., (1 + 0) / 2) when estimating the effect of *Gender* on predicted strength. Doing so introduces prediction error, however, and as demonstrated by Kristen Altenburger and Daniel Ho, this error can be especially problematic

[143] *See, e.g.*, Elizabeth A.Freeman & Gretchen G.Moisen, *A Comparison of the Performance of Threshold Criteria for Binary Classification in Terms of Predicted Prevalence and Kappa*, 217 ECOLOGICAL MODELING 48 (2008) (reviewing criteria for establishing cutoffs in ecological forecasting).

[144] This approach to de-biasing proxy variables has been advanced by several scholars. See Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AM. ECON. J. 206, 206 (2011); Crystal Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, John M. Olin Center For Law, Economics, and Business Discussion Paper No. 1019 (October 2019). We provide an example of this approach, as well as its limitations, in the Appendix.

[145] See, *e.g.*, Pope & Sydnor, supra note 144, at 212.

when the approach is deployed in common machine-learning models.[146] More troublesome, these prediction errors can themselves be systematically biased against members of a protected group who are otherwise qualified in the target. We illustrate this challenge in the Appendix, which presents a simple example showing that this "de-biasing" procedure may actually have almost no effect on the extent of bias in the final outcome.

These considerations reinforce our conclusion that any variable that fails our test should be excluded from a decision-making model. While this approach risks sacrificing some degree of predictive accuracy in favor of ensuring an unbiased decision-making process, our discussion in Part 2(C) illustrates that U.S. antidiscrimination law has long made this trade-off. Additionally, a rule of exclusion also creates obvious incentives to seek out observable variables that can more accurately capture the target variable of interest, consistent with the holding of *Dothard* that the prison should adopt a test that more directly measured applicant's strength.[147] Indeed, in the machine learning context, this history of U.S. antidiscrimination law provides an independent reason to adhere to a rule of exclusion given the capacity of machine-learning processes to analyze an ever-increasing volume of data to identify proxy variables that enhance accuracy while remaining unbiased with respect to a protected classification.

C. Challenges in Implementing the Test

Implementing the IAT faces several challenges, which we list below and then discuss in the context of the hiring test ($Height_i = \alpha \cdot Strength_i + \varepsilon_i$), where the target variable is $Strength$.

i. Unobservability of the Target Variable

The problem of an unobservable target variable of interest is always the starting point for constructing an algorithm to screen an applicant (or make some other decision), since the motivation for using statistical inference in the first place is the challenge of measuring unobservable characteristics such as creditworthiness, productivity, longevity, or threat to public safety.[148] In designing a machine-learning algorithm, the need to solve this problem arises in the training procedure, where data on a target variable are required to determine which features predict the target. In practice, the solution is to turn to historical data, which can be used to train the predictive model.[149] In the

[146] *See* Kristen M. Altenburger and Daniel Ho, *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions*, 175 JOURNAL OF INSTITUTIONAL AND THEORETICAL ECONOMICS 98 109-118 (2018).

[147] *See supra* note 138.

[148] *See* Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan and Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEG. ANALYSIS 113, 132 (2019) ("One way to think about the goal of prediction is to overcome a missing information problem.").

[149] *Id.*

employment setting, for instance, an employer seeking to predict the future productivity of job applicants could train a model with data concerning the productivity of existing employees along with data concerning the characteristics of these employees at the time of application. The data may suffer from selection bias given that the employer will not observe applicants who were not hired, which is why in both training a model and in running the IAT, one must be attendant to measurement error—a point we discuss in subsection (ii).

Nonetheless, the threshold challenge for the IAT—that the target is unobservable—is in many ways one of transparency. That is, data concerning the target variable exist (after all, these data were required to train the model), but they may not necessarily be available. As Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass Sunstein emphasize, transparency in the training data is therefore an important step in ensuring the ability to evaluate whether algorithmic decision-making facilitates discrimination.[150] We agree. The ability to examine the training data used in designing a model would allow a regulator, litigant or data scientist to conduct the empirical test we describe in this section. In the UnitedHealth example discussed in the Introduction, one could apply the IAT using actual morbidity data to assess whether the substitute measure of the target—the cost of healthcare—has a discriminatory effect. Indeed, the availability of actual morbidity data was what enabled researchers to quantify the racial bias in *Science*.[151]

Even with data on the target variable of interest, however, this last example highlights the problem of measurement error: Do the data on the target measure what they purport to measure with error? We address this problem in the following subsection.

ii. Measurement Error in the Target

In addressing the unobservability problem of the target, one can inadvertently mis-measure it. This challenge of measurement error—or what is alternatively referred to as "label bias"[152]—has been studied in the computer science and economics literatures, providing useful guidance for addressing it when applying our test.[153]

Consider, for instance, judicial bail decisions where data scientists have used past judicial bail decisions to train algorithms to decide whether a

[150] *Id*. at 114 (arguing that harnessing the benefits of algorithmic decision-making while avoiding the risk of discrimination "will only be realized if policy changes are adopted, such as the requirement that all the components of an algorithm (including the training data) must be stored and made available for examination and experimentation").

[151] Obermeyer, et al., *supra* note 7, at 447 ("Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise.").

[152] Corbett-Davies & Goel, *supra* note 75, at 3.

[153] See *id*. at 17-18.

defendant should be released on bail pending trial.[154] In many states, judges are required to consider the risk that a defendant poses for public safety, and in training the model, the target variable is often defined to be whether a defendant who was released was later arrested prior to the trial.[155] However, heavier policing in minority neighborhoods might lead to minority defendants being arrested more often than non-minorities who commit the same offense.[156] Consequently, Sam Corbett-Davies and Sharad Goel have warned that this form of label bias risks causing a model to estimate a positive relationship between a defendant's race (and correlates of race) and whether the defendant poses a risk to public safety, simply due to the correlation of race with measurement error.[157]

Likewise, as Jon Kleinberg and others have noted, an employer who seeks to measure employee productivity through the number of hours that an employee spends at work will likely be using a biased measure of productivity if there are gender differences in how efficiently an employee works at the office (for example, to attend to childcare obligations before or after work).[158] Similar to the bail example, this form of label bias is problematic because the measurement error may be correlated with a protected characteristic, in this case, gender.

These examples illustrate a more general point, which is that measurement error in a target variable will create discriminatory bias when the measurement error is correlated with membership in a protected group. This result occurs because a statistical model that seeks to estimate the predictors of a true target y that is mis-measured as $y + \mu$ will inevitably discover that the protected classification (and any correlate of it) predicts the level of the mis-measured target.

For similar reasons, when measurement error in a target variable is correlated with a protected classification, application of our test may fail to detect this bias. Returning to the *Dothard* example, imagine that we applied the IAT to *Height* as before, but we use a measure for strength, *Strength**, that has measurement error μ that is correlated with gender. Formally, the test would be:

$$Height_i = \alpha \cdot Strength_i^* + \varepsilon_i$$

which is equivalent to:

$$Height_i = \alpha \cdot (Strength_i + \mu_i) + \varepsilon_i$$

[154] *See, e.g.*, Berk et al., *supra* note 79, at 31-33.
[155] *Id.* at 31.
[156] Corbett-Davies & Goel, *supra* note 75, at 18.
[157] *Id.*
[158] *See* Kleinberg et al., *supra* note 148, at 139.

In such a setting, the IAT may fail to reveal that the errors (ε_i) are correlated with the protected classification of gender. The reason is because the unexplained variation between "true" *Strength* and *Height* is ($\mu_i + \varepsilon_i$), but the IAT will not be able to detect how gender is correlated with μ_i because it is part of *Strength**, the mis-measured target. In short, measurement error in a target variable is a critical issue to consider regardless of whether one is calibrating a model or running our test.

Recognition of this latter point is implicit in Kleinberg, Ludwig, Mullainathan, and Sunstein's argument for making training datasets transparent. Often, the data for a target will reveal fairly obvious risks that the measurement error is biased with respect to a protected classification (such as the example cited earlier when an employer uses hours-worked as a measure for productivity). At the same time, other instances when this problem arises may be less obvious. In the example we provided in the Introduction, that of UnitedHealth, it may not have been immediately obvious that patient costs—the substitute measure of the target of interest—had measurement error for the true target (severity of illness) that was correlated with race. Yet this correlated measurement error was nevertheless revealed when researchers used an alternative estimate for severity of illness.

This last example thus underscores the need to run the IAT with alternative measures of the target which may reveal problematic measurement error in the primary target data. Moreover, opening up the possibility of running the IAT with alternative measures of the target variable should also encourage the use of theory-based models of target characteristics. Theory-based estimates of target variables may be especially valuable in addressing the measurement error that arises from estimating targets based on binary outcomes. Common approaches to estimating target variables often rely on estimating a predictive model based on a binary outcome variable, such as whether a borrower defaults on a loan or whether a defendant who was released on bail later commits a crime prior to trial. Yet estimating unobservable characteristics such as "creditworthiness" or "risk" based on these binary behaviors necessarily implicates the risk of measurement error in the true target of interest.

Consider, for instance, a model that seeks to predict creditworthiness based solely on whether a borrower defaults in the training data. By construction, the training dataset consists only of those borrowers who received a loan; borrowers who do not get a loan provide no information. Thus, it is infeasible to estimate actual creditworthiness within the broader group of all applicants. This is the "selective labels" problem that has been studied in the computer science and economics literatures.[159] The literature on selective labels in training a model has suggested a process of interventions to correct the misestimations.[160] Another approach would be to implement the IAT through a structural estimation of theoretic representations of the target business necessity.[161]

Another version of the problem of measurement error comes in the context of threshold analysis. In our example, the prison asserted that it needed a minimum required level of strength. As a result, the target was not the continuous variable of strength, but the applicant possessing a strength level of at least 60, which we assumed was a legitimate business necessity threshold for a prison guard job. But what if the level of strength needed is not obvious? What if the prison erroneously thought the true level of required strength was 80? We previously referred to this setting as a mis-asserted target threshold. Cases such as *Lanning v. Southeastern Pennsylvania Transporation Authority* underscore the potential for these target thresholds to be mis-asserted in a way that results in intentional discrimination, such as when they are purposefully set at a level that will adversely affect members of a protected group.

In Figure 5, we assume that, as in Figure 3, the prison implements a physical exam that perfectly measures actual strength. If the prison mistakenly sets the minimum required strength threshold at 80 (the dashed line), the resulting problem is that more women cluster in the just-failed space (between the dashed and straight line), which is the region of between the mis-asserted target threshold relative to the true required strength level. In fact, if an employer did not want to hire women, it could intentionally

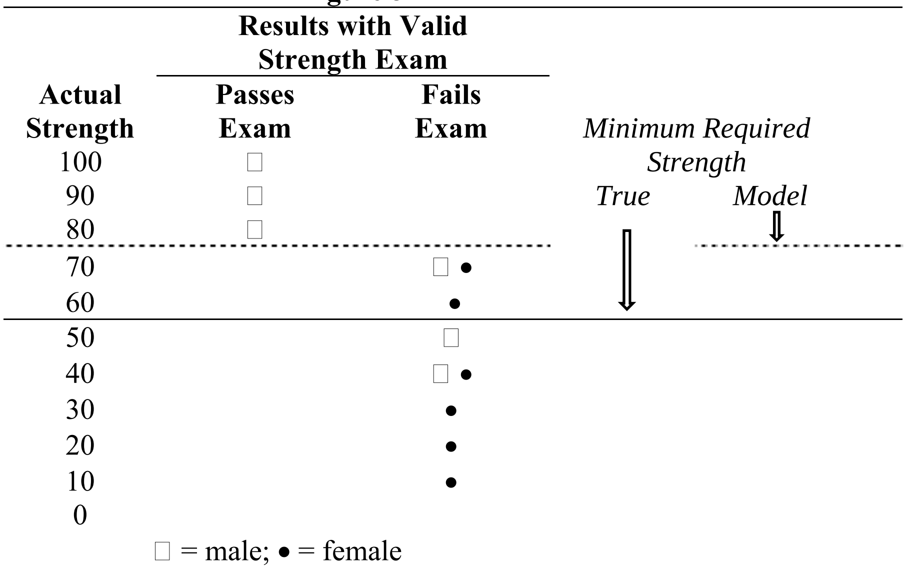
[159] *See* Himabindu Lakkaraju, et al., *The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables*, in KDD Conference Proceedings, 2017; Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q. J. ECON. 237, 256- (2018).

[160] *See, e.g.,* Maria De-Arteaga, et al., *Learning Under Selective Labels in the Presence of Expert Consistency*, (July 4, 2018), https://arxiv.org/abs/1807.00905v1 (proposing a data augmentation approach that can be used to leverage expert consistency to mitigate the partial blindness that results from selective labels).

[161] For instance, consider a credit scoring algorithm that predicts credit risk based on default rates for loans that were previously extended to a group of borrowers. A model built using these target data (i.e., whether or not a borrower defaults) suffers from bias insofar as it only includes default data for loans that were approved by a lender. This selective labels problem can result in bias if the human decision-maker who approved the loans based the approval decision on borrower characteristics that were observable to the loan officer but are unobservable to the data scientist because they do not appear in the dataset. Imagine, for instance, that a loan officer records data on a loan applicant's occupation and, for low-paying occupations, the loan officer also evaluates informally an applicant's attire, which the officer believes is associated with creditworthiness. Assume the loan officer approves loans to well-dressed applicants in occupations that would otherwise make them ineligible for a loan and that these applicants are, in fact, more creditworthy than their occupation would suggest. Training a predictive model using only default data and occupation at the time of application would therefore suggest to the model that "high risk" occupations are actually more creditworthy than they are because they default infrequently. Moreover, given racial, ethnic and gender differences in the composition of certain occupations, this model would likely be biased in addition to being inaccurate. However, evidence of this bias would become apparent in applying the IAT if one were to run the test using an estimate for creditworthiness that was based on borrowers' cash flow data as opposed to default data.

implement a mis-asserted target, knowing that more women would be excluded.

Figure 5



In this setting, the exam would pass the IAT insofar that it was unbiased with respect to gender in predicting whether an applicant had strength of at least 80. However, the employer's use of the exam would nevertheless fail our definition of accountability set forth in Part 2 because the employer has set the cut-off at a level where qualified females are systematically exluded from the position. As emphasized in *Lanning,* this example underscores the importance of supplementing the IAT with the ability to scrutinze whether a classification threshold has been set at a level that is justified by actual business necessity.

iii. Testing for "Not Statistically Correlated"

The third challenge concerns how to reject the null hypothesis that no correlation exists between a set of proxy variable residuals and a protected category. In our *Dothard* illustration, the use of *Height* as a proxy for *Strength* would pass the IAT if the unexplained variation between *Strength* and *Height* (ε_i) is uncorrelated with *Gender*, as given by the test:

Regression:	$\varepsilon_i = \beta_0 + \beta_1 Gender_i$
Null Hypothesis:	$\beta_1 = 0.$

The tradition in courts and elsewhere is to use a statistical significance level of 0.05;[162] i.e., we are willing to allow for a 5% probability of making the "Type I" error of rejecting the null hypothesis ($\beta_1 = 0$) by chance, when it is actually true. A related concept is the p-value of an estimate: the probability of obtaining an estimate for β_1 at least as far from zero as the value estimated, assuming the null hypothesis is true. If the p-value is smaller than the statistical significance level, one rejects the null hypothesis.

However, a problem with focusing on p-values is that as the sample size grows increasingly large, realized p-values converge to zero if the sample estimate for β_1 is even trivially different from the null. This is because as the sample size grows larger, the uncertainty of our estimates (usually measured by their "standard error") gets closer and closer to zero, causing any coefficient (even magnitude-irrelevant ones) to look different from an exact null of $\beta_1 = 0$ in a p-value test. In particular, a company that brings a large dataset to bear on an IAT test might be disadvantaged relative to firms with less data.

The source of the problem is the fact that in any statistical test we are actually trading off the probabilities of making two different errors: Type I errors (when we wrongly reject the null when it is, in fact, true) and Type II errors (when we wrongly fail to reject the null when it is, in fact, false). The "significance level" of a test is the probability of making a Type I error. Keeping this fixed (e.g., at 5%) as the sample size increases means that we are keeping the probability of a Type I error fixed. But at the same time, again because the standard error of our estimates is going to zero as the sample size gets large, the probability of a Type II error is actually converging to zero. If we care about both types of error, it makes sense to reduce the probability of *both* as the sample size increases, rather than fixing the probability of Type I errors and letting that of Type II errors go to zero. This point has been made forcefully by many authors, especially Edward Leamer, and a number of solutions have been proposed for adjusting the significance level as the sample size increases.[163] A full consideration of these different approaches is

[162] *See, e.g.,* Karen A. Gottlieb, *What Are Statistical Significance and Probability Values?* 1 TOXIC TORTS PRAC. GUIDE § 4:10 (2019)("Through a half century of custom, the value of 0.05 or 1 in 20 has come to be accepted as the de facto boundary between those situations for which chance is a reasonable explanation (probabilities > 0.05) and those situations for which some alternative is a reasonable explanation (probabilities < 0.05)."); *see also* Eastland v. Tennessee Valley Authority, 704 F.2d 613, 622 n. 12 (1983) (in employment discrimination lawsuit, noting that "a probability level of .05 is accepted as statistically significant" in determining whether racial disparities in pay were statistically significant).

[163] *See, e.g.,* Edward Leamer, SPECIFICATION SEARCHES: AD HOC INFERENCE WITH NONEXPERIMENTAL DATA (1978) (proposing p-value adjustment to minimize error losses associated with Type I and Type II error); I.J. Good*, Standardized Tail-Area Prosabilities*, 16 JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION 65 (1982) (proposing p-value adjustment based on a "Bayes/non-Bayes compromise"); Mingfeng Lin, Henry C. Lucas, Jr., and Galit Shmueli, *Too Big to Fail: Large Samples and the p-Value Problem,* 24 INFORMATION SYSTEMS RESEARCH 906, 908-915 (2013) (surveying approaches to adjusting p-values in large samples and recommending the reporting of effect sizes and confidence intervals and using coefficient/p-value/sample-size plots for interpreting the data along with Monte Carlo simulations); Eugene Demidenko, *The p-value You Can't Buy*, 70 THE AMERICAN STATISTICIAN 33, 34-37 (2016) (proposing use of d-values for assessing statistical inference in large datasets).

beyond the scope of this Article; however, we provide below an example of one such approach to illustrate how it can be utilized to discern when a seemingly significant result when applying the IAT is actually a function of the large sample size and not evidence of a discriminatory proxy variable.

iv. Nonlinearities or Interactions Among Proxies

Machine learning models are often focused on forming predictions based on nonlinear functions of multiple variables. In introducing the IAT, our specification focused on linear settings, but the IAT could in principle be amended to handle nonlinear models as well. For example, rather than just running the test regression once, we could run it repeatedly, with each of a set of basis functions of the explanatory variables on the left-hand side. Full consideration of this topic is beyond the scope of this Article, but in general, implementation of the IAT could be made part of the type of feature selection and feature analysis protocols that are used in practice with both linear and non-linear machine-learning processes.[164]

D. Simulation

To illustrate how the concerns of discrimination enter though proxy variables, we simulate the setting in *Dothard* of hiring a prison worker.

i. Set Up

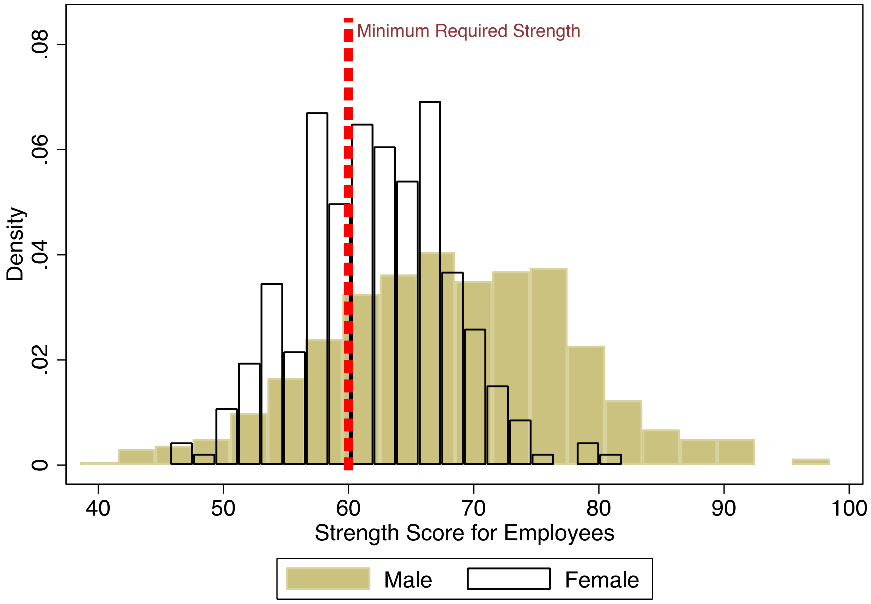
The simulation assumes that the prison has historical records for 800 employees, of which roughly one-third are female (n=256) and two-thirds are males (n=544). We further assume that the prison uses these historical records to develop a sorting algorithm for considering a pool of 1,200 applicants. The 800 employees are endowed with an *unobservable* strength level, which we model as a random variable distributed normally with (i) a mean of 68 and a standard deviation of 10 for male employees and (ii) a mean of 62 and a standard deviation of 6 for female employees. With these modeling assumptions, females have lower mean strength but a smaller standard

[164] In particular, a related literature in computer science focuses on feature selection to enhance model interpretability. *See* Datta et al. *Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems*, *Proceedings of IEEE Symposium on Security & Privacy 2016*, 598–617, 2016 (proposing a quantitative-input-influence (QII) protocol based upon Shapley values to determine the importance of features and clustering metrics to summarize feature influence); *see also* Phillipe Bracke et al., *Machine learning explainability in default risk analysis*, Bank of England Staff Working Paper No. 816 (June 5, 2019) (implementing QII method in predicting mortgage defaults). More formally, Lundberg, et al., *Consistent Individualized Feature Attribution for Tree Ensembles*, arXiv:1802.03888v3 [cs.LG], March 7, 2019 and Merrill et al., *Generalized Integrated Gradients: A practical method for explaining diverse ensembles*," ArXiv 2019, build upon game-theoretic SHAP (Shapley Additive explanation) values and propose new feature credit-assignment algorithms that can handle a broad class of predictive functions with both piecewise-constant (tree-based), continuous (neural-network or radial-basis-function based), and mixed models.

deviation, as plotted below in Figure 6. To be an effective prison guard requires a strength of 60, the business necessity. Hiring is not perfectly effective at sorting which guards will meet this threshold; therefore, even among the employees, there are guards who fall below the required strength for the job. For now, we assume that the prison can implement a costly physical exam to measure true strength for these employees. (We abstract from other aspects of effectiveness such as psychological and managerial skills needed for prison-guard work.)

We assume the strength of applicants is likewise distributed randomly. However, for obvious reasons, the applicant pool has not been previously selected for strength as employees have. Therefore, we model strength across applicants as a random variable distributed normally with a mean of 50 and a standard deviation of 10 for male employees and a mean of 44 and a standard deviation of 6 for female employees.

Figure 6



The prison managers cannot directly observe applicants' strength, and, as noted, implementing a full physical exam across applicants is costly. Therefore, the prison decides to use height as a proxy variable for an applicant's strength, since it is easily measured on applications. We model height as a sum of a baseline 50 inches (with a normally-distributed error of 4 inches) plus a concave (quadratic) function increasing in strength. Female height has the same relation to strength but a ten percent lower baseline. The resulting mean height in the employee training dataset is 5’10” with a standard deviation of 5”.

Finally, as in *Dothard*, the prison seeks to filter applicants by imposing a minimum height requirement. To determine the height cut-off, the prison runs a classification analysis. In doing so, the prison determines that they want to ascertain that an individual will be above the strength threshold with an 80% certainty, i.e., they want only a 20% risk of incorrectly classifying an applicant as eligible for hiring (above the strength threshold of 60) when the person in fact has a strength of less than 60. Based on the height and strength of the prison employees, this results in a 5’10” cut-off. The prison applies this cut-off to all 1,200 applicants.

Among the 370 female applicants, 344 (93%) fail the height test. In contrast, among the 830 male applicants, 504 (61%) fail the height test. These disparities suggest that the height cut-off may discriminate against females applicants, but we cannot definitively conclude this from the high rejection rates because, as we saw in Figure 6, females in our samples have lower strength than males on average.

ii. Applying the Input Accountability Test

Assume that in advance of deploying the height test, the prison instead decides to conduct the IAT to ensure that any disparities in hiring would be based on differences in predicted applicant strength. Table 1 presents the results from the test. To run the IAT, the prison would return to the training data it possesses regarding its employees’ actual strength and height that it used to determine the 5’10” cut-off. In panel A, we present the first step of regressing the proxy variable on employee strength, the target of interest. Because the prison is focused on using a cutoff for height, we estimate a logistic regression of whether an employee passes the height cut-off as a function of the employee’s strength. (To do so, we use as our dependent variable an indicator variable that equals 1 for employees that are at least 5’10” and 0 for all others.) Note that this indicator variable is on the left-hand side of the regression (and not strength) because we want to decompose whether an employee meets the height cut-off into two components – the part that can be predicted from an employee’s strength and the part that cannot be predicted from an employee’s strength (the “residual”). Stated differently, logistic regression effectively estimates the probability that an employee is 5’10” based on employee strength. Therefore, the residual, which is equal to one minus this predicted probability for each employee, can be viewed as the variation in whether an employee meets the height threshold of 5’10” that is unrelated to an employee’s strength. In panel B, we present the results from regressing the residual from panel A onto the indicator variable for female.

Table 1

	(1)	(2)	(3)	(4)	(5)
Panel A: First Step of IAT (DV=Column Heading)					
	Cut-Off Height	Cut-Off Muscle Mass	Muscle Mass	Job Performance	Cut-Off Muscle Mass
Strength	0.0206***	0.0377***	0.9965***		0.0387***
	[0.00155]	[0.000747]	[0.0191]		[0.0000138]
Performance Score				0.675***	
				[0.0307]	
Observations	800	800	800	800	2,000,000
[Pseudo] R-squared	0.111	0.466	0.772	0.376	0.496
Panel B: Second Step of IAT (DV=Residuals from Step 1)					
Female	-0.354***	-0.013265	-0.3552	-8.858***	-0.0013***
	[0.0327]	[0.02625]	[0.379]	[0.542]	[0.000505]
Observations	800	800	800	800	2,000,000
R-squared	0.128	0	0	0.25	0
d-value					50%

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Panel A of Column (1) reports that strength only accounts for a small part of the variation (R-squared = 0.111) for whether an employee is (or is not) taller than 5’10”. In Panel B, our column (1) results show that the residual of the first step regression has a negative, significant correlation with gender, thus failing the IAT. Females incur a penalty because the proxy variable for the business necessity of required strength has residual correlation with gender.

Imagine that the prison realizes this flaw in using a height cut-off and decides instead to consider incurring an extra cost for doing a muscle-mass index evaluation of applicants. Because the evaluation is imperfect in assessing true strength, we assume that the results of a muscle-mass index evaluation is equal to an individual’s strength plus random noise.[165] To implement this screening procedure, the prison first applies the muscle-mass index evaluation to existing employees so that it can estimate the minimum muscle mass an individual should have to be above the minimum strength

[165] We model the random noise as a randomly distributed variable with a mean of zero and a standard deviation of 5.

threshold with an 80% certainty. The classification analysis produces a muscle-mass cut-off score of 64. As above, the prison then conducts the IAT.

In column (2) of panel A we present the results of the IAT for the muscle-mass index evaluation based on the employee training data. To implement the IAT, we run the same regressions that we used for testing the height cut-off, but we substitute an indicator variable for whether an employee has a muscle mass of at least 64 for the indicator variable for whether an employee is at least 5'10". In panel A, column (2) shows that the probability that an employee has a muscle mass of at least 64 is (unsurprisingly) related to an employee's strength, resulting in a much larger R-squared. Importantly, the residual should not fail the IAT, because it has no bias against females. In column (2) of panel B, we see that this is indeed the case; the coefficient on female is statistically insignificant and small in magnitude.

In column 3, we instead consider a continuous variable version of muscle mass as a scoring variable rather than a cut-off version of the indicator variable. Perhaps the underlying job-required strength is not a threshold but a strength score that will feed into wage-setting or other profiling of individuals that focus on continuous rather than discrete measures. To implement the IAT in this context, we use the same training data that was used for column (2) of Table 1; however, the regression specification for the first step takes the form of a linear regression of employees' muscle mass scores on their measured strength. As in column (2), column (3) shows that muscle mass is a legitimate business necessity variable. In panel A, we find that muscle mass and strength are very correlated, with strength accounting for almost 80% of the variation in muscle mass. Column (3) of panel B shows that muscle mass again passes the IAT: the residual is uncorrelated with the female indicator variable.

In the final two columns of Table 1, we demonstrate the importance of the challenges we introduced in Part 3(C).

First, we use column (4) to illustrate the concern about measurement error in the target (strength). Thus far, we have been working under the assumption that the prison can take an accurate measurement via a physical exam of the training dataset employees. However, what if instead the prison cannot measure actual strength but uses a job performance assessment made by a manager. (We label this job performance measure an employee's "Performance Score"). As noted above, a central challenge in real world settings is that target variables used to train predictive models are typically estimated in this fashion and may contain measurement error that is correlated with a protected characteristic. We therefore simulate an employee's Performance Score as biased against females.[166] In this regard, the simulation

[166] In particular, for males, we model the job performance measure as strength plus random noise; however, for females, we model job performance as concave in strength (like the height variable)—a quadratic

replicates the same problem illustrated with the UnitedHealth example (where the illness severity measure was inadvertently biased against African Americans).

In addition to employees' Performance Scores, assume that the prison also has at its disposal data from the muscle measure index evaluation used in columns (2) and (3). Even without perfect data regarding employee strength, the prison can still use these data with the IAT to evaluate whether its preferred estimate of the target (an employee's Performance Score) suffers from bias. To implement this test, we treat muscle mass as an alternative measure of the target of interest (strength), and we treat the Performance Score as a proxy for strength. Accordingly, the first step of the IAT is conducted by regressing employees' Performance Scores on the muscle mass evaluation data. The results are shown in column (4) of panel A. Not surprisingly, an employee's muscle mass is closely related to an employee's Performance Score. In column (4) of panel B, we show the results of regressing the residuals from this regression on the gender variable. As shown in the table, Job Performance fails the IAT. In this fashion, the IAT can be used to test whether an estimate for a target suffers from biased measurement error, so long as one has an alternative estimate for the target (even a noisy one) that is believed to be unbiased.

The final column in Table 2 illustrates the concern of large data samples. For this column, we implement the same muscle mass test as in column (2), except that we randomly draw 2 million employees for the training dataset rather than 800 employees. (For all 2 million employees, we model their strength using the same assumptions used for the original 800 employees). For each employee, we likewise calculate muscle mass as employee strength plus a random variable distributed normally with a mean of 0 and a standard deviation of 5. Thus, in our simulated setting, muscle mass is a noisy estimate of employee strength but it has zero bias with respect to gender. Even so, however, the possibility remains that in drawing random measurement error for our sample, very slight differences may exist by chance between the average measurement error of females and males. (This is equivalent to observing that even if a coin is unbiased, it may still return more than 50% heads in a trial of 100 flips). Moreover, as we described in section 3, the p-value may converge to 0 for any small deviation, as sample sizes approach infinity. Thus, even a small (economically non-meaningful) correlation may look significant. This would create a setting of a large-dataset proxy variable failing the IAT, not because of a fundamental problem, but just because of the use of a fixed p-value. This is what we have modeled in column (5). The coefficient on female in column (5) is very small (-0.0013) but statistically

concave function of strength plus random noise. The managers evaluating females do not fairly evaluate them, especially for the stronger females.

significant, notwithstanding the fact that we modeled measurement error from a distribution that had exactly zero gender bias.

As noted in subsection 3(C)(iii), where the IAT is applied to a large dataset, it is therefore critical to check whether a proxy that fails the IAT might have failed the test simply because of the large number of observations in the sample. That the seemingly statistical finding in column (5) may be an artifact of a trivial difference within a large dataset can initially be seen by the fact that the R-squared in column (5) is 0%; if effectively no variation in the residuals can be explained by gender, how can it be that this proxy is penalizing females in a systematic fashion? Additionally, as noted previously, a number of formal solutions also exist to examine this issue more fully. Here, we illustrate one such approach using the concept of the "d-value" proposed by Eugene Demidenko.[167] Rather than focus on a comparison of group means, the d-value is designed to examine how a randomly chosen female fared under this proxy variable relative to a randomly chosen male. Specifically, in the context of the IAT, the d-value answers the question "what is the probability that members of a protected group are being penalized by the proxy?" As shown in the last row of column (5), the d-value is approximately 50%, indicating that the probability that females are penalized by the use of a muscle-mass proxy is effectively a coin-toss; that is, there is no evidence that female applicants are being systematically penalized by the use of this proxy.

This finding, of course, is hardly a surprise given that we designed the simulation to ensure that it was an unbiased proxy. In this fashion, the use of a d-value can highlight when a seemingly significant finding is a function of the large sample size and not evidence of a discriminatory proxy variable.[168]

IV. APPLICATIONS BEYOND EMPLOYMENT

The fact that the IAT is rooted in general antidiscrimination principles makes it applicable to any setting where a decision-maker relies on statistical discrimination, regardless of whether conducted by humans or algorithms.

[167] *See* Demidenko, *supra* note 163.

[168] To the extent one utilizes the d-value in this fashion, a natural question is what level of a d-value would constitute evidence of a discriminatory proxy. Given that the d-value answers the question "what is the probability that members of a protected group are being penalized by the proxy?", any result that yields a d-value deviating from 50% would presumably be evidence of a discriminatory proxy, allowing for a percentage difference to incorporate a far tail sampling draw. This conclusion follows from the conventional judicial reliance to on p-values, which likewise assumes that any finding with a p-value of less than 0.05 is evidence of discrimination. That said, in adopting such an approach, it would be important to utilize a d-value analysis only upon a finding that a proxy fails the IAT using a conventional statistical test. The reason stems from the fact that in smaller samples, even an unbiased proxy could result in a d-value that is slightly different from 50% due sample variance. For example, the d-value for column (3) is just slightly less than 51%, despite the fact that muscle mass is modeled as an unbiased proxy. However, running the same simulation with 50,000 observations produces a d-value of 50%.

Central to our argument is the idea of using a test to ascertain adherence to business necessity targets when designing a decision-making process. Indeed, even the Equal Employment Opportunity Commission subscribes to a business necessity *test* in its Uniform Guidelines on Employee Selection Procedures, stating that: "[e]vidence of the validity of a test or other selection procedure by a criterion-related validity study should consist of empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance."[169] Note, however, that even the EEOC's validity test looks only to the predictive capacity of an employment exam. But as we have emphasized throughout this Article, a simple correlation test leaves open the possibility that a test will penalize members of a protected group who are, in fact, qualified in the job-related skill in question. This, of course, was the lesson of *Dothard* and the cases examined in Part 2. In this regard, a simple method to remedy this defect when conducting a criterion-related validity study would be to incorporate the IAT.

In this section, we discuss additional implementations outside of the employment setting. We first focus on settings where a decision-maker can face liability for claims of unintentional discrimination and where a court or legislature has expressly considered what constitutes a legitimate business necessity target. We then address the application of the IAT in settings where formal liability for claims of disparate impact or other claims of unintentional discrimination are currently less clear, but where firms can use the IAT to self-regulate. Finally, given the latitude firms have to set their own business necessity targets, we conclude with an admonition that firms must be vigilant in monitoring whether a purported target is, in fact, a legitimate one to use.

A. Domains with Court-Defined Business Necessity Targets

Consider, for instance, a regulator tasked with evaluating a decision-making algorithm in one of the following domains where claims of unintentional discrimination may be possible, and where courts have expressly defined a legitimate "target" variable that can justify unintended disparities that vary across protected and unprotected groups:

Table 2

Domain:	**Legitimate Target Variable:**
Credit Determinations	Creditworthiness[170]

[169] 29 C.F.R. § 1607.5B.

[170] *See* A.B. & S. Auto Service, Inc. v. South Shore Bank of Chicago, 962 F. Supp. 1056 (N.D. Ill. 1997) ("[In a disparate impact claim under the ECOA], once the plaintiff has made the prima facie case, the defendant-lender must demonstrate that any policy, procedure, or practice has a manifest relationship to

Home Insurance Pricing	Risk of Loss[171]
Parole Determinations	Threat to Public Safety[172]
Tenant Selection	Ability to meet lease obligations,[173] pay rent,[174] and resident safety[175]
Post-Secondary School Admission	Predicted academic success[176]
Selection into Special Education	Educational ability[177]
State Merit Scholarship Eligibility	Academic achievement in high school[178]

Just as employers are permitted to make hiring decisions based on the legitimate target variable of a job-required skill, courts in these settings have likewise determined that decision-making outcomes can lawfully vary across protected and unprotected groups only if decisions are based on the target variable noted in Table 2.

the creditworthiness of the applicant…"); *see also* Lewis v. ACB Business Services, Inc., 135 F.3d 389, 406 (6th Cir. 1998)("The [ECOA] was only intended to prohibit credit determinations based on 'characteristics unrelated to creditworthiness.'"); Miller v. Countrywide Bank, NA, 571 F.Supp.2d 251, 258 (D. Mass 2008)(rejecting argument that discrimination in loan terms among African American and white borrowers was justified as the result of competitive "market forces," noting that prior courts had rejected the "market forces" argument insofar that it would allow the pricing of consumer loans to be "based on subjective criteria beyond creditworthiness.")

[171] *See, e.g.,* Owens v. Nationwide Mut. Ins. Co., No. Civ. 3:03-CV-1184-H, 2005 WL 1837959, at *9 (N.D. Tex. Aug. 2, 2005)(minimizing the "risk of loss in homeowner's insurance" was a legitimate business necessity under the Fair Housing Act that justified the use of facially neutral policy of using credit to determine eligibility for homeowner's insurance).

[172] *See, e.g.,* CAL. PENAL. CODE § 3041 (West 2017) (The Board of Prison Term "shall grant parole to an inmate unless it determines that the gravity of the current convicted offense or offenses, or the timing and gravity of current or past convicted offense or offenses, is such that consideration of the public safety requires a more lengthy period of incarceration for this individual."); *see also* Smith v. Sisto, 2009 WL 3294860 at *6 (E.D. Cal. Oct. 13, 2009) (denying claim that denial of parole constituted discrimination and concluding that "[t]he need to ensure public safety provides the rational basis for section 3041)").

[173] *See* 24 C.F.R. § 100.202(c)(1) (permitting under the FHA a landlord's "[i]nquiry into an applicant's ability to meet the requirements of ownership or tenancy").

[174] *See* Ryan v. Ramsey, 936 F.Supp. 417 (S.D.Texas 1996)(noting that under the FHA, "there is no requirement that welfare recipients, or any other individuals, secure apartments without regard to their ability to pay.")

[175] *See* Evans v. UDR, Inc., 644 F.Supp.2d 675, 683 (2009) (permitting landlord to reject tenant based on prior criminal history as the "policy against renting to individuals with criminal histories is thus based concerns for the safety of other residents of the apartment complex and their property").

[176] *See* Kamps v. Baylor University, 592 F. App'x 282 (5th Cir. 2014) (rejecting age discrimination case based on law school admissions criteria that relied on applicant's grade point average (GPA) because GPA is quantitative predictor of academic success in law school and thus a "a reasonable factor other than age").

[177] *See* Ga. State Conf. of Branches of NAACP v. Georgia, 775 F.2d 1403, 1420 (11th Cir. 1985) (finding that, in Title VI case alleging that school district achievement grouping caused disparate impact on minority students, school district's effort to classify students based on assessment of ability was justified because it bore "a manifest demonstrable relationship to classroom education").

[178] *See* Sharif by Salahuddin v. New York State Educ. Dept., 709 F. Supp. 345, 362 (SDNY 1989) (finding that state's use of SAT scores did not have a "manifest relationship … [to] recognition and award of academic achievement in high school" in Title IX claim of disparate impact alleging that state's use of SAT scores to determine student eligibility for merit scholarships had a discriminatory effect on women).

In applying the IAT in these settings, the regulator's task thus follows the same process noted in Part 3. First, the regulator must evaluate whether the decision-making process does, in fact, seek to produce outcomes based on the legitimate target variable. Second, using historical data for both the target variable and the model's full set of features, the regulator would then apply the IAT to each feature used in the model. Finally, any feature that failed the test would be required to be excluded from the model.

B. Domains Without Court-Defined Business Necessity Targets

The IAT is equally applicable to domains where antidiscrimination laws do not formally regulate decision-making processes governing disparities across protected and unprotected groups or where the legal risk for unintentional discrimination is presently unclear. We provide an example of each.

The first domain concerns insurance outside the context of home insurance.[179] As Ronen Avraham, Kyle Logue, and Daniel Schwarcz show, a number of jurisdictions do not have any laws restricting providers of automobile or life insurance from discriminating on the basis of race, national origin, or religion.[180] Nor is there a federal antidiscrimination statute applicable to insurance outside of the context of home insurance.[181] Consequently, insurers likely have considerable discretion to rely on statistical discrimination to underwrite policies, which may produce unintended disparities across protected and unprotected groups. Yet evidence that racial disparities exist in the pricing of auto loans has routinely been met by the insurance industry with assurances that premiums are based on risk. For instance, following a nationwide study by the Consumer Federation of America in 2015 that found that predominantly African-American neighborhoods pay higher auto premiums,[182] the Property Casualty Insurers Association of America responded with a declaration that "Insurance rates are color-blind and solely based on risk."[183] Thus, insurers claim to self-regulate themselves by setting risk as the business necessity target. To the

[179] As noted in Table 2, discrimination in home insurance is governed by the FHA.

[180] *See* Ronen Avraham, Kyle D. Logue & Daniel Benjamin Schwarcz, *Understanding Insurance Anti-Discrimination Laws*, 87 S. Cal. L. Rev. 195, 239 (2014).

[181] *Id.* at 241. Additionally, the few cases alleging discrimination by insurance providers under 42 U.S.C. § 1981—a Reconstruction-era statute that prohibits racial discrimination in private contracting—have required a showing of intentional discrimination. *See, e.g.,* Amos v. Geico Corp. , 2008 WL 4425370 (U.S. Minn. 2008) ("To prevail under § 1981, plaintiffs must prove that GEICO intentionally discriminated against them on the basis of race.").

[182] Consumer Federation of America, High Price of Mandatory Auto Insurance in Predominantly African American Communities (2015), available at https://consumerfed.org/wp-content/uploads/2015/11/151118_insuranceinpredominantlyafricanamericancommunities_CFA.pdf.

[183] Press Release of American Property Casualty Insurers Association of America, *Auto Insurance Rates are Based on Cost Drivers, Not Race*, November 18, 2015, available at https://www.pciaa.net/pciwebsite/cms/content/viewpage?sitePageId=43349.

extent insurers are sincere in this claim, the IAT provides them with a ready test to ensure compliance.

An example in the second domain concerns disparities in medical treatment, as motivated by our example in the Introduction concerning UnitedHealth. Discrimination in healthcare provision is covered by Title VI of the Civil Rights Act of 1964, thus making it a more regulated setting than the insurance example. However, in *Alexander v. Sandoval*,[184] the U.S. Supreme Court held that Title VI does not provide for a private right of action to enforce disparate impact claims, greatly diminishing the risk that a provider of healthcare will face a claim of unintentional discrimination. Nonetheless, the UnitedHealth algorithm was designed to determine optimal medical treatment according to an individual's level of illness. Thus, one can presume that "level of illness" is a revealed business necessity target. Here, too, the IAT can provide healthcare providers such as UnitedHealth with a means to test the proxy variables utilized in predicting their target of interest.

C. Self-Determining Business Necessity

Regardless of whether an algorithm is based on complex machine-learned insights or on conventional physical exams, the IAT can serve as an important check for consistency with the principles undergirding U.S. antidiscrimination law across a number of decision-making domains. This tool is not simply a utility for courts to evaluate claims of discrimination, but a tool for regulators and self-regulating firms seeking to detect and avoid discrimination in the first place. Before closing, however, we emphasize two considerations. First, the fact that a proxy input variable is predictive of a business necessity target is not sufficient to rule out the possibility that it systematically penalizes members of a protected group who are actually qualified in the target. This is the principle behind the IAT. Second, although we have argued above that often businesses self-regulate themselves to determine business necessity targets (e.g., *risk* for insurers, *illness intensity* for healthcare providers), businesses must be ever vigilant that a purported target is a legitimate one to use. This is especially the case when working in a domain where courts have defined what can (and cannot) constitute a business necessity target.

A case in point comes from the credit markets, whereby lenders may have incentives to deploy predictive algorithms to estimate demand elasticities across different borrowers to engage in price discrimination. Price discrimination is made possible by the fact that certain borrowers are more prone to accept higher priced loans rather than engage in price shopping. These borrowers may not shop around for a host of reasons: They might live in financial desert locations of low competition, lack the knowledge to shop

[184] 532 U.S. 275 (2001).

for the best rate, need to transact in a hurry, have a historical discomfort with financial institutions due to prior discrimination, and/or have a history of being rejected for loans in the past. Empirical studies document that loan officers and mortgage brokers are aware of variation in borrowers' interest rate sensitivity and engage in price discrimination.[185]

A loan applicant's "price sensitivity" or "willingness to shop" may therefore be an additional unobserved characteristic that is of interest to a lender. Said another way, a lender's profit margin depends on both creditworthiness (the court-determined legitimate business necessity from Table 2) and shopping profiles. A lender might therefore design an algorithm that seeks to maximize profits by uncovering credit risk and shopping profiles. Furthermore, the lender (if lending were not in a formally-regulated domain) would argue that profits are legitimate business necessity. Yet, as noted in Table 2, lending is a domain where courts have expressly held that if a lending practice creates a disparate impact, "the defendant-lender must demonstrate that any policy, procedure, or practice has a manifest relationship to the creditworthiness of the applicant."[186] That is, while differences in creditworthiness can justify disparate outcomes in lending, differences in shopping behavior cannot.

The concern of algorithmic profiling for shopping behavior is of general concern because empirical evidence, again in lending, finds that profiling on lack-of-shopping almost certainly leads to higher loan prices for minority borrowers. For instance, Susan Woodward and Robert Hall[187] as well as Mark Cohen[188] find that adverse pricing for minority borrowers has generally been the rule when it comes to lenders engaging in price discrimination. In separate work,[189] we likewise find empirical evidence that, even after controlling for borrower credit risk, "FinTech" lenders charge minority homeowners higher interest rates. We interpret these pieces of evidence as consistent with loan originators using a form of algorithmic price discrimination. Were these algorithms subject to an internal or external "accountability audit," it is likely that the proxy variables used would fail the IAT because, no matter how well the algorithm performed in detecting the profitability of a loan, the target for the test would, by law, be creditworthiness—not an outcome that included price sensitivity. In this

[185] *See, e.g.*, Susan E. Woodward, U.S. Dep't of Hous. & Urban Dev., *A Study of Closing Costs for FHA Mortgages* xi (2008), http:// www.huduser.org/Publications/pdf/FHA_closing_cost.pdf ("In neighborhoods where borrowers may not be so familiar with prevailing competitive terms, or may be willing to accept worse terms to avoid another application, lenders make higher-priced offers….")

[186] A.B. & S. Auto Service, Inc., 962 F. Supp. at 1056.

[187] Susan Woodward and Robert E. Hall, *Consumer Confusion in the Mortgage Market: Evidence of Less than a Perfectly Transparent and Competitive Market*, 100 AMER. ECON. REV. 511 (2010).

[188] Mark Cohen, *Imperfect Competition in Auto Lending: Subjective Markup, Racial Disparity, and Class Action Litigation*, 8 REV. LAW ECON. 21 (2012)

[189] Bartlett, et al., *supra* note 38.

fashion, simply asking what target variable an algorithm seeks to detect can illuminate illegitimate algorithmic discrimination.

Finally, we want to end this applications section on a positive note. In many discussions with lenders, it has become evident that, at least in the finance realm, firms want to be able to validate what they are doing or what they intend to do before they invest and commit to a predictive algorithm. As we have demonstrated throughout this Article, the standard set by an IAT-accepted environment can provide the valuable consequence of validating the use of proxy variables when their use causes no disparities except through their role in picking up business necessity leveling.

V. CONCLUSION

The era of Big Data places the antidiscrimination mandate at the heart of the Civil Rights Acts of 1964 and 1968 at a critical cross-roads. By relying on data-driven, statistical models, machine learning provides a promising alternative to the type of subjective, face-to-face decision-making that has traditionally been fraught with the risk of bias or outright animus against members of protected groups. Yet left unchecked, algorithmic decision-making can also undermine a central goal of U.S. antidiscrimination law. As we have shown throughout this Article, any decision-making rule that simply maximizes predictive accuracy can result in members of historically marginalized groups being systematically excluded from opportunities for which they are qualified to participate.

Ensuring that algorithmic decision-making promotes rather than inhibits equality thus demands a workable antidiscrimination framework. To date, however, prevailing approaches to this issue have focused on solutions that fail to grapple with the unique challenge of regulating statistical discrimination. Prominent legal approaches (such as reflected in HUD's recent proposed rule-making) have frequently prioritized predictive accuracy despite the fact that such an approach ignores the central risk posed by statistical discrimination demonstrated in our simulation. Conversely, interventions emanating from the field of computer science have largely focused on outcome-based interventions that could themselves lead to claims of intentional discrimination.

Because we derive our input accountability test from caselaw addressing statistical discrimination—in particular, the burden-shifting framework—the IAT advances a vision of algorithmic accountability that is consistent with the careful balance courts have struck in considering the decision-making benefits of statistical discrimination while seeking to minimize their discriminatory risks. By enhancing the predictive accuracy of decision-making, statistical discrimination can greatly enhance the ability of an employer, lender or other decision-maker to identify those individuals who possess a legitimate target characteristic of interest. However, cases such as

Griggs and *Dothard* underscore the danger of simply focusing on predictive accuracy because a proxy that predicts a target variable can nonetheless result in systematically penalizing members of a protected group who are qualified in the target characteristic. That such discriminatory proxies have been consistently declared to be off limits underscores the conclusion that predictive accuracy alone is an insufficient criterion for evaluating statistical discrimination under U.S. antidiscrimination law.

At the same time, our approach is also consistent with the focus in *Griggs* and *Dothard* that differences in a legitimate target can justify disparities that differ across members of protected and unprotected groups. As we show, so long as a proxy used to predict a legitimate target variable is unbiased with respect to a protected group, it will pass the IAT, even if it results in disparate outcomes. The IAT can therefore provide greater transparency into whether disparate outcomes are the result of a biased model or more systemic disparities in the underlying target variable of interest, such as credit risk. In so doing, it can provide vital information about whether the proper way to address observed disparities from an algorithmic model is through de-biasing the model or through addressing disparities in the underlying target variable of interest, such as through targeted subsidies or other transfers. More generally, because the goal of the IAT is to avoid penalizing members of a protected group who are otherwise qualified in a target characteristic of interest, our approach will also be immune to the concern informing cases such as *Ricci v. DeStefano* that our test is biased against qualified individuals.

Finally, our approach provides clear "rules of the road" for how to exploit the power of algorithmic decision-making while also adhering to the antidiscrimination principles at the heart of the Civil Rights Acts of 1964 and 1968. In particular, the IAT offers data scientists a simple test to use in evaluating the risk that an algorithm is producing biased outcomes, mitigating a key source of the regulatory uncertainty surrounding the growing use of algorithmic decision-making. Additionally, our exploration of the early caselaw considering statistical discrimination also reveals that these rules of the road encompass more general concepts to guide both data scientists and regulators when evaluating algorithmic discrimination. These include the notion that, fundamentally, algorithmic decision-making is an effort to assess an unobservable attribute, such as productivity, criminality, longevity, or creditworthiness, through the use of one or more proxy variables. Consequently, evaluating an algorithm must begin with transparency about this target characteristic. And they likewise include the fact that correlation between the unobservable characteristic and the proxy is not, by itself, sufficient to justify the use of the proxy under antidiscrimination principles.

APPENDIX
DE-BIASING PROXY VARIABLES VERSUS DE-BIASING PREDICTIVE MODELS

In this Appendix, we conduct a simulation exercise to illustrate how attempting to de-bias a proxy variable used in a predictive algorithm may do little to de-bias the ultimate predictions. The example we use assumes that a college admissions director wishes to use applicants' standardized test scores (STS) to predict college success (the criterion for allowing an application to continue to the next stage of evaluation.) For this purpose, we assume that a student's performance on the STS is a function of just two factors: *aptitude* and *family wealth*. In our simulation, wealth contributes to test performance because children of wealthier households purchase expensive test preparation classes. To keep the simulation tractable, we assume that wealth does not affect college performance; its only effect is on a student's STS.

Our simulation involves 1,000 college graduates where the admissions director has data on each student's STS at the time of application, student race, and the student's ultimate college performance (e.g., a weighted grade point average or other measure of performance). We divide the race of students, X_i^R, equally so that 500 students are Non-White ($X_i^R = 0$) and 500 are White ($X_i^R = 1$). We assume that wealth and aptitude are distributed as follows:

$$X_i^{Wealth} \sim \begin{cases} N(0,1)\ if\ X_i^R = 0 \\ N(5,1)\ otherwise \end{cases}$$

$$X_i^{Aptitude} \sim N(0,1)$$

Note that under these distributional assumptions, there is very little common support in wealth across race categories. As noted by Kristen Altenburger and Daniel Ho, it is in these settings where the effort to de-bias proxy variables can produce the largest estimation errors.[190] As noted, a student's STS (X_i^{STS}) is a function of X_i^{Wealth} and $X_i^{Aptitude}$, with each variable given equal weight:

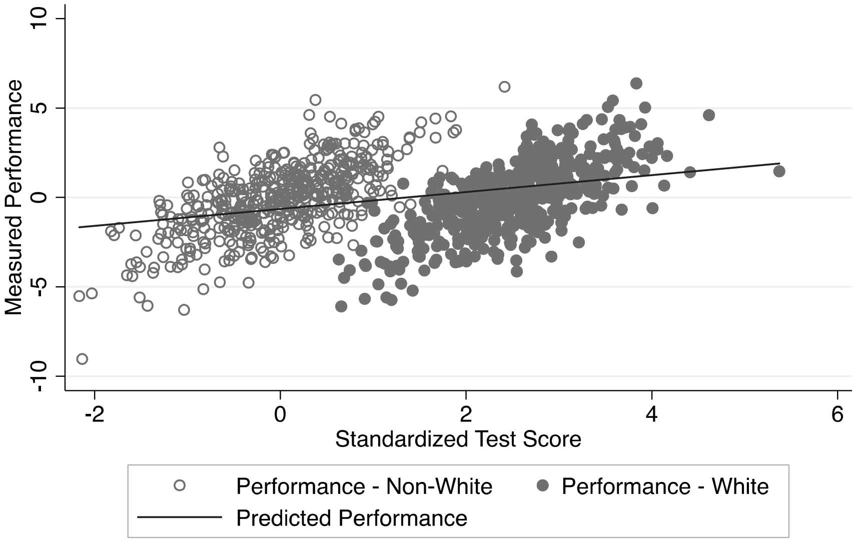
$$X_i^{STS} = 0.5\left(X_i^{Wealth}\right) + 0.5(\ X_i^{Aptitude})$$

Finally, we simulate college performance (*Performance$_i$)* to be entirely determined by aptitude multiplied by a scalar (which we assume here to be 2).

[190] *See* Altenburger & Ho, *supra* note 146, at 111. These settings arise "where sharp preexisting demographic differences may exist across groups." *Id*.

Aptitude is unobservable to the admissions director, inducing her to estimate whether she can use STS to predict college performance. In Figure 1A, we plot separately for White and Non-White graduates the relationship between college performance and STS based on data simulated using the foregoing assumptions. We also include a line that provides the predicted college performance from a simple regression of college performance on STS. As shown in the figure, White graduates had much higher STS scores on average, as would be expected from their much higher family wealth.

Figure 1A: Predicted Performance vs. Actual Performance
No Adjustment for Race

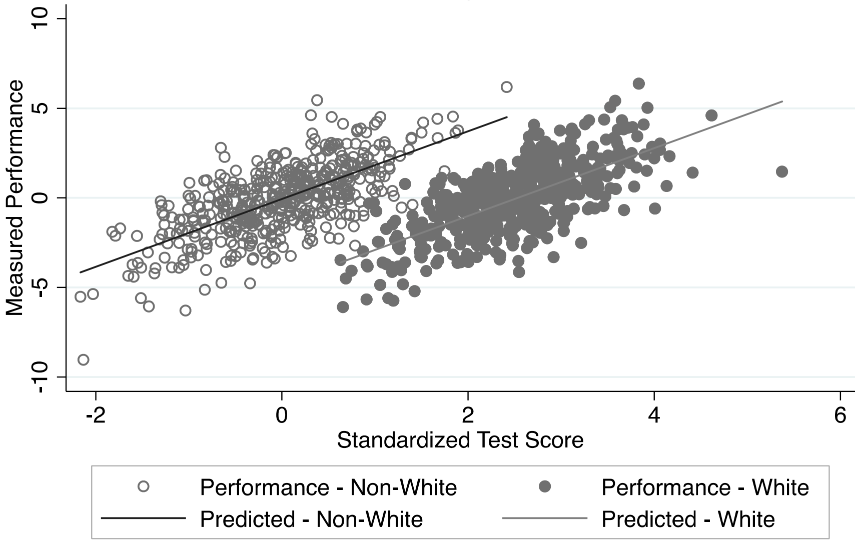


The director of admissions would like to admit students that are likely to have a positive measure of college performance (i.e., *Performance>0*). She therefore runs a simple regression of *STS* on *Performance*, which produces a regression coefficient ($\hat{\beta}^{STS}$) of 0.47. This estimate indicates that a one-point change in *STS* is associated with a 0.47 change in *Performance*. Using this regression estimate, the director generates the fitted line shown in Figure 1A, which provides a predicted measure of *Performance* based solely on *STS*. The fitted line predicts that *Performance* is zero at roughly 1.3, suggesting that using a minimum *STS* of 1.3 would admit students with an expected college performance of at least 0. However, had the admissions director applied this cut-off to these individuals, the bias in STS would result in significant bias against Non-White students owing to their lack of access to test preparation classes:

	Non-White	White
# of Qualified Candidates Predicted by Test Score	13	465

Now assume that the admissions director seeks to control for the greater wealth (and therefore, the greater test preparation bias) among White student applicants. Using the same data, the director expressly adds X_i^R as a control variable in the regression of *STS* on *Performance.* Doing so allows the director to predict *Performance* as a function of both *STS* and *Race.* The results are presented in Figure 2A.

Figure 2A: Predicted Performance vs. Actual Performance
Race Aware Regression Model

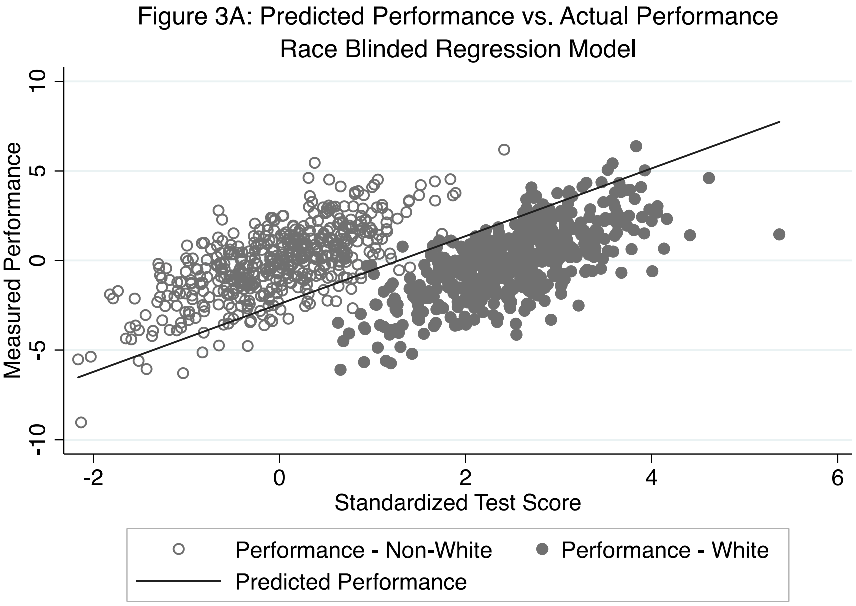


This procedure corrects for the racial bias that arises from using only *STS* to predict *Performance.* This can be seen by the two fitted regression lines, which do a much better job of predicting measured performance across the two racial groups than in Figure 1A. The reason stems from the fact that this regression specification estimates a different y-intercept for each racial category in estimating the relationship between *STS* and *Performance*. Specifically, the regression yields a y-intercept for X_i^R of -4.72, which indicates that in using *STS* to predict *Performance*, it is necessary to deduct 4.72 from the expected performance of White students. (Recall that the difference in average wealth across White and Non-White students is 5.0, so this adjustment eliminates the bias that Wealth creates when using *STS* as a measure of aptitude). With that adjustment, the regression coefficient for *STS* increases from 0.47 to 1.89 because the regression has effectively removed the confounding effect of wealth on *STS* so that it more cleanly reflects aptitude. As above, the admissions director evaluates each fitted line and determines that the fitted line for Non-White students predicts that *Performance* is zero where *STS* is also zero and that the fitted line for White students predicts that *Performance* is zero at 2.53. Applying a minimum test cut-off of 0 for Non-White students and 2.53 for White students would result in the following students being deemed qualified:

	Non-White	White
# of Qualified Candidates Predicted by Test Score	250	248

This procedure solves the racial bias created by using only *STS* to estimate *Performance*, but it is clearly problematic insofar that it requires a different minimum cut-off for White and Non-White students. This is disparate treatment. To avoid this problem the director therefore turns to the approach advanced by Devin Pope and Justin Sydnor as well as by Crystal Yang and Will Dobbie.[191] This procedure involves using the regression estimates generated for Figure 2A but treating all students as if they had the average value of race, or in this example, a race of 0.5. Making this adjustment means that every student receives a deduction of -2.36 (i.e., 0.5 x -4.72) after multiplying their exam score by the slope coefficient for *STS* of 1.89, which remains purged of the confounding influence of Wealth. This permits the director to estimate a single fitted regression line as shown in Figure 3A:

[191] *See supra* note 71.



The fitted line predicts that *Performance* is zero at approximately 1.28, which the director uses as the minimum cut-off. Had the director applied this cut-off to this group of individuals, the following results would have occurred:

	Non-White	White
# of Qualified Candidates Predicted by Test Score	15	468

In effect, the results are largely identical to those obtained by using only *STS* to predict performance. The reason stems from the lack of common support in wealth across White and Non-White students, resulting in the need for a significant negative adjustment to every White student when estimating performance from *STS*. Applying half of this negative adjustment to *every* student thus works against the de-biasing of the slope coefficient for *STS*. In short, the slope coefficient for *STS* in Figure 3A is unbiased with respect to Non-White students, but the predictive model is not. This problem was significant in this example because there was so little common support in wealth across White and Non-White students—a problem that will exist whenever there are significant demographic differences across protected and unprotected groups.

Algorithmic Accountability: A Legal and Economic Framework

Robert Bartlett (UC Berkeley Law),
Adair Morse, Richard Stanton & Nancy Wallace (UC Berkeley Finance)

Lending Footprints and Discrimination Testing

Adair Morse and Robert Bartlett

Chicago-Booth Conference on Household Finance

March, 2020

How Economists Think about Discrimination

TASTE-BASED DISCRIMINATION

- An individual dislikes members of a particular group and derives utility from discriminating against them (Becker 1957)
- Should not persist in the long run because of competition
  - *Costly*
- De facto: discretion persists

STATISTICAL DISCRIMINATION

- A decision-maker (employer, lender) does not observe a business necessity variable (productivity, creditworthiness).
- Uses a proxy for that variable, such as the average for a group of people (Arrow, 1973, Phelps, 1972)
  - *Profit Maximizing*
- De facto use of statistical discrmination term:

  Indirect stat discrmination: using averages over a non-protected group (not “black” but “high school name”) as a proxy for creditworthiness.

How the Law Thinks about Discrimination

The mapping of the law to economists' thinking is clear on the below:

1. Let's make sure to make taste-based discrimination illegal

   (And anyway, it is not profit maximizing)

2. And then let's make sure technology does not implement the direct form of Arrow/Phelps discrimination
   - i.e.: allowing lenders to score by a protected category or a "highly correlated" variable
     - Protected category: race, ethnicity, gender, etc.
     - Highly-correlated = hair styles, redlining, etc.

How the Law Thinks about Discrimination

But the law is not quite so simple as 1 and 2:

1. Let’s make sure to make taste-based discrimination illegal
2. And then let’s make sure technology does not implement the direct form of Arrow/Phelps discrimination

Rather, the law further demands that a lending process :

3. Only induces statistical discrimination for business necessity

Proxy Variables for Statistical Discrimination & Accountability

Outline

I. Law / Caselaw

II. Input Accountability Test

III. Application in Credit Data

Example: UnitedHealth (UH) - insurance co

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations, 366 SCIENCE 447 (2019)

UH used an algorithm to inform hospitals about patients' level of sickness, which the hospitals used in prioritizing intensity of care

- Purpose: Effectively allocation of resources to the sickest patients
- Problem:
  - UH had gauged sickness using historical expense data (cost of care)
  - African-American patients historically spend less for the same illnesses and level of illness
- Result: The algorithm caused African Americans to receive substandard care as compared to white patients

Setting: Algorithmic Accountability

- Two new State-level "algorithm accountability" legislations
  - **New York State law:** creates a task force to recommend procedures for deciding if automated decisions by city agencies disproportionately impact protected groups
  - **Washington State bill:** Requires the state's CIO to assess whether any automated decision system used by a state agency has a known bias, or is untested for bias
- **Absent:** Formal standard for courts or regulators to deploy in evaluating algorithmic decision-making....

i.e.: ***What exactly does it mean for an algorithm to be accountable?***

Title VII of the Civil Rights Act of 1964

An unlawful practice for an employer

1. "to … discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, sex, or national origin; or
2. to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities … because of such individual's race, color, religion, sex, or national origin."

But how do you implement this?

What do Lenders Say they do?

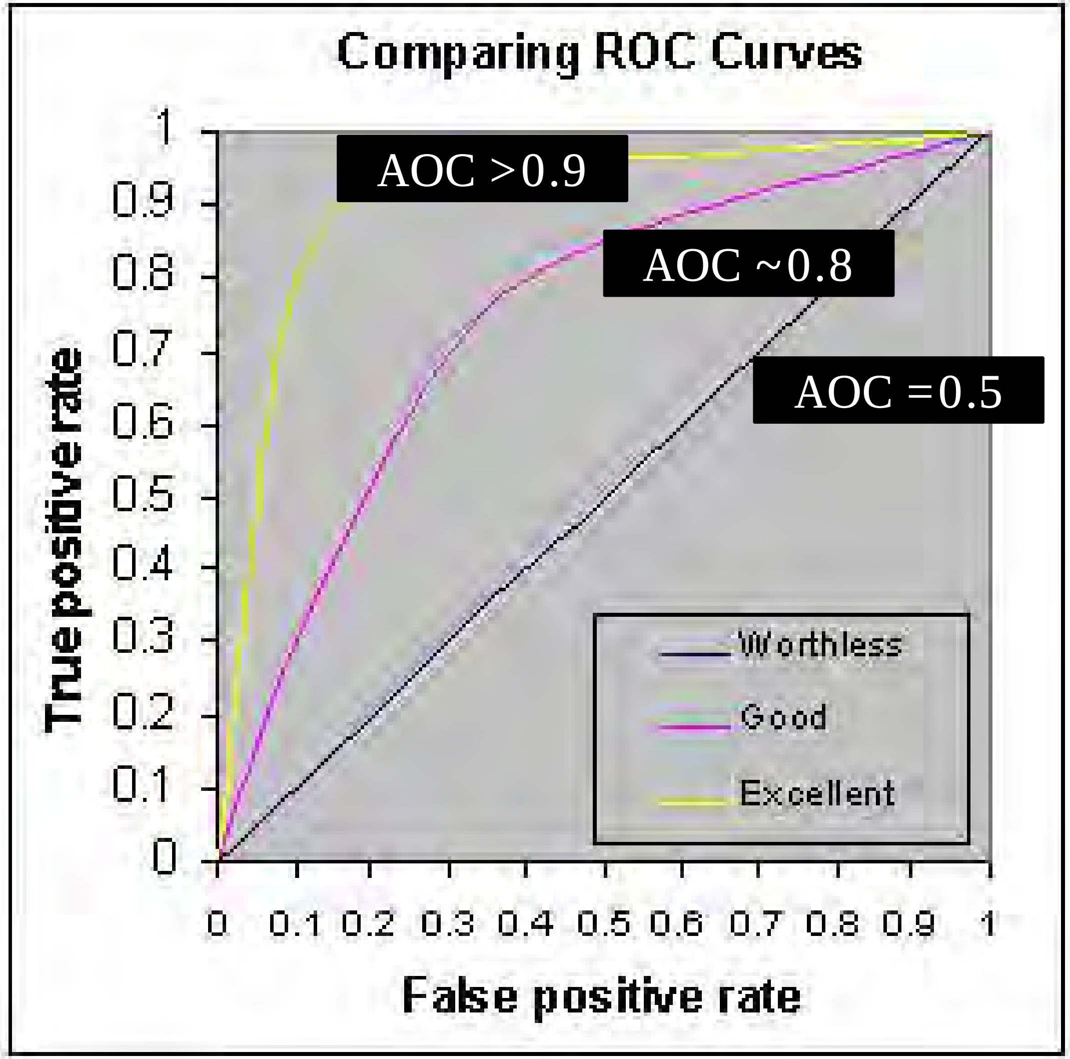
- Lender : a platform lender or bank with 1,000s of variables
- Objective: use machine learning (ML) to do credit scoring without discrimination
- Lawyers: *"To avoid discrimination, implement a `least discriminatory' approach"*
- How?

1. Define "target" (ML term) : = the business necessity need for proxy variables
    - **Courts: in lending, target = "creditworthiness"**
        - (Note: credit risk != profits; courts make this point explicit))
2. Come up with models of **predicting accuracy of default**
3. If outcomes are disparately applied against a protected category…

    lender needs to be able to show that the algorithm uses the least discriminatory predictive model for a given level of predictive accuracy

Problems with this:

Part 1: An econometrician / data scientist point of view



- ROC curves, think…
  - Run logit model of default on cash flow variables plus 1,000s of proxies for missing fundamentals
  - Calculate how predictive model is (goodness of fit)
- Imagine result…
  - *"my best predictive model generates ROC of 0.78"*
  - I can generate many models with interactions of variables /nonparametrics that have similar ROC
  - Which one has least impact on protected group?
- Problem: let's say with just pure cash flow variables the model yields ROC of 0.68.
- Does the court allow us to increase ROC by 0.10 and then apply the discrimination test?

Problems with this:

Part 2: The law…….*Burden- Shifting Framework*

Original frame from Supreme Court:

- *Griggs v. Duke Power Co*

Codified by Congress:

- *Civil Rights Act of 1991*

Important Caselaw from Supreme Court*:*

- *Ricci v. DeStefano*
- *Dothard v. Rawlinson*

Aside

- Like the Civil Rights Act of 1964, 1991 and their caselaw, original application is in context of employment decisions.
- However, credit and housing decisions adopted the interpretation of discrimination and this framework explicitly in Equal Credit Opportunity Act and Fair Housing Act

Burden- Shifting Framework

First Burden: Plaintiff must identify a specific employment practice that causes “observed statistical disparities” across members of protected and unprotected groups.

- If plaintiff successful…

Second Burden: The defendant must then “demonstrate that the challenged practice is *job related for the position in question* and consistent with business necessity.”

- If defendant successful…

Third Burden: Plaintiff must show that an equally valid and less discriminatory practice was available that the employer refused to use

Burden- Shifting Framework

First Burden: Plaintiff must identify a specific employment practice that causes "observed statistical disparities" across members of protected and unprotected groups.

- If plaintiff successful…

Second Burden: The defendant must then "demonstrate that the challenged practice is *job related for the position in question* and consistent with business necessity."

- If defendant successful…

Third Burden: Plaintiff must show that an equally valid and less discriminatory practice was available that the employer refused to use

This is where the least discriminatory approach comes from

Burden- Shifting Framework

First Burden: Plaintiff must identify a specific employment practice that causes "observed statistical disparities" across members of protected and unprotected groups.

- If plaintiff successful…

Second Burden: The defendant must then "demonstrate that the challenged practice is *job related for the position in question* and consistent with business necessity."

- If defendant successful…

But it does not excuse the defendant from satisfying Second Burden

Third Burden: Plaintiff must show that an equally valid and less discriminatory practice was available that the employer refused to use

This is where the least discriminatory approach comes from

Burden- Shifting Framework

First Burden: Plaintiff must identify a specific employment practice that causes “observed statistical disparities” across members of protected and unprotected groups.

- If plaintiff successful…

Second Burden: The defendant must then “demonstrate that the challenged practice is *creditworthiness for the loan in question* and consistent with business necessity.”

- If defendant successful…

Third Burden: Plaintiff must show that an eq tory practice was available that the employer refu

Fair Lending laws adopted burden shifting for lending… switch employment language to creditworthiness

Dothard v. Rawlinson

A California Prison wanted to hire prison guards

- Determined that a job-required necessity is strength (legitimate)
- Could not measure strength of applications, so used proxy of height
- A group of female applicants sued and won

Court:

- Indeed strength is legitimate as target and height predicts performance
- But the strength needed is a specific strength and the height measurement penalizes females beyond the business necessity

Dothard v. Rawlinson: IAT

- Econometrician Version
  - Decompose height into that which predicts the target strength and a residual
  - Test if the residual is still correlated with female:

$$Height_i = \alpha \cdot Strength_i + \varepsilon_i$$

Test: $\varepsilon_i \perp gender.....$ $\qquad$ $regress\text{:}\ \ \varepsilon_i = \beta_0 + \beta_1 gender$

Proxy height fails ⬄ $\beta_1 \neq 0$

If so, exclude height as only legitimate business necessity

We call this the *Input Accountability Test*

A fix instead of exclude?

Question: Why can’t we just fix the scoring by a protect group to de-bias?

- Pope and Sydnor (2011)

Answer: It only works on average, not for individuals. The law is about individuals

Answer: It is illegal (*Ricci v. DeStefano)*

New Haven wanted to discard the results of an “objective examination” that sought to identify city firefighters who were the most qualified for promotion because thre was statistical racial disparity in the results against a minority group. A group of white and Hispanic firefighters sued, alleging that the city’s discarding of the test results constituted race-based disparate-treatment.

Court ruled for plaintiff… no discarding

Why: Can’t use a protected class variable in a decision because (again) it could cause disparities because of the averages part

Challenges of the IAT

1. Unobservability of Target
   - Kleinberg, Ludwig, Mullainathan Sunstein (2019): *training datasets*
   - Calculating thresholds

2. Measurement Error in Target

$$Strength_i^* = Strength_i + \mu_i$$

$$Height = \alpha \cdot Strength_i^* + \zeta_i$$

$$\zeta_i = -\mu_i + \varepsilon_i$$

Note: UnitedHealth is this problem. Also, selective labels problem (De-Arteaga, et al., 2018).

Idea: Structural version

3. Standard errors as n grows large.

Implementation: “Footprints & Discrimination”

Motivation

- U.S. household debt: $14 trillion
  - Increase of $1.3 trillion from peak in 2008 (NY Fed)
  - If annual debt turnover is 15%
- Then… new float of recent years ~$2.2 trillion per year
- Of this, how much algorithmically-decided based on 1,000s of proxy variables?
  - ?, but just Quicken is $100 bn of it.
  - All big Banks have big data now.

Jeff Budzik

CTO of ZestFinance:

“The models we put into production for our customers tend to have hundreds or thousands of variables in them. We have one with 2200 variables that’s running an auto lending business”

Footprints & Discrimination

Question

Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2019),

Bartlett, Morse, Stanton, and Wallace (2019)

How can the use of machine learning in credit profiling avoid being inadvertently discriminatory?

Outline of Application :

(1) ROC Analysis

(2) IAT Tests for Gender Discrimination

Data

- Data from a consumer lender in Eastern Europe
- 300,000 consumer loans
- Loans made in stores but not collateralized, just loans for good
- Borrowers have thin credit record
- Dataset contains default (the target)

Unique:

  - 124 variables (many of the them categorical)
  - Can be made "long" into 1,000s of variables even without interactions

Step 1 – Looking for footprints

How well can we do as a ML-er?

Prediction target: Default via area under ROC assessment

Footprints of creditworthiness literature (abridged)

- Berg, Burg, Gombovic, and Puri (2019) :“digital footprints” type of device (tablet, computer, phone), operating system (Windows, iOS, Android), and email provider predicted default rates among the customers of a German lender.
- Bjorkegren and Grissen (2019) mobile phone usage data
- Vissing (2010) : Consumer goods products people buy

Types of Variables

1. Fundamentals (cash flow, wealth, cost of capital)
2. Occupation
3. Goods
4. Shelter
5. Family Life
6. Soft Info Applying
7. Soft Info Credit

Fundamental Variables

	Mean	StDev		Mean	StDev
Income	168,797	237,125	Missing data Credit Bureau	0.1350	0.3417
Credit Amount	599,028	402,494	# Outstanding Loans	4.3184	10.5095
Payment Amount	27,109	14,494	Prior Loans Delinquent %	0.0054	0.0312
payment_to_credit	0.0537	0.0225	How Delinquent, if any	0.0089	0.0851
payment_to_income	0.1809	0.0946	Ontime Prior Payments, if any	0.1371	0.2522
Homeowner	0.6937	0.4610	Percent of Prior Loans Closed, if any	0.0991	0.2089
Credit Score Max	0.6159	0.1561	Remaining Days on Last Issue	-928.0	644.8
Cedit Score Min	0.3996	0.1874	Days Since Last Issue	-419.3	526.3
# Credit Bureau Requests	0.2313	0.8568	Own Car?	0.3401	0.4737
			Age of Care, if any	0.3418	0.7508

Goods & Durables Variables

	Mean	StDev
Purchase Price of Good	538,398	369,447
LTV of Loan to Good	1.1230	0.1240

Occupation Variables

	Mean	StDev		Mean	StDev
Low Skill Worker	0.2058	0.4043	Pensioner	0.1800	0.3842
Drivers Security	0.0824	0.2749	Working - Unnamed	0.5163	0.4997
Office Worker	0.1983	0.3987	Employ Commercial	0.2329	0.4227
Manager /Skilled	0.1658	0.3719	Employment Years	5.3562	6.3202
Prof Services	0.0344	0.1821	Gives Office Phone	0.8199	0.3843

Living / Family Variables

	Mean	StDev
Civil Marriage	0.0968	0.2957
Religious Marriage	0.6388	0.4804
Widow	0.0523	0.2227
# Children	0.4171	0.7221
Rural	0.1047	0.3062
Large Metro	0.1572	0.3640

Shelter Variables

	Mean	StDev
Municipal Housing	0.0364	0.1872
Office Housing	0.0085	0.0919
Live with Parents	0.0483	0.2143
Age Building	0.2532	0.3626
N/A Age Building	0.6650	0.4720
Elevators Relative	0.0365	0.0998
N/A Elevators	0.5330	0.4989
Entrances relative	0.0741	0.1028
N/A Entrances	0.5035	0.5000

Soft Application Variables

	Mean	StDev
# Documents	0.9302	0.3443
No Documents	0.0961	0.2947
# Contacts Provided	1.5371	0.7221
Social Network: Defaulters	0.1434	0.4466
Spouse Present	0.0370	0.1887

Prior Credit Proprietary Variables

	Mean	StDev
Previous Good Loan LTV	0.960	0.255
Previous Rejection %	0.223	0.257
# Previous Apps	4.597	4.180

ROC Analysis

Logit (Default) = fundamentals + (iteratively, then all)

1. Occupation
2. Goods
3. Shelter
4. Family Life
5. Soft Info Applying
6. Soft Info Credit

Logit (default) = function of Fundamental `Permanent Income’ Variables)

	Dependent Variable: Default		
Ln Income	-0.151***	Homeowner	-0.0131
	[0.0391]		[0.0148]
Ln Credit Amount	-1.934***	Credit Score Max	-2.084***
	[0.0902]		[0.0464]
Ln Payment Amount	2.269***	Credit Score Min	-2.676***
	[0.0996]		[0.0467]
Payment_to_credit	-32.35***	# Credit Bureau Requests	-0.0112
	[1.540]		[0.00961]
Payment_to_income	-0.372*	Missing data,Credit Bureau	-0.141***
	[0.202]		[0.0245]
		Cut off the prior balances debt vars	
	Observations	307,321	
	Pseudo R-squared	0.0872	
	Area under ROC	0.7217	

ROC Analysis … Columns adding Proxies

Do the proxies add to the ROC?

How did the “Adair-Bobby-ML-Lasso Optimizing’ do?

		Variables Included: Fundamentals + ….						
	Fundamentals	Occupation	Goods	Shelter	Family Life	Soft Info App	Soft Info Credit	All
		Dependent Variable: Default						
Observations	307,321	307,321	307,045	307,321	307,321	307,321	306,302	306,026
Pseudo R-squared	0.0872	0.0944	0.0937	0.0885	0.0872	0.0916	0.0904	0.108
Area under ROC	0.7217	0.7297	0.7289	0.7232	0.7217	0.7262	0.7255	0.7434

Step 2: Which of those Proxy Variables enhancing ROC area pass the Input Accountability Test?

Regress:	Proxy = fundamentals + error
Regress:	Error = b0 + b1* female
Test:	b1 != 0

Standard Errors… with 300,000 observations, asterisks are “cheap”

- The test should be more rigorous, trying to empower firms to be able to use variables, not to say everything has asterisks (not a useful test)

- Cannot go down an “economic significance” argument because this is law. There is no sense in the law that “5 people out of 10,000 do not matter”

- d-value approach to the p-value problem as n-> large

D-value : Demidenko (2013)

“The P-value You Can’t Buy” American Statistician

- Rather than focus on a comparison of group means, the d-value is designed to examine how a randomly chosen female fared under this proxy variable relative to a randomly chosen male.

P value (under normality):

$$p = \Phi\left(-\frac{|b|}{s}\right)$$

D-value (under normality):

$$d = \Phi\left(-\frac{|b|}{s\sqrt{n}}\right)$$

Where s is the standard error: $s = \text{stdev}/\sqrt{n}$

Family Lifestyle

	(1) Civil Marriage	(2) Non-civil Marriage	(3) Widow	(4) # Children	(5) Rural	(6) Large Metro
Sign indicating disparate outcome against protected category	+	–	–	+	–	+
female	0.0174	-0.0684	0.042	-0.00596	0.0112	0.00604
	[0.00112]	[0.00177]	[0.000833]	[0.00272]	[0.00110]	[0.00136]
Observations	307,321	307,321	307,321	307,321	307,321	307,321
R-squared	0.001	0.005	0.008	0.000	0.000	0.000
Standard errors in brackets						
On d-values below: range +/- 1% around 50% is not concerning						
d-value	51.1%	47.2%	53.6%	49.8%	50.7%	50.3%
Coefficient from default logit	0.0198	-0.0999***	-0.146***	0.0155	-0.198***	0.0915***

Occupation – part 1

	(1) Low Skill Worker	(2) Drivers Security	(3) Office Worker	(4) Manager /Skilled	(5) Prof Services
Sign that would indicate disparate outcome against protected category	+	+	–	–	–
female	-0.166	-0.157	0.148	0.0571	0.0482
	[0.00150]	[0.000988]	[0.00149]	[0.00139]	[0.000685]
Observations	307,321	307,321	307,321	307,321	307,321
R-squared	0.038	0.076	0.031	0.005	0.016
Standard errors in brackets					
On d-values below: range +/- 1% around 50% is not concerning					
d-value	42.1%	38.7%	57.1%	53.0%	55.1%
Coefficient from default logit	0.195***	0.308***	-0.038	-0.0281	-0.253***

Occupation – part 2

	(1) Pensioner	(2) Working - Unnamed	(3) Employ Commercial	(4) Employment Years	(5) Gives Office Phone
Sign that would indicate disparate outcome against protected category	–	+	+	–	–
female	0.0494	-0.079	0.00753	0.323	-0.0495
	[0.00140]	[0.00187]	[0.00157]	[0.0235]	[0.00140]
Observations	307,321	307,321	307,321	307,321	307,321
R-squared	0.004	0.006	0.000	0.001	0.004
Standard errors in brackets					
On d-values below: range +/- 1% around 50% is not concerning					
d-value	56.3%	45.7%	48.3%	50.6%	43.7%
Coefficient from default logit	-2.119***	0.270***	0.168***	-0.0266***	-1.917***

Shelter – part 1

	(1)	(2)	(3)	(4)	(5)
	Municipal Housing	Office Housing	Live with Parents	Age Building	N/A Age Building
Sign that would indicate disparate outcome against protected category	+	–	+	–	–
female	0.00467	-0.00134	-0.0149	0.0146	-0.0193
	[0.000705]	[0.000349]	[0.000799]	[0.00136]	[0.00178]
Observations	307,321	307,321	307,321	307,321	307,321
R-squared	0.000	0.000	0.001	0.000	0.000
Standard errors in brackets					

On d-values below: range +/- 1% around 50% is not concerning

d-value	50.5%	49.7%	48.7%	50.8%	49.2%
Coefficient from default logit	0.105***	-0.255***	0.00486	-0.441***	-0.267***

Shelter – part 2

	(1) Elevators Relative	(2) N/A Elevators	(3) Entrances relative	(4) N/A Entrances
Sign that would indicate disparate outcome against protected category	–	+	–	+
female	0.00194	-0.0242	0.00294	-0.0263
	[0.000373]	[0.00186]	[0.000386]	[0.00187]
Observations	307,321	307,321	307,321	307,321
R-squared	0.000	0.001	0.000	0.001
Standard errors in brackets				
On d-values below: range +/- 1% around 50% is not concerning				
d-value	50.4%	49.1%	50.5%	49.0%
Coefficient from default logit	-0.255**	0.00388	-0.298**	0.0095

Goods & Durables

	(1) Goods Price	(2) Goods LTV
Sign that would indicate disparate outcome against protected category	–	+
female	5437	-0.00547
	[563.8]	[0.000463]
Observations	307,045	307,045
R-squared	0.001	0.000
Standard errors in brackets		
On d-values below: range +/- 1% around 50% is not concerning		
d-value	50.7%	49.1%
Coefficient from default logit	-5.25e-07***	0.947***

Proprietary Prior Credit

	(1) previous good loan LTV	(2) Previous Rejection %	(3) # Previous Apps
Sign indicating disparate outcome against protected category	+	+	–
female	0.0154 [0.000950]	0.0139 [0.000962]	0.439 [0.0156]
Observations	307,321	307,321	307,321
R-squared	0.001	0.001	0.003
Standard errors in brackets			
On d-values below: range +/- 1% around 50% is not concerning			
d-value	51.8%	50.4%	51.6%
Coefficient from default logit	0.213***	0.617***	-0.0109***

Soft Info – Application Variables

	(1)	(2)	(3)	(4)	(5)
	# Documents	No Documents	# Contacts Provided	Social Network: Defaulters	Spouse Present
Sign indicating disparate outcome against protected category	–	–	+	+	–
female	-0.00753 [0.00121]	0.0035 [0.00102]	-0.0121 [0.00273]	0.0111 [0.00170]	-0.0155 [0.000715]
Observations	307,321	307,321	307,321	306,302	307,321
R-squared	0.000	0.000	0.000	0.000	0.002
Standard errors in brackets					
On d-values below: range +/- 1% around 50% is not concerning					
d-value	49.6%	50.1%	49.5%	50.7%	48.3%
Coefficient from default logit	-0.317***	-0.615***	0.0515***	0.160***	-0.0492

Eliminate & Re-run Default Model

Eliminate 5 of 37 variables for bias

- buy car, previous goods loan-to-value, religious marriage, gives phone for employer, spouse present

How much area under the ROC curve / pseudo r-square is sacrificed?

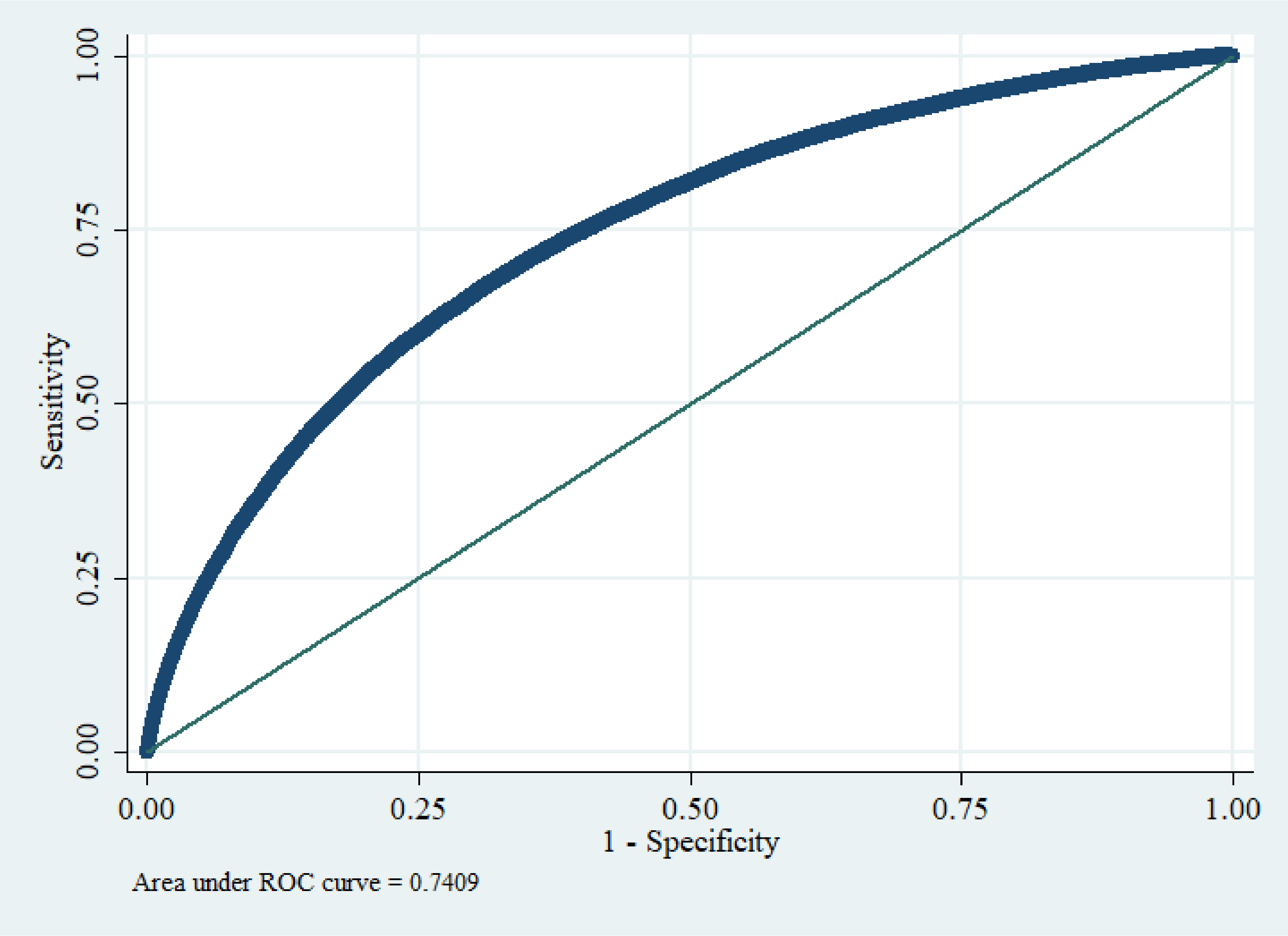
Logistic regression Number of obs = 306,026 Pseudo R2 = 0.1054

	Coef	Z-stat		Coef	Z-stat
lnamt_income_total	0.0909	2.27	occ_lowskilllabor	0.1927	8.29
lnamt_credit	-1.1398	-10.31	occ_drivers_security	0.2659	9.38
lnamt_payment	1.4618	13.44	occ_office_workers	-0.0306	-1.25
payment_to_credit	-24.1704	-14.44	occ_managers_skill	-0.0315	-1.20
payment_to_income	0.7993	3.94	occ_profserv	-0.2464	-5.00
Homeowner	0.0007	0.04	employ_pensioner	-0.2186	-5.24
Max Credit Score	-1.8938	-39.94	employ_workingunnamed	0.2696	8.49
Min Credit Score	-2.4276	-51.09	employ_commercial	0.1472	4.35
# Request Credit Bureau	-0.0141	-1.44	employed_years	-0.0268	-17.82
Missing Requests	-0.1116	-4.5	shelter_municipal	0.1044	2.86
age_car	-0.0398	-4.27	shelter_office	-0.2463	-3.00
amt_goods_price	0.0000	-9.06	shelter_parents	0.0322	1.13
ltv	0.9696	15.64	years_build_medi	-0.3906	-3.32
bb_outstanding_count	0.0005	0.48	na_years_build_medi	-0.2273	-2.51
bb_delinquent	1.8054	6.21	elevators_medi	-0.4098	-4.00
bb_howdelinquent	-0.2077	-1.98	na_elevators_medi	-0.0001	0.00
bb_ontime	-0.1306	-4.22	entrances_medi	-0.2567	-2.19
bb_succ_closed	-0.1768	-3.71	na_entrances_medi	0.0119	0.30
Days outstanding on credit	0.0002	11.93	documents_count	-0.4149	-7.73
Days outstanding on last credit	-0.0001	-2.26	documents_none	-0.7271	-11.51
prev_rej_count_pct	0.6405	20.85	contacts_personal_count	0.0414	4.25
prev_apps_HC_count	-0.0090	-4.59	Network Defaulters	0.1596	11.83

Re-running Logit (default) dropping biased proxies

Area under ROC drops from 0.7434 to 0.7409

Pseudo rsquared drops from 0.108 to 0.1054



Conclusions

Objectives:

- Get more finance research engaged in the policy debate about algorithmic use in credit scoring
- Debunk the emerging literature that AI poses not dangers because it removes discretion and any biases can be corrected

Accomplished (hopefully)

1) Demonstrated what the law dictates about inputs
2) Provided a really simple test for firms to use ex ante and regulators or courts ex post
3) Showed that at least in our application, the test provides results that are workable to firms