Detecting Heterogeneity in the MPC: A Machine Learning Approach

Last Update: 22.10.2021

Master's Thesis (Draft)

Department of Economics

University of Mannheim

submitted to:
Prof. Krzyzstof Pytka, PhD

submitted by:
Lucas Cruz Fernandez

Student ID: *********
Studies: Master of Science Economics (M.Sc.)

Adress: ******
Phone: ******
E-Mail: ******

Mannheim, ADD DATE

1 Introduction

How do households respond to income shocks and how do their responses differ given their personal characteristics and economic circumstances? These questions are not only at the centre of a wide academic debate in economics but also of major importance for policy makers. While the former revolves around verifying or neglecting the main mechanisms of the Permanent Income Hypothesis (PIH), the latter are interested in improving government transfers to more efficiently use public funds. These two sides have sparked many investigations using a wide array of approaches to quantify hosueholds' responses to income shocks. This Marginal Propensity to Consume (MPC) - at the centre of macroeconomics since Keynes introduced it at the heart of his General Economic Theory - quantifies how much households will spend on consumption of each dollar they receive from an income shock. While research has long focused on quantifying whether the MPC out of income shocks is zero and thus in line with the PIH, the literature has seen a shift of focus over the last decade. On average most studies support the notion of a zero MPC, however, more recent evidence suggests that for specific groups the response is significantly different.

Empirical research related to the MPC and its heterogeneity has used several settings to identify income shocks. One of the most prominent is to use natural experiments in which households receive exogenous income shocks. Following Parker et al. (2013) and Misra and Surico (2014), we exploit the 2008 tax rebate in the USA to estimate households' MPC using data collected by the Consumer Expenditure Survey (CEX). Similar to these two prior studies, we are able to use the rich information on consumption the CEX provides to not only identify heterogeneities in the overall MPC but also to analyse which categories of consumption goods households spend their rebate on. However, our econometric approach sets us apart as it is more sophisticated and more precise in detecting heterogeneities compared to any contribution we are aware of.

Namely, the two main channels are life-cycle dynamics and liquidity. The former is driven by a consumer's age and the associated fluctuation in income. As data consistently shows (Attanasio and Weber, 2010), consumption expenditures follow a hump shape over the life-cycle rather than being roughly constant as we would expect under the PIH. This anomaly is often referred to as the retirement-consumption puzzle and connected to an increased amount of free time which allows households to reduce the cost of their consumption. Therefore, we would expect a reduction in the measured MPC in age.

In the case of liquidity, its role is linked to the nature of the income shock and borrowing constraints. If a positive income shock is anticipated, households that are already close to or at their borrowing constraint cannot borrow new funds to smooth consumption in anticipation of a higher future income. Thus, once the shock realises, we will observe an

increase in consumption. On the other hand, saving is always possible for any household and hence we will not see a reaction once the shock realises in case of a negative anticipated shock. Thus, more liquid households react less to a positive anticipated shock in comparison with liquidity constrained ones. In contrast, in case of an unanticipated shock, we expect the opposite. Think of an agent that is temporarily out of work and has no liquid wealth at their disposal. In case of a negative shock, the agent is forced to adjust their consumption behaviour downward. Meanwhile, a positive shock will always be saved and stretched over future periods, no matter the level of households' liquidity. E.g. Bunn et al. (2018) document this asymmetry depending on the sign of the shock.
It is important to highlight that in our setting, households experience an anticipated positive income shock. The Economic Stimulus Act was signed into law by President Bush in February 2008 and payments of tax rebates started in April of the same year. Therefore, following the previous arguments, theory would expect older households to react less, but liquidity not being a major driver. However, the tax rebate was disbursed to US taxpayers during a time of national and global economic downturn. Thus, many households receiving the stimulus might have been in economic turmoil when receiving the payment and actually spend it to cover regular expenses that they otherwise would not have been able to cover (e.g. rent or utilities). However, Parker et al. (2013) emphasise that some rebates were reported to be received outside of the disbursal window, which suggests that the income shocks might not have been anticipated and only noticed after their arrival.
One major issue in the existing literature is the way how heterogeneity is measured. Most studies rely on either splitting their sample into smaller sub samples and estimating the MPC within each sample or use dummy variables that are defined by the authors based on some continuous variable. These approaches suffer from the severe issue that any heterogeneity that does not fall into this pre-defined pattern is not captured and will muddy the results of these investigations. In the worst case these procedures miss to pick up existing heterogeneity or missing patterns within these pre-defined subgroups. On the contrary, our Double Machine Learning (DML) approach allows us to estimate the conditional MPC out of the tax rebate of each individual household. Prior studies have to rely on looking at the correlation between their estimates and characteristics such as liquidity, but our setting enables us to calculate more sophisticated measures that capture the influence of each variable on the MPC.
The fine-grained consumption data of the CEX allows us to identify what kind of goods households consumed and what they spent their stimulus money on. As Kaplan and Violante (2014) note, the tax stimulus is anticiapted and is subject to these special circumstances. Therefore, one might also speak of our estimated coefficients as a 'Propensity to consume the rebate' or 'rebate coeffcient', which is not necessarily equivalent to households' overall MPC. We compare our estimates with the range found in the literature

using different income shock sources to get a grasp of whether this difference might play a role and for what households it does. However, while a government stimulus program might not be perfectly appropriate to verify theoretical models concerned with the MPC, providing evidence on their effect on individuals is of major importance for future policy making. While in some cases when economic relief is urgent broadly defined, non-targeted stimuli might be a good option to pursue, targeted transfers can play a major role in many policy settings. **(rephrase)**

this needs to be implemented somewhere By exact definition, the MPC is the reaction to an unanticipated, transitory income shock. However, in out setting, the shock cannot be fully seen as unanticipated as the tax stimulus payment was signed into law and therefore known to the public several months before the payments were conducted. However, in such cases to identify only the contemporaneous reaction of households the empirical literature (e.g. Parker et al. (2013)) .

section on results here We show that indeed both these channels play a role in the heterogeneity of households' response to the 2008 tax stimulus. Similar to the existing literature we find... However, additionally we are able to show that the heterogeneity is not only linear/indeed linear...
The rest of the paper is structured as follows: Section 2 summarizes the theoretical and empirical literature on MPC heterogeneity. Section 3 discusses the data source and challenges connected with it. The empirical methodology we use is described in Section 4, while Section 5 presents the identification and estimation results of the MPC. We further investigate sources of heterogeneity in responses in Section 6. Section 7 concludes.

2 Literature Review

The literature investigating the size of the MPC and potential heterogeneity can broadly be categorized into three different strands. The first one uses quasi-experimental settings to exploit variation in income to estimate households' MPC. The second uses surveys that explicitly question participants about their MPC - be it out of actual or hyopthetical income shocks. Lastly, a vast literature focuses on building sophisticated macroeconomic models that are calibrated to match real world data and subsequently estimate the MPC agents experience in these models. Our work falls into the first of these categories.
In this section we briefly summarize the findings in all three and additionally discuss two studies - Parker et al. (2013) and Misra and Surico (2014) - in more detail as they investigate MPC heterogeneity using the same data as we do.
Quasi-experimental settings appear all the time in the real world, e.g. in case of a specific

policy being implemented or another exogenous shock happening. Researchers interested in MPC heterogeneity focus on shocks that alter the income of a household. For example, Fagereng et al. (forthcoming) use panel data from Norwegian administrative data on winners of a state lottery in which most citizens participate. Receiving a payment from the lottery can be seen as an unanticipated income shock because the chances of winning are so low. They find that households winning the lottery spend almost half of their win within one year and 90% after 5 years. Moreover, liquidity and age are the only variables correlated with the MPC, providing evidence for the existence of the liquidity and age channels. In similar vein, Golosov et al. (2021) construct a dataset of lottery winners in the USA to estimate their MPC and labor market response. They make use of tax forms provided by the lottery winners and general income tax statements. Their main goal is to estimate the labor market responses to windfall gains in unearned income but their strategy allows them to identify the MPC as well. Using a Difference-in-Difference estimator, their estimated MPC is around 60ct out of each dollar earned on average, while labor earnings are reduced by 50ct. To investigate heterogeneity in these responses, the authors split their sample based on the quartile along the liquidity distribution. Further supporting the liqudity channel, they find that households in the highest quartile spend only 49ct while the lowest quartile spends almost 80ct of each dollar won in the lottery. However, these two lottery-based approaches suffer from the drawback that they do not measure consumption directly. Instead they have to either construct consumption out of households balance sheet data (Fagereng et al. (forthcoming)) or model consumption as a function of their observed variables (Golosov et al. (2021)).

Gelman et al. (2018) use the government shutdown in the U.S. as a transitory liquidity shock. Hence, contrary to other literature they only estimate how liquidity changes the consumption behavior and not the MPC directly. Still, their setup allows them to disentangle the pure effect a liquidity shock has on consumers spending as government workers receive a payback of their wage once the government shutdown is over. Hence, there are no changes in expected income. Meanwhile, studies using income shocks cannot quantify what effect stems from the liquidity channel and what stems from changes in expected income. Their findings highlight that low liquid households react more to a negative liquidity shock as they have no assets to fall back on. Low liquid government workers started postponing their credit card payments, while simultaneously increasing the amount spend using them. **probably only add very short inside of this as not so much related to raw MPC and little/bad heterogeneity investigation**
The second strand of literature uses survey data from field surveys that question households about potential or actually relaized income shocks and how their reaction looks like. Bunn et al. (2018) use (**use twice here**) data collected by the Bank of England to assess the asymmetry that we expect in households' reaction depending on the sign of

the income shock. As mentioned before, the liquidity channel suggests that an unanticipated shock calls for a stronger reaction if its negative. Indeed, the authors are able to provide ample evidence for such a reaction with their estimated MPC out of a negative shock being being between 5 to 12 times as high as the reaction to a positive shock. The balance sheet data their source provides also enables them to show that borrowers show a more pronounced asymmetry and reaction, which is also in line with the theoretical mechanisms of the liquidity channel. This also holds for households that face some kind of liquidity constraint. Additionally, Bunn et al. are capable of replicating their estimated MPCs in a model with households at the borrowing constraint. These findings are further underlined by Christelis et al. (2019) who use Dutch data for a similar study. Summarizing these studies strongly suggest the existence of the mechanisms related to households at their borrowing constraint and precautionatry saving motives (**the latter must be elaborated on in intro**).

2.1 2008 Tax Stimulus Studies

In using the 2008 tax rebate as the income shock and data collected by the CEX, we are by default closely related to Parker et al. (2013) and Misra and Surico (2014). The former collaborated with the Bureau of Labor Statistics (BLS) to add specific questions to the CEX and provide the first analysis using this data. Misra and Surico (2014) use the same dataset to assess the MPC out of the the tax stimulus applying Quantile Regression. This supposedly allows them to recover the whole conditional distribution of the MPC. However, there are reasons to doubt this claim, which we discuss further down.
Parker et al. (2013) estimate their rebate coefficient using OLS and Two Stage Least Squares (2SLS) estimators. We lay out their identification strategy in 5.1 and also address their motivation to instrument the amount of tax stimulus. Both estimators only allow for a limited investigation of heterogeneity. Parker et al. propopse interaction terms between tax stimulus received and proxy measures for liquidity. They create dummy variables that signal whether household i falls into the lowest, middle or the highest tercile along the liquidity distribution. However, the cutoffs for the terciles are not chosen based on the distribution of the proxy variable but such that each categroy has roughly the same number of tax rebate recipients within a given quarter. Next to simply splitting the sample based on such categorizations and estimating the MPC within each sample, this interaction based approach is quite common in the literature. However, it suffers from the major drawback that the cutoffs to identifiy specific subgroups are exogenously set by the researcher. This harbors the danger that heterogeneity patterns within these subgroups or across smaller subgroups are impossible to find. Consider the case in which heterogeneity is strongest within the lowest tercile of liquidity. Given the medium size of the

liquidity shock, not a large amount of liquid assets is necessary to borrow beforehand and smooth consumption over time. Under the LCPIH, we expect only households with very small amounts of liquid asseets or no access to these at all to not smooth consumption. Therefore, setting the cutoff of liquidity too high for the lowest group can potentially lead to missing out the strongest heterogeneities driven by liquidity.

Finally, we want to briefly address Misra and Surico (2014) and their use of Quantile Regression (QR). QR was developed by Kohnker (FIND CITATION) to estimate the conditional quantile of a distribution and the regression coefficient of variables at this quantile. To do so, QR minimizes the least absolute deviation (LAD) of some $X\beta$ from the outcome instead of the least squares deviation as in OLS or other linear regression methods. The LAD is minimized by choosing coefficients β and to find the coefficients for the τ^{th} quantile of the conditional distribution of the outcome the LAD is weighted with τ and $1-\tau$, respectively. A more detailed explanation is provided by Misra and Surico (2014) or in Kohnker's XXXX paper in which he introduced the quantile regression.

However, the QR approach by Misra and Surico (2014) suffers from a severe misinterpretation of what the QR coefficients represent as well as what underlying assumptions are made for QR to work out.

Issues with QR

- rank-invariance assumption
- the coefficient in a quantile regression shows how much variable x shifts the τ^{th} percentile of the conditional distribution of y
- problem is not that changing x by one unit moves individuals into a different quantile and therefore this doesn't show change for individual
- changing x does not move individuals away from the conditional quantile
- QR point estimate tells us by how much a one unit change in X changes the value of the τ^{th} quantile of the conditional distribution of Y
- it does not show how much INDIVIDUALS at the τ^{th} quantile react to a one unit change in X
- this is only the case when the rank-preservation/invariance condition holds
- parphrasing Angrist and Pischke in their hallmark book 'Mostly Harmless Econometrics': if the point estimate for a low decile is positive that doesn't mean that individuals with low change in consumption previously experience a strong increase in consumption. Instead it shows us that those in the lowest quantile of the distribution with treatment have a larger change in consumption than those in the lowest

quantile of the distribution that have not yet received a rebate. Thus, it does not really identify the marginal propensity to consume because unless we assume rank-invariance, the coefficient doesn't tell us how much individuals (e.g. on average) changed their consumption. The previously described coefficient is not the MPC.

3 Data

We use data collected by the Consumer Expenditure Survey (CEX) that is administered by the Bureau of Labor Statistics. Its main purpose is to provide information on the consumption preferences of US households to adjust the goods basket that is used to calculate various inflation measures (BLS, 2021). However, in an effort to understand the effects of the 2008 tax stimulus, Parker et al. (2013) added questions about these payments to the questionnaire between June 2008 and March 2009. Due to its original purpose, the CEX provides a finely grained set of information on the type of goods households consume. This enables us to analyse on what kind of goods households with a non-zero MPC spend their rebate on. In the following, we briefly outline the stimulus program and describe the CEX data.

3.1 The 2008 Tax Stimulus Program

Due to the global financial crisis and the subsequent recession, the United States government passed the Economic Stimulus Act (ESA) in February 2008. With projected costs of more than 150 billion USD it was the largest relief program passed in the history of the USA up to this point. Next to the stimulus payments, which made up roughly two thirds of the program, the ESA also enacted other steps meant to provide economic relief such as enabling government owned entities (Fannie Mae and Freddie Mac) to buy up more mortgages. However, we only focus on the effects of the stimulus payments.
The rebate was paid out to any household that filed for income taxes. Households that fell beneath the minimum amount of income required to have to file for federal income taxes had to file for taxes anyway and were eligible for the minimum amount of rebate as long as they had a minimum annual income of 3,000 USD **lots of 'minimum' here**. Eligible households received their net tax liability as their rebate, however, the payment were bounded by a minimum of 300 and a maximum of 600 USD. For couples filing jointly the limits were 600 and 1,200 USD, respectively. Parents of children under the age of 17 received additional 300 USD per child. Additionally, the rebate was capped for high income households. The rebate was reduced by 5% of the amount that the reported income exceeded 75,000 USD (150,000 USD for couples), which led the program to target mostly low to medium income households.

3.2 Consumer Expenditure Survey

The CEX is a representative survey of households in the USA interviewing households about their consumption patterns on a quarterly basis. Once a household is selected to participate, they are interviewed a total of five times. The first interview is a baseline

Figure 1: CEX quarterly rotation procedure

Interview year and month		Interview set 1	2	3	4
2015	APR	a			
	MAY	b			
	JUN	c			
	JUL	d	a		
	AUG	e	b		
	SEPT	f	c		
	OCT		d	a	
	NOV		e	b	
	DEC		f	c	
2016	JAN			d	a
	FEB			e	b
	MAR			f	c
	APR				d
	MAY				e
	JUN				f
	JUL				
	AUG				
	SEPT				

Columns show number of interview and a letter signals a specific household. Source: `https://www.bls.gov/opub/hom/cex/data.htm`

interview during which some general household characteristics, employment related variables and their stock of nondurable goods are documented.[1] The next four interviews are administered every three months and households are asked to document their expenditures over the period since the last interview. The final interview collects data on global financial variables such as amounts saved in savings or checkings accounts, which we use as our measures for liquidity. After this interview, the household is rotated out of the CEX and replaced with a new one. Hence, each month of data documented in the CEX contains a different set of households as new ones are added and others are rotated out of the survey. Figure 1 is taken from the CEX website and illustrates this procedure. Note that a household is defined as a Consumer Unit (CU), which can represent either a number of blood or legally related persons (e.g. foster children), a single individual - even if living with other people as long as the individual is financially independent - or unrelated people who are pooling their income. All information about a Consumer Units members are collected regarding their relationship to the reference person. This person is defined as the one named when asked who rents or owns the home. For personal traits such as age we follow the convention by Parker et al. (2013) and take the average of the characteristic of all CU members.

It is important to highlight the limitations set by the usage of CEX data. As mentioned, the main objective of the CEX is to assess what goods the average household consumes to create the goods basket for inflation measurements. This focus results in a lack of interest

[1]The baseline interview has only been conducted until 2015. Since then the first interview covers these questions.

in a dense documentation of household characteristics and income related variables. For example, the lack of asking for liquidity related measures in each quarter prevents us from controlling for changes in liquidity but we can only control for households overall self-reported levels of liquidity. Also, the variables collected are only crude measures for liquidity.

if doing liquidity check then refer to it in paragraph above in the end

While this is a disadvantage in comparison with other data sources, the CEX's richness in information on consumption behavior is unmatched. Keeping in mind the risk of measurement error through the self-reported consumption measurement, the CEX enables us to analyse not only the MPC for overall consumption but to dissect it and see which goods drive responses and heterogeneity seen in higher level estimates.

4 Methodology

To estimate the causal effect of tax rebate receipt on changes in consumption, we use the Double Machine Learning (DML) framework developed by Chernozhukov et al. (2017). This new kind of estimation approach allows to efficiently estimate semi-parametric models of treatment effects using Machine Learning methods. The semi-parametric approach we follow has the major advantage that it does not restrict the effect of confounders on the outcome to a specific functional form. Moreover, specific DML estimators enable us to estimate heterogeneity given observables without defining in which form the observable affects the treatment effect. Past contributions that were looking into heterogeneity had to rely on choosing the correct interactions with observables. Sophisticated DML estimators can detect these interactions without knowing them beforehand.Meanwhile, its implementation procedure deals with common biases arising in more naive estimation procedures that employ Machine Learning methods. This opens the door to combine powerful machine learning algorithms with causal inference. Many ML estimators, such as Random Forests oder Neural Nets, have proven as valuable assets in detecting complex patterns in data.

From a more theoretical perspective the DML estimator yields very efficient properties when it comes to its asymptotic behaviour. Under certain assumptions, Chernozhukov et al. (2017) are able to prove root-n consistency of the estimator, a rate of convergence not achieved in other Machine Learning based estimation approaches. However, we will not further elaborate on these details and refer the reader to Chernozhukov et al. (2017) for a more technical discussion. Instead we focus on the general idea behind the DML framework and the different estimation methods we use in our analysis.

4.1 Setup

We start with considering a Partially Linear Model of treatment and outcome

$$Y_{it} = \theta(X_{it})D_{it} + g(X_{it}, W_{it}) + \epsilon_{it} \tag{1}$$

$$D_{it} = h(X_{it}, W_{it}) + u_{it}, \tag{2}$$

where Y_{it} is the outcome, D_{it} is the treatment and X_{it} and W_{it} are observable variables. We distinct between simple confounders W_{it} which affect the outcome and also potentially the treatment and X_{it}, which additionally are considered to impact the treatment effect of D_{it} on Y_{it}. The choice of these variables is left to the researcher. We also assume that $E[\epsilon_{it}|X_{it}, W_{it}] = 0$ and $E[u_{it}|X_{it}, W_{it}] = 0$.
We are interested in $\theta(X)$, the conditional average treatment effect (CATE). In Rubin's potential outcomes framework (**citation missing**) it is defined as

$$\theta(X) = E[Y_1 - Y_0|X = x]$$

where Y_d is the outcome when treatment is $D = d$. In our setting, treatment is not binary but continuous, hence $\theta(X)$ represents the marginal CATE

$$\theta(X) = E\left[\frac{\delta Y(d)}{\delta d}\middle| X = x\right].$$

The marginal CATE measures how much a marginal increase in the continuous treatment changes the outcome for individuals that have a set of characteristics $X = x$. Note that in our setting we assume that the CATE is linear in treatment, i.e. the treatment effect is independent of the size of treatment. The task is now to find an appropriate estimator $\theta(X_{it})$.

4.2 Regularization bias and how to get rid of it - alternative title: A quest to avoid biases

As Chernozhukov et al. (2017) point out, we could come up with some seemingly straightforward approach to estimate the PLM using machine learning methods. For example, approximating the function $g(X, W)$ with a high polynomial and using a Lasso regression for regularization or use a combination of random forests for predicting $g(X, W)$ and then an OLS regression to find $\theta(X)$. However, any machine learning based approach that follows this notion will suffer from a bias due to regularization. To avoid overfitting and the resulting large variance of the estimator, machine learning methods deliberately induce a bias into their predictions. This bias does not vanish asymptotically, leading to inconsistent results.[2] However, we can deal with this regularization bias using orthogonalization. For this, we define

$$E[Y_{it}|X_{it}, W_{it}] \equiv f(X_{it}, W_{it}) \tag{3}$$

$$E[D_{it}|X_{it}, W_{it}] \equiv h(X_{it}, W_{it}) \tag{4}$$

where (4) follows from (2). It is straightforward to estimate these conditional means using any ML method of choice. Using these and the PLM defined above, we can find

$$Y_{it} - f(X_{it}, W_{it}) = \theta(X_{it})(D_{it} - h(X_{it}, W_{it})) + \epsilon_{it}. \tag{5}$$

Subtracting the conditional means from Y and D is known as orthogonalization and removes the impact of X and W on them, respectively. The residuals only contain variation that does not stem from any of the confounders. In Section 5.1 we discuss what this means in our setting in more detail. Indeed, the estimate of $\theta(X)$ retrieved from estimating the orthogonalized PLM in (5) is no longer suffering from the regularization bias. Excitingly, the authors are able to prove that even in case that the first stage estimators of $\hat{f}$ and $\hat{h}$ are converging at slower rates than root-n to the true parameter value, in the final estimator the regularization bias converges and the estimation error converges to zero at a potential rate of root-n.

In practice, the first stage of the estimation process consists of choosing an appropriate Machine Learning method, predicting the conditional expectation functions f and h and calculating residuals

$$\tilde{Y}_{it} = Y_{it} - \hat{f}(X_{it}, W_{it})$$
$$\tilde{D}_{it} = D_{it} - \hat{h}(X_{it}, W_{it}).$$

A welcome property of the DML estimation is its agnostic to the first stage estimator. Thus, it allows choosing the appropriate prediction method for the given setting.

4.3 Cross- against Overfitting

While orthogonalization takes care of the regularization bias plaguing more naive ML based estimators, it impliclty induces a new bias. Machine Learning estimators are prone to overfitting models. Instead of picking up signals in features to predict the outcome, they start interpreting noise in the training data we feed them. To avoid this behaviour, one can tune hyperparameters of the algorithm of choice to minimize this issue. Still, it is not unlikely that noise in the data is interpreted as a signal.

[2] See Appendix X.X (or only the paper?).

While orthogonalization takes care of the regularization bias plaguing more naive ML based estimators, it implicitly induces a new bias. Machine Learning methods are often prone to overfit, i.e. they start interpreting noise in the data as signals from observables. However, this same individual level noise is contained in the structural error terms of the PLM, ϵ_{it} and u_{it}. Thus, our predictions of f and h are not independent of these. Using the orthogonalized outcomes and treatments to estimate $\theta(X)$ then leads to terms such as $u_{it}(\hat{f}(X_{it}, W_{it}) - f(X_{it}, W_{it})$ to show up in the estimation error $\hat{\theta}(X) - \theta(X)$. The dependence of the structural errors and the prediction errors - both driven by the individual level noise - are then not vanishing asymptotically. Similarly to the regularization bias this lets the asymptotic variance of the estimator explode and prohibits any convergence. However, it is rather easy to resolve this issue using sample splitting - a procedure called "crossfitting."

Instead of using all observations to find the estimates of f and h and then estimate $\theta(X)$ using the whole sample, consider the case in which we split the sample into two. The first sample is used to retrieve the first stage predictions. Those are used to predict the conditional means of the second sample, which are then subsequently used for orthogonalization and the second stage estimation. Since noise is independent across individuals, the noise affecting the first stage prediction error and the structural errors coming into play in the second stage estimation, are independent as well. It is then easy to show that terms leading to problems when using the whole sample are vanishing asymptotically now. In case we are interested in the unconditional average treatment effect (ATE), this procedure is repeated with the role of the samples reversed and the resulting estimators are averaged. However, in the CATE case we are interested in individual-level point estimates. Therefore, while the role of both samples are switched, we do not average any results but keep the individual level estimates of all observations. The cross-fitting procedure for splitting up the sample into any K folds is described in Algorithm 1, which summarizes the whole DML estimation procedure.[3]

4.4 Retrieving the CATE

After retrieving the residualized outcome and treatment, the second stage estimates the conditional average treatment effect as defined in (6). We assume that it takes the following form

$$\theta(X) = \phi(X) \times \Theta, \tag{6}$$

[3]Note that Chernozhukov et al. (2017) argue that K=4 or K=5 performs reasonably well, even for smaller samples.

where Θ is the baseline treatment effect of each individual and $\phi(X)$ is a mapping of our controls X. The form of the latter depends on the estimator chosen for the second stage. In Chernozhukov et al. (2017) estimators are proposed which have a linear second stage, either using a standard OLS estimator or Lasso to regress $\tilde{Y}_{it}$ on $\tilde{D}_{it}$. In these cases, the second stage boils down to a linear regression in which the residualized outcome is regressed on interactions $\tilde{D}_{it}$ and each element of X_{it}. This implies that the treatment effect we estimate is linear in the covariates X. It is also possible to include polynomials of or interactions between different elements of X_{it}. However, we choose a simple linear mapping of X for our linear DML approach presented in Section 5. To identify nonlinearities in the CATE, we use a nonparametric approach that allows us to uncover these without defining them beforehand. Namely, we use a Generalized Random Forest estimator introduced by Athey et al. (2018). It has been developed to take advantage of the powerful random forest predictor for causal inference. Similar to DML, the GRF is an estimation framework. The GRF replaces the original objective function of the random forest algorithm (Breiman, 2001) with a moment condition containing some loss function that can be defined by the researcher. When using it for moment conditions such as (7) to identify conditional average treatment effects, the GRF is also known as a Causal Forest, which is presented in earlier work by Athey and Wager (2016). The Generalized Random Forest framework allows for causal inference as Athey et al. (2018) develop the theory that allows retrieving standard errors of the estimated coefficients. Appendix A elaborates in more detail how the Causal Forest algorithm works and how it identifies the treatment effect. In our case the moment condition is defined as

$$E\left[Y - \phi(X) \times D_{it} - \beta(x)\right] = 0. \tag{7}$$

As part of our analysis we will compare the results to check whether the relationship is indeed linear or whether we discover non-linear heterogeneities that the linear DML approach does not account for and have not been considered in the literature yet. However, note that when using a nonparametric second stage the convergence rate of the estimator declines. This implies that the Causal Forest based approach is more demanding when it comes to the number of observations.

This has to look better and be more 'algorithmic'.

Algorithm 1 Double Machine Learning Estimator

1. Split up sample into K folds.
2. To estimate $\widehat{h}$ and $\widehat{f}$ for for the k^{th} fold use observations $j \notin k$.
3. To get residuals for observations in k, calculate $\widehat{h}(X_i)$ and $\widehat{f}(X_i, W_i)$ for $i \in k$ and use to retrieve residuals.
4. Once residuals of each fold retrieved, estimate $\theta(X_i)$.

5 Estimation and Results

We implement the partially linear model as presented in Section X.X to estimate the effect of receiving a tax rebate of R_{it} on change of consumption ΔC_{it}

$$\Delta C_{it+1} = \theta(X_{it})R_{it+1} + g(X_{it}, W_{it}) + \epsilon_{it} \tag{8}$$

$$R_{it} = h(X_{it}, W_{it}) + u_{it} \tag{9}$$

Which variables are included as confounders X_{it} and W_{it} and which make up X_{it} depends on our specification and is named in the section discussing the respective results. In each specification we include monthly dummies to account for seasonality. They also capture any unobserved effects that might appear when households learn about the rebate, i.e. in line with Parker et al./Misra and Surico our estimation uncovers the effect of actually receiving the rebate and not the global rebate effect.

5.1 Identifying the Income Shock

Since we use the same data and event to estimate the MPC our identification is based on the approach by Parker et al. (2013). The main factor is the design of the stimulus rollout, which we can exploit to identify the income shock. The tax stimulus was paid out to households over several weeks as administrative and technological restrictions made it impossible to pay out all rebates at once. Instead the date of rebate receipt depended on the last two digits of tax filers' social security number. Therefore, we observe rebate receipts at different points in time, which allows us to use all other households that received their rebate in a different quarter as the control group.

the following is probably to harsh self-criticism and too much into econometric stuff; on the other hand the definition of the control group is quite crude This definition of the control group is obviously problematic. Question: deal with this here or have a paragraph in the end that deals with shortcomings of the identification section? No matter where, it should go something like this: This definition of a control group is potentially problematic as it also includes households that already received their rebate in a prior period. This might bias our results if there are long term effects of receiving the rebate that spill into the next period. Also, it leaves the consumption change from t-1 to t in the control group biased in case households receiving the rebate in t-1 actually respond positively to the rebate. Then the change in the control group is likely to be negative as in the next period households resume their smoother consumption pattern and therefore show a negative consumption change although this is due to the control group. This biases the effect of receiving the rebate upwards as even when reactions are small the divergence is larger than the reaction itself (thisreads shaky and I am not sure whether I should include such harsh criticism of my own work here). However, Parker et al. (2013) (or was it MS?) include a one period lag of rebate in their analysis and find no significant role of it in determining consumption change.

In the following, we will slightly depart from Parker et al.'s identification strategy given their findings as well as our inclusion of more control variables. They argue that using the actual amount of tax rebate received can lead to an omitted variable bias. This concern arises because of how households' stimulus payments are determined. First, they depend on the number of children as each dependent child adds 300 USD to the stimulus amount. Second, the stimulus excluding the child bonuses equals the household's net income tax liability (NTL; in the following also referred to as the net tax liability) as long as it is within the exogenously defined boundaries we discussed in Section 3.1. Parker et al. now argue that the NTL might also drive changes in consumption, rendering the treatment endogenous. Their solution is to instrument the amount received with a dummy variable that only signals whether the stimulus was received or not in the given quarter. While their results and the authors themselves suggest that this is not much of a concern, we decisively disagree with their identification approach. With respect to the first concern, the number of children is reported in the CEX and is easily controlled for as it is collected in each interview conducted. The role of the NTL is more complicated. Parker et al. do not control for any variable related to households income or salary, These variables are without a doubt directly connected to our treatment because the NTL is determined by income. Exlcuding these variables leads to an omitted variable biases that leads to inconsistent estimates. However, other than through the channel of income, we deem it highly unlikely that the net tax liability itself is driving changes in consumption. It might be possible that in other years the NTL plays a role for households income as it can be perceived as an anticipated income - or liquidity - shock ([4]) However, in 2008 the NTL affected households via their tax rebate, i.e. it does not affect the consumption change through other channels than what is captured by the tax rebate. Therefore, we argue in favor of using the actual rebate amount since it has two advantages: for one, we have an additional source of variation and second it allows us to estimate the continuous treatment effect and interpret it as the actual MPC in dollar amount.

However, one drawback of the already mentioned lack of detailed documentation of household characteristics in the CEX is the fact that once we include financial variables - most importantly liquidity and salary - our sample size shrinks because they are not consistently documented for each interviewed households. Although the DML framework achieves fast convergence rates even in cases in which the first stage predictions do not converge as rapidly, we have to keep this drawback in mind. However, contrary to Misra

[4]Households usually should know that they will have to pay this/receive this because of past experience and because the NTL is also depending on how much income tax was already paid during the previous year.

and Surico, we actually recover the conditional distribution of the MPC - obviously under the assumption that we control for any relevant confounders **(NOT A SENTENCE YET AND THINK ABOUT THIS IN MORE DETAIL)** - ****and contrary to Parker et al. we do not have to rely on defining our own cutoffs to detect heterogeneities.

5.2 Main Results

In total, we estimate four different specifications for each outcome. By specification we refer to which variables we include as confounders X and W whereas we additionally distinct between the estimator used for the second stage. Thus, we estimate each specification-outcome pair twice, once using the linear and once using the causal forest (cf) second stage. Our estimation procedure enables us to recover individual level MPCs and standard errors. Therefore, we are able to test for each household whether their MPC out of the tax rebate is statistically significantly different from zero. We analyse our results in several steps but we begin with looking at the empirical distribution of the estimated MPCs. Figure X.X shows the distribution of MPCs for the four expenditure categories Parker et al. look at: Food (FD), Strictly Non-Durables (SND) as defined by Lusardi (1996), Non-Durables (ND) and Total (TOT) expenditure. The sub-components of each category are listed in Appendix A.A. Let us point out that the difference between SND and ND consumption categories are that the non-durable category also includes components that are referred to as 'semi-durables' such as health expenditures and apparel.
A single plot of the empirical distribution in Figure X.X is retrieved as follows: we bin the range between the lowest and highest point estimate of the MPC into 20 equally large bins and calculate the share of point estimates that fall into this range. The x-axis signals the borders of the different bins and the y-axis shows the respective frequency where the blue bar signals what the total frequency of this bin is. Additionally, we calculate for each bin how many of the individual MPCs that fall into it are statistically significant at the 10% level and calculate the share of statistically significant estimates **within** this bin. This is depicted by the read overlay over the frequency bars. I.e. a completely red bar implies that all observations within this bin are statistically significant whereas a bat that is only red up to half of its height signals that only half of the point estimates within this bar are statistically significant. The vertical dashed line marks the average CATE - i.e. the average treatment effect across all households as a benchmark. The description notes whether this ATE is significant or not.
First, we have a look at common trends across all specifications and expenditure categories before we start taking a closer look at each category and estimation procedure.
Across all specifications, outcomes and estimators we find a strong pattern of heterogeneity which underlines the importance of accounting for MPC heterogeneity. The heterogeneity

looks similar to what is suggested by Kaplan and Violante (2014) and Misra and Surico, with a large number of households showing no significant response to the tax rebate and the point estimates being most densely distributed close to zero, but a certain fraction of households (CHECK WHAT RATIOS ARE!!!) in our test set are experiencing strong, significant responses. Most importantly, we find that the ATE is always very close to zero but the individual point estimates show a completely different pattern. We see that the ATE falls into bins that have the (almost) highest frequency - which makes sense by construction - but note that across all estimations the respective bin never contains more than 15% of all point estimates. This highlights the weak representativeness of the ATE and its inability to reliably assess the success of programs such as the 2008 tax stimulus. In general we see that introducing more controls to our estimation reduces the spread of the point estimates across all specification-outcome pairs, indifferent to which second-stage estimator is used. The change is the most pronounced once we add liquid assets, income and salary as confounders to the estimation.

Most notably, the responses we find for total expenditures are in part unreasonably large for a bulk of individuals. As we will later see this is probably due to the underlying composition of the total expenditure variable and our estimation approach. It is quite likely that some outliers within a specific spending category part of TOT are driving the learning behavior of our estimators. This is underlined by the fact the the causal forest estimator finds way larger responses than the linear based estimator because the causal forest cannot handle these outliers as it has not seen them before (check how much is spent on cars etc in test sample). Also, it is important to note that in both estimators the spread of the CATE is drastically reduced once we control for liquidity, salary and income - variables we expect to be closely related to the MPC. Turning to the significance of the response, we see that adding more controls also reduces the amount of significant MPCs found, suggesting that prior specifications picked up signals of the confounding factors not included and interpreting them as signals of the rebate. Concluding, it seems that our estimation procedure is quite sensitive to extreme outliers in the underlying consumption categories as for total expenditure we find extreme responses as laid out above. This notion becomes more clearly when we turn to the non-durable and strictly non-durable goods. Excluding large durable categories such as *new vehicles* immediately reduces our estimated MPCs and their spread. Moreover, looking at the ND category we see that in specification 3, which includes liquidity, the linear model fails to reject the null for all households. Interestingly though, the causal forest model still finds a small fraction of large significant MPCs. This difference suggests that there are non-linearities in the dependence of the MPC on the variables such as liquidity - and potentially with respect to their interactions - that are ignored by the linear model but picked up by the causal forest. Accounting for these non-linearities reduces the noise in the point estimates and reveals significant MPCs where the linear analysis fails to pick up any significant MPCs and it seems as if only liquidity is the underlying driver. Comparing this to the Strictly Non-Durable category there is little difference in the distribution and significance of the recovered MPCs.

6 Understanding the roots of heterogeneity

In the previous section we discussed the conditional average treatment effect of each individual given their specific set of characteristics. Similarly to prior contributions, we also looked at correlations between the significance of the estimated MPC and households characteristics to get a glimpse into which factors play a role in the MPC. However, this approach does not reliably tell us which variables really drive the response. The correlation might very well be spurious or driven by other factors that are correlated with the characteristic we are looking at. Therefore, it is more fruitful to look at measures that can help us identify what role a variable plays in our predicted MPCs. In case of specifications using the linear DML estimator, we know that this relationship is linear by construction since the CATE is defined as a linear combination of the single effects of interactions between treatment and the respective variable (see equation (6)). However, the causal forest based approach will help us reveal whether there are any non-linear patterns underlying in the effect of characteristics on the MPC without assuming any functional form of these patterns.
For this, we turn to the Machine Learning literature, which has developed a number of tools to analyse the relationship between prediction and feature. Feature is a different term for control variables. In our setting these are the variables we condition on to find the CATE, i.e. variables contained in X. Since variables in W are assumed to not impact the CATE they are not contained in the second stage and therefore play no role in predicting individuals' MPC. Machine Learning estimators such as random forests are blackboxes as they only provide predictions but stay quiet on which variables are important to arrive at this prediction. The literature has proposed multiple approaches that help quantify the role of a single feature, some of which we look at in the following.

6.1 Marginal and Partial Dependence Plots

Two popular approaches are marginal plots (M-Plots) and Partial Dependence Plots (PDPs; Friedman, 2001). Both use the same general idea to quantify the impact of some feature x_S on our predictions: we replace the value of x_S of each observation with some value v_1. Then we fit our trained prediction model to this "counterfactual" dataset and take the average over all these predictions. For example, we predict for each individual what their MPC looks like if they had a certain age and average the predicted MPC. Then we continue with $x_S = v_2$ and so on, where the values v_j are chosen from a grid along the distribution of x_S. The difference between M-Plots and PDPs is the distribution of all other features $X_C = X \setminus x_S$ we average over. In case of Marginal Plots, contrary to what the name might suggest, we use the conditional distribution of X_C given x_S, $p_{X_C|x_S}$, to obtain the impact of x_S on our prediction

$$\hat{f}_M(x_S = v_j) = \int p_{X_C|x_S=v_j} m(x_S = v_j, X_C) dX_C, \tag{10}$$

where m is our predictor and $\hat{m}_M(v_j)$ is the effect at $x_S = v_j$. On the other hand, PDPs use the marginal distribution of X_C, p_{X_C},

$$\hat{f}_{PDP}(x_S) = \int p_{X_C} f(x_S, X_C) dX_C \tag{11}$$

where $\hat{m}_{PDP}(v_j)$ is the Partial Dependence of our predictor on x_S at v_j. Using the marginal distribution of X_C effectively "marginalizes out" the effect of any other variables than x_S at some point v and therefore reveals what impact x_S has on our prediction at this point. Partial Dependence Plots are more common in the Machine Learning literature as M-Plots suffer from a severe weakness when features in X_C are correlated with x_S. However, PDPs also fail to reliably uncover the effect of x_S in such a setting.

To illustrate the issues arising in M-Plots and PDPs when features are correlated, let us consider a simply example. Lets say we have some predictive model m that only depends on two predictors x_1 and x_2, which are positively correlated. To now calculate the M-Plot of x_1 at v_1 we use the conditional distribution $p_{x_2|x_1}$. In practice, we plug in $x_1 = v_1$ for each observation that is within a specified neighborhood of $x_1 = v_1$ (e.g. observations in the same quantile). Then we predict and average to obtain the M-Plot value at $x_1 = v_1$. Repeating this procedure for other values v_j then results in the M-Plot of x_1. However, because the two variables are correlated, we do not know which variable drives the observed effect - if x_1 is increased, the values of x_2 we use for our predictions also increase because we only use x_2 of observations that are close to having $x_1 = v_j$. This problem is known as 'conflation'.

On the other hand, Partial Dependence Plots do not suffer from this problem because they use the marginal distribution of x_2. We use all observations of x_2 instead of only looking at a neighborhood in which $x_1 = v_1$ and, therefore, do average out the effect of x_2 on our predictions. Since we use the same set of x_2 values at each point v_j, we know that changes in our predictions must stem from x_1. Still, the PDPs are not a good tool when features are correlated and this is connected to the machine learning estimators we apply them to. These are nonparametric estimators that are usually very weak in predicting

outcomes based on observations they have never seen before. This extrapolation however becomes necessary when we create the "counterfactual" dataset by setting $x_1 = v_j$. By doing so, we effectviely create observations that are extremely unlikely or even impossible to observed in the real world because of the correlation the features. For example, in our data age and salary are strongly correlated, which is quite intuitive because once retired, households do not receive a salary anymore. When creating PDPs we ignore this fact and create households that have a high salary and are very old. The weakness in extrapolation leads the model to create weak predictions, which then severly bias the Partial Dependence Plots. (Apley and Zhu, 2020)
Therefore, while PDPs do not suffer from theoretical drawbacks like M-Plots, in practice they are unable to uncover the effects of x_1 on our predictions in a stable manner because of the underlying predictive estimator. If the true model is indeed linear and we use a linear prediction method with the correct specifications of any interaction terms etc., then this extrapolation issue is unlikely to occur. Moreover, by construction, a linear predictor will results in linear Partial Dependence Plots. **Remember that in our linear DML approach, we assume that the CATE we estimate is linear in features X and the second stage - the fitted model we actually investigate here - is simply a linear regression, which results in a linear PDP by construction as our predicted MPC is simply the sum of all coefficients for individual i given their characteristics. $\rightarrow$ I am not so sure about this part yet**
Indeed, results of the partial dependence plots are rather spurious. They are reported in more detail in Appendix X.X, where we look at the PDPs for non-durable consumption. The effects have a high variation and point estimates out of line of the existing literature. **this is not a good reasoning of why I don't show them because they are too close to the ALEs - I guess**

6.2 Accumulated Local Effects

To circumvent issues arising in M-Plots and PDPs from correlated features, Apley and Zhu (XXX) propose Accumulated Local Effects (ALE). The extrapolation issue PDPs suffer from is bypassed by using the conditional distribution $p_{X_C|x_S}$ as we do in M-Plots. As with M-Plots we use the conditional distribution to bypass the extrapolation issues that PDPs suffer from. The 'conflation' effect that results from this is tackled by not using average predictions at $x_S = v_j$ but rather the average marginal change in predictions at this point. We apply this by using the partial derivative of our predictor m with respect to x_S at the point v_j. Although many machine learning methods such as tree based learners have no concept of a gradient, Apley and Zhu are able to derive proofs for non-differentiable functions m (see Section X.X) and further does this not play a role when it comes to the

ALE estimation. The ALE is then defined as

$$\hat{f}_{S,ALE}(x_S) = \int_{z_{0,S}}^{x_S} E_{X_C|X_S=x_S}[\hat{f}^S(X_S, X_C)|X_S = z_S]dz_S - constant,. \tag{12}$$

Looking at this equation step-by-step reveals how the ALE recovers the effect of x_S on our predictions even when features are correlated. As already mentioned, the ALE avoids 'conflation' by using the partial derivative of m, where we have $m^S = \frac{\partial m}{\partial x_S}|_{x_S=v_j}$ as the partial derivative of m evaluated at the point we want to find the ALE for. Since we only look at an infinitesimally small change, this change in x_S will not affect the features that are correlated with it in X_C unless the correlation is extremely high. In our analysis we would want to avoid this case anyways to avoid problems in the estimation itself (e.g. multicollinearity). Once the changes in prediction are obtained for each observation, we average them over the conditional distribution, i.e. only using observations that are within a neighborhood of $x_S = v_j$ and actually exist. Now we have the average local effect, but we are interested in how x_S affects our predictions and not how it affects changes in predictions. Thus, we simply integrate over all local effects up to $x_S = v_j$, where $z_{0,S}$ is the lower bound of the distribution of x_S.

To estimate the ALE we use the following estimator, which illustrates the procedure in more intuitive terms:

$$\hat{\hat{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_{i,j}\in N_j(k)} [f(z_{k,j}, x_{i,\setminus j}) - f(z_{k-1,j}, x_{i,\setminus j})] \tag{13}$$

The intuition behind the estimator is straightforward. First, we bin our data into n_b bins based on quantiles of the distribution of x_S. To mimic the marginal change represented by the partial derivative m^S in 12 we make two predictions for each individual. For an observation i that falls in bin k, we predict its outcome with $x_S = z_k$ and $x_S = z_{k-1}$, where z_k represents the upper bound value of quantile k. We then averages over all individuals that fall within this bin k and finally accumulate all predicted differences from the lowest bin up to bin k. Only looking at individuals within a neighborhood $N(k)$ - effectively observations in the same quantile - accounts for the conditional distribution used in 12.

As a last step, Apley and Zhu propose to center the effect around the average of all ALEs such that the mean effect is zero. Thus, the ALE has to be interpreted relative to the average prediction and it shows whether for a given $x_S = v$ the effect of x_S is above or below the average prediction. I.e. whether x_S affects our predictions at $x_S = v$ more than it does on average. In practice, the *constant* in (12) is replaced by

$$averagetermhere$$

.

Note that we yet cannot say something meaningful about the statistical significance of these results. Most fields are only interested in the predictive power of machine learning methods and to understand how these predictions are achieved, but there is no notion of statistical significance in these settings. Therefore, a specific approach to quantify uncertainty of these measures has not yet been developed. A deeper look into this topic is, however, out of the scope of this work. To briefly dive into the topic of statistic significance, we use a bootstrapping based approach. We simulate the ALEs for $n_{bootstrap}$ samples. These create an empirical distribution (need at least e.g. 500-1000) on which basis we calculate pseudo-standard errors. Figure X.X reports the Confidence Intervals using the reverse percentile approach (see Appendix X.X) and the mean point estimates in each bin. We see that these bands are very wide in certain parts - especially in areas where there is a small number of observations. We strongly encourage a deeper investigation of the statistical properties of ALEs and a potential way of quantifying their uncertainty in a more rigorous way. While the ALEs show us the relationship between a specific variable and our predictions, understanding whether this relationship is statistically significantly different from playing no role would be a major improvement.

One weakness of ALE is further that they are a global measure, i.e. they do not help to uncover heterogeneity in the reaction of the prediction to a single variable. One method that can unbeil such heterogeneities are Individual Conditional Expectations, but they suffer from the same conceptual problems as Partial Dependence Plots. Next to investigating the role of uncertainty in the measures presented in this section we therefore also urge the development of theoretical foundations of such a measure. For now, we account for heterogeneity by plotting the unaveraged ALEs of each individual. To avoid overplotting, we only plot the households at the quintiles of the ALE distribution. (**delete last part of this if not doing it!!**)

6.3 Results

Figure X.X plots the Accumulated Local Effects.

7 Conclusion

(from intro 2 draft but doesn't fit that much anymore) Our contribution is twofold: for one, we estimate the conditional MPC out of the tax stimulus in the most precise and rigorous manner thus far. Second, we use an estimator that exploits the power of machine learning methods for causal inference and contribute to the wider understanding and promoition of this method among applied researchers. Machine Learning predictors are powerful tools when it comes to handling large data and/or complex relationships between variables without any specification of those.

References

Bunn, P., J. Le Roux, K. Reinold, and P. Surico (2018): "The consumption response to positive and negative income shocks," *Journal of Monetary Economics*, 96, 1–15.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2017): "Double/Debiased Machine Learning for Treatment and Causal Parameters," *arXiv:1608.00060 [econ, stat]*, arXiv: 1608.00060.

Fagereng, A., M. B. Holm, and G. J. Natvik (forthcoming): "MPC Heterogeneity and Household Balance Sheets," *American Economic Journal: Macroeconomics*.

Golosov, M., M. Graber, M. Mogstad, and D. Novgorodsky (2021): "How Americans Respond to Idiosyncratic and Exogenous Changes in Household Wealth and Unearned Income," Working Paper 29000, National Bureau of Economic Research, series: Working Paper Series.

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

Parker, J. A., N. S. Souleles, D. S. Johnson, and R. McClelland (2013): "Consumer Spending and the Economic Stimulus Payments of 2008," *American Economic Review*, 103, 2530–2553.