Dissecting Demand-Side Welfare Manipulation*

Brendon McConnell Jaime Millán-Quijano

This draft: March 25, 2025

Abstract

Optimal targeting of social aid is a fundamental issue in public policy design. A key aim is to create welfare systems that are manipulation-proof. Using rich administrative and survey data, along with a fuzzy difference-in-discontinuities design, we study manipulation of eligibility for a nationwide welfare program in Georgia. Large drops in benefits at key thresholds of welfare scores coupled with the ability to request reassessments provide households with both the means and incentives to manipulate in this setting. We (i) document substantial levels of manipulation at a key benefit discontinuity, (ii) characterize manipulating households, (iii) document how these households engage in manipulation, and (iv) provide evidence on the downstream consequences of manipulation. We find that successful manipulation crowds out labor market participation for women, but not men. Finally, we provide a lower bound estimate of the cost of manipulation at 25% of the initial expenditure on our target households.

Keywords— Welfare Eligibility, Manipulation, Public Policy Design, Bunching
JEL Codes— D1, E21, H24, H53, I38

*Author affiliations and contacts: McConnell (City St George's, University of London, brendon.mcconnell@gmail.com); Millán-Quijano (NCID and CEMR, jmillanq@unav.es). We are grateful to the evaluation team at Econometría S.A., UNICEF Georgia, GeoStat and the Social Services Agency of Georgia. We are also thankful to Dirk Foremny, Francesco Loiacono, Alan Manning, Anastasia Terskaya, and Marcos Vera-Hernández, and the participants at the University College London, Southampton University, Unviersitat de Barcelona, University of the Basque Country, NOVAFRICA, University of Gottingen, and the Institute for Fiscal Studies, for their valuable comments. The authors acknowledge financial support from the Spanish Ministry of Economy and Competitiveness Grant PID2020-120589RA-I00.

1 Introduction

Optimal targeting of social aid is a key issue in the design of public policy, irrespective of a country's level of development (Coady et al., 2004; Alatas et al., 2012). While in developed economies targeting uses rich administrative data, in developing economies proxy mean tests (PMTs) are commonly used to allocate access to social programs. In both cases, the design of targeting schemes must take into account the response of potential beneficiaries to the program features, not least the incentive to game or manipulate the system (Coady et al., 2004).

We can divide the source of social benefits manipulation into two categories. First, *demand-side* manipulation, where manipulators are the final beneficiaries and engage in manipulation for their direct gain. Examples include tax evasion (Friedberg, 2000; Saez, 2010; Kleven et al., 2011; Kleven and Waseem, 2013), access to health services (Miller et al., 2013), and local government corruption (Foremny et al., 2017). The second category is *supply-side* or *intermediary* manipulation, whereby an intermediary or service provider manipulates access to, or elements of, the program. Examples include teachers changing student grades in high stakes tests (Diamond and Persson, 2016; Machin et al., 2020), up-coding in health insurance (Geruso and Layton, 2020), vote-buying behavior of local governments (Camacho and Conover, 2011; Brollo et al., 2020), and employers and employees colluding to take advantage of social welfare schemes (Van Doornik et al., 2018; Citino et al., 2023)

In this paper we study the case of social welfare eligibility manipulation in the Targeted Social Assistance (TSA) program in the nation of Georgia. We start by providing evidence of demand-side manipulation in the TSA program – a nationwide program that uses a proxy means test (PMT) with multiple cutoffs to allocate unconditional cash transfers among low-income households.[1] When the household's situation changes – e.g., a child is born, a household member dies, a household member becomes disabled, the household purchases a car – the household is re-assessed by the Social Security Agency (SSA) and a new PMT score is calculated. We label these *SSA-initiated* reassessments. There is a second route to a welfare score reassessment – *household-initiated* reassessments – whereby households who feel their PMT score does not accurately represent their level of welfare may request a new assessment at least one year after their initial assessment. It is these household-initiated reassessments that will be the core focus of our work on welfare eligibility manipulation in this paper.

To document manipulation we present graphical evidence comparing the initial and final observed distributions of the PMT score, paying special attention to a key threshold (a PMT score of 65,000). This threshold is a focal point for manipulation, as it is associated with an unusually large drop in welfare payments, three times larger than associated drops at nearby thresholds. We supplement the visual evidence with a formal density test (Cattaneo et al., 2020). Using the rich administrative data we have available to us, we classify reassessments into two categories – household-initiated reassessments, and social security agency-initiated reassessments. This enables us to understand the *source* of any PMT score discontinuities.

[1]The PMT score is based on information from both rich, multi-agency administrative data and data from a household visit. The score incorporates information on households' demographic composition, asset holdings, income, and access to public amenities.

It is important to note at the outset that the PMT score is highly complex, and it is unlikely to be perfectly known by – or perfectly malleable to – households aiming to improve their score.[2] We thus consider our setting as one of *manipulation with imperfect control* of the welfare score. In this setting, households cannot directly determine their welfare score. They can however attempt to influence the score, by first requesting a reassessment and then hiding/changing assets prior to the household visit. Informed by the nature of manipulation attempts, we set up a model of household welfare manipulation based on the insights of the Becker model of crime (Becker, 1968). The purpose of this model is twofold. First, it informs the structure of our empirical specification. Second, it makes clear the importance of household heterogeneity in unobserved willingness to manipulate, a key point when bridging with our empirical strategy and our discussion of our marginal treatment effect (MTE) results, which we discuss in greater detail later

Based on the evidence of behavioral responses of households to discontinuities in the PMT score schedule, we set up an empirical specification in the form of a fuzzy difference-in-discontinuity (FDD) design (following Millán-Quijano, 2020).[3] Our key endogenous treatment variable is the decision to engage in welfare manipulation, and our instrument is a binary indicator for receiving an initial PMT score just above 65,000 – the focal threshold for manipulation, and thus our key threshold of interest. The reason we use a difference-in-discontinuity design is that welfare benefits also change at this threshold – hence, we have two endogenous variables changing at the focal threshold. Our FDD strategy therefore incorporates variation around a neighboring cutoff to account for the effect of changes in benefits on key outcomes. This approach enables us to cleanly estimate the causal effect of welfare manipulation attempts on household outcomes, while instrumenting for the generosity of welfare benefits.

We then restrict the data to a narrow window around the key PMT score thresholds in order to consider a relatively homogeneous group of welfare recipients. The households in our working sample are poor, with a total income (labor income plus all welfare benefit income) of roughly 100 USD per month. We provide supportive evidence for the identifying assumptions required for this approach. First, we show that our running variable is continuous through the key cutoff of interest. Second, we document the continuity of household characteristics and of key policy parameters (which may reflect both household observables and unobservables) through the cutoffs. The evidence we document provides strong support for the continuity of potential outcomes around key thresholds. We provide additional evidence that the effect of additional benefits is homogeneous over the PMT distribution, the key complementary assumption of the FDD framework. We buttress this evidence with further information regarding our ability to rule out additional threats to identification, including ruling out inspector collusion.

Our FDD approach is analogous to instrumenting for the likelihood of a manipulation at-

[2]The score is a non-linear function of ten separate indexes with multiple inputs and location-specific factor loadings. The score is also not instantaneously known by any of the involved parties, including the welfare inspector. After score inputs are recorded at the interview, the final score is released on average 34 days after the household visit. This allays concerns of inspector corruption.

[3]Our FDD design is similar to the idea behind dynamic bunching (Garbinti et al., 2023; Marx, 2024). These papers use time variation around the kink to circumvent an identification problem when using only a single kink (Blomquist et al., 2021).

tempt – proxied by the likelihood of a household-initiated reassessment – using changes in the welfare scheme at a focal welfare score threshold. We characterize the compliers in our framework – households whose manipulation status is induced by falling just above the 65,000 PMT score threshold – using the approaches of Abadie (2003) and Dahl et al. (2014). In addition to characterizing compliers across a wide set of household and property characteristics, we can directly examine where complier households fall within the baseline income distribution. For each of the non-income characteristics, we present the partial correlation between the characteristic and baseline earned income in order to assess if the characteristic in question is positively or negatively associated with income.

We then consider *how* households manipulate their scores. To do so, we exploit the richness of our administrative data, which includes every input that contributes to one of the ten indexes that compose the PMT score. We compare changes between interviews in the constituent index scores for households who request a repeat interview – our manipulation attempt proxy – with households that have repeat interviews triggered by the SSA. We use this latter group as a comparison to capture changes in the various indexes that occur over time, either due to aggregate time effects, or other random shocks. This allows us to better understand how households attempt to manipulate their final welfare score with imperfect control.

Next, we document the consequences of welfare eligibility manipulation attempts on a wide set of outcomes using both multi-agency administrative data and survey data we collected at the household level. We make use of detailed information on the timing of both interviews and all relevant outcomes. This allows us to consider only post-manipulation outcomes. We start by considering the impact of welfare eligibility manipulation on labor market outcomes – both formal and informal – and proceed to study the consequence of welfare manipulation for expenditure patterns within the household. Finally, we consider the impact of household welfare manipulation on specific household members, namely children and young people.

We estimate MTE parameters for a series of labor market outcomes. We follow Brinch et al. (2017), who show how to identify the MTE in the presence of a binary instrument using the separate MTE estimation approach developed by Heckman and Vytlacil (2007). In practice, we estimate linear MTEs using the separate approach. Our MTE analysis enables us to gain a better sense of selection into manipulation, and connects our theoretical model (where we model households' willingness to engage in manipulate) to our empirical work.

We complete the paper with two exercises. In the first exercise, we provide an estimate of cost of manipulation to the Georgian government. In the second exercise, we compare our estimates of the extent of welfare eligibility manipulation to those we would have estimated, using two distinct bunching estimators, had we been in a less data-rich environment. These comparisons enable us to get a sense of the external validity of our work, and to place our findings within the wider, bunching estimator-based, manipulation literature.

Our first key empirical finding is to document substantial manipulation of welfare eligibility. We present graphical evidence that (i) the initial distribution of PMT scores are smooth and continuous through all key benefit cutoffs and (ii) the final observed distribution has pronounced and unnatural bunching at the key PMT threshold 65,000. We confirm the visual evidence of

bunching with a CJM density test (Cattaneo et al., 2020). We then present the probability of a PMT score reassessment across the PMT score distribution. We find a large, statistically significant jump in the probability of reassessment precisely at the PMT score threshold of 65,000 and nowhere else. The probability of a reassessment is approximately 35% below the 65,000 threshold. We document a 20 percentage point jump at 65,000. Using the information on the source of the reassessment, we find that household-initiated reassessments are the exclusive driver of the total effect. The probability to be reassessed jumps precisely at 65,000 by 20 percentage points (from a base of 10 percentage points) for household-initiated cases, yet social security agency-initiated reassessments are smooth through all PMT score-based cutoffs.

Our evidence points to a case of demand-side manipulation – households' actions fully explain the discontinuity in the final PMT distribution at 65,000. We find no evidence that supply-side actors, namely the SSA or the enumerators that perform the PMT assessments, are the source of manipulation. First, the first PMT score is continuous around the cutoffs, and SSA-initiated reassessments are also continuous around the cutoffs. Hence, without the direct action of the household, the PMT distribution remains undisturbed. In addition, one could think that the SSA or the enumerators may be more yielding (or harsh) with households just above the cutoff. However, we provide evidence that both the probability of success and the probability of being caught cheating following a manipulation attempt are also continuous through the key threshold.

Our second finding uses the approaches of Abadie (2003) and Dahl et al. (2014) to characterize the complier households within our working sample.[4] We find the compliant sub-population – those who change their manipulation status based on their PMT score in relation to the 65,000 threshold – are more likely to live in rural areas. For instance, complier households are 24% more likely to own livestock, 22% more likely to have land for agriculture, 59% more likely to own a car or tractor, and 39% more likely to own a cattleshed or granary. These households are also relatively poorer in terms of baseline earned income. This finding is particularly useful from the perspective of optimal policy design, as it shines a light on the type of households who respond to the specific design of the welfare program we study. The finding suggests that manipulation attempts are economic-needs driven, a point to which we return below.

We next consider the way in which households manipulate their PMT scores at a repeat interview. Compared to households with an SSA-initiated repeat interview, households who initiate the repeat interview themselves have significantly fewer moveable rural assets (livestock) and significantly less movable property (cars and tractors) at the second interview. When combined with our work characterizing manipulating household, which highlights that complier households are more likely to be based in rural areas, the evidence here suggests that selling, or misreporting rural assets, is likely a key strategy for welfare score manipulation. We find no such evidence of between-interview changes in harder to move/misreport assets, such as land, conditions of household premises, or household demographics. Interestingly, the results from this section also highlight that households that initiate a repeat interview do not have any

[4]In our setting, a complier household is one which attempts to manipulate their welfare scores if their initial PMT score just exceeds the focal PMT threshold of 65,000, but does not if their score is just below this threshold.

significant changes between interviews in income or employment indexes. This makes clear that it is not a negative labor market shock that is driving the manipulation attempt, but rather a response to falling just above the focal threshold.

We then present evidence on the consequences of welfare manipulation on post-manipulation outcomes. Our FDD results show that women in households that engage in welfare manipulation work *more* in the formal labor market, but their income, and the household total income, do not increase significantly.[5] This is a somewhat surprising finding. To unpack this result, we split households by those with successful and unsuccessful manipulation attempts. We find that it is women in households with unsuccessful manipulation attempts that drive the increase in labor market participation. We do not find such effects for women in households where the manipulation attempt is successful, pointing to a crowding out effect of welfare income on labor market engagement. This finding once again suggests that the decision to attempt welfare manipulation is needs-based – women in households with unsuccessful attempts supply more labor. This concept of manipulation being economic needs-based is further buttressed by our MTE findings. Here we document negative selection (into a manipulation attempt) on unobserved gains in labor market outcomes suggest the women are not pulled into the labor market based on the high (unobserved) gains, but rather pushed into doing so based on the need to bolster household income.

We additionally study the expenditure response to welfare manipulation. A key element of our household survey involved collecting detailed household expenditure data. The FDD estimates for total expenditure are positive, but imprecisely estimated. Where we find the largest increases in expenditure is on children – this total expenditure figure comprises increases in child clothing. We find no effects on food expenditure, including eating out, nor do we find any effect of increased expenditure on alcohol and tobacco.

Given that households that manipulate their welfare eligibility status spend their additional income almost exclusively on children, we focus our attention for the remainder of the paper on the outcomes of children and young people in the household. Our setting is interesting from the perspective of childhood skill investment in that, as a consequence of manipulation, there are two countervailing forces present at the household level. Manipulating households spend more on children, yet women in the household work more. There is an active literature focusing on the child consequences of us such "time versus money" trade-offs (Caucutt et al., 2020; Agostinelli and Sorrenti, 2021; Nicoletti et al., 2023; Mullins, 2022).

We combine our administrative health data and survey data to investigate changes in early childhood investment. These investments take the form of health and time investments. We find no changes in vaccination rates of children aged 0-5. Using our survey data, we do not find evidence of drops in the number of health check-ups. Whilst we have information on child-related time use of parents, the FDD estimates for these outcomes are imprecisely estimated, and we run into issues with the strength of our instrument.[6] We additionally focus on later

[5]We find no labor market impacts for men.

[6]In Section A.2.4, we document very clearly that these instrument strength issues are a consequence of the reduced sample size with which we have to work when using our survey data – the first stage estimate is remarkably constant across all our data settings.

child investments, in the form of high school and university attendance of older children and young adults in the household. We do not find any changes in post-compulsory high school attendance, neither we do find increases in university attendance for 18-23 years old living in the household.

In concluding the paper, we present two exercises. The first is to present a conservative lower bound estimate of the cost of welfare manipulation to the state coffers. Welfare eligibility manipulation is costly, amounting to an additional 25% of the initial welfare expenditure on our target group of welfare recipient households. The second exercise is to compare our estimates of welfare score manipulation – based on our FDD approach – with estimates from two generations of bunching estimators. This exercise serves as an external validity check of the approach we take in this paper. Our prevalence estimates are as good as identical to those from bunching estimators. It is worth noting that bunching estimates would have only allowed us to quantify the size of the problem and would have required further assumptions to explain the effects of manipulation. However, due to the richness of our data sources and our empirical strategy, we are able to break new ground in dissecting welfare manipulation attempts – describing in detail *which* households are more likely to manipulate, showing precisely *how* they manipulate, and documenting the downstream *consequences* of manipulation on key economic outcomes – without needing to impose strong assumption on households' behavior or relying on the estimation of counterfactual distribution of the PMT score.

Our work contributes to three strands of literature. First, we contribute to the literature on optimal public policy targeting, with a specific focus on demand-side manipulation. Our main contribution to this literature, given our unique data and our novel empirical approach, is that we can work more directly with welfare manipulation. Using the reasons for a repeat assessment, we can isolate the key margin on which manipulation occurs. Based on our characterization of compliers, we have a considerably better sense of observable profile of households susceptible to manipulation. Our data setting additionally allows us to observe how households attempt to manipulate their scores and to directly trace out the consequence of manipulation attempts. Our work ties in with the recent developments in the bunching literature, including work by Garbinti et al. (2023); Marx (2024) on dynamic bunching methods, and Diamond and Persson (2016) who combine bunching estimators and discontinuities in observable characteristics and outcomes to analyze the characteristics and effects of manipulation.[7,8]

A noteworthy aspect of the framework we employ in our study is that it applies to a broader

[7]Our approach differs from the modern bunching methods in several distinct ways. First, we do not rely on counterfactuals – we observe the initial, untainted welfare score. Second, we directly observe the action that leads to manipulation – the household initiation of a repeat interview. Accordingly, we estimate a local average treatment effect (LATE) as opposed to an intention to treat (ITT). This means we can characterize compliers, and extend our IV-based approach to consider marginal treatment effects. We note that our setting of two endogenous variables changing at a given threshold precludes our ability to use such methods – the FDD approach we implement allows us to just-identify the key parameters of interest.

[8]Additionally, Gelber et al. (2020) introduced a method to estimate bounds when there is manipulation of the score in RDD, which is used in work including that of Deshpande et al. (2021); Howell (2022); Britto et al. (2022). Miller et al. (2013) estimates the implicit non-manipulated score to estimate the effect of access to subsidized health. A similar approach is used by Best et al. (2020) when analyzing the effect of changes in the interest rate on the mortgage market in United Kingdom.

literature that focuses on effort or output manipulation in the face of discontinuities, notches, and kinks, such as those found in various contexts in public economics – any setting where there is both (i) the *incentive* to manipulate behavior, and (ii) the *scope* to do so. Such scenarios include tax schemes, unemployment insurance programs, and targeted subside schemes.

Second, we add to the academic and public debate on household responses to cash transfers, notably on whether such transfers have negative or positive effects on key outcomes. Targeted welfare may lead to welfare traps, whereby households are discouraged from labor market participation and making productive investments in order to keep receiving benefits.[9] Conversely, cash transfers can relieve households' liquidity constraints, thereby allowing them to search for better jobs or to invest in their children's education or productive ventures (for example Gertler et al., 2012; Carneiro et al., 2021). While we do not find negative effects from receiving the cash transfer itself, we do document a crowd-out effect of successful manipulation for female labor market outcomes (relative to unsuccessful manipulation attempts).

Finally, we add to the childhood skill investment literature investigating the return of different types of parental inputs at different stages of childhood (Cunha and Heckman, 2007; Caucutt et al., 2020; Agostinelli and Sorrenti, 2021; Nicoletti et al., 2023; Mullins, 2022). Making full use of both our administrative health and education data, as well as our rich survey data on time use and health investments, we are able to document the impact on children and young people in the household of a setting where adults within the household have more available income, but less available time. Despite not finding statistically significant changes in children and youth outcomes, this is valuable, as much of the work that considers the competing roles of parental time versus income investments do so within the context of developed economies.

2 Institutional Framework and Data

2.1 Georgia's TSA Program

In 2008, Georgia faced a deep crisis due to both the effects of the international financial crisis and the conflict with the Russian Federation in Ossetia. In response to the social consequences of this crisis, the government initiated the Targeted Social Assistance program – an unconditional cash transfer – as part of the social safety net (World Bank, 2018). The objective of the program was to alleviate poverty by direct cash transfers to households for a country where over one third of the population lay below the poverty line. The management of the TSA program is in the hands of the SSA.

In 2015, the SSA introduced major changes to the TSA program. First, the Agency commenced targeting using a PMT score.[10] To do so, the Agency interviewed all households registered in the United Database for Socially Unprotected Families (UDSUF).[11] The PMT measures households welfare using data on income, consumption, expenditure, assets, and household com-

[9] Although theoretically possible (Banerjee et al., 2017), there is no empirical evidence of such a response.

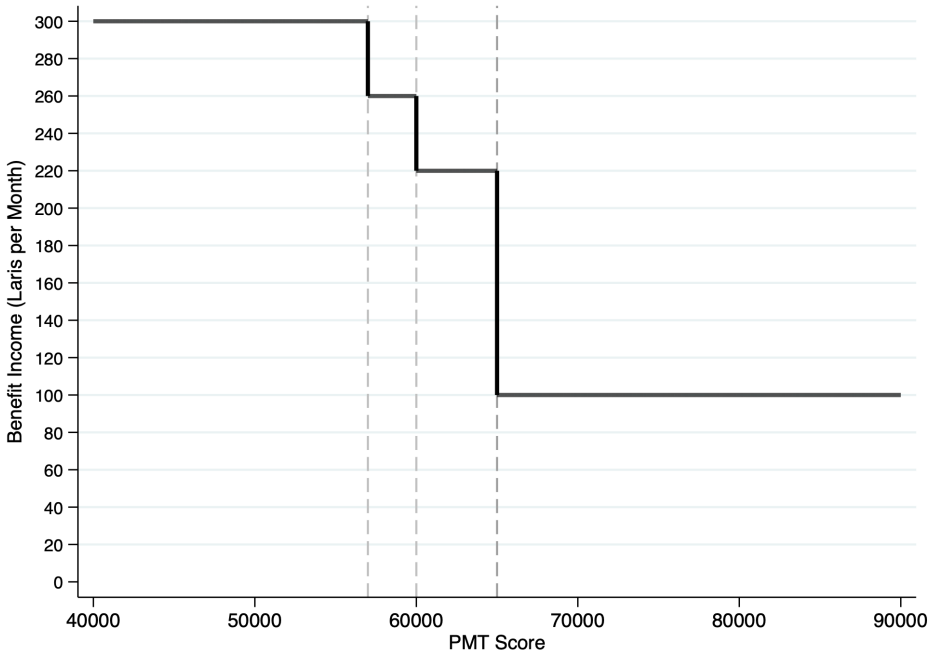
[10] The PMT formula was approved by the Resolution No. 758 (December 31, 2014) of the Government of Georgia

[11] From 2008, every household who wished to apply for receive social benefits was registered in this database.

position.

Second, the TSA allocated benefits as a step-wise decreasing function the PMT score. Third, by recommendation of UNICEF, the TSA introduced an additional benefit per child. Initially the benefit was 10 Lari per child month, but from January 2019 this increased to 50 Lari per child per month. After a household is assessed by the SSA and receives a score, they receive a monthly cash transfer (deposited in a bank account) based on their household composition and PMT score. After a household assessment, it takes just over a month on average (the mean is 34 days) to release the final PMT score. Hence, the score is not known at interview either by the welfare inspector or the household. Figure 1 provides the TSA benefit schedule for the sample median household composition. Table A1 in Appendix A provides a full summary of the TSA benefit scheme. It is clear from Figure 1 why the 65,000 is a focal point for welfare manipulation attempts – the reduction in benefits is considerably larger than the reduction in benefits in previous cutoffs.

Figure 1: The TSA Benefit Schedule



Notes: Benefit Income-PMT score schedule for the sample median household structure of two adults, two children. See Table A1 in Appendix A for the full schedule.

Furthermore, households may be reassessed for various reasons. For example, changes in household composition, changes in income (observed by the SSA), or changes in household's location.[12] In addition, if a household feels that their PMT score does not accurately represent their welfare, they can request an additional interview after one year of being assessed. In each case, the SSA will re-interview the household and calculate a new score, which may be larger or smaller than the original score, and will adjust the benefits accordingly.

[12]The SSA has access to data from different governmental sources in order to follow the TSA beneficiaries. For example, births, deaths, children dropping out school, increases in formal labor market income, disability claims.

As we note above, the PMT score is highly complex, and is unlikely to be perfectly known by households aiming to improve their score. The score comprises ten separate indexes. Each index comprises multiple inputs. Each index is calculated first by multiplying a specific input by a location-specific factor loading, then summing these input scores and finally by applying a non-linear transformation. The final PMT score sums the indexes, applies the exponential function, then divides by another index related to equivalized household size.

2.2 Data

We combine multiple sources of administrative data with information from a household survey that we conducted on a block-random subset of households.[13] Our core data is the universe of all PMT interviews conducted by the SSA from April 2015 – the start of the new TSA regime – to June 2019. This allows us to track every interaction a household has with the SSA and the benefits they receive since 2015. The PMT interview data contains new entrants in the welfare system, as well as existing welfare recipients, who were interviewed in order to calculate their PMT score.

For households with multiple interviews, we also observe the reason why an additional reassessment occurred. Using this information, we are able to observe if a reassessment was initiated by the household in the form of a request for a repeat interview, or was automatically triggered due to a change in the demographic or economic situation of the household.

Once a household receives a new PMT score, the previous score is annulled by the SSA. The SSA may also cancel the welfare payments to a household if (i.) the Agency finds out that the household hid changes that could alter their PMT score or cheated in any other way, or (ii.) if the household refuses a reassessment. In these cases the PMT score allocated to the household is annulled. We observe the PMT score status for each household-interview couplet.

We match households in the PMT interview database to three other administrative data sets. First, for every adult aged 18 to 64 years old we match in labor income and labor market participation information from the Revenue Service database from the Ministry of Finance. This covers only the formal sector. Using this data we can observe the extensive margin of formal labor supply, and the associated income with this job. We observe this information at four points in time – August 2018, February and August 2019, and February 2020. In our analysis, we only use formal labor market tranches of the data that fall after the last observed interview of the household.

Second, we use administrative data from the Ministry of Education on school attendance for children aged 5 to 18 years old. We observe in which grade they enroll in September 2017 to September 2019. Primary and secondary education in Georgia is free and compulsory (grades 1 to 10). At 16, teens are expected to enroll in high-school for grades 11 and 12, where school is still free but no longer compulsory. In addition, we have information on college attendance and college graduation for individuals aged 16 to 25 years old still living within the household.

Third, we use information from the Ministry of Health regarding vaccinations for children within the household.

[13] We randomly surveyed households at specific parts of the PMT score distribution – these are the "blocks".

We supplement the wealth of administrative data with a household survey conducted in the Fall of 2019, which surveyed a random sample of 7,392 households with children in 41 primary municipalities. The survey includes information about income, expenditure, labor market participation (in both the formal and informal sectors), education, health and childcare.[14]

2.3 Sample Selection

We focus on households who have children when initially assessed by the SSA as a matter of internal consistency – the household survey we conducted only interviewed households with children. The analysis we present in the paper focuses on the structure of the welfare payment system in Georgia, specifically the discrete cutoffs in the welfare payment-PMT score schedule, which creates incentives for PMT score manipulation. We focus our analysis on the 65,000 cutoff because as shown in Figure 1, the changes in benefits is larger there. This, along with the fact that households can apply for reassessments result, as we will show shortly, 65,000 is the only cutoff where we find clear evidence of PMT score manipulation. For this reason, we use households with an initial PMT score in the range 60,000 - 70,000. In addition, in order to have a groups of households where benefits change but not the probability of manipulation, we add to the sample those whose initial PMT score is around the 57,000 cutoff (from 54,000 to 60,000). These households are key to disentangle the effect of score manipulation as we will explain in the following section.

Furthermore, additional interviews play a key role in our analysis. For this reason we exclude households whose first interview was after December 31 2017, to allow that all the households in our sample have the opportunity to request a second interview, within the time frame for which we have all necessary data. Given that we have detailed information on the *reason* for a repeat interview, we omit from our core sample households with repeat assessments triggered solely by the SSA. The purpose of this sample restriction is to avoid conflating manipulation with a random demographic or labor market shock. Finally, we exclude households receiving Internal Displaced People (IDP) benefits the first time they were interviewed as these households receive a different set of benefits from the SSA.

Our final administrative data sample, once we apply all relevant sample selection restrictions, contains 11,972 households. Our final survey data sample contains 1,682 households.[15] Table 1 summarizes the main characteristics of our sample of analysis.

In our sample, 18% of households request an additional interview. Over the course of time that we observe these households, they average 2.8 interviews. Many households are multigenerational, with an average of 5 household members – 3 adults and 2 children. Most children attend school, only 7.3% of households have at least one child not attending. The households in our working sample are poor – baseline income for these households is 288 Lari (about 100 USD). 62% of households in our working sample own some form of estate (for example, garage, additional housing), 61% have agricultural land, and 41% have some livestock.

[14] More details of the questionnaire and sample selection in Econometría (2020)

[15] The fact that the survey data is an order of magnitude smaller than the administrative data leads to sample size-based power issues in some of our later analysis. Throughout our empirical work we balance the size and accuracy of the administrative data, with the richness of the survey data.

Table 1: Summary Statistics

	Mean	Standard Deviation
Household-Initiated Repeat Interview	.181	.385
Number of interviews[1]	2.85	.99
Household Composition		
Household size	4.54	1.52
Adults	2.7	1.2
Children	1.84	.852
Child Not in School	.073	.26
Pensioner in the Household	.0352	.184
Household Head Characteristics		
Age	50.4	15.2
Female	.391	.488
Single Mother	.0333	.179
Income and Expenditure		
Total income (Lari per Month)	288	274
Utility Bills (Lari per Month)	18.9	15.1
Housing Characteristics		
Number of Rooms	3.24	1.52
Good Quality Floor	.774	.418
Assets		
Owns any Estate	.618	.486
Owns a Car or Tractor	.0399	.196
Owns Agricultural Land	.612	.487
Owns any Livestock	.408	.491
Observations	11,972	

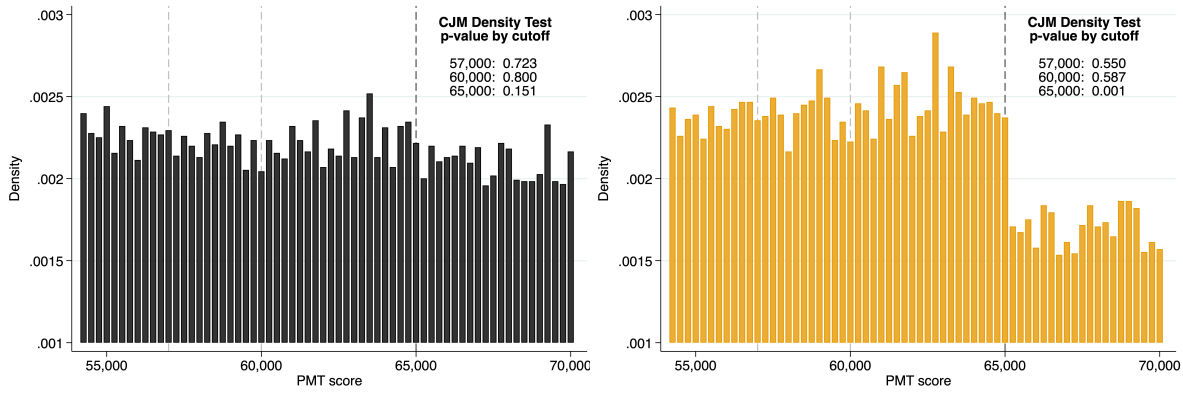
Notes: [1] Conditional on requesting at least one additional interview. Household characteristics as measured at the time of the initial interview. Data source: PMT Interview Data.

3 Evidence of Manipulation

We start by providing initial evidence of welfare manipulation, in order to motivate both the theoretical model and the empirical specification that follows. To do so, we consider a wider range of PMT scores than used for our main analysis – specifically 40,000-90,000. We first present the distribution of initial PMT scores in Figure 2(a). A visual inspection suggests that the distribution is smooth and continuous through the 57, 60 and 65 thousand cutoffs. This is confirmed by the associated p-values from a CJM density test (Cattaneo et al., 2020). In Figure 2(b) we present the analogous figure for the final PMT score distribution of households, allowing for reassessments. The difference between the two distributions is stark. There is clear visual evidence of unnatural bunching of households to the left of the 65,000 threshold, and a large discontinuity precisely at 65,000.[16] The CJM density test confirms the presence of manipulation, with a p-value of 0.001. In contrast, for the 57,000 and 60,000 cutoffs the distribution remains continuous.

[16]With perfect control over the welfare score, all households with initial scores above 65,000 that engage in manipulation would likely bunch just below 65,000. This is not the case in our setting, where households have only limited and imperfect control over their realized future welfare score.

Figure 2: The PMT Distribution at Initial and Final Interview



(a) First Interview (b) Final Interview

Notes: Bin size of 500. Panel (a) shows the distribution of the PMT score for the first interview each household had. Panel (b) shows the distribution of the last PMT score each household received. The box in both figures contains CJM Density Test p-value from the Cattaneo et al. (2020) manipulation test using households with scores between the cutoff above and below each cutoff in the estimation, a polynomial of order 2, and data driven bandwidths, around each cutoff. Sample: all households, excluding households with initial interview after December 31, 2017, and households receiving IDP benefits.

Households may be assessed multiple times for a variety of reasons, and these reassessments may be initiated by both households and the SSA. In Figure 3 we present evidence that is highly consistent with household-initiated requests for PMT score reassessment being the key driver of the discontinuity we document in Figure 2(b). We start by presenting Figure 3(a), which shows the unconditional probability that a household will have multiple interviews in the period of analysis. The probability jumps by approximately 20 percentage points, or just under 60%, precisely at 65,000. We then make use of the rich administrative data we have available, and separate between reasons for a reassessment. We respectively plot the probability of household-initiated and SSA-initiated reassessments in Figure 3(b) and Figure 3(c). The discontinuity at 65,000 is driven solely by household-initiated reassessment requests.

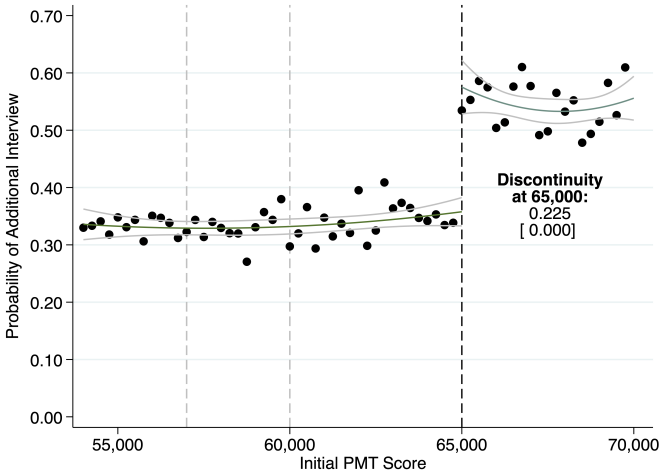
Following this evidence, in the following section we present a model in which we explain manipulation as the result of households optimally choosing whether or not to request an additional interview.

4 Modelling Welfare Eligibility Manipulation

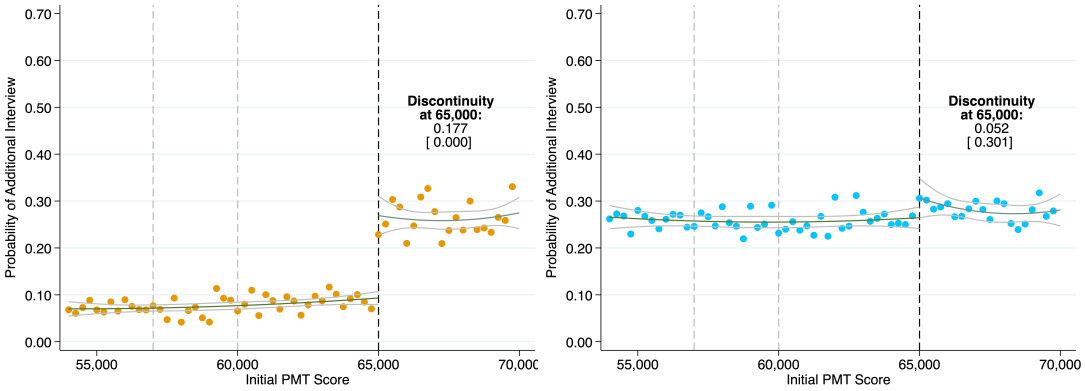
4.1 A Becker Model of Manipulation

We model welfare eligibility manipulation – here the decision to request a repeat interview – through the lens of the Becker-Ehrlich model (Becker, 1968; Ehrlich, 1973). According to this approach, an individual will choose to engage in welfare eligibility manipulation if the expected value of manipulation (V_R) exceeds that of accepting their initial benefit level (V_A). When requesting a repeat interview, the individual may receive a higher benefit level B^+ with exogenously determined probability p, or may receive the same benefit level as their initial allocation B^0. The cost of requesting a repeat interview is C. This cost captures the administrative and time cost of requesting a repeat interview, as well as the time cost involved in the repeat inter-

Figure 3: The Probability of an Additional Interview



(a) All Cases



(b) Household-Initiated Reassessment (c) SSA-Initiated Reassessment

Notes: Each figure shows the probability of having an additional interview by the first PMT score each household obtained. Panel b plots the probability that at least one additional interview was asked by the household. Panel c plot the probability that all PMT reassessments were initiated by the SSA. We include in each figure the resulting RD estimate and p-value in brackets, following Calonico et al. (2014).

view itself. With probability q the SSA discovers that the individual is falsifying information and imposes a sanction – a suspension of welfare payment for at least one year. B^- is the expected value of potential sanctions including any additional costs the individual may face, for example, loss of social capital due to engaging in welfare fraud (Williams and Sickles, 2002), or debt-related issues such as high interest payments, if households expect they may fall behind on bills or other payments if sanctioned.

Combining these factors, we can write an expression for the expected utility of requesting a reassessment – $E(V_R) = p(1-q)U(B^+) + (1-p)(1-q)U(B^0) + qU(B^-) - C$. The expected value of accepting the initial PMT score is $E(V_A) = U(B^0)$ Equating $E(V_R)$ and $E(V_A)$ allows us to characterize the point at which an individual is indifferent between requesting a repeat interview and accepting their initial PMT score. By defining the possible utility gain of requesting an additional interview as $\Delta U^+ = U(B^+) - U(B^0)$, the possible utility loss of an additional interview as $\Delta U^- = U(B^0) - U(B^-)$, and rearranging yields, we know that a given household will request a reassessment if:

$$p(1-q)\Delta U^{+} - q\Delta U^{-} > C. \tag{1}$$

Equation 1 has the following implications. First, both ΔU^{+} and ΔU^{-}, depend on B_0, which in turn depends on the initial PMT score. Hence, households incentives to manipulate vary along the PMT distribution. Second, and in contrast to the definition of manipulation in the existing literature, households have only imperfect control over the result of their manipulation attempt – (i) they likely do not fully comprehend the formula that maps inputs into the final score and (ii) the success of their manipulation attempts is subject to two external, exogenous shocks (p and q). Finally, households may differ in C – for the same level of B_0, some households may decide to manipulate and others not. These three features are the key to linking our theoretical model with our data and empirical approach.

4.2 Requesting a Reassessment and Empirical Specification

We now map our theoretical model onto a specification that we will estimate with our data. For a given household i, ΔU^{+} and ΔU^{-} are functions of the household's initial PMT score, $z_{0,i}$, and some limited household characteristics H_i, which determine B_0. Thus, the left hand side of Equation (1) can be written as:

$$p(1-q)\Delta U^{+} - q\Delta U^{-} = f(H_i, z_{0,i}) \tag{2}$$

C_i is a function of a broader set of observable variables, X_i, which encompasses H_i, and an unobservable component μ_i, which captures household-level tendency towards welfare eligibility manipulation:

$$C_i = k(X_i) - \mu_i \tag{3}$$

We denote $R_i = 1$ when a household requests a repeat interview. Assuming $f()$ and $k()$ are linear in X_i, and given that B_0, B^{+} and B^{-} depend on $z_{0,i}$ and a cutoff k, from equation 1, we write down a latent variable model for requesting a repeat interview:

$$R_i^{*} = X_i^{'}\beta + g(z_{0,i}, k) + \mu_i \tag{4a}$$

$$R_i = 1 \quad \text{if } R_i^{*} > 0 \tag{4b}$$

$$R_i = 0 \quad \text{if } R_i^{*} \leq 0 \tag{4c}$$

Having defined the details of a manipulation attempt, the obvious next step is to estimate the effect of R_i on an outcome variable Y_i. Using the initial PMT score as a running variable and the discontinuity in R_i at 65,000 could be used to estimate the effect of manipulation attempts on Y_i in a Regression Discontinuity Design (RDD). There is, however, an issue that precludes the straightforward estimation of this effect. Namely, at 65,000 there are discontinuities in *two* variables that can affect final outcomes: the probability of requesting an additional interview (R_i) and the value of the initial household benefit (B_0). For this reason we use variation around other cutoff, using a difference-in-discontinuities design (Grembi et al., 2016), in order to

disentangle the effect of manipulation attempts from the effect of initial benefits. The approach we take in this work, which we outline below, involves an instrumental variables approach to a difference-in-discontinuities design, which we refer to as a fuzzy difference-in-discontinuities (FDD) design (Millán-Quijano, 2020).

Empirical Specification

In order to operationalize our FDD approach, we include in our analysis sample households located around a second cutoff of 57,000. Around this lower cutoff we do not find evidence of manipulation (see Figure 2(b)), nor do we document any changes in the probability of requesting an additional interview (Figure 3(b)). The only change of note at this cutoff is that initial benefits drop by 10 Lari per person per month as shown in Figure 1. Our FDD approach means we use variation in Y around the auxiliary threshold to identify the impact of B_0 on the outcome of interest, and then use additional variation in Y_i around the focal threshold (65,000) to identify the impact of a manipulation attempt on the outcome.

Following the difference-in-discontinuities literature, we define a dummy A_i that takes the value of one (1) for households whose first PMT is in the neighborhood of 65,000 ($A_i = \mathbb{1}[60,000 < z_{0,i} \leq 70,000]$). We also define D_i to takes the value of 1 for households above their respective cutoff. Thus, $D_i = \mathbb{1}[z_{i,0} > 57,000 \ \& \ A_i = 0]$ or $D_i = \mathbb{1}[z_{i,0} > 65,000 \ \& \ A_i = 1]$. Then, we can write the outcome Y as function of B_0 and R as:

$$R_i = \omega_1 D_i + \omega_2 A_i + \omega_3 A_i \times D_i + g_R^{D,A}(z_{0,i}) + X_i^{'}\omega + \mu_{R,i} \tag{5a}$$

$$B_{0,i} = \gamma_1 D_i + \gamma_2 A_i + \gamma_3 A_i \times D_i + g_B^{D,A}(z_{0,i}) + X_i^{'}\gamma + \mu_{B,i} \tag{5b}$$

$$Y_i = \theta_R R_i + \theta_B B_{0,i} + \theta_3 A_i + g_Y^{D,A}(z_{0,i}) + X_i^{'}\theta + \mu_{Y,i} \tag{5c}$$

The latest system of equations can be also understood as a two-stage least squares (2SLS) system with two endogenous variables (R, B_0), and two instruments $(D, A \times D)$. In order to identify the causal effect of manipulation attempts R_i on outcomes, we require several identifying assumptions to be met – two assumptions common in RD-based designs, and another two use in difference in discontinuity designs.[17] First, we consider the core RDD assumptions around both cutoffs. Figure 2(a) shows that $z_{0,i}$ is continuous through the cutoffs at 57,000 and 65,000, which means that D_i is as good as randomly assigned. In addition, Figure A1 in Appendix A.2 shows that the RDD continuity assumption holds for a large set of observable variables X, using information from the households' initial PMT interviews.

Next, following Grembi et al. (2016) and Millán-Quijano (2020), we outline two further, FDD-based identifying assumptions. We provide detailed evidence to suggest that these assumptions do hold in our setting, thereby enabling us to proceed with our FDD design. First, we require that the effect of B_0 is constant across the two PMT areas: $A = 0$ and $A = 1$. We use two cutoffs close to each other, thus, after controlling by A, assuming that $E(\theta_B|A = 0) = E(\theta_B|A = 1)$ is plausible, as θ_B represents the effect of one additional Lari. Figure A2 in

[17] More details on this in Appendix B

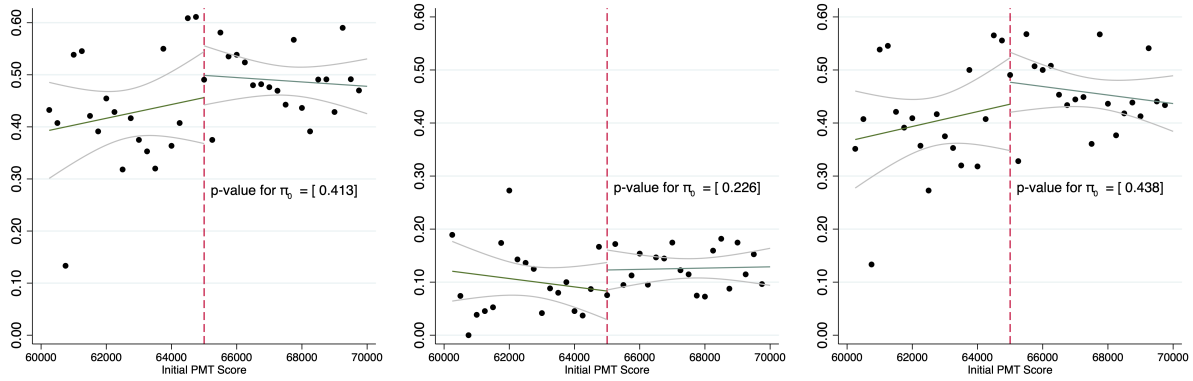
Appendix A.2.2 shows that the impact of an additional Lari is not statistically different for a set of labor market outcome around the three cutoffs where we do not find evidence of manipulation (around 30,000, 57,000 and 60,000). One potential concern is that the change in benefits around 65,000 is three times the change in benefits at 57,000. However, the scale of the jump in benefits is taken into account by γ_1 and γ_3 in equation 5b. Figure A3 in Appendix A.2.3 shows that $\gamma_1 + \gamma_3 = 3\gamma_1$ regardless of the estimation sample.

Second, in order to isolate the effect of R, the change in manipulation attempts only happens around one of the cutoffs, in our case around 65,000. We have already provided evidence in support of this assumption (that manipulation only occurs at 65,000): the distribution of the final PMT is continuous around 57,000 (Figure 2(b)), and manipulation attempts are also continuous around 57,000 (Figure 3(b)).

Potential Threats to Identification

We may still be concerned about the role household unobservables play in the reassessment process. We address this concern directly by providing evidence that there are no discontinuities at the thresholds for the two key dimensions governing the success of a manipulation attempt – the probability that a repeat interview will reduce a household score, and the probability of being caught and sanctioned by the SSA. These are, respectively, the parameters p and q from the theoretical model we present in Section 4.1. Using detailed data about the final status of each household, we can plot these two probabilities against our running variable. We do so in Figure 4, which shows that both p and q are continuous through the cutoff. The p-values that we present in the graphs are based on the null that there is no discontinuity at the threshold. Our statistical tests confirm what a visual inspection of the figures tells us – p and q are continuous through the cutoff.

Figure 4: The Continuity of Key Model Parameters Through the Focal Threshold



(a) p (b) q (c) $p \times (1-q)$

Notes: Panel (a) plots the probability that a reassessment leads to a reduction in the PMT score. Panel (b) plots the probability that the household's last PMT score is recorded as invalid by the SSA. Both figures only take into account households who requested at least one additional interview. Panel (c) represents the probability of a successful manipulation attempt. In each graph we present the respective p-value for the parameter π_0 from a regression of the form $y_i = \pi_0 D_i + g_1^D(z_{0,i})$, where $D_i = \mathbb{1}[z_{0,i} > 65,000]$ and $g_1^D(z_{0,i})$ is a polynomial of order 1 in $z_{0,i}$ above and below the cutoff.

Such null results as those we present in Figure 4 allay concerns about the continuity of

household unobservables around the focal threshold – an ability to game the welfare system (p) or the ability to deceive, or collude with, the welfare assessor (q). In addition, the evidence in Figure 4 additionally assuages concerns over differential welfare assessor treatment of households just above and just below the focal threshold. Were this differential treatment to occur, we would capture this in jumps in either p or q. Finally, the absence of any perturbations to the initial distribution of welfare scores (Figure 2(a) in Section 3) further cements the evidence in favor of no collusion between welfare assessors and households. If there was such collusion, we would expect to see pronounced discontinuities in this distribution as well. We suspect the complexity of the welfare score formula, coupled with the long delay between household inspection and score release, is at least part of the reason we do not find such collusion.

5 Results

In Section 3 we provide evidence that given (i) the structure of the benefit scheme we study and (ii) the availability of household-initiated reassessment creates the incentives to manipulate welfare eligibility. In this section we seek to answer three related questions regarding welfare eligibility manipulation. First, *what type* of households attempt to manipulate their welfare eligibility status? Second, *how* do these households manipulate their score? Third, what are the downstream consequences of welfare eligibility manipulation?

5.1 Who are the Compliers?

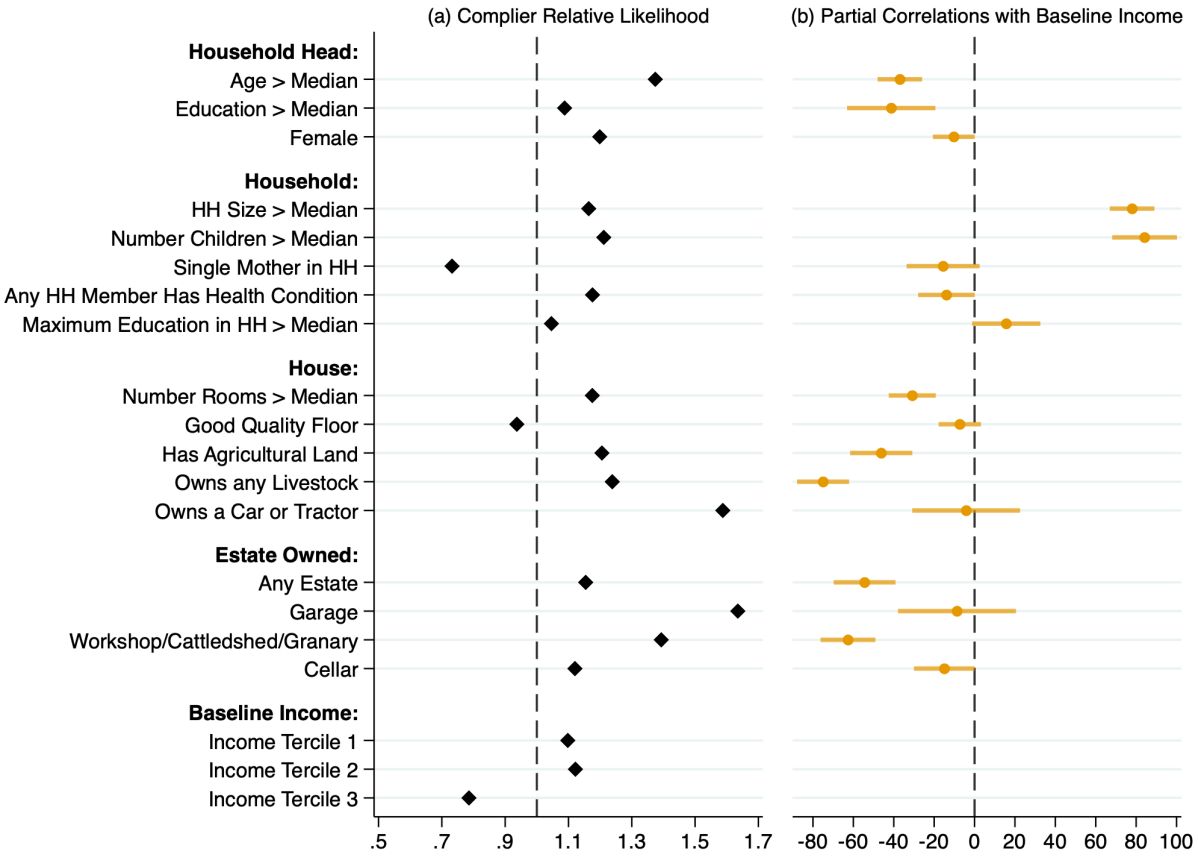
In this section we focus on the decision to manipulate welfare eligibility, which leads us to zoom in on the first-stage equation for R_i, Equation 5(a). Our FDD design is founded on an instrumental variables framework, which we exploit here to consider the compliers on the manipulation margin – households whose manipulation status is induced by falling above the 65,000 PMT score threshold. Whilst we cannot directly identify the compliant sub-population, we can characterize these households.

To do so, we follow the approaches of Abadie (2003) and Dahl et al. (2014) in characterizing compliers in an IV-based framework. Our target statistic is the complier relative likelihood of having a given Bernoulli-distributed characteristic, x_{1i}, which we express as $P[x_{1i} = 1 \mid R_{1i} > R_{0i}] / P[x_{1i} = 1]$. For continuous characteristics, we binarize the variable.[18] R_{1i} and R_{0i} denote the potential outcomes of R_i when $A_i \times D_i = 1$ and $A_i \times D_i = 0$ respectively. We present a series of complier relative likelihoods in Figure 5(a).

Complier households are more likely to be headed by a woman, and have an older, slightly more educated head of household than average. These households are larger in terms of total size and number of children, are more likely to have a household member with a health condition and are less likely to have a single mother present. Compliers appear to live in more rural settings, as they are more likely to own a workshop, granary or cattleshed, to have agricultural land, to own both livestock, and a car or tractor. They are also more likely to have a garage. Finally, we document that complier households are poorer than the average household in the

[18] We calculate the relative likelihood using Bayes' Rule and by taking the ratio of the first-stage coefficient for the sub-group with $x_{1i} = 1$ divided by the first stage coefficient for the full sample, $P[R_{1i} > R_{0i} \mid x_{1i} = 1] / P[R_{1i} > R_{0i}]$.

Figure 5: Characterizing Complier Households



Notes: Panel (a) – We characterize compliers by presenting the ratio of the first stage coefficient on the instrument for each binary (or binarized) characteristic to the overall first stage coefficient. By Bayes' rule this ratio of first stage estimates – which we can express as $P[R_{1i} > R_{0i} \mid x_{1i} = 1]/P[R_{1i} > R_{0i}]$ – yields the complier relative likelihood of a given characteristic, $P[x_{1i} = 1 \mid R_{1i} > R_{0i}]/P[x_{1i} = 1]$. PMT range: 60,000-70,000. Panel (b) – We report the coefficient and 95% confidence interval for each characteristic from a regression where the dependent variable is household (own) income at baseline. We condition on a common set of household-level control variables (our baseline covariates), region-by-quarter and interview time fixed effects. The estimation sample is based on a PMT range of 55,000-60,000 and 70,000-75,000, i.e., bands of 5,000 on either side of our range of interest.

PMT score range of 60,000-70,000 – they are more likely to fall in the lower two terciles, and much less likely to be in the upper tercile of baseline income.

In Figure 5(b) we present estimates for each characteristic from a regression where the dependent variable is household earned income at baseline.[19] Combining the information in Figure 5(a) and 5(b), we can better understand the baseline economic status of the compliant households – 5(b) informs us of the partial correlation between a given characteristic and baseline income, whilst Figure 5(a) informs us of the relative likelihood a complier household will have the characteristic. With the exception that complier households are typically larger than average, all other characteristics of these households are correlated with lower economic status at baseline, particularly those related to the more rural setting in which complier households appear to be based.

[19]We condition on a common set of household-level control variables (our baseline covariates, described below), region-by-quarter and interview time fixed effects.

5.2 How do Households Manipulate Their PMT Scores?

We next aim to understand how households manipulate their PMT scores, in order to achieve a lower score and thus receive higher benefit income. To do so, we exploit the richness of our administrative data, which includes each and every input into the combined score that yields a PMT score. Hence, for households with multiple interviews we observe which PMT inputs changed between the first and last interview. We present changes for each of the relevant consumption-based index scores, as well as the needs index – the denominator in the calculation of the welfare score.

We focus in households with initial PMT scores between 65,000 and 70,000. We calculate the change in each of the PMT score input components between the last and the first interview each household has. We use the SSA-initiated interviews to establish a baseline for changes that are natural between interviews, or that are likely to come from random shocks.[20] The aim of this exercise is to identify which variables are more likely to change between interviews for households with household-initiated compared to SSA-initiated re-interviews.[21]

Figure 6 shows the results of this analysis. Panel (a) the average change in each component for both household- and SSA-initiated reassessments, i.e., the group-based differences. One can see that many components change similarly for both types of reassessments, while for other components the average change after a new interview is different when the source of the request for re-interview comes from the household, not the SSA. In Panel (b) we present the difference across these two groups. Clear patterns emerge, which provide suggestive evidence of how households that engage in welfare eligibility manipulation do so in practice.

First, the index associated with movable agricultural property – labelled cattle, but including cattle, beehives, and poultry – decreases significantly, both statistically and economically so. We also find a large between-group decrease in the index related to movable property (vehicles or trucks, tractors or seed-planting machines). When combined with the evidence we provide in Section 5.1, which highlights that complier households are more likely to be based in rural areas, the evidence here suggests that selling, or misreporting, rural assets is likely a key strategy for welfare score manipulation. This stands in contrasts to harder to move/misreport assets, such as land, properties, or the quality of the properties, for which we do not find changes.[22]

5.3 The Labor Market and Welfare Eligibility Manipulation

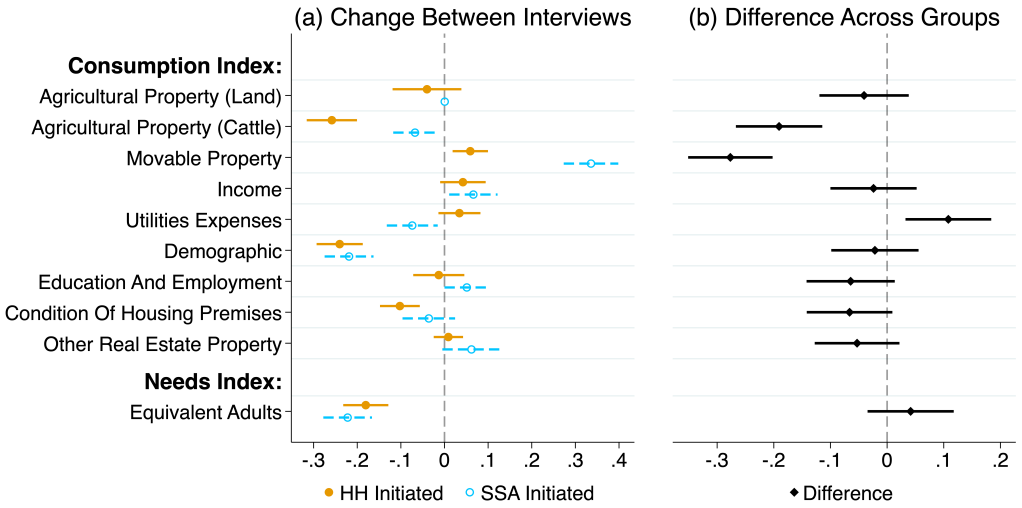
In order to understand the labor market consequences of welfare eligibility manipulation, we make use of two data sources. First, administrative data on formal labor market activity and earnings from four periods of time. We supplement this administrative data with survey data

[20] We cannot utilize the machinery of other approaches in a wider neighborhood around the 65,000 threshold, for example Diamond and Persson (2016). This is because household benefits, B_0, also change at the focal threshold. This complication is what motivates the FDD design that we employ in this study.

[21] We cannot present this evidence using an FDD design, our primary estimation strategy, as we only observed between-interview changes in welfare score inputs for households with multiple interviews.

[22] Note that we find between-groups changes statistically significantly different from zero in several other cases that are generated by SSA-initiated repeat interviews, which we would expect – if households situations change significantly, this automatically triggers a new (SSA-initiated) interview. Examples of this include the purchase of a new car, or a change in household utility payments.

Figure 6: Changes in Index Scores Between Interviews



Notes: We center and studentize all index scores to create a common scale for graphical presentation. Panel (a) – We plot the average difference in the value of each variable from the last to the first interview for households with more than one interview and whose first interview PMT score is between 65,000 and 70,000. The line represents the 95% confidence interval. Panel (b) – We plot the average difference between household-initiated and SSA-initiated in the change between the last and the first interview. The line represents the 95% confidence interval.

that contains information on both formal and informal labor market activity. The combined use of both data sources permits us to capture a broad and comprehensive view of the labor market consequences of welfare eligibility manipulation. Using data on the timing of both household interviews – specifically the date of the final interview – and labor market outcomes, we restrict our attention to post-final interview outcomes only.[23]

We first consider formal labor market outcomes. In Table 2 we provide evidence of the labor market consequences of attempted welfare eligibility manipulation. Although the welfare eligibility manipulation occurs at the household level, the heterogeneity in formal labor market responses by gender means it is instructive to consider the results by gender. As we note above, we only consider labor market outcomes that occur after the households' final interview, thus we interpret the results in this section as the downstream consequences of welfare eligibility manipulation.

We start with men. The OLS results indicate a negative relationship between a welfare eligibility manipulation attempt and labor market outcomes. This suggests that the act of manipulation could be driven by need – those who attempt to manipulate their eligibility have lower formal labor market income and are less likely to be employed at least once – our least stringent measure of labor force attachment. This interpretation of the negative coefficient is consistent with what we document for household income in Figure 5(a). An alternative explanation for the negative OLS estimates is selection bias – those that attempt to manipulate welfare eligibility have unobservables negatively correlated with labor market outcomes. Comparing the OLS estimate of an eligibility manipulation attempt to the mean of the outcome variable for those with a first PMT score below the 65,000 cutoff ($\overline{Y}_0$ at the base

[23]We find no differences between interviews in labor market outcomes for household that attempt to manipulate their welfare scores – this can be clearly seen in Figure 6.

of the table), we see those attempting to manipulate their scores earn 24% less in the formal labor market and are 11% less likely to be minimally attached to the formal labor force.

Given that our 2SLS estimates reflect a local average treatment effect (LATE) based on the compliant sub-population, and deal with the problem of endogenous welfare eligibility manipulation attempts, we present complier-reweighted OLS estimates to bridge between our OLS and 2SLS estimates (Bhuller et al., 2020). The complier-reweighted OLS estimates are broadly in line with our main OLS estimates, which suggests that treatment effect heterogeneity is unlikely to be a primary concern when interpreting our 2SLS estimates.

Finally, we turn to the 2SLS estimates and document that the impact of a manipulation attempt are still negative but statistically insignificant. Hence, once we use the exogenous variation in the probability to manipulate, we conclude that manipulation attempts do not lead to changes in the probability to work in the formal sector, or in the income earned.

In columns 4-6 of Table 2 we present analogous results for women. Both the OLS and complier-reweighted OLS estimates suggest little correlation between a household welfare eligibility manipulation attempt and labor market outcomes for women. The 2SLS estimates paint a very different picture however. In columns 4 and 5, we document a positive effect of a manipulation attempt on the labor force participation of women. They are statistically significantly more likely to work at least once, and in all periods, following a household manipulation attempt. This increased labor market attachment leads to higher income, although this effect is not statistically significantly different from zero.[24]

In Figure A5 in Appendix A.3, we provide evidence that the labor market response of households appears strategic. If households who receive welfare benefits increase their monthly earned income by more than 175 Lari, this income change automatically triggers a further, SSA-initiated repeat interview. Figure A5 highlights that income changes stay within these limits for the vast majority (93%) of households.

5.3.1 Does success in the manipulation attempt matter?

So far, we have documented the effect of a welfare manipulation *attempt* – proxied by household-initiated repeat interview – on labor market outcomes. In Figure 4 we show that approximately half of household-initiated repeat interview lead to a PMT score reduction sufficient to increase benefit income ($p(1-q)$ from our conceptual model). This observation gives rise to the possibility of treatment effect heterogeneity, which relates to our labor market findings above. Do women in households that attempt to manipulate their welfare eligibility work more due to the removal of (non-labor) income-based constraint that previously prevent them from working? Such a response would rely on a credit-constraint channel of some form. Or is it the converse, i.e., manipulation attempts are needs driven, and a failed manipulation attempt leads women in the household to shift towards formal labor market participation?

[24] At the household level, the results by gender lead to an increase in formal labor market participation but a null effect on total labor market income (Table A2, Appendix A.3). For informal labor supply, we do not find any significant changes (Table A3, Appendix A.3).

Table 2: Welfare Eligibility Manipulation and Formal Labor Market Engagement

	(1)	(2)	(3)	(4)	(5)	(6)
	Men			**Women**		
	Employed At Least Once	Employed All Periods	Mean Income	Employed At Least Once	Employed All Periods	Mean Income
OLS						
Repeat Interview	–0.027**	0.009	–24.637***	–0.011	0.004	–2.920
	(0.012)	(0.007)	(7.264)	(0.010)	(0.006)	(3.348)
CW-OLS						
Repeat Interview	–0.010	0.012	–14.898*	–0.009	0.007	–1.344
	(0.013)	(0.008)	(7.774)	(0.010)	(0.006)	(3.578)
2SLS						
Repeat Interview	–0.112	0.202	–127.498	0.312*	0.257**	64.521
	(0.311)	(0.186)	(186.933)	(0.181)	(0.110)	(75.602)
SW F-Statistic: R.I.	20.435	20.435	20.435	44.620	44.620	44.620
$\overline{Y}_0$	0.251	0.070	102.779	0.179	0.053	44.251
Observations	11,220	11,220	11,220	14,544	14,544	14,544

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Labour market outcomes measured after the last household interview. Each column summarizes the results for the respective outcome variable following Equation 5(c) for the OLS and the system of equations 5(a), 5(b), and 5(c) for the 2SLS. All estimations control for the individual's age (using a quadratic function), for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects. The CW-OLS calculation follows Bhuller et al. (2020).

To make progress on answering this question, we first briefly recap the ordering of events. (1) All households receive an initial interview. (2) Some households request an additional interview. We find that those just above the PMT threshold of 65,000 do so at a much higher rate, consistent with welfare eligibility manipulation. (3) After a repeat interview is requested, the SSA visits the household to reassess their need and recalculates the household PMT score. The indicator p takes the value 1 if the household moves down a PMT score category, and therefore increases their benefit income, and takes value 0 otherwise. (4) At the repeat assessment, the interviewer completes a 9-point checklist of questions to indicate the reliability of the information supplied by the household – this is the indicator variable q, which takes value 1 if the household is adjudged to have provided unreliable information and 0 otherwise. A value of $q = 1$ leads to all SSA benefits being suspended for at least 12 months. Accordingly we classify a manipulation attempt to be a success if $p(1-q) = 1$ and unsuccessful otherwise. (5) Approximately one month after the interview the score is released. There is no difference in the time-lag of score creation between household- and SSA-initiated repeat interviews.

We split our sample into two, non-mutually exclusive groups: first we consider households with only a single interview plus households with successful attempts. Second we consider the same group of single-interview households plus households with unsuccessful attempts. We present two pieces of evidence to support the validity of this approach. In Figure 4(c) in Appendix A.2 shows that the probability of success is continuous through the cutoffs, with a mean of approximately 45%. In Figure A6 in Appendix A.4 we present evidence in support of the continuity assumption, showing that observable variables are continuous at the cutoffs for the two split-samples. We then present the 2SLS results for formal labor market outcomes for the two split-samples in Table 3.

Once again, we find no impact for men, irrespective of the success of the manipulation attempt. When we turn to women, we find that it is women in households with unsuccessful manipulation attempts who drive our core labor market results that we document in Table 2. Following an unsuccessful manipulation attempt, women supply more labor in the formal sector. There is no statistically significant change in labor market engagement for women in households with successful manipulation attempts. This finding highlights an additional costs that welfare manipulation imposes upon the government – successful manipulation attempts *crowd out* labor market participation, reducing tax revenue, and perpetuating a cycle of reliance on the welfare system. As a final point, even though our estimates are imprecise, it is worth noting that the income of women increases by 120 Lari per month, which is equivalent to the expected TSA benefit increase after a manipulation attempt for a household of four members.[25]

Table 3: Formal Labor Market Outcomes by Success Status of Manipulation Attempts

	(1)	(2)	(3)	(4)	(5)	(6)
	Men			**Women**		
	Employed At Least Once	Employed All Periods	Mean Income	Employed At Least Once	Employed All Periods	Mean Income
(a) Unsuccessful Manipulation Attempts						
2SLS						
Repeat Interview	–0.046	0.416	–38.626	0.468*	0.375**	119.573
	(0.547)	(0.342)	(328.086)	(0.255)	(0.156)	(106.478)
SW F-Statistic: R.I.	9.281	9.281	9.281	33.717	33.717	33.717
$\overline{Y}_0$	0.255	0.072	105.508	0.181	0.054	45.054
Observations	10,434	10,434	10,434	13,485	13,485	13,485
(b) Successful Manipulation Attempts						
2SLS						
Repeat Interview	–0.200	0.343	–134.008	0.316	0.296	8.906
	(0.441)	(0.264)	(264.137)	(0.358)	(0.210)	(147.866)
SW F-Statistic: R.I.	27.158	27.158	27.158	27.701	27.701	27.701
$\overline{Y}_0$	0.251	0.069	102.792	0.177	0.051	43.159
Observations	10,130	10,130	10,130	13,244	13,244	13,244

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Labour market outcomes measured after the last household interview. Successful manipulation is defined as (i) having a final PMT score in a category below the initial assigment (increase in benefits) and (ii) a valid score from the SSA. Unsuccessful manipulators are either those who did not get an increase in benefits after the reassessment or whose score was cancel by the SSA due to unreliable information. Each column summarizes the results for the respective outcome variable the system of equations 5(a), 5(b), and 5(c) for the 2SLS. All estimations control for the individual's age (using a quadratic function), for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects.

5.3.2 Marginal Treatment Effects

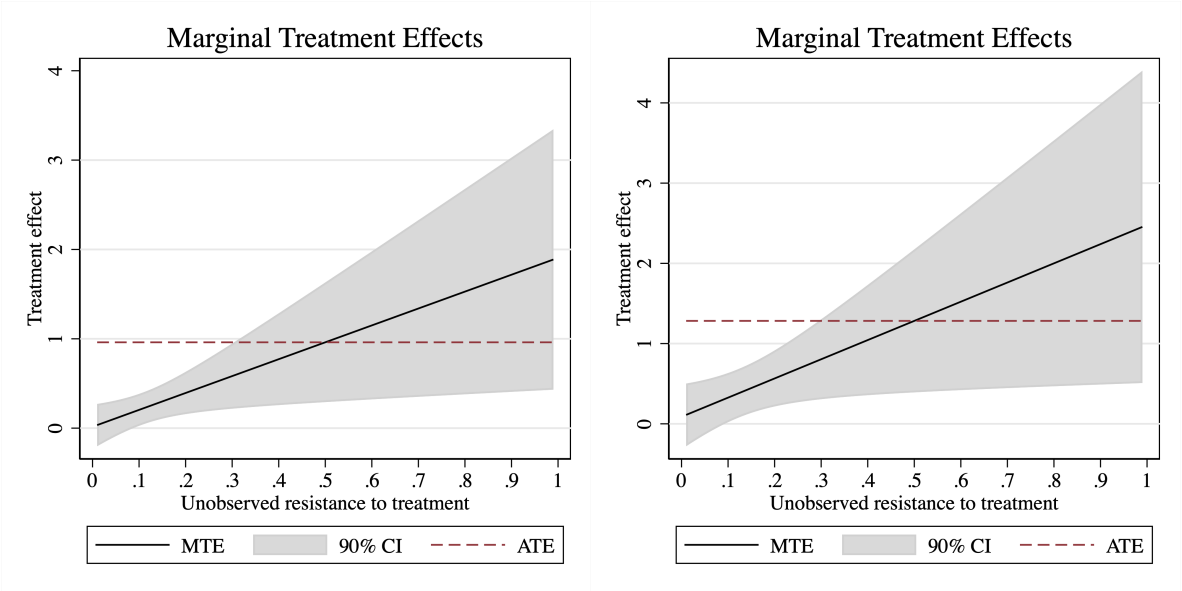
To conclude the analysis of the effects of manipulation on labor market outcomes we present the results from an MTE approach. Given we have a binary instrument, we follow Brinch

[25]The analysis at the household level shows a similar pattern. Formal labor market income increases by 135 Lari in households were the manipulation attempt was unsuccessful. However, this estimate is not statistically significant (See Table A4 in Appendix A.4).

et al. (2017) and specify a linear MTE using the separate MTE estimation approach. We provide details of this in Appendix A.4.1. Our primary impetus to conduct the MTE analysis is to better understand the extent to which households select into manipulating their welfare eligibility status due to their unobserved labor market gains.

We present the full set of MTE results in Table A6 in Appendix A.4.1, and highlight the key information we glean from this exercise in Figure 7. For women, for all three labor market outcomes, we document negative selection (into manipulation) on the unobservable (labor market) gains. For both employment outcomes, this selection on unobservable gains, inferred from the slope of the MTE curve, is statistically significantly different from 0 at the 10% level. Put differently, those most likely to select into manipulation have least to gain from doing so. Such a finding echoes that of Cornelissen et al. (2018).

Figure 7: Marginal Treatment Effect Evidence For Women



(a) Working All Periods (b) Working at Least One Period

Notes: MTE following Brinch et al. (2017) using a linear function. The gray area represents the 90% confidence interval.

This finding – of negative selection on unobserved gains – is a new insight gleaned from the MTE analysis, and enables us to form a clearer picture of what drives selection into a manipulation attempt. Our first observation is that, given that the unobserved gains from selection are negative, households do not ex-ante have the downstream labor market consequences in mind when engaging in manipulation. The second observation comes from combining the results of our MTE analysis with our investigation of the heterogeneous labor market effects of manipulation by success status. The combined results reinforce the notion that successful manipulation attempts exert a crowd-out effect on labor market engagement, at least for women – women only engage more in the labor market following an unsuccessful manipulation attempt. That the (heterogeneous) returns to labor market engagement are so low for this sub-group suggests that women in households with an unsuccessful manipulation attempt are not pulled into the labor market based on the high (unobserved) gains, but rather pushed into doing so based on

the need to bolster household income.

5.4 Linking our Findings to our Conceptual Framework

At this stage, we reflect on our findings and link these back to our conceptual model of household manipulation. In Equation 3, we specify the cost of manipulation (C_i) to have two components. The function $k(X_i)$ is the part of the cost that we observe and depends only on observable characteristics. According to both (i) our characterization of manipulators (Section 5.1) and (ii) the evidence we provide on how such household manipulate their welfare eligibility status (Section 5.2), we conclude that households are more likely to manipulate their welfare eligibility status by reducing rural assets (either by selling, hiding, or misreporting these assets). This points to a lower cost of manipulation in the rural context – hiding small livestock is easier than altering the characteristics of a city apartment. This is likely why we find manipulator households to be more likely based in rural settings.

The second component of our cost function is μ_i, which is unobservable to the econometrician. This unobservable term can be viewed as akin to the unobserved resistance to treatment variable, which plays a central role in the MTE analysis. That we document negative selection into manipulation on unobserved gains in the labor market suggests that manipulation behavior is likely needs-driven. This interpretation – of economic needs-driven manipulation – is bolstered by our finding in Section 5.1, where we document that compliers are relatively low income households within PMT score band. A second finding that supports this view is that we only find statistically significant increases in labor market engagement following an unsuccessful manipulation attempt – women in the household only increase their labor market engagement when a manipulation attempt does not yield additional welfare income.

5.5 Expenditure Responses to Welfare Manipulation Attempts

In Section 5.3 we provided evidence of the impact of welfare eligibility manipulation on labor market outcomes, documenting a positive effect of welfare eligibility manipulation on employment for women and no significant impacts for men. We provided further evidence highlighting the mediating role that the success of the manipulation attempt plays in driving these results. Ultimately, households that engage in welfare eligibility attempts are better off financially – successful manipulation attempts bring in more welfare income, and unsuccessful attempts end up yielding more labor income due to the labor market response of women to failed manipulation attempts.

Having taken stock of the evidence in the previous section, a natural question to ask is how do these welfare-manipulating households spend the extra income? In Table 4 we present expenditure patterns based on our survey data. The penultimate row of this table displays the outcome variable for those with an initial PMT score below the 65,000 cutoff ($\overline{Y}_0$). These statistics are particularly useful to gain a sense of expenditure patterns for a control set of households.

Both the OLS and complier-weighted OLS estimates are negative for almost every single expenditure group, and typically statistically indistinguishable from zero. The 2SLS estimates, however, tell a different story. We find that households that engage in welfare eligibility ma-

nipulation target their additional income on expenditure on children. We find no changes in expenditure on other areas, such as expenditure on food, eating out, tobacco or alcohol. We document a 88 Lari increase in total expenditure on children, the lion's share of which is on clothing, and a smaller share on increased education spending. Although the effect for total expenditure is statistically insignificant, we can see that the increase in spending on children is approximately 90% of the total expenditure response. A different way to benchmark the increase in child spending is to use the baseline total expenditure (Column 1, penultimate row), in which case the increase in child spending is 21% of baseline expenditure.

Summarizing the results, we find that a manipulation attempt leads to a significant increase in labor supply for women. In addition, in Table 4 we document that children are the primary beneficiaries of the corresponding increase in household spending. As our gaze turns now to child outcomes, we note that the evidence we document so far identifies two, countervailing forces on the child skill production function within households that attempt to manipulate their welfare eligibility. The increase in income, and concomitant expenditure on children, should have a positive impact on childhood skill production, whereas the fact that moms are now more likely to work and thus have have less time available will likely lead to a decrease in the production of childhood skills (Cunha and Heckman, 2007; Caucutt et al., 2020; Agostinelli and Sorrenti, 2021; Nicoletti et al., 2023; Mullins, 2022).

We investigate the impact of a household manipulation attempt on a battery of child outcomes for a variety of different ages, using administrative and survey data at the household and children level. For children 0 to 5 years old we consider the possible effects of manipulation on health investments such as vaccinations and check ups, and parental investments measured in time spent with their children. We additionally consider the impact of manipulation attempts on educational attendance for children 15 to 18, and for enrollment into tertiary education for young adults (18 to 23 years old). The details of these analyses are in Appendix A.5. We do not find any statistically significant changes in any children related outcomes, in many cases due to small sample size.

Table 4: Welfare Eligibility Manipulation Attempts and Household Expenditure

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
						Children			
	Total	Food	Food Outside of House	Alcohol Tobacco	Adult Clothing	Total	Clothing	Education	Childcare
OLS									
Repeat Interview	–40.523	–11.749	–0.647**	–3.983	–3.041***	–2.133	0.440	–1.848*	–0.725
	(33.057)	(9.162)	(0.293)	(2.689)	(0.811)	(2.423)	(1.878)	(1.068)	(0.494)
CW-OLS									
Repeat Interview	–40.229	–12.754	–0.716**	–3.906	–3.091***	–2.785	0.267	–2.267*	–0.784
	(35.897)	(9.584)	(0.301)	(2.776)	(0.839)	(2.538)	(1.925)	(1.176)	(0.539)
2SLS									
Repeat Interview	96.333	–49.255	0.516	–51.534	0.064	87.949*	55.881*	25.879	6.188
	(330.187)	(139.620)	(4.691)	(50.765)	(12.353)	(47.617)	(31.343)	(17.387)	(9.904)
SW F-Statistic: R.I.	8.771	8.771	8.771	8.771	8.771	8.771	8.771	8.771	8.771
$\overline{Y}_0$	414.302	142.252	0.565	17.864	5.656	30.813	23.956	5.986	0.871
Observations	1,670	1,670	1,670	1,670	1,670	1,670	1,670	1,670	1,670

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Expenditure data from a detailed survey to a random sample of households. Each column summarizes the results for the respective outcome variable following Equation 5(c) for the OLS and the system of equations 5(a), 5(b), and 5(c) for the 2SLS. All estimations control for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects. The CW-OLS calculation follows Bhuller et al. (2020).

6 Discussion

6.1 What is the Cost of Manipulation?

Having documented both the extent and the consequences of welfare manipulation, we next pose the question: how costly is welfare eligibility manipulation to the Georgian government. The figures we provide here establish a conservative lower bound on the true cost of manipulation – for instance, we do not factor in the crowding out effect that a successful manipulation attempt has on formal labor supply, nor the concomitant loss in income tax revenue.

We present our estimates of the costs of welfare manipulation attempts in Table 5. We consider two primary sources of costs to the government. First, the costs from additional welfare payments to households with successful manipulation attempts. Second, the administrative costs of additional interviews. First, in row [1] we calculate the number of successful manipulators around the 65,000 cutoff. These are households whose initial PMT score was just above 65,000, whose final score was below 65,000, and whose request for a repeat interview did not lead to a welfare suspension. Row [2] presents the total increase in benefits of those households who succeed in their manipulation attempt. To compute the calculation of the administrative cost of reassessments, we start in row [3] with the average number of (household-initiated) requested reassessments in a month. The costs we associate to each reassessment is the sum of a 3 hour wage of an enumerator, which we assume is the median wage in Tbilisi.[26] We add 40 Lari per interview as additional administrative costs.

Table 5: The Direct, Financial Cost of Welfare Manipulation

(a) The Cost of Manipulation	
Additional Benefit Payments	
[1] Additional Households at Higher Benefit Level	403
[2] Additional Monthly Benefit Payments (Lari)	106,150
Additional Interview Costs	
[3] Additional Visits to Households	38
[4] Additional Visit Costs	3,116
[5] Total Additional Costs (Lari) Due to Manipulation	109,266
(b) Baseline Costs as a Benchmark	
[6] Number of Households at Baseline	4,854
[7] Total Baseline Benefit Costs (Lari)	441,800
[8] **Cost of Manipulation as a Percentage of Baseline Costs**	**24.73%**

Notes: The table summarizes the monthly cost of manipulation. Rows [1] and [2] are based on the difference between households with successful manipulation attempts with initial PMT scores between 65,000 and 70,000 and those with PMT scores between 60,000 and 65,000. For row [4] we assume that an interview takes three hours, enumerators are paid the median wage in Georgia (14 Lari per hour), plus an administrative cost of 40 Lari. The total additional cost associated with manipulation [5] = [2] + [4]. For row [6] we sum all households with an initial PMT score between 65,000 and 70,000. Row [7] is the associated monthly benefit payments for these reference households at baseline. Row [8] = [5]/[7].

[26] From https://www.salaryexpert.com/salary/area/georgia/tbilisi

Combining these two costs, we document that the Georgian government loses 109,000 Lari per month (about 38,000 USD per month) to welfare eligibility manipulation. To benchmark this cost, we consider the welfare expenditure on households with initial PMT scores of 65,000-70,000. The cost of manipulations is one quarter of the initial welfare payments to this group. This bench-marking exercise highlights the substantial costs of welfare manipulation to the Georgian government, as well as to those on low incomes living in Georgia. This money could be spent to increase the generosity of welfare payments to existing recipients, or to extend the range of PMT scores that yield welfare benefits.

Given this observation, it is worth recapping on why we see such manipulation in the first place – the specific design of the welfare system at this time in Georgia. First, there is a step-wise benefit schedule with a particularly large drop at 65,000 (Figure 1). Second, households are freely able to request a repeat interview, at no direct economic cost. Given the costs we document of welfare manipulation, it is useful to consider how one may address this issue. Removing the ability to request an additional interview may lead to poor households not having the ability to directly address genuine errors, which does not seem appropriate. Removing the extreme steps in the PMT schedule however, seems like a much more direct and simple approach to obviate the large welfare manipulation we document in this paper. This could be done by replacing the large steps multiple smaller steps, or by "smoothing through" these steps, with, for instance, a reverse cumulative distribution function. Such an approach would eliminate from the system any points of the welfare score distribution with a disproportionately large, discrete drop in welfare payments, thereby removing a focal point for manipulation attempts.

6.2 How do our Results Compare With Those From Bunching-Based Methods?

We complete this section by comparing the estimates we derive from our FDD approach with alternative approaches – specifically different bunching methods – that are typically used in the literature. Given the data to which we have access, we have been able to directly measure the extent of welfare manipulation, with the use of an FDD design. This can most easily be seen in Figure 2 and Figure 3. Standard bunching estimators approach the topic from a different, less direct perspective. A useful starting point for this analysis is to recall the distributions of PMT scores we present in Figure 2. The bunching methodologies we implement will estimate the "missing" mass of final PMT scores just above the 65,000 threshold.

In Table 6, we present the results of a comparison between standard bunching estimators and our approach. For the bunching estimator approaches we present the proportion of missing households in the final PMT score distribution as a proxy of those households that successfully manipulated their score. We use two different bunching approaches. The first generation bunching estimator follows the approach of Chetty et al. (2011) and Foremny et al. (2017). We use the distribution of the final PMT score outside a range around 65,000, where households are more likely to be passing from above to below the cutoff (exclusion area), to estimate a counterfactual distribution without bunching.[27] The proportion of missing households are the

[27]We select the exclusion area that minimizes the difference between the missing households above the cutoff and the additional households below the cutoff.

difference between the observed and counterfactual households just above 65,000 within the exclusion area. For the second generation bunching estimator, we follow Zwiers (2021) and take advantage of our data to use the initial PMT score distribution to estimate the counterfactual distribution, having already documented an absence of manipulation in this initial PMT score distribution. Finally we present the first stage coefficient of our FDD estimation (ω_3 from Equation 5(a)), which represents the proportion of households who attempted to manipulate their welfare eligibility due to falling just above the 65,000 cutoff. We present this estimate (row [3*a*]) as an intermediate estimate – it is not the correct estimate for this exercise. We present this here for completeness, as this is the estimate we have used elsewhere in the paper. Row [3*b*] presents the correct estimate for this exercise – the first stage estimate when we consider *successful* manipulation attempts, i.e., manipulation attempts that lead to the types of bunching we document in Figure 2(b).

What is particularly striking about the estimates of the degree of manipulation around the 65,000 threshold that we present in Table 6 is how similar the estimates are from the different methods. Both generations of bunching estimators and our FDD approach yield as good as identical results. This is surprising given the different manners in which these methodologies estimate the degree of manipulation. Such a finding locates our approach within a more familiar terrain for estimating manipulation in response to cutoffs and kinks in public economics.

Table 6: Bunching Estimates – A Comparison Across Existing Methods

	(1)
[1] **First Generation Bunching Approach**	0.133
à la Chetty et al. (2011); Foremny et al. (2017)	[0.050]
[2] **Second Generation Bunching Approach**	0.114
à la Zwiers (2021)	[0.031]
[3*a*] **FDD Approach**: Welfare Manipulation Attempt	0.254
	(0.026)
[3*b*] **FDD Approach**: Successful Welfare Manipulation Attempt	0.117
	(0.018)

Notes: The table shows the proportion of missing households resulting from bunching estimates using bins of 500 points and PMT scores from 40,000 to 90,000. Rows 1 and 2 show the estimated proportion of missing households above the 65,000 cutoff as the difference between the counterfactual and the observed number of households divided by the number of households according to the counterfactual distribution. Row 1 estimates the counterfactual distribution using data from the last PMT score following Chetty et al. (2011); Foremny et al. (2017). We choose the polynomial degree and the exclusion window that minimizes the difference between the excess of households below the cutoff and the missing households above the cutoff. Row 2 uses the first PMT score to estimate the counterfactual following Zwiers (2021). Bootstrap standard error for 1000 repetitions in brackets. Row 3a presents the estimated change in manipulation attempts ω_3 from Equation 5. Analogously, Row 3b presents the estimated change in successful manipulation attempts. Robust standard error in parentheses.

7 Conclusion

In this work, we study a large, nationwide welfare program in Georgia. The program uses a typical form of targeting for a developing country – proxy means testing – and has prominent discontinuities in the schedule that links benefit income to PMT scores. Coupled with the

fact that households may request repeat PMT score assessments, the program gives households incentives to manipulate their welfare eligibility. We start by showing that such manipulation is extensive at a particular threshold, a threshold with a particularly large benefit discontinuity.

We develop a Becker-style model of household manipulation, which we use to inform our empirical approach – a fuzzy difference-in-discontinuities design. We provide extensive evidence of the causal effects of welfare manipulation on labor market engagement, household expenditure, and outcomes of children and young people within the household. We find that women in manipulating households work more and find null effects for men and for the total household income. By probing this finding, we document evidence of welfare benefits crowding out labor market participation for our complier households. We document that also that households spend more in their children. Given our setting, where we have quasi-experimental variation that leads to a simultaneous drop in parental time and a rise in parental income, we study the consequences of household manipulation behavior on a battery of child outcomes, spanning from health and time use investments for children aged 0-5, to educational investments for older children and young adults. We do not find changes in children's outcomes.

We conclude our work by providing a conservative lower bound for the cost of welfare manipulation, which we find to be substantial, amounting to roughly 25% of initial welfare expenditure on our target group of households. In a follow-up project, we are working on alternative benefit schedule designs that can maintain similar levels of benefit payments to target households, yet avoid the large discontinuities that give rise to large welfare eligibility manipulation incentives.

We locate our approach to estimating the prevalence of manipulation within the wider terrain of bunching estimators more commonly used in the public economics literature. Our prevalence estimates coincide almost perfectly with those from two different bunching estimator methodologies.

References

Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231–263.

Agostinelli, F. and G. Sorrenti (2021): "Money vs. time: family income, maternal labor supply, and child development," *University of Zurich, Department of Economics, Working Paper*.

Alatas, V., A. Banerjee, R. Hanna, B. A. Olken, and J. Tobias (2012): "Targeting the poor: evidence from a field experiment in Indonesia," *American Economic Review*, 102, 1206–40.

Andresen, M. E. (2018): "Exploring marginal treatment effects: Flexible estimation using Stata," *The Stata Journal*, 18, 118–158.

Banerjee, A. V., R. Hanna, G. E. Kreindler, and B. A. Olken (2017): "Debunking the stereotype of the lazy welfare recipient: Evidence from cash transfer programs," *The World Bank Research Observer*, 32, 155–184.

Becker, G. S. (1968): "Crime and Punishment: An Economic Approach," *The Journal of Political Economy*, 76, 169–217.

Best, M. C., J. S. Cloyne, E. Ilzetzki, and H. J. Kleven (2020): "Estimating the elasticity of intertemporal substitution using mortgage notches," *The Review of Economic Studies*, 87, 656–690.

Bhuller, M., G. B. Dahl, K. V. Løken, and M. Mogstad (2020): "Incarceration, Recidivism, and Employment," *Journal of Political Economy*, 128, 1269–1324.

Blomquist, S., W. K. Newey, A. Kumar, and C.-Y. Liang (2021): "On bunching and identification of the taxable income elasticity," *Journal of Political Economy*, 129, 2320–2343.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125, 985–1039.

Britto, D. G., P. Pinotti, and B. Sampaio (2022): "The effect of job loss and unemployment insurance on crime in Brazil," *Econometrica*, 90, 1393–1423.

Brollo, F., K. Kaufmann, and E. La Ferrara (2020): "The political economy of program enforcement: Evidence from Brazil," *Journal of the European Economic Association*, 18, 750–791.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust nonparametric confidence intervals for regression-discontinuity designs," *Econometrica*, 82, 2295–2326.

Camacho, A. and E. Conover (2011): "Manipulation of Social Program Eligibility," *American Economic Journal: Economic Policy*, 3, 41–65.

Carneiro, P., L. Kraftman, G. Mason, L. Moore, I. Rasul, and M. Scott (2021): "The impacts of a multifaceted prenatal intervention on human capital accumulation in early life," *American Economic Review*, 111, 2506–2549.

Cattaneo, M. D., M. Jansson, and X. Ma (2020): "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, 115, 1449–1455.

Caucutt, E. M., L. Lochner, J. Mullins, and Y. Park (2020): "Child Skill Production: Accounting for Parental and Market-Based Time and Goods Investments," Working Paper 27838, National Bureau of Economic Research.

Chetty, R., J. N. Friedman, T. Olsen, and L. Pistaferri (2011): "Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from Danish tax records," *The quarterly journal of economics*, 126, 749–804.

Citino, L., K. Russ, and V. Scrutinio (2023): "Manipulation and selection in unemployment insurance," *mimeo*.

Coady, D., M. Grosh, and J. Hoddinott (2004): "Targeting outcomes redux," *The World Bank Research Observer*, 19, 61–85.

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2016): "From LATE to MTE: Alternative methods for the evaluation of policy interventions," *Labour Economics*, 41, 47–60.

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2018): "Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance," *Journal of Political Economy*, 126, 2356–2409.

Cunha, F. and J. Heckman (2007): "The technology of skill formation," *American economic review*, 97, 31–47.

Cunha, F. and J. J. Heckman (2008): "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation," *Journal of human resources*, 43, 738–782.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, 78, 883–931.

Dahl, G. B., A. R. Kostøl, and M. Mogstad (2014): "Family Welfare Cultures," *Quarterly Journal of Economics*, 129, 1711–1752.

Deshpande, M., T. Gross, and Y. Su (2021): "Disability and distress: The effect of disability programs on financial outcomes," *American Economic Journal: Applied Economics*, 13, 151–78.

Diamond, R. and P. Persson (2016): "The long-term consequences of teacher discretion in grading of high-stakes tests," Tech. Rep. 22207, National Bureau of Economic Research.

Econometría (2020): "Impact Evaluation of Targeted Social Assistance (TSA) in Georgia. Final Report," Tech. rep., Econometría Consultores. Evaluation summoned by UNICEF.

Ehrlich, I. (1973): "Participation in illegitimate activities: A theoretical and empirical investigation," *Journal of political Economy*, 81, 521–565.

Foremny, D., J. Jofre-Monseny, and A. Solé-Ollé (2017): "'Ghost citizens': Using notches to identify manipulation of population-based grants," *Journal of Public Economics*, 154, 49–66.

Friedberg, L. (2000): "The Labor Supply Effects of the Social Security Earnings Test," *The Review of Economics and Statistics*, 82, 48–63.

Garbinti, B., J. Goupille-Lebret, M. Muñoz, S. Stantcheva, and G. Zucman (2023): "Tax design, information, and elasticities: evidence from the French wealth tax," Tech. rep., National Bureau of Economic Research.

Gelber, A. M., D. Jones, and D. W. Sacks (2020): "Estimating adjustment frictions using nonlinear budget sets: Method and evidence from the earnings test," *American Economic Journal: Applied Economics*, 12, 1–31.

Gertler, P. J., S. W. Martinez, and M. Rubio-Codina (2012): "Investing cash transfers to raise long-term living standards," *American Economic Journal: Applied Economics*, 4, 164–192.

Geruso, M. and T. Layton (2020): "Upcoding: evidence from Medicare on squishy risk adjustment," *Journal of Political Economy*, 128, 984–1026.

Grembi, V., T. Nannicini, and U. Troiano (2016): "Do fiscal rules matter?" *American Economic Journal: Applied Economics*, 8, 1–30.

Heckman, J. J. and E. J. Vytlacil (2007): "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 4875–5143.

Howell, A. (2022): "Impact of a guaranteed minimum income program on rural–urban migration in China," *Journal of Economic Geography*.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez (2011): "Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark," *Econometrica*, 79, 651–692.

Kleven, H. J. and M. Waseem (2013): "Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan," *The Quarterly Journal of Economics*, 128, 669–723.

Machin, S., S. McNally, and J. Ruiz-Valenzuela (2020): "Entry through the narrow door: The costs of just failing high stakes exams," *Journal of Public Economics*, 190, 104224.

Marx, B. M. (2024): "Dynamic bunching estimation with panel data," *Journal of Econometric Methods*.

Millán-Quijano, J. (2020): "Fuzzy difference in discontinuities," *Applied Economics Letters*, 27, 1552–1555.

Miller, G., D. Pinto, and M. Vera Hernandez (2013): "Risk protection, service use, and health outcomes under Colombia's health insurance program for the Poor," *American Economic Journal: Applied Economics*, 5, 61–91.

Mullins, J. (2022): "Designing Cash Transfers in the Presence of Children's Human Capital Formation," Working Papers 2022-019, Human Capital and Economic Opportunity Working Group.

Nicoletti, C., K. G. Salvanes, and E. Tominey (2023): "Mothers working during preschool years and child skills: does income compensate?" *Journal of Labor Economics*, 41, 389–429.

Saez, E. (2010): "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy*, 2, 180–212.

Van Doornik, B., D. Schoenherr, and J. Skrastins (2018): "Unemployment insurance, strategic unemployment, and firm-worker collusion," *Central Bank of Brazil, Research Department*.

Williams, J. and R. C. Sickles (2002): "An Analysis of the Crime as Work Model: Evidence from the 1958 Philadelphia Birth Cohort Study," *The Journal of Human Resources*, 37, 479–509.

World Bank (2018): "Georgia—First, Second and Third Development Policy Operations," Tech. Rep. 125186, The World Bank Independent Evaluation Group, Washington, DC.

Zwiers, E. (2021): "The Lifecycle Fertility Consequences of the Great Depression and WWII: Evidence from the Netherlands," .

Appendix

A Robustness and Ancillary Results

A.1 The TSA Benefit Schedule

Table A1: TSA benefits by PMT score (Lari per month)

PMT score	Benefit per household member	Benefit per child
0 to 30,000	60	50
30,001 to 57,000	50	50
57,001 to 60,000	40	50
60,001 to 65,000	30	50
65,001 to 100,000	0	50
100,000 or more	0	0

Notes: Payment scheme from January 2019. 1 USD $\approx$ 2.89 Lari.

A.2 Support for the Identifying Assumptions

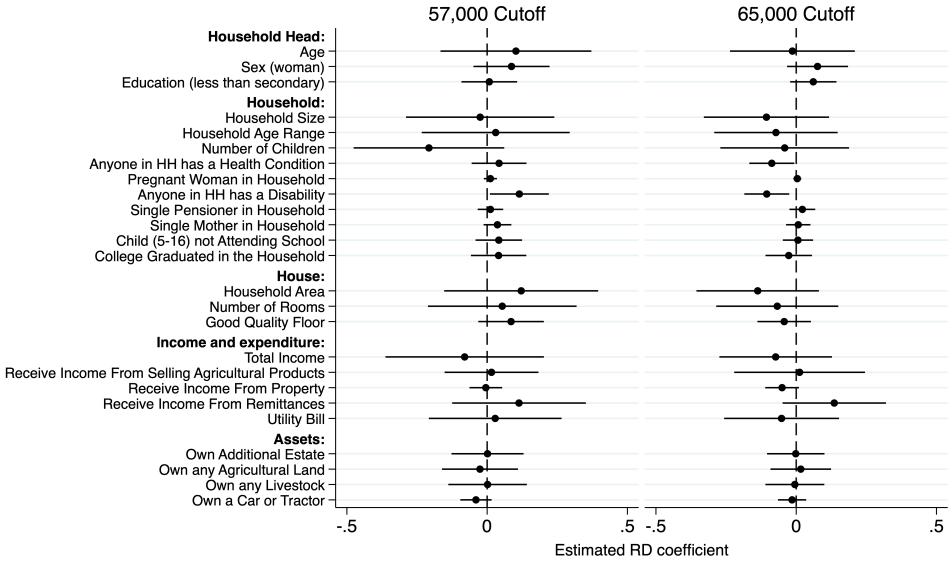
A.2.1 The Continuity Assumption

The key identifying assumption in a RD design is the continuity assumption, which states that the potential outcomes ($Y_{0,i}$ and $Y_{1,i}$) are smooth functions of the running variable $z_{0,i}$ through the cutoff, κ. In order to provide support for this assumption, we implement a series of covariate balance tests, estimating the following specification:

$$X_i = \lambda D_i + g_1^D(z_{0,i}) + \upsilon_i \tag{6}$$

where $g_1^D(z_{0,i})$ is a polynomial of order 2 in $z_{0,i}$ on either side of the cutoff, κ. We present the estimates of λ from these regressions in Figure A1 below. In order to give support to out difference-in-discontinuities strategy, we show continuity over the 57,000 and 65,000 cutoffs. Of the 25 covariates we consider, only a single estimate is significantly different from 0 at the 5% level. We take these results as strong supportive evidence in favor of the continuity assumption.

Figure A1: The Baseline Balance of Covariates

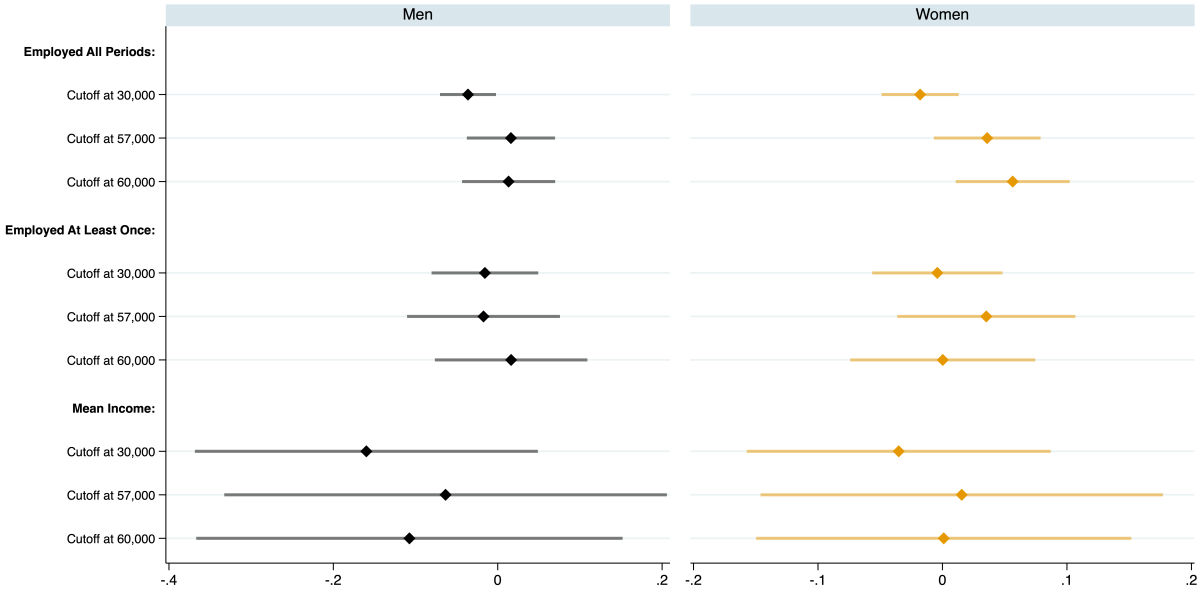


Notes: The figures shows the resulting λ coefficients from Equation 6 for each X variable.

A.2.2 FDD Assumption 1: Homogeneity of the Labor Market Impact of Initial Benefits (B_0) Across Cutoffs.

In addition to the RD assumptions, the FDD method requires two additional assumptions. We use the 57,000 cutoff to estimate the effect of one additional Lari in B_0 on outcome Y. The idea is that changes in outcomes around 57,000 are only driven by B_0, thus, we can use this variation to separately identify the effect of B_0 from the effect of manipulation attempts R, around the 65,000 cutoff. For this approach to be credible, we require that the effect of B_0 on Y is homogeneous across cutoffs. In order to provide evidence in support of this assumption, we estimate the effect of B_0 on labor market outcomes using administrative information from the Revenue Service database around the 30,000, 57,000 and 60,000 cutoffs, using an FRD-IV approach, where being above each cutoff is used to instrument for changes in benefits. One can see that even for the 30,000 cutoff, where households will be considerably poorer than around other cutoffs, the effect of B_0 on formal labor market participation and income, are not meaningfully or statistically different from one other.

Figure A2: The Homogeneous Effect of Benefit Income on Labor Market Outcomes Across Cutoffs

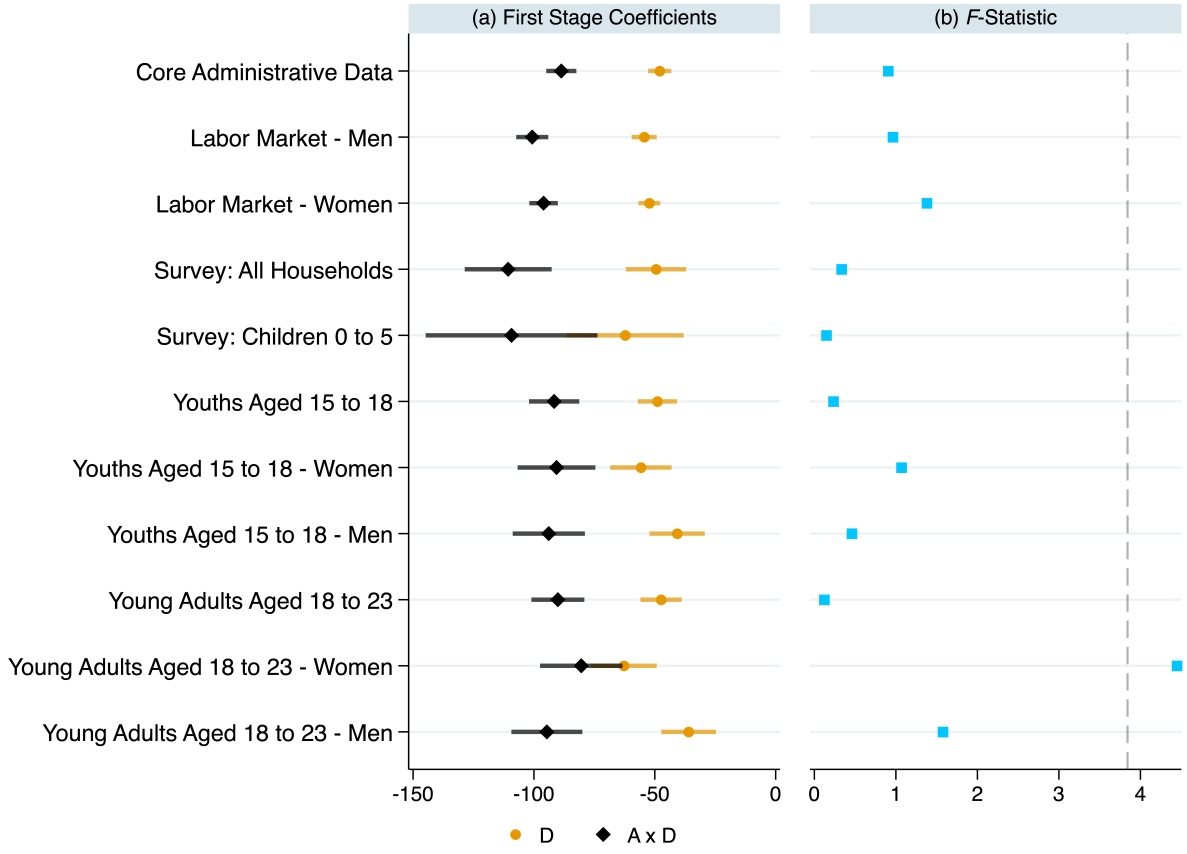


Notes: The figure shows the resulting FRD-IV coefficient and 95% confidence interval of the effect of initial benefits – B_0, on each outcome variable for estimations around each cutoff. Labor market outcomes measured after the last household interview. All estimations control for the individual's age (using a quadratic function), for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects.

A.2.3 FDD Assumption 2: First Stage Coefficients Account for Cutoff-Specific Benefit Changes

The second additional FDD assumption is that our first stage equation for initial benefits adequately accounts for different levels of benefit changes at the respective cutoffs. When a household scores just above 57,000 it only loses 10 Lari per person with respect to a household just below 57,000. However a household scoring just above 65,000 loses 30 Lari per person. Hence, we require that the coefficients that represent the change in benefits for being above 65,000 ($\gamma_1 + \gamma_3$ in Equation 5b) is three times the change in benefits around 57,000 (γ_1). Figure A3 shows that both $\hat{\gamma_1}$ and $\hat{\gamma_3}$ are stable across different data settings. It also shows that estimated coefficients capture the differences in B_0 between both cutoffs, which is necessary for identifying the effect of manipulation attempts on household outcomes.

Figure A3: Testing the First Stage Coefficients for Cutoff-Specific Benefit Changes

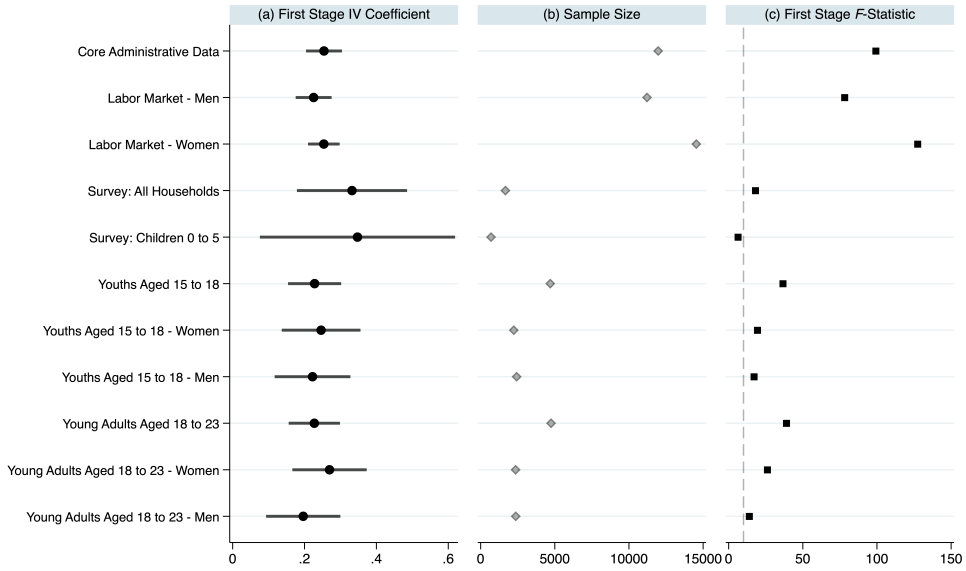


Notes: In Panel (a) we plot the first stage coefficients and 95% confidence interval of D and $A \times D$ for the endogenous variable B_0 for different data sets we used over the paper. Panel (b) represents the F test for the null hypothesis H_0 : $\gamma_1 + \gamma_3 = 3\gamma_1$ following equation 5(a). Dashed line at 3.84. All estimations control for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects.

A.2.4 First Stage Statistics for Manipulation Attempts

For both endogenous treatment variables, the estimated first stage coefficient on the instrument $A \times D$ is extremely stable across all data settings. Figure A4(a) relates to our primary endogenous variable, showing the first stage coefficient for $A \times D$ on the probability of requesting for a reassessment (ω_3 in Equation 5a) across various data settings we use in our analysis.[28] The coefficients are stable, however, precision depends on sample size (Panel (b)), which can clearly be seen in Panel (c) as the F-Test for this first stage only decreases when the sample used decreases.

Figure A4: The Stability of The Effect of $A \times D$ on R Across Samples



Notes: Panel (a) shows the first stage coefficient and the 95% confidence interval of $A \times D$ for the endogenous variable R for different data sets we used over the paper. Panel (b) shows the sample size of each estimation. Panel (c) shows the SW F-Statistic of $A \times D$ on R, and a dashed line at 10. All estimations control for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects.

[28]Figure A3 provides evidence on the stability of the first stage estimates across samples for our auxiliary endogenous variable, initial benefits.

A.3 Additional Labor Market Results

Table A2: Formal Labor Market Engagement at the Household Level

	(1)	(2)	(3)	(4)	(5)
	At Least One Adult Employed at Least Once	All Adults Employed at Least Once	At Least One Adult Employed All Periods	All Adults Employed All Periods	Mean Labor Income of Household
OLS					
Repeat Interview	–0.069***	0.009	–0.009	0.007	–55.305***
	(0.012)	(0.007)	(0.009)	(0.004)	(8.805)
CW-OLS					
Repeat Interview	–0.064***	0.014*	–0.004	0.012***	–51.457***
	(0.014)	(0.008)	(0.010)	(0.005)	(9.660)
2SLS					
Repeat Interview	0.131	0.032	0.439**	0.023	26.746
	(0.269)	(0.139)	(0.202)	(0.057)	(194.333)
SW F-Stat: Repeat Interview	30.216	30.216	30.216	30.216	30.216
$\overline{Y}_0$	0.364	0.076	0.128	0.021	153.207
Observations	11,695	11,695	11,695	11,695	11,695

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Labour market outcomes measured after the last household interview. Each column summarizes the results for the respective outcome variable following Equation 5(c) for the OLS and the system of equations 5(a), 5(b), and 5(c) for the 2SLS. All estimations control for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects. The CW-OLS calculation follows Bhuller et al. (2020).

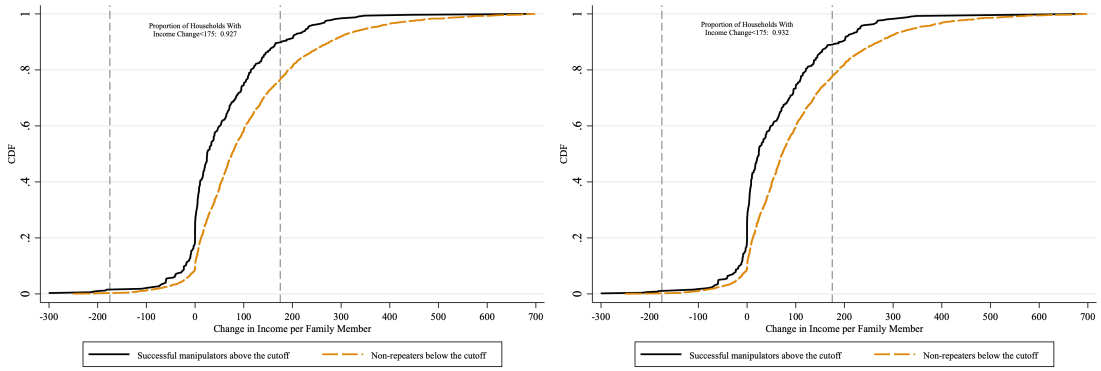
Table A3: Informal Labor Market Participation

	(1)	(2)	(3)	(4)	(5)
	Employed			**Hours per Week**	
	Informal Waged Work	Self-Employed	Unpaid Home Work	For a Wage (Informal)	Self-Employed
A. Men					
OLS					
Repeat Interview	0.014	0.003	0.006	0.219	0.826
	(0.017)	(0.033)	(0.005)	(0.620)	(1.023)
CW-OLS					
Repeat Interview	0.025	–0.001	0.007	0.710	0.414
	(0.023)	(0.042)	(0.005)	(0.908)	(1.315)
2SLS					
Repeat Interview	–0.452	–0.640	–0.016	–16.460	–17.738
	(0.308)	(0.557)	(0.028)	(11.621)	(16.729)
SW F-Statistic: R.I.	8.152	8.152	8.152	8.152	8.152
$\overline{Y}_0$	0.044	0.392	0.004	1.539	9.320
Observations	1,833	1,833	1,833	1,833	1,833
B. Women					
OLS					
Repeat Interview	0.002	–0.012	0.023	0.397	–0.448
	(0.011)	(0.020)	(0.029)	(0.594)	(0.455)
CW-OLS					
Repeat Interview	0.010	–0.008	0.003	1.012	–0.518
	(0.015)	(0.024)	(0.034)	(0.835)	(0.528)
2SLS					
Repeat Interview	–0.180	0.126	0.686	–6.883	–4.357
	(0.172)	(0.340)	(0.522)	(7.687)	(8.339)
SW F-Statistic: R.I.	9.202	9.202	9.202	9.202	9.202
$\overline{Y}_0$	0.032	0.157	0.522	1.393	3.120
Observations	2,146	2,146	2,146	2,146	2,146

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Labour market outcomes measured after the last household interview. Each column summarizes the results for the respective outcome variable following Equation 5(c) for the OLS and the system of equations 5(a), 5(b), and 5(c) for the 2SLS. All estimations control for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects. The CW-OLS calculation follows Bhuller et al. (2020).

A.3.1 Manipulation and Subsequent Income Changes

Figure A5: Formal Labor Income Changes and the 175 Lari per Month per Person Rule



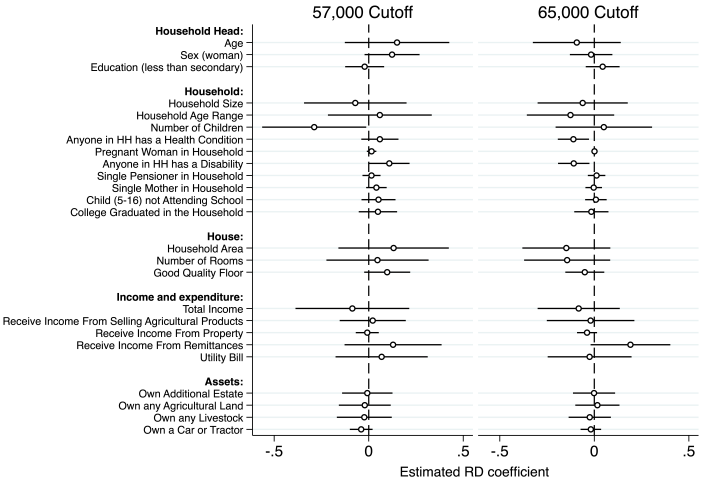
(a) Unweighted (b) Complier Reweighted

Notes: Only households-months that reported income in the Revenue Service data, with first PMT between 60,000 and 70,000. Vertical dashed lines at -175 and 175 Lari. Change in income is measured as the difference between the income earned by the household in the Revenue Service data and the Income from "Salary (including all other types of remuneration)" in the last PMT declaration filled by the household. Successful manipulators are households above the 65,000 cutoff that asked for an additional interview and the final result is a score below 65,000. Non-manipulators are households below the cutoff that did not request an additional interview.

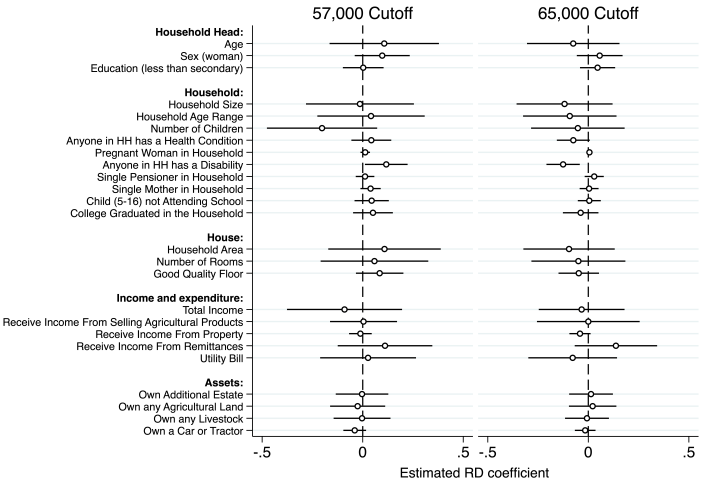
A.4 Sub-Sample Analysis by Manipulation Success Status

Prior to presenting our split-sample analysis by the success statys of the household manipulation attempt, we show that the continuity assumption inherent in our discontinuity-based design hold for both the split-samples (Figure A6).

Figure A6: Baseline Covariate Balance by Manipulation Success Status



(a) Sample With Successful Manipulators



(b) Sample With Unsuccessful Manipulators

Notes: The figures shows the resulting λ coefficients from Equation 6 for each X variable.

We present the analysis on labor market outcomes at the household level in Table A4. We also show the results of the effect of the manipulation attempt on household expenditure on Table A5. However, in this case, given that we are using survey data and splitting the sample, results may not be reliable as the first stage is weak.

Table A4: Household-Level Labour Market Outcomes by Manipulation Success Status

	(1)	(2)	(3)	(4)	(5)
	At Least One Adult Employed at Least Once	All Adults Employed at Least Once	At Least One Adult Employed All Periods	All Adults Employed All Periods	Mean Labor Income of Household
(a) Unsuccessful Manipulation Attempts					
2SLS					
Repeat Interview	0.193	0.052	0.590**	0.049	135.202
	(0.355)	(0.183)	(0.269)	(0.075)	(260.298)
SW F-Statistic: R.I.	24.627	24.627	24.627	24.627	24.627
$\overline{Y}_0$	0.372	0.076	0.131	0.021	157.591
Observations	10,811	10,811	10,811	10,811	10,811
(b) Successful Manipulation Attempts					
2SLS					
Repeat Interview	0.123	0.126	0.713*	0.021	–28.986
	(0.550)	(0.282)	(0.419)	(0.104)	(390.210)
SW F-Statistic: R.I.	18.864	18.864	18.864	18.864	18.864
$\overline{Y}_0$	0.363	0.074	0.126	0.021	152.342
Observations	10,517	10,517	10,517	10,517	10,517

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Labour market outcomes measured after the last household interview. Successful manipulation is defined as (i) having a final PMT score in a category below the initial assigment (increase in benefits) and (ii) a valid score from the SSA. Unsuccessful manipulators are either those who did not get an increase in benefits after the reassessment or whose score was cancel by the SSA due to unreliable information. Each column summarizes the results for the respective outcome variable the system of equations 5(a), 5(b), and 5(c) for the 2SLS. All estimations control for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects.

Table A5: Household Expenditure By by Manipulation Success Status

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
						Children			
	Total	Food	Food Outside of House	Alcohol Tobacco	Adult Clothing	Total	Clothing	Education	Childcare
(a) Unsuccessful Manipulation Attempts:									
2SLS									
Repeated Interview	427.3	-68.60	0.643	-75.15	3.035	146.0*	92.23*	41.64	12.17
	(521.1)	(222.4)	(7.970)	(80.91)	(19.60)	(82.84)	(53.60)	(29.65)	(15.98)
SW F-Statistic: R.I.	7.059	7.059	7.059	7.059	7.059	7.059	7.059	7.059	7.059
$\overline{Y}_0$	419.767	143.654	0.626	17.438	5.838	30.673	23.727	6.039	0.907
Observations	1,490	1,490	1,490	1,490	1,490	1,490	1,490	1,490	1,490
(b) Successful Manipulation Attempts:									
2SLS									
Repeated Interview	356.5	17.70	2.191	-70.01	11.08	170.3*	104.8	51.05	14.44
	(614.0)	(237.3)	(8.864)	(89.08)	(22.55)	(103.2)	(65.13)	(36.71)	(18.87)
SW F-Statistic: R.I.	4.603	4.603	4.603	4.603	4.603	4.603	4.603	4.603	4.603
$\overline{Y}_0$	416.284	141.011	0.598	18.106	5.715	31.033	24.041	6.116	0.877
Observations	1,529	1,529	1,529	1,529	1,529	1,529	1,529	1,529	1,529

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Expenditure data from a detailed survey to a random sample of households. Successful manipulation is defined as (i) having a final PMT score in a category below the initial assignment (increase in benefits) and (ii) a valid score from the SSA. Unsuccessful manipulators are either those who did not get an increase in benefits after the reassessment or whose score was cancel by the SSA due to unreliable information. Each column summarizes the results for the respective outcome variable following the system of equations 5(a), 5(b), and 5(c) for the 2SLS. All estimations control for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects.

A.4.1 Marginal Treatment Effects

In this section we estimate marginal treatment effects (MTEs) based on our binary instrument. We present a brief overview of the MTE approach in order to anchor what we do here, then proceed to discuss how we may identify the MTE when we have only a binary instrument. We follow the exposition of Andresen (2018), and additionally use his Stata command – `mtefe` – in order to estimate the MTEs.

The starting point for all MTE approaches is the generalized Roy model, which takes the form of:

$$Y_j = \mu_j(X) + U_j = X\beta_j + U_j \qquad \text{for } j = 0, 1 \tag{7}$$

$$Y = DY_1 + (1 - D)Y_0 \tag{8}$$

$$D = \mathbb{1}[\mu_D(Z) > V] = \mathbb{1}[Z\gamma > V] \qquad \text{where } Z = (X, Z_-) \tag{9}$$

Y_0 and Y_1 are respective outcomes in untreated (no manipulation) and treated (manipulation) states.

The unobservable, V, in (9) – which we view as the unobserved resistance to treatment status – plays a core role in the interpretation of MTEs. If V is continuously distributed, we can rewrite (9) as $D = \mathbb{1}[P(Z) > U_D]$, where $P(Z)$ is the propensity score and U_D represents quantiles of V and is uniformly distributed.

We follow Brinch et al. (2017), who show how to identify the MTE in the presence of a binary instrument using the separate MTE estimation approach developed by Heckman and Vytlacil (2007). This involves separately estimating the conditional expectation of the outcome variable for the treated and untreated samples separately, combined with an appropriate function for the conditional expectation of the error terms:

$$\mathbb{E}(Y_0 \mid X = x, D = 0) = x\beta_0 + E(U_0 \mid U_D > p) = x\beta_0 + K_0(p) \tag{10}$$

$$\mathbb{E}(Y_1 \mid X = x, D = 1) = x\beta_1 + E(U_1 \mid U_D \leq p) = x\beta_0 + K_1(p)\,, \tag{11}$$

where $p = P(Z)$. Brinch et al. (2017) describe in detail the identification challenges posed when attempting to use a binary instrument to estimate MTEs, show how the separate approach with linear specifications for K_0 and K_1 enables identification of the MTE, and present a graphical representation of the geometry of the linear MTE. We leave the interested reader to view that paper for further detail.

In order to operationalize the MTE approach – which considers a single endogenous regressor – in our FDD setting of two endogenous regressors and two instruments, we take the following steps.

MTE Bootstrap Routine

Step 1: we estimate the first stage equation for our nuisance endogenous variable, $B_{0,i}$ – this is Equation 5(b). We then calculate predicted values of the outcome from this first stage: $\hat{B}_{0,i}$.

Step 2: we implement the MTE procedure, using $\hat{B}_{0,i}$ in place of $B_{0,i}$, and thus effectively

reducing our set of FDD equations to a simpler case of a fuzzy RD design with a single endogenous variable, for which the MTE approach is better suited.

Because we include a predicted regressor in the MTE procedure, our standard errors will be incorrect. Additionally, as the output associated with the `mtefe` package notes, several of the components of the MTE procedure are based on estimated objects – the propensity score, the means of the X vector, and the treatment effect parameter weights. For these reasons, we implement a bootstrap procedure that wraps around both Step 1 and Step 2. We use 500 bootstrap iterations of this bootstrap procedure to provide standard errors on the key output in Table A6 below.

A key output of most author's MTE analysis is a plot of the MTE estimate along the distribution of U_D – the quantiles of the unobserved resistance to treatment. For an example, see the discussion of empirical applications in Cornelissen et al. (2016), where the slope of the MTE curve takes center stage in the discussion of the two papers' results. There is good reason for this – the slope of the MTE curve informs us about patterns of selection on unobservable gains to treatment. Where the MTE curve is downward-sloping, this indicates positive selection on unobservable gains – those with the highest marginal treatment effect are those with the lowest unobserved resistance to treatment, hence conditional on observables, the most likely to take up treatment. The converse applies.

In Table A6, we present the key outputs of our MTE analysis for formal labor market outcomes considering three different functional form specifications for the manipulation model (Equation 9). Other dimensions on which authors often consider heterogeneity analysis – the orders of polynomials used in the specification of $K()$ and the type of MTE approach – are not permissible in our binary instrument case. Here we must use the separate approach, and $K()$ must be linear (Brinch et al., 2017). In lieu of 18 MTE curves, we extract the slope of the term in p, and the p–value of this term from each of our MTE specifications. This is a sufficient statistic for the key information contained in the linear MTE. As one can see from Table A6, for women there is a uniform pattern of negative selection on unobservable gains to treatment across all choice models and all labor market margins. When we estimate a linear propensity score – which most closely mimics the linear first stage of the 2SLS approach – we find that this negative selection is statistically significantly different from zero at conventional levels for the extensive labor supply margins that we consider. We discuss the ramifications of this patterns of negative selection on unobservables gains in Section 5.3.2, and reflect on what we can learn from this at a broader level in Section 5.4.

Table A6: Marginal Treatment Effects for Formal Labor Market Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	Men			**Women**		
	Manipulation Model Function:			**Manipulation Model Function:**		
	LPM	Probit	Logit	LPM	Probit	Logit
A. Employed At Least Once						
ATE	0.353	–0.512*	–0.430**	1.283**	0.344*	0.285*
	(1.096)	(0.271)	(0.216)	(0.600)	(0.207)	(0.169)
ATT	–0.115	–0.093	–0.080	0.264	0.113	0.088
	(0.326)	(0.156)	(0.132)	(0.198)	(0.109)	(0.095)
ATUT	0.448	–0.597*	–0.501*	1.482**	0.389	0.324
	(1.308)	(0.322)	(0.257)	(0.708)	(0.244)	(0.199)
LATE	–0.084	–0.135	–0.119	0.355*	0.147	0.118
	(0.340)	(0.159)	(0.136)	(0.201)	(0.108)	(0.094)
p-value: Essential Heterogeneity	0.674	0.143	0.129	0.089	0.348	0.338
Selection on Unobservable Gains	Negative	Positive	Positive	Negative	Negative	Negative
B. Employed All Periods						
ATE	–0.229	–0.258	–0.207	0.961**	0.160	0.128
	(0.702)	(0.167)	(0.133)	(0.433)	(0.146)	(0.118)
ATT	0.206	0.067	0.054	0.162	0.007	0.002
	(0.216)	(0.097)	(0.082)	(0.109)	(0.071)	(0.062)
ATUT	–0.317	–0.324	–0.259	1.117**	0.190	0.153
	(0.839)	(0.199)	(0.159)	(0.515)	(0.173)	(0.139)
LATE	0.184	0.039	0.029	0.217**	0.014	0.006
	(0.223)	(0.101)	(0.087)	(0.110)	(0.070)	(0.061)
p-value: Essential Heterogeneity	0.532	0.070	0.072	0.069	0.369	0.369
Selection on Unobservable Gains	Positive	Positive	Positive	Negative	Negative	Negative
C. Mean Income						
ATE	–327.1	–115.2	–97.7	465.4*	131.6	109.8
	(564.0)	(155.1)	(126.5)	(260.1)	(86.4)	(68.9)
ATT	–141.5	–125.6	–104.1	85.2	47.1	41.5
	(209.9)	(100.1)	(84.7)	(88.1)	(45.0)	(38.9)
ATUT	–364.6	–113.0	–96.4	539.6*	148.1	123.1
	(675.0)	(185.6)	(151.4)	(308.4)	(102.7)	(81.8)
LATE	–160.4	–131.3	–109.6	104.4	44.5	37.0
	(212.9)	(101.5)	(86.6)	(85.0)	(44.3)	(38.7)
p-value: Essential Heterogeneity	0.741	0.974	0.993	0.156	0.438	0.456
Selection on Unobservable Gains	Positive	Negative	Negative	Negative	Negative	Negative

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Labor market outcomes measured after the last interview each household has. Standard errors are bootstrapped using 500 iterations. The bootstrap routine incorporates the first stage regression for our auxillary endogenous variable – benefit income – the prediction of benefit income, and the MTE routine using predicted benefit income. All estimations control for the individual's age (using a quadratic function), for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects.

A.5 The Impact of Welfare Eligibility Manipulation on Child Outcomes

In Section 5.3 we find that a manipulation attempt leads to a significant increase in labor supply for women. In addition, in Table 4 we document that children are the primary beneficiaries of the corresponding increase in household spending. As our gaze turns now to child outcomes, we note that the evidence we document so far identifies two, countervailing forces on the child skill production function within households that attempt to manipulate their welfare eligibility. The increase in income, and concomitant expenditure on children, should have a positive impact on childhood skill production, whereas the fact that parents now have less time available will likely lead to a decrease in the production of childhood skills (Cunha and Heckman, 2007; Caucutt et al., 2020; Agostinelli and Sorrenti, 2021; Nicoletti et al., 2023; Mullins, 2022).

A.5.1 Early Childhood Investments

We first consider child outcomes in the first six years of life, a key period for childhood interventions if there are dynamic complementarities in investments in children across their life cycle (Cunha and Heckman, 2007, 2008; Cunha et al., 2010). We bring two data sources to bear to study this early childhood investment – administrative data on vaccinations, and survey data on health and time investments in children. The data we have available to us will predominantly reflect time costs, rather than money costs.

Once again, when working with the survey data we face a very small sample size. The consequence of this can be seen again by viewing the first-stage F statistic in Table A7, which is large for the administrative data sample, but below standard thresholds for the survey data sample. This failure of the rank condition when using the survey data occurs when we use the same specification, and considering the same PMT score range, as we do with the administrative data, so we are confident that this loss in significance reflects the small sample size of the survey data sample.

We first show in Columns 1 and 2 that a manipulation attempt has no impact on vaccinations. Given that the main parental cost of such health investments are time-based, these results are informative of household responses to changing labor supply patters as a consequence of a manipulation attempt. When we turn to the survey data, we still do not find any effect on health investments (columns 3 and 4), neither on the time parents spend with their kids.

A.5.2 Mid- and Late-Period Childhood Skill Investments

We now shift our attention to the later periods of childhood skill investments, using administrative educational data to study outcomes at two key educational margins – high school and university attendance.[29] The administrative data we use contains information on school/university attendance for the previous three years. Using this information, along with child age, we can observe if school-age children are still attending school. For 19 and 20 year olds, we can observe if they attended school in the previous years. Combining the available information, we construct an indicator for high school attendance during ages 15-18.

[29]Post-compulsory education – both secondary and university education – is free in Georgia.

Table A7: Early Childhood Investments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Administrative Data		**Survey Data**				
	Full Vaccines	Full exc. DTaP/ IPV/ Hib/ HepB	Any Health Check-ups	Number of Health Check-ups	Screen-time	Time Spent Together – Total	Time Spent Together – Reading
OLS							
Repeat Interview	–0.014	–0.029	–0.047	–0.170	–5.533	–7.714	1.023
	(0.020)	(0.022)	(0.046)	(0.450)	(6.069)	(5.906)	(2.088)
CW-OLS							
Repeat Interview	–0.019	–0.029	–0.051	–0.415	–4.305	–10.343*	0.190
	(0.023)	(0.025)	(0.047)	(0.461)	(6.146)	(5.868)	(2.060)
2SLS							
Repeat Interview	0.015	–0.190	0.567	3.445	–20.803	7.604	–4.377
	(0.327)	(0.395)	(0.557)	(5.628)	(83.031)	(74.238)	(28.642)
SW F-Statistic: R.I.	12.390	12.390	3.554	3.554	3.554	3.554	3.554
$\overline{Y}_0$	0.205	0.288	0.807	5.457	62.157	76.044	15.845
Observations	3,148	3,148	701	701	701	701	701

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Each column summarizes the results for the respective outcome variable following Equation 5(c) for the OLS and the system of equations 5(a), 5(b), and 5(c) for the 2SLS. Vaccination's indicators (columns 1 and 2) based on the official vaccines-age schedule. Columns 3 and 4 are refer to all health check-ups since birth. Columns 5 to 7 measured in hours in the week prior the survey. All estimations control for the child's age and gender, for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects. The CW-OLS calculation follows Bhuller et al. (2020).

Columns 1-3 show the impact of a household manipulation attempt on high school attendance of children ages 15-18 within the household. Both the standard and complier re-weighted OLS estimates are both very close to zero. The 2SLS estimates are imprecise but positive. The imprecision of the estimates is not driven by small sample sizes or a weak instrument – there appears to be little effect of a household manipulation attempt (which, given what we show in previous sections, is likely best thought of as a bundle of outcomes) on high school attendance for older children within the household. Columns 4-6 present estimates for university attendance for teens and young adults age 18-23 still living at home. The OLS show that the correlation between a manipulation attempt and university attendance is negative. The complier re-weighted OLS highlight the lack of treatment effect heterogeneity among the compliant sub-population. The 2SLS estimates tell a different story, as the coefficient for men is still negative but it is positive for women. The effect on women is large but imprecise and not statistically significant.

Summarizing, we document that manipulation attempts lead to an increase in women labor market participation (less time with children) with null effects on households' income, and an increase in children related expenditure (more money to children). We then explore what is the effect of this trade off on children's welfare, but do not find any improvements driven by a manipulation attempt.

Table A8: High School and Unversity Attendance

	(1)	(2)	(3)	(4)	(5)	(6)
	High School			**University**		
	All	Males	Females	All	Males	Females
OLS						
Repeat Interview	–0.007	–0.017	0.002	–0.033*	–0.039	–0.031
	(0.014)	(0.022)	(0.019)	(0.020)	(0.028)	(0.032)
CW-OLS						
Repeat Interview	–0.003	–0.013	0.018	–0.033	–0.033	–0.030
	(0.015)	(0.024)	(0.020)	(0.022)	(0.030)	(0.035)
2SLS						
Repeat Interview	0.248	0.299	0.312	–0.045	–0.724	0.318
	(0.361)	(0.581)	(0.427)	(0.342)	(0.688)	(0.393)
SW F-Statistic: R.I.	15.630	7.229	9.046	23.614	6.055	20.969
$\overline{Y}_0$	0.813	0.793	0.835	0.281	0.213	0.352
Observations	6,764	3,502	3,251	4,749	2,366	2,357

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Each column summarizes the results for the respective outcome variable following Equation 5(c) for the OLS and the system of equations 5(a), 5(b), and 5(c) for the 2SLS. Outcomes measured after the last PMT interview. High school attendance (columns 1 to 3) for children 15 to 18. University attendance (columns 4 to 6) for individuals 18 to 23. All estimations control for the individual's age, for the first PMT first score above and below each cutoff, household's characteristics such as households size, age, gender and education of the household head, dummies for presence of a pregnant woman, a person with a physical condition, a person with a disability, a pensioner, and a single mother, maximum level of education in the household, and region-by-quarter and interview time fixed effects. The CW-OLS calculation follows Bhuller et al. (2020).

B Identification of θ_R

We start from Equation 5c:

$$Y_i = \theta_R R_i + \theta_B B_{0,i} + \theta_3 A_i + g_Y^{D,A}(z_{0,i}) + X_i'\theta + \mu_{Y,i}$$

We take expectations of Y with respect to our instruments $(D, A \times D)$, conditional on the score (z), and the observable (X). We define $E[Y_i|D = d, A = a, X_i, z_i] = Y^{da}$, $E[B_{0,i}|D = d, A = a, X_i, z_i] = B_0^{da}$, and $E[R_i|D = d, A = a, X_i, z_i] = R^{da}$. Taking into account that X is continuous around the cutoffs, and under the FDD assumption that R does not change around the $57,000$ cutoff ($R^{10} = R^{00}$), we can identify θ_B by subtracting $Y^{10} - Y^{00}$:

$$\theta_B = \frac{Y^{10} - Y^{00}}{B_0^{10} - B_0^{00}}$$

As explain before, θ_B is identified using the variation around the $57,000$ cutoff. Now, using variation around $65,000$, by subtracting $Y^{11} - Y^{01}$:

$$Y^{11} - Y^{01} = \theta_R(R^{11} - R^{01}) + \theta_B(B^{11} - B^{01})$$

Under the assumption that the effect of one additional Lari is the same around $57,000$ and $65,000$, we can plug the estimate for θ_B on the latest equation. After reorganizing, we get the following expression:

$$\theta_R = \frac{(Y^{11} - Y^{01}) - (Y^{10} - Y^{00}) \times \left(\frac{B_0^{11} - B_0^{01}}{B_0^{10} - B_0^{00}}\right)}{R^{11} - R^{01}}$$

As in Grembi et al. (2016) and Millán-Quijano (2020), to identify θ_R we use the difference in the variation around $65,000$ minus the difference around $57,000$. However, given that the change in B_0 is different around $57,000$ and $65,000$, we weight the difference in outcomes by $\left(\frac{B_0^{11} - B_0^{01}}{B_0^{10} - B_0^{00}}\right)$, that takes into account the difference in the change in B_0.