CERGE
Center for Economic Research and Graduate Education
Charles University

Explorations into Behavioral Phenomena

William Morris Appleman

Dissertation

Prague, October 2021

Dissertation Committee:

Fabio Michelucci (Università Ca' Foscari and CERGE-EI, Chair)

Nikolas Mittag (CERGE-EI, Local Chair)

Daniel Munich (CERGE-EI)

Patrick Gaule (University of Bristol)

Referees:

Michael McBride (University of California, Irvine)

Giuseppe Attanasi (Université Côte d'Azur)

Contents

Acknowledgements

First, I would like to thank my PhD supervisor/advisor/chair, Fabio Michelucci, for all his patience, diligence, endurance, and excellent guidance for the far too many years it took to complete this dissertation. My eternal gratitude also goes out to Nikolas Mittag, Randall Filer, Patrick Gaule, Zurab Abramishvili, Levan Bezhanishvili, and Radek Janhuba for all their time, effort, comments, and contributions throughout this process.

Furthermore, my appreciation goes out to those who impacted the individual chapters of this dissertation. Regarding chapter one, I am grateful to all those above as well as Giuseppe Attanasi, Gary Charness, Martin Dufwenberg, Michael McBride, and John Duffy for all their suggestions. Special thanks go to my amazing and magnanimous programmer, Martin Fibiger; spectacular lab assistant, Jan Vavra; lab manager, Tomas Miklanek; the professor that inspired the research, Peter Katuscak; my grant administrator, Michael Jetton; and the entire grant administration and finance staff at CERGE-EI. With respect to the second chapter, my coauthors and I additionally thank Stepan Jurajda, Alena Bicakova, Simon Clark, and Maxym Brukhanov for all their insights and detailed comments and Daniel Munich for his contributions, including access to the data. For the third chapter, my coauthors and I send our additional appreciation to Milan Scasny and Michael McBride for their useful input.

I would also like to recognize the great assistance and training provided by our fantastic ASC staff at CERGE-EI: Deborah Novakova, Paul Whitaker, Andrea Downing, Grayson Krueger, and former member Robin-Eliece Mercury. My thanks to all my professors and the entire faculty at CERGE-EI for all you have taught me and for the amazing environment you created. Acknowledgement and gratitude are due to CERGE-EI and the Global Development Network (the first chapter was supported by a grant from the CERGE-EI Foundation under a program of the Global Development Network (RRC16+32)) and GEMCLIME (research visitation to UCI supported by the H2020-MSCA-RISE project GEMCLIME-2020, GA no. 681228) for all the funding that allowed me to conduct and complete this dissertation.[1]

Lastly, but most of all, I thank my family and friends for their love and support. This would have been impossible without the encouragement, sacrifice, and fortitude of my wife, Nataliya. My deepest love and thanks go out to her; my son, Joshua; my mother, Zdenka; and the rest of my family.

Prague, Czech Republic — William Morris Appleman

October 2021

[1] All opinions expressed herein are those of the author(s) and have not been endorsed by the CERGE-EI Foundation, the GDN, or GEMCLIME. Responsibility for any errors remains with the author(s).

Abstract

Taxation impacts social welfare in an intricate manner. Currently employed tax instruments throughout the world possess inherently different salience characteristics. Tax salience effect refers to the optimization error occurring when agents do not fully account for taxes that do not appear in posted prices. Purchase size effect (PSE), refers to how the tax salience effect changes based on the size of the monetary stake of the purchase. The first chapter of this dissertation aims to assess whether participants in a 'quasi-field' laboratory experiment with an online shopping environment, real goods, low and high price/type goods, and with conspicuous taxation cues will make tax salience "errors" and if they will make less of them when purchasing more expensive goods. This is the first experiment in the literature to focus on the PSE dimension of tax salience and to divide participants into low and high income-/wealth-analogous budget levels.

The results confirm consistent and significant tax salience effects and reveal the PSE despite the cues. Tax salience effect is most driven by high-budget participants while low-budget participants exhibit several more significant PSE difference estimates. Such a combination suggests that those with more constraining budgets were less likely to be making a tax salience error generally and even lesser likely with more expensive goods, even though the high-budget group faced higher utility stakes. This may imply the dominance of one of the budget adjustment mechanisms, which is predicted to lead to a positive social welfare outcome.

Supplemental, informative findings indicate that these behaviors are not gender specific and appear to be an intentional/strategic choice, substantially mediated by one's budget constraints and shopping speed. These results suggest that rational inattention drives the tax salience effect. With correlation to income/wealth measures, these findings imply an innate progressivity in less salient consumption tax instruments.

The second chapter examines the presence of parental preference for a particular gender of children. We test for the predominance of the two main explanations for the existence of such preference, namely differences in the costs of raising sons and daughters versus the gender bias (corresponding to parental utility derived from a child's gender or from characteristics exclusive to that gender). First, we use recent EU-SILC data from several Balkan and Scandinavian countries to confirm that the gender of the firstborn predicts the likelihood of a given family having three children or more—a common measure of parental gender preference. We confirm son preference in certain Balkan countries and daughter preference in Scandinavian countries. Both having a first child of the preferred gender and of the more costly gender can decrease the probability of having three or more children because parents may already be content or may lack sufficient resources, respectively. Next, we use information on household consumption to differentiate the two explanations. We argue that under the differential cost hypothesis, parents of children of the more costly gender should spend more on goods for children and less on household public goods as well as on parental personal

consumption. In contrast, having children of the preferred gender should increase spending on household public goods since such families have higher marriage surplus and are more stable. Our evidence corroborates the cost difference explanation in countries exhibiting daughter preference.

The third chapter evaluates a one-time, immediate policy initiative enacted by the Republic of Georgia that shifted public office working hours from 10:00-19:00 to 9:00-18:00, which affected the work schedules of government employees. Although the policy affected approximately 200,000 employees, it has never been evaluated; and to our knowledge, nor has any similar policy in any economic literature. The effects of the policy impact gender and family types asymmetrically, relating this paper most closely to work-family conflict literature, which provides a framework for making two opposing predictions based on gender similarity and difference models. Furthermore, we examine how the policy impacts gender inequality through female labor participation.

Given that the policy did not affect the private sector, we employ a difference-in-differences approach using the National Statistics Office of Georgia Households Incomes and Expenditures Survey from 2013-2016. We find that the policy primarily produces a significant reduction in the average level of working hours of full-time employees with children, directly in line with the prediction following the gender similarity model. We also find a significant increase in average work hour engagement by women without children. However, the placebo effect analysis identifies this as an already existing trend and the short-term analysis indicates that this is an ordinal reaction to the reduction of engagement by full-time employees with children. We conclude that this increase is a secondary, indirect effect and that the policy did not directly cause an increase in female labor participation. Furthermore, since men with children were most negatively affected and women picked up the gains, the policy may also indirectly increase overall gender equality.

Abstrakt

Zdanění ovlivňuje společenský blahobyt složitým způsobem. Současné daňové nástroje používané v různých zemích světa mají ze své podstaty odlišné charakteristiky daňové salience. Efekt daňové salience je druh chyby rozhodování spotřebitelů, kteří plně nerozpoznávají daně v případech, kdy nejsou explicitně uvedené v ceně. Efekt velikosti nákupu (PSE) představuje změnu efektu daňové salience v závislosti na ceně nákupu. Cílem první kapitoly této disertační práce je zjistit, zda se účastníci „kvazi-field" laboratorního experimentu, v prostředí online nákupu reálného zboží s nízkou nebo vysokou cenou a s nápadnými daňovými náznaky, dopouštějí chyb daňové salience a zda se jich dopouští méně při nákupu dražšího zboží. Jedná se o první experiment v literatuře, který se zaměřuje na PSE v daňové salience a rozděluje účastníky do skupin s nízkým a vysokým rozpočtem analogicky k jejich příjmu/bohatství.

Výsledky potvrzují konzistentní a signifikantní efekt daňové salience a odhalují PSE i navzdory významně nápadným daňovým náznakům. Efekt daňové salience je nejvíce hnán účastníky s vysokým rozpočtem, zatímco účastníci s nízkým rozpočtem vykazují větší efekt velikosti nákupu. Tato kombinace naznačuje, že se účastníci s omezenějším rozpočtem obecně méně často dopouštěli chyby daňové salience. U dražšího zboží se jí dokonce dopouštěli ještě méně, přestože skupina s vysokým rozpočtem měla v sázce vyšší užitek. To může ukazovat na dominanci jednoho z mechanismů přizpůsobení rozpočtu, který by dle očekávání měl vést k pozitivnímu výsledku společenského blahobytu.

Další doplňující výsledky studie naznačují, že toto chování se neliší podle pohlaví a navíc se zdá, že jde o záměrnou/strategickou volbu, která je podstatně ovlivněná rozpočtovým omezením a rychlostí nakupování. Tyto výsledky naznačují, že hnací silou efektu daňové salience je racionální nepozornost. V kombinaci s měřítky příjmu/majetku tato zjištění naznačují přirozenou progresivitu v méně salientních nástrojích spotřební daně.

Druhá kapitola zkoumá preference rodičů pro určité pohlaví dětí a testuje, které ze dvou hlavních vysvětlení takové preference převažuje, a to zda je příčinou rozdíl v nákladech na výchovu synů a dcer nebo se jedná o genderovou předpojatost (odpovídající užitku rodičů odvozenému od pohlaví dítěte nebo od charakteristik výlučných pro toto pohlaví). Nejprve využíváme aktuální data ze šetření EU-SILC v několika balkánských a skandinávských zemích, abychom potvrdili, že pohlaví prvorozeného dítěte predikuje pravděpodobnost, že daná rodina bude mít tři a více dětí, což je běžně používaný ukazatel genderové preference rodičů. Potvrzujeme existenci preference synů v některých balkánských zemích a dcer ve skandinávských zemích. Narození prvního dítěte jak preferovaného, tak nákladnějšího pohlaví, může snížit pravděpodobnost, že rodiče budou mít tři nebo více dětí, jelikož již mohou být spokojeni či mohou čelit nedostatku prostředků. V další části studie využíváme informace o spotřebě domácností k rozlišení obou vysvětlení. Tvrdíme, že podle hypotézy rozdílných nákladů by rodiče dětí nákladnějšího pohlaví měli více utrácet za zboží pro děti a méně za

společné statky domácnosti i za osobní spotřebu rodičů. Naopak, pokud mají děti preferovaného pohlaví, měly by se zvýšit výdaje na společné statky domácnosti, jelikož takové rodiny mají vyšší manželský přebytek a jsou stabilnější. Naše zjištění jsou v souladu s vysvětlením založeným na rozdílu v nákladech v zemích, které vykazují preferenci dcer.

Třetí kapitola hodnotí jednorázovou a okamžitou reformu, ve které Gruzínská republika posunula pracovní dobu státních úřadů z 10:00-19:00 na 9:00-18:00, což ovlivnilo pracovní hodiny státních zaměstnanců. Přestože se tato reforma dotkla přibližně 200 000 zaměstnanců, nebyla nikdy dříve vyhodnocena, a pokud je nám známo, tak v ekonomické literatuře nebyla hodnocena ani žádná podobná reforma. Tato reforma má asymetrický dopad dle pohlaví a typu rodin, což tento článek zařazuje nejblíže k literatuře zabývající se konflikty mezi prací a rodinou, která poskytuje rámec pro vytvoření dvou protichůdných teoretických výsledků na základě modelů genderové podobnosti a genderové rozdílnosti. Dále zkoumáme, jak tato reforma ovlivňuje genderovou nerovnost prostřednictvím zapojení žen na trhu práce.

Vzhledem k tomu, že reforma neměla vliv na soukromý sektor, používáme metodu rozdílu v rozdílech na datech z šetření National Statistics Office of Georgia Households Incomes and Expenditures Survey z let 2013-2016. Výsledky ukazují, že reforma přináší především signifikantní snížení průměru odpracovaných hodin zaměstnanců na plný úvazek, kteří mají děti, což je v souladu s očekáváním vyplývajícím z modelu genderové podobnosti. Zjišťujeme také významné zvýšení průměrně odpracovaných hodin u žen bez dětí. Analýza placebo efektu však tento jev identifikuje jako součást již existujícího trendu a krátkodobá analýza naznačuje, že se jedná o reakci navazující na snížení angažovanosti zaměstnanců na plný úvazek s dětmi. Došli jsme k závěru, že toto zvýšení je sekundárním, nepřímým efektem a že reforma přímo nezpůsobila zvýšení zapojení žen na pracovním trhu práce. Navíc vzhledem k tomu, že zatímco muži s dětmi byli ovlivněni nejvíce negativně a ženám intervence přinesla pozitivní efekt, může reforma také nepřímo zvýšit celkovou rovnost žen a mužů.

Preface

In the second half of the twentieth century, empirical and experimental studies uncovered a myriad of stark behavioral differences to standard economic theory throughout all branches of economics. In response, behavioral economics was founded and developed along two main dimensions: cognitive psychology and cognitive limitation. Kahneman and Tversky are often recognized as the founding fathers of the former dimension, which explains human behavior through mental shortcuts such as heuristics, cognitive biases, and psychological and social influences. On the other hand, bounded rationality models assume various forms of cognitive limitation to explain behavior that is inconsistent with standard theory. However, these concepts overlap and coexist with a fuzzy border, in which people may strategically invest time and attention needed for thorough rational thought into areas they consider important, based on their individual conditions, constraints, and preferences, with the remaining behavior relying more on instincts and subject to "errors", following the definition of standard models. In fact, Kahneman eventually conceptualized a framework, dual-system theory, that cast these two dimensions as two coexisting human thought systems.

This dissertation examines the above dichotomy in the context of certain behavioral phenomena in public economics and gender economics. The first chapter contributes to literature that shows how purchase decisions are affected by the salience of a tax. In particular, it explores the underlying mechanisms and dynamics of this phenomenon by focusing on one mechanism of the bounded rationality model that is strongly related to the fuzzy border between "rational" behavior and "error".

The other chapters focus on behavioral phenomena with gender as the core element. The second chapter goes to the heart of this gender element, examining the potential causes of parental preferences for the gender of their children. Two main competing explanations for gender preferences are rational-based differences in the costs and benefits of children of a given gender versus psychological- and emotional-based bias for a given gender. The third chapter of this dissertation evaluates how public sector employees react to a one-time, immediate policy that changes their working hours by shifting them earlier by one hour, which causes asymmetrical effects along gender and family types. Predicted reactions are hypothesized along work-family conflict frameworks, following the more rationally based gender similarity model and the more socially and psychologically influenced gender difference model.

1 Making Salience More Salient: Purchase Size Effect

1.1 Introduction

Taxation is a leading cause of deadweight loss, yet it is indispensable to public finance, long-term socio-political stability, and economic growth, and it is the predominant solution to several major economic issues, including public goods provision and negative externalities. Optimal tax theory aims to minimize excess burden and incentive distortions while maximizing the benevolent social planner's aggregated welfare problem, which includes the use of tax tools with particular traits, such as Pigouvian taxes, to achieve a second-best solution. Associated literature illustrates that tax salience could be another useful tool for policymakers to achieve optimal taxation aims. Currently employed tax instruments throughout the world possess inherently different salience characteristics. Empirical and experimental evidence demonstrates a consistent and significant behavioral "error"—a salience effect—influencing how people incorporate costs into their economic decisions based on price framing.[2] Tax salience effect refers to the optimization error that occurs when agents do not fully account for taxes that are not included in posted prices in their purchase decisions.

Key papers of this literature include Chetty, Looney, and Kroft (2007 & 2009) (further referred to as CLK), Goldin (2015), and Reck (2016). CLK normatively assess tax salience using a measure of inattention/underreaction to consumption taxes not posted in purchase prices, θ, which compares the effects on demand elasticities from price changes versus tax changes.[3] With the aim of properly identifying an optimal tax system structure, CLK also propose a positive, bounded-rationality model in which consumers face cognitive costs from calculating final prices when taxes are not included in price tags. CLK's theoretical and empirical findings, as well as those of the other studies, indicate potentially large-scale welfare consequences from the tax salience effect.[4] Accordingly, understanding the inner workings of tax salience is important to economic theory and practical policy application.

This paper concentrates on a specific dimension of tax salience: purchase size. The purchase size effect (PSE) refers to how the tax salience effect changes based on the monetary stake (price) of a purchase. Understanding whether and how the PSE affects the tax salience effect

---

[2] Examples from empirical, theoretical, and experimental studies have shown wide ranging effects, including the dominance and robustness of "add-on" (less salient) pricing (Gabaix & Laibson, 2006), electronic toll payment systems causing 20 to 40% increases in toll rates over manual payment systems (Finkelstein, 2009), higher revenue for sellers that "shroud" or do not include shipping & handling costs in the auction price (Hossain & Morgan, 2006), and even the significant increase of labor supply under a consumption tax versus an equivalent income tax (Blumkin, Ruffle & Ganun, 2012).

[3] In particular, they define this key measure, θ, as the ratio between the demand elasticity of changes in less-than-salient taxes with respect to demand elasticity of changes in price (and/or taxes fully included in the price). θ is used to measure an individual's incorporation of a less-than-salient tax—where zero is equal to the complete neglect of the tax into the total price and one is equal to the complete incorporation of the tax in the total price—as well as at the aggregate level to denote the ratio of consumers that pay attention and incorporate the less-than-fully-salient tax into their purchase decision.

[4] See section 1.2 for a detailed description.

is important for several reasons. First, it will help explicate the behavioral relationships involved, such as the strategic (in)attention element of the bounded-rationality model, which may clarify the underlying dynamics of tax salience. Second, as noted above, since the implications of the tax salience effect on social welfare could be considerable, comprehending the impact and mechanisms of the PSE may prove critical. Third, while CLK, Reck (2016), and others touch upon purchase size in their models and implications, it remains an underexplored[5] element that may be consequential to theoretical outcomes[6] as well as pragmatical concerns.[7] Furthermore, determining whether the error is intentional, and how, could also help shape policy for optimal practical application.

To that end, the research aim herein is to assess whether participants in a laboratory experiment with conspicuous cues regarding taxation will make tax salience "errors", and whether they will make fewer errors when purchasing more expensive goods.[8] The laboratory experiment employs an innovative, 'quasi-field' methodology; it uses a pseudo-realistic, online shopping website and remuneration mainly in the form of real goods that participants ordered. Moreover, the laboratory setting offers the notable advantages of stronger randomization allowing for better causality inference; the ability to include zero-quantity shopping choices as observations and the associated construction of more suitable variables of choice;[9] more information and data about participants (shoppers) and their behavior and characteristics; and greater setting control, which, in combination with the increased data about shoppers, results in the ability to perform expanded analyses. While the external validity of laboratory experiments is considered weaker than field experiments, this 'quasi-field' methodology preserves the advantages of the laboratory setting while mitigating external validity concerns.[10] Furthermore, the ability to perform expanded analyses allows this study to examine several model-extrinsic elements that may also influence tax salience, finding unexpected and consequential insights, such as weaker-than-anticipated correlation with cognitive abilities and a pivotal relationship with shopping speed. In addition, the experiment features a novel goods selection procedure that increases both subject choice and test power, and it is the first

[5] Until now, no study has focused on PSE, and it was only an auxiliary element in Taubinsky & Rees-Jones (2018).
[6] Fundamentally so in Reck's (2016) model. Reck focuses on tax size; only noting PSE could produce the same effect. In contrast, implications for CLK's budget adjustment mechanism are unforeseen and meaningful.
[7] For example, if PSE is strategic and naturally reduces the tax salience effect, then that diminishes the serious distortionary income effect argument that underlies the potentially socially-negative findings of CLK (2009) and Reck (2016) and may mean that less-than-salient taxes are mostly welfare enhancing.
[8] It is not the aim herein to test merely a difference in the monetary stake of some neutral, objective good, but to follow the good-type conjecture made by CLK, and disputed by Gamage & Shanske (2010 & 2011), which specifies the effect on welfare in terms of first choice and income effects and the resulting budget adjustments from overspending on taxes for luxuries and then being able to afford necessities. Moreover, the experiment aims to present the shopping environment as realistically as possible to the participants in order to attempt to coax more natural shopping behaviors.
[9] This is elaborated upon in further detail in section 1.4.
[10] Please see footnote 25 for further discussion.

experiment in this branch of research to divide participants into low- and high-budget shoppers.[11]

The results show a consistent and significant tax salience effect along all measures and reveal the PSE despite the experiment's conspicuous presentation of taxation elements.[12] Specifically, participants, on average, make 3% fewer tax salience "errors" in their shopping choices with 9.13 times more expensive goods, which is a nontrivial figure. Incorporating the different elasticities of the goods, the tax salience effect measure, θ, reveals about two-thirds as much attention to taxes for less expensive goods than for more expensive goods, representing a material reduction in the tax salience effect. This adds credence to the CLK bounded-rationality model, as PSE is one of the model's comparative static predictions. Moreover, informative findings from the supplemental analyses (see sections 1.5.2 and Appendix A6) about participant intention, understanding, and shopping speed indicate that the 'underreacting' or disregarding of taxes not posted in prices is presumably intentional and probably reflects a strategic allocation of attention, further supporting the model's appropriateness. This finding, combined with the findings that budget constraints consequentially affect salience-related shopping behaviors[13] and that income/wealth heterogeneity may be correlated with the tax salience effect, imply the existence of inherent progressivity in less-than-salient consumption taxes.

High-budget participants displayed substantially stronger and more significant tax salience effects[14] while low-budget participants displayed more significant purchase size effects. This finding suggests that those with less to spend were less likely to make a tax salience error generally and even lesser likely with more expensive goods. This is further compelling considering the well-documented effect in economic experiments that higher stakes cause participants to pay greater attention to an activity than their lower-stake counterparts. Together, these discoveries may provide clarification on the budget adjustment mechanism of the CLK model, implying the dominance of one of the model's predictions leading to a strictly positive-to-welfare outcome.

---

[11] While this statement is true, prior to publication, Huseynov et al. (2019) also had such a division and findings. Budget constraints may be a point of substantial interaction at different purchase size levels.

[12] The experiment directly explained the two types of tax systems involved in the instructions, stated (with photos) which goods would be taxed before every period, in one form or another tax-inclusive prices were visible at all times in the two main pages of the e-shop, and shopping choices were costlessly changeable. Such information and cues are not found in close proximity to shopping decisions in the real world, and thus the tax salience effect and PSE are expectedly muted and represent a lower envelope for how strong the effects may actually be. The result is an overall $\theta = 0.50$, compared to CLK's stronger $\theta = 0.35$. Note that experiments tend to result in much higher values of θ from less-than-salient taxes (RST) than empirical works, which usually find quite low values, e.g. CLK (2009) found $\theta = 0.06$.

[13] For example, the results of a supplemental analysis on budget constraint influence imply that budget constrained shopping behavior may be related more to real purchasing power than the experiment's exogenously imposed budget constraints. This discovery is important, and it would probably be unattainable outside of the controlled laboratory setting.

[14] In fact, most of the results throughout the paper were mainly driven by the larger magnitude and statistical significance of high-budget participant output.

1.2 Literature Review

CLK (2007 & 2009) combine a field experiment, an empirical study, and a theoretical assessment of tax salience. For all three portions of the paper, CLK develop elasticity-based formulas for empirical assessment and welfare analysis. They use the formulas to assess efficiency costs related to agents imperfectly optimizing due to less-than-salient taxes. Their field experiment was run in three of five large grocery stores in Northern California and involved placing tax-inclusive price tags on a selection of commonly purchased items for three weeks. It revealed a θ of 0.35, which means that aggregate consumer demand was reduced by 35% of the amount it would have been with an equivalent fully-salient-tax/price change.[15]

Their empirical study focused on how demand for beer changed based on variance in price through excise taxes (a fully-salient tax included in the price on the price tag) and retail sales taxes (not included in the price tag but added at the register). They found a significant reduction of consumer demand behavior equivalent to nearly the full, price-elastic amount when the price tag included the tax compared to almost no reduction when the tax was a retail sales tax. Specifically, they found a θ of 0.06.

In the theoretical portion, CLK use Bernheim and Rangel's "refinement" method to calculate consumer welfare and excess burden (CLK, 2009). The results show that salience decreases excess burden (deadweight loss) through a reduction in consumer choice distortion and increases excess burden through distortionary income effects. Consumers sustain second-order losses through income effects from ignoring some taxes, yet there are first-order gains for government revenue and overall social welfare. How tax salience (θ) affects consumer welfare depends on their utility functions and budget adjustment methods. If their utility function is quasilinear in nature or individuals adjust by reducing future purchases of both taxed and untaxed goods, then the tax salience effect is strictly positive to welfare. However, with arbitrarily-separable general utility functions, cognitive costs, and budget adjustment predominantly of the untaxed good, the tax salience effect can cause distortionary income effects greater than choice substitution utility gains. CLK propose a bounded-rationality model with cognitive costs that can explain tax salience effects as well as their other stylized facts. Comparative statics of that model predict four types of consumer behavior, one of which indicates that consumers will pay more attention to less-than-salient taxes as purchase size increases. Extrapolating to include heterogeneity and purchase size variation into the model would result in a situation where agents have varying tax salience effects at different purchase sizes, which could be relative to those agents based upon their individual budget constraints. This could inform CLK's budget adjustment dynamics and narrow their theoretical conclusions about how tax salience affects welfare as well as imply that less-than-salient tax instruments may be inherently progressive while reducing consumer choice distortion and excess burden.

---

[15] The strategy CLK employed, Strategy 1 (CLK, 2009, p. 1150), is the strategy used in the PSE laboratory experiment of this paper. Thus, the resulting θ figures should be reasonably comparable.

Goldin (2015) and Reck (2016) build their analyses directly on CLK. Goldin (2015) focuses on how salience can improve a government's tax revenue collection system as well as the social welfare. Notably, his second proposition shows that at the fully-salient margin, consumers are strictly better off from a reduction of salience in their taxes (an increase in the tax salience effect). Intuitively, Goldin explains that by reducing the salience of a given tax from full salience, the government can reduce excess burden, as it is able to raise the same amount of tax revenue at a lower level of distortionary effect on consumption. Even though this causes consumers to unintentionally over-consume the taxed good in relative terms, the welfare cost of the optimization error (which depends on the differences in marginal utilities) is small in comparison, especially near the fully-salient margin. Reck (2016) considers the implications of agents adjusting their economic decisions from neglecting less-than-salient taxes to incorporating them fully into their decisions when the utility stakes are large enough (through high tax rates or purchase size), which he dubs "debiasing". He shows that if agents debias, the appeal of utilizing biases for social welfare improvements is significantly undermined.

To summarize, Goldin (2015) shows that at the initial margin, individuals are better off with less-than-salient taxes than not, while Reck (2016) shows that at a certain point, individuals will debias and will face a higher marginal excess burden at that point than under the tax-inclusive counterpart. Moreover, Reck finds that total excess burden may potentially be worse under a tax system that employs less-than-salient taxes when there are high cognitive costs and when a large enough portion of the population debiases. Consequently, it is necessary to know if and how debiasing occurs along the main comparative static dimensions within tax salience, notably tax size and purchase size.

Tax size has been explored in at least three experiments, all of which find large changes in tax size (e.g. tripling current US retail sales taxes) to be a relatively weak or inconclusive cause of debiasing.[16] Despite the consequential impact purchase size may have on social welfare outcomes through debiasing or CLK's budget adjustment mechanism, nearly no attention has been given to PSE in the literature. As of yet, one paper found explicit evidence of PSE.

[16] Most studies that explore tax size effect find very modest evidence of its existence. On the high side, Taubinsky & Rees-Jones (2018) find that tripling normal retail sales taxes in the US (commonly around 7-8%) results in consumers paying twice as much attention to the taxes—which is a decrease in the tax salience effect, as debiasing predicts. Conversely, evidence from one tax size effect experiment (Feldman, Goldin, & Homonoff, 2018) indicates a greater salience effect from higher tax rates. This, however, may possibly be explained as being the unintentional outcome of how the prices were presented. To maintain the same final price including taxes for each of the tax size variations, the fully-salient price tags featured lower prices for higher tax rates. This could mean that individuals who were not fully incorporating the less-than-salient taxes into their purchase decisions may have been paying greater attention to what they thought were very good prices compared to the real-world prices with which they were familiar, which may have further exacerbated the tax salience effect during the rounds in which taxes were higher, as those prices would have appeared that much more attractive. Regarding empirical studies focused on tax size effect, Zheng et al. (2019) mostly find that increases in excise (salient) tax rates cause steeper drops in demand than increases in sales (less-than-salient) tax rates. Concerning cigarettes, Goldin & Homonoff (2013) find that lower income consumers begin to debias due to the tax size effect, and Hyunjin (2019) find weak evidence of a converging drop in demand between excise and sales tax. Chuang (2019) and Tiezzi & Verde (2016) find tax size effects for plane tickets and gasoline, though their larger fluctuations in price muddles the comparison and the analyses did not include variation in excise tax.

Taubinsky & Rees-Jones (2018) ran a large-scale (N=2998) online shopping experiment that focused on tax salience and its innerworkings by using variation in tax size, consumer heterogeneity, and prices of goods. They find statistically significant evidence of a reduction in the tax salience effect between products with prices of less than $5 and greater than $5. Their experiment employed a Becker-DeGroot-Marshak (BDM) willingness-to-purchase shopping mechanism. It has been argued that BDM offers advantages in terms of reservation price and incentive-compatible demand elicitation, but that it also drastically differs from normal shopping and can introduce new behavioral/psychological effects. However, Karni and Safra (1987) show that BDM is not incentive compatible with lotteries, and Horowitz (2006) finds that it is not incentive compatible even under no uncertainty. Incentive compatibility in BDM requires an expected utility maximization assumption. This paper relaxes that assumption and eliminates the potential behavioral effects by removing the intermediary apparatus.

1.3 Experimental Design

As opposed to the Taubinsky & Rees-Jones (2018) design, this experiment reproduces a natural shopping environment (e.g. weekly grocery shopping from a local e-grocer). To make the experimental environment more realistic to the participants, the e-shop was programmed in the familiar guise of a common online shopping portal (www.itesco.cz). It included an assortment of less expensive grocery goods and more expensive department store goods, and remuneration consisted of the actual goods the subjects ordered, as well as cash from their unused budget—collected together one week after participating. Data collection and analysis design were derived from the CLK assumptions: "(A1) consumption is a sufficient statistic for utility: tax design affects utility only through its effect on the agent's consumption" and "(A2) when tax-inclusive prices are fully-salient, the agent chooses the same allocation as an optimizing agent" (CLK, 2009, p. 1170). Hence, all relevant data comes directly from participant consumption choices.

Participants were randomly divided into low- and high-budget groups as they signed in to the experiment (they were unaware of the existence of different budget groups) and repeated the shopping activity a total of fifteen times in two shopping environments, consisting of lower and higher priced goods, which only differed by type of taxation. One shopping environment employed a fully-salient consumption tax (included in the price tag) while the other employed a consumption tax that was less salient (excluded in the price tag, but charged at the checkout page). The shopping environments were made up of three different marketplaces: a grocery store, a department store, and a superstore (all goods together). The null hypothesis herein is that agents are susceptible to tax salience equally between lower and higher price-level goods.

The following summarizes what a random subject faced during the experiment. For more specifics about the design, see the detailed design subsections below. A participant signed in

to the lab and listened to oral instructions read from a script (first page of Appendix A1) by the experimenter. Subsequently, they signed in and completed the online instruction section, in which the two taxation environments were described, and participant understanding was tested in a comprehension quiz. Next, they proceeded to a goods selection section, in which they chose 10 items they preferred from a list of 50 available goods from both the grocery and department marketplaces (for a total of 20 goods).

Thereafter, the participant began the main shopping section comprised of two control and two treatment periods, each consisting of the three marketplace rounds. Before each period, an instruction page displayed all goods subject to taxation in that period. Each market round featured a budget bar that displayed the tax-inclusive subtotal of their cart and their remaining budget, as well as a line-item-accounting checkout page that specified all taxes.[17] Both the order of the periods and the three rounds within each period were randomized. The participant then faced three final superstore rounds constituting the 'welfare control' section for a total of fifteen shopping rounds. Subsequently, the subject answered the experiment's built-in demographic and internal control survey. Once the subject finished, the computer randomly chose a single round of the experiment for remuneration and displayed the goods from the round the computer chose on the earnings page. Next, the participant was given a short paper questionnaire with CRT and math questions. Once completed, it was collected, and the participant was asked to sit quietly until all other participants had finished. Both the purchased goods and cash earnings were collected a week later.

1.3.1 Treatment Variables

To assess the PSE, three main treatment variables were employed. First, the key variable necessary to assess the PSE was a good price/type binary[18] variable for purchase size. The lower-cost-per-item grocery-store goods take a value of 0 and the higher-cost-per-item department-store goods take a value of 1. In particular, the grocery goods consisted of low cost, common household items, with mostly short- to medium-term perishability. The department goods consisted of much higher priced[19] items that are not common necessities[20] (e.g. certain electronic goods, gadgets, perfumes, colognes, jewelry, etc.; see Appendix A4 for

[17] Given such prominent taxation cues, the findings should be considered a lower bound of the effect.
[18] A continuous analysis of good price was also conducted and is discussed in Appendix A6.3.10.
[19] On average, there was a 9.13 times cost difference between the low- and high-cost good prices. The low-cost minimum was CZK 1.9, while the maximum was CZK 35.16. The high-cost minimum was CZK 99, and the maximum was CZK 319. Therefore, the most expensive high-cost good cost 167.89 times more than the most expensive low-cost good, and the least expensive high-cost good cost 2.82 times more than the most expensive low-cost good.
[20] This division is based both on the purchase size effect as well as the purchase type argument made by CLK and the counterarguments of Gamage & Shanske and Goldin. Specifically, CLK (2007 & 2009) use the example of the purchase of an automobile and the resulting budgetary effect; that is, agents can overspend on luxuries and then not afford necessities. Both Goldin (2015) and Gamage & Shanske (2010 & 2011) contend that this argument will not hold and that consumers will more likely respond to low salience taxes through the reduction of luxury goods rather than necessities.

a complete list). This treatment variable was presented through three randomly ordered shopping "rounds"—a grocery store marketplace round, a department store marketplace round, and all goods together in a single 'superstore'[21] marketplace round—comprising one experimental "period".

The second variable was the traditional tax salience binary variable of consumption tax type. That variable takes the value of 0 when the shopping environment employed a fully-salient (included in the price tag) excise or value added tax (VAT) versus a value of 1 when the environment employed a less-than-salient (excluded in the price tag but charged at the checkout counter) retail sales tax (RST). Following the literature, VAT corresponds to tax-inclusive pricing and RST corresponds to tax-exclusive pricing. The third variable was a binary variable indicating whether a good was taxed (value of 1) or not (value of 0). More specifically, participants in the experiment faced consumption choices between taxed and untaxed goods. During one half of the main shopping section of the experiment, half of the goods were randomly subject to taxation, followed by the other half being taxed during the other half of the main shopping section. All the findings are assessed based on whether or not and how subjects substitute between untaxed goods and taxed goods between the control and treatment periods.[22] The difference-in-difference of the latter two variables reveals the tax salience effect occurring in the experiment. Altogether, the three variables reveal the PSE through the difference in tax salience behavior between the good types.

1.3.2 Participants and Instructions

The experiment was conducted at the "LEE" lab at VSE (University of Economics) in Prague, Czech Republic over a two-week period during the last two weeks of April 2017 with 192 subjects.[23] Participants were mostly students, visiting students, or alumni of VSE and Charles University. Subject pool statistics are available in Appendix Figure A5.1. Special care was taken to ensure that all the subjects faced as close to the exact same experimental circumstances as possible. Experimental sessions were held at the exact same time, 15:00-16:45, on Tuesdays, Wednesdays, and Thursdays[24] of the two weeks. Earnings were always given out exactly seven days after the session at the same time, 12:00-14:30.

---

[21] The combination removes the experimenter-imposed budget and other marketplace divisions, but theoretically reduces the power of the test. It also provides evidence related to the purchase type issue argument as well as reduces remuneration expenses and helps prevent corner choices.

[22] Per the CLK assumptions (and branch literature), subjects reveal the tax salience effect through the difference in their consumption choices between a fully-salient tax (i.e. the VAT tax of the control periods) and a less-than-fully-salient tax (i.e. the RST of the treatment periods). The tax salience effect is revealed when, *ceteris paribus*, consumers/participants purchase a greater portion of taxed goods under the less-than-fully-salient tax than under the fully-salient tax.

[23] Originally, the goal was 200 subjects, but due to no-shows combined with experimental budget constraints, the experiment was concluded with a total of 192 subjects.

[24] Studies have shown that day-of-the-week effects are strongest on Mondays and Fridays. Moreover, analyses showed that clustering at the session level had almost no affect upon significance and was not worthwhile, while clustering at the student and shopper-type levels was relevant and applied.

Each participant was signed in to the lab by a lab assistant to preserve anonymity. Once all were seated, a scripted page of oral instructions was read aloud to them by the experimenter. Then the participant signed in to the actual experiment, which began with the instruction section. There, they read the written instructions, familiarized themselves with the e-shop environment and its navigation through annotated screenshot visualizations, answered a short instruction comprehension quiz, and completed the section by reading the revealed correct answers to the comprehension quiz. Please see Appendix A1 for a copy of the instructions. Subjects were randomly divided into two "income" levels of low and high (see Table 1); they were unaware of this division.

To alleviate any potential notion that the results may be caused by any lack of information, machination, or subject unfamiliarity, the goods subject to taxation in a particular shopping "period" were announced before every "period", and the difference between value added tax (VAT) and retail sales tax (RST) was detailed in the instructions. In fact, due to the conspicuousness of this information in the instructions and recall cues that are not found commonly in tax salience experiments or in such close approximation to shopping decisions in the real world, the results herein represent a lower envelope for how strong the effects may actually be.[25]

1.3.3 Goods Selection

Between the instruction section and the shopping section, subjects were asked to go through the many items offered in the grocery and department marketplaces and select ten items out of a total of fifty from each market they would be interested in purchasing during the experiment. The goods were listed without brands and prices—participants knew they would be facing real prices from the oral instructions. This selection stage, which had not been used in any previous experiment (to the best knowledge of this author), was introduced to reduce noise and increase the power of the test, while still ensuring that the items available for purchase would be in the subject's utility preference space. Several follow-up questions were in the experiment's survey and questionnaire (described in section 1.3.6) that ensured this assertion could be scrutinized through robustness checks. Increasing the likelihood of the goods being in the participant's utility preference space should reduce the salience-related costs of paying attention while

[25] Regarding external validity, the use of windfall income may be faulted as diminishing the external validity through a transformation of subject behavior (as it does in the dictator and ultimatum games). However, this design is much more complex than the dictator and ultimatum games. Furthermore, an initial game (i.e. the market game) or a productivity exam were considered as a way to have subjects arrive at endogenously created income, but this was abandoned in favor of productivity correlation, fatigue, and deception concerns. Also problematic may be the exogenously separated budgets in the grocery and department store rounds of each period, but the fully combined superstore rounds are in the design to allow the analysis of more free form shopping without any experimenter-imposed divisions. Furthermore, while some observers may note that there may be deficiencies in terms of the mitigation of certain biasing effects, it is the overall balance of learning effect, experimenter effect, order effect, flexibility, and fatigue in the experimental design that should prevail over individual issues.

shopping, as they would be interested in purchasing those goods and should optimize accordingly.

1.3.4 Shopping Environment

Participants faced a combined control-and-treatment main shopping section composed of four shopping "periods". Each period consisted of three rounds, each corresponding to a market type discussed above: a grocery store market (with 10 subject-selected goods), a department store market (with 10 subject-selected goods), and a full combination of the two in a superstore market (meaning the full combination of all goods, budgets, and time limits of the grocery and department store markets). There was a total of twelve rounds of shopping in this main section.

The four periods consisted of two control and two treatment periods. The only differences between those sets of periods were the set of taxed/untaxed goods and the tax treatment (VAT vs. RST). More specifically, after the goods selection section, the computer randomly divided each set of goods in half (once for all four periods). Within the two control and two treatment periods, the participants faced one period with one half of the goods taxed. In the other period, they faced the other half of the goods being taxed.[26] That is, in one of the two control (and in one of the two treatment) periods, half of the goods of each market were taxed while the other half remained untaxed. In the second period, the goods that had been taxed in the first period were tax-free and the goods that had not been taxed were now taxed. The order of the four periods was randomized as were the order of the three rounds within each period. See Table 1 below for a visualization of the period structure.

After the main treatment-control section, participants completed an additional, 3-round shopping segment, herein dubbed "welfare control" but inconspicuously presented to participants as just another shopping period, which aimed to capture the subjects' preferences and local elasticities.[27]

Each round had a shopping page and a checkout page. Participants were able to see a full breakdown of the taxes by good and were able to seamlessly go between them. In addition, each round had a budget (see Table 1) and time limit (165 seconds for shopping and 45 seconds for checkout during the grocery and department store rounds; double of each during the

[26] Having half the goods taxed one period and their counterparts taxed the next ensures that the purely randomized division does not conceal behavior by coincidentally dividing the total set such that stronger or weaker individual preferences between the sets obscure tax elasticity choices. In addition, such period pairs provided another robustness check and another layer of analysis that could have further illuminated within- and between-subject behavior.

[27] This portion was composed of three superstore style rounds, and shopping was stylistically just as in a treatment (VAT) period. In the first round, all the goods were for sale tax-free, and the subjects chose the quantity of each item they would like to purchase at the given price within their given budget. Next, the prices were all raised to the standard tax-inclusive level and the subjects chose quantities again at the new price levels. Finally, in the last round, they faced the tax-inclusive prices once again, but with a compensated budget; altogether, with the aim of revealing their underlying preferences through their full Slutsky decomposition.

superstore rounds), representing consumer opportunity costs. Screenshots of this section are available in Appendix A1.

Table 1: Breakdown of the period structure

Each period has three rounds:	Low-income subject budget	High-income subject budget	Notes
Necessity Market	400Kč (CZK)	800Kč (CZK)	• Rounds were randomly ordered each period • Budgets were constant, except for welfare control round 3 (compensated budget)
Luxury Market	400Kč (CZK)	800Kč (CZK)	
Superstore (Combined) Market	800Kč (CZK)	1600Kč (CZK)	

1.3.5 Remuneration and Incentives

Remuneration consisted of both cash and real goods. Both forms of remuneration were collected together exactly seven days after participation in order to mitigate strategic risk avoidance behavior choices. As mentioned above in section 1.3.4, opportunity costs of agent economic decisions were represented through time limits in each round and with the remaining budget being linked directly with remuneration. Fifty percent[28] of the unused portion of a subject's budget—which represents a quasi-linear opportunity cost of other, undefined future consumption—made up the cash portion. A relatively higher value of the goods combined with the goods being chosen based upon individual preferences during the goods selection section resulted in most subjects not choosing a "click through" corner solution of purchasing zero goods and just taking cash.[29] The real goods portion was composed of the actual goods purchased by the participant in the one and only randomly chosen round by the computer—a method that incentivizes each shopping round individually.

As the experiment is an individual decision problem, participants were able to complete the experiment at their own pace, which should have helped alleviate participant fatigue. At the same time, the participants needed to remain in the lab until all participants from that session completed the experiment. This should have further reduced "click through" corner solution

[28] Fifty percent was not a random choice, but the result of a main focus of the pilot study, in which 30%, 50%, and 70% were tested. T-tests showed that 50% was significantly different than the other two in both low- and high-budget groups, that 30% and 70% were not significantly different from one another, and that 50% resulted in the most balanced shopping. The result was decidedly akin to a Laffer curve. In addition, the 50% group from the pilot had the most significant main analysis results.

[29] For example, if the computer randomly chose a round with a budget of CZK 800, each unspent crown would be worth .5 real CZK. Participants that chose to simply "click through" earned CZK 400.

incentives and introduced motivation to pay greater attention to the instructions and the shopping as they were not able to leave the lab earlier even if they rushed through.

The checkout page clearly communicated exactly which goods were taxed and participants were able to costlessly click between shopping and checkout pages and adjust their shopping choices. Moreover, even within the shopping page, the final price of goods including taxes was visible in the budget bar. Therefore, participants had ample opportunity to observe final tax prices even in the treatment periods. All of this combined with the prominent taxation cues discussed earlier supplement the contention that the results are more tenably a lower bound.

1.3.6 Survey and Questionnaire

After the shopping portion of the experiment, each subject was required to complete a programmed, anonymous survey. It included several questions about education, occupation, age, gender, personal income, household income, nationality, personal wealth, family wealth, and several additional control questions. The survey contained a total of 13 demographic question and 11 control questions. The full survey is presented as Appendix A2.

After the survey, the purchases from the randomly chosen round were displayed on the earnings page, which concluded the programmed part of the experiment. Finally, this was followed by a one-sheet paper questionnaire that had two additional experimental control questions about preferred brands and budget constraints. Neither the survey nor the questionnaire was directly incentivized.[30] It also asked three Cognitive Reflection Test (CRT) questions developed by Frederick (2005) to assess a specific form of cognitive ability—a participant's capability or inclination to reflect on a question versus answering intuitively. Essentially, this test should assess whether certain participants make decisions more intuitively or with more cerebral reflection, which could potentially be highly correlated with the tax salience effect. The questionnaire concluded with three arithmetic questions that directly corresponded to the tax calculations that would be required of the participants for the least expensive good price, the most difficult mid-level price, and the most expensive good price. These questions intended to assess a specific form of individual productivity that is directly related to the cognitive costs associated with tax salience. Appendix A3 is a reproduction of the questionnaire.

[30] However, the results of the pilot study interviews (48 paid participants) revealed a unanimous consensus that the much higher compensation level of this experiment compared to normal lab experiments was very attractive and motivated the participants to comply fully even with the indirectly remunerated portions of the experiment.

1.4 Methodology

All the regression equations and their associated variables used to conduct the analyses are based on the difference-in-difference-in-differences (DDD) estimator—as in CLK (2007 & 2009), Gruber (1994), etc.—and the classic difference-in-differences (DD) estimator. The following table describes all the variables from the primary regression equations. There are five indices in use: i per individual subject, r per individual round, c for category of good, g for individual good, and m for the type of market.

Table 2: Description of regression variables

Variable	Description
$Choice_{ircgm}$	*Choice* is the choice of quantity of each good selected in a round; it takes a pure quantity form, a revenue (monetary equivalent) form, and ratios of those forms with total quantity, total spend, and budget
TR_{ir}	*TR* stands for Tax Revenue; it is the amount of taxes paid/collected
α_{ir}	α is the basket-level intercept
PS_{ircg}	*PS* stands for Purchase Size; it is a dummy variable for the market type and price of a good (necessity = 0 vs. luxury = 1)
$Taxed_{ircgm}$	*Taxed* is a dummy variable for whether a good was taxed (=1) or not (=0) in a particular round
$Treat_{ir}$	*Treat* is a dummy variable for whether a round was control (VAT = 0) or treatment (RST = 1)
ε_{ircgm}, u_{ircgm}, υ_{ircgm}	ε, u, and υ are the error terms

The three main regression equations:

(1) Tax salience DD:

$$Choice_{icrgm} = \alpha_{ir} + \delta_1 Treat_{ir} + \delta_2 Taxed_{ircgm} + \delta_3 Treat_{ir} * Taxed_{ircgm} + \varepsilon_{ircgm}$$

(2) Purchase size effect DDD:

$$Choice_{icrgm} = \alpha_{ir} + \beta_1 Treat_{ir} + \beta_2 Taxed_{ircgm} + \beta_3 PS_{ircg} + \beta_4 PS_{ircg} * Taxed_{ircgm} + \beta_5 PS_{ircg} * Treat_{ir} + \beta_6 Taxed_{ircgm} * Treat_{ir} + \beta_7 PS_{ircg} * Taxed_{ircgm} * Treat_{ir} + u_{ircgm}$$

(3) Tax Revenue DD:

$$TR_{ir} = \alpha_{ir} + \gamma_1 Treat_{ir} + \gamma_2 PS_{ircg} + \gamma_3 PS_{ircg} * Treat_{ir} + \upsilon_{ircgm}$$

Choice consists of the four different output variables used by CLK. The first is the discrete quantity of goods (y) purchased. The second is the log version of y. The third is the revenue (or monetary equivalent) of the discrete quantity of goods (x) purchased—it is the product of the tax-inclusive price of a purchased good and the number of units chosen of that good. The fourth is the log version of x. As in CLK, the log versions offer the advantage of being better suited for comparing across goods/categories/baskets and marketplace types with naturally differing quantities sold, which is preferable in a comparison of shopping choice at distinctly different price levels. However, logs feature the distinct disadvantage of eliminating observations with zero quantity sold, which is a distinct advantage the laboratory experiment setting has over the field setting.

Therefore, three additional output variables were added that preserve the advantage without the said disadvantage: a set of ratios of choice out of total choice. As this paper examines how participants adjust their shopping choice ratio between taxed goods and untaxed goods by taxation type, these ratio variables specifically capture this crucial scale and make it comparable across type/price of goods. The first is the ratio of the pure discrete quantity of goods (y) divided by the total discrete quantity of goods purchased in a given round (y:TQ). The second is the monetary equivalent version of the previous ratio: revenue divided by the total amount spent in a given round (x:TS). Finally, the third is the x:TS ratio divided by the budget in a given round (x:TS:B). As the "all participant" results combine the x:TS ratios of the low- and high-budget participants and since the superstore market rounds have double the separated market budgets, this last variable provides a weighted average figure by participant budget level.

In the regression specifications above, the coefficients capture the change in shopping choice by tax type ($\widehat{\delta_1}$, $\widehat{\beta_1}$, & $\widehat{\gamma_1}$), tax-type-invariant purchase differences between taxed and untaxed goods ($\widehat{\delta_2}$ & $\widehat{\beta_2}$), and differences in shopping choice between the types/prices of goods ($\widehat{\beta_3}$ & $\widehat{\gamma_2}$). Interaction terms at the second-level control for the tax-type-invariant changes in purchase between taxed and untaxed goods by type/price of good ($\widehat{\beta_4}$ & $\widehat{\gamma_3}$), differences in choice between the types/prices of goods by type of taxation ($\widehat{\beta_5}$), and the good-invariant differences in the changes in shopping choice by tax type on goods that are taxed ($\widehat{\delta_3}$ & $\widehat{\beta_6}$). $\widehat{\beta_6}$ ultimately reveals the tax salience effect figure for the grocery goods as the use of the PS_{ircg} dummy variable controls out the difference. The third-level interaction term ($\widehat{\beta_7}$) captures the changes in shopping choices of taxed goods by type of tax and type/price of good, which is the main treatment effect of the experiment, PSE. $\widehat{\gamma_3}$ also captures a form of the PSE through the amount of taxes paid, *TR*. As $\widehat{\beta_6}$ captures the grocery good tax salience effect and $\widehat{\beta_7}$ reveals what effect the larger purchase size/type has on the tax salience effect, it is preferable for solid identification when both $\widehat{\beta_6}$ and $\widehat{\beta_7}$ are significant. The more negative the $\widehat{\beta_7}$ coefficient, the greater the PSE; or in terms of Reck (2016), the greater the debiasing.

1.5 Results

Before examining the PSE dimension, the existence of the tax salience effect must be verified. The results of the tax salience analyses (Appendices A7 and A8) indicate a strong and significant tax salience effect despite the conspicuous taxation cues throughout the experiment. Table 3 presents the findings from all the main shopping rounds for all participants all at once, herein dubbed the "overall general case".

Table 3: Tax Salience Effect (DD) Results of the Overall General Case

Choice	$\widehat{\delta_3}$	P-value	N	R-squared
Quantity (y)	0.609**	(0.006)	6144	0.021
Log of Quantity (log y)	0.0843**	(0.007)	4667	0.008
Revenue (x)	30.35**	(0.009)	6144	0.018
Log of Revenue (log x)	0.107^{+}	(0.058)	4667	0.004
Quantity/Total Quantity (y:TQ)	0.0564**	(0.006)	6144	0.066
Revenue/Total Spend (x:TS)	0.0629**	(0.008)	6144	0.038
Revenue/Total Spend/Budget (x:TS:B)	0.0000966*	(0.023)	6144	0.024

Notes: The first column lists the seven *Choice* variables as described in section 1.4 (y is the discrete quantity of goods, x is the revenue or monetary equivalent of the purchases, log y and log x are the logarithm versions of them, y:TQ is the discrete quantity of (taxed) goods of the total quantity of all goods purchased, x:TS is the total spend on (taxed) goods of the total spent in a given round, and x:TS:B is the x:TS of each round divided by the given budget the participant had in that round) and the second column lists the (increased) amount of each of those measures when facing the less-than-salient RST tax against the salient VAT tax.

The results displayed in Table 3 highlight the usefulness of the additional ratio variables even at the tax salience effect level. While the logarithm scale may provide superior figures for comparison across the quantity and price differences across the marketplace types, when comparing with the new choice variables, the log figures seem to be inflated from the loss of the 1477 observations that represent the choice of zero quantity of goods of a given type. R-squared figures also happen to be higher for the ratio variables than for the absolute and log versions. The overall general case showed an average increase of 0.609 more units of taxed goods purchased, equivalent to CZK 30.35 more spent on taxed goods under a less-than-salient tax (RST) than under a fully-salient tax (VAT). Log figures indicate an increase of 8.43% more units and a 10.70% increase in spending, though the last figure was only marginally significant. All three ratio variables were statistically significant and showed increases of 5.64% more units of taxed goods out of the total quantity chosen, 6.29% more spent on taxed goods out of the total spent, and a ratio of 0.0000966 (equates to an increase of 7.73%[31]) when the spend on taxed goods out of the total spent was weighted by participant budgets.

[31] This figure was found by multiplying the coefficient by 800, which is the average budget amount in the general case—from the averages of 533.33 for low- and 1066.67 for high-budget participants.

Compared to the CLK results as well as other literature from this branch, it is evident that the taxation info and cues reduced the tax salience effect compared with those found in experimental environments and natural experiments. Using the method described in CLK's "Strategy 1", the overall general case results in a ρ value of 0.61 and a θ of 0.50, which is considerably higher than the $\theta = 0.35$ that CLK (2007 & 2009) found. As purchase size is a modifier of tax salience, its effect is expectedly much smaller than the tax salience effect, so under a reduced tax salience effect, it is not surprising that it was only detected intermittently.

1.5.1 Main Analysis

1.5.1.1 Comparison of Means

As this experiment investigates PSE, which CLK (2007) theoretically predicted, and follows the CLK experimental approach (though in a lab experiment setting), this paper also presents the main findings per CLK. Their analysis begins with a simple comparison of means of the quantity sold (the variable dubbed "y" in section 1.4) at the basket level. The basket level herein refers to the full sum of all units (including zero units) of purchased goods divided into four categories: taxed vs. untaxed and grocery vs. department good.[32] Basket level analyses capture all individual-preference-based choice substitution across the basket when switching from taxed to untaxed goods.[33]

Before presenting the main results, two distinctions between the CLK comparison of means and the one presented below should be noted. First, Table 4 employs the ratio variable of quantity to total quantity (y:TQ) instead of only quantity (y), as in CLK. A direct result of the built-in-by-design price levels is that it is unclear whether or not the DDD coefficient is statistically significant for the absolute quantity difference (the third D) between the tax salience effect with grocery goods and department goods, simply because each unit of department good takes up a materially larger portion of the shopping budget. This is further described in Appendix A6.1. Therefore, a meaningful comparison of means requires the use of an outcome variable that resolves this issue. The most appropriate is the ratio of quantity to total quantity (y:TQ), as it relevantly compares the portion of the shopping basket that is allocated between taxed and untaxed goods between the two forms of taxation. The other difference in CLK's comparison of means is that their treatment was tax-inclusive (VAT = 1) pricing, whereas it is control here (VAT = 0).

---

[32] A basket, then, could have up to four different values: only taxed & untaxed quantities in the grocery and department rounds, but in the superstore round, those would also be broken down by market type.

[33] Nevertheless, comparative analyses were conducted at the good, category, category group,* and basket level (see Appendices A13 and A14). They further helped confirm that the basket level was most appropriate due to the considerably higher R-squared values of the regression outputs.

**Department goods are sorted into 9 loose categories of substitute goods. Grocery goods are sorted into 25 direct substitute categories as well as 9 looser categories. The "category" level is the aggregate of the 9+25 categories; "category group" level is the 9+9 looser categories.*

The results of the unabridged PSE analysis (available at the top of Appendix A10) confirm the existence of PSE along the ratio of quantity to total quantity amongst the superstore marketplace rounds with a statistically significant difference-in-difference-in-difference coefficient of -0.318, with a p-value of 0.024. However, an unexpected amount of noise was identified in the data resulting from participants accidentally clicking the checkout button before having meant to do so.[34] When removing these observations, the results become somewhat stronger and more significant as well as presumptively more in accordance with participant shopping choice intentions. Hence, the following comparison of means and all the analyses were conducted after removal. For completeness, the main and supplemental analyses were duplicated with those removed observations (see Appendices A7-A20). Comparing the unabridged and abridged output tables, there are no major differences in statistical output, effect direction, or interpretation.

The following comparison of means table displays the average portion sold of untaxed and taxed goods by row and the difference between the forms of taxation by column. For grocery goods (the upper panel), the average difference between taxed and untaxed goods proportions that constituted the shopping basket fell from 20.2% to 13.1% between the control and treatment periods. This means when taxes were fully salient (VAT), the taxed-to-untaxed ratio of average purchased portion of a basket was 28.9%:49.1% (a difference of 20.2%). When the tax was less-than-fully salient (RST), the same ratio was 33%:46.1% (a difference of 13.1%). A narrowed difference is exactly how one would identify the impact of PSE on tax salience effect. As the average proportional difference between the tax regimes for department goods was 18.3% and 14.2%, it is clear that the tax salience effect was smaller for department than grocery goods.

Below the comparison of means table are two useful visualizations of participant shopping data that further demonstrate the difference in the tax salience effect between grocery and department goods. Figure 1 plots the mean share of taxed goods purchased by treatment (tax type) and good type for all participants. Linear fit lines are generated to show the differences amongst the shopping choices. As the order of the participants is irrelevant, the directions of the lines mean nothing. More important is the area between the treatment and control fit lines, which visually identifies the tax salience effect. The discernably larger area between the grocery good choice and the department good choice fit lines illustrates the PSE. Figure 2 is a bar graph that displays the aggregated mean for all the participants for each good and treatment type as well as their differences. In this figure, the "diff" bars are the difference in the

[34] During the pilot trial of the experiment, a single participant noted during the pilot debriefing session that since the confirmation button is in the same position as the cart button, they once accidentally checked out before they had intended. The other pilot participants were asked about it, but no others experienced it. A fix to this small issue was explored, but the result was a surprisingly lengthy and expensive programming solution that would have broken binding budget and time constraints. As it was considered a minor issue that only 1 of 48 participants encountered, it was decided that it would be controlled for with an additional question in the programmed post-experimental survey. In the end, however, a much larger portion of participants fell victim to this issue than in the experimental trial. About 15% reported having accidentally checked out earlier than intended at least once.

aggregated mean between treatment and control, i.e. the tax salience effect. PSE is evident in the difference between the grocery and department "diff" bars. Appendix A6 provides equivalent figures for low- and high-budget groups as well as additional graphic figures that depict shopping speed, budget usage, and correlated variable distributions.

Table 4 — Purchase Size Effect of Tax-Inclusive versus Tax-Exclusive Prices: DDD Analysis of Mean Quantity Ratio (Quantity Taxed/Total Quantity Sold)

Goods	Control (VAT)	Treatment (RST)	Difference
Panel A. Grocery Goods			
Goods with no Tax (0%)	0.4906	0.4607	-0.0299
	(0.0079)	(0.0089)	(0.0115)
	[648]	[648]	[1296]
Goods with Tax (21%)	0.2885	0.3300	0.0415
	(0.0065)	(0.0072)	(0.0090)
	[648]	[648]	[1296]
Difference	-0.2022	-0.1307	$\boldsymbol{DD_{GG}}$ **= 0.0715**
	(0.0071)	(0.0123)	(0.0149)
	[1296]	[1296]	[2592]
Panel B. Department Goods			
Goods with no Tax (0%)	0.4211	0.3969	-0.0242
	(0.0082)	(0.0094)	(0.0089)
	[648]	[648]	[1296]
Goods with Tax (21%)	0.2380	0.2553	0.0173
	(0.0075)	(0.0121)	(0.0139)
	[648]	[648]	[1296]
Difference	-0.1831	-0.1416	$\boldsymbol{DD_{DG}}$ **= 0.0415**
	(0.0146)	(0.0106)	(0.0176)
	[1296]	[1296]	[2592]
DDD Estimate			**-0.0300**
			(0.0114)
			[5184]

Notes: Each cell displays the mean quantity of untaxed/taxed goods out of the total quantity of goods sold per round, under the two forms of taxation: VAT/RST. The standard errors related to the point estimates are in parentheses and the number of observations (each round has 2 – 4 basket-level observations per respective marketplace).

Figure 1: Mean Share of Taxed Goods by Treatment and Good Type

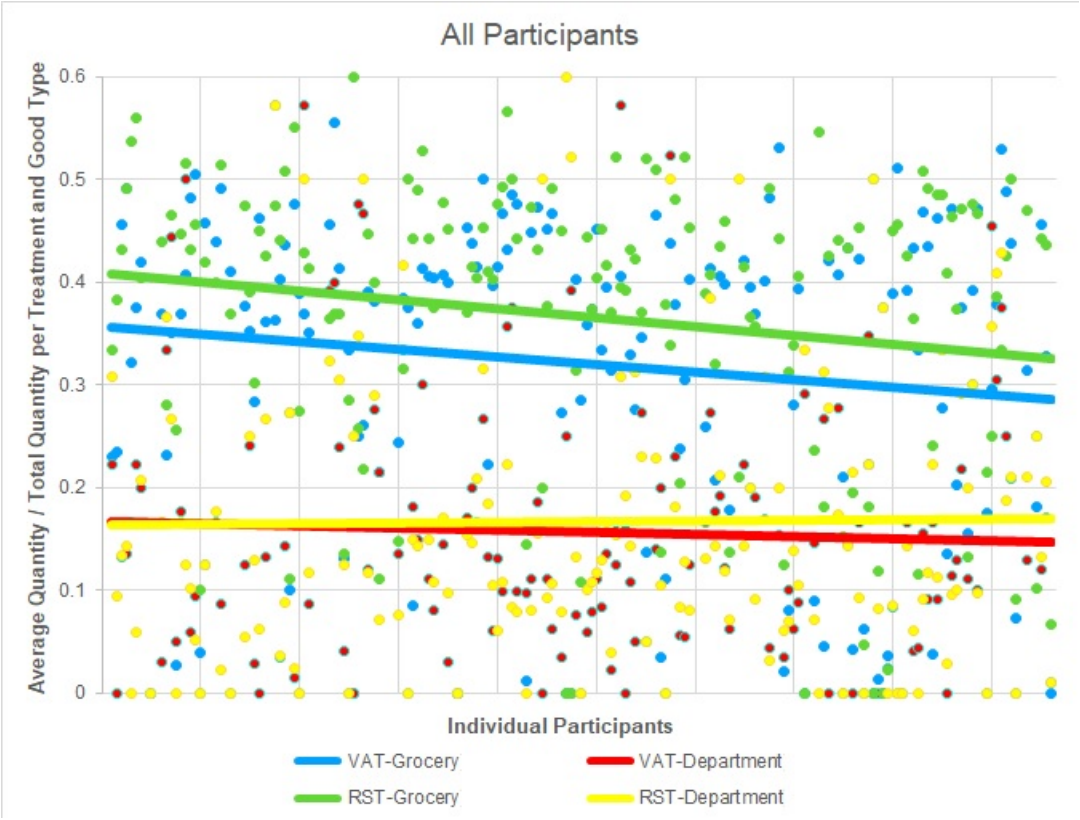
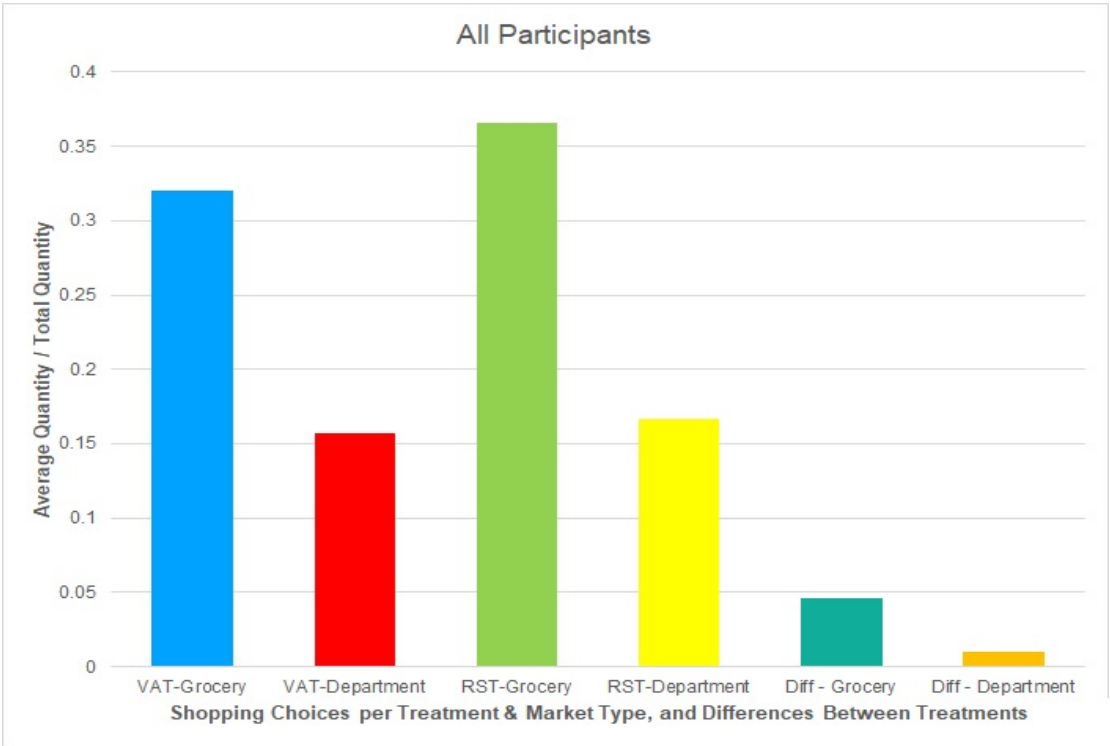


Figure 2: Aggregated Mean Share by Treatment and Good Type and Their Differences



1.5.1.2 Purchase Size Effect Results and Discussion

Table 4 shows that the difference between these differences reveals a PSE significantly occurring (with a p-value of 0.024) in the amount of -0.0300 for all marketplace- and budget-types combined (the overall general case). This DDD estimate and many more that constitute the main analysis results, broken down by output variable and participant type, are available in the first table of Appendix A9. The DDD estimate from Table 4 signifies that there is a 3% reduction in the tax salience effect in the average shopping basket for department goods versus grocery goods. This result successfully rejects the null hypothesis that agents exhibit the tax salience effect equally between low and high price-level goods. This result confirms a PSE occurring in the CLK expected direction—participants (consumers) paid more attention to less-than-salient taxes of more expensive goods.

While a 3% difference in shopping behavior between the more and less expensive goods seems to be a concrete representation of PSE on tax salience effect, it actually remains abstract until its impact on θ is assessed. Following the methods from CLK's "Strategy 1", a value of the normalized tax visibility effect, ρ,[35] was calculated for both types of goods—specifically, ρ_{gg} = 0.71 and ρ_{dg} = 0.37—and then was divided by the elasticity of each type of good to find the individual thetas. Using all the shopping data from all the rounds from the entire experiment that had tax-inclusive pricing, the elasticity regressions of grocery and department store goods were found: $\varepsilon_{gg,p}$ = 0.641 and $\varepsilon_{dg,p}$ = 0.899; which correspond to some estimated ranges from the literature on elasticity by type of good. The final result was θ_{gg} = 0.46 and θ_{dg} = 0.70. These thetas show a considerable diminishing of the tax salience effect between the less expensive grocery goods and the more expensive department goods, which clearly indicates the existence, direction, and consequential implication of a purchase size effect.

If it were possible to extrapolate linearly using the fact that there was a 9.13:1 ratio of average price between the marketplaces, the DDD estimate from Table 4 would indicate that goods with a pre-tax price of CZK 6653 (at the time of the experiment, that amount exchanged to about USD 266; using OECD estimated PPP,[36] that equates to about USD 451) under a retail sales tax system with a 21% tax rate should no longer exhibit any tax salience effect. Despite the high tax rate, an educated estimate would find it unlikely that such an amount would cause an absolute elimination of the tax salience effect. If that estimation is correct, then it would mean that the PSE is unlikely to be linear. Of course, purchase size is certainly not the only element that influences tax salience. Several other influences, including heterogeneity, are explored in section 1.5.2 and Appendix A6.

---

[35] From CLK (for taxed goods): ρ = -(log quantity(VAT) - log quantity(RST))/log(1+t)
[36] Source: https://en.wikipedia.org/wiki/Purchasing_power_parity#OECD_comparative_price_levels

1.5.1.3 Tax Incidence and Revenue Discussion

This section analyzes the differences in tax incidence between low- and high-budget participants as well as tax revenue in the experiment. As mentioned in the introduction, there is an a priori stakes effect likely at play that, *ceteris paribus*, incentivizes high-budget participants to pay greater attention to their shopping choices than their counterparts. However, there is also another effect likely at play that increases attention through more constraining budgets. In the first table in Appendix A9, the results from the main analysis show that while high-budget participants were predominantly exhibiting considerably stronger and more significant tax salience effects in both the grocery goods case ($\widehat{\beta_6}$) and in the general case[37] ($\widehat{\delta_3}$ from the first table in Appendix A8), the opposite is true for low-budget participants with regard to the differences ($\widehat{\beta_7}$) in the tax salience effect between the grocery goods and department goods. Simply put, the main analysis shows that tax salience results are considerably more significant for high-budget participants than low-budget participants. In addition, low-budget participants exhibit several more significant PSE difference estimates than high-budget participants, which implies that those with more binding budget constraints begin to account for less-than-salient taxes at lower price points than their counterparts. In other words, those with less to spend (i.e. are more budget constrained) are probably not making as many tax salience errors and probably even fewer with more expensive goods (i.e. PSE), despite any opposing influence from the higher stakes effect. This indicates that PSE is dominated by the former effect, which implies that the tax salience effect may be naturally reduced by tighter budget constraints (i.e. less income/wealth).

The effect on tax revenue (Appendices A11 and A12) provides further insight. Again, the built-in budget difference issue precludes sole reliance on a direct comparison of pure tax revenue figures between the two marketplaces. With tax revenue, no ratio is possible, and thus the natural log of tax revenue is relied on to best elucidate the difference from PSE. Table one in Appendix A11, which displays the general results, reveals a reduction in the log of tax revenue for higher priced goods. This difference is larger and more significant for low-budget participants, which means that the tax revenue collected was reduced by PSE more so for low-budget participants.

Altogether, this may represent a small, yet important point of clarification into the CLK bounded rationality model in that it elaborates on the mechanism of how purchase size impacts tax salience through relative budget constraints. Based on the evidence above, when consumers find themselves under a tighter budget constraint (perhaps after disregarding a less-than-salient tax) they are inclined to pay more attention to less-than-salient taxes (debias), which suggests a lower marginal cognitive cost during such purchases. If this is a consistent mechanism, then the crucial dynamic of budget adjustment in the CLK model would lead consumers to intrinsically reduce future purchases of both untaxed and taxed goods, thus

[37] Huseynov et al. (2019) also find that high-budget participants display a greater tax salience effect.

avoiding the most negative impacts of the distortionary income effects and leading to the strictly positive-to-welfare outcome CLK predict.

The vast majority of the results of the supplementary analyses are also driven by the higher statistical significance and greater magnitude effects of high-budget participants. While not a direct conclusion from the data, such results suggest that the debiasing distribution of the low-budget participants is thicker at lower price levels and may even be first order stochastically dominant over the debiasing distribution of the high-budget participants, which would further signify the positive-to-welfare outcome. If this extrapolates to real-world budget constraints,[38] then this would imply the potential for an inherently progressive tax system with significantly reduced excess burden from tax distortion.

1.5.2 Supplemental Analyses

Beyond the main analysis, several sub-analyses (i.e. by taxed good counterparts, substitute good categories, and continuous version of choice) as well as survey-answer-based dissection, placebo, robustness check, and heterogeneity analyses were conducted. The dissections tackle some of the elements that possibly cause or contribute to the tax salience effect—and alter PSE—such as participant characteristics and abilities, their naiveté/sophistication, their internal states/motivations, environmental issues, etc. The most interesting ones are summarized herein. It should be noted that many of the dissections have a less rigorous foundation and should be understood as more informative than definitive. For detailed explanations of their rationale, results, and discussions, see Appendix A6.3. Corresponding output tables are presented in Appendices A7-A26.

Initial shopping speed results pointed to a significant correlation. Thus, an in-depth examination that featured complete analyses of three forms of speed (faster speed in a given shopping round = "faster shopping"; faster speed for the total of all rounds for a given participant = "faster shoppers"; and faster completion of the experiment = "faster participants") were conducted for both slower and faster groups in each dimension as well as a heterogeneity analysis and an analysis that broke down shopping speed by type of marketplace (a "speed-type" categorization analysis). All three forms of speed are correlated with large increases in the tax salience effect, especially for high-budget participants. Since participant learning is expected to lead to later shopping rounds that are faster and with similar or more optimal shopping behavior, that should result in a lower tax salience effect for "faster shopping" than "faster shoppers". Instead, the revealed pattern shows the opposite, indicating that shopping speed is likely driving the tax salience effect. Moreover, the categorized analysis results reveal that the greatest increases of the tax salience effect in relation to speed are from department market goods, signifying that shopping speed may impact tax salience more than purchase size.

---

[38] Goldin & Homonoff (2013) find evidence of this in their empirical analysis of tax salience for cigarette sales, where only low-income consumers reacted to less-than-salient (RST) taxes.

Alternatively, there is an implication of a correlation with a loss or lack of learning, which, if this resembles the real world in such a persistent manner, could mean that consumers may not learn to or would "forget" to pay attention to less-than-salience taxes over time. At the same time, the categorized analysis also indicates that there are participants shopping faster in grocery rounds and slower in department rounds strategically to avoid taxes at higher stakes, reflecting cognitive PSE behavior and suggesting a rationally inattentive choice of shopping speed. As there is insufficient evidence herein to disentangle these interpretations, such work remains an avenue for future research.

The results of comprehension analyses (based on participant answers to the comprehension quiz at the end of the instruction section), familiarity (based on the correlation between local nationality and familiarity with the tax rate), and participant-subjective rationale (based on participant answers to if and why they changed any shopping choices between rounds) indicate that ignoring less-than-salient taxes, especially for grocery goods, seems to be intentional and not due to a lack of understanding. This lends further credence to CLK's bounded rationality model. The crux seems to be a shopper choosing how much attention to allocate to taxes given the stakes versus given how constrained they are by their budget, which directly impacts speed and tax salience.

The results of the Cognitive Reflection Test (CRT) analysis (based on how many of the three CRT questions the participants answered correctly) show that participants who display more cognitive reflection ability exhibit a considerable increase in the tax salience effect with grocery goods and a moderate increase in PSE. The evidence indicates that they are rationally choosing a cognitive cost (debiasing) threshold between the grocery and department price levels. The arithmetic ability analysis (based on how many of the three arithmetic questions, which were reproductions of the tax calculations from the experiment, the participants answered correctly) and the combined CRT-with-arithmetic-ability analysis indicate that those with greater such abilities choose not to pay attention to taxation even though these participants have a theoretically lower cognitive cost. This again implies that the reason is not opportunity cost based, but strategic, which again supports the CLK bounded rationality model.

Budget analyses (based on budget usage in relation to budget constraint as well as binding budget constraints from changes in which goods were taxed) imply that actual budget usage is not consequentially related to tax salience and purchase size effects other than through budget constraints. The budget constraints imposed by the experimental design do not somehow falsely create tax salience and purchase size effects. Rather, they appear to lead most low-budget participants to strategically maximize their experimental endowment, decreasing tax salience effect and ambiguously affecting PSE results. In total, these analyses reinforce the main findings herein (which were mostly driven by high-budget participants), uncover the existing influence of the stakes effect, and inform future experimental designs in this area.

The interaction between speed and budget usage analysis, while again reinforcing the findings already discussed, does potentially provide another glimpse into how the two opposing effects

on strategic (in)attention (stakes versus budget constraints) may be operating. Low-budget participants that shopped faster and used more of their budget appear to be maximizing both spending and tax avoidance, while their slower shopper counterparts are only maximizing spending and ignore taxes. Faster shopping, high-budget participants that used less of their budget have increased PSE, which may imply the dominance of budget constraints over utility stakes on attention.

Several additional analyses were conducted, but the results neither point to any further triggers of attentiveness as those above nor provide much further insight into tax salience, PSE, or this experiment. These analyses are available in Appendix sections A6.3.7-A6.3.10 (they included analyses of goods preferences, shopping familiarity, hunger, substitute goods, and robustness and placebo checks). It is noteworthy that the goods preferences analyses (Appendix A6.3.7) show that some substitution effect is occurring from taxed to untaxed goods in both grocery and department marketplaces, but to a lesser extent in the former. This is probably because a larger number of items could be bought in the grocery marketplace. Another is that the downplaying of brands from the experiment seems to decrease the opportunity costs of substitution and thus slightly lower tax salience effect and increase PSE. Furthermore, it should be noted that the result of the placebo and "AB" analyses (Appendix A6.3.10) support the methods and findings of this paper.

Finally, the design of this experiment includes a programmed survey and questionnaire to analyze related dynamics of tax salience, including how differences along dimensions of consumer heterogeneity impact tax salience and PSE, which is intended for a follow-up experiment. While the heterogeneity of the participants of that experiment would be crucial for a proper analysis, a set of analyses were nevertheless run across dimensions of heterogeneity from this experiment (Appendix A.6.4.). The results indicate that men and women do not behave differently with respect to tax salience and there is mixed evidence related to productivity (proxied by education and arithmetic ability), income, and wealth measures. Education results suggest a regressivity in the tax instrument, while income and wealth measures suggest progressivity. The latter evidence again points to an innate progressivity of this tax instrument, which motivates future research.

1.6 Conclusion

The results herein reject the null hypothesis that agents are susceptible to tax salience equally between low and high purchase level goods, revealing a purchase size effect. On average, it amounts to a decrease in the tax salience effect of 3% from grocery to department goods, which are 9.13 times more expensive on average. Combined with the elasticities of the different types of goods, this purchase size effect results in a grocery good θ_{gg} of 0.46 and a department good θ_{dg} of 0.70. This corroborates CLK's bounded rationality model and prediction that agents will more likely pay the cognitive cost of tax salience as price increases. Furthermore, this work

replicates the main findings of the tax salience literature despite the intensity and proximity of information and cues about taxation presented in the instructions as well as throughout the quasi-field environment of this experiment (which would be much weaker in the real world and, thus, cause the θ values presented herein to be considered upper bounds, i.e. lower bounds of the tax salience effect), finding an overall tax salience effect of $\theta = 0.50$ (CLK found $\theta = 0.35$ in their experiment and $\theta = 0.06$ in their empirical research).

The dichotomized budget endowment also provided some insights into opposing forces at work within the tax salience experimental setting. As potential payoff stakes in a given economical experiment increase, a participant is more likely to take the activity more seriously. In the case of this experiment, this effect would increase participant attention on their purchasing behavior, which would be in opposition to salience effects. Nevertheless, the results herein reveal that the tax salience effect is considerably more significant for high-budget participants than low-budget participants. Moreover, low-budget participants have more significant PSE results than high-budget participants. Together, this implies that more binding budget constraints lead participants to pay more attention to less-than-salient taxes and make fewer tax salience errors and even more so as the price level increases. This may represent a point of clarification for the CLK bounded rationality model indicating an innate budget adjustment mechanism of both taxed and untaxed goods, leading to CLK's strictly positive-to-welfare outcome. The results of the tax revenue and supplemental analyses further collaborate this finding. If this extrapolates to the real world, it would mean that the tax salience effect is inherently reduced by less income/wealth and less-than-salient taxes would be intrinsically progressive while causing considerably lower deadweight loss than salient instruments.

The dissections into the possible causes of and contributors to tax salience effect reveal some insights of an informative nature. In particular, participants who understood better, or were more "rational" (per CRT results), or more proficient arithmetically all exhibited greater tax salience effect. Additionally, the discovery that the tax salience effect behavior of disregarding some or all taxes that are not posted within a price is presumptively intentional and strategic lends even greater credence to the bounded rationality model. The results of the extensive shopping speed analysis show that the tax salience effect is most driven by department market speeds, implying that it is probably the shopping speed itself that causes the tax salience effect rather than vice versa, as speed appears to be a rationally inattentive choice to avoid paying the cognitive cost even at the higher price level. Alternatively, there is an implication of a correlation with a lack or loss of learning, which, if this corresponds to the real world in a persistent manner, could mean that tax salience behavior may remain dominant in the long term. The results of the heterogeneity analysis uncover a lack of difference in the tax salience effect between genders and indicate tax instrument progressivity along income and wealth dimensions. However, considering that heterogeneity was not the main focus of this experiment and that actual subject heterogeneity was not as varying as would be desired, this additional progressivity evidence is ambiguous, but should impel further research.

This paper introduces a novel experimental design to study the purchase size effect, an element of tax salience not focused on in earlier literature. It is also the first experiment in this branch that divided participants into low- and high-budget groups to ascertain how budget constraints may correlate with tax salience (and PSE) and with the many tax-salience-related dissections herein. Regarding improvements in methodology, it is the first experiment to incorporate a goods selection methodology that reduces noise by improving participant choice mapping while increasing the power of the test. It also offers the stronger randomization and greater degree of control only offered by a lab experiment, but with enhanced external validity thanks to the use of pseudo-realistic settings and actual subject-choice remuneration.

Several insights from the supplemental analyses may motivate research that would further elucidate the forces at play in the salience phenomenon, and the budget usage analysis could help inform the designs of such future research. For example, a similar style experiment, with large budgets, that concentrates on shopping speed by varying the timing constraints, perhaps attempting to elicit the value that participants would place on loosening such time constraints, could reveal some useful insights that may translate into welfare enhancing policies and/or business practices. In addition, the link between speed and CRT, and possibly arithmetic ability, could be explored by incorporating Rubinstein's Contemplative Index into a future experiment. Furthermore, the rejection of the null hypothesis, despite conspicuous taxation cues, may motivate a broader scope of research into the purchase size effect in a much wider range of price levels, perhaps through a natural experiment. Implications of shopping speed and progressivity by heterogeneity should incite additional research into other dimensions of tax salience, which may be crucial in assessing how a less-than-salient tax tool affects social welfare and who may be most impacted and how. Optimistically, these findings could help form the foundation of a practical set of guidelines that would shape future tax policy.

References

Blumkin, T., Ruffle, B. J., & Ganun, Y. (2012). Are income and consumption taxes ever really equivalent? Evidence from a real-effort experiment with real goods. *European Economic Review*, 56(6), 1200-1219.

Chetty, R., Looney, A., & Kroft, K. (2007). Salience and taxation: Theory and evidence (No. w13330). *National Bureau of Economic Research.*

—Later merged with Chetty (2009) and published: (2009) *American Economic Review, 99(4), 1145–77.*

Chetty, R. (2009). The simple economics of salience and taxation (No. w15246). *National Bureau of Economic Research.*

Chuang, S. H. (2019). Salient or Not? The US Air Travel Taxes. *The US Air Travel Taxes (April 26, 2019).*

Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of personality and social psychology*, 27(1), 100.

Feldman, N., Goldin, J., & Homonoff, T. (2018). Raising the stakes: Experimental evidence on the endogeneity of taxpayer mistakes. *National Tax Journal*, 71(2), 201-230.

Figlio, D. N., & Rueben, K. S. (2001). Tax limits and the qualifications of new teachers. *Journal of Public Economics*, 80(1), 49-71.

Finkelstein, A. (2009). E-ztax: Tax salience and tax rates. *The Quarterly Journal of Economics*, 124(3), 969-1010.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25-42.

Gabaix, X., & Laibson, D. (2006). Shrouded attributes, consumer myopia, and information suppression in competitive markets. *The Quarterly Journal of Economics*, 121(2), 505-540.

Gamage, D., & Shanske, D. (2010, November). The Case for Reducing the Market Salience of Taxation. In *National Tax Association, Proceedings of the 103rd Annual Conference*.

Gamage, D., & Shanske, D. (2011). Three Essays on Tax Salience: Market Salience and Political Salience. *Tax L. Rev.*, 65, 19.

Goldin, J. (2015). Optimal Tax Salience. *Journal of Public Economics*, 131, 115-123.

Goldin, J., & Homonoff, T. (2013). Smoke gets in your eyes: Cigarette tax salience and regressivity. *American Economic Journal: Economic Policy*, 5(1), 302-336.

Gruber, J., & Poterba, J. (1994). Tax incentives and the decision to purchase health insurance: Evidence from the self-employed. *The Quarterly Journal of Economics*, 109(3), 701-733.

Hendricks, M. D. (2014). Does it pay to pay teachers more? Evidence from Texas. *Journal of Public Economics*, 109, 50-63.

Horowitz, J. K. (2006). The Becker-DeGroot-Marschak mechanism is not necessarily incentive compatible, even for non-random goods. *Economics Letters*, 93(1), 6-11.

Hossain, T., & Morgan, J. (2006). ... plus shipping and handling: Revenue (non) equivalence in field experiments on ebay. *The BE Journal of Economic Analysis & Policy*, 6(2).

Huseynov, S., Palma, M. A., & Segovia, M. (2019). Distributional Effects of Price Salience on Reservation Wages and Food Choices.

Hyunjin, Yun. (2019). Rising Salience of Sales Tax with Its Rate [Doctoral dissertation, *Department of Economics, The Graduate School Seoul National University*].

Karni, E., & Safra, Z. (1987). "Preference reversal" and the observability of preferences by experimental methods. *Econometrica: Journal of the Econometric Society*, 675-685.

Reck, D. H. (2016). Taxes and Mistakes: What's in a Sufficient Statistic?. *Available at SSRN 2268617.*

Rubinstein, A. (2016). A typology of players: Between instinctive and contemplative. *The Quarterly Journal of Economics*, 131(2), 859-890.

Taubinsky, D., & Rees-Jones, A. (2018). Attention variation and welfare: theory and evidence from a tax salience experiment. *The Review of Economic Studies*, 85(4), 2462-2496.

Tiezzi, S., & Verde, S. F. (2016). Differential demand response to gasoline taxes and gasoline prices in the US. *Resource and energy economics*, 44, 71-91.

Zheng, H., Huang, L., & Ross Jr, W. (2019). Reducing Obesity by Taxing Soft Drinks: Tax Salience and Firms' Strategic Responses. *Journal of Public Policy & Marketing*, 38(3), 297-315.

Appendix

The first six chapters of the appendix (below) cover the instructions, experimental details, and in-depth supplemental analyses and discussions. The remaining chapters present the regression outputs and are available for perusal and download through the following link.

https://www.dropbox.com/s/uxyp7bja6tqn6io/MSMS_PSE_Appendices.pdf?dl=0

APPENDIX A1: Experimental Instructions

0. Oral instructions (read prior to starting and short notes written on board)

Now that we are all here, I will present a few initial notes and instructions. After that, everyone will push button F5 and we'll begin the experiment.

Please note that you will have to stay in the lab until everyone completes the experiment, so please take your time to understand the instructions and with the experiment itself.

Please raise your hand, and keep it raised, at any time if you have any questions as well as once you finish the experiment (which concludes on the "Earnings" page). Please do not be shy if you have any trouble understanding any part of the instructions, even if you are simply unsure about a single word, please raise your hand, as it is important that you understand the instructions. And please note that you will be given a short questionnaire to be written by hand at the very end of the experiment.

The experiment is about individual decision-making and is in the form of an e-shop. Your earnings will be made up of real goods and cash, collected in a week from today in this very room between 12:00 and 14:30. Please remember that all grocery goods will be freshly delivered that day! Also, please know that the prices you will face are real market prices taken directly from the supplier websites *(Tesco & Alza)*.

And please do not push the browser's back button, backspace, or the back button on the side of the mouse. If you do and something strange happens with the experiment, such as you find yourself back at goods selection, please do nothing and just raise your hand!

Finally, you will find a stack of papers on your desk. It has a small piece of paper, a blank (and lined) sheet for your personal notes, and a printed version of the online instructions on your desk. You may refer to them throughout the experiment.

Please pick up the small piece of paper on your stack. Now you may press F5. The registration number on the left side of the screen is yours *(you may ignore the right side of the screen)*. As soon as I finish these instructions, please create your own password, push the "new registration" button, and write down your registration number and the password you created on the provided piece of paper and/or into your phone/tablet/etc. You will need this information to collect your earnings. As soon as that is complete, all electronic devices must be put away for the remainder of this experiment. You may use them again once you have completed the experiment and final questionnaire, while you are waiting for the other participants to finish.

Thank you for your attention. Please begin.

1. Introduction

This is an experiment designed to imitate a common online or in-store shopping environment. A research foundation has provided funds for conducting this research. Your earnings will depend solely on your decisions in this experiment and will not depend on the decisions of any other participants. Nonetheless, please do not communicate with any other participant during the experiment.

These instructions are in 5 sections. After this introduction section, section 2 describes the shopping environment and section 3 provides details about your task, earnings, and a summary of main points. Section 4 presents a set of screenshots that explain how to navigate the e-shop and section 5 is a short comprehension questionnaire about these instructions.

The experiment is made up of independent shopping decisions in 3 types of marketplaces with slightly different environments. After these instructions and before proceeding to the marketplaces, you will be asked to select a set of goods from a larger set. These goods will then make up the available goods in the marketplaces you will face throughout the remainder of the experiment as well as your actual earnings.

After that you will begin the main portion of the experiment. It will be made up of 5 periods consisting of 3 rounds each, for a total of 15 shopping rounds. Each round will have its own budget and time limit. After you complete the shopping rounds, you will fill out a survey, and then the computer will select a single round at random for your earnings. The real goods purchased during that chosen round will be purchased and ready for you to collect in 7 days, or as otherwise instructed, in an announced location.

As this experiment is conducted at the individual level, every participant will complete the experiment at his or her own pace. The expected average individual experimental time including these instructions is about 50 minutes. However, laboratory rules dictate that no participant is allowed to leave until all participants finish. Therefore, this experimental session is expected to take about 75 minutes.

PLEASE DO NOT PUSH THE BROWSER'S BACK BUTTON throughout the experiment. If you do, you may have to repeat that given round/portion.

2. Shopping Environment

Please engage with the shopping environment as you normally would when making standard shopping decisions, as your earnings will be the direct combination of your shopping choices and a portion of your unused budget.

In each round, you will have a shopping budget. You may consider it as a standard amount that you have allocated from your income towards an average shopping activity or simply the cash you have in your pocket when going to the store. Each round will have its own budget. It does not carry over or accumulate in any fashion. There is no borrowing or credit between rounds. And you will not be able to add any goods into your shopping cart that surpass the budget.

You will shop in 3 types of marketplaces that involve repetitive, independent, and identical shopping decisions. Each period is made up of 3 rounds, one of each type of marketplace: a grocery store style round made up of perishable food goods; a department store style round made up of electronics, cosmetics, and small luxury items; and a superstore style round—a full combination of grocery and department markets, including their budgets and time limits. An exception is that the last 3 rounds will be made up of just superstore style market rounds.

Each round has a shopping page and a cart/checkout page. You may alternate between the pages with the click of a button. Each round will also have time limits (for example, 165 seconds for shopping and 45 seconds to checkout). If you run out of shopping time, you will not be able to return to the shopping page. If you run out of checkout time, you will not be able to return to the cart page. If you run out of time in both, the items in your shopping cart at that moment will be automatically purchased.

In the experiment you will go through two market regimes that differ in form of consumption tax. You may consider the two forms as marketplaces in two different countries (i.e. Czech Republic and USA).

In one market regime, some of the available goods are taxed under a Value Added Tax system, where the taxes are included in the price of the goods. In the other market, some of the available goods are taxed under a Retail Sales Tax system, where the taxes are not included in the price of the goods, but are added at checkout. Specifically:

Value Added Tax (VAT) is a consumption tax added to a product's sales price. For example, imagine that a bag of coffee has a cost of 100Kc and is subject to a 21% VAT tax. The coffee would have a price tag of 121Kc and you would pay a final price of 121Kc at check out.

Retail Sales Tax (RST) is a consumption tax added to a product's final price at checkout. Now imagine that a bag of coffee has a cost of 100Kc and is subject to a 21% RST tax. The coffee would have a price tag of 100Kc and you would pay a final price of 121Kc at check out.

Please note that for any good that is taxed, the tax amount and final sales prices under both tax systems are identical. Also, before each period and last 3 rounds, you will be shown a complete list of the goods that will be taxed throughout that upcoming period.

3. Description of the Task

After this instruction section, you will choose the goods that make up the 3 marketplaces. Specifically, you will choose 10 goods out of an available 50 goods from the grocery market and 10 goods out of an available 50 goods from the department market. The 20 chosen goods will automatically make up the superstore selection. These goods will make up the available goods in the marketplaces you will face throughout the remainder of the experiment.

A major part of your final earnings will be your actually purchased basket of goods, so it is important that your good selection be based upon your personal preferences. After the goods selection portion, your task will simply be to shop.

For each shopping round, please choose the quantity (minimum of 0, maximum limited by the round budget) of each type of good that you wish to purchase, proceed to the cart, and complete your purchases for that round.

PLEASE BE CAREFUL NOT TO ACCIDENTALLY CLICK THE "PROCEED TO CHECKOUT" BUTTON ON THE CART PAGE as that concludes the round. You may go back and forth between the shopping and cart pages as long as your time limit there has not run out.

Description of the Earnings

Your earnings will be the real goods purchased and 50% of any unspent budget (in cash) from a single round chosen at random by the computer at the end of the experiment. The computer will notify you which round was chosen for your earnings. Your earnings, made up of both cash and goods, will be collected in 7 days or per other oral instructions/agreements. All grocery goods will be freshly delivered on collection day.

Please note that only one round will be chosen at random for your earnings. Therefore, the shopping decisions made during each separate round are individually important and equally likely to result in ALL your actual earnings.

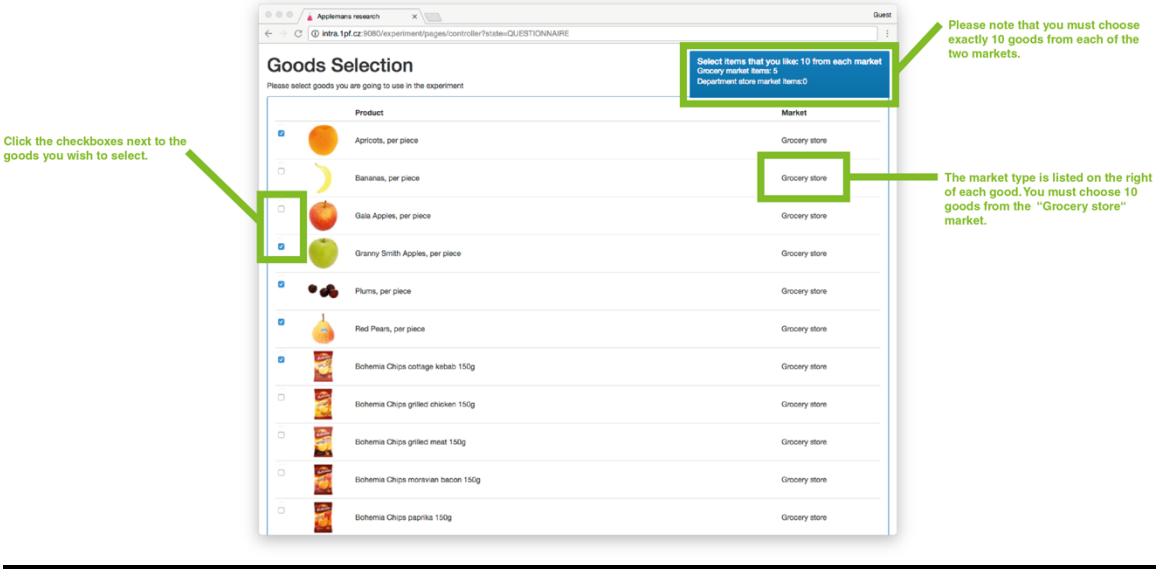
Example: At the end of the experiment, the computer randomly chooses a round with an 800Kc budget. In that round, the participant purchased several goods for a total price of 700Kc. The final earnings amount is the actual goods selected and 50Kc (50% of 100Kc, which is the remaining budget amount).

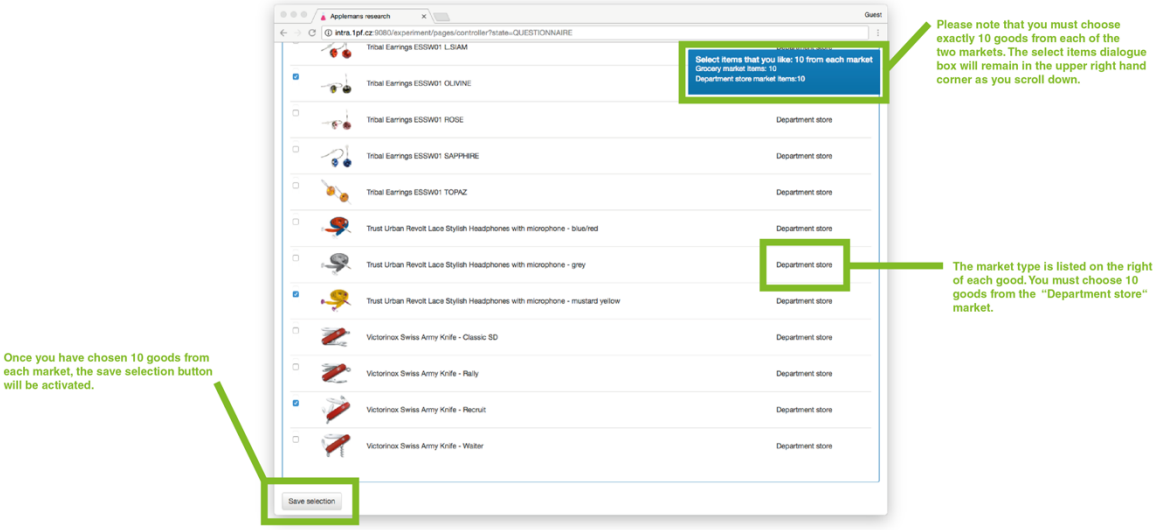
Summary of Main Points

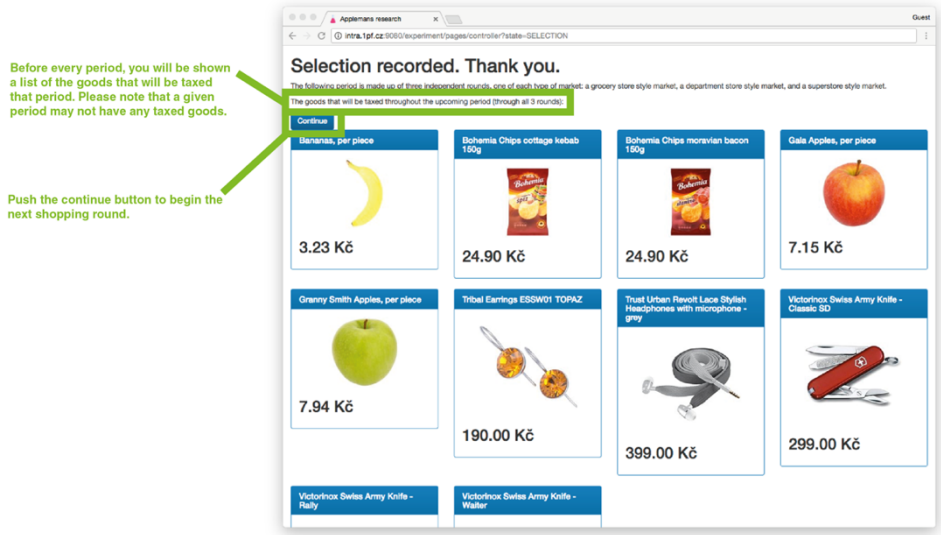
- Total of 15 shopping rounds: the first 4 periods have 3 rounds each, one of each type of marketplace, and the last 3 rounds are all superstore style rounds
- You will choose the goods that make up the 3 marketplaces (10 out of 50 grocery goods and 10 out of 50 department goods; those 20 goods make up the superstore)
- Each round will have an independent budget and time limit
- The superstore rounds are a full combination of the goods, time limits, and budgets of the grocery store and department store rounds
- If you run out of time, the items in your cart will be automatically purchased
- Your earnings will be the real goods purchased and 50% of any unspent budget from a single round chosen at random by the computer; to be collected in 7 days; all grocery goods will be freshly delivered on collection day

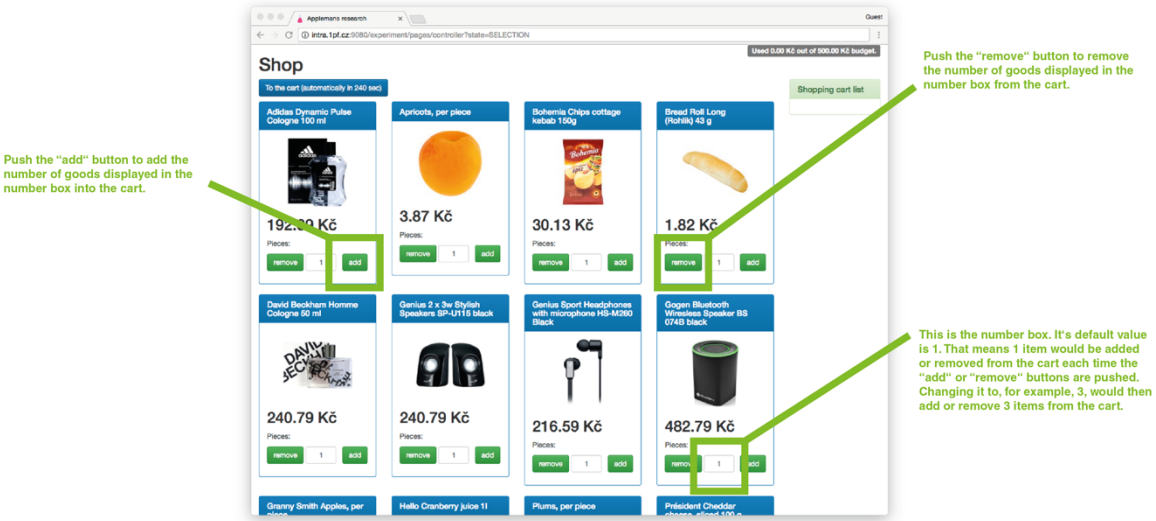
- Only one chosen round means that the shopping decisions made during each separate round are important and equally likely to result in ALL your earnings
- PLEASE DO NOT PUSH THE BROWSER'S BACK BUTTON
- PLEASE BE CAREFUL NOT TO ACCIDENTALLY CLICK THE "PROCEED TO CHECKOUT" BUTTON

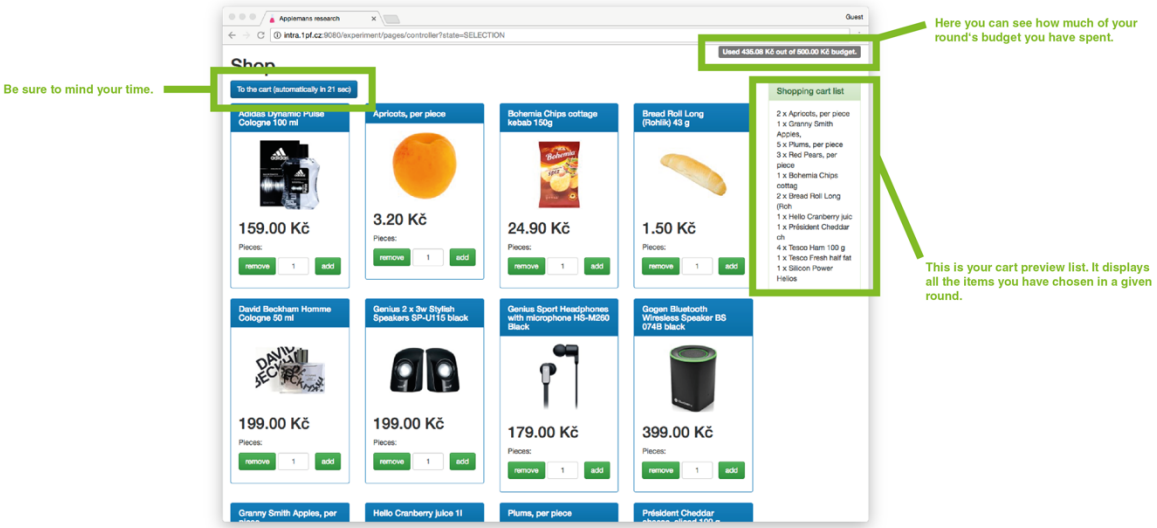
4. Shopping Environment Navigation

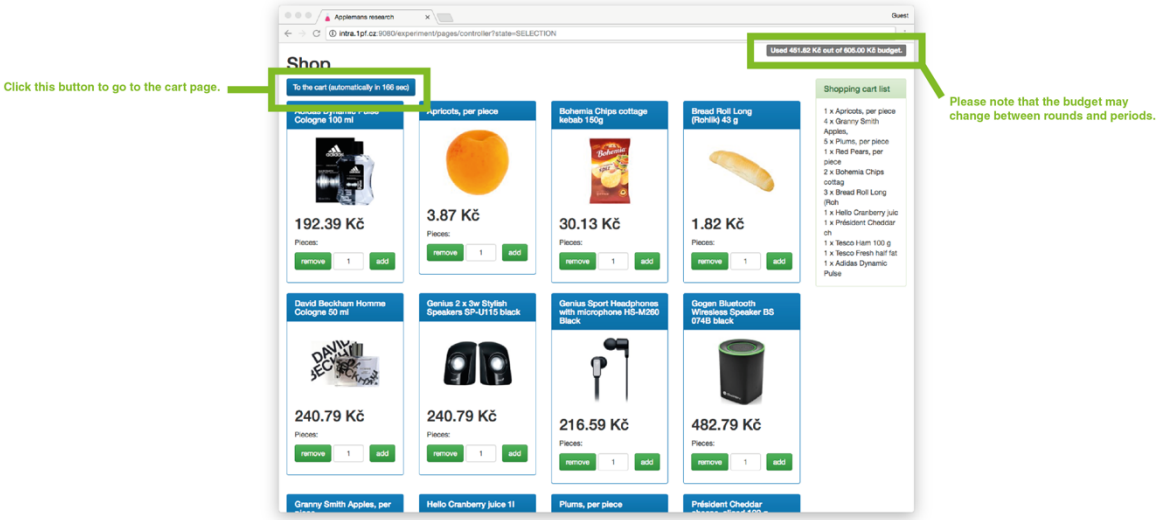


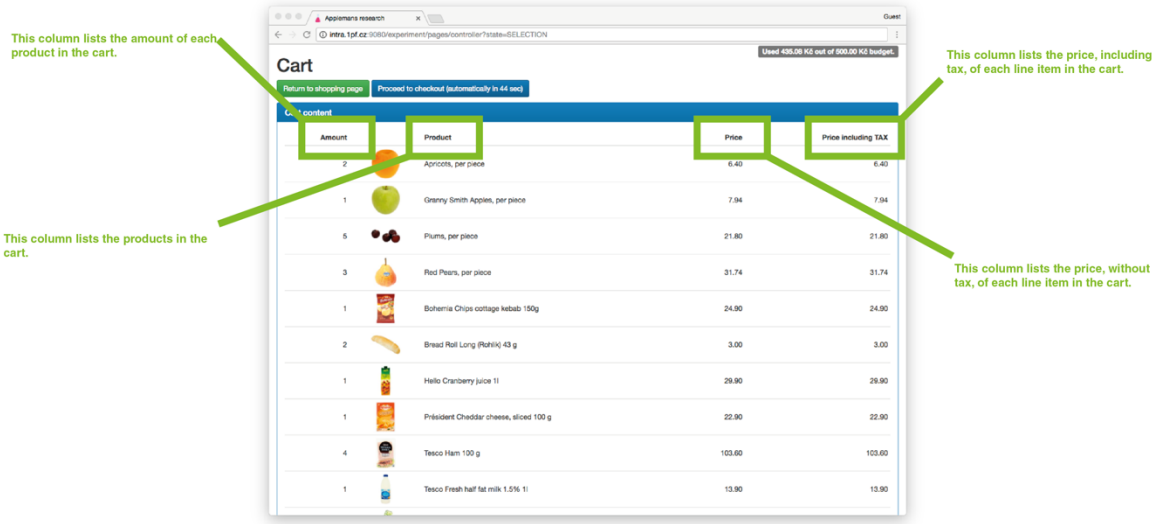


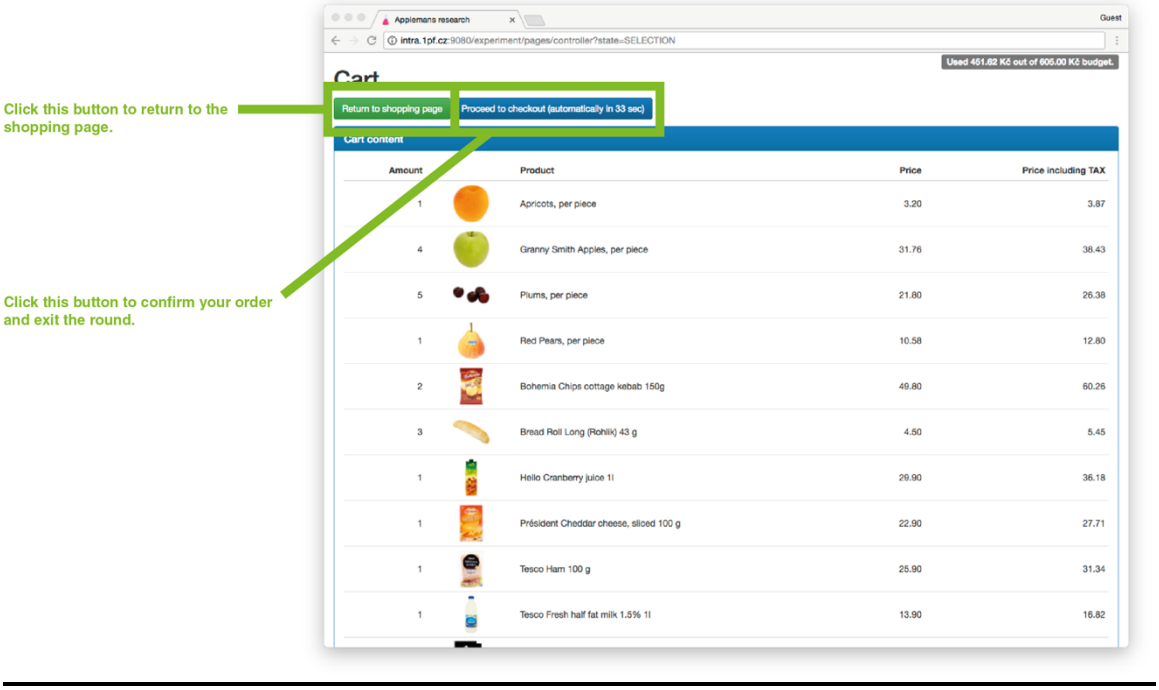

Before every period, you will be shown a list of the goods that will be taxed that period. Please note that a given period may not have any taxed goods.
Push the continue button to begin the next shopping round.
Selection recorded. Thank you.
The goods that will be taxed throughout the upcoming period (through all 3 rounds):
Continue
Bananas, per piece
3.23 Kč
Bohemia Chips cottage kebab 150g
24.90 Kč
Bohemia Chips moravian bacon 150g
24.90 Kč
Gala Apples, per piece
7.15 Kč
Granny Smith Apples, per piece
7.94 Kč
Tribal Earrings ESSW01 TOPAZ
190.00 Kč
Trust Urban Revolt Lace Stylish Headphones with microphone - grey
399.00 Kč
Victorinox Swiss Army Knife - Classic SD
299.00 Kč
Victorinox Swiss Army Knife - Rally
Victorinox Swiss Army Knife - Waiter


Push the "add" button to add the number of goods displayed in the number box into the cart.
Push the "remove" button to remove the number of goods displayed in the number box from the cart.
This is the number box. It's default value is 1. That means 1 item would be added or removed from the cart each time the "add" or "remove" buttons are pushed. Changing it to, for example, 3, would then add or remove 3 items from the cart.
Shop
Used 0.00 Kč out of 500.00 Kč budget.
To the cart (automatically in 240 sec)
Shopping cart list
Adidas Dynamic Pulse Cologne 100 ml
Apricots, per piece
3.87 Kč
Bohemia Chips cottage kebab 150g
30.13 Kč
Bread Roll Long (Rohlík) 43 g
1.82 Kč
David Beckham Homme Cologne 50 ml
240.79 Kč
Genius 2 x 3w Stylish Speakers SP-U115 black
240.79 Kč
Genius Sport Headphones with microphone HS-M260 Black
216.59 Kč
Gogen Bluetooth Wireless Speaker BS 074B black
482.79 Kč
Granny Smith Apples, per
Hello Cranberry juice 1l
Plums, per piece
Président Cheddar


Be sure to mind your time.
Here you can see how much of your round's budget you have spent.
This is your cart preview list. It displays all the items you have chosen in a given round.
Shop
To the cart (automatically in 21 sec)
Used 435.08 Kč out of 600.00 Kč budget.
Shopping cart list
Adidas Dynamic Pulse Cologne 100 ml
159.00 Kč
Apricots, per piece
3.20 Kč
Bohemia Chips cottage kebab 150g
24.90 Kč
Bread Roll Long (Rohlík) 43 g
1.50 Kč
David Beckham Homme Cologne 50 ml
199.00 Kč
Genius 2 x 3w Stylish Speakers SP-U115 black
199.00 Kč
Genius Sport Headphones with microphone HS-M260 Black
179.00 Kč
Gogen Bluetooth Wireless Speaker BS 074B black
399.00 Kč
Granny Smith Apples, per
Hello Cranberry juice 1l
Plums, per piece
Président Cheddar


Click this button to go to the cart page.
Please note that the budget may change between rounds and periods.
Shop
To the cart (automatically in 166 sec)
Used 481.82 Kč out of 805.00 Kč budget.
Shopping cart list
Adidas Dynamic Pulse Cologne 100 ml
192.39 Kč
Apricots, per piece
3.87 Kč
Bohemia Chips cottage kebab 150g
30.13 Kč
Bread Roll Long (Rohlík) 43 g
1.82 Kč
David Beckham Homme Cologne 50 ml
240.79 Kč
Genius 2 x 3w Stylish Speakers SP-U115 black
240.79 Kč
Genius Sport Headphones with microphone HS-M260 Black
216.59 Kč
Gogen Bluetooth Wireless Speaker BS 074B black
482.79 Kč
Granny Smith Apples, per
Hello Cranberry juice 1l
Plums, per piece
Président Cheddar


This column lists the amount of each product in the cart.
This column lists the products in the cart.
This column lists the price, including tax, of each line item in the cart.
This column lists the price, without tax, of each line item in the cart.
Cart
Return to shopping page
Proceed to checkout (automatically in 44 sec)
Amount
Product
Price
Price including TAX


Click this button to return to the shopping page.
Click this button to confirm your order and exit the round.
Cart
Return to shopping page
Proceed to checkout (automatically in 33 sec)
Cart content
Amount
Product
Price
Price including TAX
1 Apricots, per piece 3.20 3.87
4 Granny Smith Apples, per piece 31.76 38.43
5 Plums, per piece 21.80 26.38
1 Red Pears, per piece 10.58 12.80
2 Bohemia Chips cottage kebab 150g 49.80 60.26
3 Bread Roll Long (Rohlík) 43 g 4.50 5.45
1 Hello Cranberry juice 1l 29.90 36.18
1 Président Cheddar cheese, sliced 100 g 22.90 27.71
1 Tesco Ham 100 g 25.90 31.34
1 Tesco Fresh half fat milk 1.5% 1l 13.90 16.82

5. Comprehension questionnaire

1) The superstore marketplace is a full combination of which of the following elements from the grocery and department store marketplaces?

a) Only the goods

b) Only the budget

c) Only the time limit

d) All of the above (goods, budget, and time limit)

2) True or false: you can save and use unused budget amounts in another round?

a) True

b) False

3) A good that has a pre-tax price of 100Kc, which is subject to a 21% RST tax has:

a) A price tag of 100Kc

b) A price tag of 121Kc

c) A final price paid of 100Kc

d) The same amount on the price tag as the final price paid

4) True or false: you can navigate the shopping environment using the browser's back button?

a) True

b) False

5) The computer will randomly choose one, and only one, round at random for your earnings, so shopping decisions made during each round are individually important, because what will make up your earnings from this experiment?

a) The goods selected in that one and only chosen round.

b) The goods selected throughout all three rounds from a chosen period.

c) The remaining, unused budget from that one and only chosen round.

d) The goods selected and 50% of the remaining, unused budget from that one and only chosen round.

Correct answers to the comprehension questionnaire

1) The superstore marketplace is a full combination of which of the following elements from the grocery and department store marketplaces?

~~a) Only the goods~~

~~b) Only the budget~~

~~c) Only the time limit~~

d) All of the above (goods, budget, and time limit)

2) True or false: you can save and use unused budget amounts in another round.

~~a) True~~

b) False; the budget does not carry over or accumulate in any fashion and there is no borrowing or credit between rounds.

3) A good that has a pre-tax price of 100Kc, which is subject to a 21% RST tax has:

a) A price tag of 100Kc; and a final price of 121Kc would be paid.

~~b) A price tag of 121Kc~~

~~c) A final price paid of 100Kc~~

~~d) The same amount on the price tag as the final price paid~~

4) True or false: you can navigate the shopping environment using the browser's back button.

~~a) True~~

b) False; if you push the browser's back button, you may have to repeat that given round/portion.

5) The computer will randomly choose one, and only one, round at random for your earnings, so shopping decisions made during each round are individually important, because what will make up your earnings from this experiment?

~~a) The goods selected in that one and only chosen round.~~

~~b) The goods selected throughout all three rounds from a chosen period.~~

~~c) The remaining, unused budget from that one and only chosen round.~~

d) The goods selected and 50% of the remaining, unused budget from that one and only chosen round.

Pre-period Instructions

Selection recorded. Thank you.

The following period is made up of three independent rounds, one of each type of marketplace: a grocery store style round, a department store style round, and a superstore style round.

The goods that will be taxed throughout the upcoming period (throughout all 3 rounds):

Generated list of goods (not to be written on the page)

Pre-period Instructions (for welfare control)

Selection recorded. Thank you.

The following round will be a superstore style market round.

The goods that will be taxed throughout the upcoming round:

Generated list of goods (not to be written on the page)

Between-round Instructions

Selection recorded. Thank you.

Once you are ready to begin the next round, please click the button below.

Survey Instructions

Thank you for completing the shopping portion of the experiment.

Please complete the following survey.

Once complete, the computer will randomly choose one of the rounds you completed during the shopping portion of the experiment and present you with your earnings info.

Earnings

You have successfully finished the experiment. Thank you for your time today.

Please raise your hand so that one of the researchers may give you the final questions to fill out. Thereafter, you may once again use your electronic devices, but please make sure they are set to silent, and wait quietly while the other participants complete their tasks.

At the appointed date and time, please come collect your cash and goods.

Have a nice day!

APPENDIX A2: Survey Questions

1. What is your gender?

Male

Female

2. What is your age?

Below 18

18-19

20-21

22-23

24-25

26-27

28-29

30-32

33-35

36-39

40-44

45-49

50-54

55-59

60-69

70+

3. What is the highest level of education you have completed?

some high school

high school graduate

trade/technical/vocational training

some undergraduate study

bachelor's degree

some postgraduate study

master's degree

doctoral candidate

doctoral degree

4. Are you currently enrolled as a student?

Yes, full time

Yes, part time

Yes, distance learning

No

5. What is your official nationality?

American (USA)

Armenian

Canadian

Czech

Georgian

Ethiopian

Israeli

Mexican

Russian

Slovak

Ukrainian

Other

6. Were you hungry at all during the experiment?

Yes, I was already hungry at the start of the experiment.

Yes, I became hungry during the first half of the experiment.

Yes, I became hungry during the second half of the experiment.

No, I was not hungry throughout the entire experiment.

7. Who does the shopping in your household?

Me

Mostly me

Mostly somebody else

Somebody else

8. What is your household status?

living with parents/family/other and not paying rent

renting a room/property with others (including school housing)

renting a property as head of household (alone or with a significant other)

living in a property I own as head of household (alone or with a significant other)

homeless

9. What is your marital status?

single/never been married

married

separated

divorced

widowed

10. Do you have any children (please include biological, adopted, and step)? If yes, how many and do they reside with you?

Yes, my one and only child resides in my household full time

Yes, my one and only child resides in my household part time

Yes, my one and only child does not reside in my household

Yes, my two children reside in my household full time

Yes, my two children reside in my household part time

Yes, my two children do not reside in my household

Yes, my three or more children reside in my household full time

Yes, my three or more children reside in my household part time

Yes, my three or more children do not reside in my household

No, I do not have any children

11. Which of the following categories best describes your employment status?

Employed, working 1-39 hours per week

Employed, working 40 or more hours per week

Self-employed

Out of work and looking for work

Out of work but not currently looking for work

A homemaker and not otherwise working

A full time student and not otherwise working

Military

Retired

Disabled and unable to work
Other

12. How much money did YOU personally earn in 2016? Please include ALL sources of income (i.e. wages from jobs; net income from business, farm, or rent; pensions; dividends; interest; social security payments; stipends; money given or loaned to you by family or friends; and any other money received by YOU). Please report the total amount of money received from any and all sources.

0,-Kč – 49.999,-Kč	(0,-Kč – 4.166,-Kč monthly)
50.000,-Kč – 99.999,-Kč	(4.167,-Kč – 8.333,-Kč monthly)
100.000,-Kč – 199.999,-Kč	(8.334,-Kč – 16.666,-Kč monthly)
200.000,-Kč – 299.999,-Kč	(16.667,-Kč – 24.999,-Kč monthly)
300.000,-Kč – 399.999,-Kč	(25.000,-Kč – 33.333,-Kč monthly)
400.000,-Kč – 499.999,-Kč	(33.334,-Kč – 41.666,-Kč monthly)
500.000,-Kč – 749.999,-Kč	(41.667,-Kč – 62.499,-Kč monthly)
750.000,-Kč – 999.999,-Kč	(62.500,-Kč – 83.333,-Kč monthly)
1.000.000,-Kč – 1.999.999,-Kč	(83.334,-Kč – 166.666,-Kč monthly)
2.000.000,-Kč or more	(166.667,-Kč or more monthly)

13. How much total combined money did all members of your HOUSEHOLD earn in 2016? Please include ALL sources of income (i.e. wages from jobs; net income from business, farm, or rent; pensions; dividends; interest; social security payments; stipends; money given or loaned to you by family or friends residing outside of the household; and any other money received by all members of your HOUSEHOLD that are EIGHTEEN (18) years of age or older). Please report the total amount of money received from any and all sources.

0,-Kč – 99.999,-Kč	(0,-Kč – 8.333,-Kč monthly)
100.000,-Kč – 199.999,-Kč	(8.334,-Kč – 16.666,-Kč monthly)
200.000,-Kč – 299.999,-Kč	(16.667,-Kč – 24.999,-Kč monthly)
300.000,-Kč – 399.999,-Kč	(25.000,-Kč – 33.333,-Kč monthly)
400.000,-Kč – 499.999,-Kč	(33.334,-Kč – 41.666,-Kč monthly)
500.000,-Kč – 749.999,-Kč	(41.667,-Kč – 62.499,-Kč monthly)
750.000,-Kč – 999.999,-Kč	(62.500,-Kč – 83.333,-Kč monthly)
1.000.000,-Kč – 1.499.999,-Kč	(83.334,-Kč – 124.999,-Kč monthly)
1.500.000,-Kč – 2.999.999,-Kč	(125.000,-Kč – 249.999,-Kč monthly)
3.000.000,-Kč or More	(250.000,-Kč or more monthly)

14. Growing up, did your parents own a car and/or real estate property(ies)?

No, my parents neither owned a car nor a property

Yes, my parents did own the property we lived in, but did not own a car

Yes, my parents owned multiple properties, but did not own a car

Yes, my parents owned one car, but no property

Yes, my parents owned one car and the property we lived in

Yes, my parents owned one car and multiple properties

Yes, my parents owned at least one car each, but no property

Yes, my parents owned at least one car each and the property we lived in

Yes, my parents owned at least one car each and multiple properties

15. Do you and/or your parents currently own a car and/or real estate property(ies)?

No, neither my parents nor I own a car or a property

Yes, my parents and/or I do own the property we live in, but do not own a car

Yes, my parents and/or I own multiple properties, but do not own a car

Yes, my parents and/or I own at least one car, but no property

Yes, my parents and/or I own at least one car and the property we live in

Yes, my parents and/or I own at least one car and multiple properties

Yes, my parents and I own at least one car each, but no property

Yes, my parents and I own at least one car each and the properties we live in

Yes, my parents and I own at least one car each and multiple properties

16. Out of the 50 grocery store goods that were available during the goods selection portion, how many were you interested in purchasing?

0

1-3

4-6

7-9

10+

17. Out of the 50 department store goods that were available during the goods selection portion, how many were you interested in purchasing?

0

1-3

4-6

7-9

10+

18. Out of the 10 grocery store items you faced during the shopping portion, how many of them were you seriously interested in purchasing (please exclude all items you chose as just maybe interesting or needed to achieve the required 10)?

0

1

2

3

4

5

6

7

8

9

10

19. Out of the 10 department store items you faced during the shopping portion, how many of them were you seriously interested in purchasing (please exclude all items you chose as just maybe interesting or needed to achieve the required 10)?

0

1

2

3

4

5

6

7

8

9

10

20. Out of the 10 grocery store items you faced during the shopping portion, how many were your strongly preferred brand (the one you usually buy / want)?
(Definition of brand: a brand is a distinguishing symbol, mark, logo, name, word, sentence or a combination of these items that companies use to distinguish their product from others in the market.)

0
1
2
3
4
5
6
7
8
9
10

21. Out of the 10 department store items you faced during the shopping portion, how many were your strongly preferred brand (the one you usually buy / want)?
(Definition of brand: a brand is a distinguishing symbol, mark, logo, name, word, sentence or a combination of these items that companies use to distinguish their product from others in the market.)

0
1
2
3
4
5
6
7
8
9
10

22. Please think about a grocery store good that you chose in one round, but then did not choose (or chose less of) in the same round of a later period. What was the main reason you did not choose (chose less of) that good?

Price

Taxes

Price including taxes

Other factors (brand, appearance, quality, etc.)

I always chose the same grocery store goods in every grocery store and superstore rounds

23. Please think about a department store good that you chose in one round, but then did not choose (or chose less of) in the same round of a later period. What was the main reason you did not choose (chose less of) that good?

Price

Taxes

Price including taxes

Other factors (brand, appearance, quality, etc.)

I always chose the same department store goods in every department store and superstore rounds

24. Did you ever and, if so, how many times did you accidentally click the checkout before you wished to complete a round?

0

1

2

3

4

5

6

7

8

9

10+

APPENDIX A3: Post Experiment Questionnaire

Exact time now: _____ Your registration number: _____

1) Did you recognize any of your preferred and/or strongly preferred brands during the experiment? If yes, please list them below by marketplace type.
(Definition of brand: a brand is a distinguishing symbol, mark, logo, name, word, sentence or a combination of these items that companies use to distinguish their product from others in the market.)

a) I didn't recognize any of my preferred or strongly preferred brands

b) I recognized the following brands:

a) Preferred grocery brands: _____

_____

_____

b) Strongly preferred grocery brands: _____

_____

_____

c) Preferred department brands: _____

_____

_____

d) Strongly preferred department brands: _____

_____

_____

2) Were you ever <u>forced</u> (i.e. constrained by your budget) to purchase less of any good in any round that you were able to purchase in another round of the same type <u>solely</u> due to an increase in price from taxes? *(Please choose all that apply.)*

a) No, I always consciously chose the quantities purchased per round conditions

b) Yes, during a grocery store round

c) Yes, during a department store round

d) Yes, during a superstore round

e) Yes, but only during the last three superstore rounds

Comments: _____

Exact time now: _____

3) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents

4) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ minutes

5) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____ days

Exact time now: _____

6) Please complete the following arithmetic exercises:

a) 1.90 * 1.21 = _____

b) 35.16 * 1.21 = _____

d) 319.00 * 1.21 = _____

Exact time now: _____

APPENDIX A4: List of Goods

	Price (CZK)	Translated name	Category	Brand
Grocery Store				
1	1.9	Bread Roll Long (Rohlík) 43 g	1	Penam
2	2.2	Bread Roll (Houska) 50 g	1	Penam
3	10.9	Rustic bread loaf, dark, 100 g	2	Inter Europol Pekarna Szwajcarska
4	9.9	Rustic baguette, buckwheat, 120 g	2	La Lorraine
5	8.9	Grandma's strawberry yoghurt 150 g	3	Ehrmann
6	8.9	Grandma's sour cherry yoghurt 150 g	3	Ehrmann
7	25.9	Fresh Eggs - S, 10 pieces	4	Tesco
8	34.9	Fresh Eggs - M, 10 pieces	4	Tesco
9	30.9	Mayonnaise, 225ml	5	Hellmann's
10	33.9	Mayonnaise delicate, 225ml	5	Hellmann's
11	24.9	Fresh half fat milk 1.5% 1l	6	Kunin
12	26.9	Fresh whole milk 4%, 1l	6	Kunin
13	9.58	Pears, green, per piece	7	Tesco
14	9.44	Pears, red, per piece	7	Tesco
15	19.9	Bundle of Carrots	8	Tesco
16	14.9	Bag of Carrots, 1Kg	8	Tesco
17	32.9	Broccoli stalk	9	Efes
18	29.9	Cauliflower stalk	9	Efes
19	33.9	Bag of Potatoes, 2Kg	10	BROP
20	33.9	Bag of Red Potatoes, 2Kg	10	BROP
21	7.58	Apples, Red Delicious, per piece	11	Tesco
22	6.98	Apples, Granny Smith, per piece	11	Tesco
23	5.24	Oranges, per piece	12	Tesco
24	5.08	Bananas, per piece	12	Tesco
25	18.9	100% Apple juice 1l	13	Tesco
26	21.9	100% Orange juice 1l	13	Tesco
27	29.9	Shaved Ham Off the Bone, 100g	14	Le & Co
28	29.9	Shaved Turkey Breast, 100g	14	Le & Co
29	23.9	Eidam cheese slices, 100g	15	Natur Lander
30	26.9	Swiss (Emental) cheese slices, 100g	15	Natur Lander
31	19.9	Potato Chips, salted, 77g	16	Lay's
32	16.7	Potato Chips ridged, salted, 70g	16	Lay's
33	24.9	Jumbo Raisins, 200g	17	Tesco
34	27.9	Prunes, 200g	17	Tesco
35	18.9	Penne pasta, 500g	18	Rosicke
36	18.9	Fusili pasta, 500g	18	Rosicke
37	21.9	Long grain rice, 4 cooking bags, 480g	19	Lagris
38	24.9	Parboiled rice, 4 cooking bags, 480g	19	Lagris
39	22.9	Mushrooms, garden champignons, 250g	20	Tesco
40	34.9	Mushrooms, brown champignons, 250g	20	Tesco
41	28.9	Mini Muffins Marble (chocolate & vanilla) 180 g	21	Delasheras

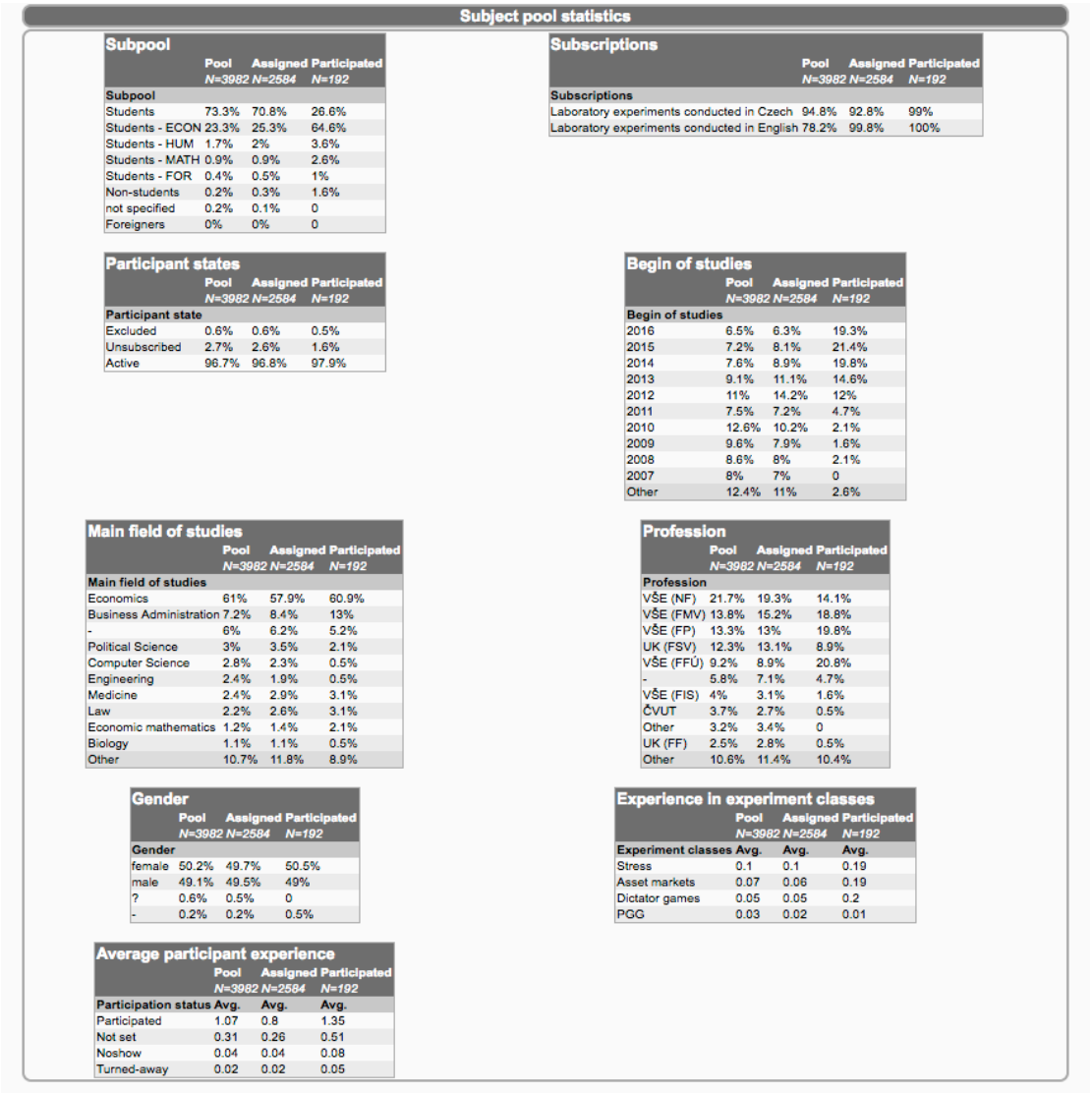
42	28.9	Mini Muffins Doucle Choc (with chocolate bits) 180 g	21	Delasheras
43	19.9	Cream Cheese, 80g	22	Gervais
44	24.9	Cream Cheese whipped, 120g	22	Gervais
45	32.12	Fresh Chicken Legs (spodní stehna), per piece	23	Tesco
46	35.16	Fresh Chicken Thighs (horní stehna), per piece	23	Tesco
47	29.9	Fresh Baby Tomatoes, 250g package	24	Tesco
48	29.9	Fresh Cherry Tomatoes, 250g package	24	Tesco
49	23.9	Fresh Salad Mix Venetian, 160g, (frisée lettuce, endive, radicchio, rucola)	25	Tesco
50	26.9	Fresh Salad Mix Italian, 150g, (frisée lettuce, endive, radicchio, romaine lettuce)	25	Tesco
	Price (CZK)	Translated name	Category	Brand
Department store				
1	119	External multiple memory card reader, white, USB 2.0, 4-slot SD / microSD / MS / M2	101	Axagon
2	219	Multi connector data cable 1.5m, white, USB-C, microUSB, Lightning, 2.4A	101	PowerCube
3	189	Powerbank Credit Card, 2000mAh, 62x96x7mm, black	101	Omega
4	159	Audio headphone splitter, retractive, 2 x 3.5 mm jack, gold-plated connectors, max. cable length 0.9 m	101	Retrax
5	129	Cordless mouse, optical, 800dpi, USB, black	101	Hama
6	179	Multimedia gaming keyboard, slim, black-red, CZ/SK, USB, minimal joint impact technology	101	C-Tech
7	299	Gaming headphones, black-red, noise-canceling, 20 Hz - 20 kHz, 110 dB 32 Ohm, 3.5mm jack, 2m cable with volume control	102	Zalman
8	199	Earbuds with microphone and hands-free controls, black, ergonomic design, frequency range 20-20000Hz, sensitivity of 102 db, Impedance 32Ohm, 1.2 m cable, 3.5mm jack	102	Gogen
9	269	Sport earphones, blue, waterproof (coverage IPX2), sensitivity of 112 db, impedance 23Ohm	102	Panasonic
10	159	Stylish earbuds, retractable, neon blue-green, three sizes of earplug, gold-plated 3.5mm jack, 1.2 m	102	Retrax
11	299	Wireless speaker, silver, 3 watt, Bluetooth 3.0, FM, microSD, 350mAh battery (lasts 3-4 hours), aluminum	102	Omega
12	199	Portable Speakers, 2.0, 2x2W, USB, black & orange	102	Defender
13	169	Power strip with switch, gray, 5 socket, 5m cord, child proof, flip-out hanger, light indicator	103	Connect IT
14	269	Electronic luggage scale with pressure gauge, max 50 kg, silver & black, 2 units, function TARA, overload indicator and low battery	103	Orava

15	299	Personal scale, sunset motif, 150Kg, 100g increments, 6mm tempered safety glass, LCD display	103	Gallet
16	275	Digital battery tester for round & button batteries from 1.2 to 12V, LCD display, displays residual capacity in % (2xLR44 button batteries included)	103	Voltcraft
17	149	Surge protector with 2 usb ports, power 2200W, load current of 10 A, energy 175 J, 1 socket	103	Defender
18	199	Flash Disk, stylish with crystal, 8GB, USB 2.0, red	104	Silicon Power
19	229	Flash Disk, 16GB, USB 2.0, black	104	Silicon Power
20	169	Micro SDHC 8GB Class 10	104	Adata
21	279	Micro SDHC 16GB Class 10 + SD adapter + USB reader	104	Kingston
22	99	Firestarter: Tinder on a Rope	105	LMF
23	109	Firestarter: Magnesium Flint Striker	105	Frendo
24	219	Pocket knife, stainless steel 420, blue plastic sides, blade lengths 70mm & 90mm, 12 mm blade width, 2 openers, weight 59 grams	105	Mikov
25	279	Multifunction folding keychain, 5 tools in one, 53x35x7mm, material stainless steel	105	Munkees
26	319	Multifunction car & camping light, hand-crank, usb charging port, magnets, seat belt knife, glass hammer, for emergencies (5 minutes crank = 5 hours power)	105	Evolveo
27	139	Aluminum Water Bottle, 600ml, green	106	Frendo
28	129	Jump rope with mechanical counter, black & green	106	Lifefit
29	149	Butterfly trainer, adjustable resistance, metal arms in soft foam (durable & comfortable), easily stored, length 53 cm, weight 420 g (suitable for men, women, and children)	106	Spokey
30	139	Hand grips, mechanical counter, variable resistance, ergonomic grip	106	Lifefit
31	209	Fitness mat with a non-slip surface, dimensions 173 x 61 x 0.4 cm, moisture resistance, high-quality non-toxic material	106	Lifefit
32	189	Twister rotating disk, for the whole family, helps eliminate fat from the waist without burdening the spine, strengthens the legs and massages the feet	106	Lifefit
33	99	Mascara, black, 8ml, curved brush	107	Rimmel London
34	99	Nail polish gel effect, 12ml, gives effect of a gel manicure, extends life of nail polish (up to 3 times), and leaves nails shiny, smooth, and resistant to abrasion and fraying	107	Eveline

35	99	Nail polish rapid dry, accelerates drying, keeps nails strong and beautifully shiny, works within 45 seconds	107	Dermacol
36	255	Nail Polish Remover Rose 120ml	107	Alessandro
37	279	Eyeshadow palette with 12 shades of brown and beige tones, includes a two-sided applicator	107	Maybelline
38	159	Eyeshadow, Smoky Eyes Trio, Violet Romantic, 4,5 g	107	Bourjois
39	129	Facial Creamy Scrub Apricot, 150ml, natural peeling composition of apricot extracts, vitamins C, E, and minute particles of walnut shells	107	Freeman
40	109	Make-up removal lotion, 200 ml, with aloe vera & hibiscus extract, suitable for very sensitive eyes, gently removes makeup and impurities	107	Biopha
41	149	Cologne 100ml, spicy scent; Head: grapefruit, bergamot, anise, and peppermint; Heart: geranium, patchouli, and sandalwood; Basis: musk and tonka	108	Adidas
42	159	Cologne 100ml, woody scent; Head: lavender, vetiver; Heart: bergamot; Basis: jasmine, woody notes	108	Cuba
43	149	Perfume 30ml, fruity scent; Head: Black currant, strawberry; Heart: cyclamen, freesia; Basis: Musk	108	Adidas
44	229	Perfume 30ml, green tea scent spray; Head: notes of cumin, rhubarb crisp citrus fruit, bergamot; Basis: oak moss, musk, white amber; Heart: peppermint, jasmine, carnation, fennel, celery seed	108	Elizabeth Arden
45	249	Crystal heart earrings, silver, purity: 925/1000 RH, weight 0.2 g, diameter: 7 mm	109	Swarovski
46	249	Bracelet with heart pendant, material: stainless steel, length: 21 cm, width: 6 mm	109	Tribal
47	299	Pearl earrings, material: material: silver, fineness: 925/1000 RH, weight: 3.2 g, length: 20 mm, diameter: 10 mm (bijouterie pearls)	109	Swarovski
48	259	Butterfly pendant material: silver, purity: 925/1000 RH, weight 0.9 g, length: 20 mm, width: 17 mm	109	Lola Aura
49	290	Bracelet golden brown and shiny leather, decorated with small glittering stones and metal accessories, magnetic clasp	109	Tamaris
50	290	Bracelet beige leather, decorated with small sparkling clear rhinestones and metal studs, firm clasp, coils twice around the wrist	109	Tamaris

APPENDIX A5: Further Information on Experimental Participants and Design

Figure A5.1: Subject Pool Statistics



Subject pool statistics

Subpool

Subpool	Pool N=3982	Assigned N=2584	Participated N=192
Students	73.3%	70.8%	26.6%
Students - ECON	23.3%	25.3%	64.6%
Students - HUM	1.7%	2%	3.6%
Students - MATH	0.9%	0.9%	2.6%
Students - FOR	0.4%	0.5%	1%
Non-students	0.2%	0.3%	1.6%
not specified	0.2%	0.1%	0
Foreigners	0%	0%	0

Subscriptions

Subscriptions	Pool N=3982	Assigned N=2584	Participated N=192
Laboratory experiments conducted in Czech	94.8%	92.8%	99%
Laboratory experiments conducted in English	78.2%	99.8%	100%

Participant states

Participant state	Pool N=3982	Assigned N=2584	Participated N=192
Excluded	0.6%	0.6%	0.5%
Unsubscribed	2.7%	2.6%	1.6%
Active	96.7%	96.8%	97.9%

Begin of studies

Begin of studies	Pool N=3982	Assigned N=2584	Participated N=192
2016	6.5%	6.3%	19.3%
2015	7.2%	8.1%	21.4%
2014	7.6%	8.9%	19.8%
2013	9.1%	11.1%	14.6%
2012	11%	14.2%	12%
2011	7.5%	7.2%	4.7%
2010	12.6%	10.2%	2.1%
2009	9.6%	7.9%	1.6%
2008	8.6%	8%	2.1%
2007	8%	7%	0
Other	12.4%	11%	2.6%

Main field of studies

Main field of studies	Pool N=3982	Assigned N=2584	Participated N=192
Economics	61%	57.9%	60.9%
Business Administration	7.2%	8.4%	13%
-	6%	6.2%	5.2%
Political Science	3%	3.5%	2.1%
Computer Science	2.8%	2.3%	0.5%
Engineering	2.4%	1.9%	0.5%
Medicine	2.4%	2.9%	3.1%
Law	2.2%	2.6%	3.1%
Economic mathematics	1.2%	1.4%	2.1%
Biology	1.1%	1.1%	0.5%
Other	10.7%	11.8%	8.9%

Profession

Profession	Pool N=3982	Assigned N=2584	Participated N=192
VŠE (NF)	21.7%	19.3%	14.1%
VŠE (FMV)	13.8%	15.2%	18.8%
VŠE (FP)	13.3%	13%	19.8%
UK (FSV)	12.3%	13.1%	8.9%
VŠE (FFÚ)	9.2%	8.9%	20.8%
-	5.8%	7.1%	4.7%
VŠE (FIS)	4%	3.1%	1.6%
ČVUT	3.7%	2.7%	0.5%
Other	3.2%	3.4%	0
UK (FF)	2.5%	2.8%	0.5%
Other	10.6%	11.4%	10.4%

Gender

Gender	Pool N=3982	Assigned N=2584	Participated N=192
female	50.2%	49.7%	50.5%
male	49.1%	49.5%	49%
?	0.6%	0.5%	0
-	0.2%	0.2%	0.5%

Experience in experiment classes

Experiment classes	Pool N=3982 Avg.	Assigned N=2584 Avg.	Participated N=192 Avg.
Stress	0.1	0.1	0.19
Asset markets	0.07	0.06	0.19
Dictator games	0.05	0.05	0.2
PGG	0.03	0.02	0.01

Average participant experience

Participation status	Pool N=3982 Avg.	Assigned N=2584 Avg.	Participated N=192 Avg.
Participated	1.07	0.8	1.35
Not set	0.31	0.26	0.51
Noshow	0.04	0.04	0.08
Turned-away	0.02	0.02	0.05

Table A5.1: Table with Main Elements of the Experiment & Explanations

Experimental Element	Explanation/Mapping
Three different rounds: -Grocery store (necessity) -Department store (luxury) -Superstore (combined)	The grocery store and department store separation allow the direct comparison of purchase size effect under the same conditions with the only variable changed being the cost/type of good. The existence of this effect will also lend credence to the bounded rationality cause of the effect. The superstore allows the participant to purchase

	amongst the necessity and luxury goods without an externally imposed budget separation. This both gives the participant the more natural ability to allocate their budget as they see fit as well as provides some indication of budget allocation behavior (i.e. will subjects end up spending "excessively" on taxed luxury goods as in Chetty's car example?).
Two types of control sections: 1) An exact duplicate of the treatment section except the taxes are included in the price tag (fully salient) 2) Three superstore rounds -A round with no taxed goods -A round with all goods taxed -A round with all goods taxed and compensated budget	The first control section allows a direct within- and between-subject assessment. A single variable that differs between control and treatment sections (the salience of the taxes) is optimal per the scientific method. The second control section should capture each participant's preferences and purchase strategy methods, which potentially allows for individual welfare evaluations.
Categories of goods (25 direct grocery substitute categories, and 9 loose department & 9 loose grocery close substitute categories)	The use of categories of goods allows for the assessment of how subjects behave not only between taxed and untaxed goods, but also between taxed and untaxed substitute goods. This difference should provide evidence that indicates whether people behave as if their utility functions were quasilinear or separable in nature. That is, if they behave the same between any taxed and untaxed goods as they do with substitute taxed and untaxed goods, then this points towards quasilinear behavior. Such quasilinear behavior also intuitively indicates that people would follow the budget adjustment rules that reduce future purchases of both the taxed and untaxed goods, which would mean that tax salience is strictly positive per CLK. As Reck (2014, p.9) describes, "intuitively, the individual acts like her income is reduced in lump-sum fashion during budget adjustment."
Two internal wealth/income categories	Internal wealth/income categorization of subjects allows another check of heterogenic behavior imposed within the experiment. While this originally had more to do with the second focus/paper, it was decided to keep it in the experiment since it seemed the effect power would

	be strong enough to reach significance even with the fewer subjects and as there were indications from the pilot data that it may provide other insights into the purchase size behavior than solely as an internal heterogenic test.
Survey with characteristic and control questions (i.e. hunger, shopping, etc.)	The survey came just before the remuneration page and collected all the info needed to analyze the second focus: heterogeneity effect. For this paper, that info serves for control purposes. In addition, it includes useful questions about a participant's state during the experiment, such as their hunger, as well as their familiarity with shopping and prices, for robustness checks.
Instruction comprehension questions	Allows for the standard understanding robustness check. The page with the correct answers adds credibility to the contention that it helped clarify the experiment ex-post and, thus, supports the idea that even those who did not get all the questions correct understood thereafter.
Goods selection	Narrows the shopping choices for the remainder of the experiment, thus reducing noise and increasing the power of the testing, while still securing that the items involved would be in the subject's utility preference space.
Single chosen remuneration round	Standard element in experimental economics to motivate participants to "play" each and every round as an equally important, equally likely chosen, round and, thus, should eliminate extra-round strategic behaviors.
E-shop environment	Increases external validity. This type of environment has now been employed in several experiments, including within this specific branch of study.
Real goods and cash remuneration (both collected at the same time in 7 days)	Real goods and cash remuneration increases external validity and collection at the same time removes any time-inconsistent strategic behavior (i.e. choosing to take only cash since it could be obtained immediately, while they would have to wait for the goods).
Portion of unused budget (pilot evidence resulted in 50% being the fraction employed in the final experiment)	Increases likelihood of purchasing within the experiment. Alternatively, large discounts could be given on the goods to increase in-experiment purchases, as done in other experiments, but the

	real price method may enhance external validity and the, pragmatically, the discount method is more expensive…
Alternating period methodology	Having half the goods taxed one period and their counterparts the next ensures that purely randomized division does not conceal behavior. Further, period pairs will directly show if participants are responding to salience effects already within period pairs, which permits another layer of analysis and may further illuminate within- and between-subject behavioral variance.
Individual game	As it is an individual game, subjects would be able to complete the experiment at their own pace, which should help diminish fatigue concerns.
All subjects must wait in the lab until all are finished	As the subjects will need to remain in the lab until all subjects from that session complete the experiment, this should significantly reduce "click through" corner solution incentives. Along with the previous element, it creates a well-balanced environment that allows individual pace to reduce fatigue and still reduces the incentive of racing through the experiment simply to obtain the remuneration without exuding effort.
Round and period order randomization	Either the control and treatment sections would have had to be ordered differently amongst the participants or all the periods would have had to be randomly ordered to reduce order and learning effects.

APPENDIX A6: Additional Result Details

A6.1. COMPARISON OF MEANS WITH DISCRETE QUANTITY

The standard variable of quantity (y)—also applies to revenue (x)—produces a questionable illustration of PSE, because the associated quantities—and spends—differ considerably between the two forms of goods simply since each unit of department good takes up a materially larger portion of the shopping budget. This can be seen in the table below. Nevertheless, the table visibly reveals a PSE occurring (discernible by comparing the untaxed and taxed as well as control and treatment difference figures). Unfortunately, it is not possible to ascertain whether the DDD coefficient is statistically significant only because there is a level difference between the grocery goods difference figure and the department goods difference figure or not.

Table A6.1 — Purchase Size Effect of Tax-Inclusive versus Tax-Exclusive Prices: DDD Analysis of Mean Quantity Sold

Goods	Control (VAT)	Treatment (RST)	Difference
Panel A. Grocery Goods			
Goods with no Tax (0%)	8.890	8.468	-0.423
	(.424)	(0.391)	(0.247)
	[648]	[648]	[1296]
Goods with Tax (21%)	5.323	6.088	0.765
	(0.282)	(0.252)	(0.129)
	[648]	[648]	[1296]
Difference	-3.568	-2.380	$\boldsymbol{DD_{GG}}$ **= 1.188**
	(0.320)	(0.208)	(0.362)
	[1296]	[1296]	[2592]
Panel B. Department Goods			
Goods with no Tax (0%)	1.631	1.556	-0.0756
	(0.110)	(0.126)	(0.0278)
	[648]	[648]	[1296]
Goods with Tax (21%)	0.931	1.012	0.0818
	(0.081)	(0.0902)	(0.0320)
	[648]	[648]	[1296]
Difference	-0.701	-0.543	$\boldsymbol{DD_{DG}}$ **= 0.157**
	(0.0646)	(0.0651)	(0.0511)
	[1296]	[1296]	[2592]
DDD Estimate			**-1.031**
			(0.322)
			[5184]

Notes: Each cell displays the mean quantity of untaxed/taxed goods sold per round, under the two forms of taxation: VAT/RST. The standard errors related to the point estimates are in parentheses and the number of observations (each round has 2 – 4 basket-level observations per respective marketplace).

A6.2. AVERAGE MEAN VISUALIZATIONS FOR LOW AND HIGH BUDGETS

Figure A6.1: Mean Share of Taxed Goods by Treatment & Good Type, Low Budget

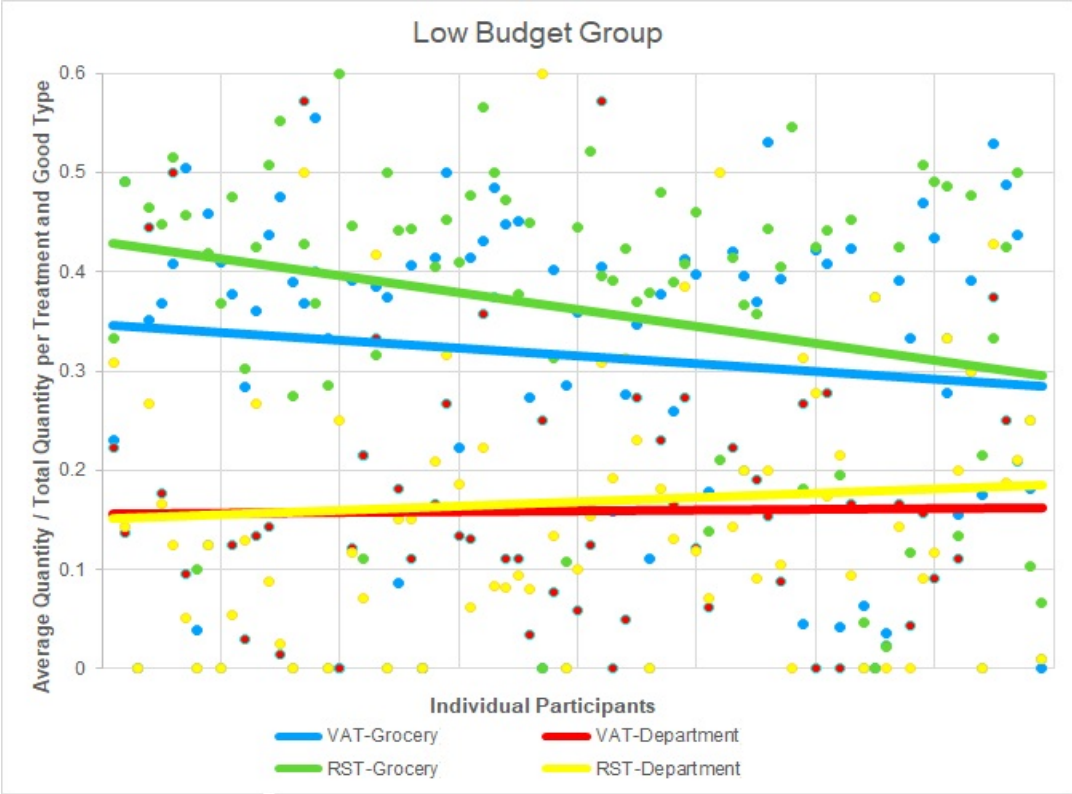


Figure A6.2: Aggregated Mean Share by Treatment and Good Type, Low Budget

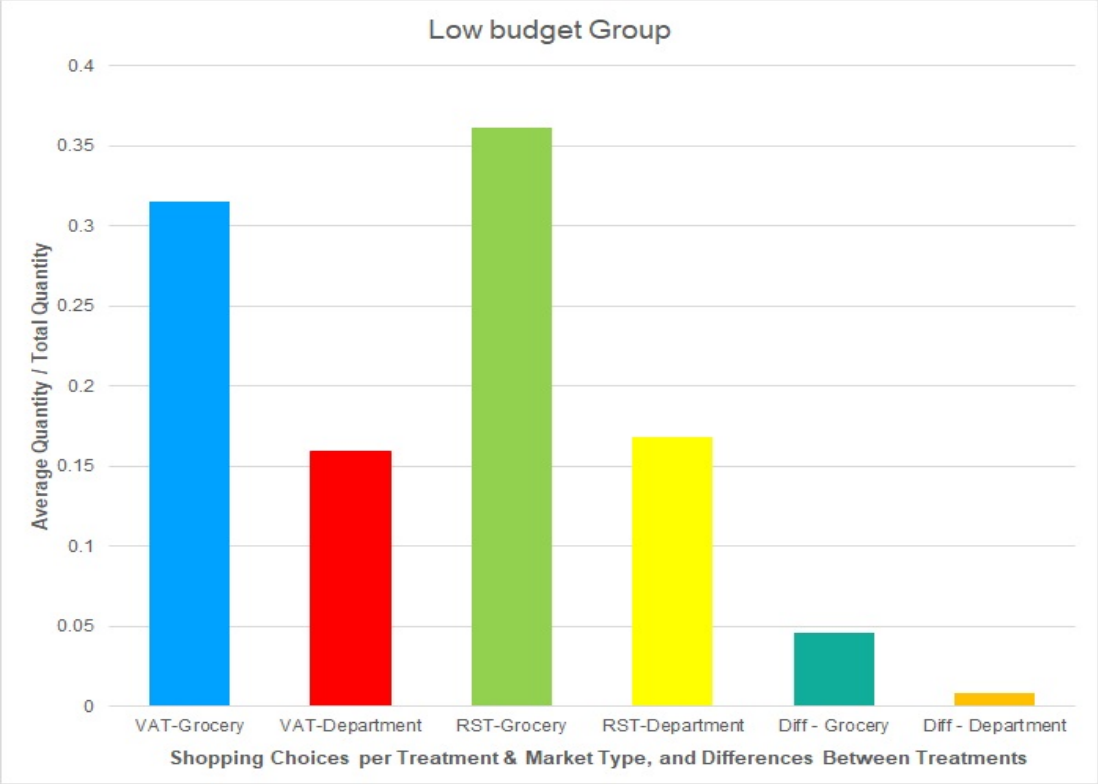


Figure A6.3: Mean Share of Taxed Goods by Treatment & Good Type, High Budget

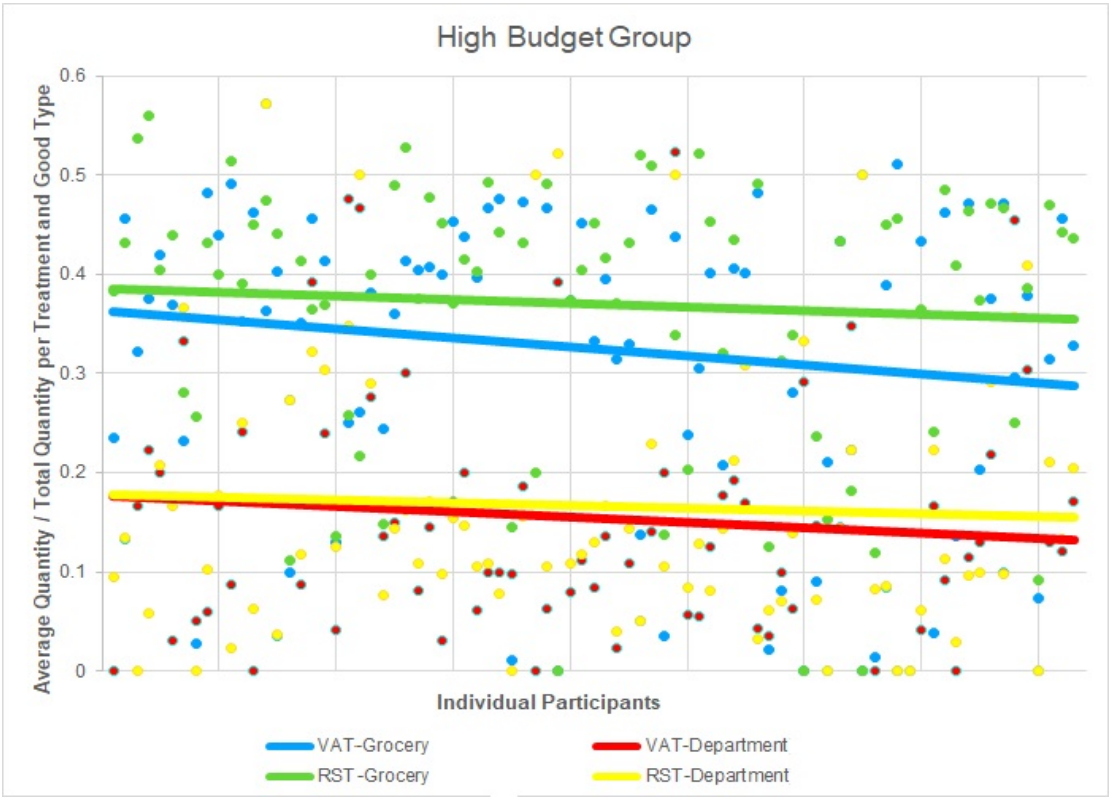
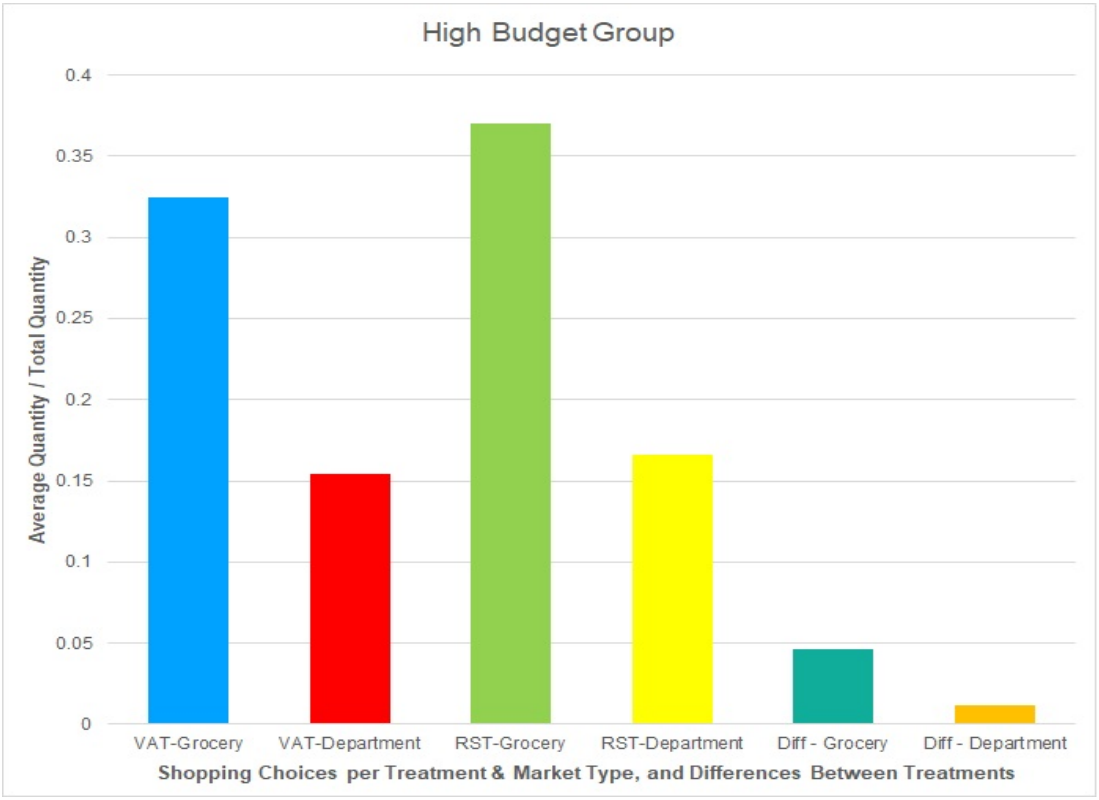


Figure A6.4: Aggregated Mean Share by Treatment and Good Type, High Budget

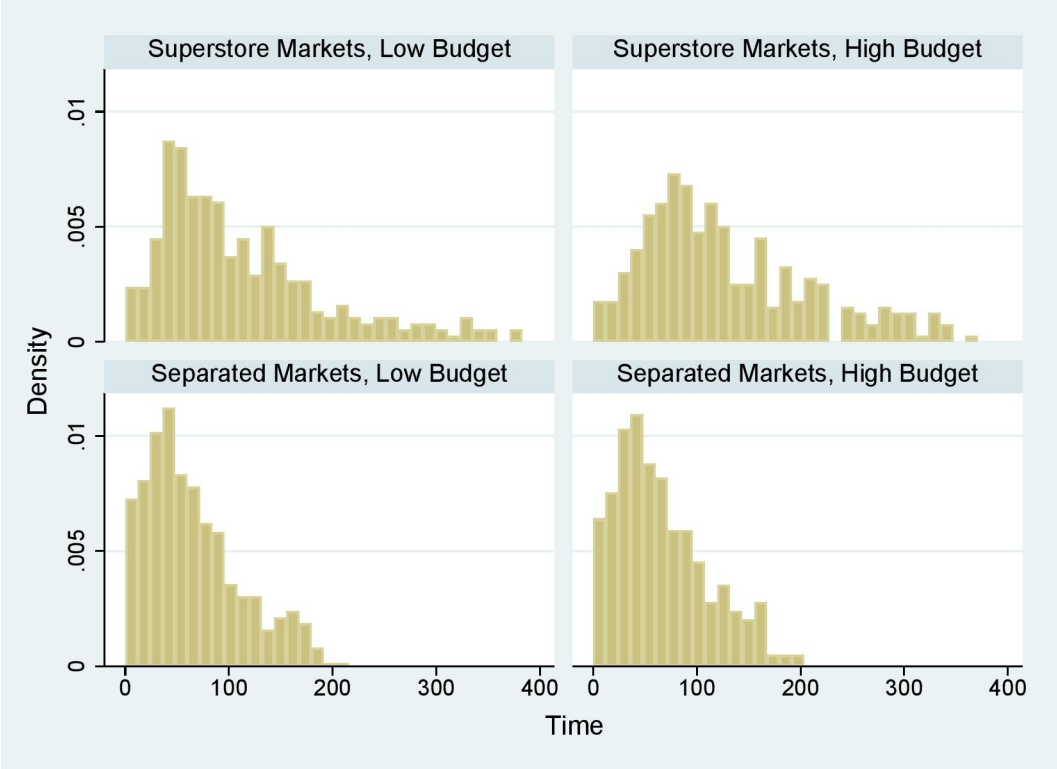


A6.3. SUPPLEMENTAL ANALYSES – DETAILED

A6.3.1. DETERMINANTS OF TAX SALIENCE EFFECT

Of all the supplemental analyses conducted herein, speed of shopping and participation were most strongly correlated with the tax salience effect. For this reason, an extended analysis was conducted. The time cost related to the cognitive cost of calculating less-than-salient taxes is certainly a possible cause or contributor to the tax salience effect. It has been shown that lack of time causes great differences in behavior.[39] The figures below display the distribution of individual round times by types or markets, budgets, and participant responses.

Figure A6.5: Total round time by super/separated market types and budget



[39] For example, the famous seminary experiment conducted at Princeton (Darley & Batson, 1973) showed that time had the greatest affect upon "good Samaritanism", the helping of a person in probable physical need (slumped on the ground); much more so than religious affiliation or religious prompting (even if they are hurrying to go give a talk about the "good Samaritan" topic!).

Figure A6.6: Grocery round time by budget

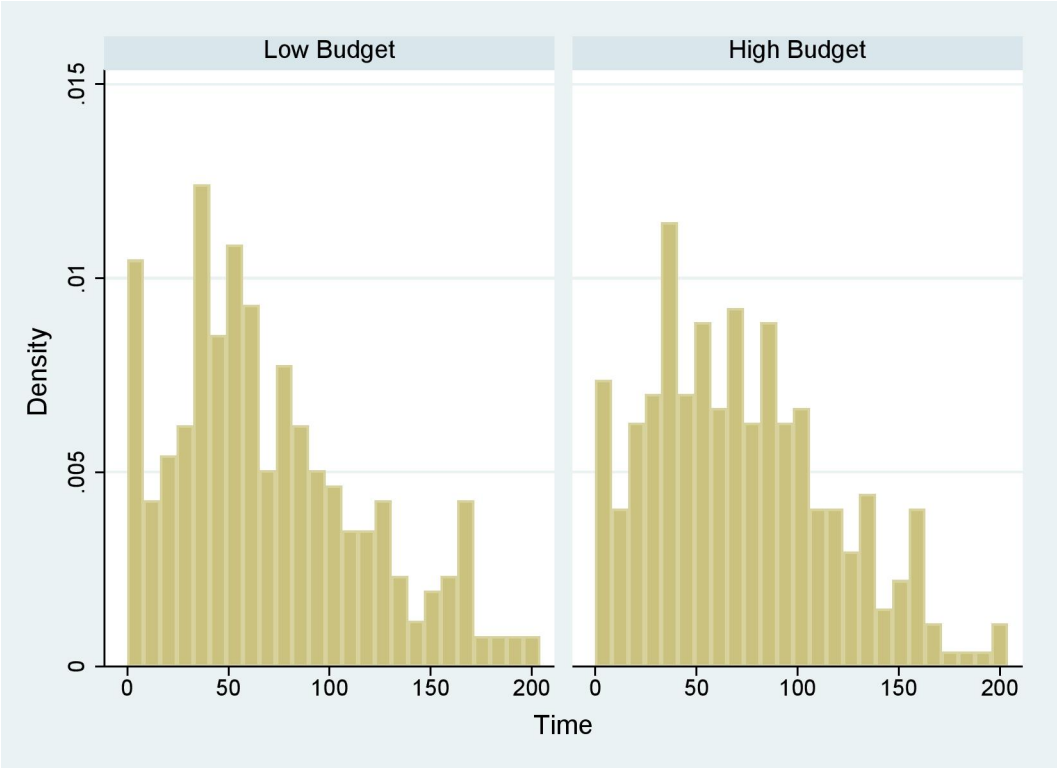


Figure A6.7: Department round time by budget

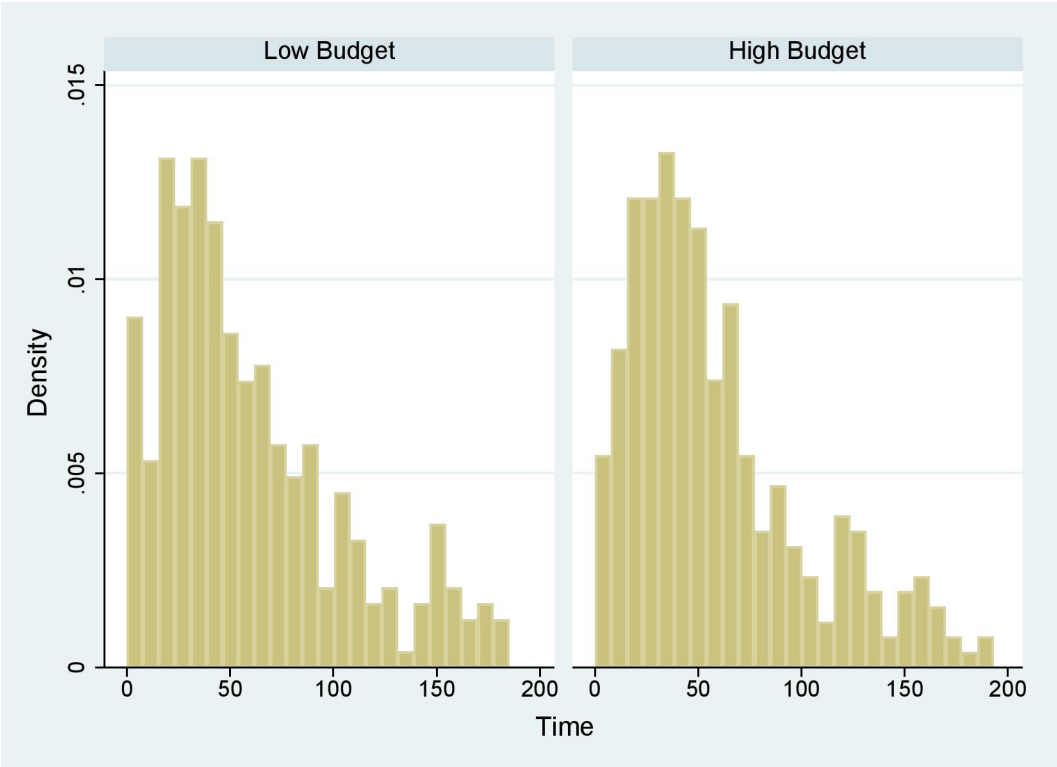


Figure A6.8: Total round time by correct CRT answers

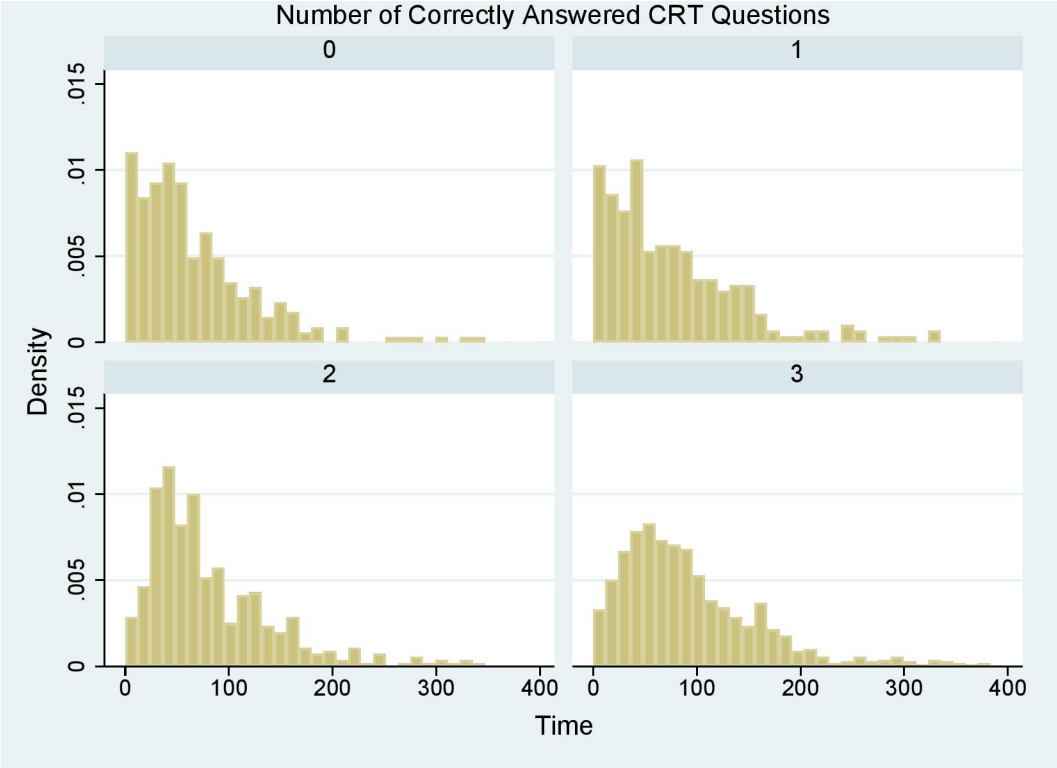
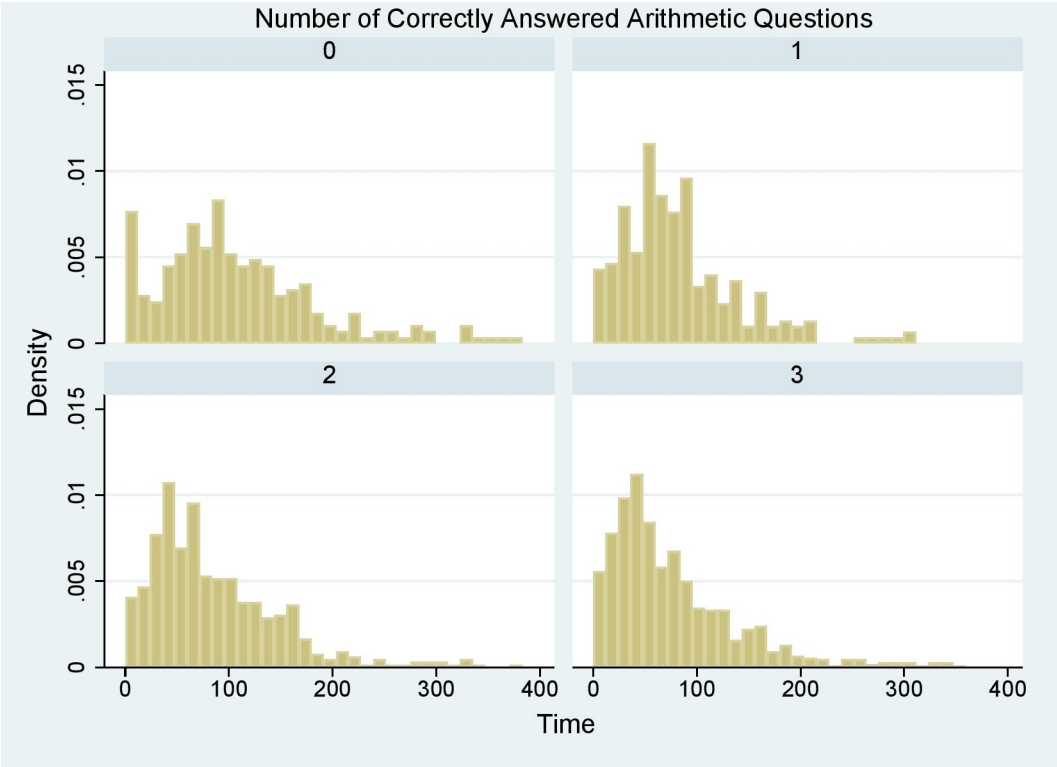


Figure A6.9: Total round time by correct Arithmetic answers



The following information about time/speed characteristics is useful in understanding the analysis results. Time limits were not set up to be constraining to participants, but rather to ensure that total experiment time was not excessive. Indeed, shopping and cart time limits became constraining for less than 1% of rounds for the entire experiment (with the exception of cart time for separated market rounds, where participants reached the last 5 seconds of cart time in 1.5% of rounds). As can be seen from the figures above, the distributions of time spent by both low- and high-budget participants in both the separated and superstore marketplaces is very similar and moderately skewed to the left. This holds for the grocery and department marketplaces when separated out. No distinctively different distribution correlations were found between speed and participant characteristics—especially income, wealth, and education—that were not directly due to small sample size in a given category. One exception is that correct CRT answer distributions were correlated with speed (those that scored lower also shopped faster on average), which echoes Rubinstein's (2016) typology mapping between instinctive versus contemplative behaviors and speed.

Another exception is that those who scored zero correct on the arithmetic questions had a bunching at the fastest speed, which may indicate that there was a small subsample of participants that was either incapable, in terms of arithmetic ability, of calculating the less-than-salient taxes or did not make a serious effort throughout the experiment. This similarity to CRT correlation may imply that Rubinstein's mapping could even be related to arithmetic ability. Removing those that scored zero resulted in marginally greater tax salience magnitude and slightly lower statistical significance on average. When broken down by budget type, there was more of a magnitude increase for lower budget participants and marginally increased statistical significance, while the high budget measures were about the same on average. PSE and tax revenue results remained about the same across the board, with only a small marginal increase in magnitude for the overall and lower budget groups. These differences were not very consequential, but they may imply that the zero-scoring participants paid slightly less attention than their counterparts, which only made a difference for the low budget group. Such an implication lends support to the notion that the low-budget participants had lower tax salience and higher PSE due to a conscious decision to pay more attention (discussed further in Section A6.3.2).

Using three dimensions of speed, the analysis dissects how tax salience and PSE change between faster and slower shopping behaviors. The first dissection looks at shopping time by round. Using the sum of the time spent on the shopping and checkout pages in a given round, all rounds were split into two groups: "slower" and "faster shopping" (the median values were 55 seconds for the separated market rounds—with 63 seconds for grocery rounds and 47 seconds for department rounds, a reasonable difference considering the difference in numbers of units purchased in each—and 99 seconds for superstore rounds). Next, the round times for a given participant were summed up and the participants were then split into two groups: "slower" and "faster shoppers" (the median value was 1182 seconds). Third, the total time between the start of the experiment and the conclusion of the shopping portion of the

experiment was also timed and the participants were also split into two groups by this measure: "slower" and "faster participants" (the median value was 56 minutes for both groups). A full set of analyses (tax salience, PSE, and tax revenue) were conducted for both slower and faster groups in each dimension as well as a heterogeneity analysis for each dimension; all can be found in Appendix A19-A26. Finally, an in-depth examination revealed that 30.86% of participants were below the median shopping speed (fast) during both grocery and department market rounds, 30.86% were both above (slow), 19.14% were fast/slow, and 19.14% were slow/fast. An additional analysis was conducted for this "speed-type" participant categorization.

Under all three dimensions, faster speeds revealed considerable increases in the tax salience effect in terms of magnitude across the board, especially driven by the high-budget participants. This could be a matter of reverse causality, with the choice to not spend the time paying attention to taxes causing the shopping and participation time to be shorter. However, of the three dimensions, "faster shopping" led to the greatest increases in the tax salience effect, while "faster shoppers" and "faster participants" were rather similar, especially for high-budget participants.[40] Repetitive shopping allows a participant to learn from extra time spent on decisions in earlier rounds. Therefore, later rounds are expected to be faster as well as feature more optimal shopping behavior, which would show up in the "faster shopping" observations. That would result in a lower tax salience effect than "faster shoppers". However, "faster shopping" displays more tax salience effect than "faster shoppers" and "faster participants". Combined with the categorized analysis results showing that department market speeds are the strongest driver of the tax salience effect, this suggests that it is probably the shopping speed causing the tax salience effect rather than vice versa.

These results imply that shopping speed itself may be an even more important factor than PSE, because this form of speed seems to reveal the intentional choosing of not paying the cognitive cost, even for the department-price-level goods. In fact, tax salience effect results for grocery goods either only slightly increased or remained about the same, especially for "faster shopping" with high budgets, further corroborating that the increase in the tax salience effect is coming most from the more expensive goods. Alternatively, if it is a loss or lack of learning driving the result, the indication would be that such behavior seems persistent. If such behavior corresponds to the real world, this would imply an overall reduction or erasing of learning effects resulting in a continued dominance of the tax salience effect in the long term. It is even possible that both explanations may be valid at the same time. As it is not possible to confirm either the learning or the speed-changing behavior interpretations herein, they remain in the realm of future research.

[40] High-budget "faster participants" had only a very slightly higher magnitude on average, though low-budget participants clearly had moderately higher magnitude.

While this evidence shows a strong connection between hurried shopping and tax salience, the slower shopping speed tables do still show that tax salience and purchase size effects continue to occur at a lesser degree amongst slower, probably-more-deliberate shopping speeds. The tax revenue analyses exhibit a tax revenue reduction from faster, low-budget participants as well as from slower, high-budget participants.

In terms of PSE, there was an overall diminishing of the effect for higher speed shoppers/shopping and an increase for slower ones. Several results stood out showing low-budget "faster shopping" participants and high-budget "faster shoppers" each displayed reversals of the PSE, while the PSE of "faster participants" was almost unchanged in comparison with the main analysis. A general decrease in the PSE or reversed PSE for faster shoppers/shopping corroborates the above conclusions regarding speed and the tax salience effect. Such results suggest that even low-budget participants continued to not pay more attention to the taxes of the more expensive goods when they shopped faster than their counterparts. Heterogeneity analyses corroborate the evidence above in lesser detail, showing an increase in the tax salience effect mostly by high-budget participants and similar differences in the PSE as described.

The additional "speed-type" categorized analysis that separated participants into four groups by slower/faster shopping and by grocery/department markets confirms that those who shopped fast in both had the most tax salience effect, while slow-grocery/fast-department shoppers had the second most and fast-grocery/slow-department shoppers had no significant results, though the sample size is too small to confirm a null result. This could indicate that some participants are strategically shopping faster in the grocery rounds and slower in the department rounds as an aim to avoid the taxes when the stakes are higher as CLK predicted. Indeed, shoppers who shopped slowly in both marketplaces exhibited the most PSE. Unfortunately, however, the PSE results of the slow/fast and fast/slow shoppers were all insignificant. Either this is a result of a dearth of observations resulting in too low statistical significance, or it may indicate that tax salience and PSE are being caused mostly by shopping speed, with "speed changers" strategically adjusting their time to avoid tax salience errors as the stakes increase.

Are shoppers simply not paying attention to taxes and prices, especially with relatively inexpensive goods[41] and basing their shopping decisions more on other factors or perhaps even out of pure habit? Questions 22 and 23 from the programmed survey directly asked if and why a participant decided to change any of their shopping choices from any one round to another. These questions provide a participant-subjective dimension of determinant rationale. One would likely assume that participants who focused on prices and/or taxes over other factors—such as brand, appearance, perceived/known quality, etc.—would probably show less tax salience effect and/or more PSE. 69.14% responded that they focused on prices and/or taxes.

[41] One could argue that even the more expensive department goods herein could still be considered relatively inexpensive in the real world; some results herein imply that some participants did.

They exhibited a moderate increase in the tax salience effect and PSE on average across all participants and only a marginal reduction in significance on average. Those responders appear to have focused on these elements less for grocery goods, where a moderate increase in the tax salience effect seems to constitute most of the overall increase. This check suggests that those who focus on prices and/or taxes (or believe they did so retrospectively) are taking into account less-than-salient taxes less so for grocery goods and slightly more so as prices increase. This result supports the intentional behavior findings below.

A6.3.2. INTENTIONAL BEHAVIOR, UNDERSTANDING, FAMILIARITY

Directly following the instruction section of this experiment, 5 comprehension questions (see Appendix A1) were asked on important topics such as the shopping environment, shopping budget, experimental earnings, and the difference in taxation systems (which was explained in detail in the instruction section). Despite the considerably fewer observations (67.28% with all 5 questions answered correctly), most of the tax salience effects considerably strengthened while only becoming slightly less significant. The number of statistically significant PSE results and their individual significance values decreased considerably and all but disappeared. Where comparable, the effect remains in the same direction and approximate magnitude. This combination of results signals that those who better understood the instructions (and therefore, probably the experiment as well) paid less, rather than more, attention to less-than-salient taxes and, in comparison, increased their attention to the taxes on more expensive department goods at a similar or lesser degree than their counterparts. Such deliberate behavior may indicate that consumers are consciously disregarding (or regarding to a lesser extent) the taxes, especially for the lower price-level goods.[42]

Other experiments in this branch of literature have asked participants about familiarity with local consumption taxes. This is less of a concern for this experiment as it does not focus on tax rates, it is conducted in a country with a fully-salient consumption tax system, and the tax rate was borrowed from the local market with the intent to enhance external validity. Nonetheless, a random subsample of participants was asked if they knew Czechia's main VAT rate when they came to collect their earnings.[43] While Czechs were familiar with the local tax system, participants from other countries were not as familiar with the exact details. Therefore, an extra robustness check was run that kept only Czech participants, who accounted for 74.48% of the participants. They exhibited a slight increase in tax salience effect magnitude, a slight decrease in PSE magnitude, and a slight decrease in statistical significance across the board. This check indicates no impact on the overall results in terms of familiarity with tax rates or Czech origin.

---

[42] The fact that the tax salience effect and PSE magnitudes increased between the unabridged and abridged results further implies that the effects are more related to intentional behavior rather than some coincidental error.

[43] A lab colleague pointed out that the experiment's many survey questions lacked this simple control question. Thus, it was supplemented through ex-post interviews during remuneration collection.

All in all, these analyses suggest that the behavior of ignoring of less-than-salient taxes is probably intentional in nature and is not the result of a lack of understanding or familiarity with the tax system or the shopping environment, which further qualitatively supports the CLK bounded rationality model of the tax salience effect.[44] These results in combination with the above conclusions further support the CLK bounded rationality model's logic in regard to the existence and rationale of the PSE, the innate progressivity potential of less-than-salient taxes, and the need for further research directed at these determinants and heterogeneity.

A6.3.3. COGNITIVE REFLECTION AND COST

Questions 3-5 of Appendix A3 constitute the Cognitive Reflection Test (CRT), a succinct test developed and introduced by Frederick (2005) to capture an individual's ability to reflect and think logically/rationally versus not reflecting and answering intuitively. One may presume that cognitive reflection would result in a lowering of the tax salience effect in general or in an increase in the PSE if consumers "rationally" choose a cognitive cost threshold (i.e. a strategic attention or debiasing point) that is between the grocery and department good levels. Tabulating the results of the CRT exam found a rather large difference between the number of participants who answered all three correctly and those who answered at least two of the questions correctly. Consequently, the analysis was conducted with both sets of participants.

A total of 48.15% of the participants answered all three CRT questions correct and had reduced statistical significance. These participants exhibited a moderate increase in tax salience effect on average driven more so by low-budget participant point estimates. The PSE results became rather insignificant, but where comparable, there is a moderate increase in grocery good tax salience and a marginal increase in PSE magnitude. When only removing the participants who answered at least two CRT questions correctly (72.22% of the observations remained), the tax salience effect results were still stronger than, yet much closer to, the main analysis on average. Regarding PSE, the effects were actually more pronounced, both in terms of significance and a moderate increase in magnitude than with all three correct. In fact, where comparable to the main analysis, there was an increase in significance for low-budget participants, but the magnitude only very slightly increased or remained about the same on average. These results imply that the participants who scored higher on the CRT exam may have indeed been rationally choosing a cognitive cost threshold somewhere between the grocery and department goods levels. That is, these participants chose to account for the total costs for less-than-salient taxes for more expensive department goods more so than for grocery goods, where the

[44] This kind of intentional neglect could stem from an indifference to the additional monetary cost the taxes amount to or a reliance on the automatic calculation of the taxes provided on the checkout page. Both are directly related to the avoidance of cognitive cost—it could not be that the participants are actually paying the cognitive cost and still choosing to purchase as they do, because then they would make the same (i.e. indistinguishably similar) shopping choices as in the control rounds.

cognitive cost would be proportionally much greater as predicted in the CLK bounded rationality model.

A similar situation occurred with the results of the arithmetic test (question 6 on the paper survey). A total of 46.30% answered all three questions exactly correctly. Relaxing the threshold to a distance of 1.00, a total of 67.90% answered "close" to all three questions. By supposition, those better at arithmetic face a lower cognitive cost in calculating the total price paid for an RST taxed good. As with CRT, that would likely result in such participants having a lower overall tax salience effect or a larger purchase size effect if they debias between the two price levels. Arithmetic productivity (ability) could conceivably cause a lower debiasing point than CRT rationality due to its direct relation to the involved cognitive cost. Participants who answered all three correctly exhibited a moderate increase in tax salience effect, only a marginal increase in grocery good tax salience effect, and only a slight decrease in statistical significance on average. The PSE results had almost no statistical significance. Results from participants who were "close" to all three answers followed the same trends as those who answered all three correctly, except the differences were not quite as large and were more driven by high-budget participants (corroborated by the heterogeneity analysis results; see section A6.4 and the arithmetic table in Appendix A20). The loss in significance in the PSE seems to be more related to an increased tax salience effect for all goods, rather than for mostly grocery goods. This evidence implies that the participants who were more arithmetically proficient were less inclined than their counterparts to pay the cognitive cost at either price level. Perhaps this means that these participants had a higher debiasing threshold than the department goods price level, or that the above predictions were simply incorrect, or it may be due to another unrelated reason. In any case, it is not possible to disentangle here with the data from this experiment.

Finally, an extra analysis was conducted that examined how those who scored highly on both the CRT and Arithmetic tests behaved. Remarkably, the 24.69% of participants who answered all CRT and Arithmetic questions correctly had considerably greater tax salience effect coefficients. While magnitudes most increased with low budget participants, statistical significance was not much higher than for the groups that had all CRT or all Arithmetic questions correct individually. High budget participants also had considerably increased magnitudes, though not as much as the low budget participants, yet did have moderately increased statistical significance. Those who had at least two CRT questions answered correctly and were close (within 1.00) to all three Arithmetic answers represented 48.77% of the participants. These participants had moderately greater magnitudes of tax salience effect for the overall and low budget groups than either group individually. The high budget group had only slightly higher magnitude. Statistical significance was about the same across the board, with slightly increased significance for the low budget group. The PSE almost does not show up significantly for the all-correct group. There is one measure for the high budget group with higher magnitude, which could indicate a greater effect of higher stakes. For the close scoring group, it shows up with moderately greater magnitude somewhat more for overall and

high budget groups, but not very often. These results seem to bolster the previous conclusions in this section: higher scoring individuals seem to be choosing to strategically not pay attention to the taxation despite their likely lower cognitive costs. Again, this may be due to a higher debiasing threshold, utility stakes, or some other reason that is not identifiable herein.

A6.3.4. BUDGET USAGE AND CONSTRAINTS

The following graphical depictions show budget usage by marketplace and participant budget level. They feature a small spike at the zero budget-used point and a large bunching towards the full budget-used area, especially for low budget participants.

Figure A6.10: Percentage of budget used by super/separated market types and budget

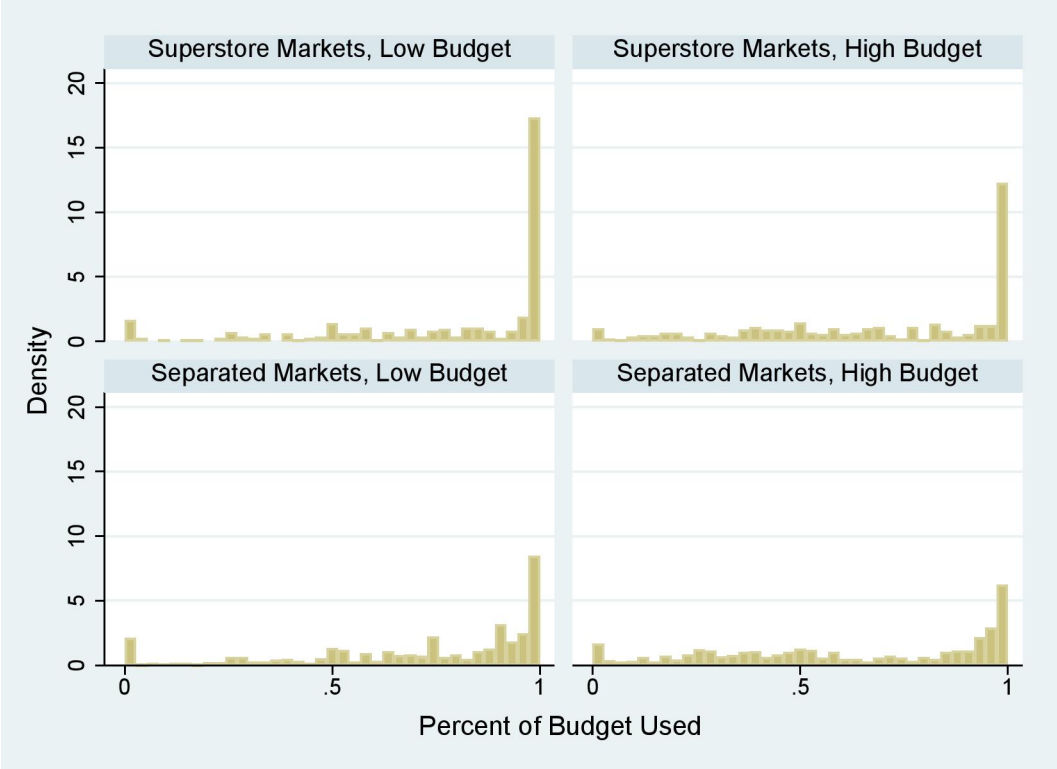
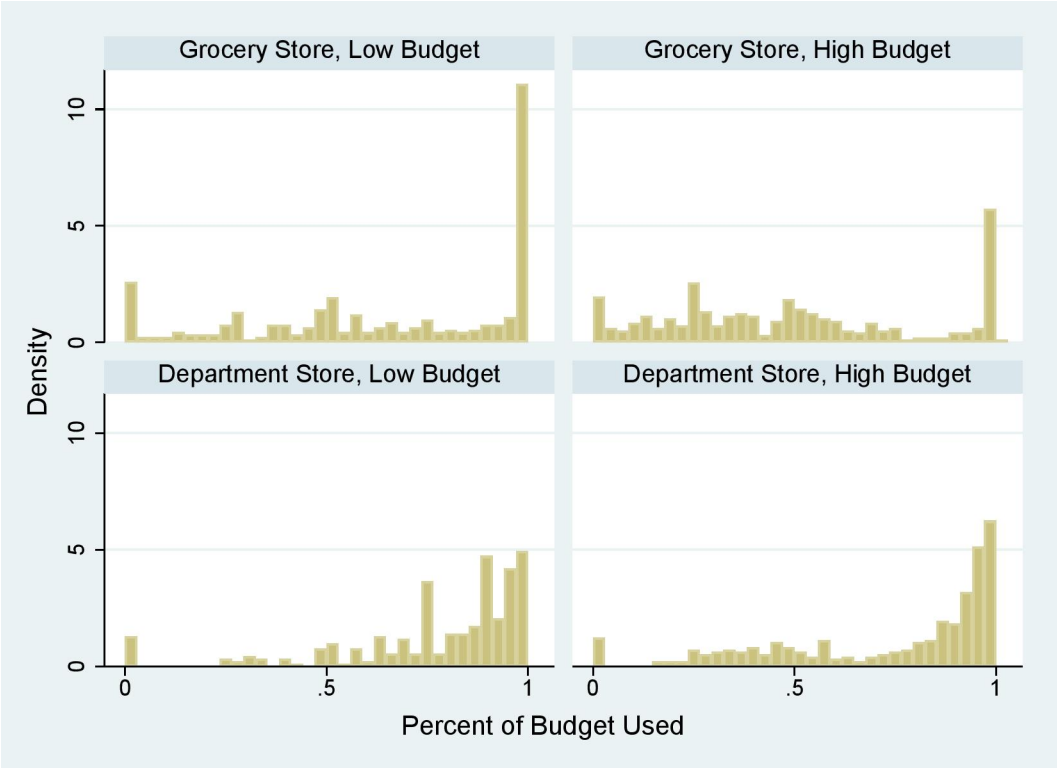


Figure A6.11: Percentage of budget used by grocery/department market types & budget



To explore how budget constraints and usage may have affected the results herein, rounds were divided into halves by the median budget usage percent for each budget type and amount (CZK 400, CZK 800 low budget, CZK 800 high budget, and CZK 1600). The resulting median amounts were 82.00% for CZK 400, 95.88% for CZK 800 low, 62.25% for CZK 800 high, and 76.34% for CZK 1600. The analyses results were actually rather similar in terms of magnitude and statistical significance between the higher- and lower-budget-usage groups, and to the main analysis. In particular, the results of the individual variable measures were sometimes slightly higher/lower in comparison with the main analysis, but there was little consistency in the direction of those effect differences within and amongst the two budget usage groups.

Only the heterogeneity effect analyses provided a coherent result that showed a decrease in the tax salience effect and PSE for those who used less of the budget. This difference was mostly driven by low-budget participants, which is not surprising when taking into consideration the substantially different median amounts between low- and high-budget participants. Most probably, these differences are related to the binding budget constraint issue explored below. Such a decreased tax salience effect and PSE may imply that those with lower budgets more strategically used their more limited budgets to obtain the most out of their experimental endowment. Such an implication would reinforce the findings of this paper, which are mostly driven by high-budget participants. Furthermore, such strategic shopping may point out why the effects were so dampened for the low-budget participants, which would further suggest that real-world effects of tax salience and PSE would be greater than those found herein.

An additional analysis was conducted to explore and compare the behavior of the high budget participants in the higher-budget-usage group that remained once removing the highest budget users in the group (Appendix A25). This was accomplished by dropping the highest budget users per the much higher median values of the low budget participant group (82.00% usage for separated marketplaces and 95.88% usage during superstore rounds). These amounts left 59.64% and 65.76% of high budget observations, respectively. The tax salience effect results were slightly greater in magnitude, though slightly lower in statistical significance. The PSE results were moderately stronger and tax revenue was considerably lower. This indicates that even some higher budget participants were strategically shopping to maximize their endowment per the higher-stakes hypothesis. The remaining higher budget participants may represent the most realistic behaviors (i.e. in relation to consumers in the real world facing such price-level goods) as they were most likely not shopping to maximize their endowment and they were not constrained by a, perhaps, overly tight budget. These remaining participants were paying less attention to the taxes at the grocery level and more attention at the department level, further supporting previous findings. This analysis helped to expose the influence of the higher stakes effect and reinforce the notion that tighter budget constraints may be distorting results in other laboratory experiments on tax salience.

Question 2 on the paper survey asked participants if they were ever not able to purchase any quantity of goods that they were able to purchase in a different round solely due to an increase in price from taxes (causing a binding budget constraint). While the majority of the "yes" responses confirmed that this concern was much more related to department goods than grocery goods, to properly compare behavior when not bound by budget constraints caused solely by tax changes, all participants who responded that they faced such a constraint during grocery, department, or superstore rounds were dropped, leaving 46.91% of observations for this check. There was a moderate decrease in statistical significance across the board, especially for low-budget participants (who were more likely to be affected by budget constraints), and, on average, the magnitude of the tax salience effect moderately decreased. PSE significance nearly disappeared. In one comparable measure, the PSE moderately increased in effect magnitude for high-budget participants. Perhaps due to the influence of the statistically insignificant low-budget-participant point estimates, the all-participant results had two significant point estimates of purchase size in the opposite direction. Nonetheless, by and large, this check implies that tax salience effect and PSE are naturally occurring based on shopping choices and are clearly not due to an experimental design with artificially imposed budget constraints.

A6.3.6. INTERACTION OF SPEED AND BUDGET USAGE

Since shopping speed and budget usage were the two most consequential variables, an additional analysis was conducted that examined their interaction (Appendix A26). Higher budget users who were faster shoppers represented 17.13% and slower shoppers 33.02% of the

observations. Lower-budget users who were faster shoppers represented 32.87% and slower shoppers 16.98% of the observations. It is clear from the tables that lower-budget users who shopped slower featured the least amount of tax salience effect, especially for low budget participants. Faster-shopper, lower-budget users had a much stronger tax salience effect (both in terms of magnitude and statistical significance) than slower-shopper, higher-budget users; even more so than faster-shopper, higher-budget users; and especially so for high budget participants in both comparisons. This may indicate that it is speed (or the related lack of attention) combined with budget constraints more so than the overall spending behavior that are most related to the tax salience effect. This seems to reinforce the discussions above about the two opposing effects on participant behavior: both effects appear to be occurring simultaneously, but budget constraints seem to intensify one over the other.

The PSE results are weak in terms of statistical significance and, with such sample sizes, often inconclusive. According to the results, it appears that higher-budget usage increased the PSE for low budget participants who shop faster. These participants seem to be maximizing both in terms of spend and tax avoidance, while their slower-shopping-higher-budget-using counterparts have only a few PSE results going in the other direction, which would indicate that these shoppers were maximizing only spend and utility and ignoring taxes. At the same time, lower-budget usage seems to increase the PSE for high budget participants, especially for the faster shopping ones, which once again indicates a dominance of strategic inattention due to budget constraints over potential utility stakes. The only significant tax revenue results are for the low-budget-using-slower shoppers, who had reduced tax spending compared to all their counterparts. All in all, this analysis seems to reinforce previous interpretations and potentially indicates a slightly more detailed perspective on how the discussed incentive forces may be working.

A6.3.7 GOODS PREFERENCES

Although participants were able to choose 10 goods out of 50 from each marketplace, the concern remained that all or most of the 50 goods available could be out of their preferential space. Therefore, questions 16 through 19 asked participants how many of the grocery and department goods they were interested in, and seriously interested in, purchasing. A significant majority of participants were interested in 1-6 of the 50 department store goods and 4-10+ of the 50 grocery goods. They were also seriously interested in 2-5 out of their chosen 10 department goods and 3-7 as well as all 10 of their chosen grocery goods. Interest in a lower number of department goods was expected and not a problem (as long as they were interested in at least two or more goods), because only a small quantity of them were affordable under the budget constraints the participants faced. Dropping all participants who answer 0 or 1-3 to questions 16 and 17 as well as those who answered 0 or 1 to questions 18 and 19, resulted in 51.23% of participants remaining for the general “interest in goods” analysis. One would expect that when more of the goods are in the participant’s preferential space, substituting from

taxed goods to untaxed goods would involve a lower opportunity cost and/or increase their strategic allocation of attentiveness for those goods, resulting in a lower tax salience effect and/or higher PSE. Nonetheless, the tax salience effect in general and only for grocery goods increased in effect magnitude and only slightly decreased in statistical significance. This breaks down to a considerable increase for low budget participants combined with a slight decrease on average for high budget participants (except in grocery goods, where the effect for high budget participants remained about the same or had a marginal increase). In fact, the low budget participant increase was so large that they were now exhibiting a moderately greater magnitude of tax salience effect on average than the high budget group.

In terms of PSE, low budget participants displayed a moderate increase in average effect magnitude with the same or marginally greater statistical significance even with the major decrease in observations. High budget participants hardly changed at all except for some small mixed movements in significance. The most likely explanation for this is that the expected increase in substitution towards untaxed goods did occur in both marketplaces, especially for low budget participants. However, the desire for more variance amongst grocery goods (where many could be purchased) outweighed the increased cost from the taxes, whereas only a small number of department goods could be purchased in a given round and the opportunity costs are much smaller. Indeed, a comparison of means showed that this group of participants purchased more goods on average with a large portion of that increase going towards the purchase of untaxed goods of both types; but, of course, with a greater increase of taxed goods of both types under treatment. With regard to strategic attentiveness, participants may have either chosen to simply disregard the less-than-salient taxes and "save" on cognitive costs or to consciously evaluate the cost-variance benefits and purchase those desired goods deliberately despite the extra tax cost. Unfortunately, that is not discernable from the data from this experiment. Nevertheless, as the comparison of means identifies only a minor difference between the main group and this subset of participants, it points to the latter option.

Another area that could affect tax salience and confound the PSE is the positive value and preferential association people have with brands. To combat this potential issue, all brand names were removed from item descriptions (as mentioned in section 1.3.3) and inconspicuous or generic brands were chosen for the vast majority of the goods. However, item pictures, though small, were not edited or "debranded" in any manner, so participants may have been able to recognize a preferred brand from the item image. Questions 20 and 21 on the programmed survey and question 1 on the paper survey controlled for this issue. The former questions inquired about how many strongly preferred brands they had out of their 10 chosen goods of each type, while the latter question asked about any brand recognition. For both of the following analyses, only those participants who answered "0" to both 20 and 21 or those that did not recognize any brands in question 1 remained in the analysis. The tables labeled "No strong brand preference" represent the former check and constitute 68.52% of all participants. Tables labeled "No brand recognition" represent the latter check and constitute 53.70% of all participants. The latter saw a considerable reduction in statistical significance in

the tax salience effect and a marginal increase in purchase size effect significance. Effect magnitude of tax salience slightly decreased on average with high budget participants exhibiting a larger decrease and low budget participants remaining about the same or even exhibiting a marginal increase on average. PSE magnitude displayed exactly the opposite change as that of tax salience. The "No strong brand preference" check was not as dramatic, with a tax salience effect that moderately increased, which mostly came from low budget participants, as well as a slight increase in the PSE on average. In total, it is clear that brands do consequentially affect shopping decisions, tax salience, and the PSE. A lack of brand influence seems to have lowered the opportunity cost of substituting goods, thus reducing overall tax salience effects and increasing the PSE, especially for lower budget participants.

A6.3.8 SITUATIONAL CONDITIONS

Questions 6 and 7 from the programmed survey were conceived to assess how a participant's familiarity with prices and shopping as well as hunger could affect the tax salience [purchase size] effect. One would likely assume that familiarity would likely reduce [inflate] the effect due to greater savvy from experience. Hunger has been known to affect shopping choices such that more food is generally purchased and with less attentiveness to prices (Gilbert et al., 2002). Such an effect could result in more purchases of grocery goods with less care, thus increasing [increasing] the effect. With 81.48% of participants responding that they do all or most of the shopping for their household, there were some marginal changes in statistical significance in both directions throughout the tables (somewhat more for low budget participants, where it mostly increased). Familiarity seems to have affected the magnitudes of both effects very little or not at all on average. In regard to hunger, 66.67% of participants responded that they were not hungry throughout the experiment. When retaining them, there was a moderate decrease in statistical significance amongst all participants. As anticipated, these expectedly less hungry shoppers, on average, display a slight decrease in the tax salience effect for grocery goods, a marginal increase in the overall tax salience effect on average, and a slight to moderate increase in the PSE. All in all, neither familiarity nor a lack of hunger changed the main results in any meaningful way.

A6.3.9 SUBSTITUTES

One of the intents of the design of this experiment was an analysis of how substitutes may affect tax salience and the PSE. Two analyses were conducted and are presented in Appendices A17 and A18. The first analysis examined the tax salience effect within the substitute level. It was conducted within the 25 direct substitute category and the 9 loose-substitute category group levels for the grocery goods as well as within the 9 loose-substitute category (group)

level for the department goods.[45] Most of the results did not show up significantly or consistently, with a few exceptions. The most significant and consistent substitute effect comes from category 22: cream cheese and whipped cream cheese. With slightly lower magnitude and significance was the substitute effect from category 7: red and green pears. Then slightly stronger but less significant was broccoli and cauliflower (category 9), weaker but more significant was sport equipment (category 106), and even weaker and less significant was baby and cherry tomatoes (category 24). Regarding the looser grocery category groups, a weaker effect showed up from dairy (group 206), vegetables (group 203), and fruits (group 202).

The second analysis was a re-running of the main analysis, except with the substitutes removed from the regressions. This was accomplished by only keeping the item within each category with the highest quantity chosen (in order to conserve as much statistical significance as possible). Tax salience effect results exhibited a moderate decrease in statistical significance (except for low budget participants, where it slightly increased), marginal changes in magnitude (usually reduced, except for the standard variables of y and x), and a moderate increase in magnitude and moderate decrease in significance for grocery goods. Results from the PSE tables displayed considerably more significance and magnitude across the board, especially for high budget participants. All in all, it is apparent that some substitute goods do affect tax salience, seemingly as expected, by decreasing the salience effect. Removing the substitutes appears to impact tax salience more for grocery goods than department goods. This could indicate that participants understand which goods are taxed, but are not caring at lower price levels, which further upholds the intentionality and strategic inferences. Alternatively, this could simply be a coincidental artifact of the asymmetrical categories in this experiment. Disentangling this "substitute" effect is impossible herein. Further clarification will remain the domain of future research.

A6.3.10 ROBUSTNESS & PLACEBO CHECKS & MISSING ANALYSES

An "AB" analysis was conducted to assess whether the effects or statistical significance strengthened when the analysis was limited to only the same-taxed-goods group. As described in section 1.3.4, in half of the main control-treatment section, one half of the chosen goods were taxed in two of the periods and the other half in the other two periods. Herein, we dub those groups of periods by taxed goods as "A" and "B". As can be seen from the tables in Appendix A15, the tax salience effect (both average and grocery) figures for group "A" exhibited a moderate increase in both effect magnitude and statistical significance, while group "B" had a moderate decrease in effect magnitude and a considerable decrease in statistical significance (except in the case of the high budget participants). The PSE results were once again mostly insignificant and incomparable. The differences in statistical significance are

[45] Unfortunately, since direct substitute categories became impractical for the department marketplace, the full substitution analysis that had been intended was not possible to conduct.

almost certainly due to the halving of the observations. In total, the "AB" analysis revealed no surprising results between the groups that could call into question the results of the main analyses, nor did it provide any further insight.

A placebo analysis was conducted similarly to the "AB" analysis, except that the "A" and "B" groups replaced the control (VAT) and treatment (RST) groups. Appendix A16 demonstrates that the placebo analysis succeeded in its aim of showing that the results related to the main control-treatment division are not spurious, since for the first time herein, the tax salience estimates are mostly insignificant and in the wrong direction.

Lastly, some comments about missing analyses are important here. First, a continuous version of the PSE analysis was conducted by comparing the differences in point estimates from OLS regressions of the *Choice* variable of log y against the variable of interest, log of price, between the four grocery/department and VAT/RST groups, amongst taxed goods. The results exhibited a purchase size effect, with a 26.97% reduction in grocery goods under treatment and 68.42% in department goods under treatment. Unfortunately, three of four (both department store goods and one grocery store goods) results were statistically insignificant. Second, a robustness check that removed all regression structuring and allowed for full heteroskedasticity was conducted for the overall general case. Statistical significance decreased across the board. Tax salience results were mostly still significant, remaining within 95% confidence levels. The PSE results were lower than the standard significant levels usually adhered to in empirical analyses, with many remaining within 80% confidence levels, which is considered still relevant in experimental economics. Third, a welfare analysis was not conducted, because the designed welfare control period failed to accomplish its aim. While noise was expected, only 9.90% of participants chose the exact same baskets between the base and compensated rounds, and only 22.92% had an absolute difference of two or less, which could have still been used to assess the indirect utility function needed for the "refinement" method.

A6.4. HETEROGENEITY EFFECT

Heterogeneity effect analyses used the difference-in-difference-in-difference (DDD) regression (see regression (2) section 4), replacing purchase size with given dimensions of heterogeneity to assess the effects on tax salience and a DDDD[46] regression—as in Hendricks (2014) and Figlio & Rueben (2001)—to enrich the full DDD purchase size effect analysis, employed above, with the entire additional difference of heterogeneity. The experimental design for the full heterogeneity study associated with this research calls for much more actual

[46] $Choice_{icrgm} = \alpha_i + \lambda_1 HE_i + \lambda_2 Taxed_{ircgm} + \lambda_3 Treat_{ir} + \lambda_4 PS_{ircg} + \lambda_5 HE_i*Taxed_{ircgm} + \lambda_6 HE_i*Treat_{ir} + \lambda_8 HE_{ig}*PS_{ircg} + \lambda_9 Taxed_{ircgm}*Treat_{ir} + \lambda_{10} PS_{ircg}*Taxed_{ircgm} + \lambda_{11} PS_{ircg}*Treat_{ir} + \lambda_{12} HE_i*PS_{ircg}*Treat_{ir} + \lambda_{13} HE_i*PS_{ircg}*Taxed_{ircgm} + \lambda_{14} HE_i*Taxed_{ircgm}*Treat_{ir} + \lambda_{15} PS_{ircg}*Taxed_{ircgm}*Treat_{ir} + \lambda_{16} HE_i*PS_{ircg}*Taxed_{ircgm}*Treat_{ir} + \sigma_{ircgm}$

heterogeneity in the participation pool than was obtained for this experiment. Nevertheless, the following highlighted[47] results provide a glimpse into the heterogeneity effect and its potential impact and implication on tax salience. Full output of the heterogeneity effect analyses is presented in Appendices A19 and A20. The only straightforward and clear result is that the gender analysis returned a virtual absence of difference in the tax salience effect between men and women.

Mixed evidence on income differences indicates that higher individual income is mostly positively correlated with the tax salience effect, but this only became evident as the difference threshold was increased from 100.000Kc to 200.000Kc annual salary, driven mostly by the high-budget participants. However, this reversed in the opposite direction for household income. The magnitude of this reduction became moderately larger and more significant as the household income threshold increased from 400.000Kc to 500.000Kc annual income. Differences in PSE were nearly nonexistent for the individual income measures and slightly mixed in direction, but overall may indicate a small increase in the PSE as the threshold increased. For household income, the PSE decreased for low budget participants, increased for high budget participants, and the difference all but disappeared for both groups as the threshold increased.

The tax salience effect appears to increase with wealth experienced during youth, driven almost exclusively by high budget participants, though the differences all but disappeared as the threshold of wealth increased. Higher current wealth seems to have nearly no impact on the tax salience effect. The PSE is reduced, and considerably so, by higher youth-wealth individuals, though the differences became all but insignificant as the threshold increased. Higher current wealth participants mostly displayed an increased PSE, which all but disappeared and even changed direction in one point estimate above the threshold. Altogether, these income and wealth related outcomes seem to imply a difference in an individual's debiasing threshold that may be more related to the relationship with money developed during their youth.

More education is correlated with a somewhat reduced tax salience effect as well as increased PSE. However, as the educational threshold was raised from attending some undergraduate

---

[47] Probably the most important aspects of heterogeneity that can affect social welfare through tax salience are income, wealth, and productivity (through the proxies of education and arithmetic proficiency). This is because any politically-viable, real-world tax system has to adhere to a legal framework that intrinsically limits who, how much, and by what measures a person may be taxed. Moreover, the average taxpayer is motivated to avoid taxes by any legal means necessary (sometimes even by illegal means). Optimal tax theory often ignores this in favor of economic, mathematically-based theoretical concepts, such as productivity, that are impossible to accurately measure for tax purposes in the real world. Instead, real tax systems need to rely upon a person's (or firm's) actual monetary holdings and flow. Income and wealth are direct measures of this kind. The productivity proxies mentioned above also tend to be directly, positively correlated with income and wealth. The analyses were conducted by setting dummy thresholds along median answers to the programmed survey questions. Measures for wealth, income, and education had two median answers and analyses were conducted on both and compared. Furthermore, questions were asked (and analyses conducted) about both personal and household income levels as well as proxies of wealth from youth and the present.

higher education to having completed a bachelor's degree, the tax salience effect differences all but disappeared and the PSE difference considerably weakened and became moderately less statistically significant. The arithmetic proficiency measure showed an increase in tax salience effect for those who scored higher on the arithmetic question in the survey, mostly driven by high budget participants. When the threshold was raised from answering three questions closely to answering all three correctly, the increased tax salience effect remained, but it moderately weakened and became slightly less significant, especially for high budget participants. This may indicate that participants are choosing not to pay the cognitive cost (debias) as a matter of strategic attentiveness, informed by one's budget constraint. Differences in PSE were essentially insignificant across the board.

In total, bearing in mind the heterogeneity limitations noted above, the evidence seems to point to regressivity in a less-than-salient tax system by education and progressivity for wealth and income (except at the household level, though its reduced PSE may imply that it still points towards progressivity in aggregate through a higher debiasing threshold). The evidence is mixed, especially amongst the income and wealth measures that are directly connected with the cash flows of income and returns on capital and, therefore, tax revenue and its impact on social welfare. Hence, the potential, consequential existence of an inherent progressivity demands further exploration on the topic of the heterogeneity dimension of the tax salience effect.

A6.5. ADDITIONAL PARTICIPANT DISTRIBUTION FIGURES

Figure A6.12: Shopping time by super/separated market types and budget

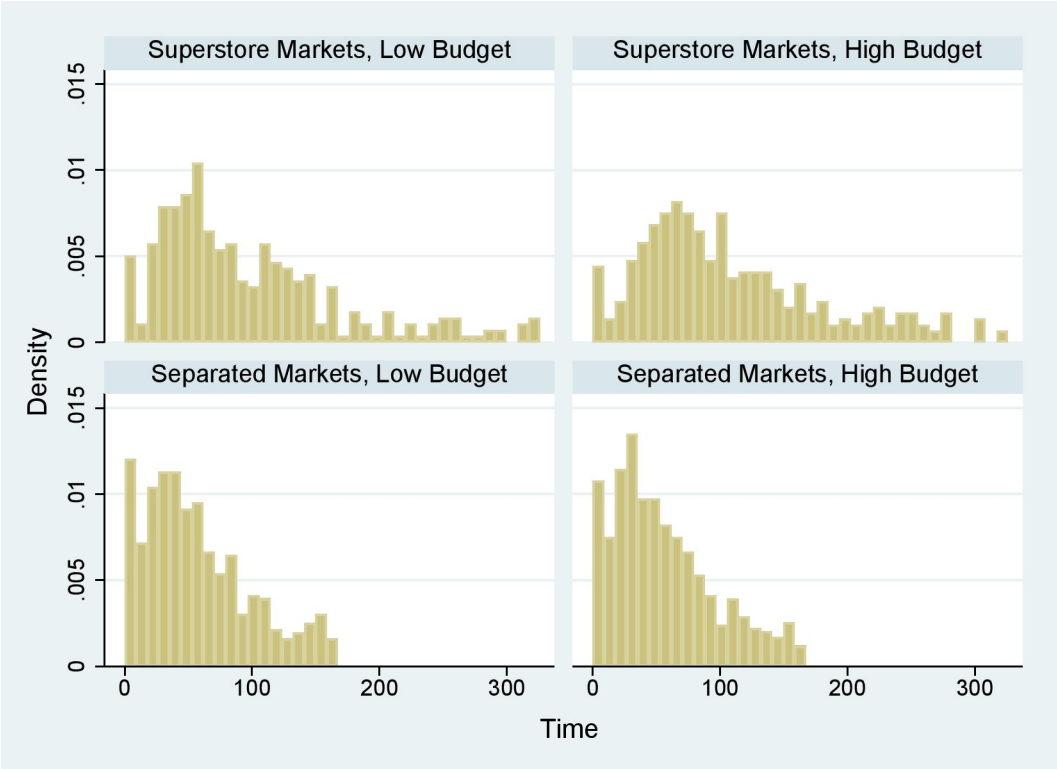


Figure A6.13: Cart time by super/separated market types and budget

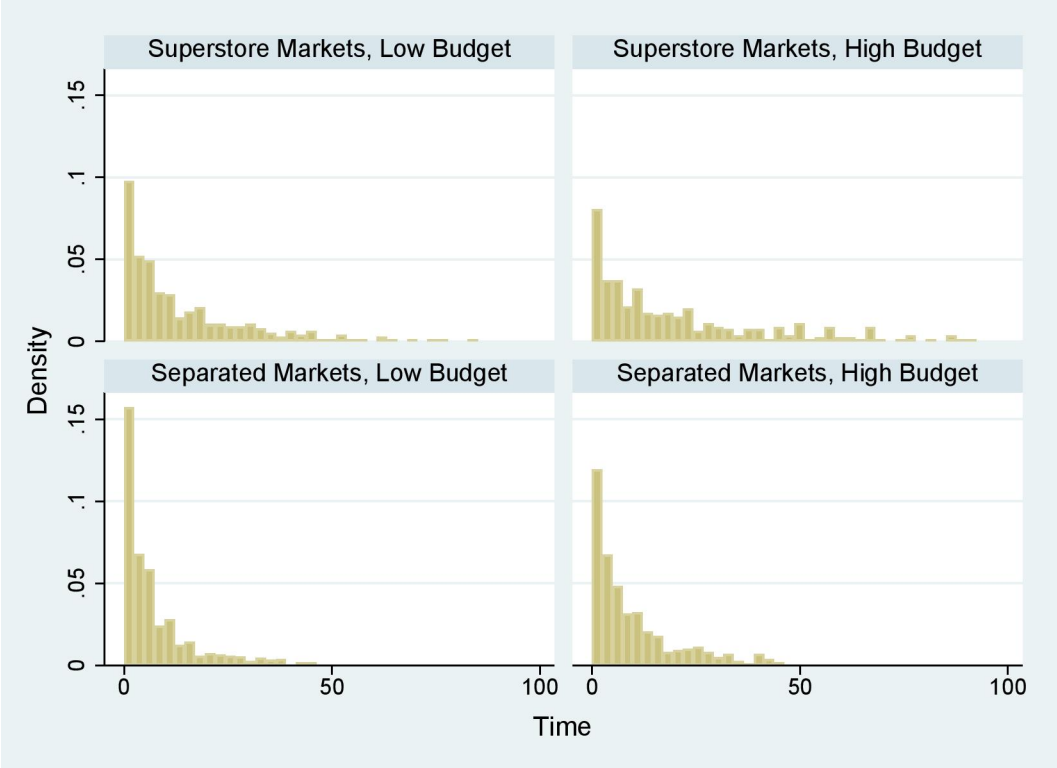


Figure A6.14: Grocery shopping time by budget

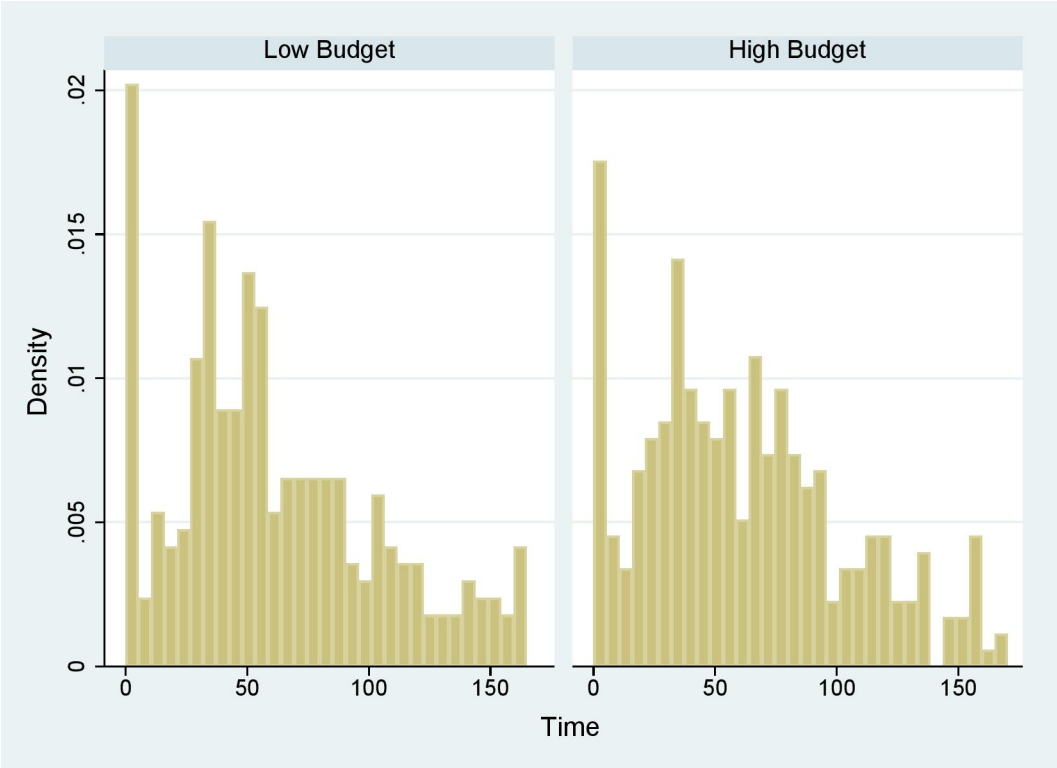


Figure A6.15: Grocery cart time by budget

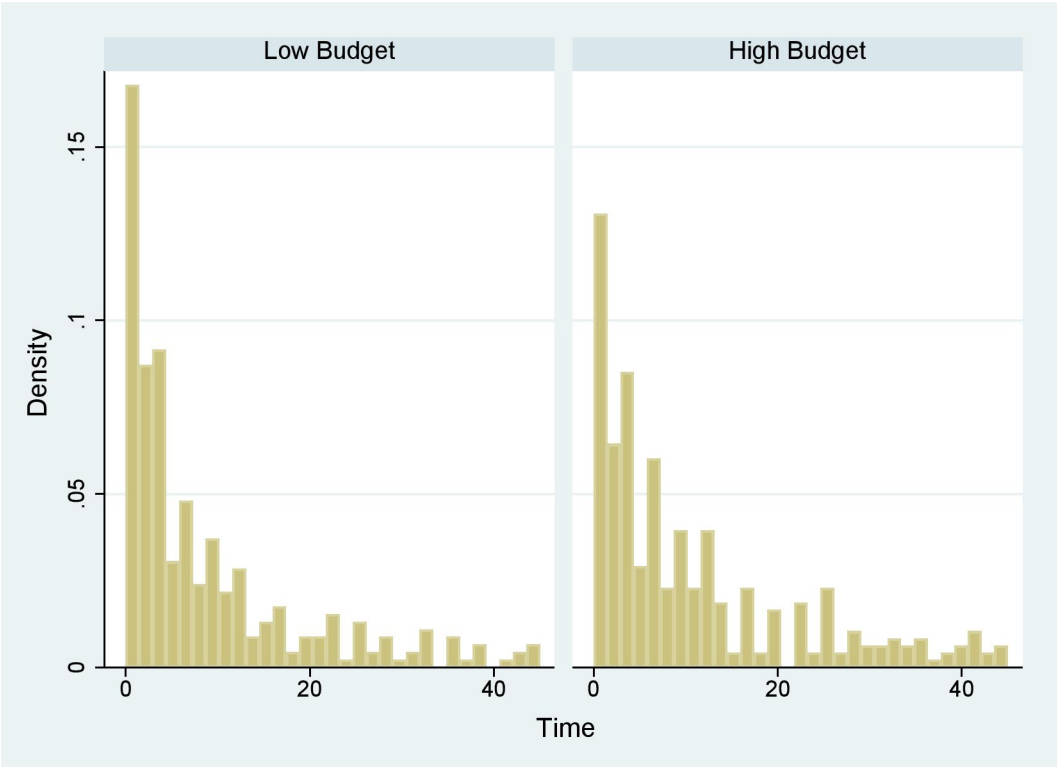


Figure A6.16: Department shopping time by budget

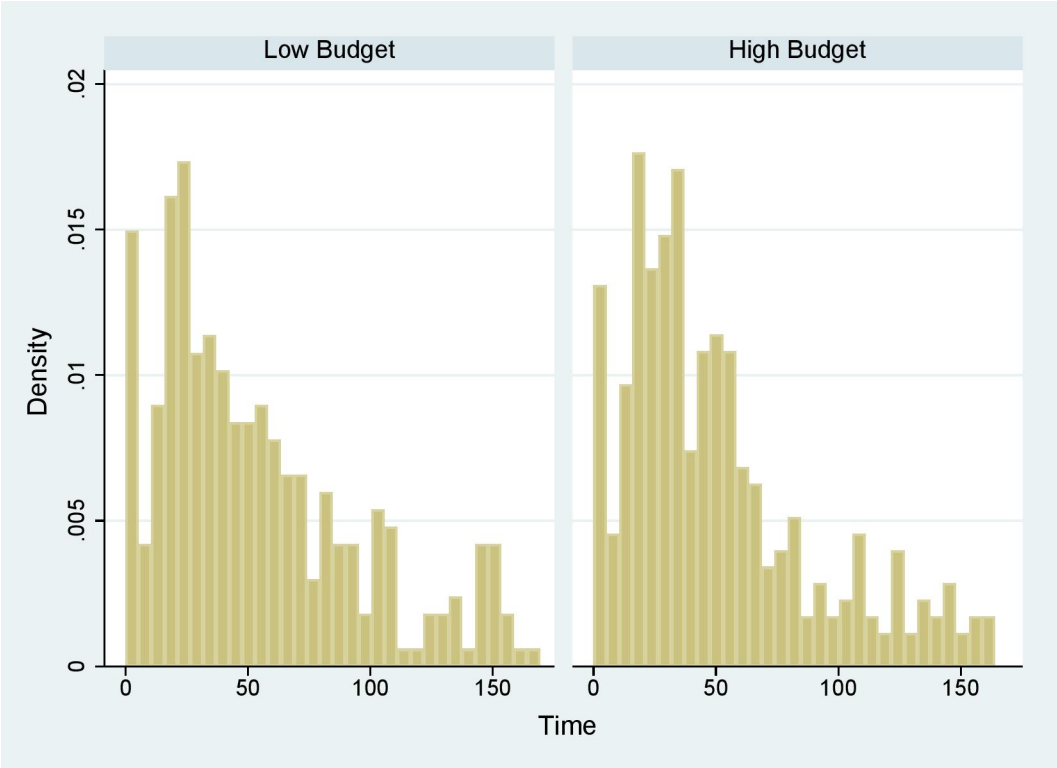


Figure A6.17: Department cart time by budget

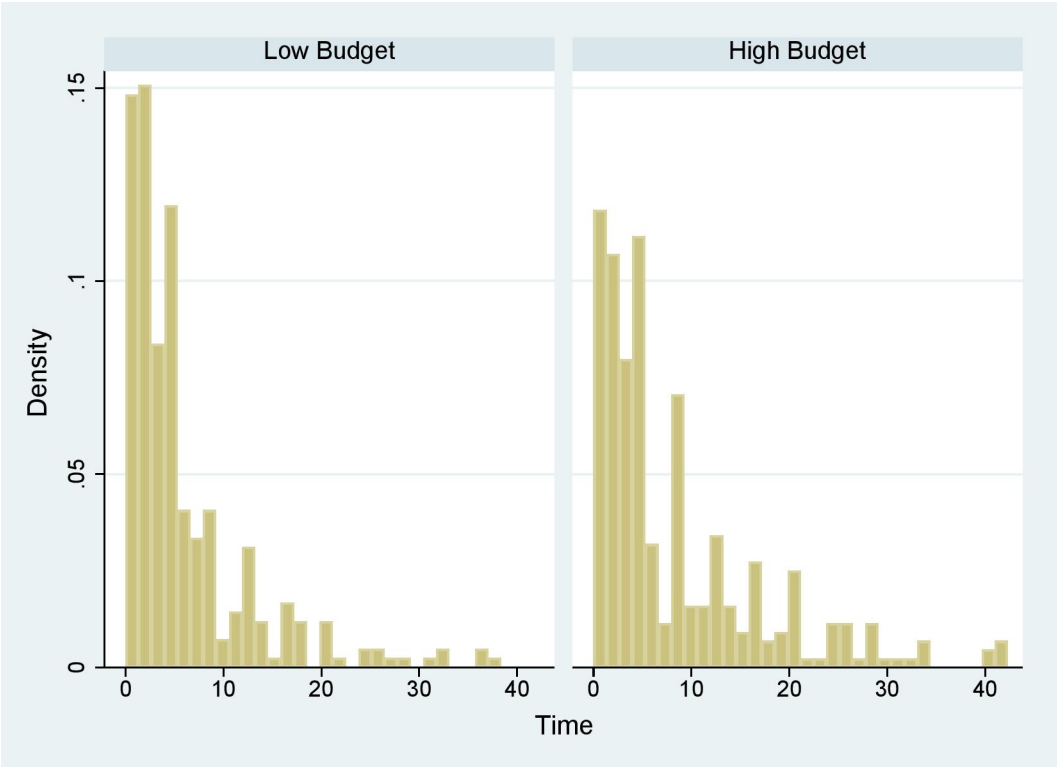


Figure A6.18: Budget used by super/separated market types and budget

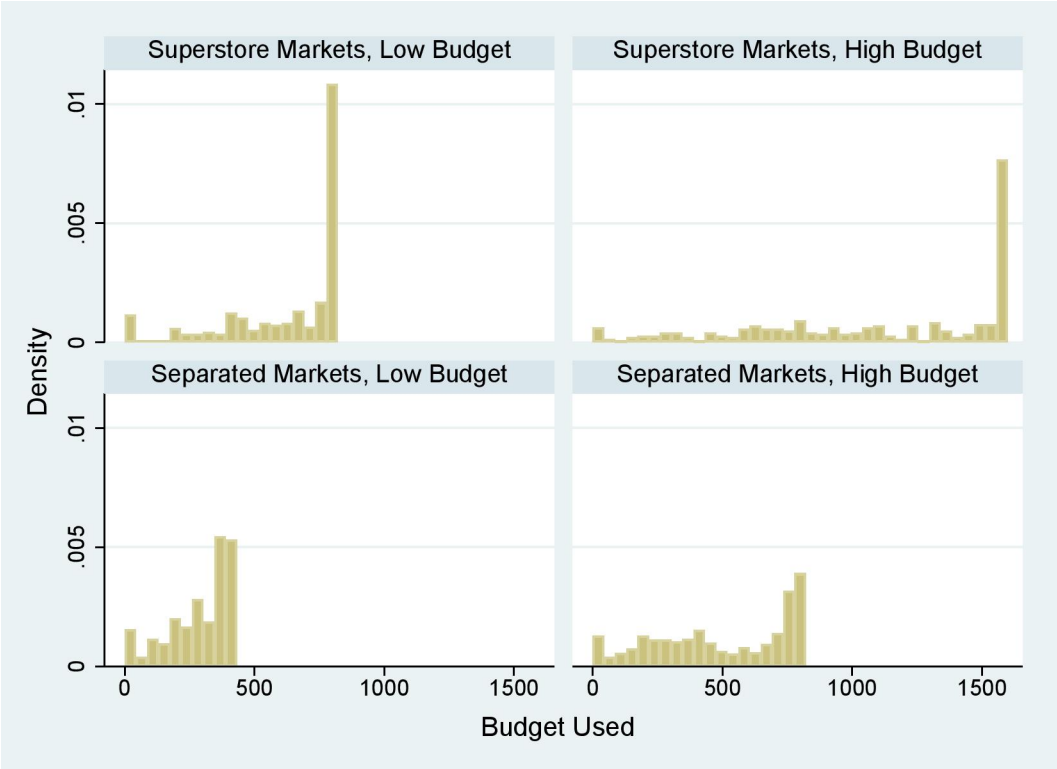
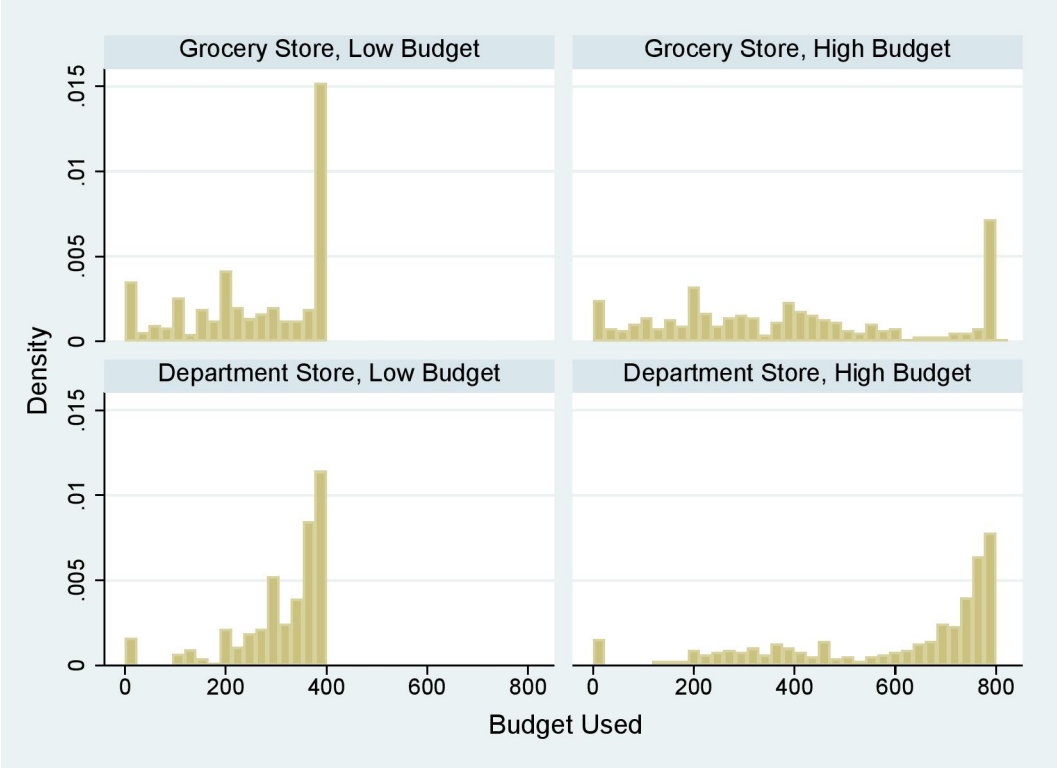


Figure A6.19: Budget used by grocery/department market types and budget



2 Parental Gender Preference in the Balkans and Scandinavia: Gender Bias or Differential Costs?

Coauthored with Sergii Maksymovych and Zurab Abramishvili

2.1 Introduction

The impact of the gender of the first-born child on the number of children in a family has been repeatedly observed in many countries. We confirm son preference using the parity-three progression method applied to a pooled 2004-2015 European Union Statistics on Income and Living Conditions (EU-SILC) cross-sectional sample from four Balkan countries: Bulgaria, Croatia, Slovenia, and the Republic of Serbia.[48] We also confirm daughter preference for three Scandinavian countries, i.e. Denmark, Norway, and Sweden, which had been identified previously by Andersson et al. (2006) and Hank and Kohler (2000). Two possible causes of gender preference considered in the literature are parental bias in favor of one or another gender and different costs[49] of raising sons and daughters (Ben-Porath and Welch, 1976; Lundberg, 2005). There is a consequential difference at the core of the two causes. Gender bias is gender-based asymmetric parental utility related to cultural, psychological, and biological origins. Differential cost has a more impartial, parsimonious foundation based on gender-equitable parental utility with heterogeneous costs of child human capital and their expected returns. Their diverging results reflects a changing world that is becoming more socially, legally, and fiscally equitable in terms of gender.[50] This paper aims to identify which of the two causes is more prevalent in Balkan and Scandinavian countries. Each explanation implies a distinctive relationship between the gender of children and the allocation of household resources. We test for the predominance of the two explanations by checking which relationships hold for the household-level data.

We find that Balkan households with more female children replace furniture less frequently than households with fewer female children. Moreover, in households with more female children, mothers report a lower ability to spend on themselves. Additionally, for Balkan countries we find no difference in parental investment in male and female children and no impact of the gender composition of children on the ability to make ends meet or the minimum

---

[48] These countries are covered by EU-SILC and had the highest SIGI son bias component in Europe according to the OECD: https://www.genderindex.org/ranking/sonbias/

[49] While we test for the difference in costs of children, it is actually the difference in "prices" of sons and daughters in which we are primarily interested. The price of a child is the commitment of resources required to raise a child of given 'quality'. At the same time, the cost of a child is a measure of the actual amount of resources committed to child-raising (Bradbury, 2004). Thus, the cost of children is deliberately chosen by parents and, in principle, is measurable. We refer to this as the "cost" of children, i.e. their human capital, or as parental "investment", "outlays", or "expenditures" on children. In most theoretical models related to the subject, the price of children equals cost, because parents are assumed to pay the full life-time prices of children once they are born or the per-period price every period. Hence, from a model perspective, the terms are interchangeable and are considered so throughout this paper.

[50] These evolving preferences, their explanations, and how they relate to this research are elaborated upon on pages 92-93 in the literature review and pages 120-121 in the Appendix.

amount of money needed to make ends meet. We argue, based on earlier studies, that these findings are consistent with the gender bias explanation and not with the differential cost explanation. For Scandinavian countries, we find no impact of the gender composition of children on replacing furniture or on consumption of other household public goods, and we find significantly larger parental investment in households with more female children. Moreover, we do not find a systematic impact of the gender of their children on parental consumption. We argue, based on conclusions in Lundberg (2005) and Lundberg and Rose (2003b), that these findings are not consistent with the gender bias explanation, but are in line with the differential cost explanation. Supplementary analyses of the top-income-decile sub-sample and of cross-country relationships between gender preference, parental investment, and conventional measures of gender equality support our argument.

2.2 Literature Review

The evidence on the impact of parental gender preference pertains to developing economies (Barcellos et al., 2014; Jiang et al., 2016; Altindag, 2016) and developed economies (Dahl and Moretti, 2008; Andersson et al., 2006; Pollard and Morgan, 2002; Brockmann, 2001). Authors attribute this impact to parental preference for a particular gender of children. In developing economies, parents usually have more children (progress to higher parities) when their firstborn is a daughter (Filmer et al., 2009; Arnold, 1992). The interpretation of such behavior is that they have a son preference, so they continue producing children until they reach a desired number of sons or the upper limit of the desired family size. At the same time, in some developed economies, parents also exhibit son preference (Dahl and Moretti, 2008; Choi and Hwang, 2015), but daughter preference in others (Andersson et al., 2006; Brockmann, 2001).[51] The consequences of parental gender preference have mostly been researched for developing economies. The main consequence is that girls, on average, have more siblings and receive a lower share of household resources (Vogl, 2013; Jensen, 2003; Basu and De Jong, 2010). Consequences include a shorter breastfeeding period for girls (Jayachandran and Kuziemko, 2011), worse health and nutritional status of girls (Arnold, 1992), and biased sex ratios (e.g., Jayachandran, 2017; Guilmoto and Duthe, 2013). In more developed economies, Kippen et al. (2006) and Dahl and Moretti (2008) argue that a son preference increases fertility in Australia and the US. Edlund (1999) demonstrates theoretically that gender preference combined with availability of gender selection technology[52] could lead to a female "under-class", because poorer parents would prefer daughters and richer ones would prefer sons (Trivers and Willard, 1973). Another possible consequence in the setting developed by Edlund (1999) is the

[51] Sandstrom and Vikstrom (2015) provide evidence for the existence of son preference in Germany in the second half of the 19th century, which faded later, while Outram (2015) finds evidence for son preference in Edwardian England.
[52] Such technologies may include infanticide, sex-selective abortion, or poorer health care.

existence of a "backlog" of unmarried men (Gupta, 2014) with ensuing ramifications, such as polygamy (Economist, 2018; Seidl, 1995). This is because changes in the socio-demographic structure lead to the "adoption of adequate institutions" (Seidl, 1995), which is evident, e.g., in the falling marriage-market value of young men across commuting zones in the US (Autor et al., 2017) accompanied by rising acceptance of polygamy in the US recorded by Gallup pollster (Economist, 2018). Any policy that mitigates the effects of gender preferences would need to take into account the causes behind the observed behavior (Lundberg, 2005). Two possible causes considered in the literature are parental bias in favor of some gender and different costs of raising sons and daughters (Ben-Porath and Welch, 1976; Lundberg, 2005). This paper studies which of the two causes is more prevalent across selected European countries. Each explanation implies a distinctive relationship between the gender of children and the allocation of household resources. We test for the predominance of the two explanations by checking which relationships hold for the household-level data.

Regarding parental gender bias, there are several definitions in the economic literature. The first is that some gender brings more direct utility or has a utility premium. This definition is used in most papers on the subject (e.g., Jayachandran and Kuziemko, 2011; Dahl and Moretti, 2008; Yoon, 2006). Authors either forgo explaining possible mechanisms behind the gender bias and take the gender-biased fertility behavior as their starting point (Jayachandran and Kuziemko, 2011) or explain it by a predilection (Dahl and Moretti, 2008) or cultural and biological factors (Yoon, 2006). Scholars in demographic and sociological literature elaborate more and offer further explanations for gender bias, such as expansion of the self, affiliation, stimulation, accomplishment or social comparison (Hank, 2007), as well as the emotional value of children (Sandstrom and Vikstrom, 2015). Moreover, mothers and fathers can perceive the extent to which sons and daughters fulfill these expectations differently (Hank, 2007). Finally, the definition proposed in Lundberg (2005, p. 344) encompasses the aforementioned elements, stating that "parents have child-gender preferences if the marginal value of an additional male child differs, *ceteris paribus*, from the marginal value of an additional female child, or if the marginal utility of increments in boy quality is not equal to the marginal utility of girl quality." Here 'quality' means child outcomes that are outputs of a household production process in which inputs are parental time and market goods and services. This definition incorporates two different cases. In the first case, parental valuation of the gender of children or accompanying outcomes does not relate to parental outlays on children (beyond providing for a minimal subsistence level). In the second case, child outcomes are closely dependent on parental inputs until these inputs reach significant values. The second case is not consistent with previous definitions since the gender is not preferred *per se*, but because it makes the technology of producing a certain quality cheaper, i.e. it is only one means of reaching a specific discrete end. In this paper, we understand gender bias as in the first case, as the predilection for such gender-intrinsic characteristics of children that depend neither in extent nor intensity on parental outlays. Therefore, the gender bias does not mean that parents prefer a son or daughter because s/he will bring higher returns to their investments. Instead, it means

that they want a child of a particular gender because of its predetermined characteristics. If gender bias, as we understand it, were the only determinant of the family size connected to the gender of children, two relationships for household outcomes would likely hold. First, parents who desire boys but have a girl or vice versa anticipate having more children in the future and might start saving or work more to support a larger family (Barcellos et al., 2014). Second, parents who have children of a preferred gender should spend more on household public goods, because their marriage is more stable, as the preferred gender child generates higher surplus (Lundberg, 2005). Therefore, in countries where firstborns of the preferred gender have, on average, fewer siblings, parents of firstborns of this gender should work less, save less, and spend more on household public goods. Moreover, if sons directly increase the utility of fathers, then the standard bargaining model of the household predicts a shift of household resources from fathers to mothers. This redistribution could be observable as increased leisure among mothers of sons, or increased consumption of private commodities typically consumed by women (Lundberg and Rose, 2003b).

Turning to the difference in costs of raising sons and daughters, the literature considers two cases. The first is when sons and daughters have constant, albeit not necessarily equal, cost. The assumption of constant costs of children is taken in much, if not most, of the applied studies on the topic (van Praag and Warnaar, 1997), which frequently calculate so-called normative budgets.[53] Nominal expenditures or normative budgets, however, do not equal total expenditures on children. The latter also include time costs of childcare and exclude the value of children's contribution to household production. Still, the monetary outlays *per se* do not fully reflect the quality of inputs. Another issue is whether parents take into account net flow of future transfers from children (Blacklow, 2002; Adda et al., 2016). Available empirical evidence suggests that parental expectations are important for parental spending (Hao and Yeung, 2015). These assumptions describe the case in which parents rely on some rules of thumb when deciding about outlays on children. These rules of thumb, in turn, are based on perceptions about optimal living arrangements in a given society in a given time (Kornrich and Furstenberg, 2007). Then, to calculate the gender difference in costs of children, studies in the literature employs two methods. The first, the Rothbarth method, measures the adult-good equivalent of child cost. This method, unlike normative budgets or discretionary equivalence scales (van Praag and Warnaar, 1997), is theoretically plausible (Deaton and Muellbauer, 1986). This method estimates the difference in the consumption of private adult goods or leisure time (Bradbury, 2004) between parents having first-born sons and first-born daughters. The second method measures gender difference in costs of children relying on the subjective scales method (Leyden approach) proposed and substantiated in van Praag and Warnaar

[53] For example, the U.S. Department of Agriculture (USDA) has provided estimates of expenditures on children since 1960. Forensic economists use these figures in wrongful death and birth cases, as well as in child support cases (Lino and Carlson, 2010). The constant cost of children is also assumed in, e.g., Dahl and Moretti (2008); Hazan and Zoabi (2015); Leung (1991); Sienaert (2008); Bojer (2002); and Raurich and Seegmuller (2017).

(1997). The second case considered in the literature regarding the difference in costs of sons and daughters is when the cost consists of fixed and variable components. This case is captured by models like those in, e.g., Galor (2011); de la Croix and Doepke (2003); and Hazan and Zoabi (2015). In this case, either fixed (one-time costs) or variable components (cost of human capital) of the child cost could differ. Differences in fixed costs are revealed by parental outlays during the early childhood years. At the same time, differences in the variable component are revealed by the differences in availability of parental investment items. Children with lower human capital costs will receive higher outlays and have fewer siblings due to the substitution of quality for quantity (Galor, 2011; Aaronson et al., 2014).[54] For a better understanding of how the two explanations of parental gender preference differ regarding costs of child human capital discussed above, please see figure A1 and the discussion of the distinction between the gender bias and differential costs concepts in the Appendix.

We use a set of home items as measures of parental investment (Cunha et al., 2010), i.e. as proxy variables for parental outlays on children. Parents buy more of such items when they bring more parental utility per unit of expense for a gender and will have fewer children after having a firstborn of that gender. In our analysis, we assume the costs of children per the latter case, when the costs include fixed and variable components, so that it is consistent with economic theory. Thus, if the differential cost explanation is true, parents of a child of the more expensive gender should have fewer children thereafter, spend less on themselves (both parents), spend less on adult public goods, and spend more on children. Moreover, parents of a "more expensive" child should report higher sums needed to make ends meet. However, if the gender bias explanation specified above is correct, parents will report lower sums, because they should spend more on household public goods which exhibit economies of scale in consumption.[55] The restriction on child age applied in our analysis ensures that a child's financial contribution to a household does not confound the estimates obtained. We analyze only households in which the oldest child is, at most, 12 years old, which is compulsory school age in all European countries.

The two causes considered, the gender bias and the differential costs, might actually be in play simultaneously, but our testing aims to determine which cause most drives the estimates. We expect to find support for gender bias and no impact of differential costs, because cost difference should play a lesser role in European economies compared to developing economies characterized by a pronounced son bias (Brockmann, 2001). However, we observe higher parental expenditure on daughters in countries with daughter preference, which is consistent with a lower cost of child human capital for daughters, whereas the son bias drives son

[54] It could be that either items for some gender are cheaper or produce more parental utility through child human capital. One more case is possible when items generate little human capital and thus, more of them are bought (i.e., the demand for them is inelastic). However, it is unlikely that this effect would be stronger in countries with more gender-equivalent attitudes as Figure A3 in the Appendix shows.

[55] Tables A5 and A6 in the Appendix include a summary of key household allocation decisions that illustrate whether a given parental gender preference is based upon gender bias or cost difference.

preference in countries where we observe it, outweighing the effects of a higher cost of daughters (which is, however, not as high as in daughter-preferring countries). Moreover, the cross-country correlation between our estimates of the gender preference and the cost difference is stronger than the correlation between our estimates of the gender preference and the conventional measures of gender equality (GGI, GDI, etc.), which arguably approximate gender bias. All these findings taken together indicate that gender preference across countries is more strongly determined by the cost difference than by gender bias. Therefore, a policy intended to neutralize gender preference effects would subsidize the costs of human capital for sons from families which are less well off.

2.3 Data and Sample Statistics

We use a data set from the European Union Survey of Income and Living Conditions (EU-SILC) for 2004-2015. The data set is collected annually by national statistical offices in cooperation with Eurostat from nationally representative samples, which covered the EU-28 and several non-EU countries in 2015. In 2004, only 15 countries were covered by the survey. Our analysis is based on data from four Balkan countries and three Scandinavian countries. The Balkan countries are Bulgaria, Croatia, Slovenia, and the Republic of Serbia.[56] The Scandinavian countries are Denmark, Norway, and Sweden.[57] A primary goal of EU-SILC is to collect cross-sectional and longitudinal microdata using a rotational four-year panel scheme on income, poverty, social exclusion, and living conditions (Eurostat, 2017). The longitudinal component is not used in our research. The reference population in EU-SILC includes all private households and their current members residing in the territory of the respective countries at the time of data collection. All household members are surveyed, but only those aged 16 and older are interviewed. The data set for each year after 2004 consists of two groups of variables: primary and secondary. Primary variables are collected annually. Secondary variables are collected approximately every five years in so-called ad-hoc modules. A variable may include information at the household or personal level about specific topics. The primary variables convey information on household demographic composition, incomes, living conditions, and labor market activity. The secondary variables used in the current research were collected in 2009, 2010, and 2013-2015 in ad-hoc modules on material deprivation.

[56] These are Slavic-speaking Balkan countries covered by the EU-SILC survey. When we extend the set of Balkan countries to include Greece and Romania, the estimates of gender preference do not change qualitatively.

[57] These groupings of countries have been frequently used in previous studies. For instance, Estrin and Uvalic (2014) use a similar grouping of Balkan countries and conduct regression analyses on the pooled sample of data under the assumption that regression parameters do not differ between these countries. Similarly, Baranowska-Rataj and Matysiak (2016) and Ragan (2013) use the mentioned grouping of Scandinavian countries. Both studies assume that the considered characteristics of those economies (model parameters) are similar across Scandinavian countries. In a similar vein, Filmer et al. (2009) pool HNS data into six sub-samples by parts of the world and assume no difference in parameters between countries within groups.

These secondary variables contain more in-depth information on material deprivation in the household than the annual primary variables. Eurostat calculates cross-sectional household and individual weights to correct for non-random sampling and non-responses (Eurostat, 2015).[58]

Two main advantages of this data set are important for our analysis. First, it contains information on the age and gender of all adults and their children living in the household. Second, the ad-hoc modules from 2009, 2010, and 2013-2015 contain detailed information on material deprivation of adults and children in the household. There are also two significant drawbacks. First, not all children might be present in the household at the time of the survey for some reason (e.g., because they study or work elsewhere). We cannot be sure that the firstborn child lives in the household. Second, the information on material deprivation of children is available only for all children in the household together and not for each child separately.[59] To correct for the first drawback, we limit our sample to data where we can claim with high certainty that the firstborn child is still in the household. Specifically, following other studies in the literature (Dahl and Moretti, 2008; Karbownik and Myck, 2017; Ananat and Michaels, 2008), we limit the analysis to mothers aged between 18 and 40 who had their first child at the age of 16 or older. The limit for the age of the oldest child is set at 12 years.[60] Our calculated sex-ratio for firstborns is 1.057, close to the commonly accepted value of 1.06 (Grech et al., 2002).[61] To correct for the second drawback, we connect the material condition of children in the household to the gender composition of children (i.e., the share and presence of daughters among children are instrumented with a dummy for the first child being a girl). Since the gender of children influences household composition, we limit our analysis, for the most part, to married and cohabiting couples. Table 1 contains descriptive statistics for selected household socio-demographic characteristics separately for all families and for cohabiting couples. Table 2 presents descriptive statistics on variables characterizing different aspects of the household material condition. We use variables in Table 2 as dependent variables and variables in Table 1 as covariates. Amongst adult and household material deprivation characteristics, Table 2 also presents the average frequency of the ten home environment items

---

[58] More detailed information on the dataset is available at the following link: http://ec.europa.eu/eurostat/web/microdata/overview

[59] For example, the answer to the question: ”Do children have books at home suitable for their age?” should be ”Yes” if all children have books and ”No” if at least one child does not have books.

[60] The sample bias is likely to be very small because the minimal age of leaving school in all European countries is above 16. Other studies (Dahl and Moretti, 2008; Karbownik and Myck, 2017) use the threshold of 12 years. Karbownik and Myck (2017) use this threshold since it corresponds to the grouping of expenditure information on clothing. We need a broader range of ages because we aim to control for the age of children (which was not done in other studies). Dahl and Moretti (2008) find the 12-year cutoff conservative while Ichino et al. (2011) and (Ananat and Michaels, 2008) use 15-year and 17-year cutoffs, respectively. Importantly, our chosen threshold ensures that child earnings do not confound our results because this threshold is below the compulsory schooling age in all European countries. At the same time, when we estimate our models on the entire sample, the estimates preserve signs and statistical significance but reduce in size.

[61] This fact also suggests that gender-selective abortion or gender difference in early childhood treatment should be too rare to show up in the data.

for children along with girl-boy differences. One can readily see that girls are more likely to have books, have an opportunity to invite friends, and to host celebrations. These differences are small, however, and hover around one percent of the standard deviation of the corresponding items. This is less than reported by Xu (2016). The largest differences between all families and intact, i.e. married and cohabiting, families appear to be in food and clothing. Specifically, the girl-boy difference is significant for all families, but disappears for intact families. This could be explained by more limited resources of non-intact families.[62] Otherwise, the intact families do not appear to differ systematically from all families along the considered characteristics, which supports our decision to focus the analysis on intact families.

[62] This result is consistent with the Trivers-Willard hypothesis. Further exploration of this question is beyond the scope of this study.

Table 1: Descriptive statistics – demographics and labor market information.

	Balkan countries				Scandinavian countries			
	All families		Married couples		All families		Married couples	
Selected household characteristics	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference
Living without father	0.114	-0.005	-	-	0.106	0.003	-	-
	(0.318)	(0.003)	-	-	(0.308)	(0.003)	-	-
Number of children	1.855	0.047	1.872	0.046	1.996	0.004	2.016	0.005
	(1.047)	(0.010)***	(0.996)	(0.010)***	(0.839)	(0.007)	(0.832)	(0.007)
First-born girl	0.481	-	0.484	-	0.487	-	0.487	-
	(0.500)		(0.500)		(0.500)		(0.500)	
Age of mother at first birth [a]	26.44	0.035	27.06	0.03	28.91	0.04	29.09	0.07
	(7.35)	(0.07)	(5.36)	(0.06)	(5.36)	(0.04)	(4.79)	(0.04)
Age of mother	34.68	0.001	35.4	0.0009	37.44	0.002	37.56	0.05
	(7.40)	(0.07)	(6.12)	(0.06)	(6.26)	(0.05)	(5.80)	(0.05)
Mother having tertiary education	0.178	-0.005	0.195	-0.007	0.363	0.002	0.402	0.004
	(0.382)	(0.003)	(0.396)	(0.004)	(0.481)	(0.004)	(0.490)	(0.004)
Mother employed	0.606	0.000	0.650	-0.003	0.746	-0.001	0.821	0.001
	(0.489)	(0.004)	(0.477)	(0.005)	(0.435)	(0.003)	(0.383)	(0.003)
Mother's weekly hours of work	28.100	-0.106	28.738	-0.159	27.985	0.341	28.001	0.340
	(19.424)	(0.183)	(19.141)	(0.186)	(14.872)	(0.122)**	(14.851)	(0.123)**
Father employed	-	-	0.805	0.004	-	-	0.924	-0.005
	-	-	(0.396)	(0.004)	-	-	(0.264)	(0.002)
Father's weekly hours of work	-	-	37.156	0.082	-	-	37.810	-0.165
	-	-	(16.689)	(0.162)	-	-	(12.762)	(0.106)
Household disposable income (euros)	20,469.770	265.421	20,982.732	214.079	64,070.609	325.596	65,957.259	450.271
	(15,431.683)	(141.036)	(15,550.905)	(150.036)	(57,680.462)	(447.583)	(59,032.599)	(483.734)
Living in urban area	0.137	0.003	0.131	0.002	0.347	0.000	0.341	0.002
	(0.344)	(0.003)	(0.337)	(0.003)	(0.476)	(0.004)	(0.474)	(0.004)
Ownership of accommodation	0.767	-0.003	0.763	-0.004	0.920	-0.005	0.929	-0.004
	(0.423)	(0.004)	(0.425)	(0.004)	(0.271)	(0.002)**	(0.257)	(0.002)**
N of hhds	24,951		22,027		28,352		25,294	

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Note: The statistics were calculated for the subsample of intact families with children. Columns one and three show means and standard deviations while columns two and four show differences between mean values for girls versus boys. Values in parentheses in even numbered columns correspond to t-test standard errors.

[a] These statistics were calculated only for families in which the mother is younger than 41 and older than 17 and had her first child at the age of 16 or older and child ages are in the range 0–12.

Table 2: Availability of selected items in the home environment for girls and boys.

	Balkan countries				Scandinavian countries			
	All families		Married couples		All families		Married couples	
Dependent variables	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference
Household-level material condition characteristics[a]								
Amount of money needed to make ends meet	1,486.629	11.577	1,507.179	8.119	4,725.007	44.569	4,823.201	84.074
	(830.649)	(7.705)	(831.329)	(8.141)	(13,992.615)	(115.330)	(14,112.091)	(122.862)
Ability to make ends meet	0.215	0.004	0.225	0.002	0.776	0.002	0.798	0.006
	(0.411)	(0.004)	(0.418)	(0.004)	(0.417)	(0.003)	(0.402)	(0.003)**
Replacing worn-out furniture	0.278	-0.008	0.290	-0.005	0.888	-0.006	0.905	-0.004
	(0.448)	(0.007)	(0.454)	(0.007)	(0.316)	(0.005)	(0.293)	(0.005)
Adult-specific material condition characteristics[b]								
Ability to spend a small amount of money on oneself (women)	0.522	0.000	0.533	-0.000	0.399	0.017	0.381	0.016
	(0.500)	(0.007)	(0.499)	(0.007)	(0.490)	(0.007)**	(0.486)	(0.007)**
Ability to spend a small amount of money on oneself (men)	0.540	0.003	0.573	0.005	0.383	-0.013	0.408	-0.014
	(0.498)	(0.007)	(0.495)	(0.007)	(0.486)	(0.007)*	(0.492)	(0.007)**

Table 2 (continued)

	Balkan countries				Scandinavian countries			
	All families		Married couples		All families		Married couples	
Dependent variables	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference
Availability of two pairs of properly fitting shoes (women)	0.615 (0.487)	-0.003 (0.007)	0.627 (0.484)	-0.001 (0.007)	0.437 (0.496)	0.017 (0.007)**	0.411 (0.492)	0.015 (0.007)**
Availability of two pairs of properly fitting shoes (men)	0.597 (0.490)	-0.000 (0.007)	0.634 (0.482)	0.002 (0.007)	0.408 (0.492)	-0.012 (0.007)*	0.435 (0.496)	-0.012 (0.007)*
Replace worn-out clothes (women)	0.540 (0.498)	0.003 (0.007)	0.555 (0.497)	0.004 (0.007)	0.415 (0.493)	0.013 (0.007)*	0.393 (0.488)	0.011 (0.007)
Replace worn-out clothes (men)	0.535 (0.499)	0.002 (0.007)	0.571 (0.495)	0.003 (0.007)	0.396 (0.489)	-0.012 (0.007)*	0.422 (0.494)	-0.013 (0.007)*
Get together with friends/family at least once a month (women)	0.552 (0.497)	0.004 (0.007)	0.565 (0.496)	0.005 (0.007)	0.429 (0.495)	0.018 (0.007)**	0.405 (0.491)	0.017 (0.007)**

Table 2 (continued)

	Balkan countries				Scandinavian countries			
	All families		Married couples		All families		Married couples	
Dependent variables	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference
Get together with friends/family at least once a month (men)	0.551	-0.002	0.586	-0.001	0.401	-0.016	0.426	-0.016
	(0.497)	(0.007)	(0.493)	(0.007)	(0.490)	(0.007)**	(0.495)	(0.007)**
Regularly participate in a leisure activity (women)	0.233	-0.006	0.244	-0.006	0.322	0.010	0.307	0.010
	(0.423)	(0.005)	(0.430)	(0.006)	(0.468)	(0.007)	(0.462)	(0.007)
Regularly participate in a leisure activity (men)	0.254	-0.005	0.276	-0.006	0.317	-0.009	0.338	-0.009
	(0.435)	(0.006)	(0.447)	(0.006)	(0.465)	(0.007)	(0.473)	(0.007)
children home environment items[d]								
Replacing worn-out clothes	0.822	-0.007	0.843	-0.005	0.986	0.000	0.987	0.001
	(0.382)	(0.007)	(0.363)	(0.007)	(0.118)	(0.003)	(0.113)	(0.003)
Two pairs of properly fitting shoes	0.845	0.006	0.867	0.007	0.983	0.000	0.986	-0.002
	(0.362)	(0.006)	(0.340)	(0.006)	(0.128)	(0.003)	(0.118)	(0.003)

Table 2 (continued)

	Balkan countries				Scandinavian countries			
	All families		Married couples		All families		Married couples	
Dependent variables	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference
Fresh fruits and vegetables once a day	0.866	-0.010	0.885	-0.006	0.982	-0.003	0.983	-0.003
	(0.341)	(0.006)	(0.319)	(0.006)	(0.134)	(0.003)	(0.127)	(0.003)
One meal with fish, chicken or meat (or vegetarian equivalent) at least once a day	0.842	-0.003	0.862	-0.001	0.988	0.003	0.989	0.002
	(0.365)	(0.006)	(0.345)	(0.006)	(0.108)	(0.002)	(0.103)	(0.002)
Books at home suitable for children's ages	0.844	0.006	0.863	0.009	0.983	0.006	0.984	0.005
	(0.363)	(0.006)	(0.344)	(0.006)	(0.131)	(0.003)	(0.126)	(0.003)
Outdoor leisure equipment	0.821	-0.001	0.841	0.004	0.987	-0.002	0.990	-0.003
	(0.383)	(0.007)	(0.366)	(0.007)	(0.112)	(0.002)	(0.102)	(0.002)
Indoor games	0.875	-0.002	0.891	0.000	0.995	-0.000	0.996	-0.001
	(0.331)	(0.006)	(0.312)	(0.006)	(0.072)	(0.001)	(0.066)	(0.001)
Regular leisure activity	0.503	0.010	0.518	0.009	0.776	0.017	0.779	0.019
	(0.500)	(0.009)	(0.500)	(0.009)	(0.417)	(0.008)**	(0.415)	(0.009)**

Table 2 (continued)

	Balkan countries				Scandinavian countries			
	All families		Married couples		All families		Married couples	
Dependent variables	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference	Mean	Girl-boy difference
Celebrations on special occasions	0.867	-0.002	0.884	0.000	0.981	0.001	0.983	0.002
	(0.339)	(0.006)	(0.320)	(0.006)	(0.137)	(0.003)	(0.129)	(0.003)
Invite friends over to play	0.790	0.002	0.807	0.005	0.959	0.002	0.959	0.002
	(0.408)	(0.007)	(0.395)	(0.007)	(0.198)	(0.004)	(0.198)	(0.004)

Note: The statistics were calculated for the subsample of intact families with children. Columns one and three provide means and standard deviations while columns two and four provide differences between mean values for girls versus boys. Values in parentheses in even numbered columns correspond to t test standard errors.

[a] The amount of money needed to make ends meet and the ability to make ends meet are primary variables collected annually while replacing worn-out furniture was collected in ad-hoc modules in years 2009 and 2013-2015.

[b] Adult-specific material condition characteristics were collected in ad-hoc modules in years 2009 and 2013-2015.

[c] This variable and the three next variables were collected in 2010.

[d] Children's home environment items were collected in ad-hoc modules in 2009 and 2013-2015.

2.4 Empirical Analysis

Our analysis tests for the predominance of the two alternative explanations for parental gender preference. Each has different implications for household economic behavior. The gender bias hypothesis implies that households with a first-born child of the desired gender save less (Barcellos et al., 2014)[63] and spend more on household public goods (Lundberg, 2005). As we do not have a direct measure of household savings, we use the capacity to face unexpected financial expenditures as a proxy variable. Here we rely on the intuitively appealing assumption that greater savings mean higher capacity to deal with unexpected expenditures. Regarding the measure of household public goods, we use replacing worn-out furniture. Other measures, like good nutrition and quality of leisure or availability of appliances and cars, are more likely to have a direct impact on child well-being and thus might be not invariant to the gender of children. Moreover, more household public goods available should also result in a greater ability to make ends meet and less money needed to make ends meet, because the consumption of household public goods exhibit returns to scale. At the same time, the differential costs hypothesis implies that parents of a child of the preferred gender (i.e., of the more expensive gender, resulting in fewer additional or total births) work more, save less, and spend less on adult public goods. Parents of more expensive children should report a lower ability to make ends meet along with higher sums needed to make ends meet.

One possible way to test our hypotheses is to compare families with different child gender composition. This is the approach taken by Bogan (2013), who explores the relationship between household financial-asset-market participation and the gender of children. Specifically, Bogan estimates a regression in which the dependent variable is stock or bond ownership while the explanatory variables are dummies for only female and only male children or a proportion of female children in the household. However, since the explanatory variable in both specifications (the dummies for same-gender children and share of daughters) might be decided by households and, thus, may be endogenous, such estimates cannot be taken as evidence of a causal relationship between the variables in question.[64] Similarly, in the case of our analysis, more daughter-preferring parents could also derive more utility from the well-being of their children and, thus, tend to create better material conditions for them. To address these concerns, we use the gender of the firstborn as the explanatory variable. Our identification strategy is to assume that the gender of the firstborn is randomly determined. This assumption has been made in other studies that use the gender of firstborns as an

[63] These authors also mention that in such households, mothers end their maternal leave earlier. Evidence from the US, however, suggests that fathers of sons tend to work less. At the same time, many authors find sons preferred in the US. The descriptive statistics for the pooled EU-SILC sample show that mothers of daughters actually work more when daughters are the preferred gender. Nevertheless, a comprehensive testing of this implication for the EU-SILC data is beyond the scope of this paper.

[64] More daughter-preferring families, for instance, are more likely to have all daughters: they self-select into having all daughters because son-preferring families who have only daughters are more likely to continue having more children until they have a son. At the same time, daughter-preferring families could be less risk-averse and, consequently, more inclined to participation in financial asset markets.

instrument for household characteristics. Some of these characteristics are the bargaining power of women in China (Li and Wu, 2011), the number of children in a family (Dahl and Moretti, 2008), the occurrence of divorce (Bedard and Deschenes, 2005; Ananat and Michaels, 2008), and the area of accommodation (Dujardin and Goffette-Nagot, 2009).[65]

To test the hypotheses above, we proceed in three steps. First, we estimate gender preference across European countries using the third-parity method. Second, we verify the validity of the gender bias explanation by testing its aforementioned implications in daughter-preferring countries and son-preferring countries, respectively. That is, in countries where we observe daughter preference, parents of a first-born daughter should be less capable of dealing with unexpected financial expenditures (because they save less), spend less on themselves, be more likely to replace worn-out furniture, be more able to make ends meet, and need less money to make ends meet. The same predictions should hold for parents of first-born sons in son-preferring countries. Third, we verify the validity of the differential costs explanation by testing its implications in daughter- and son-preferring countries. We do this in two stages. In the first stage, we assume constant costs of sons and daughters (e.g., Dahl and Moretti, 2008; Jayachandran and Kuziemko, 2011; Leung, 1991). In the second stage, we relax this assumption and, instead, assume the cost of children to consist of two components, fixed and variable (e.g., Galor, 2011; Aaronson et al., 2014; de la Croix and Doepke, 2003). In the latter case, we determine whether the difference is driven by the fixed or the variable component.

The baseline specification of the regression model takes the following form:

$$y_i = \beta(\textit{First child girl})_i + \alpha \boldsymbol{X}_i + \epsilon_i \quad (1)$$

where y_i stands for either the progressing to parity three (having three children) or a child's material conditions indicator for a household i and $\boldsymbol{X}_i$ is a vector of household i socio-demographic and economic characteristics. The *First child girl* indicator takes value 1 if the first-born child was a girl and 0 if a boy. Within a given country, the residual values, ϵ_i, can be correlated. The specific set of variables that make up $\boldsymbol{X}$ depends on the particular regression equation specification. We use this form at each of the three steps of the hypothesis testing.

To test for gender preference, we put the third parity progression on the left-hand side. Progression to the third parity has been the most widely used indicator in the literature to test for gender preference. There are two main reasons it is better to use parity-three progression rather than parity-two progression to measure the gender preference. First, it is likely that the desire for a gender-mix of children (to have at least one son and one daughter) coexists with the gender bias towards one gender (Dahl and Moretti, 2008). In that case, parents who have

[65] The second Appendix subsection describes additional considerations and reservations about using this instrument.

a bias towards any gender will progress to parity two independently of the gender of their firstborn. That is why the causal effect of the gender of the firstborn on the progression to parity two is not likely to be significant. The second reason is that first-born twins would distort the estimates for parity two progression. Still, we also report second parity progression and total number of children. We choose covariates that have been used in similar studies: gender of the first two children, cubic polynomial of mother's age, squared polynomial of mother's age at first birth, length of cohabitation of spouses, mother's education, father's education, mother's employment, father's employment, household disposable income, and living in an urban area (Dahl and Moretti, 2008; Hank and Kohler, 2000; Haughton and Haughton, 1998; Larsen et al., 1998; Clark, 2000; Basu and De Jong, 2010). We include higher degree polynomials in the mother's age to account for the conclusions reached by Yamaguchi and Ferguson (1995), who argue that the probability of giving birth for women is lower at a younger age, then increases, and then again decreases. Such a relationship is best fit by the third-degree polynomial in age. Finally, we include the family's occupied accommodation tenure along with year and country dummies. We estimate the models with OLS, as do most other studies on the subject, because this method yields consistent estimates of the coefficient on the dummy for the gender of the firstborn. The linear probability model may be an especially good choice because right-hand side variables are mostly dummies (of 23 covariates only 7 are continuous variables) and the unboundedness problem is less acute in this case (Wooldridge, 2002, p.456). Nevertheless, we also run Probit estimations to check for consistency with the OLS-based results.[66] Since we expect observations not to be iid, but correlated within countries, we cluster the standard errors at the country level.

In regard to testing for differential costs of sons and daughters, we assume that the cost of children consists of two components: constant (one-time cost) and the variable (outlays on human capital). Researchers commonly use this assumption in models featuring parental investment in children. The fixed component of child cost primarily represents the time cost of rearing children during infancy, whereas the variable component represents parental expenditures on child human capital. Thus, if our analysis finds that parental outlays on children of one gender are larger, there could be two causes: larger one-time costs or lower cost of human capital (parental discounted utility derived from child human capital). The mechanism behind the second cause is that of substitution of quality for quantity of children. For example, parents may spend more on daughter "quality" and have fewer children after daughters. If this explanation is true, daughters in daughter-preferring countries should receive more parental investments. One measure of parental investments used in the literature[67] is the availability of conditions and items at home which are necessary for normal child development (Cunha et al., 2010; Todd and Wolpin, 2007; Juhn et al., 2015).[68] The expected effects of the

[66] The Probit estimates correspond to OLS estimates in terms of impact direction and statistical significance.
[67] The most common measure is years of schooling, conditional on household income.
[68] These variables are described in more detail in the second Appendix subsection.

first-born daughter are systematically presented in Appendix Table A7. We use the 2009/2010 and 2013-2015 EU-SILC data on the availability of such items in households to test if daughters tend to have better material conditions in daughter-preferring countries and sons, respectively, in son-preferring countries. Under this assumption, parents having a child of the more expensive gender, in addition to having a lower progression ratio, should also have lower expenditures on private consumption and household public goods, be less able to deal with unexpected financial expenditures, be less able to make ends meet, and need more money to make ends meet. The ability to make ends meet is measured by a binary variable taking value 1 when a household is able to make ends meet. The aforementioned predictions follow from the fact that they have fewer financial means left after making outlays on children than parents with a child of the cheaper gender. The method of measuring the cost of children through comparing the amount of money needed to make ends meet reported by families having children of different gender was proposed and used by van Praag and Warnaar (1997).[69]

2.5 Results

2.5.1 Estimates of Parental Gender Preferences for Children

Table 3 presents coefficients on the gender of the firstborn for different specifications of the dependent variable in Equation 1 estimated on data from Balkan countries. These results resemble those obtained by Dahl and Moretti (2008) in the US. The first column indicates that families in which the first child is a girl ultimately have more children than families in which the first child is a boy, although the difference is not significant. In line with the expectations discussed above, the impact of the gender of the firstborn on progression to parity two in column (2) is much less statistically significant and lower than the impact on progression to parity three and has a much lower percent effect. The numbers in column (3) show the probability of having three or more children is 1.3 percent higher when the first child is a girl, which is an order of magnitude higher than the result obtained by Dahl and Moretti (2008) in the US. In other words, first-born-girl families are 17% more likely to have three or more

[69] One way to conceptually unify the aforementioned gender differences in the costs of raising children is to interpret them as differences in constraints associated with raising sons and daughters (Lundberg, 2005). In that case, intact families have a comparative advantage in raising a child of a preferred gender provided that, in the vast majority of cases, mothers have custody of children (Dahl and Moretti, 2008). Specifically, in the case of paternal comparative advantage in raising sons, intact families have a comparative advantage in raising sons over single-mother-headed families. In the case of differential costs, an intact family also has a comparative advantage in raising a child with a lower cost of human capital, because it has more resources at its disposal thanks to economies of scale, even if the total nominal incomes of family members remain the same whether it is intact or not. Here the economy of scale means that the opportunity cost of raising a child of a gender with more costly human capital (in terms of utility forgone if the child were the gender with lower cost of human capital) increases with family income. This is true, for instance, when a marginal return to parental investment in children is constantly higher for one gender. The proposed unification of child gender differences in the costs of children, along with the previous reasoning, has several implications for household allocation, which are presented in Appendix Table A5.

children compared to first-born-boy families. We also find significant positive effects for the probability of four or more and five or more children when the first-born child is a girl. The positive effect of the first-born daughter on progression to parity three has also been found by Filmer et al. (2009) in Central Asia, South Asia, Middle East, and North Africa. It is this result which is most commonly interpreted in the literature as a manifestation of son preference.

Table 3: The firstborn-child gender and family size in the Balkans.

Breakdown by number of children					
	(1) Total number of children	(2) Two or more children	(3) Three or more children	(4) Four or more children	(5) Five or more children
First-born child being a girl	0.030	-0.001	0.013	0.011	.003
	(0.010)***	(0.008)	(0.005)***	(0.002)***	(0.001)***
Controls	Yes	Yes	Yes	Yes	Yes
First boy baseline	1.57	0.483	0.077	0.011	0.002
Percent effect	0.019	-0.002	0.17	0.18	0.50
R-sq	0.26	0.39	0.13	.04	.02
Observations	19,807	-	-	-	-

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: S.E. are given in parentheses and are clustered at the country level. Estimates are based on the 2004-2015 EU-SILC samples for Bulgaria, Croatia, Serbian Republic, and Slovenia. The sample consists of households formed by one cohabiting couple, their children, and, occasionally, other relatives. The mother of children in the household is younger than 41 and older than 17 and had her first child at the age of 16 or older, and children's ages are in the range 0–12. The estimation method used is weighted OLS with probability weights reflecting non-random sampling within and between countries. The table presents estimated effects of the firstborn being a daughter compared with the baseline case of the firstborn being a son. The effect is a ratio of the estimated OLS coefficient on the firstborn's gender dummy to the baseline value of the dependent variable. The dependent variables are the total number of children and a set of binary indicators for specific numbers of children. The control variables, besides the gender of the firstborn: the dummy for a first-born daughter, gender of the first two children, cubic polynomial of mother's age, squared polynomial of mother's age at first birth, length of cohabitation of spouses, mother's education, father's education, mother's employment, father's employment, household disposable income, living in urban area, tenure status, year and country dummies.

Table 4 presents estimates analogous to those in Table 3, but for Scandinavian countries. These results are notably different from the results for Balkan countries. First, the impact of a first-born daughter on progression to parity three in column (3) is negative and statistically significant. Despite having a similar absolute value, the effect is half of the Balkan effect, because a larger share of Scandinavian families progresses to parity three. Second, impacts of a first-born daughter on the total number of children and on progression to other parities have small absolute magnitudes and are not statistically significant. The parity three progression results in column (3) are in line with those obtained by Andersson et al. (2006), for each of the Scandinavian countries separately. This alone suggests that gender bias is probably not the only mechanism behind these results, because they would then also be similar for progressions to higher parities.

Table 4: The firstborn-child gender and family size in Scandinavia.

Breakdown by number of children					
	(1) Total number of children	(2) Two or more children	(3) Three or more children	(4) Four or more children	(5) Five or more children
First-born child being a girl	-0.009	0.002	-0.013	0.002	0.0002
	(0.010)	(0.006)	(0.005)***	(0.002)	(0.0002)
Controls	Yes	Yes	Yes	Yes	Yes
First boy baseline	1.82	0.64	0.16	0.02	0.003
Percent effect	0.005	0.003	0.08	0.1	0.07
R-sq	0.29	0.38	0.22	0.05	0.01
Observations	25,227	-	-	-	-

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: Estimates are based on the 2004-2015 EU-SILC samples for Denmark, Norway, and Sweden. For details about sampling and estimates presentation, see the notes under Table 3.

In Appendix Figure A6 and Table A3, we present gender preferences across EU countries. Our results are broadly consistent with those obtained in previous literature (Hank and Kohler, 2000). We also attempt to evaluate how our results would differ if there were no family disruptions caused by child gender, which is frequently reported in the literature (see, e.g., Lundberg (2005) for a review). Estimates obtained for that counterfactual scenario, however, do not differ qualitatively and do not differ much quantitatively from those reported here. Absence of rank correlations between the country-level impacts of the firstborn's gender on progression to parity two and parity three suggests different driving causes behind these impacts.

2.5.2 Testing Predominance of the Gender Bias and Differential Cost Explanations

The gender bias explanation implies two patterns in household-level allocations.[70] First, expenditures on household public goods should be higher when the firstborn is of the preferred gender (Lundberg, 2005). Specifically, if a son increases marital surplus more than a daughter, then the birth of a son reduces the probability of divorce and increases the incentive of partners to invest further in the marriage, i.e. the family as a whole (Lundberg and Rose, 2003b). Second, saving should be less, because parents anticipate fewer births in the future (Barcellos et al., 2014). To test the first implication, we estimate the impact of a first-born daughter on the frequency of replacing furniture in the household. Lundberg and Rose (2003b) consider

[70] Table A6 shows the results of testing for the predominance of the gender bias and the gender-specific constraints explanations for the Balkans and Scandinavia separately. The rounded cells in Table A6 indicate that data corroborate the gender bias explanation for the Balkans and the differential constraints explanation for Scandinavia. The ensuing discussion clarifies which specific form the differential constraints are most likely to take. The current section further explains that it is the gender difference in cost of child human capital.

furniture an important household public good along with automobiles and housing conditions as proxies for housing expenditures. Spending on automobiles and housing, however, can be directly influenced by child gender composition. As Lundberg and Rose (2003b) note, observed differences in housing spending could influence the need for greater space to accommodate the size and activity of sons or the desire for a higher quality neighborhood to reduce the probability of risky behavior by boys or probability of crimes against girls. Concerning automobiles, possessing one might be more prevalent when a couple has sons, who are possibly expected to be more skillful with cars and for whose socialization access to an automobile may be considered more important than for daughters (Peters, 1994). Meanwhile, expenditures on furniture do not appear to be directly influenced by the gender of children.

Table 5: Impact of a first-born girl on availability of household public goods across countries grouped by observed gender preference

Countries	(1) Replacing worn-out furniture	(2) Capacity to deal with unexpected expenditures	(3) Ability to make ends meet	(4) Lowest monthly income to make ends meet	(5) Availability of home items
Balkan	-0.020	0.0019	0.008	-0.671	0.017
	(0.011)*	(0.007)	(0.006)	(9.848)	(.015)
Scandinavian	-0.006	0.005	0.005	152.7	0.035
	(0.007)	(0.005)	(0.004)	(142.2)	(0.018)**

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: The standard errors of estimates on sub-samples for Balkan and Scandinavian countries are clustered at the country level. Estimates in columns (2), (6), and (7) are based on the 2009 and 2013-2015 EU-SILC ad-hoc modules, while the estimates in the remaining columns are based on the 2004-2015 EU-SILC primary modules. The sample consists of households formed by one cohabiting couple, their children, and, occasionally, other relatives. The mother of children in the household is younger than 41 and older than 17 and had her first child at the age of 16 or older, and children's ages are in the range 0–12. The estimation method used is weighted OLS with probability weights reflecting non-random sampling within and between countries. Dependent variables for columns (1) and (3)-(7) are binary indicators taking value 1 when a household has the indicated condition and value 0 otherwise. The table presents estimated effects of the firstborn being a daughter compared with the baseline case of the firstborn being a son. Other control variables: the dummy for a first-born daughter, gender of the first two children, cubic polynomial of mother's age, squared polynomial of mother's age at first birth, length of cohabitation of spouses, mother's education, father's education, mother's employment, father's employment, household disposable income, living in urban area, tenure status, year and country dummies.

Column (1) of Table 5 contains estimates of the firstborn's gender impact on the replacement of worn-out furniture in the household. The negative and statistically significant estimate for Balkan countries confirms the prediction from the son bias explanation of the observed gender preference. To support the daughter bias explanation for Scandinavian countries, the estimate would need to be positive, which is not the case. Regarding the prediction that savings should be less in families with a firstborn of the preferred gender, we test this by estimating the impact of the firstborn's gender on the ability to deal with unexpected expenditures. Assuming that households with higher savings are more likely to respond positively to this question, the

estimate should be positive in Balkan countries and negative in Scandinavian countries. The estimates obtained in column (2), however, are small in magnitude and not statistically significant. For Balkan countries, this result could be reconciled with son preference by the fact that common savings are also a household public good and respond positively to the arrival of a child of the preferred gender, countering the negative effect of reduction in expected number of children.

Higher expenditure on household public goods may also be consistent with the comparative advantage a father has in raising sons, i.e. the so called "technology" explanation, according to Dahl and Moretti (2008). The gender bias and technology explanations have different implications for consumption patterns of fathers and mothers. The gender bias explanation suggests lower consumption of mothers of daughters while the technology explanation implies it to be higher. Specifically, if sons directly increase the utility of fathers, then a standard bargaining model of the household predicts a shift of household resources from fathers to mothers. This redistribution could be observable through lower consumption of private commodities by mothers of daughters.

Table 6: Impact of a first-born girl on employment consumption of mothers and fathers in the Balkans

	(1) Being employed	(2) Weekly hours of work	(3) Ability to spend on oneself	(4) Two pairs of shoes	(5) Replacing clothes	(6) Get together with friends	(7) Regular leisure activity
Mothers	-0.011	-0.369	-0.0233	-0.007	-0.007	0.006	0.032
	(0.006)*	(0.265)	(0.0117)**	(0.01)	(0.01)	(0.01)	(0.024)
Fathers	-0.006	-0.328	0.004	-0.002	-0.003	0.002	0.049
	(0.005)	(0.228)	0.011	(0.01)	(0.01)	(0.01)	(0.024)**

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: The standard errors of estimates on the sub-sample for Balkan countries are clustered at the country level. For details on sampling and estimation see the note under Table 5.

The negative impact of the mother's ability to spend on herself in Balkan countries in column (3) of Table 6 is in line with the gender bias explanation. In addition, two more facts hold for intrahousehold allocations in Balkan countries. First, mothers of daughters are less likely to be employed. Second, fathers of daughters report more time spent on leisure. The first could be explained by self-selection into unemployment of mothers whose comparative advantage in raising daughters results in an even greater opportunity cost than for similar mothers of sons (otherwise, first-born daughters would also negatively impact the intensive margin of mother's employment). Still, such self-selection of mothers into employment would not undermine our results, because the "technology" explanation implies lower progression to parity three when fathers have a sufficiently high comparative advantage in raising sons and a sufficiently wide wage gap in favor of men (Gugl and Welling, 2012). Despite the existence of a wide gender

wage gap, our estimates do not support the existence of a sizable comparative advantage of fathers in raising sons in the Balkans, which would be evident from fewer hours of work and higher personal consumption reported by fathers with first-born sons, as explained earlier. Finally, the fact that fathers have more leisure could be explained by longer hours of housework done by daughters.[71] Thus, the obtained results are consistent with the gender bias explanation for Balkan countries. For Scandinavian countries, there is no firstborn gender effect on either furniture replacement or the ability to deal with unexpected expenditures (the first two columns of Table 5). Moreover, the estimates of the firstborn's gender impact on parental consumption in Table 7 do not differ between fathers and mothers, which would be in line with parental comparative advantage.

Table 7: Impact of a first-born girl on employment and consumption of mothers and fathers in Scandinavia

	(1) Being employed	(2) Weekly hours of work	(3) Ability to spend on oneself	(4) Two pairs of shoes	(5) Replacing clothes	(6) Get together with friends	(7) Regular leisure activity
Mothers	0.005	0.439	-0.002	0.005	0.001	0.013	-0.060
	(0.005)	(0.185)**	(0.008)	(0.006)	(0.007)	(0.007)*	(0.031)**
Fathers	-0.007	-0.357	0.004	(0.004)	0.005	0.0003	-0.032
	(0.003)	(0.156)**	(0.008)	(0.006)	(0.007)	(0.007)	(0.030)

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: The standard errors of estimates on the sub-sample for Scandinavian countries are clustered at the country level. For details on sampling and estimation see the note under Table 5.

In other words, the difference in parental consumption between fathers and mothers points to the parental comparative advantage explanation.[72] This is because mothers of sons should redirect household resources to fathers to keep them in the family due to their important role in raising sons (Lundberg, 2005). At the same time, estimates of the impacts on the ability of mothers to meet with friends and family and to have regular leisure activity do not contradict the gender bias explanation *per se*. However, the estimated impacts on father consumption should be positive according to the gender bias explanation and it is not. Fathers of daughters work fewer hours, but they do not redirect that time to leisure. Moreover, the fewer hours worked by fathers of daughters is not likely to drive the observed daughter preference because similar effects were found in the US and West Germany (Lundberg and Rose, 2002; Choi et

[71] This is true for the 2010 ad-hoc sample from Romania and Bulgaria. The question about hours of housework was included in the 2010 EU-SILC ad-hoc module. However, since this was an optional question, and national statistical agencies chose whether or not to include it in the survey presented to their residents, this data is available only for 10 EU countries.

[72] It cannot be the main driving cause of the observed gender preference in Scandinavia because the gender wage gap should be in favor of women (Gugl and Welling, 2012) and that is not the case. Still, this result is consistent with a comparative advantage of intact families with daughters in producing "child quality".

al., 2008), which exhibit son preference (Dahl and Moretti, 2008; Hank and Kohler, 2000). All in all, the data does not support the gender bias explanation for Scandinavian countries.

The differential cost hypothesis is not confirmed by household-level estimates for Balkan countries. There are no statistically significant results in the last three columns of Table 3 for Balkan countries. Moreover, if expenditures on sons were higher, explaining the lower progression after a first-born son, parents of daughters would have more resources to spend on themselves. This is in contrast with the negative impact of the first-born daughter on private expenditure of mothers in column (3) of Table 5.

The Scandinavian results do show the expected higher outlays on daughters consistent with the differential cost explanation. Households with first-born daughters are more likely to have the entire set of ten important children consumption items. However, neither the ability to make ends meet nor the minimum amount of money to make ends meet depend on the gender of the firstborn. Nevertheless, for the top income decile, the minimum amount of money needed to make ends meet is larger for families with a first-born daughter.[73] Mothers of daughters appear to more frequently forgo regular leisure activity and substitute it with apparently less costly socialization through meeting with friends and family. Moreover, more hours worked by mothers of daughters suggest that they are willing to substitute leisure for outlays on daughters. At the same time, fathers of daughters tend to work less than fathers of sons. When Lundberg and Rose (2002) reported a similar effect for fathers from the US, they offered an explanation based on the son bias idea but did not formally test it. Our testing, however, does not support the son bias explanation. Furthermore, Norwegian data indicates that paternal leave has more pronounced positive effects for daughters than sons (Cools et al., 2015). That could be a reason fathers in Scandinavian countries substitute time spent on work for time spent on children (rather than leisure).[74] All in all, the differential expenses explanation of daughter preference in Scandinavian countries is supported by the data. Appendix Figure A2 and Figure A3 show cross-country relationships between gender preference, gender gap in parental investment, and conventional measures of gender equality. These relationships are in line with our previous points.[75]

[73] The argument as to why this should be true is developed in the Appendix (Figure A4 illustrates this idea).
[74] Examining data from detailed time-use surveys could shed more light on this issue.
[75] Specifically, Appendix Figure A3 shows that daughters tend to receive greater parental investment in countries with higher indicators of gender equality. This suggests that child household items for daughters are either cheaper or more useful in more gender-equal countries. Both situations are consistent with a lower cost of child human capital for daughters in countries with greater gender equality. Meanwhile, if the gender equality indicators at hand reflect a degree of gender bias and gender bias drives parental gender preference, Figure A2 should show negative relationships, which is not the case.

2.6 Conclusion

We find evidence that parental gender preferences in different countries are caused by different reasons. In Balkan countries, the observed son preference is likely driven by gender bias towards sons. In Scandinavian countries, the observed daughter-preference is likely driven by a lower cost of daughter quality, which incorporates gender-specific personal characteristics and their usefulness for parents. To measure the effect of the gender difference in the cost of children precisely, we would need to observe its random variation. Evidence of a lower cost for female human capital is most pronounced in more gender-equal societies, in line with trends of institutional change in modern societies in favor of women (Roberts and Baumeister, 2011). If this is not compensated by policies that reduce the cost of human capital for sons in less well-off families, the consequences mentioned in Edlund (1999) and Seidl (1995) might be realized.

References

Aaronson, D., Lange, F., and Mazumder, B. (2014). Fertility Transitions Along the Extensive and Intensive Margins. *American Economic Review*, 104(11):3701– 3724.

Adda, J., Dustmann, C., and Stevens, K. (2016). The Career Costs of Children. Working Paper 6158, CESifo.

Altindag, O. (2016). Son Preference, Fertility Decline, and the Nonmissing Girls of Turkey. *Demography*, 53(2):541–66.

Ananat, E. O. and Michaels, G. (2008). The effect of marital breakup on the income distribution of women with children. *Journal of Human Resources*, 43(3):611 – 629.

Andersson, G., Hank, K., Ronsen, M., and Vikat, A. (2006). Gender Family Composition: Sex Preferences for Children and Childbearing Behavior in the Nordic Countries. *Demography*, 43(2):255–267.

Angrist, J. D. and Evans, W. N. (1998). Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *American Economic Review*, 88(3):449–477.

Arnold, F. (1992). Sex Preference and Its Demographic and Health Implications. *International Family Planning Perspectives*, 18(3).

Arnold, F., Choe, M. K., and Roy, T. K. (1998). Son preference, the family-building process and child mortality in India. *Population Studies*, 52(3):301 – 315.

Autor, D., Dorn, D., and Hanson, G. (2017). When work disappears: Manufacturing decline and the falling marriage-market value of young men. Working Paper 21173, National Bureau of Economic Research. 34

Baranowska-Rataj, A. and Matysiak, A. (2016). The Casual Effects of the Number of Children on Female Employment—Do European Institutional and Gender Conditions Matter? *Journal of Labor Research.*

Barcellos, S. H., Carvalho, L. S., and Lleras-Muney, A. (2014). Child Gender and Parental Investments in India: Are Boys and Girls Treated Differently? *American Economic Journal: Applied Economics*, 6(1):157–189.

Basu, D. and De Jong, R. (2010). Son Targeting Fertility Behavior: Some Consequences and Determinants. *Demography*, 47(2):521–536.

Bedard, K. and Deschenes, O. (2005). Sex Preferences, Marital Dissolution and the Economic Status of Women. *Journal of Human Resources*, 40(2):411–434.

Ben-Porath, Y. and Welch, F. (1976). Do sex preferences really matter? *Quarterly Journal of Economics*, 90(2):285 – 307.

Blacklow, P. (2002). Intertemporal Equivalence Scales: Measuring the Life-Cycle Costs of Children. Memorandum 8, University of Oslo Department of Economics.

Bogan, V. L. (2013). Household investment decisions and offspring gender: parental accounting. *Applied Economics*, 31(45):4429–4442.

Bojer, H. (2002). The Time Cost of Children and Equivalent Full Incomes. Memorandum 8, University of Oslo Department of Economics.

Bond, T. N. and Lang, K. (2013). The Evolution of the Black-White Test Score Gap in Grades K-3: the Fragility of Results. *The Review of Economics and Statistics*, 95(5):1468–1479.

Bradbury, B. (2004). The Price, Cost, Consumption and Value of Children. Sup- porting children: English-speaking countries in international context, workshop presentation paper, Princeton University.

Bradley, R. H. and Caldwell, B. M. (1980). The Relation of Home Environment, Cognitive Competence, and IQ among Males and Females. *Child Development*, 51:1140–1148. 35

Bradley, R. H. and Caldwell, B. M. (1981). The HOME Inventory: A Validation of the Preschool Scale for Black Children. *Child Development*, 52:708–710.

Bradley, R. H. and Caldwell, B. M. (1984). The Relation of Infants' Home Envi- ronments to Achievement Test Performance in First Grade: A Follow-up Study. *Child Development*, 55:803–809.

Brockmann, H. (2001). Girls Preferred? Changing Patterns of Sex Preferences in the Two German States. *European Sociological Review*, 17(2):189–202.

Burman, D. D., Bitan, T., and Booth, J. R. (2008). Sex Differences in Neural Processing of Language Among Children. *Neuropsychologia*, 46(5):1349–1362.

Cameron, E. Z. and Dalerum, F. (2009). A Trivers–Willard effect in contemporary humans: male-biased sex-ratios among billionaires. *PLoS ONE*, 4(1).

Choi, E. J. and Hwang, J. (2015). Child Gender and Parental Inputs: No More Son Preference in Korea? *American Economic Review: Papers and Proceedings*, 105(5):638–643.

Choi, H.-J., Joesch, J. M., and Lundberg, S. (2008). Sons, daughters, wives, and the labor market outcomes of West German men. *Labour Economics*, 15:795–811.

Clark, S. (2000). Son Preference and Sex Composition of Children: Evidence from India. *Demography*, 37(1):95–108.

Cools, A. and Patacchini, E. (2017). Sibling Gender Composition and Women's Wages. Discussion Paper 11001, IZA.

Cools, S., Fiva, J. H., and Kirkeboen, L. J. (2015). Causal effects of paternity leave on children and parents. *Scandinavian Journal of Economics*, 117(3):801 – 828.

Cunha, F., Heckman, J. J., and Schennach, S. M. (2010). Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica*, 78(3):883–931.

Dahl, G. and Moretti, E. (2008). The Demand for Sons. *The Review of Economic Studies*, 75:1085–1120.

Davis, D. L., Gottlieb, M. B., and Stampnitzky, J. R. (1998). Reduced Ratio of Male 36 to Female Births in Several Industrial Countries: A Sentinel Health Indicator? *Journal of American Medical Association*, 279(13):1018–1023.

de la Croix, D. and Doepke, M. (2003). Inequality and Growth: Why Differential Fertility Matters. *American Economic Review*, pages 1091–1113.

Deaton, A. and Muellbauer, J. (1986). On measuring child costs: with applications to poor countries. *Journal of Political Economy*, 94:720–745.

Dujardin, C. and Goffette-Nagot, F. (2009). Does public housing occupancy increase unemployment? *Journal of Economic Geography*, 9:823–851.

Economist (2018). Why polygamy breeds civil war. https://www.economist. com/blogs/economist-explains/2018/03/economist-explains-16. Accessed: 2019-07-04.

Edlund, L. (1999). Son Preference, Sex Ratios, and Marriage Patterns. *Journal of Political Economy*, 107(6).

Estrin, S. and Uvalic, M. (2014). FDI into transition economies. Are the Balkans different? *Economics of Transition*, pages 281–312.

Eurobarometer, S. (2009). Gender equality in the EU in 2009.

Eurostat (2015). *ESS handbook for quality reports*. Luxembourg: Publications Office of the European Union.

Eurostat (2017). *European Union Statistics on Income and Living Conditions*. http://ec.europa.eu/eurostat/web/microdata/ european-union-statistics-on-income-and-living-conditions [Accessed: 18.11.17].

Filmer, D., Friedman, J., and Schady, N. (2009). Development, Modernization, and Childbearing: The Role of Family Sex Composition. *World Bank Economic Review*, 23(3):371–398.

Galor, O. (2011). *Unified growth theory*. Princeton : Princeton University Press.

Grech, V., Savona-Ventura, C., and Vassallo-Agius, P. (2002). Unexplained differences in sex ratios at birth in Europe and North America. *BMJ*, 324(7344):1010– 37 1011.

Gugl, E. and Welling, L. (2012). Time with sons and daughters. *Review of Economics of the Household*, 10:277–298.

Guilmoto, C. Z. and Duthe, G. (2013). Masculinization of births in Eastern Europe. Population & Society: Monthly bulletin 506, The French National Institute for Demographic Studies.

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, Gender, and Math. *Science*, 320:1164–1165.

Gupta, B. (2014). Where have all the brides gone? son preference and marriage in India over the twentieth century. *Economic History Review*, 67(1):1 – 24.

Hank, K. (2007). Parental Gender Preferences and Reproductive Behavior: A Review of the Recent Literature. *Journal of Biosocial Sciences*, 39:759–767.

Hank, K. and Kohler, H.-P. (2000). Gender Preferences for Children in Europe: Empirical Results from 17 FFS Countries. *Demographic Research*, 2(1).

Hao, L. and Yeung, W.-J. J. (2015). Parental Spending on School-Age Children: Structural Stratification and Parental Expectation. *Demography*, (52):835–860.

Haughton, J. and Haughton, D. (1998). Are simple tests of son preference useful? An evaluation using data from Vietnam. *Journal of Population Economics*, 11:495– 516.

Hazan, M. and Zoabi, H. (2015). Sons or Daughters? Sex Preferences and the Reversal of the Gender Educational Gap. *Journal of Demographic Economics*, 81:179–201.

Husain, M. and Millimet, D. L. (2009). The mythical 'boy crisis'? *Economics of Education Review*, 28(1):38 – 48.

Ichino, A., Lindstrom, A.-A., and Viviano, E. (2011). Hidden Consequences of a First-Born Boy for Mothers. Discussion Papers 5649, IZA.

Jacobsen, R., Møller, H., and Engholm, G. (1999). Fertility rates in Denmark in relation to the sexes of preceding children in the family. *Human Reproduction*, 14(4):1127–1130.

Jayachandran, S. (2017). Fertility Decline and Missing Women. *American Economic Journal: Applied Economics*, 9(1):118–139.

Jayachandran, S. and Kuziemko, I. (2011). Why Do Mothers Breastfeed Girls Less Than Boys? Evidence and Implications for Child Health in India. *The Quarterly Journal of Economics*, 126:1485–1538.

Jensen, R. T. (2003). Equal Treatment, Unequal Outcomes? Generating Sex In- equality Through Fertility Behavior. Working paper, Harvard University John F. Kennedy School of Government.

Jiang, Q., Li, Y., and Sanchez-Barricarte, J. J. (2016). Fertility Intention, Son Preference, and Second Childbirth: Survey Findings from Shaanxi Province of China. *Social Indicators Research*, 125:935–953.

Juhn, C., Rubinstein, Y., and Zuppann, C. A. (2015). The quantity-quality trade- off and the formation of cognitive and non-cognitive skills. Working Paper 21824, National Bureau of Economic Research.

Kanazawa, S. (2007). Beautiful parents have more daughters: A further implication of the generalized Trivers–Willard hypothesis (gtwh). *Journal of Theoretical Biology*, 244(1):133 – 140.

Karbownik, K. and Myck, M. (2017). Who gets to look nice and who gets to play? Effects of child gender on household expenditures. *Review of Economics of the Household*, 15:925–944.

Kippen, R., Evans, A., and Gray, E. (2006). Parental Preference for Sons and Daughters in a Western Industrial Setting: Evidence and Implications. *Journal of Biosocial Science*, 39:583–597.

Kornrich, S. and Furstenberg, F. (2007). Investing in Children: Changes in Parental Spending on Children, 1972 to 2007. Unpublished paper.

Larsen, U., Chung, W., and Das Gupta, M. (1998). Fertility and son preference in Korea. *Population Studies*, 52:317–325.

Leung, S. F. (1991). A Stochastic Dynamic Analysis of Parental Sex Preferences and Fertility. *Quarterly Journal of Economics*, pages 1063–1088.

Li, L. and Wu, X. (2011). Gender of Children, Bargaining Power, and Intrahousehold Resource Allocation in China. Journal of Human Resources, pages 295–316.

Lindstrom, E.-A. (2013). Gender Bias in Parental Leave: Evidence from Sweden. *Journal of Family and Economic Issues*, 34:235–248.

Lino, M. and Carlson, A. (2010). Estimating Housing Expenses on Children: Comparison of Methodologies. *Journal of Legal Economics*, 16(2):61–79.

Lundberg, S. (2005). Sons, Daughters, and Parental Behavior. *Oxford Review of Economic Policy*, 21(3).

Lundberg, S., McLanahan, S., and Rose, E. (2007). Child Gender and Father Involvement in Fragile Families. *Demography*, 44(1):79–92.

Lundberg, S. and Rose, E. (2002). The Effects of Sons and Daughters on Men's Labor Supply and Wages. *The Review of Economics and Statistics*, 84(2).

Lundberg, S. and Rose, E. (2003a). Child Gender and the Transition to Marriage. *Demography*, 40(2):333–349.

Lundberg, S. and Rose, E. (2003b). Investments in Sons and Daughters: Evidence from the Consumer Expenditure Survey. Unpublished manuscript, Joint Center for Poverty Research.

Mammen, K. (2008). The Effect of Children's Gender on Living Arrangements and Child Support. *American Economic Review: Papers and Proceedings*, 98(2).

McKenzie, D. J. (2005). Measuring inequality with asset indicators. *Journal of Population Economics*, 18:229–260.

Norberg, K. (2004). Partnership status and the human sex ratio at birth. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1555), 2403-2410.

Outram, Q. (2015). Son Preference Among the Edwardian Middle Classes. *Available at SSRN 2694948.*

Pabilonia, S. W. and Ward-Batts, J. (2007). The Effect of Child Gender on Parents' Labor Supply. *American Economic Review: Papers and Proceedings*, 97(2):402– 406.

Peters, J. F. (1994). Gender socialization of adolescents in the home: Research and discussion. *Adolescence*, 29(116), 913.

Pollard, M. S. and Morgan, S. P. (2002). Emerging parental gender indifference? sex composition of children and the third birth. *American Sociological Review*, 67(4):600 – 613.

Ragan, K. (2013). Taxes and Time Use: Fiscal Policy in a Household Production Model. *American Economic Journal: Macroeconomics*, pages 168–192.

Raurich, X. and Seegmuller, T. (2017). Growth and bubbles: the interplay between productive investment and the cost of rearing children. Working Paper 26, AMSE.

Roberts, R. and Baumeister, R. (2011). Baumeister on gender differences and culture.

Sadowski, M. (2010). Putting the "Boy Crisis" in Context. *Education Digest: Essential Readings Condensed for Quick Review*, 76(3):10–13.

Sandstrom, G. and Vikstrom, L. (2015). Sex preference for children in German villages during the fertility transition. *Population Studies*, 69(1):57–71.

Seidl, C. (1995). The Desire for a Son Is the Father of Many Daughters: A Sex Ratio Paradox. *Journal of Population Economics*, 8(2).

Sienaert, A. (2008). Some Child Cost Estimates for South Africa. Working Paper Series 15, CSAE.

Todd, P. E. and Wolpin, K. I. (2007). The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps. *Journal of Human Capital*, 1(1):91–136.

Trivers, R. L. and Willard, D. E. (1973). Natural selection of parental ability to vary the sex ratio of offspring. *Science*, 179(4068):90–92.

van Praag, B. M. and Warnaar, M. F. (1997). Chapter 6 the cost of children and the use of demographic variables in consumer demand. In Rosenzweig, M. and Stark, O., editors, *Handbook of Population and Family Economics*, volume 1, pages 241 – 273. Elsevier.

Vogl, T. S. (2013). Marriage institutions and sibling competition: Evidence from south asia. *Quarterly Journal of Economics*, 128(3):1017 – 1072.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.

Xu, J. (2016). Patriarchy, Gendered Spheres, or Evolutionary Adaptation? A Cross-National Examination of Adolescent Boys and Girls Access to Home Resources. *Chinese Sociological Review*, 48(3):209–247.

Yamaguchi, K. and Ferguson, L. R. (1995). The Stopping and Spacing of Childbirths and their Birth-History Predictors: Rational-Choice Theory and Event-History Analysis. *American Sociological Review*, 60:272–298.

Yoon, Y. J. (2006). Gender Imbalance: The Male/Female Sex Ratio Determination. *Journal of Bioeconomics*, 8:253–268.

Appendix

The distinction between the gender bias and differential costs concepts

In the literature, there is neither a clear-cut definition of what we have designated as gender bias nor a conventional term for labeling it. In some cases, gender bias is readily recognizable. For example, Arnold et al. (1998) assert that some Indian parents prefer sons for reasons connected with religious beliefs and kinship descent, whereas Jacobsen et al. (1999) argue that women's need for companionship leads to daughter preference in Denmark. Characteristics, like continuing the family name or providing the same-gender companionship to parents, are intrinsically pertinent to the gender of a child and their utility does not directly depend on the parental outlays on children. Preferences for such characteristics are captured by the first part of Lundberg's 2005 definition, because a son has a greater marginal value in the first case and a daughter in the second. This understanding is consistent with other previously provided definitions. In other situations, the gender bias is less recognizable. One possible example is the case of a man who wants a son because the boy may be a player in his favorite soccer team. Yet, the father cannot do much to bring this about beyond encouraging him or taking him to a local soccer academy. Had this man had a daughter instead of a son, he would likely have done not much less for her physical development. Similarly, parents might want a daughter, because she can become a soprano singer. These examples are captured by the second part of the aforementioned definition. That is, the man values a son's soccer skills more than a daughter's, because they increase the chances of the son becoming a player in the father's favorite team. While in the second example, parents value a daughter's singing skills more than a son's, because the son's soprano will eventually disappear. In both cases, parents would not need to invest much (parental time and tuition at a soccer academy or music school), provided the children have sufficient aptitude. A common feature of these examples is the absence of a close relationship between the parental investment of time and market goods on one side and child quality (desired characteristics) on the other side beyond some relatively low level of investment.

An alternative example could be parents who want a household member to know a foreign language. One way to proceed is to have a child who would learn that language. On average, it would be cheaper with a daughter, because girls are known to be better at picking up foreign languages (Burman et al., 2008). Here, the more parents invest in a child's language learning, the better the result (hours with tutors, educational trips abroad, etc.). Keeping other things equal, these parents are likely to invest significantly more in a daughter's language learning, because of greater marginal returns on their investment. We understand such situations as cases of differences in costs of children.

Considerations about using the gender of the firstborn as the instrumental variable

Some authors claim that the gender of the firstborn is not random. For example, Norberg (2004) reports that children who were conceived when their mother was living with a partner were 14% more likely to be boys than siblings conceived when the parents were living apart. This finding aligns with the falling gender ratio in a set of industrialized countries (Davis et al., 1998). One possible explanation for these findings is the evolutionary advantage of species that can adjust the gender ratio of offspring in response to changes in conditions affecting the relative reproductive success of males and females (Trivers and Willard, 1973). Furthermore, the wealthiest individuals in societies tend to have sons born more frequently (Cameron and Dalerum, 2009). To address these concerns, we repeat our analysis on the sample of partners cohabiting at the time when the firstborn arrived, control for country fixed-effects, and repeat the analysis after dropping the top 1% of the wealthiest households in each country from the sample.[76] At the same time, the gender of the firstborn might impact marital stability (Lundberg and Rose, 2003a; Mammen, 2008; Lundberg et al., 2007), family size (Hank and Kohler, 2000; Angrist and Evans, 1998), and parental time allocation (Lundberg and Rose, 2002; Lindstrom, 2013; Choi et al., 2008). This makes "exclusion restrictions a priori unpersuasive" (Lundberg, 2005). To solve this problem, we focus our analysis on the sample of intact families, instrument the number of children with twin-births, and argue that the impact of the gender of the firstborn on parental employment does not notably alter our estimates or their statistical significance.

The documented impact of the gender of firstborns on parental employment differs across countries. For example, a first-born son increases a father's work hours in the US by 3% of the mean male work hours more than for fathers with a first-born daughter (Lundberg and Rose, 2002). However, Pabilonia and Ward-Batts (2007) find one third of the same effect and not at a statistically significant level. An even larger effect, almost 5% of mean annual male work hours, was found in West Germany (Choi et al., 2008). Meanwhile, Ichino et al. (2011) find a negative impact of a first-born son on a mother's working hours and employment in the US, UK, and Italy. This is still smaller than the previously mentioned effect for fathers and hovers across the countries at around 1% of the mean. Lindstrom (2013) finds that a first-born son increases paternity leave by 0.6 days (1.5%) and decreases maternity leave by a similar amount. In our analysis, we do find that the gender of a firstborn affects the employment status of mothers. However, we do not find an effect on their work hours or on father's employment status or work hours. The negative effect of a first-born son on a mother's employment is approximately 1% of mean female employment. This is in line with previously reported estimates from the literature. However, when we multiply this effect on employment with its

[76] One study (Kanazawa, 2007) reports that physically more attractive parents are significantly more likely to have a daughter. We are not aware of other studies confirming this finding.

coefficient, the final effect on the variable of interest is by an order of magnitude smaller than the direct effect of the first-born gender variable. This is why, following Karbownik and Myck (2017), we believe that the impact on employment does not undermine our estimates of interest and as such we keep the employment status and workload of parents as covariates.

A description of the material deprivation measures

The EU-SILC ad-hoc modules on material deprivation from 2009 and 2014 each contain thirteen questions about the availability of child items and amenities (the module from 2009 contained questions on 22 items, but the recent module was reduced). Each question corresponds to a variable that indicates the presence of a specific item or amenity. Specifically, the variables are: replace worn-out clothes; two pairs of properly fitting shoes; fresh fruit and vegetables once a day; one meal with fish, chicken, or meat (or vegetarian equivalent) at least once a day; books at home suitable for children's age; outdoor leisure equipment; indoor games; regular leisure activity; celebrations on special occasions; invite friends home to play and eat from time to time; participate in school trips and school events that cost money; suitable place to study or do homework; and go on holiday away from home at least 1 week per year. In our analysis, we primarily only use the first ten questions, because they are available for nearly all children in the sample, while the last three are available only for school-age children. These questions do not completely correspond to the questions from other surveys on material conditions of children that have been analyzed in the literature, e.g., NLSY79-CS HOME-SF module (Cunha et al., 2010; Todd and Wolpin, 2007; Juhn et al., 2015) and PISA-2000 Xu (2016). Those surveys are more extensive. Instead, the ten questions we consider largely overlap with the resources-spent and time-with-child subcomponents defined by Juhn et al. (2015) based on the NLSY79 survey. For instance, all questions in the resources-spent and some questions from the time-with-child subcomponents of Juhn et al. (2015) are contained in EU-SILC ad-hoc modules from 2009 and 2014. All in all, the EU-SILC ad-hoc modules considered here could be seen as extended versions of the two subcomponents mentioned above, and since elements in these two subcomponents were highly correlated with child development (Bradley and Caldwell, 1980, 1981, 1984) and strongly influencing it (Cunha et al., 2010), the raw score of the EU-SILC ad-hoc modules should also be correlated with and have an impact on child development. Furthermore, the responses from the PISA-2000 survey analyzed by Xu (2016) contain more detailed information, but correspond directly with the EU-SILC questions on participating in regular leisure activity, availability of a suitable place to study, and having books at home. Xu argues that precisely those items are important for a child's adult outcomes and supports the point by referring to multiple related studies.

To test for a gender-gap in children's material conditions at home, we use five alternative dependent variables in equation 1 for measuring material condition. The first is a pure sum of

the binary indicators of the presence of the first ten material conditions listed in the previous paragraph. This sum corresponds to the so-called HOME index used in the literature. One problem with this variable is susceptibility to monotonous transformations, also known as the scaling problem (Bond and Lang, 2013). Another problem is that all the items in that dependent variable are assigned equal weights in summation, which means that those with larger variance contribute more to the estimated effect. We attempt to overcome these problems by constructing four other measures of material condition. First, we conduct the principal component analysis (PCA), where the first principal component (the one with the most variance) obtained from this analysis is used as an alternative dependent variable. In this way, we follow Cools and Patacchini (2017) who also construct a measure for material conditions of children albeit based on a different dataset, using different indicators, and addressing a different research question. The rationale behind the method is elaborated, for example, by McKenzie (2005). He applies this method to measuring household wealth inequality based on responses about the availability of different items. Importantly, he demonstrates that there is invariance of this measure across linear transformations. Additionally, we use ordered probit and Poisson models with the raw sum of ten indicators as the dependent variable. In this case, however, we assume that households acquire the most necessary child items first. The probit and the Poisson regressions measure the probabilities of acquiring the next most necessary items. Finally, the frequency histogram of the raw sum of indicators shows that around one-half of households possess all ten items. Therefore, we introduce one more binary alternative dependent variable. It takes a value of 1 for households which possess all specified items and a value of 0 for the other households. This specification of the dependent variable is the most intuitively appealing to us and we rely on it in the main analysis. Nevertheless, under all specifications of the dependent variable, the results of the analysis are qualitatively similar and the estimated coefficients of primary interest are statistically significant.

Cross-country comparison of gender preference and parental investment

Table A3 displays the results of estimating gender preference by country. The geographical pattern of the gender preference at birth is depicted in Figure A6. Our results are broadly consistent with those previously obtained in the literature. Similarly to Hank and Kohler (2000), we find son preference in Italy and France and daughter preference in Portugal and Lithuania. Similar to Andersson et al. (2006), we also find daughter preference in Norway, though not in Sweden.[77] We also attempt to evaluate how our results would differ if there were no family disruptions caused by the gender of children, which is frequently reported in the

[77] Still, our estimates are correlated with (ρ=0.6) and statistically significantly predict comparable estimates to Hank and Kohler (2000)

literature (see, e.g., Lundberg, 2005, for a review). The results are presented in Table A1. Son preference becomes statistically significant in Slovenia and stops being statistically significant in Croatia. However, the estimates obtained after including Slovenia and excluding Croatia from son preferring countries do not differ qualitatively and do not differ much quantitatively from those reported here. The rank correlations between the country-level impacts of the firstborn's gender on the selected household fertility outcomes are presented in Table A3. The absence of a strong correlation between estimated impacts on progression to parity two and parity three suggests different factors driving these impacts, as we expected above.

Two measures of the same variable should be correlated, yet the correlations between second-parity coefficients and third-parity coefficients are quite low (Tables A2 & A3). Still, the last two sets of coefficients are strongly correlated with coefficients for the total number of children. This might spur an examination of whether it is proper to use third parity progression for measuring gender preference, which is a frequent practice in the literature.

To rationalize the estimates obtained, we plot the coefficients against several existing measures of gender inequality. As Figure A2 shows, the estimates do not exhibit a strong relationship with those measures. Only the coefficients from the third-parity equation exhibit a negative relationship with our gender equality score based on Eurobarometer data and with the proportion of households reporting balanced decision-making. At the same time, neither the coefficients for the total-number nor the second-parity equations exhibit any such relationship. This fact again suggests that second parity progression and third parity progression actually measure different kinds of preferences. This is why we use third parity progression results in Figure A6 and beyond.

In addition, the fact that parents tend to invest more in daughters as measured by the presence of home items[78] hold for the pooled EU-SILC sample. To test for the gender gap in parental investment, we estimate Equation 1 with several alternative measures of child material conditions on the LHS. We primarily focus on the specification with the binary home indicator (the dummy variable for all 10 items) on the LHS. Table A8 displays estimates for this specification on a pooled sample. The results suggest that daughters, on average, receive more parental investment in terms of home items. For example, the number in column 1 means that families with first-born girls are 1.5% more likely to have all 10 items. This estimate is robust to the alternative sets of covariates, as can be seen in the rest of Table A8. Still, this effect is not large, remaining between 1.7% and 2% of the standard deviation of the binary home indicator. Results of this scale are typical in the literature on gender effects. Meanwhile, the gender preference pattern established before holds for the sub-sample of households from the highest income decile. These results might suggest that society as a whole is attaching increasingly positive significance to female children, an idea that has appeared in previous

[78] Availability of these indicators has been frequently used in the literature as a measurement of parental investment. A more detailed discussion is presented in the previous Appendix section as well as in Sections 2.2 and 2.4.

studies, such as Brockmann (2001) and Andersson et al. (2006). A daughter may assume both the role of a breadwinner and that of a caregiver.[79] As Brockmann (2001, p. 199) puts it, "in the future, the average girl may well wish to become the mother of a one-daughter family."

As with the estimates of the preference for gender of children at birth, we relate the estimates of the gender gap in parental investment to specific country-level measures of gender inequality. The impact of the gender of the firstborn on material conditions exhibits a much stronger relationship with conventional measures of gender inequality than the impact on parity progression. Figure A3 displays the three strongest relationships. Most importantly, there is a strong relationship with the Global Gender Gap (GGG) score, calculated by the World Economic Forum (we used the most recent 2016 data). This index is also strongly related to the gender gap in PISA math achievement (Guiso et al., 2008).

However, Xu (2016) did not find any strong relationship between the gender gap in the home environment measure (similar to ours) and the GGG, though he measured the gender gap by the difference in the unconditional mean between genders.

Moreover, as explained earlier, our measure is preferable to the one used in Xu (2016). Therefore, the gender gap in child material conditions more closely corresponds to conventional gender-inequality measures than the gender gap in the number of younger siblings.[80] Nevertheless, the latter is commonly used as a measure of parental gender preference.

[79] In this regard, some authors speak about the "boy crisis" (Husain and Millimet, 2009; Sadowski, 2010).

[80] A similar and statistically significant relationship also holds between the first-daughter coefficient in the material-conditions regression and two other indexes: the GDI (it highly correlates with the GGG) and the SIGI (though it is available only for seven countries from our sample).

Table A1: Coefficients corrected for selection bias

Cntrs.	Coefs.	Cntrs.	Coefs.	Cntrs.	Coefs.	Cntrs.	Coefs.
AT	0.006	EE	-0.0007	IS	-0.003	PL	-0.003
BE	0.0003	EL	-0.006	IT	0.011***	PT	-0.017***
BG	0.0217***	ES	-0.001	LT	-0.006	RO	0.024***
CH	0.002	FI	0.004	LU	0.003	RS	0.029**
CY	-0.016*	FR	0.007	LV	-0.002	SE	0.010
CZ	0.002	HR	0.027*	MT	-0.010	SI	0.012**
DE	0.006	HU	-0.008*	NL	-0.004	SK	0.010
DK	-0.017**	IE	0.007	NO	-0.018**	UK	0.0007

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: The estimates contained in this table do not differ from those in the third column of Table A3 except in the sample characteristics and omission of father-related control variables (which have little explanatory power). The sample also includes incomplete families with simulated numbers of additional children—simulated under the assumption that those divorced because of the gender of children are characterized by bias towards that gender and do not stop producing more children until they have a child of the desired gender.

Table A2: Effects of firstborn gender on selected measures of fertility

Explanatory var-s	(1) Total number of children	(2) Two or more children	(3) Three or more children	(4) Four or more children	(5) Five or more children
First child a girl	-0.0050**	-0.0073***	0.0011	0.0004	0.0005*
	(0.0025)	(0.0017)	(0.0012)	(0.0006)	(0.0003)
Controls	Yes	Yes	Yes	Yes	Yes
First boy baseline	1.54	.406	.106	.0248	.00462
Percent effect	-.00323	-.0179	.0102	.018	.109
R-sq	.27	.235	.137	.0491	.0163
Observations	265,507	265,507	265,507	265,507	265,507

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: S.E. are given in parentheses and are clustered at the country level. Estimates are based on the 2004-2015 EU-SILC samples for Bulgaria, Croatia, Republic of Serbia, and Slovenia. The sample consists of households formed by one cohabiting couple, their children, and, occasionally, other relatives. The mother of children in the household is younger than 41 and older than 17 and had her first child at the age of 16 or older, and child ages are in the range 0-12. The estimation method used is weighted OLS with probability weights reflecting non-random sampling within and between countries. The table presents estimated effects of the firstborn being a daughter compared with the baseline case of the firstborn being a son. The effect is a ratio of the estimated OLS coefficient on the firstborn gender dummy to the baseline value of the dependent variable. The dependent variables are the total number of children and a set of binary indicators for specific numbers of children. The control variables, besides the gender of the firstborn, are the dummy for a first-born daughter, gender of the first two children, cubic polynomial of mother's age, squared polynomial of mother's age at first birth, length of cohabitation of spouses, mother's education, father's education, mother's employment, father's employment, household disposable income, living in urban area, tenure status, and year and country dummies.

Table A3: Effects of firstborn gender on selected measures of fertility

Countries[a]	(1) Total number of children	(2) Two or more children	(3) Three or more children	(4) Four or more children	(5) Five or more children	Obs
AT	-0.0181	-0.0245*	0.0083	-0.0050	0.0015	6,574
BE	-0.0074	-0.0139	0.0054	0.0007	0.0004	7,694
BG	0.0206	-0.0112	0.0222**	0.0096*	0.0011	3,509
CH	0.0353	0.0364**	0.0013	0.0013	-0.0017	4,461
CY	-0.0422*	-0.0330**	-0.0125	0.0032	0.0002	5,675
CZ	-0.0123	-0.0167*	0.0037	-0.0002	0.0001	10,329
DE	-0.0141	-0.0179*	0.0060	-0.0012	-0.0010	9,790
DK	-0.0183	-0.0023	-0.0178*	0.0012	0.0007	7,889
EE	-0.0147	-0.0091	-0.0032	0.0027	-0.0017	6,594
EL	-0.0040	-0.0075	-0.0065	0.0045	0.0041***	8,147
ES	-0.0292**	-0.0277***	-0.0030	0.0003	0.0008	16,054
FI	-0.0027	-0.0031	0.0070	-0.0000	-0.0011	13,145
FR	0.0209*	0.0102	0.0072	0.0005	0.0029**	14,496
HR	0.0878**	0.0507*	0.026**	0.0127	0.0031	1,742
HU	-0.0082	0.0057	-0.0137**	-0.0027	0.0015	11,281
IE	0.0002	0.0094	0.0030	-0.0074	-0.0007	5,636
IS	-0.0059	0.0009	-0.0022	-0.0028	-0.0014	5,711
IT	0.0091	-0.0032	0.0121***	-0.0004	0.0002	21,486
LT	-0.0352	-0.0096	-0.0090	-0.0098**	-0.0040*	3,742
LU	-0.0068	-0.0069	0.0022	0.0020	-0.0029*	8,084
LV	-0.0172	-0.0204	-0.0020	0.0028	0.0008	5,102
MT	-0.0170	-0.0013	-0.0118	-0.0019	-0.0013	2,872
NL	0.0021	0.0039	-0.0033	-0.0001	0.0001	11,942
NO	-0.0385**	-0.0210*	-0.0191*	0.0006	0.0007	8,108
PL	0.0049	-0.0037	-0.0008	0.0023	0.0035**	18,374
PT	-0.0794	-0.0486***	-0.0216***	-0.0074**	-0.0008	6,044
RO	0.0293	0.0028	0.0218**	0.0075*	-0.0027	4,948
RS	0.0619	0.0378	0.0214	0.0044	-0.0017	1,221
SE	0.0240	0.0112	0.0114*	0.0019	-0.0006	9,228
SI	0.0140	-0.0147	0.0113	0.0093***	0.0036***	10,544
SK	0.0191	-0.0025	0.0093	0.0072*	0.0018	5,802
UK	-0.0155	-0.0104	0.0034	-0.0085*	-0.0012	9,288

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: See notes for Table A2 for data samples, variable definitions, and included control variables. The columns contain estimated country-level effects of firstborn daughters on the corresponding variables in the column headings.

[a] Table A4 contains names of countries corresponding to the abbreviations.

Table A4: Abbreviations for countries

Abbrev.	Countries	Abbrev.	Countries	Abbrev.	Countries	Abbrev.	Countries
AT	Austria	EE	Estonia	IS	Iceland	PL	Poland
BE	Belgium	EL	Greece	IT	Italy	PT	Portugal
BG	Bulgaria	ES	Spain	LT	Lithuania	RO	Romania
CH	Switzerland	FI	Finland	LU	Luxembourg	RS	Republic of Serbia
CY	Cyprus	FR	France	LV	Latvia	SE	Sweden
CZ	Czech Republic	HR	Croatia	MT	Malta	SI	Slovenia
DE	Germany	HU	Hungary	NL	Netherlands	SK	Slovak Republic
DK	Denmark	IE	Ireland	NO	Norway	UK	The United Kingdom

Source: Eurostat

Table A5: Impact of the first-born daughter on selected household allocation decisions under two alternative explanations of the parental gender preference

Allocation decisions	Bias		Intact family advantage	
	toward sons	toward daughters	in raising sons	in raising daughters
Household public goods expenditure	-	+	.	.
Savings	+	-	.	.
Personal well-being of a father	+	-	-	+
Personal well-being of a mother	-	+	+	-

Notes: The sign "+" means a positive impact and the sign "-" means a negative impact. The rationale behind the predictions is explained primarily in the Introduction and also in Sections 3 and 4.

Table A6: Impact of the first-born daughter on selected household allocation decisions under two alternative explanations of the parental gender preference

Allocation decisions	Balkan countries		Scandinavian countries	
	Bias towards sons	Intact family comparative advantage in raising sons	Bias towards daughters	Intact family comparative advantage in raising daughters
Household public goods expenditure	−	.	+	.
Savings	+	.	-	.
Personal well-being of a father	+	-	-	+
Personal well-being of a mother	−	+	+	−

Notes: The sign "+" means a positive impact and the sign "-" means a negative impact. The rationale behind the predictions is explained primarily in the Introduction and also in Sections 3 and 4.

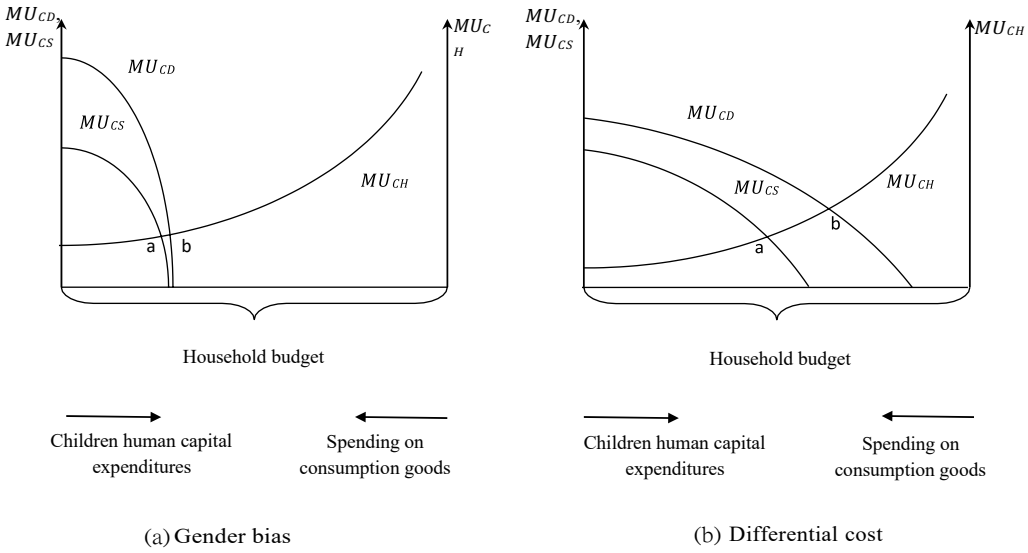


Figure A1: Graphical distinction between cases of gender bias and differential costs

Notes: The graphs show marginal parental utilities of human capital expenditures on children, MU_{CS} and MU_{CD}, together with accompanying marginal utility of household consumption expenditures, MU_{CH}. An underlying unitary household model is assumed. Human capital expenditure is on the horizontal axis and household marginal utility is on the vertical axis. Marginal utility of household consumption increases as expenditures on household consumption decrease, which occurs along the horizontal axis as human capital expenditures on children increase. On the left graph, marginal utilities of human capital expenditures on children plummet quickly, and parental investments are low and do not differ significantly between genders. At the same time, the difference in parental utility derived from children of different genders is significant. This is a graphically depicted example of gender bias. On the right graph, the marginal utility of investment in a child of some gender is notably larger along a broad range of possible investment volumes. The optimal volumes of investment differ considerably between children of different genders. This is a graphically depicted example of differential cost.

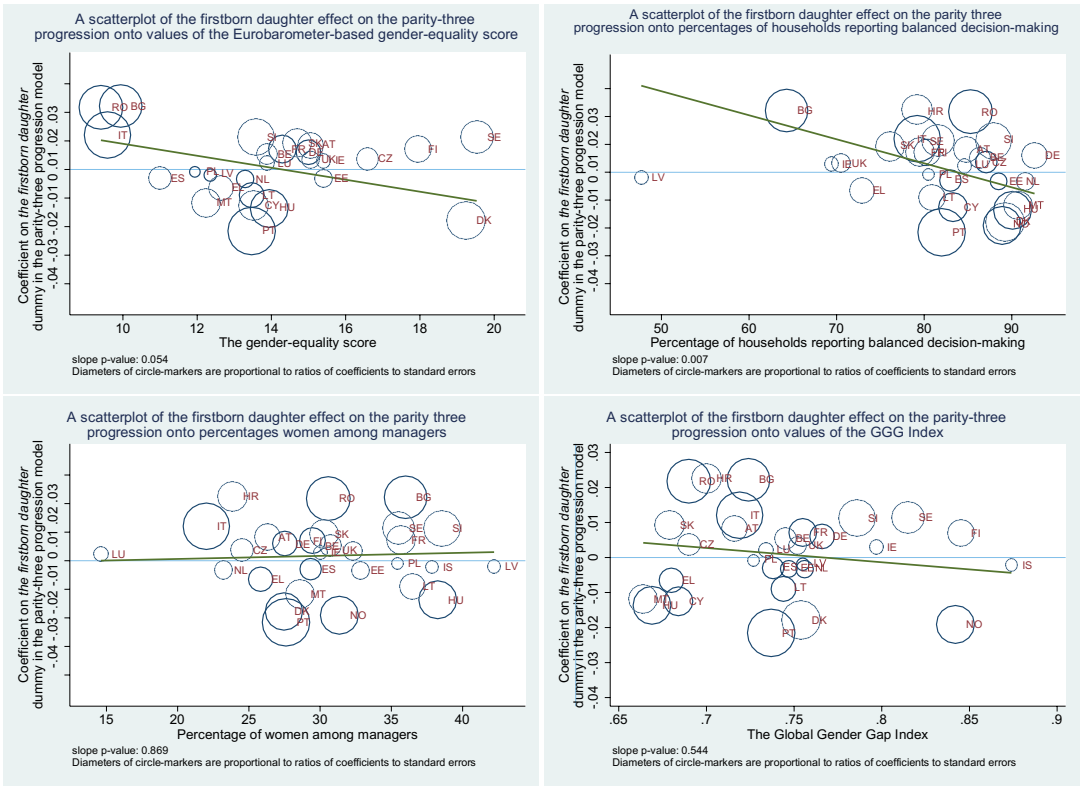


Figure A2: The relationship between the effect of first-born daughters on third parity progression and specific gender-equality measures across countries.

Notes: We calculate the Eurobarometer-based gender equality score for a particular country as a sum of the country's ranks in responses to questions about attitudes towards gender equality. These responses were collected in the 2009 Eurobarometer special survey (Eur, 2010). For each question, countries were ordered according to shares of respondents who report an existence/wish to exist in gender-egalitarian conditions in a specified realm of life. The country with the highest share of such respondents was assigned rank 1 for the corresponding question. We then calculated the sums of such ranks across all 13 pertinent questions and our gender-equality score. Please note that we do not have scores for Switzerland, Croatia, Iceland, Norway, and the Republic of Serbia, because the Eurobarometer survey was not conducted in those countries. Percentages of households reporting balanced decision-making were taken from the data of the Health and Demographic Survey collected by the World Bank in multiple years and from the Survey of Income and Living Conditions collected by Eurostat in 2010. The percentage of women managers was obtained from the data of the Enterprise Surveys, conducted by the World Bank in multiple years. The Global Gender Gap Index was calculated by the World Economic Forum in 2016.

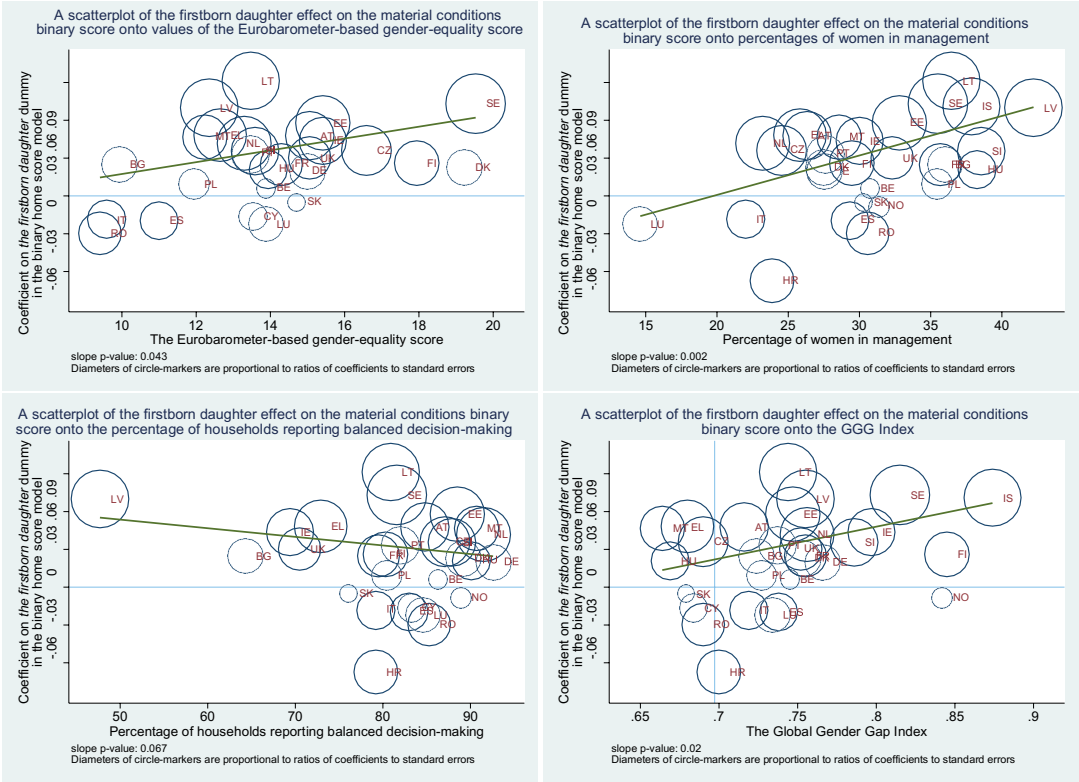


Figure A3: The relationship between the effect of first-born daughters on child material conditions and specific gender-equality measures across countries.

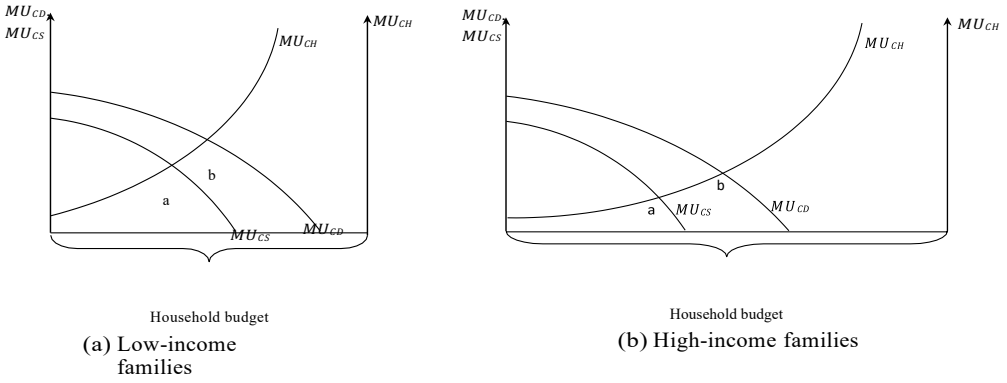


Figure A4: Differences in expenditures on children between low-income and high-income households

Notes: See the note to Figure A1 for explanation.

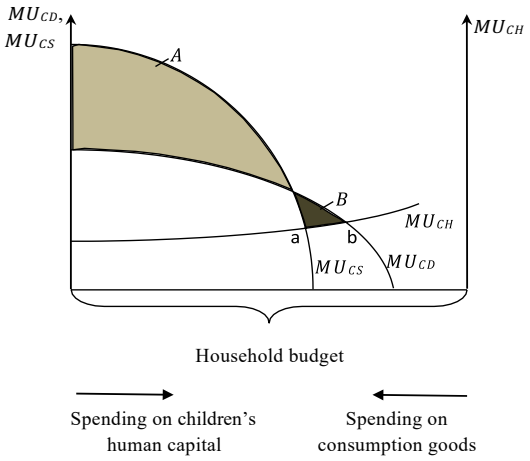


Figure A5: Coexistence of gender (son) bias and differential cost with the gender bias effect on fertility prevailing.

Notes: See the note to Figure A1 for explanation.

Table A7: Spearman's rank correlations between country-level effects of first-born daughters on selected measures of fertility

	Total number of children	Progression to parity two	Progression to parity three	Progression to parity four	Progression to parity five
Total number of children	1				
Progression to parity two	0.8380***	1			
Progression to parity three	0.7878***	0.4765***	1		
Progression to parity four	0.4758***	0.2753	0.3680**	1	
Progression to parity five	0.0037	-0.1334	-0.0169	0.2834*	1

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: Spearman's rank correlations are based on estimates for 32 European countries covered in the EU-SILC survey during 2004-2015. The estimates are contained in Table A3.

Table A8: The impact of the firstborn gender on the binary material deprivation indicator

	The binary material deprivation indicator on the LHS			
Explanatory var-s	(1) OLS	(2) OLS	(3) OLS	(4) IV
First child a girl	.015***	.0148***	.0168***	.0172***
Number of children		.0896***	.0797***	-.0231*
Covariates	No	No	Yes	Yes
R-Square	.000225	.0191	.168	.146
N obs	51,087	51,087	49,922	49,922

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Notes: The standard errors of estimates on pooled EU-SILC sample are clustered at the country level. The table presents estimated effects of the firstborn being a daughter compared with the baseline case of the firstborn being a son. Estimates are based on the 2009 and 2013-2015 EU-SILC ad-hoc modules, while the estimates in the remaining columns are based on the 2004-2015 EU-SILC primary modules. The sample consists of households formed by one cohabiting couple, their children, and, occasionally, other relatives. The mother of children in the household is younger than 41 and older than 17 and had her first child at the age of 16 or older, and children's ages are in the range 0–12. The estimation method used is weighted OLS with probability weights reflecting non-random sampling within and between countries. The dependent variable is the binary indicator taking value 1 when a household has all 10 material condition items listed earlier 0 otherwise. Other control variables are the dummy for a first-born daughter, gender of the first two children, cubic polynomial of mother's age, squared polynomial of mother's age at first birth, length of cohabitation of spouses, mother's education, father's education, mother's employment, father's employment, household disposable income, living in urban area, tenure status, and year and country dummies. The estimates in the fourth column are obtained using the 2SLS method from a regression-model in which the number of children is instrumented with twin-birth. The first stage F-statistic value for this model is above two thousand.

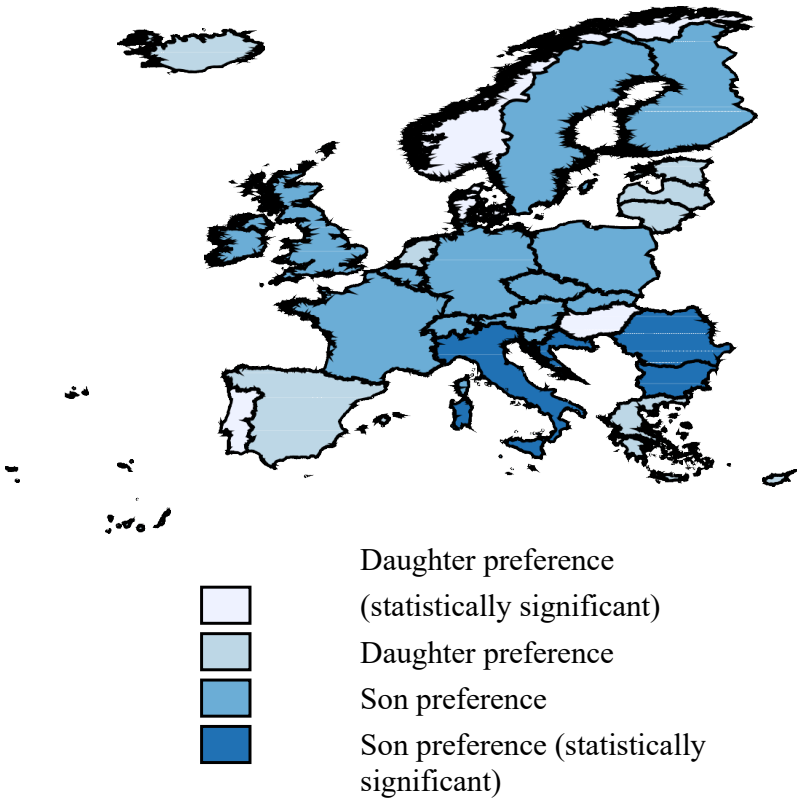


Figure A6: Gender Preferences of Children in 31 EU-SILC countries

3 Was a One Hour Adjustment in Georgian Public Sector Working Hours "Family Friendly" and Did It Increase Female Labor Participation?

Coauthored with Levan Bezhanishvili and Zurab Abramishvili

3.1 Introduction

On August 1, 2014, the prime minister of the Republic of Georgia announced a countrywide initiative to shift the working hours in the public sector[81] from 10:00-19:00 to 9:00-18:00 (Khunashvili, 2014). There was no parliamentary pushback, no protests by citizens or public employees, no journalistic coverage beyond the announcement, and no related Google search keyword trends. Within a month, it passed through parliament and was enacted on September 1, 2014. This example is as close to a theoretical one-time, immediate policy shift as practically possible. Officially, the rationale of the new policy was to adjust the working hours of public offices to those more common in "the modern world". In fact, the policy was one of several that aligned Georgia more with practices of OECD countries. Unofficially, there is anecdotal evidence that some parliament members also thought the new hours could improve public service efficiency and encourage women to participate more in the labor market.[82] The new policy affected approximately 200,000 public sector workers or 13.4% of the total workforce, a nontrivial amount; and yet, the consequences of the policy have never been studied. Importantly, this is the only policy specifically affecting public employees in the years before and after its implementation.

Work hours have a considerable influence on our personal lives and on a myriad of economic areas. A number of studies address work hours and their relationship to productivity (Golden, 2006), efficiency (Hanse, 1993), types of employment (Wasserman, 2015), wage inequality (Carr, 2011), educational outcomes (Baffoe-Bonnie & Golden, 2007), benefits of flexibility (Bird, 2015), work-life balance (Holly & Mohnen, 2012), intra-household bargaining (Rangel, 2003), gender differences in market and home labor (Goldin, 2014), gender wage gap (Blau & Kahn 2017), impact on health (Dawson et al., 2005), impact on happiness (Galay, 2007), and impact on the environment (Knight et. al., 2013). As far as we are aware, however, no study evaluates the effects of a policy that exogenously shifts the working hours of a major cross-section of workers, eliminating the common self-selection bias issues faced in many work

[81] The policy affected most offices of the government, ministries, the national bank, the national statistical office, among other public offices.

[82] Based on anecdotal evidence gathered from discussions with government officials. It is not surprising that Georgian parliament members would have such concerns in mind, since, despite the recent history of many progressive policy initiatives promoting female labor participation in Georgia, traditional gender roles remain culturally dominant for both men and women (Kachkachishvili et al., 2014). Moreover, as a signatory participant of the 1995 World Women's Conference Platform of Action (Jashi, 2005), Georgia should initiate and assess such policies.

hours studies. This advantage combined with the novelty of the policy render this examination worthwhile and informative to several topics within the working hours literature. Moreover, the dearth of directly related literature appears to be due to the uniqueness of the policy, since there are several adjacent areas of research examining how work hours and work schedules affect economic, physical, emotional, psychological, social, familial, and vocational well-being. The effects of this policy logically, and by intention (of at least some members of parliament), impact genders and family types asymmetrically, relating this paper most closely with work-family conflict, gender inequality, and intra-household bargaining and resource allocation literature. Work-family conflict literature identifies two types of conflict, family interference with work (FIW) and work interference with family (WIF). Outcomes of conflicts are informed by two models, the gender similarity model and the gender difference model. While these conflict types seem to have some conceptual overlap and are not objectively defined in a rigorous manner, since they are psychological concepts and subjective in nature, the potential effects of this policy may still be conjectured along these frameworks.

A prediction more in line with FIW and the gender difference model would be that this policy could relieve familial conflict with (potential) work for mothers. In Georgia, the vast majority of family-related household activities are conducted in the evening and by females.[83] As later working hours were considered a source of personal-professional scheduling conflict for women with household responsibilities, it is understandable why some members of parliament believed the new initiative might be more "family friendly", i.e. convenient for successfully combining economic and family activities, thus removing barriers for women with families seeking employment in the public sector. Moreover, the policy was put into effect with family schedules in mind; those with children aged 12 or younger were given a half hour of flexibility in their work schedules to relieve the resulting burden from the convergence of their new starting time and the legally-mandated-universal school starting time of 9:00 a.m. (Farulava, 2014), which generally extends to preschool as well as formal childcare.[84] In addition, public office employment in OECD countries generally offers stability, reasonable financial security,

---

[83] Winett et al. (1982) found that the introduction of flexitime programs for working parents in two US federal agencies, which allowed them to shift their work schedules by up to one hour, also led to parents spending more evening time with family. This paper, which still suffers from self-selection bias, and Orpen (1981), which uses randomization in a flexitime experiment with 64 female clerical employees to assess effects of flexitime on satisfaction and performance, represent the closest examples of anything resembling equivalence to our paper.

[84] All public primary and secondary schools in Georgia are mandated to start at 9:00 a.m. and most also offer late pickup times for working parents. Though not mandated, most private primary and secondary schools follow the same pattern. Preschools, as well as formal and informal childcare, also tend to start at 9:00 a.m. or earlier and are accessible and affordable to the population, with 84.2% of urban children and 67.7% of rural children attending preschool (National Statistics Office of Georgia, 2019). Informal childcare, especially familial, is more common in rural settings, while formal childcare outside of preschool is more common in urban settings. For both settings, formal childcare is more expensive than preschool, but tends to still make economic sense for those with average or better wages. In addition, of note in terms of familial conditions, Georgians tend to marry in their mid-to-late 20s across urban and rural settings (Hakkert, 2017), with an average difference of about 1 year later for urbanites, and tend to start families soon after marriage. Due to the immediate and transient nature of the effects of the policy we study, we do not believe that the policy would reversely impact familial conditions in any statistically significant manner.

and some flexibility, which tends to attract women (Wasserman, 2015; Goldin, 2014; Gicheva, 2013). In Georgia, the public sector holds a greater place in the economic hierarchy than in most OECD countries, with public sector employees earning well above median wages, possessing higher average levels of education, and enjoying a generally esteemed position in society. Thus, it is clear that the policy could make public sector employment more attractive for mothers and increase female labor participation.

On the other hand, a prediction more in line with WIF and the gender similarity model would be that this policy could create work-caused conflict for both mothers and fathers, perhaps more so for fathers as they tend to work longer hours on average. Families with young and school-age children are probably less flexible in their daily program than couples without children, single people, or older people with or without older children. For working parents (or those considering entering the labor market), the policy could result in conflicts with their established household itinerary that even the added flexibility and increased evening time with the family would not resolve. Such parents could find it more difficult to participate in the labor market under the new schedule.

Furthermore, though outside of the work-family conflict framework, it is also ambiguous whether those without younger children would find this time shift attractive or not. Some younger, single people might find the possibility of having more free time in the evenings to pursue social activities appealing, while others may be used to sleeping longer in the mornings and find the change objectionable. We hypothesize that the policy disparately impacts affected populations by gender, marital status, family type and size, and along other individual characteristics. Our hypothesis leads us to assume the effects of the policy will be heterogenous across characteristics and circumstances, informing the main aim of this paper: to determine the dominant effects of the policy, along which dimensions it was most impactful, and related behavioral insights. While the policy could give rise to many compelling research questions that fit our aim, we concentrate on how it may impact gender inequality through female labor participation in government jobs.

Since the policy had no effect on the private sector (where the standard working hours largely remained at 10:00 a.m.-7:00 p.m.), we are able to employ difference-in-differences (DD) methodology to compare public and private sectors, before and after policy implementation. Given the circumstances and data, DD is the optimal methodology to identify the precise effects of the policy on labor engagement as it separates out all other effects experienced by both control and treatment groups. We find that the policy does not increase female labor participation through an increase in women entering the public sector. In fact, the policy appears to have no significant effect at the extensive margin of employment in either direction. Instead, it primarily leads to a material reduction in the average level[85] of hours worked by full-

[85] The GeoStat survey data employed does not offer exact numbers of hours worked by participants, but provides intervals of weekly hours worked (1-20 hours, 21-40 hours, 41-60 hours, and more than 60 hours).

time employees with children; the outcome in line with the WIF and gender similarity model prediction. At the same time, there is also a significant increase in average work hour engagement by women without children. However, the placebo effect analysis identifies this as an already existing trend and the short-term analysis indicates that this is an ordinal response to the reduction of engagement by full-time employees with children. Altogether, we conclude that this increase is a secondary, indirect effect and that the policy did not directly cause an increase in female labor participation. Furthermore, since men with children were most negatively affected and women picked up the gains, the policy may have also indirectly increased overall gender equality.

3.2 Literature Review

According to Sayer (2005), since the 1960s, there has been an ongoing convergence in the manner in which men and women use their time. Females gradually shifted from being focused primarily on unpaid, household labor increasingly towards paid, labor-market employment. Despite the cultural shift, women remain primarily responsible for most household activities. Based on US household data, although men have been spending an increasing amount of time on household activities such as cleaning, childcare, and cooking, women remain the main contributors (Sayer, 2016). This convergence appears to lead to a reduction in leisure time for females without significantly affecting leisure time for males. Analogous data is found in the United Nations' survey "Men and Gender Relations in Georgia" (Kachkachishvili et al., 2014).

In Georgia, the government enacted initiatives that have been promoting gender equality since 1997 (Jashi, 2005) and there has been a steadily increasing female labor participation rate over the last decade (ILO, 2019). Nonetheless, 89% of the UN's gender relations survey respondents, comprising of both men and women, agree that "a woman's main responsibility is to take care of the family": 50.8% of Georgian respondents were identified as having a "negative" attitude toward gender equality, while only 3.7% had a "positive" attitude, and even the "positive" group maintained a 'patriarchal' pattern of gender-divided household duties (Kachkachishvili et al., 2014). The report concludes that any recent changes in the distribution of household tasks are "quite superficial" with only a limited amount of actual behavioral and attitudinal modification, while the underlying culturally rooted gender biases have not changed. Georgia is experiencing similar, or more severe, trends as those identified by Sayer (2016).

Gender inequalities in unpaid, household work are known to be directly related to gender inequalities in the paid labor market in terms of participation, engagement, type of employment, vulnerability, career progression, wages, retirement savings, and more (Ferrant et al., 2014). Bearing the majority of responsibility for household duties and the need to coordinate those with paid economic activities results in female "occupational downgrading", accepting worse conditions, and below-skill-level employment (Hegewisch and Gornick, 2011). Empirical studies of female labor participation are innumerous and results typically

point, in one manner or another, to the relationship between household and market labor. Vlasblom and Schippers (2004) identify “low education” and “having children” as the most important barriers to female participation in the labor market. Cortes and Pan (2017) conclude that females that anticipate difficulties in balancing career and family are more likely to exit the labor market and specialize at home than their male peers. Herr and Wolfram (2012) claim that an inflexible work environment is a major force driving women to opt out of the labor market at motherhood. Similarly, women might respond to greater occupational time demands by shifting to more family-friendly occupations or by withdrawing from the labor force (Cortes & Pan, 2017). Thus, the time demands of a given occupation seem to, on average, predominantly affect women, who already have a tendency to work less than men, causing women to switch into positions with more flexible time requirements to be able to combine professional and household activities (Wasserman, 2015; Goldin, 2014; Gicheva, 2013).

In the work-family conflict literature, Gutek et al. (1991) were the first to combine research from work-and-family sociologists and development psychologists by bifurcating the work-family conflict into two types of conflict: family interference with work (also known as family-work conflict, FWC) and work interference with family (also known as work-family conflict, WFC). Outcomes of conflicts are informed by two models: the rational view or gender similarity model versus the gender role or gender difference model. The rational view or gender similarity model is predicated on the notion that we have only so much time available to us to split between our work and family roles and as the time spent on them increases, conflict will be perceived regardless of gender. It predicts a convergence of attitudes towards conflict and the balance of work and family (Keene & Quadagno, 2004). The gender role or gender difference model is based on traditional gender roles resulting in normative differences between the genders, with asymmetric boundaries, expectations, responsibilities, and perceptions of balance. This model predicts that men and women will react to role conflict differently as more time spent in one’s gendered domain is perceived as less of a burden (Gutek et al., 1991).

An examination of why the prevailing attitudes towards gender roles endure reveals a complex, psychological web of attitudes, socialized beliefs, evolutionary differences, and individual thresholds and proclivities that commingle to result in individual, group, and societal standards. For example, in many cultures, the nature of the female gender is perceived as more fluid than that of the male gender. In the context of labor division, this means that it is more acceptable for women to adopt "masculine" behaviors, such as taking up paid work, than it is for men to adopt "feminine" behaviors, such as doing unpaid domestic work (Sayer, 2016). By not doing unpaid work, or at least minimizing their involvement in such activities, men may have (perhaps subconsciously) emphasized their masculinity and reinforced their social power (Brines, 1994; Risman, 1999). Extrapolating, it may follow that women performing a greater amount of domestic labor, even under changing socioeconomic conditions, regardless of how much time they spend in paid employment, could persist as a culturally accepted norm. This

has, so far, been reflected in what Sayer (2016) finds from five U.S. time use datasets from 1965 to 2012.

Alberts et al. (2011) put forth a compelling theory of domestic labor division that addresses and infuses several single-explanation theories into a more complex, yet rational-based framework. Their theory helps to explain why many contrary phenomena persist in household labor division, including why many full-time employed wives still do a majority of domestic work, why even men that earn less and/or work fewer hours still do not do more domestic work, and why both genders tend to view the currently unequal distribution as equitable. The theory explains that small differences in traits, informed by evolutionary biological differences and biosocial conditioning,[86] result in disparities in responses to stimuli. Divergent self-organizing systems and response thresholds[87] cause the repetition of minute behaviors that lead to "expertise" and large behavioral differences that become ingrained over time and across contexts. Moreover, few couples and dyads explicitly discuss domestic labor division and, instead, default to individual response thresholds, social norms, and habits to guide their behavior and only address issues explicitly once discord occurs. According to Alberts et al. (2011) women, on average, have lower innate thresholds for domestic disorder, certain biological characteristics, and different competencies from gendered socialization, which typically lead to higher standards of cleanliness and frequency of household task performance. This puts women at a disadvantage, on average, in the formation and long-term organization of domestic labor. Thus, this theory may substantially explain why the majority of household task responsibility and performance remains with women, despite labor market trends.

Regardless of the underlying cause, the contemporary global labor market is a diverse place, characterized by individual, occupational, local, national, and regional variation in work cultures, work-life balances, standard working hours, and gender-based differences. Moreover, it is clear from the above literature that household characteristics and circumstances affect labor market outcomes, especially impacting women, but there are often conflicting conclusions about the direction and mechanism. The policy being evaluated here, while not revolutionary by any means, imposes a foreign cultural timing onto a significant percentage of the population in an economic ecosystem that was built up, over time, in a local culture. Economic actions cause interactions and externalities. Institutional working hours may have, in part, led to the establishment of specific working hours elsewhere, such as directly related service providers,

---

[86] Women, through survival, have developed a better sense of smell as well as more attention and sensitivity to household cleanliness, combined with reinforcement from more time spent in the home due to childbearing (Alberts et al., 2011).

[87] Self-organization systems, evident throughout the living world, explain how local, individual interactions lead to group-level attributes (Camazine et al., 2001). "Convergent" self-organization is when the behaviors of individuals become more alike. "Divergent" self-organization is when the behavior of an individual causes the same behavior to be less likely in others and the act of performing the behavior also reduces stimulus-level-causing responses. Response thresholds are "the perceived stimuli that must exist for an individual to decide to perform a task (Theraulaz, Bonabeau, & Deneubourg, 1998). Like Hrdy's (1999) responsive mothers, individuals with low response thresholds for a specific task are moved to perform the task earlier than individuals who have a higher threshold for the task (Breshers & Fewell, 2001; Robinson & Page, 1989)" [Alberts et al., 2011, page 7].

associated private sectors, schools and childcare facilities, and restaurants, afterwork, and nightlife venues that follow employee schedules. It is reasonable to expect that even a one-hour shift in work hours disrupts a steady-state element of the Georgian society and could cause behavioral adjustments at the individual and/or organizational level. As the data indicates that there were no significant effects at the organizational level, the evaluation of this unique policy may shed light on how individuals and households react to such seemingly minor changes and provide insights into how situational and familial composition affect labor participation behaviors and gender equality, as well as illustrate nuances related to domestic division of labor and intra-familial/intra-household bargaining.

To assess how the policy impacts labor participation in the affected sector, we turn to the difference-in-differences method using the affected public sector as the treatment group and the unaffected private sector as control. We begin by confirming that the private sector is, in fact, unaffected and that the consequent adjustments are at the individual level. Next, we assess the differences between the employee behaviors in the two sectors following policy implementation. While this paper may be the first to evaluate such a working hour shift policy, it is not the first to use DD methodology to assess outcomes between affected and unaffected sectors. Some recent examples include the specific use of public and private sector employees to evaluate the impact of a Taiwanese pension policy shift to identify the effect employer-sponsored pensions have on household saving (Yang, 2020), the use of sector-specific import tariff increases to estimate their impact on U.S. export growth (Handley et al., 2020), and the use of differences in implementation of anti-smoking regulations amongst sectors and countries across Europe to determine the economic effect on restaurants, bars, and cafes (Pieroni & Salmasi, 2017). In a recent paper closely related to our topic, Angelov et al., (2016) employ DD methodology to assess the long-term effects of entering parenthood on the gender wage gap and female labor participation, though they did not use sectors as an instrument to evaluate a policy. Generally, DD methodology is common in labor economics research and we believe it is appropriate and optimal for the purposes of our analysis.

3.3 Data

3.3.1 Primary Dataset

The primary data used in this study is publicly available on the web site of the National Statistics Office of Georgia (GeoStat). In particular, we utilize individual level data from the Households Incomes and Expenditures Survey for the four calendar years 2013, 2014, 2015, and 2016. Every quarter, GeoStat surveys approximately 3,500 Georgian households and aims to have each randomly selected household participate in the survey four consecutive times. The outcome is a piecemeal panel dataset composed of repeated individual observations for up to a one-year history of a household's socio-economic, gender, and geographical characteristics.

As true for any survey dataset, the household budget survey data is expected to contain some measurement error. Each respondent reports detailed information regarding household or private socio-economic and geographical information for the past quarter, which can lead to recall and other inaccuracies while reporting numbers. According to GeoStat documentation as well as direct discussions with data collectors, the collection process uses a best-practice methodological approach supervised by the statistical department and the collected data is a population-representative sample with a small margin of error. All things considered, there seems to be no evidence that the measurement error would not be random.

Table 1 shows the variables we used in our analyses.

Table 1: GeoStat household survey variables and their descriptions

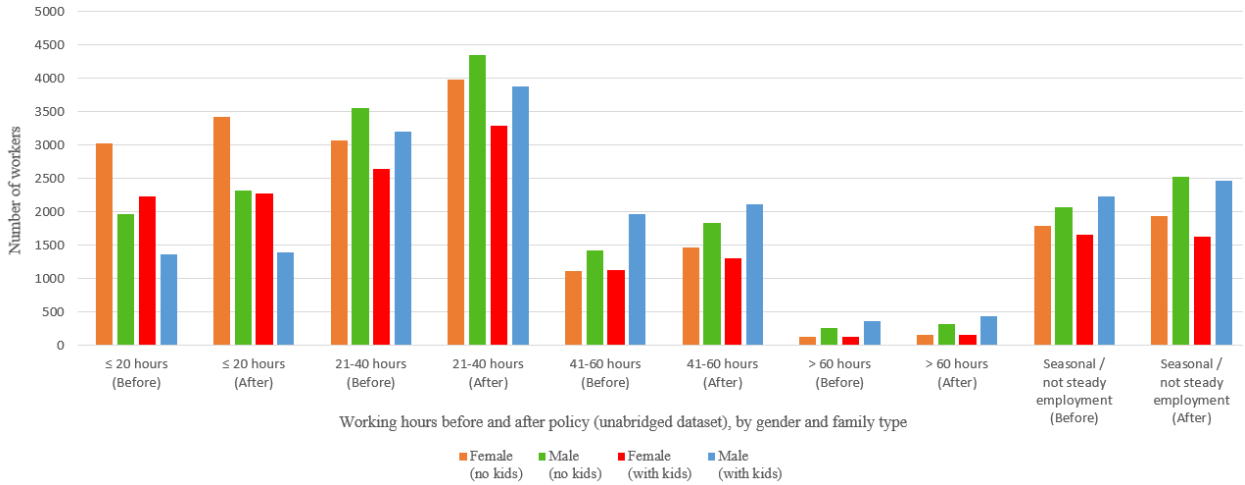
Variables	**Description**
Weekly working hours (intervals)	The number of working hours during the week. Categorical variable: "20 hours and less; 21-40 hours; 41-60 hours; Depends on a period (season); More than 60 hours."
Activity	Economically active according to the ILO strict criteria. Binary variable: "Yes; No"
Urban or Rural	Rural\Urban Classification. Binary variable: "Rural; Urban"
Owner of home	Owner of the dwelling (ownership type). Categorical variable: "Belongs to the household; Mortgaged; Rented; Used without payment"
Assistance	Whether the household received assistance or any kind of advantage or not. Binary variable: "Yes; No"
Age	Age of an individual
Family size	Number of household members
Education	Categorical variable: "Illiterate; Does not have primary education; Lower secondary education; Primary education; Secondary professional program; Higher professional program; Upper secondary education; Vocational program; Bachelor; Master; Doctor."
Small kids	Number of children (0-7 years old)
Big kids	Number of adolescents (8-15 years old)
Working man	Number of working age men (16-64 years old)
Working woman	Number of working age women (16-59 years old)
Duration in the living place	Duration of living at this address
Dwelling selling price	The amount in local currency that the household would pay to buy a dwelling similar to theirs.

Change in financial condition	Financial condition of the household has changed during the past 12 months (subjective evaluation). Categorical variable: "improved very much; not changed; slightly improved; slightly worsened; worsened very much"
Attending any professional courses	Whether the household member attended any courses for learning new professions/skills during the past three months. Binary variable: “Yes; No”
Never worked before	Whether a household member has never worked. Binary variable: “Yes; No”
Economic condition based on income	Economic condition of the household based on household income (subjective evaluation). Categorical variable: “Very bad; Bad; Satisfactory; Middle; Good”
Profession	Profession or specialty defined by a diploma, certificate or other document or gained another way. Categorical variable, listed more than 200 professions at the level of 4-digit code adopted to International classification of ISCO-88
Marital status	Categorical variable: "Single; Married; Non-registered marriage; Divorced; Widowed"
Migration	If a household member moved. Categorical variable: "From other country; From another region in Georgia; From the same region in Georgia"
Economic condition based on property	Economic condition of the household based on household property (subjective evaluation). Categorical variable: "Extremely poor; Poor; Middle; Rich; Well-off"
Reason for not applying for assistance	Reason the household has not applied to the Social Service Agency. Categorical variable: "I don't hope to get the assistance; Our family doesn't require social assistance; It’s difficult to answer; I don't know where to apply; I can’t do it myself and there is nobody to whom I can address for help; I consider it being humiliating for our family"
Special status	Special status of the household member. Categorical variable: "Chronic patient; Disabled (I group); Disabled (II group); Disabled (III group); IDPs"
Area of dwelling	Total area of the dwelling (in square meters)
Pensioner man	Number of pension age men (65 years and older)
Pensioner women	Number of pension age women (60 years and older)
Mobile phone	Quantity of the owned durable good
Additional activity	Secondary employment

Notes: Variable names adjusted for ease of comprehension. For example, “Weekly working hours (intervals)” is actually “TimeDuration”.

The following figures present an examination of our dataset, beginning with a breakdown of weekly working hours before and after policy implementation, delineated by gender and sector.

Figure 1: Weekly working hours (intervals), by gender, sector, and implementation



Within our dataset, 13.7% of all working people are employed in the public sector and 86.1% work in the private sector. Segregating by gender, 56.2% of public sector employees are female and 43.8% are male. Unlike the government sector, the number of men exceeds the number of women working in the private sector. Men account for 53.6% and women 46.4% of workers in the private sector. On average, over the entire period of the dataset, 19.54% of the public sector employees worked 20 hours and less per week, 48.79% worked 21-40 hours per week, and 25.95% worked 41-60 hours per week. Only 1.96% were employed in a seasonal/not steady public sector position and 3.76% worked more than 60 hours per week. Partitioning this information further by whether employees had children gives us the next three figures (for all workers, private sector workers, and public sector workers). In Appendix Tables A60-A67, we provide this and additionally delineated descriptive statistics in table form as well as partition the public sector observation numbers by monthly mean, maximum, minimum, and standard deviation. Given that it is a representative sample, the tables show a reasonably balanced division amongst the subsample groups.

Regarding the balance between the sectors, Appendix Table A61 shows that the distribution amongst the working hour intervals between the two sectors is fairly similar but diverges most amongst the full-time and seasonal employment figures. In terms of structural differences between the sectors, it is important to recall that in Georgia, public sector employees earn above median wages, considerably more than their private sector counterparts, on average; have much higher average levels of education, most have at least a master's degree; and carry a high level of respect in society. The public sector distribution has less variance than the private sector, because it simply does not offer many seasonal employment opportunities and mostly does not hire people without a higher education, which immediately removes teen and early adult employees from the variance, who tend to work the fewest hours. When we remove the

seasonal workers from the numbers, the distributions become much more similar across the intervals. Prior to the policy, the distribution is almost identical, slightly skewed to greater work hour engagement in the public sector, though this does not reflect top-down organizational differences between the sectors, but rather the natural, bottom-up difference in sector breadth and variance noted above. Moreover, while there is essentially no change in the distribution amongst the intervals in the private sector before and after the policy, we see a 7.5% increase in the 21–40-hour interval in the public sector post policy, accompanied by a direct decrease in the intervals just below and above, but especially below. This foreshadows the findings of this paper.

Figure 2: Weekly working hours, by gender, parental status, and implementation

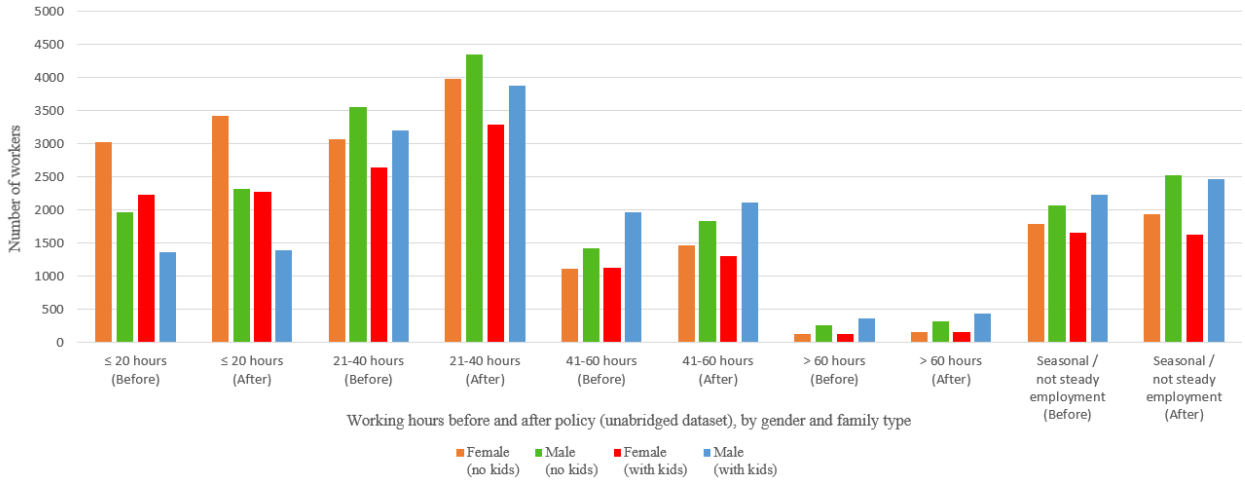


Figure 3: Private sector weekly working hours, by gender, parental status, and implementation

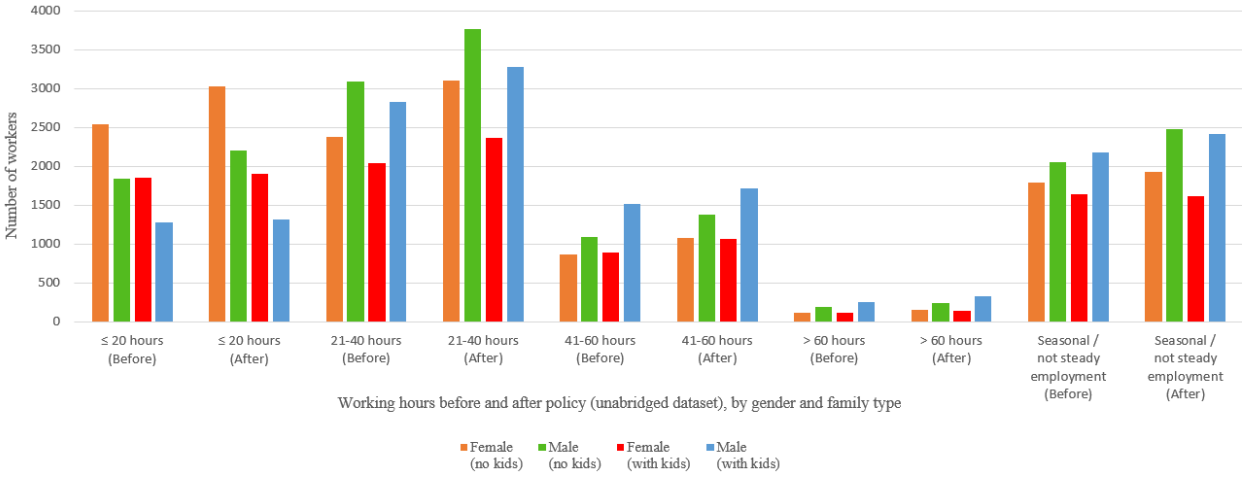
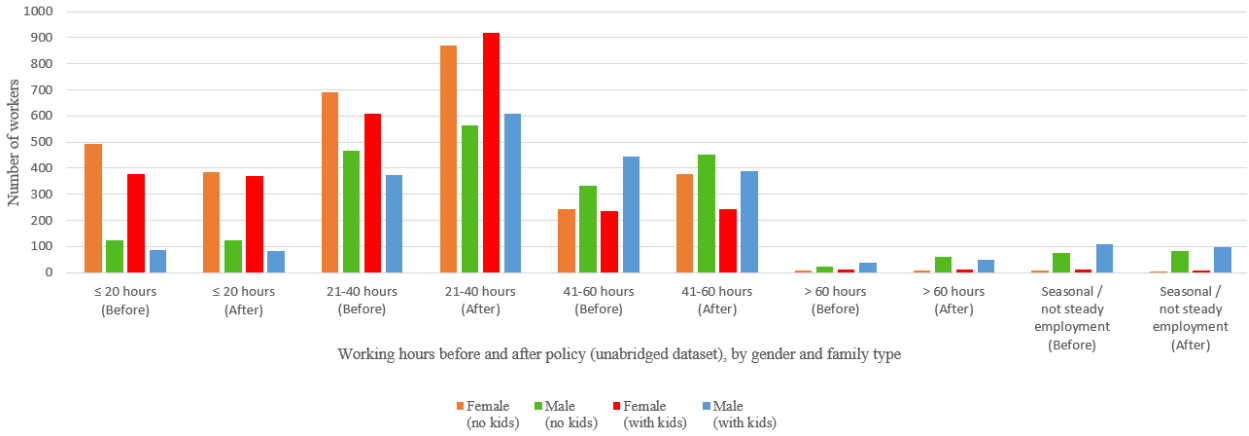


Figure 4: Public sector weekly working hours, by gender, parental status, and implementation



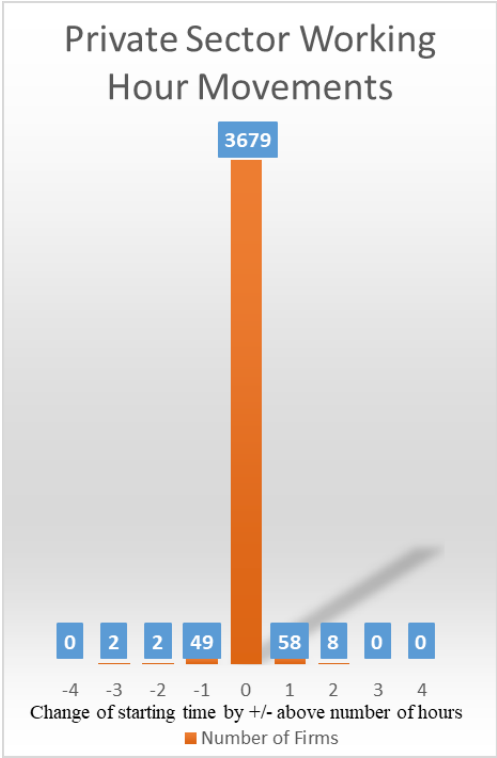
Figures 2, 3, and 4 visually communicate the total number of weekly working hours (intervals) observations in our dataset before and after the policy implementation broken down by gender, parental status, and sector. In total, a slim majority of employees working 40 hours or less are females, while employees working overtime hours are mostly male. On the face of the data, it seems that there is a significant increase in the number of female (and male) employees with children working 21-40 hours in the public sector, which some might claim as evidence of "family friendliness" and increased female labor participation. However, increases are also present for their male, no children, and private sector counterparts, hence the need for the methodology described in the next section to conduct a proper evaluation of the policy. For example, several such regressions without covariates return positive gains, especially for women with small children. However, after including covariates that control for alternative sources of this increased employment, the policy's effect is weakened and becomes statistically insignificant. One notably important revelation is the very small number of observations in our dataset of females working 60 hours or more in the public sector. Such a small sample size is insufficient for reliable inference regarding female labor participation around the 60-hour threshold. While not as impactful to inference, another questionable sample size revealed by the descriptive statistics is the relatively small sample size of men with older children working more than 60 hours.

3.3.2 Supplementary Dataset

A supplemental, firm-level database is used to check whether the implementation of the government's new policy led the private sector to adjust working hours for their employees. The Business Information Agency (BIA) is a leading data collector of company profiles operating in Georgia. Their database consists of statistical information for more than 45,000 active companies. Each firm's general information (e.g. trademarks, products, registration date, VAT number, business activity, legal address, website, and working hours) is publicly

available on BIA's webpage. We extracted and analyzed the data for a subsample of firms that had observations recorded before and after the policy implementation between 2013 and 2016. We found 3802 firms with observations both before and after the policy implementation between 2013 and 2016. Only 3.2% of those firms changed their business hours after the policy had been applied by the government. Moreover, as evidenced by Figure 5, the changes were normally distributed around the mean and mode of zero change. Additionally, we analyzed the shift of working hours for the placebo threshold of one year before as well as one year after the policy to check that the trend holds for the other periods. The results show that only 4.2% and 3.2% of firms shifted their business-operating hours, respectively, and in a similarly distributed manner. The following figure visually demonstrates the lack of direct effect on private sector working hours from the policy.

Figure 5: Distribution of private business starting time movements post policy



Notes: The bar chart shows the relative amounts visually, and the actual numbers above the bars, of private sector businesses that changed their starting times after the policy was implemented, and by how many hours (from -4 to +4).

3.4 Methodology

Having confirmed that the working hours of the private sector were in no way systematically affected by the policy change that directly altered them in the public sector, we now detail how the difference-in-differences method is utilized to determine how the new government policy affected participation in the labor market. According to Angrist and Pischke (2008), the method estimates the effect of the treatment (i.e., an explanatory variable or an independent

variable) on the outcome (i.e., the response variable or the dependent variable) by comparing the average change over time in the outcome variable for the treatment group, compared with the average change over time for the control group. We designate the private sector as the control group and the public sector as the treatment group. In formal terms, *s* denotes sector (either public or private) and *t* denotes time period. As the policy was implemented on September 1, 2014, with essentially no notice, we believe that any direct effect of the policy change on labor market participation would not occur before 2 months at the earliest due to established employment notice periods for leaving a position, the time it takes to process and hire a new employee, and the time it takes for managers and employees to assess the policy's actual effects and permanently adjust work hour schedules internally. This assumption, further discussed in section 3.5.1, is by and large confirmed by the findings of the short-term-effect and September-threshold analyses, which are described at the end of this section. Accordingly, the main analysis time threshold was set as November 1, 2014. In formal terms, this outcome variable takes the form:

Y_{ist} = 1, if an individual is working a specified range of hours per week

Y_{ist} = 0, if an individual is working an alternate range of hours per week

In particular, Y_{ist} equals zero (below) or one (above) across the specific binary extensive margin threshold of 0 hours and more than zero hours (including seasonal / not steady employment), and the following intensive margins (which do not include those working 0 hours nor seasonal / not steady employment): above and below 20 hours, above and below 40 hours, above and below 60 hours, and pairwise[88] amongst the individual weekly working hour values. DD regression equations take two conventional forms (ending up with the same result). We opt for the interaction term form:

$$Y_{ist} = \alpha + \beta_1 Treatment_{is} + \beta_2 Time_{it} + \beta_3 Treatment_{is} * Time_{it} + \beta_4 X_{iat} + \varepsilon_{ist} \qquad (1)$$

Where $Treatment_{is}$ is a dummy variable that equals one if the observed individual is in the public sector, $Time_{it}$ is a dummy variable that equals one if the time of the observation occurred in November 2014 or later, α is a constant, and X_{ist} is a set of covariates that includes an individual's characteristics and answers to other survey questions that are correlated with the outcome variable. The resulting coefficient, β_3, expresses the post-policy correlation difference between the control and treatment groups, making it the only consequential and

[88] The pairwise thresholds represent a supplemental analysis that aims to approximate how local the policy-induced working hour movements are (just a few hours across the nearest threshold or larger jumps) as well as provide an enhanced picture of the movements just around the thresholds.

relevant coefficient to this research and the only coefficient reported in the output tables. As the weekly working hours replies are in 20-hour intervals, the DD regressions with the constructed thresholds are specifically capturing the changes between the average number of workers below/above a given threshold.[89] We attempt to further distinguish the specific correlation of the policy on Y_{ist} by executing the regression using three sets of covariates[90] that increase the precision of the coefficient and the explanatory power of the regression. Furthermore, we also aim to increase the precision by more accurately defining the treatment threshold.[91] All tables in section 3.5 and the Appendix display only the coefficients with the full covariate schedules, broken down by increasing particularization of the treatment group.

To support causality inferences of β_3 covariates, we provide parallel trends analyses to assess whether the two groups had similar trends over time prior to the policy implementation, which then diverged due to the effect of the policy on the treatment group. In addition, we check causality by conducting placebo effect analyses, counterfactually changing the time threshold to twelve months prior to the actual change. A resulting lack of a statistically significant β_3 bolsters the notion that effects found from the difference-in-differences regressions were specific to the policy change and not just random noise. Consequently, we consider a strongly statistically significant β_3 coefficient that holds in the most stringent control configuration, is part of a parallel trend that diverges post-policy, and does not produce a placebo effect, to be a credible substantiation of a causal effect of the policy on those treated.

Since we employ a two-month lag from the actual initiation of the policy, we further supplement the main analyses with DD regressions of the main thresholds using the September 1, 2014 threshold. Furthermore, we run short-term analyses of the effects for three months, six

---

[89] As there are, essentially, no changes along the extensive margin (see section 3.5.1.1), the constructed threshold regressions are not (or are minimally) capturing new or leaving employees on either side of a given threshold. Moreover, any changes in the average number of employees in a given interval are not captured as part of the DD regression unless they are across a given threshold. That is, a change in average number of employees between the 1-20-hour interval and the 21-40-hour interval is not captured as a difference at the 40-hour threshold.

[90] First, we run the regressions without controls. Next, we add several substantial covariate controls for individual, household, and professional attributes, including age, education, family size, number of working age people in the household, number of children, living in an urban or rural area, length of time living there, owning their own home, several objective and subjective measures of income and wealth, if they are economically active, and if they have ever been unemployed. Finally, we add all remaining covariate controls that had any statistically significant correlation from the DD regression, including marital status, migration history, profession category, additional wealth measures, number of retired family members in the household, and disability status. While only the full covariate results are presented in the body and online appendix of this paper, a full appendix with all results is available by request. Across the regressions of the main thresholds, the covariates that were consistently most correlated with Y_{ist}, which is evident through their statistically significant coefficients (available in Appendix Tables A56 – A59) were urban location, years in this city, wealth and ownership measures, and age.

[91] The baseline is all public employees as treatment and private as control. However, as noted in the introduction, not all public employees were affected by the policy. Therefore, in the second specification, we move the employees from the entirely unaffected public fields, such as public education employees (teachers, school administrators, etc.), to the control group. In the third specification, we move expectedly unaffected public employees to the control group as well. That is, while the expected majority of public employees in specific professions should not be affected, such as dentists, some may happen to be affected by the policy due to certain idiosyncratic peculiarities (such as office location) or the ambiguous nature of certain professions. Hence, they are included only in the final specification.

months, and twelve months from both the November and September thresholds. These results reveal an ordinal nature of the effects of the main analysis, with some of the effects beginning around three to six months after policy initiation only to have the strength of those effects depleted by the end of 2016, while others begin later and grow stronger and more statistically significant through 2016. Lastly, though we control for age and type of location throughout the analyses, we also run partitioned analyses by separating the sample by urban versus rural locations and dividing it in half by median female age (49 years old) and median male age (45 years old) as well as their interaction to assess whether the policy had age-specific and/or location-specific implications.[92] Full result tables are presented in Appendix Tables A1 – A53.

3.5 Results

In this section, we present and discuss tables that highlight the most significant and relevant regression findings from all three treatment specifications and binary thresholds listed in section 3.4. Only select subsample groups from the main analyses that give an overview or provide statistically significant results, or their counterparts, are featured herein. Complete full covariate control output tables that exhibit all results for every subsample group and supplemental analysis are presented in the Appendix.

3.5.1 Main Results

3.5.1.1 Extensive Margin

The first output table we present is the 0 hours and more than zero hours threshold (the extensive margin between working and not working; including seasonal/not steady employment). As can be seen in Table 2, all of the resulting β_3 coefficients are weak and statistically insignificant. A detailed analysis of the extensive movements confirms an overall lack of changes at this margin. The analysis uncovered that there were only 303 [311] extensive margin moves out of 5964 [5667] total panel observations and only 102 [102] extensive margin moves involving the public sector around the threshold [lagged] of the policy implementation. This resulted in only a 20 [20] net employee gain in the public sector (an insignificant difference). Appendix figures A7 and A8 visualize the extensive margin data points for both the official implementation timing and the lagged threshold used in the DD regressions. These movements are meager (see Figure A7) and not statistically different from the extensive margin moves from the placebo thresholds of one year prior (see Figure A8). Furthermore, the September, short-term, age, and location analyses all return weak and insignificant results.[93]

---

[92] An analysis by education was also explored, but since nearly all public employees have higher education degrees and the vast majority have a master's degree or higher, a DD with the private sector population would be biased.

[93] The younger [and urban] groups had a few β_3 coefficients at the 10% [and 5%] level for males without kids, especially when not married. However, all significance disappeared in the combined young & urban analysis,

Therefore, we conclude that the policy did not have a statistically significant effect on the extensive margin of employment.

Table 2: DD regression results for weekly working hours (intervals), extensive margin

Subsample	Gender	(1)	(2)	(3)	N	R^2
All	All	-0.00233	-0.0006	0.000152	85523	0.34
		(0.005)	(0.006)	(0.006)		
	Male	-0.00764	-0.00817	-0.00389	45627	0.32
		(0.009)	(0.009)	(0.01)		
	Female	0.0016	0.006	0.00398	39896	0.38
		(0.006)	(0.007)	(0.008)		
With kids	All	0.00169	0.00352	0.00656	40124	0.31
		(0.008)	(0.009)	(0.009)		
	Male	-0.00514	-0.00712	-0.00172	21979	0.26
		(0.013)	(0.013)	(0.014)		
	Female	0.00675	0.0146	0.02	18145	0.41
		(0.009)	(0.011)	(0.012)		
Without kids	All	-0.00864	-0.00687	-0.00561	45399	0.36
		(0.007)	(0.008)	(0.009)		
	Male	-0.0163	-0.0143	-0.00898	23648	0.37
		(0.012)	(0.012)	(0.013)		
	Female	-0.0012	0.00186	-0.00168	21751	0.36
		(0.008)	(0.009)	(0.011)		

Notes: 10%, 5%, 1%, and 0.1% levels of confidence are indicated by (+), (*), (**), and (***), respectively. Standard errors are in parentheses. Column labels: (1) is the pure sector division of public and private as treatment and control, respectively; (2) adds employees from entirely unaffected public fields, such as public education, into the control group; (3) adds employees with professions where the expected majority would not be affected, into the control group.

3.5.1.2 Intensive Margin

Further analysis revealed statistically significant movement within the intensive margin of labor participation through the weekly working hours (intervals) variable. We assessed those effects by creating specific binary thresholds using the survey's interval responses to the question of how many hours each employed individual works to construct the thresholds of above and below 20 hours, above and below 40 hours, and above and below 60 hours. As there is a lack of women working more than 60 hours and since the 20- and 40-hour thresholds represent standard part-time and full-time working hours with most employment bunched there, those two thresholds are most pertinent herein. We begin with the 20 hours or less versus 21+ hours threshold in Table 3.

---

which may expose the previous results as spurious or be caused by the reduced sample size, though the latter argument is uncertain with the associated number of observations. Combined with the findings of the detailed analysis, these anomalous results do not change our overall conclusion, but they may indicate that some young, unmarried, urbanite males found the new hours unattractive, as hypothesized in the introduction.

Table 3: DD regression results for weekly working hours (intervals), 20-hour threshold

Subsample	Gender	(1)	(2)	(3)	N	R^2
All	All	0.0396***	0.0187+	0.0190+	60234	0.15
		(0.009)	(0.01)	(0.011)		
	Male	0.011	0.00611	0.00822	30740	0.13
		(0.013)	(0.013)	(0.014)		
	Female	0.0597***	0.0292*	0.0323+	29494	0.14
		(0.013)	(0.015)	(0.017)		
With kids	Female	0.0335+	0.0102	0.00403	13154	0.14
		(0.019)	(0.022)	(0.025)		
Without kids	Female	0.0854***	0.0414*	0.0575*	16340	0.16
		(0.018)	(0.02)	(0.023)		
Without kids (family size>1)	Female	0.0830***	0.0381+	0.0631**	14363	0.14
		(0.019)	(0.021)	(0.024)		
Without kids (family size>1, married)	Female	0.0933***	0.0518+	0.0821**	9472	0.12
		(0.024)	(0.027)	(0.031)		

Notes: 10%, 5%, 1%, and 0.1% levels of confidence are indicated by (+), (*), (**), and (***), respectively. Standard errors are in parentheses. Column labels: (1) is the pure sector division of public and private as treatment and control, respectively; (2) adds employees from entirely unaffected public fields, such as public education, into the control group; (3) adds employees with professions where the expected majority would not be affected, into the control group.

Regression results for the entire subsample indicate that there is a positive correlation between the DD identified policy effect and weekly working hour interval, but when dividing the subsample by gender, it is evident that correlation is heavily driven by the change in average female working hour engagement. When further dividing that sample into those with and without children, it becomes clear that those who have most increased their engagement are women without children. When again dividing women without children into married and not married subsample groups, it becomes clear that the increase is driven by married women without children. This is an unexpected result, especially for those who believed that changing public office working hours would break down barriers for women with children. Moreover (visible in the unabridged table in Appendix Table A2), at the 20 hour-working-week threshold, there is no significant difference if the children are small (0-7 years old) or big (8-15 years old). According to the output of the short-term analyses, these differences begin to become evident and significant about 8 months post policy initiation and strengthen through the end of 2016.

The supplemental age analysis reveals that these effects were more consistently occurring for older employees, even for men at a much weaker level, across the treatment specifications, including strong, positive effects by singles who were the sole member of their household, though the sample size may already have become an issue there. By location, the effects are stronger and more consistent for rural female employees, though the subpopulations that were most strongly affected were older, rural women without children, followed by younger, urban women without children. We also witness moderate, negative effects on younger, urban

women with children and older, rural women with older children. The refined subpopulation analyses also uncovered some otherwise elusive effects that were averaged out in the larger sample groups: moderate, positive effects of the policy on younger, rural women with children and older, urban women with older children[94] as well as fairly sizable, positive effects for urban males with children, especially older children, and for older, urban males without children.

Table 4 displays the results for the 40 hours or less versus 41+ hours threshold. From the full subsample results, it is evident that the effect is strong at this threshold. Dividing it by gender reveals that both men and women are affected at this threshold, but especially men. This gender difference decreases as the subsample is further reduced to include only those with children. Those with younger children seem most likely to reduce their work engagement across this threshold in general, though men with older children seem more affected than their female colleagues. The lack of an effect on women from the full sample population seems to be due to the countering effect from women without children increasing working hour engagement at this threshold, especially those who are part of a household of two or more people and married.

Further refinements are revealed by the age and location analyses. While women with children were similarly affected across the age groups, urban women with children were much more impacted than their rural counterparts. Urban and rural men with children were similarly negatively impacted, though slightly more so in rural locations. Across locations, older men with children were much more impacted than their younger counterparts. Older men and women with younger children were the most negatively affected at this margin, while of those with older children, only older men were affected and not as strongly. For the younger group, the opposite is true, with the greatest negative effects experienced by women with older kids as well as men with younger kids. Though the positive effects for women without children were universal amongst the partitioned subsamples, the vast majority were experienced by younger women, especially urbanites. Regarding men, only older, urban males without children exhibited positive effects from the policy at this margin.

The short-term and September analyses (from 6 to 14 months post policy implementation) consistently display slightly stronger and more statistically significant results for those with kids at the 40-hour threshold than many of the full-data, November-threshold results above. This indicates that the effects on working hours at this margin are primary and early ordinal results of the policy. It seems these effects at this margin were, on average, greatest and most significant about 12-14 months after the policy went into effect and then began to decline over time. Given that the policy impacted individual (and by interaction, household) schedules by 30-60 minutes, it seems logical that they would have a transitory nature and be more intense in

[94] The placebo effect output related to the older, urban women with older children returns at a rather significant level, though the placebo trend was in the opposite direction. While this disqualifies this finding as fully credible, it also does not indicate an already occurring trend. Since it is a minor finding, we decided to keep it herein.

the short-term and then dissipate as a new steady state is achieved. For women without kids, the effects begin to become significant 12 months post policy and then strengthen.

Table 4: DD regression results for weekly working hours, 40-hour threshold

Subsample	Gender	(1)	(2)	(3)	N	R^2
All	All	-0.0289***	-0.0365***	-0.0488***	60234	0.16
		(0.009)	(0.009)	(0.01)		
	Male	-0.0541***	-0.0543***	-0.0596***	30740	0.17
		(0.014)	(0.014)	(0.015)		
	Female	-0.00835	-0.0168	-0.0353**	29494	0.17
		(0.011)	(0.012)	(0.013)		
With kids	All	-0.0898***	-0.105***	-0.120***	27868	0.16
		(0.013)	(0.014)	(0.015)		
	Male	-0.109***	-0.106***	-0.116***	14714	0.17
		(0.02)	(0.021)	(0.022)		
	Female	-0.0682***	-0.0879***	-0.110***	13154	0.15
		(0.016)	(0.019)	(0.021)		
With small kids	All	-0.118***	-0.127***	-0.116***	12630	0.17
		(0.02)	(0.021)	(0.023)		
	Male	-0.140***	-0.122***	-0.107***	6994	0.17
		(0.03)	(0.03)	(0.032)		
	Female	-0.0829**	-0.108***	-0.0898**	5636	0.18
		(0.026)		(0.033)		
With big kids	All	-0.0683**	-0.0861***	-0.139***	9727	0.18
		(0.021)	(0.024)	(0.026)		
	Male	-0.106**	-0.125***	-0.149***	4787	0.2
		(0.037)	(0.038)	(0.041)		
	Female	-0.0302	-0.0309	-0.115***	4940	0.17
		(0.026)	(0.03)	(0.035)		
Without kids	Female	0.0452**	0.0432**	0.0272	16340	0.2
		(0.014)	(0.015)	(0.017)		
Without kids (family size>1)	Female	0.0364**	0.0344**	0.0208	29354	0.17
		(0.012)	(0.013)	(0.014)		
Without kids (family size>1, married)	Female	0.0610***	0.0535**	0.0521*	9472	0.16
		(0.018)	(0.02)	(0.022)		

Notes: 10%, 5%, 1%, and 0.1% levels of confidence are indicated by (+), (*), (**), and (***), respectively. Standard errors are in parentheses. Column labels: (1) is the pure sector division of public and private as treatment and control, respectively; (2) adds employees from entirely unaffected public fields, such as public education, into the control group; (3) adds employees with professions where the expected majority would not be affected, into the control group.

Table 5: DD regression results for weekly working hours, 60-hour threshold

Subsample	Gender	(1)	(2)	(3)	N	R^2
All	All	-0.0117**	-0.0108**	-0.00870+	60234	0.03
		(0.004)	(0.004)	(0.005)		
	Male	-0.0181**	-0.0131+	-0.0102	30740	0.03
		(0.007)	(0.007)	(0.007)		
	Female	-0.00588	-0.0068	-0.00534	29494	0.03
		(0.004)	(0.005)	(0.005)		
With kids	All	-0.0189**	-0.0196**	-0.0135+	27868	0.03
		(0.006)	(0.007)	(0.007)		
	Male	-0.0271**	-0.0207+	-0.0104	14714	0.04
		(0.01)	(0.011)	(0.011)		
With small kids	Male	-0.0417**	-0.0394*	-0.0398*	6994	0.05
		(0.015)	(0.016)	(0.017)		

Notes: 10%, 5%, 1%, and 0.1% levels of confidence are indicated by (+), (*), (**), and (***), respectively. Standard errors are in parentheses. Column labels: (1) is the pure sector division of public and private as treatment and control, respectively; (2) adds employees from entirely unaffected public fields, such as public education, into the control group; (3) adds employees with professions where the expected majority would not be affected, into the control group.

At the 60 hours or less versus more than 60 hours threshold, presented in Table 5, the results are somewhat similar to those of the 40-hour threshold, except they are weaker and less statistically significant. Due to the lack of female representation of public employees working more than 60 hours, it is not remarkable that both women with children and women without children have completely insignificant and low magnitude results at this threshold.[95] Men, however, are seemingly quite affected at this margin. In particular, younger men in rural areas with children have the largest reduction in engagement.[96] All effects at this margin with large enough sample sizes for realistic inference are negative and mostly driven by younger workers.

From the September and short-term analyses, we learn that the effects of the policy are even more immediate than at the 40-hour margin, being felt within three months of the commencement of the policy and, therefore, already partially captured in the period prior to the two-month lag of the November threshold, resulting in them being slightly diluted in the main DD comparison. Considering the schedules of those who worked more than 60 hours per week, it is logical that they would be so immediately impacted by this exogenous schedule change. As with the 40-hour margin, the effect seems to peak somewhere around 12 months after the policy went into effect and then dispersed into 2016. The most notable difference between Table 5 and the related September and short-term analyses is that men without

[95] Despite the problematic sample size, there is a modicum of evidence that a negative impact on urban women with children exists at this margin, though this is likely a remnant of the placebo effect discussed in section 3.5.2.1. A complete lack of evidence characterizes their rural counterparts. Moreover, the inference related to men with children at this margin is disputable and assumably does not hold due to a rejected parallel trends assumption.
[96] The placebo checks would counter this finding, but those results have too small sample sizes to be reliable.

children also display a significant, negative effect (amounting to about 2-3%) from the policy at this margin.

3.5.2 Robustness Checks

As noted in the methodology section, the legitimacy of difference-in-differences regression results rests on certain underlying assumptions, which can be substantiated through parallel trends and placebo effect analyses.

3.5.2.1 Placebo Effect

Placebo effect analysis helps to confirm that the identified effect is actually directly related to the effect of the policy and not some other cause. This is generally conducted by changing one of the difference points in the DD regression to something that should not be causing an effect similar to the policy. When the resulting β_3 coefficient is statistically insignificant, that supports the contention that statistically significant β_3 coefficients from the actual DD analyses are caused by the policy and not some other phenomenon. In our case, we elected to use the fairly standard placebo threshold of one year prior to the threshold used for the main analysis. The complete results of the main placebo effect analyses are in Appendix Tables A40-A42.

Table 6 displays the output of the placebo effect analysis for the complete sample population at each main threshold as well as all statistically significant findings from the main threshold analyses. Most of the β_3 coefficients from all the placebo analyses are weak and statistically insignificant, confirming that the vast majority of the main analysis results are not caused by some other effect. There are sporadic β_3 coefficients below and in the supplemental placebo analyses that come out as statistically significant, but do not counter the findings and conclusions from the main analysis. These coefficients are from the subsamples of women with young children at the 40-hour threshold, unmarried people at the 60-hour margin, men with older children at the 60-hour threshold, urban women with older children at the 20-hour margin, and younger women with younger kids at the 40-hour threshold.

One major exception to this is the strong and statistically significant positive β_3 coefficients from women without children (including those from families composed of two or more people, both married and unmarried) at the 40-hour threshold. This indicates that women without children were gaining more working hours in public sector jobs than their counterparts in the private sector prior to the policy implementation, ruling out the policy as the explicit cause. Instead, the policy may have aided in the continuation of this trend by providing additional hours for women without children in the public sector to acquire. This interpretation is echoed in the ordinal findings of the short-term analyses.

Table 6: Placebo analysis results for weekly working hours (intervals), multiple thresholds

Subsample	Gender	(1)	(2)	(3)	N	R^2
	All	0.00222	0.00929	0.00607	25051	0.16
		(0.015)	(0.016)	(0.017)		
All	Male	-0.00419	0.00944	0.0103	12834	0.15
(20-hour threshold)		(0.02)	(0.021)	(0.022)		
	Female	0.000653	0.00637	0.00512	12217	0.15
		(0.021)	(0.024)	(0.026)		
	All	0.0154	0.00634	0.015	25051	0.18
		(0.013)	(0.014)	(0.016)		
All	Male	-0.0078	-0.0101	0.0128	12834	0.2
(40-hour threshold)		(0.021)	(0.021)	(0.023)		
	Female	0.0279+	0.0182	0.0197	12217	0.17
		(0.016)	(0.018)	(0.021)		
With small kids	Female	0.0474	0.0479	0.111*	2316	0.2
(40-hour threshold)		(0.041)	(0.046)	(0.052)		
Without kids	Female	0.0567**	0.0668**	0.0483+	6652	0.2
(40-hour threshold)		(0.021)	(0.023)	(0.026)		
Without kids	Female	0.0588**	0.0731**	0.0564*	5858	0.19
(family size>1)		(0.022)	(0.025)	(0.028)		
(40-hour threshold)						
	All	2.12E-05	-0.00143	-0.00378	25051	0.03
		(0.006)	(0.006)	(0.007)		
All	Male	-0.00248	-0.00576	-0.00584	12834	0.04
(60-hour threshold)		(0.01)	(0.011)	(0.011)		
	Female	0.000567	0.00231	-0.00149	12217	0.03
		(0.006)	(0.007)	(0.008)		
With big kids	Male	0.0559*	0.0735**	0.037	2108	0.07
(60-hour threshold)		(0.025)	(0.026)	(0.028)		
Without kids	All	-0.0291*	-0.0325*	-0.0298*	4194	0.05
(family size>1, unmarried)		(0.013)	(0.014)	(0.015)		
(60-hour threshold)						

Notes: 10%, 5%, 1%, and 0.1% levels of confidence are indicated by (+), (*), (**), and (***), respectively. Standard errors are in parentheses. Column labels: (1) is the pure sector division of public and private as treatment and control, respectively; (2) adds employees from entirely unaffected public fields, such as public education, into the control group; (3) adds employees with professions where the expected majority would not be affected, into the control group. Time threshold set to one year prior to lagged threshold used in main analysis. All observations up until implementation used to assess placebo effects.

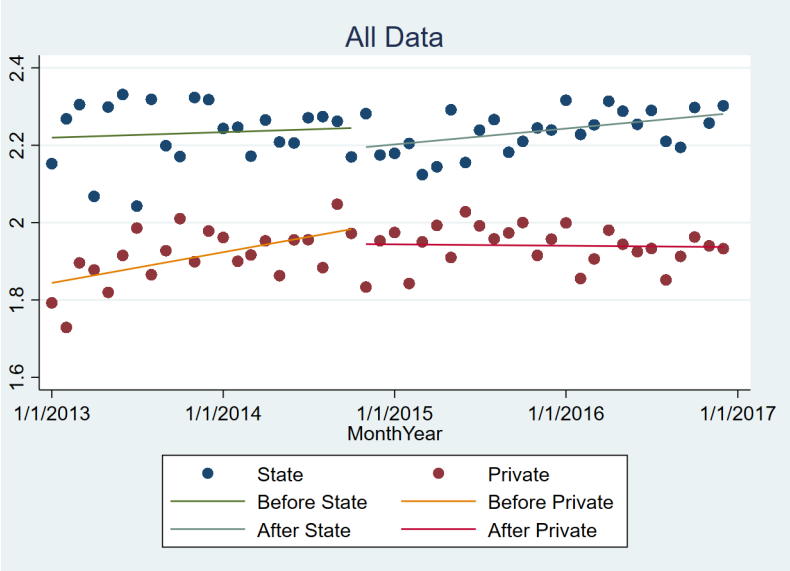
3.5.2.2 Parallel Trends

Parallel trends analyses assess whether the control and treatment groups were on a trend prior to the implementation of the policy in question and diverged thereafter so that the difference experienced between the groups after implementation can be identified as causal. Figure 6 below is a visualization of mean working hours for the entire sample data, using ordinal

integers to represent the intervals, split by the policy implementation threshold. Figures 7 and 8 break that down by gender. The trends are represented by linear best fit lines for the public and private sector groups for the period prior to and post threshold. As the name, parallel trends, suggests, the ideal validation is when the two lines (or the pattern in the data points) prior to the threshold are reasonably parallel to one another to substantiate that the two groups were on a similar trajectory prior to the policy. Post threshold, there should be a level change in the data and/or the lines (or the pattern in the data points) should diverge to confirm that the policy altered their trajectories.

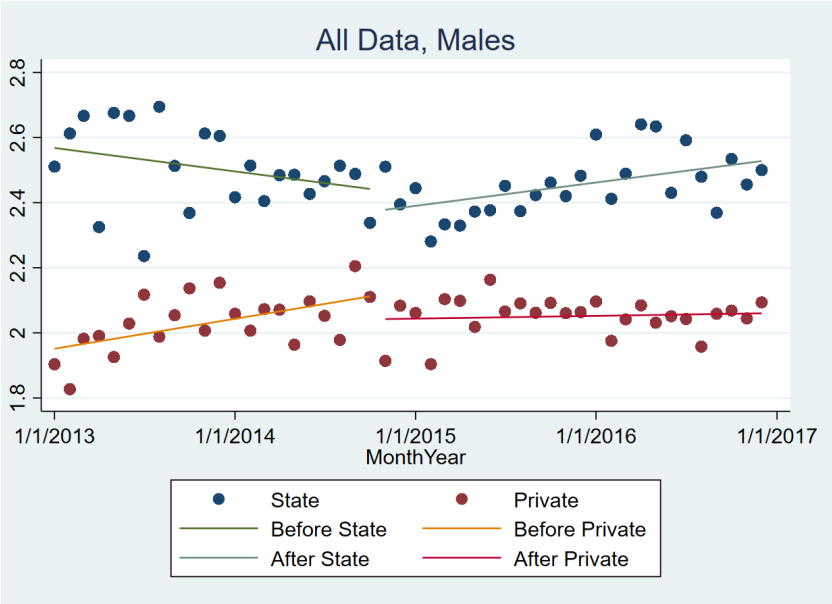
Figure 8 represents a good example of a corroborating parallel trend graph. It is clear that women, overall, between the private and public sectors were following a similar general trend prior to the policy and then diverged thereafter. Figure 6 is a less perfect example but seems to still conform to expectation. Undoubtedly, linear best fit lines are imperfect, and thus latitude on their similarity is expected. Moreover, Figure 6 is almost certainly less perfect than Figure 8 due to the influence of the questionable Figure 7. The diverging trend in Figure 7 is evident, but the pre-policy fitted line for the treatment group seems to be skewed up and to the left by a bunching of some early data points. Sometimes, patterns in the data points, which may not necessarily match the fitted lines due to outliers or bunches, are visually discernible, as is the case with Figure 7. Moreover, there appears to be a similar upwards sloping pattern throughout 2014 to that of the control group. When we examine the trend just one year prior to the threshold, then the pattern does become much more parallel.

Figure 6: Parallel trend scatter plot with linear fitted lines, all data, all observations



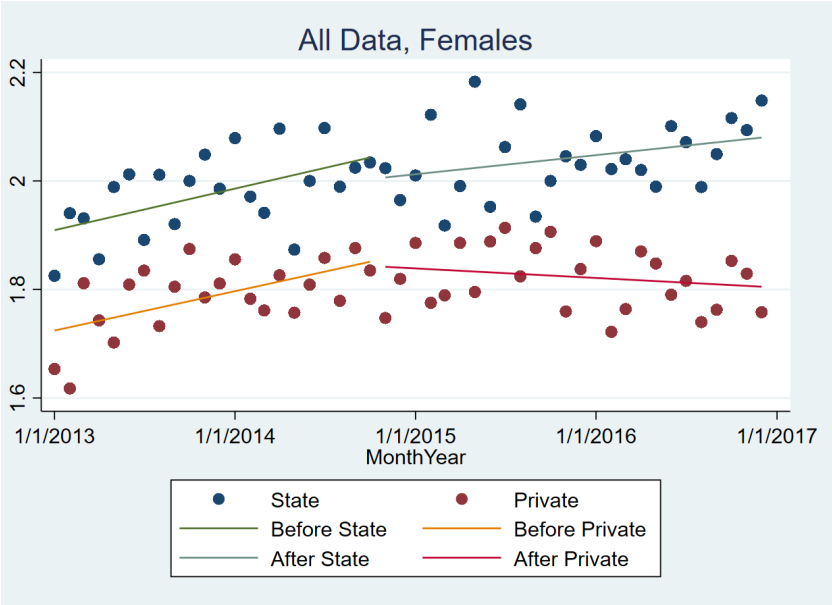
Notes: "Private" includes all workers in the private sector as well as the certainly unaffected public field workers. "State" includes all remaining public sector workers. Threshold is set at two months post policy implementation.

Figure 7: Parallel trend scatter plot with linear fitted lines, all data, males



Notes: "Private" includes all workers in the private sector as well as the certainly unaffected public field workers. "State" includes all remaining public sector workers. Threshold is set at two months post policy implementation.

Figure 8: Parallel trend scatter plot with linear fitted lines, all data, females



Notes: "Private" includes all workers in the private sector as well as the certainly unaffected public field workers. "State" includes all remaining public sector workers. Threshold is set at two months post policy implementation.

While the overall parallel trends are generally informative, when we examine the subsample populations that constitute the main findings of this paper and focus on their trends around the policy implementation threshold, the picture deviates considerably from the overall trends. For example, the trends of the male subsample groups tend to become generally more corroborating, while the female subsamples deviate. Parallel trend figures for every subsample

regression that led to the main findings discussed above are available in Appendix Figures A1-A6. At the 20-hour threshold, both women with and without children have fairly corroborating parallel trend figures. At the 40-hour threshold, the women with children graph, while not perfect, still seems to validate the assumption. The figure for females without children at the 40-hour threshold does not feature very parallel fit lines prior to implementation, though this may not be surprising given the placebo effect analysis outcome. Due to a lack of decisive support for the parallel trends analysis based on the graphs alone, we turn to an alternative method for validation.

We construct a regression based on the DD methodology, but instead of utilizing a single dummy variable for time before and after policy implementation, we create dummy variables for every month in our dataset as well as dummy variables for each month interacting with the single treatment variable. By taking the resulting interaction term coefficients from the period prior to policy implementation and running an F-test on their joint significance, we are able to assess whether they were jointly significant and reject the null hypothesis that they are equal to zero and the groups are the same. As with the visual analysis, this does not represent a perfect confirmation of the parallel trends assumptions but does provide a more rigorous method of assessing whether the parallel trends assumption is broken. See Appendix Table A55 for the F-test significance results analyzing the main subsample groups that constitute the main findings above. Notably, the results show that we cannot reject the parallel trends assumption for women without children at the 40-hour margin and that the parallel trends assumption for men with children at the 60-hour margin is rejected. Due to the latter, the inference in section 3.5.2 is not substantiated and assumably does not hold. Otherwise, we conclude that the parallel trends visual analysis and the supplemental joint significance F-test analysis support the parallel trends assumption for the main analysis findings discussed below.

3.5.3 Discussion

An overall depiction of the effect of the policy on labor participation has been revealed from the main and supplemental analyses. Despite the beliefs and intentions of some members of the Georgian parliament, public office working hours do not seem to have been a "family friendly" barrier to female labor participation and the policy did not cause any statistically significant increase in the extensive margin of employment. Moreover, employees with children reacted to the policy by mostly reducing their working hour engagement. Primarily, the policy negatively affected the ability of full-time employees with children to work the longer hours that they had been working prior to the implementation of the policy. This result echoes the prediction from the WIF conflict type and gender similarity model framework. Secondarily, the engagement of women without children, predominantly married women, substantially increased across both the 20- and 40-hour thresholds. Women without children were probably able to take up most of those hours given up by parents with children because of having more flexible schedules than their colleagues with children, with married women

having flexibility to an even greater extent perhaps due to a more settled-down personal life than their single colleagues. We do not portend to know the exact causes of these behaviors and leave that to the realm of future research.

The ordinal findings of the short-term and September analyses show negative effects on working hour engagement beginning to occur much earlier and dissipating into 2016 and the positive effects beginning to occur later on and accumulating into 2016. These findings indicate that the policy, which caused a 30-60-minute impingement on individual and, by proxy, household schedules, is transient in nature, with stronger effects in the short term that disperse over time as a new steady state is attained.

Delving into the supplemental analyses offers insights of an informative nature. The results indicate that the effects of the policy were not uniform across family types but, as hypothesized in the introduction, were circumstantially disparate, differing in magnitude and direction amongst age- and location-based subsample populations. For example, though the policy mostly resulted in reductions of working hours for parents with children, it seems positive work hour engagement effects at the 20-hour margin were experienced by older, urban and younger, rural women with children as well as urban males with children. This represents the only evidence of any positive, "family friendly" effects resulting from the policy. Furthermore, women without children may not be the only ones who increased work engagement as a result of the policy; older, urban males without children appear to have done so at the 20- and 40-hour thresholds as well.

In addition, the age and location analyses uncovered further inconsistent patterns of effects that may reveal informative insights into those differences. At the 40-hour margin, the negative impact on women with children was almost exclusive to urbanites, which may reflect a more modern trend in domestic arrangements in urban areas. While men with children were negatively affected at the 40-hour margin, those in rural locations were somewhat more so, which may be related to the respectively greater travel distances and inferior social infrastructure. Older men and women with younger children were the most negatively affected at the full-time employment margin, perhaps reflective of the impact of unanticipated, later-in-life fecundity on families. A number of other conflicting patterns across age- and location-partitioned groups indicates that the opposing hypothesized effects and incentives identified in the introduction all seem to be at play. For example, across age groups, it appears that older people with children bore the greater brunt of the negative effects at the 20-hour threshold, younger women without children experienced most of the positive effects at the 40-hour threshold, and younger men were most negatively impacted at the 60-hour threshold. When including location differences, it appears that the resulting positive work engagement effects were most experienced by younger, urban women and by older, urban men without kids at the 40-hour margin.

3.5.4 Further Investigation

We attempt to further enhance the perception of the working hour engagement movements by analyzing the working hour intervals pairwise in order to reduce noise from average changes in both directions throughout the entire sample. We also continue to use the methodological setup to further investigate additional subsample groups to see if we can uncover any more circumstantially specific effects of the policy. One circumstance that we conjecture as potentially influential on one's decision to increase or decrease labor participation at work is if they happen to be working in multiple jobs. Another circumstance is related to the composition of a household. Specifically, we hypothesize that married couples with one partner in the treatment group and one in the control group may face a greater strain upon their previously established status quo. Moreover, this may be especially true for couples with a single vehicle. Full results and discussion of these analyses are available in Appendix 3.1. The results of the pairwise analysis, for the most part, parallel those of the main analysis, implying that the changes are mostly local across the thresholds. No evidence of any effect of multiple employment on the main analysis findings was found with a single discrepancy at the 20-hour threshold, revealing a statistically significant increase for women with children, especially young children (a finding that had been only suggested by the main analysis results, but found at significant levels in this and the pairwise analyses). Regarding mixed sector couples, males had negative effects in terms of work hour engagement, particularly so when the head of the household was in the public sector and when the family had only one vehicle. Females in mixed sector couples in which the spouse was in the public sector showed positive effects in work engagement. Such a combination of results implies gains by women in intra-familial bargaining or a modernization of social norms.

3.6 Conclusion

On September 1, 2014, the country of Georgia enacted a unique policy moving the working hours of public office employees from 10:00-19:00 to 9:00-18:00, impacting the working hour schedules of all affected employees. While not the official or main reason for implementing such a policy, some members of parliament had believed that the new hours would be "family friendly", making it easier for women to balance household and professional responsibilities, and thus increase female labor participation. Thanks to access the Georgian government provides to their household data survey, combined with the fact that the policy did not affect the private sector, we were able to implement a difference-in-differences methodology to accurately analyze whether the policy increased female employment and gender equality. This policy affected an estimated 200,000 employees, yet the impact of this policy had never been evaluated. Moreover, we were unable to find any literature evaluating any policy that exogenously adjusted the working hours of a significant portion of employees in an economy. Nevertheless, since the effects of the policy variously impacted employees across multiple

characteristics, especially by gender and family type, this study is most closely related to work-family conflict, gender inequality, and intra-household bargaining and resource allocation literature. Based on concepts from the work-family conflict literature, we arrived at two opposing predictions for the possible effects of the policy on employees with families.

The results discussed in section 3.5 of this paper reveal that the policy had no significant effect on the extensive margin and, instead, directly and primarily led to a substantial decrease in working hour engagement by full-time employees with children. This result is in accordance with the prediction based on the gender similarity model and WIF conflict type. Although there is some evidence of a modest increase in engagement by part-time employees with children, it does not come close to the magnitude of the negative effect on full-time employees with children. Therefore, we assert that the policy did not directly lead to an increase in female labor participation. While we also found a greater expansion in engagement in the public sector than in the private sector by women without children, the placebo effect analysis discovered that this was a trend already occurring prior to policy implementation and the short-term analyses confirmed that this effect was ordinally second. We infer that those hours gained by married women without children and, to a lesser extent, unmarried women without children, were a subordinate result of the negative effect on working hours of full-time employees with children. Thus, it could be argued that the policy did indirectly increase female labor participation. Furthermore, as the majority of the negative engagement effects fell on male employees and positive effects on female employees, the policy also indirectly improved gender equality by increasing the female side of the gender balance equation of the labor force.

Moreover, there were several additional, informative insights gained into the effect of the policy. As hypothesized in the introduction, the policy caused heterogenous effects with considerable variance in size and direction that were often strongly informed by circumstance, age, and location. For example, the analyses revealed that male employees with older children appear to be both those that had the largest general negative effect on their engagement when working 40+ hours, especially in rural locations, and the largest general positive effect on engagement for part-time employees working 20 hours or less. Hence, despite the female-focused intentions of certain parliament members, the policy seems to have directly affected male employees on both sides of the spectrum more than female employees. We also find that there were especially negative effects on the engagement of male employees who were part of a mixed sector couple, especially when they had only one vehicle. There are even indications that unmarried men without children had some modest negative effects on engagement from the policy.

Of course, women were certainly affected by the policy as well. The negative effects on urban females with children was substantial, especially for those with younger children. On the more positive side, part-time female employees in a mixed sector couple with zero or one vehicle showed considerably positive effects on their engagement. This may also be true for full-time female employees in couples where the head of the household is in the private sector and the

spouse in the public sector, but this result has a questionable sample size. All in all, the additional insights may imply increased female intra-familial bargaining power or that Georgian fathers and husbands (especially in urban areas) have begun to participate more in household duties and are open to more modern feminist outcomes than the UN gender survey found. Both the former and latter explanation imply an occurring or future evolution in social norms.

Our work contributes to the vast literature on working hours in several dimensions. First, it is the first paper that evaluates such a work hour shift policy. Second, it may contribute to the gender inequality, intra-household, and work-family conflict literatures. And third, to a lesser extent, this unique exogenous policy and the multi-dimensional findings of this study may be useful to those with research areas related to work hours and shifts, such as work-life balance, benefits of flexibility, etc., as well as research bodies dedicated to the field, such as "The Shift Project". For instance, the indirect effects of this policy that affected workers differently by their familial conditions may likely provide insights for future research into the myriad work schedule effects on workers and their families, or practical identification of diverse "family friendly" policies as pursued by Saltzstein et al. (2002). Regarding future research, the policy appears to have revealed evolving social norms and affected the amount of time and manner in which family members spend time together. For example, the policy may have caused WIF spillover leading some families to spend less time together, which may negatively influence family well-being, especially for spouses, as, *ceteris paribus*, the more time spouses spend together, the more satisfying the marriage (Kingston & Nock, 1987). Given that Alberts et al. (2011) find that intrahousehold division of labor may be rather universally human in nature, some of our conclusions may directly extend to counterfactual situations around the world. Nevertheless, we only conjecture that the random disruption of a steady state in working hours will likely result in generally negative consequences for employees with children, at least in the short- to mid-term, probably because disruption of household schedules causes work interference with family conflict for both mothers and fathers. For policymakers considering a similar work hour shift, to ensure fewer negative effects, we would recommend that any such policy be accompanied by even greater flexibility, daycare, and/or other WIF-conflict-reducing support for employees with children.

References

Alberts, J. K., Tracy, S. J., & Trethewey, A. (2011). An integrative theory of the division of domestic labor: Threshold level, social organizing and sensemaking. *Journal of Family communication*, 11(1), 21-38.

Angelov, N., Johansson, P., & Lindahl, E. (2016). Parenthood and the gender gap in pay. *Journal of Labor Economics*, 34(3), 545-579.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Baffoe-Bonnie, J., & Golden, L. (2007). Work-study: time use tradeoffs, student work hours and implications for youth employment policy. *Student Work Hours and Implications for Youth Employment Policy* (December 26, 2007).

Bird, R. C. (2015). Why Don't More Employers Adopt Flexible Working Time. *W. Va. L. Rev.*, 118, 327.

Business Information Agency, BIA (2019). BUSINESS INFORMATION ABOUT 49,516 ACTIVE COMPANIES. Retrieved from: https://www.https://www.bia.ge/

Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3), 789-865.

Breshers, S. N., & Fewell, J. H. (2001). *Annual Review of Entomology*, 46, 313–340.

Brines, J. (1994). Economic dependency, gender, and the division of labor at home. *American Journal of Sociology*, 100(3), 652-688.

Camazine, S., Deneubourg, J. L., Franks, N. R., Sneyd, J., Theraulaz, G., & Bonabeau, E. (2001). *Self-organization in biological systems*. Princeton, NJ: Princeton University Press.

Carr, M. D. (2011). Work hours and wage inequality: Evidence from the 2004 WERS. *The Journal of Socio-Economics*, 40(4), 417-427.

Cortes, P., & Pan, J. (2017). Cross-country evidence on the relationship between overwork and skilled women's job choices. *American Economic Review*, 107(5), 105-09.

Dawson, J., Linsell, L., Zondervan, K., Rose, P., Carr, A., Randall, T., & Fitzpatrick, R. (2005). Impact of persistent hip or knee pain on overall health status in elderly people: a longitudinal population study. *Arthritis Care & Research: Official Journal of the American College of Rheumatology*, 53(3), 368-374.

Farulava. (2014, September 2). Part of the public servants will start working at 09:30am instead of 09:00am. *liberali*. Retrieved from http://liberali.ge/news/view/11809/sajaro-mokheleta-natsili-mushaobas-0900-saatis-natsvlad-0930ze-daitsyebs

Ferrant, G., Pesando, L. M., & Nowacka, K. (2014). Unpaid Care Work: The missing link in the analysis of gender gaps in labour outcomes. *Boulogne Billancourt: OECD Development Center.*

Galay, K. (2007). *Patterns of Time Use and Happiness in Bhutan: Is there a relationship between the two?*. Institute of Developing Economies.

Gicheva, D. (2013). Working long hours and early career outcomes in the high-end labor market. *Journal of Labor Economics*, 31(4), 785-824.

Golden, L. (2006). How long? The historical, economic and cultural factors behind working hours and overwork. *Research companion to working time and work addiction*, 36-57.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, 104(4), 1091-1119.

Gutek, B. A., Searle, S., & Klepa, L. (1991). Rational versus gender role explanations for work-family conflict. *Journal of applied psychology*, 76(4), 560.

Hakkert, R. (2017). Population dynamics in Georgia. *An overview based on the 2014 general population census data*, 60-69.

Handley, K., Kamal, F., & Monarch, R. (2020). Rising Import Tariffs, Falling Export Growth: When Modern Supply Chains Meet Old-Style Protectionism. *NBER Working Paper*, (w26611).

Hanse, G. D. (1993). The cyclical and secular behaviour of the labour input: Comparing efficiency units and hours worked. *Journal of Applied Econometrics*, 8(1), 71-80.

Hegewisch, A., & Gornick, J. C. (2011). The impact of work-family policies on women's employment: a review of research from OECD countries. *Community, Work & Family*, 14(2), 119-138.

Herr, J. L., & Wolfram, C. D. (2012). Work environment and opt-out rates at motherhood across high-education career paths. *ILR Review*, 65(4), 928-950.

Holly, S., & Mohnen, A. (2012). *Impact of Working Hours on Work-Life Balance* (No. 465). DIW Berlin, The German Socio-Economic Panel (SOEP).

Hrdy, S. B. (1999). *Mother nature: Maternal instincts and how they shape the human species. New York*: Ballantine Books.

ILO (International Labour Organization). (2019). ILOSTAT database.

Jashi, C. (2005). Gender economic issues: the case of Georgia. Sida.

Kachkachishvili, I., Nadaraia, K., & Rekhviashvili, B. (2014). Men and Gender Relations in Georgia".

Keene, J. R., & Quadagno, J. (2004). Predictors of perceived work-family balance: Gender difference or gender similarity?. *Sociological Perspectives*, 47(1), 1-23.

Kingston, P. W., & Nock, S. L. (1987). Time together among dual-earner couples. *American Sociological Review*, 391-400.

Khunashvili. (2014, August 1). Prime Minister is taking the initiative to start working at 9am in the public sector. *Georgian Press*. Retrieved from http://georgianpress.ge/com/news/view/8422

Knight, K. W., Rosa, E. A., & Schor, J. B. (2013). Could working less reduce pressures on the environment? A cross-national panel analysis of OECD countries, 1970–2007. *Global Environmental Change*, 23(4), 691-700.

National Statistics Office of Georgia, Employment Indicators. (2015). Distribution of employed by institutional sector. Retrieved March 28, 2019 from: https://www.geostat.ge/en/modules/categories/128/databases-of-2009-2016-integrated-household-survey-and-2017-households-income-and-expenditure-survey

National Statistics Office of Georgia. (2019). *Georgia Multiple Indicator Cluster Survey 2018, Survey Findings Report.* Tbilisi, Georgia: National Statistics Office of Georgia.

Orpen, C. (1981). Effect of flexible working hours on employee satisfaction and performance: A field experiment. *Journal of Applied Psychology*, 66(1), 113.

Pieroni, L., & Salmasi, L. (2017). The Economic Impact of Smoke-Free Policies on Restaurants, Cafés, and Bars: Panel Data Estimates From European Countries. *Journal of Policy Analysis and Management*, 36(4), 853-879.

Rangel, M. A. (2003). Alimony Rights and Intrahousehold Bargaining: Evidence from Brazil. *UCLA CCPR Population Working Papers*.

Risman, B. J. (1999). *Gender vertigo: American families in transition*. Yale University Press.

Robinson, G. E., & Page, R. E. (1989). Genetic basis for division of labor in an insect society. *The genetics of social evolution*, 61-80.

Sayer, L. C. (2016). Trends in Women's and Men's time use, 1965–2012: Back to the future? *In Gender and couple relationships* (pp. 43-77). Springer, Cham.

Sayer, L. C. (2005). Gender, time and inequality: Trends in women's and men's paid work, unpaid work and free time. *Social forces*, 84(1), 285-303.

Theraulaz, G., Bonabeau, E., & Denuebourg, J. N. (1998). Response threshold reinforcements and division of labour in insect societies. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1393), 327-332.

Vlasblom, J. D., & Schippers, J. J. (2004). Increases in female labour force participation in Europe: Similarities and differences. *European Journal of Population/Revue Européenne de Démographie*, 20(4), 375-392.

Wasserman, M. (2015). Hours constraints, occupational choice and fertility: Evidence from medical residents. MIT.

Winett, R. A., Neale, M. S., & Williams, K. R. (1982). The effects of flexible work schedules on urban families with young children: Quasi-experimental, ecological studies. *American Journal of Community Psychology*, 10(1), 49.

Yang, T. T. (2020). The effect of workplace pensions on household saving: Evidence from a natural experiment in Taiwan. *Journal of Risk and Insurance*, 87(1), 173-194.

Appendix

This policy evaluation has two appendices. The appendix below offers extended analyses of the further investigations summarized in section 3.5.4. Please follow the link below to the output appendix with all cited tables, figures, and full covariate results. An unabridged appendix with all results is available by request.

Appendix 3.2: https://www.dropbox.com/s/w747ii45rpxm5s5/GPSWHPS.pdf?dl=0

Appendix 3.1. Further Investigation

A3.1.1. Pairwise Analyses

Given the nature of the methodology employed, only the positive β_3 coefficients from the lowest interval pair and the negative β_3 coefficients from the highest interval pair have undeniable value for interpretation, because only those movements are bounded by absolute frontiers (zero hours and all hours greater than 60). These are presented below. As all other pairwise output is not necessarily capturing movements across the given threshold, those results may only be implicative. Nevertheless, the pairwise analyses may provide some additional insight even at the middle margins and are discussed below. The output tables are presented in Appendix Tables A5-A10. Table 7 examines the movements in the weekly working hours variable from the 20 hours or less interval to/from the 21-40-hour interval.

From this pairwise analysis, just above and below the 20-hour threshold, an enhanced picture of the effects of the policy at this margin has emerged. The β_3 coefficients follow the same pattern as the 20-threshold analysis but have become stronger and more statistically significant. Moreover, this perspective also reveals the positive effects on working hour engagement experienced by parents with children, which are most consistent across the treatment specifications for men with children, especially driven by men with older children. Furthermore, both women without children who are married and unmarried seem to be experiencing positive effects from the policy change, though the effect is more consistent for the married ones.

Table 7: DD regression results for weekly working hours, pairwise, 20 hours or less ↔ 21-40 hours

Subsample	Gender	(1)	(2)	(3)	N	R^2
All	All	0.0656***	0.0406**	0.0433**	45937	0.11
		(0.012)	(0.013)	(0.015)		
	Male	0.0383*	0.0278	0.0326	22024	0.11
		(0.019)	(0.02)	(0.022)		
	Female	0.0774***	0.0465**	0.0542**	23913	0.10
		(0.016)	(0.018)	(0.02)		
With kids	All	0.0549**	0.0450*	0.0426*	20275	0.11
		(0.017)	(0.02)	(0.022)		
	Male	0.0565*	0.0585*	0.0584+	9846	0.11
		(0.028)	(0.029)	(0.031)		
	Female	0.0577*	0.0368	0.0376	10429	0.11
		(0.023)	(0.027)	(0.03)		
With small kids	Female	0.0781*	0.0471	0.0777	4453	0.11
		(0.037)	(0.043)	(0.047)		
With big kids	Male	0.102*	0.0915+	0.108+	3313	0.13
		(0.05)	(0.052)	(0.056)		
Without kids	Female	0.0914***	0.0480*	0.0721**	13484	0.11
		(0.021)	(0.024)	(0.027)		
Without kids (family size>1)	Female	0.0852***	0.0421+	0.0779**	11742	0.10
		(0.023)	(0.025)	(0.028)		
Without kids (family size>1, married)	Female	0.0830**	0.046	0.0842*	7994	0.09
		(0.028)	(0.032)	(0.035)		
Without kids (family size>1, not married)	Female	0.0872*	0.0508	0.0536	3748	0.16
		(0.04)	(0.045)	(0.048)		
Just singles (family size=1)	Female	0.179*	0.131	0.0262	1742	0.20
		(0.081)	(0.093)	(0.106)		

Notes: 10%, 5%, 1%, and 0.1% levels of confidence are indicated by (+), (*), (**), and (***), respectively. Standard errors are in parentheses. Column labels: (1) is the pure sector division of public and private as treatment and control, respectively; (2) adds employees from entirely unaffected public fields, such as public education, into the control group; (3) adds employees with professions where the expected majority would not be affected, into the control group.

Table 8: DD regression results for weekly working hours, pairwise, 41-60 hours ↔ more than 60 hours

Subsample	Gender	(1)	(2)	(3)	N	R^2
All	All	-0.0247+	-0.0162	-0.00682	14297	0.05
		(0.014)	(0.014)	(0.015)		
	Male	-0.0165	-0.00259	0.00477	8716	0.05
		(0.019)	(0.019)	(0.02)		
	Female	-0.0218	-0.023	-0.017	5581	0.07
		(0.02)	(0.021)	(0.024)		
Just singles (family size=1)	Male	-0.592*	-0.592*	-0.601*	155	0.72
		(0.291)	(0.291)	(0.288)		

Notes: 10%, 5%, 1%, and 0.1% levels of confidence are indicated by (+), (*), (**), and (***), respectively. Standard errors are in parentheses. Column labels: (1) is the pure sector division of public and private as treatment and control, respectively; (2) adds employees from entirely unaffected public fields, such as public education, into the control group; (3) adds employees with professions where the expected majority would not be affected, into the control group.

Table 8 shows a closer view of the 60-hour threshold. This time the results are weak in magnitude and statistically insignificant across the board of all subsample divisions. It may be that any substantial policy-caused effects of the 60-hour threshold are captured in the pairwise analyses of Appendix Tables A7 and A9. There is a noteworthy result for men who make up the whole of their household. The effect appears to be an extreme decrease in working hours across these intervals. However, the sample size is minuscule, which means the result is almost certainly spurious.[97]

Appendix Table A6 examines the pairwise intervals of 20 hours or less and 41-60 hours. Here only the positive effects experienced by women without children are significant. While not as strong as the effects experienced by married women, unmarried women from households with two or more members now also exhibit statistically significant effects consistently across treatment specifications. The pairwise analysis between less than 20 hours and more than 60 hours in Appendix Table A7 does not have many β_3 coefficients with statistical significance and does not reveal much new information. A consistent negative effect, though neither strong in magnitude nor statistical significance, seems to be occurring for women with children. However, this is a spurious result given the lack of women in the sample who work more than 60 hours. Moreover, the results from the subsample groups of younger and older children are both insignificant. Another noteworthy result in Appendix Table A7 is that men who are the only members of their household display a distinct increase in working hour engagement across this pair, though the sample size is already rather small and probably also indicates only a spurious outcome of happenstance.

As in Tables 7 and 8, Appendix Table A8 is a pairwise analysis that provides an enhanced depiction of one of the main thresholds: just above and just below the 40-hour threshold. It mostly echoes the 40-hour threshold analysis with a strong negative effect on working hours for all people with children, especially for men, and while the effect is more balanced across genders with small children, it is more pronounced for men with older children. Furthermore, women without children continue to display a strong positive effect, driven by women who are married and part of a household of two or more people.

Similar to the 60-hour threshold, the 21-40 hour and more than 60 hours interval pair in Appendix Table A9 shows only negative effects upon work engagement, driven by men with children, especially those with younger children. However, unmarried men without children in households that are made up of two or more people also display a modest negative effect here. Expectedly, between this pair of intervals, women have almost uniformly insignificant and low magnitude results, with the positive effect women without children have at the lower thresholds completely disappearing in terms of magnitude and significance. While women with older children exhibit a single statistically significant, negative β_3 coefficient at the

[97] The negative placebo effect for single household males at the 60-hour threshold, also with a small sample size, further supports the supposition that the findings for single household males in Appendix Table A7 and Table 8 are spurious.

strictest treatment specification, it is another spurious result due to the small sample size of women working more than 60 hours.

A3.1.2. Multiply Employed

The GeoStat survey asked participants if they held secondary employment. By dividing those who answered "yes" and "no" into two different subsample groups, we then evaluated how each group was affected by the policy. Every main analysis table in the Appendix includes the multiply employed subsample. Appendix Table A54 breaks down the two subsample groups into those with one job and those with more than one job, by threshold and gender. Across the thresholds and genders, the results tend to follow the main results with the singly-employed group having stronger, more statistically significant coefficients and the multiply-employed group displaying much weaker, insignificant results. Furthermore, results from the stricter treatment-specification groups generally tend to be reflective of those in the pure public/private sector specification. One result that stands out in opposition to both of these trends is that of female workers holding multiple jobs at the 20-hour threshold for the stricter treatment specifications.

Since the effect at the 20-hour threshold is undeniable, not opposed by the placebo analysis, and the sample size large enough, we further explore this group in Appendix Table A11 by subdividing it into the family-type subsample groups used throughout the analyses. However, as the sample sizes tend to become rather small here, we must weigh the results carefully. Women without children holding multiple jobs continue the previous pattern of increasing their working-hour engagement at the 20-hour threshold. The results also provide another example of a small indication that women with children also increased engagement at this threshold. Additionally, there is a consistent result amongst the treatment specifications showing men without children holding multiple jobs give up working hours as a result of the policy, which is the lone example of such a finding at the 20-hour margin and in opposition to the gains found for older, urban males in the age- and location-based analyses. The magnitude and statistical significance are both strong and the placebo analysis finds no opposing results. However, the size of the sample renders the finding plausible but inconclusive.

A3.1.3. Mixed Sector Couples

To evaluate how mixed sector couples may have been affected by the policy, we limited the subsample to only married couples. We identified which couples had one partner in the private sector and one in the public sector.[98] Next, we summed up all the automobiles, trucks,

[98] This may not be exactly treatment versus control, as evidenced by our treatment specifications. However, technical limitations and sample sizes resulted in this division. Moreover, this analysis is beyond the scope of our main research question and we consider this close enough to satisfy curiosity and possibly inspire future research.

minibuses, and motorcycles into a single variable we dubbed "vehicle" and divided the mixed sector couples into groups that had zero, one, or more than one vehicle. We also then further divided the mixed sector couples into smaller subsamples by which spouse was in the private sector and which was in the public sector. For the vast majority of the couples in the dataset, the "head" of the household in a married couple is the husband and the "spouse" is the wife. Finally, we again divided these subsamples by those who had zero, one, or more than one vehicle. It is presumably not surprising that our sample sizes sometimes dropped far below a minimal level for the central limit theorem to reasonably be in effect. Nonetheless, we present all the results of these analyses as part of every main analysis table in the Appendix.

At the 20-hour threshold (Appendix Table A2), the results of mixed-sector-couples reflect the findings of the corresponding main sample analysis at a generally lower statistical significance. It seems that females in couples without vehicles increase their engagement most at this threshold. While there are not too many divergent results, one that stands out is for men who are part of a couple in which the head of the family is in the public sector and the spouse in the private sector. Their hours seem to be severely reduced below the 20-hour threshold because of the policy. However, given the sample size, this result is probably spurious.

Appendix Table A3 may indicate several new insights in addition to those from the 40-hour threshold in the main analysis, though the sample sizes in the majority of the further divided subsample groups tend to be unreliably small. One finding that does seem to come with a large enough sample size for proper inference is that men in mixed sector couples reduce engagement more than their full subsample counterparts (at a substantially increased percent compared to the main analysis), especially for those in couples with just one vehicle. However, there is a modestly statistically significant effect found in the placebo analysis for the mixed sector couple males with just one vehicle, and thus the policy may be exacerbating an underlying trend. The output also indicates that the effect is driven mostly by men in a mixed sector couple in which the head of the family is in the public sector, but here the sample size is already too small to consider this a reliable inference.

One result for women that may be approaching a large enough sample size is the strong, positive effect displayed by women who are part of a mixed sector couple in which the head is in the private sector and the spouse is in the public sector. Furthermore, the negative effect experienced by males in mixed sector couples is driven mostly by men in couples in which the head of the family is in the public sector. Moreover, the considerable increase in female working hours for mixed sector couples is most driven by women in couples with more than one vehicle and in couples in which the spouse is in the public sector with only one vehicle. Of course, these findings come with the consequential caveat that the sample size is very small in the detailed subsamples.

The 60-hour threshold by mixed sector couple analysis in Appendix Table A4 expectedly returns almost no statistically significant β_3 coefficients, except for mixed sector couples with

the head of the family in the private sector and the spouse in the public sector, but with a dubious sample size.

A3.1.4. Interpretation of Further Investigations

Supplementing the main threshold analyses with the pairwise analyses both confirmed and enhanced many results from the threshold analyses as well as further revealed new findings. Tables 7, 8, and Appendix Table A8 examine the intervals just below and above each of the thresholds in the previous section. Altogether, they reinforce the conclusions above as well as confirm the existence of the few positive, but weak, "family friendly" effects on parents at the 20-hour threshold, especially on men with older children. The analysis also indicates that unmarried women without children increased engagement across both the 20- and 40-hour margins, which had not been evident from the full sample threshold analyses. Moreover, the results suggest that the vast majority of the changes across the engagement thresholds were local, meaning that effects on working hours were most commonly to the adjacent interval rather than causing major gains or losses, which seems echoed in the lack of extensive margin movement. Furthermore, such local movements imply that the interval nature of the data is not capturing the full effects of the policy in the intensive margin of working hours, which may indicate an avenue for future research to elaborate further.

The multiply employed analysis found that the only divergence from the main analysis results (for women) occurred at the 20-hour threshold. Results from the pairwise analysis for those holding multiple jobs at the 20 hours or less versus 21-40-hour interval pair further lent support to this finding. It may be that many of this subpopulation who worked 20 hours or less in the public sector held multiple jobs out of necessity, adding or shifting hours to their public sector jobs once it became possible. Further delving into the 20-hour threshold revealed that female workers without children who held multiple jobs continued the previous pattern of increasing their working hours at the 20-hour threshold. However, we know from the placebo effect analysis that this is probably not directly caused by the policy. There is also fairly strong evidence from a potentially large enough sample size showing men without children who held multiple jobs giving up working hours due to the policy. It could be that the new hours conflicted with their other job(s) and, therefore, they reduced their hours in the public sector job to adjust.

The household composition analyses seem to indicate that being part of a mixed sector couple does appear to make a material difference to those affected by the policy. For full-time male employees, especially those who have only one vehicle, the effects are substantially more negative. For part-time female employees, the effect may be moderately more positive for those in a mixed sector couple with one or zero vehicles, though the latter is probably more related to a lack of wealth and income than to transportation difficulties (i.e. indicative of an employee who will work more hours if the opportunity arises). There seems to be a prescient combination of negative effects experienced by men in mixed sector couples (especially with

just one vehicle and with the head in the public sector) and positive effects experienced by women in mixed sector couples in which the spouse is in the public sector. Likewise, there is some evidence that full-time female employees in couples in which the spouse is in the public sector experience much stronger positive effects, though it is unclear whether the inference is reliable due to sample size. Altogether, these findings may signify an overall change in social norms or female gains in intra-familial bargaining, perhaps affecting resource distribution and household division of labor. This would be quite contrary to the findings of Kachkachishvili (2014).