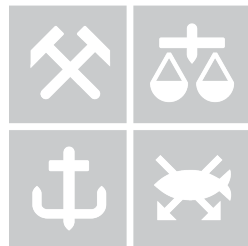




**ESSAYS ON BEHAVIOURAL  
ECONOMICS AND TAX  
COMPLIANCE**

NHH



**ANDREAS OLDEN**

DEPARTMENT OF BUSINESS AND MANAGEMENT SCIENCE

NHH Norwegian School of Economics

A thesis submitted for the degree of

*Philosophiae Doctor (PhD)*

BERGEN 2022

## Acknowledgements

I would like to express my deep gratitude to my supervisor, Jarle Møen. Jarle has provided me with continuous support and guidance, an open door, and wonderful opportunities for learning. My co-supervisor Alexander Cappelen has also been a tremendous support, and included me in the FAIR Choice Lab family.

Furthermore, I am grateful for the support and interactions I have had with faculty and staff at the Norwegian Centre for Taxation and the Department for Business and Management Science. From seminars to lunches, professional advice, and generous time use, the excellent social and professional level at the department, and the research center, are surely one of the reasons for me finishing and the high satisfaction of getting a PhD. In particular I want to mention my co-authors, Aija, Jonas, and Jarle, great friends and colleagues such as Steffen, and all of the PHD-students at the department. A special thanks goes out to the administration, Charlotte, Kristin, Natalia, Stein, Torill and Turid, surely the world would stop spinning without you.

As a part of the extended FAIR Choice Lab family I have attended many great seminars, met great faculty and guests. Mathias Esktröm and Erik Sørensen both deserve a special mention. Further, the PHD-students at NHH have also been of, and continue to be of great importance to me, in particular Ole, Erling, Ingar, Lars and Aija.

I have been fortunate to travel to many great workshops, conferences and

courses. I am grateful for all feedback, interactions, new friends and colleagues, as well as funding. Spending a semester at the University of California, Berkeley was also a great experience, where Shachar Kariv was generous to host me.

This list could go further, but the ones that really deserve my gratitude are family and friends, for their continuous support and great patience, in particular my wife Tiril, and daughter Hedvig.

# Contents

<b>Introduction</b>	<b>1</b>
<b>1 What Do You Buy When No One's Watching? The Effect of Self-Service Checkouts on the Composition of Sales in Retail.</b>	<b>5</b>
1.1 Introduction . . . . .	6
1.2 Literature Review . . . . .	8
1.3 Data and Variable Definitions . . . . .	11
1.4 Identification Strategy . . . . .	18
1.5 Analysis . . . . .	20
1.6 Conclusion . . . . .	29
<b>2 The Triple Difference Estimator</b>	<b>35</b>
2.1 Introduction . . . . .	36
2.2 The triple difference literature . . . . .	37
2.3 The triple difference estimator . . . . .	41
2.4 The difference between two difference-in-differences . . . . .	44
2.5 Identifying assumptions . . . . .	45
2.6 Triple difference as difference-in-differences . . . . .	50
2.7 Inference . . . . .	52

2.8	Concluding remarks . . . . .	53
<b>Appendices</b>		<b>56</b>
2.A	Tables . . . . .	56
2.B	Simulations . . . . .	60
<b>3</b>	<b>Fraud detection by a multinomial model: Separating honesty from unobserved fraud</b>	<b>73</b>
3.1	Introduction . . . . .	74
3.2	The model . . . . .	78
3.3	Monte Carlo Study . . . . .	81
3.4	Conclusions . . . . .	90
<b>4</b>	<b>Fraud Concerns and Support for Economic Relief Programs</b>	<b>92</b>
4.1	Introduction . . . . .	93
4.2	Sample and experimental design . . . . .	96
4.3	Results . . . . .	102
4.4	Concluding remarks . . . . .	107
<b>Appendices</b>		<b>109</b>
.A	Appendix tables and figures . . . . .	112
.B	English translation of experimental instructions . . . . .	115
.C	Screenshots of the experiment in Norwegian . . . . .	119
.D	Pre-analysis plan . . . . .	127
<b>References</b>		<b>128</b>

# INTRODUCTION

This thesis contains four chapters in behavioral and public economics. The chapters use a range of statistical and empirical methods, having in common that they expand on the classical fields and methods of public economics. Since the seminal paper on Prospect theory (Kahneman and Tversky (1979)), the field of behavioral economics has established itself as a distinct field of economics. With its rise, the influence has also been seen in public economics, with papers such as *Testing behavioral public economics theories in the laboratory* (Alm, 2010) and *Using behavioral economics in public economics* (Alm and Sheffrin, 2017), before Bernheim and Taubinsky (2018) combines theoretical and empirical recommendations to get optimal policy under *behavioral* assumptions, in a full integration of the fields. This thesis weaves between the two fields as well.

The first chapter studies how new technology influences demand, through self-service and stigma costs. While mainly business relevant, it gives arguments to tax or subsidize technology for items whose demand is affected by technological change to achieve policy goals, i.e. it shows how technology might be used as a policy tool through an integration of public economics and behavioral economics. While studying the grocery retail market, one obvious extension is to the use of pharmacies, which carries multiple stigma items. In a sense the finding is akin to Chetty et al. (2009), which show how tax salience influences purchasing behavior, which expands on the tra-



ditional public economics understanding of demand through a behavioral mechanism.

The second chapter is a methodological companion piece, showing the identifying assumptions and properties of the triple difference estimator, used in the first chapter. The triple difference estimator is an extension of the difference-in-differences estimator, and as documented in our paper, has grown to be a very common policy evaluation tool.

The third chapter uses Monte Carlo simulations to explore how the EM-algorithm might be useful in a well-known issue concerning (tax) audits, i.e. imperfect detection rates. There has been a sharp rise in the use of audit data to study tax compliance and tax gaps, see Slemrod (2019) and Alm (2019). In any research using audit data to understand tax evasion, neglecting the imperfect detection rates will cause bias in the coefficients and standard errors. We investigate the performance of correcting for imperfect detection rates with the EM-algorithm and bootstrapped standard errors in a Monte Carlo simulation study.

The fourth, and final chapter studies how support for economic relief programs and trust in tax administrations are influenced by the perceived audit rates, in the context of the COVID-19 crisis. In the beginning of the pandemic most countries reduced audit rates to ease the burden on taxpayers and reduce disease transmission risks in the case of on-site audits, which was also recommended by the OECD (OECD, 2020). While reducing audit rates as a policy tool is novel, the concern of real effects of audits is not. Audits might tie up key employees for weeks, which might be detrimental for particularly small firms in times of hardship, something that is partly confirmed by Belnap et al. (2020). Further, it was unclear whether the public would be

sympathetic to these concerns and would support economic relief when the policy was enacted. The latter concern is the focus of chapter four.

The title of the chapters are:

**Chapter 1** What Do You Buy When No One's Watching? The Effect of Self-Service Checkouts on the Composition of Sales in Retail

**Chapter 2** The Triple Difference Estimator

**Chapter 3** Fraud detection by a multinomial model: Separating honesty from unobserved fraud

**Chapter 4** Fraud Concerns and Support for Economic Relief Programs

**Chapter 1: What Do You Buy When No One's Watching? The Effect of Self-Service Checkouts on the Composition of Sales in Retail.** This chapter, single-authored, uses a novel dataset to study real purchasing behavior in the face of new technology in retail. Using a gradual, as-good-as-random roll-out of the technology, including different adaptations, I study sales-pattern and document how certain items experience a sales increase when introducing the technology. The data suggests that the main mechanism is that there is a stigma cost which is eliminated or reduced, and shows how demand is not fully captured by prices and taxes.

**Chapter 2: The Triple Difference Estimator** is a methodological companion piece to Chapter 1, which uses the triple difference estimator to identify the main effects. The chapter is co-authored with Jarle Møen. The triple difference estimator has grown in use and importance, but the identifying assumptions relied mostly on intuition, and to a lesser extent on formal

derivation. We present the identifying assumptions of the estimator, and run simulations to investigate properties of inference.

**Chapter 3: Fraud detection by a multinomial model: Separating honesty from unobserved fraud** is co-authored with Jonas Andersson and Aija Rusina. A common feature of (tax) audit data, is that there is an imperfect detection rate, which can lead to biased estimates, invalid inference or poor predictions. Imperfect detection rates can be treated as a misclassification error, and taking this approach, we run Monte Carlo simulations on the Expectation-Maximization algorithm of Dempster et al. (1977), as shown in Caudill et al. (2005). Using bootstrapped standard errors, we investigate the performance of the methods performance.

**Chapter 4: Fraud Concerns and Support for Economic Relief Programs.** The final chapter is co-authored with Ingar K. Haaland. We use a survey information provision experiment to investigate attitudes and beliefs concerning a novel policy response to economic downturns. In the spring of 2020, when the COVID-19 crisis was at peak levels, most tax administrations responded by reducing or eliminating on-site audits, as shown and suggested by the OECD (OECD, 2020). The idea was to reduce disease transmission, as well as to ease the burden on the taxpayers. We show how information about fewer audits reduces support for economic relief programs, lower trust in the tax administration, as well as reduces the perceived tax fraud detection rate.

## Chapter 1

# What Do You Buy When No One's Watching? The Effect of Self-Service Checkouts on the Composition of Sales in Retail.

Andreas Olden\*

### Abstract

Buying items that are unhealthy or are of a private nature may carry a stigma and cause embarrassment. I analyze whether the anonymity provided by self-service checkouts changes customers' shopping patterns in grocery stores. I look at a natural experiment where two stores in a grocery-chain implement self-service checkouts. Using a triple difference estimator, comparing the sales of stigma items to the sales of mundane items and to the sales of a group of control stores, I find that the sales of stigma items increase by 10 percent. The

---

\*Olden: Department of Business and Management Science, NHH Norwegian School of Economics. This project has many to thank for its current form. I would especially like to thank Jarle Møen, Mathias Ekstrøm, Einar Blix Huseby and Alexander Cappelen for invaluable feedback. The project has been presented at 68 Degrees North Conference on Behavioral and Experimental Economics (2017) and the Annual Congress of the European Economic Association (EEA) in Lisbon (2017). I thank the conference attendees for their feedback.

increase comes from the product lines *Microwaveable food, frozen pizza, intimate hygiene, potato chips, and alcohol*. Also, fully converting to self-service seems to scare away some customers and decreases overall sales. (JEL D12, D90, M20, L81)

## 1.1 Introduction

Imagine standing in line at your local convenience store. Your basket contains frozen pizza, soda and chocolate for Friday movie night. Maybe you are planning to purchase candy, condoms or tampons. The content of this basket is not something you want to share with others. Yet, in many purchasing situations, this is what you are forced to do. First, you stand in line with people around, then you have to display your items at the registry for scanning. Having your items displayed like this may cause a mild embarrassment. Goldfarb et al. (2015) call this embarrassment a social friction, and show that it can change your behavior. Dahl et al. (2001) show that even an imagined social presence may cause embarrassment, while Ariely and Levav (2000) argue that in social settings, people choose options that undermine their personal preferences because of self-presentation goals.

Self-service checkouts allow you to move items anonymously from basket to bag, removing the social component. This should increase the sales of items prone to social frictions. I suggest two reasons for why a product may be a stigma product and cause a social friction; first that the item is unhealthy, second that the consumption of the item is of a private nature. To investigate these mechanisms, I look at a natural experiment where two stores in a grocery-chain implement self-service checkouts.

The simplest approach to check if self-service increased sales of stigma items would be to compare the sales before and after the systems were introduced. However, this may pick up general sales trends. I control for this by comparing sales growth to the sales growth of control stores that did not introduce self-service. Another problem is that self-service checkouts may change sales in other ways than through reduced friction, for instance, by making the store more efficient, thus increasing all sales. To control for such effects, I compare the changes in the stigma category to changes in a reference category of items that are not stigmatized. To incorporate both controls in one estimation framework, I use the triple difference estimator.

All stores are in urban Scandinavian areas, and in the same country. I have daily information on quantities sold by detailed product codes. The chain coordinates prices and campaigns nationally for all the products used in the analysis, and all products used are available in all stores. The data spans more than four years, starting in January 2010. Treatment store 1 introduced the system in the summer of 2011 and kept no traditional cash registers. Treatment store 2 introduced the system in the summer of 2012, but kept one manned cash register. Both of the stores that introduced self-service introduced the systems at, or close to the original registers, and no major rebuilding occurred.

The estimate for the overall effect of the sales of stigma items is a 10 percent increase. This corresponds to an additional sale of about 150-200 units per day in the stores analyzed, and is driven mostly by the product lines *Microwaveable food, frozen pizza, intimate hygiene, potato chips, and alcohol*. There are also clear interactions between the choice of technology, items with an age limit and their physical presence in the store. Considering the size of the

estimates, knowledge of these types of effects may increase profit directly through the investment decision, or imply re-optimizing the product-cost cross-subsidization. Also, due to the size of the market and the nature of the products, the findings may warrant public policy responses. The analysis is restricted to the grocery market, but the suggested mechanisms imply that there are other markets that may experience similar frictions, in particular pharmacies.

## 1.2 Literature Review

There is a relatively large literature in psychology and other social sciences that document a feeling of embarrassment that arises relating to different situations and products. However, those studies do not document changes in real behavior. To remedy this, Engelbrecht et al. (2021) has created a computer-simulated shopping experiment tool to conduct evaluations of in-store interventions to support more sustainable food choices.

One notable exception is Goldfarb et al. (2015) who considers two case studies where retailers have changed the retail format towards less social interaction. They find that in both cases, there is an increase in the sale of items that are possibly embarrassing to buy, and they attribute this increase to the more anonymous purchasing situation.<sup>2</sup>

The first case study in Goldfarb et al. (2015) is a field experiment in Systembolaget, Sweden's government-run alcohol retail monopoly, conducted in the 1990s. Systembolaget has a monopoly on retailing beverages with more than 3.5 % alcohol. Initially, all items were behind a manned counter, but then

---

<sup>2</sup>See Goldfarb et al. (2015) for a more comprehensive literature review of social stigma.

some of the stores were allowed to convert to self-services stores. The stores that converted experienced an increase in sales, and the increase was larger for items with names that were hard to pronounce. The authors attribute this increase to the removal of the customers' fear of seeming unsophisticated by pronouncing the names incorrectly. The study has two main limitations. First, the authors cannot rule out that the increase is just reduced search costs, meaning that it became easier and less costly for customers to view the full selection. Second, they cannot rule out that customers avoid items with difficult names to avoid misunderstandings with the clerk. My study does not suffer from these issues, as the items do not change place and the social interaction is visual, and not through dialogue.

The second case study in Goldfarb et al. (2015) is a pizza chain that introduced online orders. The main finding is that relative to phone orders, online orders increased the number of ingredients and calories. The authors attribute this to people not wanting to seem finicky, ordering strange combinations or toppings, and that they are afraid of seeming unhealthy ordering items such as extra cheese.

The disadvantage of the second case study is that there may be unobserved selection of customers into the web platform, which they cannot completely rule out. In addition, ordering over a phone may induce a feeling of urgency due to the social interaction with the clerk, while the web platform allows for time to reflect. This may affect the orders. While they do control for the role of miscommunication by looking at the errors in the orders, miscommunication may also be an issue. In my study, these problems are not relevant, as there is no time pressure when shopping, and there are limited miscommunication possibilities.



By addressing several of the limitations in Goldfarb et al. (2015), I substantiate that social interaction has real consequences in retail. Studying the grocery market, I also provide evidence from a large market that uses a different physical infrastructure from that of Goldfarb et al. (2015).

Other related works include Dahl et al. (1998) who suggests that buying condoms is embarrassing, and that this embarrassment may reduce condom sales and have negative public health consequences. Furthermore, they also show that embarrassed people have a tendency to buy from vending machines. Dabholkar et al. (2003) also finds some evidence that people use self-service to avoid social interaction. For more references concerning potential stigma items, see Section 3.4 on category selection.

There are also related work on how technology changes sales. For instance Adamopoulos et al. (2020) shows how an Internet-of-Things (IoT) purchasing situation, with fewer frictions, increases sales. Harris-Lagoudakis (2021) shows how online shopping increases healthier purchasing patterns, however, as they quote in their introduction this might be because it reduces impulse shopping behavior, which is nicely symmetrical to this paper. Further, Gavett (2015), a Harvard Business Review interview with Ryan Buell, claims that McDonald's with self-service kiosks experience increased order values and more upselling.

There are also several papers that show differential adaptation of new technologies which might be relevant for explaining differential effects between technologies. Chandrasekhar et al. (2018) uses a field setting to explore how social stigma and signaling can inhibit learning. While signaling dominates, the shame effect is strong in socially close pairs, for instance network distance and caste co-membership. Nunan and Di Domenico (2019) iden-

tifies a research gap when it comes to older consumers and adaptation of new technologies. Yin et al. (2019) show how culture might play a part in the adaptation of new technology. Finally, Mickeler et al. (2021) studies how anonymity affects seeking information in a lab setting, and finds that psychological costs are particularly pronounced for women.

## **1.3 Data and Variable Definitions**

### **1.3.1 Data and Data Cleaning**

The dataset contains daily sales in quantities of all products in 14 grocery stores, from January 2010 to June 2014. Treatment store 1 introduced self-service in the summer of 2011, while treatment store 2 introduced self-service in the summer of 2012. In total, there are about 27 million observations. The main analyses are conducted on subsets of these data, containing about 38 000 observations, where sales are aggregated to the daily sales per store, but split on stigma items and reference items.

I want items that are available in all stores and have therefore removed locally produced items. Only positive sales were included, but the negative observations were few in number, and probably stemming from inventory. I have chosen to err on the inclusive side when choosing items, meaning that the categories provided by the grocery chain remain mostly intact, except for clear errors.

There is no price data available. Fortunately, prices and campaigns are nationally coordinated, so all stores have the same prices. Among the chosen categories, the only exceptions are items that are about to expire, but

these items are likely to be few in numbers and unrelated to other store characteristics.

### **1.3.2 The Stores**

There are 14 stores in the dataset, and they are all located in medium to large cities within a Scandinavian country. There are two treatment stores. Treatment store 1 did not keep any manned cash registers, but converted fully to the automated option. Treatment store 2 kept one manned cash register. The different implementation strategies may have different effects, which warrants a distinction between them in the analysis. The two stores did not change much except for this conversion, and the automated option is either at, or close to the original location of the manned cash registers. The self-service checkout area is self-contained and has several automated registers, in which you place your basket on one side of a scanner, you pick up an item, scan it, and put it straight into a plastic bag on the other side.

Both stores automatically calls on an employee if you scan a product with an age limit. The employee then comes over, checks your ID, and leaves again. This introduces some social interaction, and for this reason I make a distinction between items with and without an age limit. There may also be people that are uncertain about how the age verification process works, and avoid self-service when buying items with an age limit, which again might yield interactions with whether the store converts fully. Based on this, one might expect the stigma effect to be smaller for items with an age limit. However, many of the items that are the most likely to be stigmatized, such as cigarettes, have an age limit, so the total effect is unclear.

As of late 2013, it became possible to register your fingerprint for age verification purposes. Since I do not have data on the implementation, and it only affects a short period at the end of the dataset, I do not analyze it explicitly. However, one should expect that this technology increases the sales of stigma items with an age limit.

### **1.3.3 The Market**

The stores have a recognized brand name, and they are well known in the regions they operate. Their market is urban, inhabited areas, competing mostly for every day grocery shopping. Consumer research for the country in question shows that the most important factor for choosing grocery stores is proximity to home and quality of products. These findings are stable over time (Lavik and Schjoll, 2012). The survey by Lavik and Schjoll (2012) shows that 55 percent of the respondents ranks proximity to home as very important, and 83 percent use the closest store for their week-day shopping. Also, only 6 percent use more than 3 stores regularly, and 72 percent use only 1 or 2 stores. The market is further characterized by 57 percent of the customers shopping at least 3 times per week. Taken together, this implies a market that is strongly proximity- and habit driven, with many repeat customers going to the same stores, close to home.

### **1.3.4 Selection of Categories**

A combination of broad criteria from the literature, and a survey to further validate the selection process, was used to choose the products to include in the study, as well as practical considerations. There are 181 categories of

items, all made available to me by the chain. Some categories are very large. Others are very small. The average store has about 3000-5000 different items at any given time, but it changes depending on season, trying new product lines etc. The categories are mostly consistent, but there are inconsistencies and mistakes, so one must manually go through them. All categories can be seen in Figure 1.1.

### **Selection of the Stigma Categories**

Two criteria were used to select products for the stigma category. First, that a product is unhealthy, second, that a product is of a private nature.

Unhealthiness as a marker of stigma was chosen for many reasons. First of all, there seems to be strong indicators that people believe that "you are what you eat" (Stein and Nemeroff, 1995), and that you are judged negatively for eating unhealthy foods. Second, it seems that stereotypes concerning food are used for impression management (Vartanian et al., 2007). Third, people know what is healthy or not according to public dietary guidelines (Povey et al., 1998).

Items of a private nature are somewhat more difficult to define. There is a broad range of items that may have some stigma attached to them. Dahl et al. (1998) suggests that buying condoms is embarrassing, while (George and Murcott, 1992) argues that products related to female hygiene are stigmatized. Pharmaceuticals are arguably also products of a private nature.

Combined, the two criteria yielded fifteen potential stigma categories, namely cigarettes, sanitary items, other tobacco products, micro-waveable food, cake, candy, frozen pizza, potato chips, cookies, intimate hygiene products, alcohol,

ice cream, flavored milk, soda, and pharmaceuticals.

### **Selection of the Reference Categories**

There are several reasonable expectations for sales changes when a store introduces self-service checkouts. The normal cash registers may be bottlenecks. Removing them makes the store more efficient, attracting more customers and increases all sales. It is also possible that introducing new technology may scare away some customers. In order to isolate the effect on stigma sales caused by the new and anonymous purchasing situation, I compare the sales of stigma items to a reference category. The criteria for choosing reference items is that; they are not stigmatized, common, mundane, have a stable sales pattern, and are sold in all stores. This gives a representative picture of the customer base that visits each store, as well as an indicator of the effects on the composition of sales.

This selection criteria yielded eleven categories, namely: light bulbs, cereals, spices, storing products (for food), cleaning products, milk, candles and napkins, shampoo, children's food, other children related articles, and diapers.

### **Survey refinement**

To validate the categories a survey was conducted: The results can be seen in Figure 1.1.<sup>3</sup> Inspired by Goldfarb et al. (2015), the respondents were given

---

<sup>3</sup>100 responses were collected, 79 of them were complete. The survey was conducted on a student pool. Note that all stores are in urban areas, where the inhabitants on average are younger and better educated than a representative sample of the Norwegian population. From this perspective the student pool might be an appropriate choice. The survey was conducted in Norwegian. The respondents were given a random subset of 20 categories,

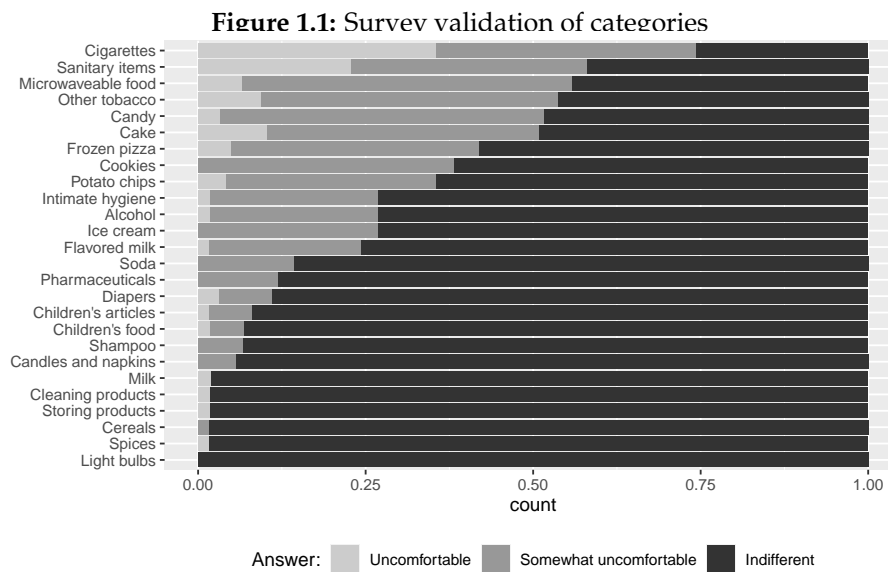
the following instructions:

Imagine that you are an average Norwegian on your way to your regular grocery store. The store is in your neighborhood and you often go there. In the store, you sometimes meet neighbors, and you have had repeated interaction with the employees. On the next page you are asked to give your opinion on what you think an average Norwegian thinks about buying a range of products in this situation. All answers are anonymous, and you don't have to answer all questions.

The respondents were given the option of answering whether they are indifferent, somewhat uncomfortable, or uncomfortable. Figure 1.1 clearly shows that the categories are appropriate. However, some of the reference categories, namely candles and napkins, shampoo, children's food, other children's related articles, and diapers, are not entirely clear. This might be because of the age of the respondents. They probably do not have children. I will conduct robustness checks on these products, running the main analysis both with and without them.

---

and the order was randomized.



All potential categories for the analyses are included. The answer alternatives are whether the respondent is indifferent, somewhat uncomfortable or uncomfortable. The respondents are told: "Imagine that you are an average Norwegian on your way to your regular grocery store. The store is in your neighborhood and you often go there. In the store, you sometimes meet neighbors, and you have had repeated interaction with the employees. On the next page you are asked to give your opinion on what you think an average Norwegian thinks about buying a range of products in this situation. All answers are anonymous, and you don't have to answer all questions." They are then asked to rank a random subset of the categories.



## 1.4 Identification Strategy

### 1.4.1 Triple Differences

### 1.4.2 Triple Differences

I use a triple difference estimator in order to account for both sales trends that affect all stores, and effects that are not due to social frictions, such as increased efficiency. For a full exposition of the triple difference estimator and its properties, see Olden and Møen (2020). The estimator is presented in Equation 1.1. One way to view the triple difference in this setting is that it produces one difference-in-differences for the reference items, one difference-in-difference for the stigma items, and then takes the difference between the two. The final difference estimate, capturing the additional sales coming from stigma items relative to the reference items, while accounting for counterfactual trends in the control stores, is given by the coefficients on  $TreatmentStore_s * TreatmentPeriod_t * StigmaProduct_i$  in Equation 1.1.  $TreatmentStore$  is a treatment store specific dummy variable,  $TreatmentPeriod$  is a treatment period specific dummy variable,  $StigmaProduct$  is a stigma product dummy variable. The  $s$ ,  $t$ , and  $i$  denotes store number, time period and product category. The  $\alpha$ 's are store fixed effects. Unless otherwise specified, the outcome variable is log-transformed daily sales. This gives the coefficients an approximate percentage increase interpretation.

$$\begin{aligned}
\text{Log}(Qty_{sti}) = & \alpha_s + \text{controls} + \\
& \beta_1(\text{TreatmentStore}_s * \text{TreatmentPeriod}_t * \text{StigmaProduct}_i) + \\
& \beta_2(\text{TreatmentStore}_s * \text{TreatmentPeriod}_t) + \\
& \beta_3(\text{TreatmentStore}_s * \text{StigmaProduct}_i) + \\
& \beta_4(\text{TreatmentPeriod}_t * \text{StigmaProduct}_i) + \\
& \beta_6\text{TreatmentPeriod}_t + \beta_7\text{StigmaProduct}_i + \epsilon_{sti}
\end{aligned} \tag{1.1}$$

The control variables included are year and month fixed effects, and store-stigma interacted fixed effects which accounts for heterogeneity in product category size. There are also interactions variables between the treatment store dummies and dummies for the last two months prior to implementation, and then for the three months after implementation. The final pre and post controls allow for fire sales, rebuilding, implementation issues and some response time for customer adoption.

It is well known that difference-in-difference type estimations with few clusters can have significantly biased error terms (Bertrand et al., 2004). According to Cameron and Miller (2015) this can be accounted for by using the Wild Cluster bootstrap with Rademacher weights.<sup>4</sup> I implement this procedure across most analyses, with additional specifications as robustness.

A potential worry with the identification strategy is that some of the measured effect may come from store rebranding, and not stigma. Simply put, the store change may attract or scare away a particular type of customer

---

<sup>4</sup>The procedure is described in Cameron and Miller (2015) and available in several R packages. Simply told, it transforms clustered residuals by doing a random draw from the values [-1, 1], each with probability p= 0.5, and multiplying it with the residuals.

based on their perception of the store. This may in turn change the composition of items in the average basket. This will not bias the coefficient estimates, but it will change the interpretation. The sales gain or loss from converting to self-service will then only apply to the early adapters of the technology, and would be driven by substitution between competing stores. Given the market structure described in section 3.3, with proximity to home and product quality being the most important factors for the choice of grocery store, I do not believe this is an important issue, particularly not for store 2 which keeps a manned cash register. In store 1, however, some customers might react to only being able to use the automated option. I return to this issue when discussing the empirical findings.

## **1.5 Analysis**

### **1.5.1 Descriptive Statistics and Parallel Trends**

Descriptive statistics are presented in Table 1.1, and the development over time of the most important outcome variables are shown in Figure 1.2.

From Table 1.1, we see that there are clear level differences between the various stores and product categories. This is to be expected. Total sales of stigma products ranges from a daily average of 1209 to 1581, depending on whether we look at one of the treatment stores or at the average of the control stores. The average daily sales of the reference items range from 124 to 257.

For a difference-in-differences estimation to have a causal interpretation, we need the parallel trend assumption to hold (Angrist and Pischke, 2008). The parallel trend assumption states that in the absence of treatment, the

treatment group would have had the same sales trend as the control group. The assumption is typically validated by comparing the pre-treatment trend of the treatment stores with the pre-treatment trend of the control stores. The more similar they are, the more reasonable the assumption is.

The triple difference estimator is in essence, the difference between the estimated effect from two different difference-in-difference analyses. However, rather than requiring a parallel trend for each of the two categories, it only requires that the difference between the two categories follows a parallel trend. To see why, imagine a general sales increase affecting only the treatment stores. Individually, this would bias both the difference-in-difference estimator for the stigma items, and the estimator for the reference items. However, since they are both affected in the same way, deducting the first difference-in-difference from the second, the final difference will not contain the general sales increase, and therefore not be biased.

Figure 1.2 shows the development over time for the most important product categories, decomposed to the mean of the control stores, and for each of the two treatment stores. The first insight is that there is substantial seasonality. We also see level differences between the stores, and a large drop around implementation for treatment store 2. The regression analysis controls for all these issues.

The vertical lines in Figure 1.2 show when the two treatment stores introduced self-service checkouts. Evaluating the trend before the implementation of self-service gives some cause for concern. For treatment store 2, there seems to be an increase in the log-stigma sales leading up to the implementation of self-service. This increase is not seen in the control stores. However, the reference category in treatment store 2 exhibits the same behavior. As

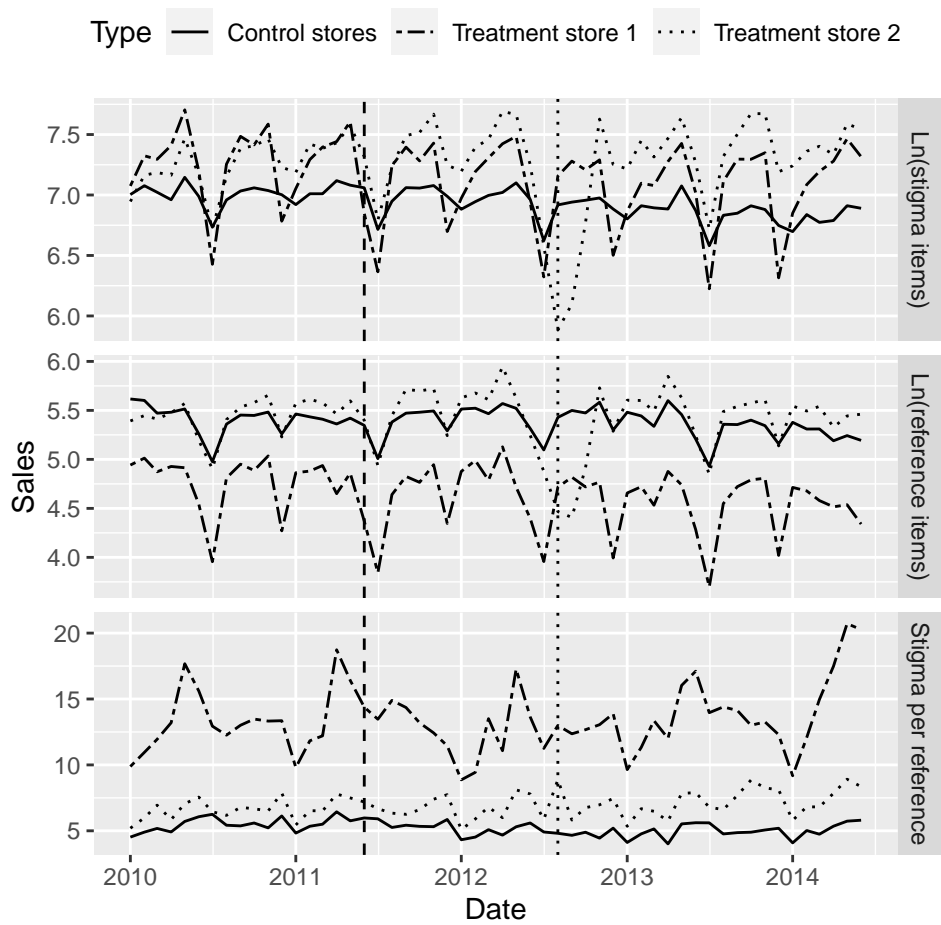
mentioned, what is important is that the difference between the stigma items and the reference items follow a parallel trend. To further underline this point we can look at the third graph from the top, which shows the development of the sales of stigma items relative to the sales of the reference items. It does not give cause for concern since it has a development over time that is very similar to the control stores. This suggests that the trends in the difference between the stigma items and the reference items are quite similar in the treatment and control stores.

Furthermore, the increase is in the peaks, not in the dips, and most of the observations are not from the peaks. In addition, the final peak before implementation is explicitly controlled for as it is quite possibly caused by a fire sale. Ignoring the final peak, the picture changes substantially, and it does not look as if there is differential pre-treatment trending.

**Table 1.1:** Descriptive Statistics of Quantities Sold per Day

Statistic	N	Mean	St.Dev	Min	Max
<b>Control stores</b>					
Stigma items	1385	1208.6	344.3	325.6	2402.0
Stigma without an age limit	1386	742.8	182.4	175.6	2132.6
Stigma with an age limit	1386	466.5	211.9	68.0	1438.7
Reference items	1385	256.8	61.6	74.6	760.9
<b>Treatment store 1</b>					
Stigma items	1338	1463.6	666.7	2.0	4136.0
Stigma without an age limit	1340	1003.3	449.1	1.0	2073.0
Stigma with an age limit	1340	458.1	331.2	1.0	2773.0
Reference items	1340	124.7	62.8	1.0	460.0
<b>Treatment store 2</b>					
Stigma items	1377	1581.0	639.3	15.0	8557.9
Stigma without an age limit	1377	1013.5	435.1	6.0	7877.9
Stigma with an age limit	1377	567.5	333.3	4.0	4896.0
Reference items	1377	246.4	96.1	1.0	804.0

**Figure 1.2:** Graphical Representation of Sales



All figures are plots of monthly averages of daily sales. The control stores is the average of all twelve of them. The vertical lines are the implementation of the self-service checkouts for the treatment store with the same line type.

## 1.5.2 Results

The main empirical results are given in Table 1.2. From Column 1 we see that introducing self-service checkouts increases the sales of stigma items relative to the reference items by 10.1 percent. This is given by the coefficient at the interaction term  $STIG*TS*TP$ . The estimate is highly significant. Since there are two treatment stores, this is an average of the effects in the two stores.

In Column 2, I estimate the effect in the two treatment stores separately. The division is achieved by interacting the stigma dummy with dummies for each treatment store, and their respective treatment periods. The estimated effect for treatment store 1, the store that converted to self-service only, is 9.7 percent. The estimated effect for treatment store 2, the store that kept a manned register, is 10.5 percent. A back of the envelope calculation suggests that these effects correspond to a relative increase of 150 and 200 additional units sold of stigma items per day.

The coefficients from  $TS1*TP1$  and  $TS2*TP2$  give the estimated effects of introducing self-service on all items, excluding the stigma effect. The estimated effect for treatment store 1 is -18.5 percent, meaning that there has been a significant decrease in sales after introducing the self-service checkouts. Total sales will be moderated by the positive stigma effect, but is still negative. In treatment store 2, there is a positive and statistically significant estimated effect of 9.8 percent. In total, this store has experienced a positive and significant sales increase.

Note that the decrease in sales in the store that converted fully to self-service is consistent with the embarrassment idea. Fear of potential embarrassment of not being able to use or understand how the self-service checkouts work

may scare away customers, just as the anonymity of the purchasing situation may attract them. Further, this decrease is not seen in the store that kept a manned cash register, which also fits well. Instead of avoiding that store, a customer can simply choose the manned cash register instead of the self-service alternative.

Given the proximity- and habit driven market structure discussed in section 3.3, the general sales decrease in treatment store 1 is likely caused by the technology aversion described above, and not other effects like store rebranding. However, if the customers with technology aversion are not representative, the sales decrease makes the interpretation of the stigma sales estimate for treatment store 1 a bit uncertain, as it could be partly driven by a change in the customer base. Note, however, that the estimated stigma sales effect in treatment store 1 is smaller than the estimated effect in treatment store 2. Hence, the technology aversion effect seems to go in the opposite direction of the stigma effect, causing the stigma effect to be underestimated in treatment store 1.

There are some additional points to make on how the effect of the self-service checkouts are generated. First, it is unlikely that the effect is generated by substitution between items. This is because I would expect the substitution to be either within stigma category, for instance by buying a regular soda instead of diet soda, or by substitution between stigma categories, for instance by switching from flavored milk to soda. While there may be such substitution, neither is captured in my analysis. This is because it is conducted on aggregated stigma sales in quantities, so that changing product, increasing content size of product and changing between categories will still only add the same quantity as the original product. This makes my approach



a conservative estimate of the stigma effect.

Column 3 estimates a difference-in-differences model, using the ratio of daily sales of stigma items to the daily sales of the reference items as the outcome variable. The difference-in-differences estimation provides a slightly different functional form, the modeling framework is somewhat simpler, and the parallel trend assumption is easier to assess as it directly corresponds to the third graph from the top in Figure 1.2, which is discussed earlier. The downside of the model is that it produces less information than the triple difference, and that extreme values might become more influential (if dividing by a number close to zero).

The treatment effect is now given by the interaction between the treatment stores and their respective treatment periods. The estimated effect for treatment store 1 is an increase in the ratio of stigma items to reference items of 1.1. The average ratio for the whole period is about 12. A back of the envelope calculation holding the reference sale constant, suggests that the estimated effect corresponds to an increase of about 150 units, consistent with previous findings. For treatment store 2, the estimated increase is 0.5, which with an average ratio of about 6 suggests an effect about 10 percent, which is again, is similar to the main specification.

Table 1.2, Column 4, is a triple difference comparing the sales of stigma items without an age limit to the sales of the reference category. The reason for this split is that within the stigma categories, products without an age limit may be less stigmatized, suggesting a smaller effect. On the other hand, the social interaction is not completely removed when buying items with an age limit. This goes in the opposite direction and suggests a larger effect.

The estimate for treatment store 1 is a 6.8 percent, which is smaller than the overall effect, and weakly significant. The estimate for treatment store 2, which did not convert fully, is 11.4 percent. This is stronger than the overall effect. Taken together, no clear conclusion emerge, with respect to the role of age limits.

### **Robustness**

Table 1.3 re-estimates Table 1.2, but without the categories of which there was some uncertainty, see section 1.3.4 and Figure 1.1. The main results, hold, albeit with changes in the size of the coefficients. For instance, the stigma effect is estimated to 7.4 percent in treatment store 1, while it is estimated to 14 percent in treatment store 2, while the original results were about 10 percent for both. Some of this might be because the reference items category is smaller, but it also suggests that the estimated effect is not uniform across categories. This, of course, should not be expected, as the product lines might have different stigma (governed by purchasing behavior).

Table 1.4 re-estimates Column 1 of Table 1.2. However, Table 1.4 Column 1 uses classic cluster robust standard errors, Column 2 drops all control variables, Column 3 clusters standard errors on store and year, while Column 4 introduces a store-specific linear time trend as robustness (Angrist and Pischke, 2008, p. 238-241). Overall, there are no material changes, and the results seem very robust.

## Decomposing the Findings on Stigma Categories

Table 1.5 presents triple difference estimates for each of the 15 categories of stigmatized items, all of them compared to the same reference category as in the main specification, and with the full reference category. Table 1.6 is the same table, but excluding the uncertain categories from the reference category. The most interesting aspect to look at is which categories drive the main results. Note, however, that the categories are much smaller individually, which gives more variance and uncertainty than in the main results.

Defining a consistent category as one for which there is a positive and significant (ten percent level) coefficient for both treatment stores, and using all the reference categories, the only consistent category is *frozen pizza*. Excluding the uncertain reference categories, we can add: *Microwaveable food*, *potato chips*, *intimate hygiene*, and *alcohol*. A weaker consistency criterion, looking only at whether both coefficients are positive, and using the full set of reference categories, the consistent categories are: *Microwaveable food*, *potato chips*, *intimate hygiene*, and *alcohol*. Excluding the uncertain reference categories, we can add: *cookies* as well. Note that all categories except *cookies* have the same sign regardless of whether we include the uncertain categories in the reference category or not. I will therefore not include *cookies* as a main driver of the effect.

There are two additional points to make here. First, seven categories are positive for treatment store 1, while 10-11 are positive for treatment store 2, depending on whether we use all the reference categories or not. This is not surprising since the main results suggested that full automatization scares away some customers. Second, among the products with an age limit,

only alcohol have a positive coefficients. Interestingly, alcohol is the only product with an age limit that is physically available in the store. Tobacco and pharmaceuticals have to be ordered in the check-out zones, and this might be a barrier for increased sales.

**Table 1.2:** The Effect of Self-Service on Daily Sales

	Ln(qty) 1	Ln(qty) 2	(Stig/ref) 3	Log( <i>qty</i> <sub>no age limit</sub> ) 4
STIG*TS*TP	0.101*** (0.019)			
STIG:TS1*TP1		0.097*** (0.028)		0.068* (0.028)
STIG:TS2*TP2		0.105*** (0.026)		0.114*** (0.026)
TS1*TP1	-0.187*** (0.019)	-0.185*** (0.021)	1.087*** (0.090)	-0.202*** (0.021)
TS2*TP2	0.099*** (0.017)	0.098*** (0.019)	0.510*** (0.077)	0.090*** (0.019)
Num.Obs	38058	38058	19029	38061
Method	Triple	Triple	DID	Triple
SEs	Wild Bootstrap	Wild Bootstrap	Wild Bootstrap	Wild Bootstrap
Cluster	Store	Store	Store	Store
Controls	Yes	Yes	Yes	Yes
R2	0.873	0.873	0.711	0.810
R2 Adj.	0.873	0.873	0.710	0.809
Log.Lik.	-11079.119	-11079.097	-31823.436	-11235.214

+ p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001

*Note:* Columns 1-3 uses all stigma items and reference items. Column 4 uses only stigma items without an age limit. *Qty* is quantity measured as number of items sold. *STIG*, *TS*, and *TP* are dummy variables for *stigma products*, *treatment store*, *treatment period*, while 1 and 2 indicates the treatment store and the respective treatment period. Standard errors are Wild Cluster bootstrapped using Rademacher weight and 1 000 replications.

## 1.6 Conclusion

I find ample evidence suggesting that when a store converts to self-service, it sells more stigmatized items, meaning items that are private or unhealthy, relative to non-stigmatized items such as milk and laundry detergents. I attribute this increase to the removal of social friction in the form of the

**Table 1.3:** The Effect of Self-Service on Daily Sales, without uncertain categories

	Ln(qty) 1	Ln(qty) 2	(Stig/ref) 3	Log( $qty_{no\ age\ limit}$ ) 4
STIG*TS*TP	0.109*** (0.020)			
STIG:TS1*TP1		0.074** (0.028)		0.045 (0.028)
STIG:TS2*TP2		0.140*** (0.027)		0.149*** (0.027)
TS1*TP1	-0.177*** (0.019)	-0.160*** (0.022)	1.242*** (0.138)	-0.176*** (0.022)
TS2*TP2	0.077*** (0.017)	0.061** (0.019)	0.936*** (0.117)	0.054** (0.019)
Num.Obs	38058	38058	19029	38061
Method	Triple	Triple	DID	Triple
SEs	Wild Bootstrap	Wild Bootstrap	Wild Bootstrap	Wild Bootstrap
Cluster	Store	Store	Store	Store
Controls	Yes	Yes	Yes	Yes
R2	0.893	0.893	0.685	0.842
R2 Adj.	0.893	0.893	0.684	0.842
Log.Lik.	-12074.027	-12072.546	-39874.987	-12125.305

+ p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001

*Note:* Columns 1-3 uses all stigma items and reference items. Column 4 uses only stigma items without an age limit. *Qty* is quantity measured as number of items sold. *STIG*, *TS*, and *TP* are dummy variables for *stigma products*, *treatment store*, *treatment period*, while 1 and 2 indicates the treatment store and the respective treatment period. Standard errors are Wild Cluster bootstrapped using Rademacher weight and 1 000 replications.

**Table 1.4:** The Effect of Self-Service on Daily Sales, robustness checks

	1: Ln(qty)	2: Ln(qty)	3: Ln(qty)	4: Ln(qty)
STIG*TS*TP	0.101*** (0.014)	0.101*** (0.014)	0.101*** (0.025)	0.101*** (0.019)
TS1*TP1	-0.187*** (0.036)	-0.178*** (0.026)	-0.187*** (0.048)	-0.394*** (0.049)
TS2*TP2	0.099** (0.031)	0.045 (0.028)	0.099* (0.043)	-0.129*** (0.030)
R2	0.873	0.845	0.873	0.876
R2 Adj.	0.873	0.845	0.873	0.876
Log.Lik.	-11079.119	-14894.132	-11079.119	-10600.376
Num.Obs	38058	38058	38058	38058
Method	Triple	Triple	Triple	Triple
SEs	Cluster robust	Cluster robust	Cluster robust	Wild Bootstrap
Cluster	Store	Store	Store+year	Store
Controls	Yes	No	Yes	Yes
store*date trend	No	No	No	Yes

+ p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001

*Note:* Columns 1-4 estimate the same model as in Column 1 in Table 1.2. *Qty* is quantity measured as number of items sold. *STIG*, *TS*, and *TP* are dummy variables for *stigma* products, *treatment store*, and *treatment period*, while 1 and 2 indicates the treatment store and the respective treatment period. Standard errors are mixed between classical cluster robust standard errors and Wild Cluster bootstrapped using Rademacher weight and 1 000 replications, as indicated in the table. Column 4 also includes a date-store interaction which is a store specific time trend.

**Table 1.5:** Triple Difference Estimates of the Effect of Self-Service on Daily Sales for Individual Stigma Categories with Uncertain Categories

	Cigarettes	Sanitary items	Tobacco, other	Micro-waveable	Cake	Candy	Frozen pizza	Potato chips	Cookies	Intimate hygiene	Alcohol	Ice-cream	Flavored milk	Soda	Pharmaceuticals
stigma:ts1:tp1	-0.157+ (0.084)	-0.083 (0.085)	-0.080 (0.084)	0.173* (0.085)	-0.280*** (0.084)	-0.065 (0.084)	0.172* (0.084)	0.007 (0.084)	0.309*** (0.084)	0.133 (0.099)	0.022 (0.085)	0.069 (0.087)	-0.105 (0.084)	-0.035 (0.084)	-0.083 (0.086)
stigma:ts2:tp2	-0.266*** (0.078)	0.112 (0.079)	-0.088 (0.079)	0.090 (0.079)	0.241** (0.079)	0.049 (0.079)	0.148+ (0.079)	0.064 (0.079)	-0.014 (0.079)	0.132 (0.084)	0.123 (0.080)	-0.048 (0.080)	0.126 (0.079)	0.095 (0.078)	-0.188* (0.080)
R2	0.215	0.094	0.169	0.094	0.178	0.312	0.187	0.167	0.096	0.157	0.335	0.092	0.114	0.330	0.100
R2.Adj.	0.214	0.094	0.169	0.094	0.178	0.312	0.186	0.167	0.096	0.157	0.335	0.092	0.114	0.330	0.099
Num.Obs	210736	210530	210733	210637	210734	210736	210730	210731	210691	207576	210308	210630	210719	210734	210347
Method	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple
SEs	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB
Cluster	store	store	store	store	store	store	store	store	store	store	store	store	store	store	store
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

+ p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001

*Note:* The dependent variable is the log of the number of items sold of each product category. "STIG", "TS" and "TP" are dummy variables for "stigma products", "treatment store" and "treatment period", while 1 and 2 indicates the treatment store and the respective treatment period. All estimates include the full set of triple difference variables, uses the full reference category, and have a full set of control variables. Standard errors are Wild Cluster bootstrapped using Rademacher weight and 1 000 replications.

**Table 1.6:** Triple Difference Estimates of the Effect of Self-Service on Daily Sales for Individual Stigma Categories without Uncertain Categories

	Cigarettes	Sanitary items	Tobacco, other	Micro-waveable	Cake	Candy	Frozen pizza	Potato chips	Cookies	Intimate hygiene	Alcohol	Ice-cream	Flavored milk	Soda	Pharmaceuticals
sigma:ts1:tp1	-0.136+ (0.074)	-0.062 (0.075)	-0.059 (0.074)	0.192* (0.076)	-0.259*** (0.075)	-0.044 (0.074)	0.193** (0.074)	0.027 (0.075)	0.330*** (0.075)	0.152+ (0.088)	0.042 (0.076)	0.089 (0.079)	-0.084 (0.075)	-0.015 (0.074)	-0.063 (0.076)
sigma:ts2:tp2	-0.239*** (0.070)	0.139* (0.071)	-0.061 (0.070)	0.117+ (0.070)	0.268*** (0.070)	0.076 (0.070)	0.175* (0.070)	0.092 (0.071)	0.014 (0.070)	0.159* (0.075)	0.150* (0.072)	-0.021 (0.073)	0.154* (0.070)	0.122+ (0.070)	-0.164* (0.071)
R2	0.297	0.202	0.240	0.182	0.251	0.418	0.260	0.232	0.170	0.334	0.443	0.150	0.182	0.441	0.226
R2 Adj.	0.296	0.201	0.239	0.182	0.251	0.418	0.260	0.231	0.170	0.334	0.443	0.149	0.181	0.440	0.225
Num.Obs	130454	130248	130451	130355	130452	130454	130448	130449	130409	127294	130026	130348	130437	130452	130065
Method	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple	triple
SEs	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB	WCB
Cluster	store	store	store	store	store	store	store	store	store	store	store	store	store	store	store
Controls	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes

+ p < 0.1, \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001

Note: The dependent variable is the log of the number of items sold of each product category. "STIG", "TS" and "TP" are dummy variables for "stigma products", "treatment store" and "treatment period", while 1 and 2 indicates the treatment store and the respective treatment period. All estimates include the full set of triple difference variables, excludes the uncertain categories from the reference category, and have a full set of control variables. Standard errors are Wild Cluster bootstrapped using Rademacher weight and 1 000 replications.



purchasing anonymity provided by self-service checkouts. The estimated effect is about a 10 percent increase in stigma items.

There are two business models to choose from when converting to self-service. Either the store can keep some manned cash registers, or it can convert to self-service only. My findings suggest that the former may be the preferred strategy. The store that converted fully experienced a net reduction in total sales compared to the control stores. There is no reduction in sales for the store that kept a manned cash register, suggesting that converting fully scares away some customers, not self-service in itself. The two business models also have differential effects on individual categories of items. The main drivers for both technologies are the product lines *Microwaveable food, frozen pizza, intimate hygiene, potato chips, and alcohol*. There seems to be an interaction effect between whether the product has an age limit, and whether the products is ordered in the checkout process, or physically available in the store. Among the stigma items with an age limit, only alcohol sells more after the introduction of self service checkouts. This is the only product with an age limit that is physically available in the store.

## Chapter 2

# The Triple Difference Estimator

Andreas Olden      Jarle Møen \*

### Abstract

Triple difference has become a widely used estimator in empirical work. A close reading of articles in top economics journals reveals that the use of the estimator to a large extent rests on intuition. The identifying assumptions are neither formally derived nor generally agreed on. We give a complete presentation of the triple difference estimator, and show that even though the estimator can be computed as the difference between two difference-in-differences estimators, it does not require two parallel trend assumptions to have a causal interpretation. The reason is that the difference between two biased difference-in-differences estimators will be unbiased as long as the bias is the same in both estimators. This requires only one parallel trend assumption to hold. )

---

\*Affiliation of all authors: Department of Business and Management Science, NHH Norwegian School of Economics. This paper is a methodological companion paper to Olden (2018). We thank two anonymous referees for very helpful and valuable comments. We are also grateful to Erik Øiolf Sørensen and Håkon Otneim for useful discussions and comments.

## 2.1 Introduction

The triple difference estimator is widely used, either under the name “Triple difference” (TD) or the name “difference-in-difference-in-differences” (DDD), or with minor variations of these spellings. Triple difference is an extension of double differences and was introduced by Gruber (1994). Even though Gruber’s paper is well cited, very few modern users of triple difference credit him for his methodological contribution. One reason may be that the properties of the triple difference estimator are considered obvious. Another reason may be that triple difference was little more than a curiosity in the first ten years after Gruber’s paper. On Google Scholar, the annual number of references to triple difference did not pass one hundred until year 2007. Since then, the use of the estimator has grown rapidly and reached 928 unique works referencing it in the year 2017.<sup>2</sup>

Looking only at the core economics journals *American Economic Review*, *Journal of Political Economy* and *Quarterly Journal of Economics*, we have found 32 articles using triple difference between 2010 and 2017, see Table 2.1 in the appendix. A close reading of these articles reveals that the use of the triple difference estimator to a large extent rests on intuition. The identifying assumptions are neither formally derived nor generally agreed on. We fill this void in the literature and give a complete presentation of the triple difference estimator.

The triple difference estimator can be computed as the difference between

---

<sup>2</sup>More details on the historical development of the use of the triple difference estimator can be found in the working paper version, Olden and Møen (2020), Figure 1. In the working paper we also analyse naming conventions and suggest that there is a need to unify terminology. We recommend the terms ‘triple difference’ and ‘difference-in-difference-in-differences’.

two difference-in-differences estimators. Despite this, we show that the triple difference estimator does not require two parallel trend assumptions to have a causal interpretation. The intuition is that the difference between two biased difference-in-differences estimators will be unbiased as long as the bias is the same in both estimators. In that case, the bias will be differenced out when the triple difference is computed. This requires only one parallel trend assumption, in ratios, to hold. In fact, the sole purpose of subtracting the second difference-in-differences is to remove bias in the first. Gruber (1994) states the identification requirement verbally, but the result has not been fully formalized in the econometric literature, and it is overlooked in most of the recent applications.

The rest of the paper is organized as follows: Section 2 gives a short overview of the use of the triple difference estimator. Section 3 derives the triple difference estimator. Section 4 shows that the triple difference estimator can be viewed as the difference between two difference-in-differences estimators. Section 5 derives the identifying assumptions. Section 6 shows that the triple difference estimator can also be viewed as a difference-in-differences using a ratio between two outcome variables. Section 7 discusses some issues related to inference. Section 8 provides concluding remarks.

## **2.2 The triple difference literature**

The most authoritative and formal treatment of the triple difference estimator was for many years an unpublished NBER summer institute lecture note on difference-in-differences estimation by Imbens and Wooldridge (2007). In the introductory “Review of basic methodology” chapter they included a

brief exposition of the triple difference estimator.<sup>3</sup> The formula for the triple difference estimator is now available in two econometrics books by Frölich and Sperlich (2019, p. 242) and Wooldridge (2020, p. 436). We complement these recent books by providing a more detailed discussion of the estimator, and in particular by deriving the assumptions needed to identify a causal effect.<sup>4</sup>

Other authoritative sources have treated the topic only in passing. In their famous text book, *Mostly Harmless Econometrics*, Angrist and Pischke (2008, p. 242) write that "A modification of the two-by-two DD setup with possibly improved control groups uses higher-order contrast to draw causal inference". The authors then go on to explain the basic setup using Yelowitz (1995) as an example. They do not discuss or present the estimator, nor the identifying assumption. They simply conclude that "This triple-difference model may generate a more convincing set of results than a traditional DD analysis".

Lechner (2011, p. 3) follows a similar avenue in his survey *The estimation of causal effects by difference-in-differences methods*. He uses Yelowitz (1995) as an example of triple difference, and states that "the basic ideas of the approach of taking multiple differences are already apparent with two dimensions. Thus, we refrain from addressing these higher dimensions to keep the discussion as focused as possible."

A look at Yelowitz (1995) reveals that he does not go into depth on the estimator and the identifying assumptions. Instead, he cites Gruber (1994) and Gruber and Poterba (1994). Gruber and Poterba (1994), however, refer

---

<sup>3</sup>Imbens and Wooldridge (2007) start out with a setup that is identical to ours in all respects except notation (compare their Equation 1.3 to our Equation 2.1) However, the estimator presented in their Equation 1.4, contains an error as it lacks the last term in our Equation 2.4. This was corrected already in the 2008-version of the lecture notes, but unfortunately, later versions have been less widely distributed.

<sup>4</sup>We are grateful to an anonymous referee for making us aware of the two recent books.

back to Gruber (1994).

In his single-authored 1994 article, Gruber analyzes the labour market effects of mandated maternity benefits. Gruber explains the setup as follows:

I compare the treatment individuals in the experimental states to a set of control individuals in those same states and measure the change in the treatments' relative outcomes, relative to states that did not pass maternity mandates. The identifying assumption of this "differences-in-differences-in-differences" (DDD) estimator are fairly weak: it simply requires that there be no contemporaneous shock that affects the relative outcomes of the treatment group in the same state-years as the law".

We have also looked at all articles applying triple difference (using one of the six most common ways of referencing the estimator) in *American Economic Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics* between 2010 and 2017. As seen in Table 2.1 in the appendix, we found a total of 32 articles, 16 articles in AER, five in JPE and 11 in QJE. Of these articles Muehlenbachs et al. (2015), Hornbeck (2010), and Shayo and Zussman (2011) show some version of the estimator itself, indicating that it is not entirely obvious. In a similar spirit, Walker (2013) shows the error term of the triple difference estimator and uses it for discussion of robustness. Only Nilsson (2017) cites Gruber (1994).

We will later show formally that a parallel trend assumption very similar to the difference-in-differences approach is needed for the estimated effect to have a causal interpretation. The parallel trend in DDD is, however, on a differential between two categories. In some applications this is stated verbally. Walker (2013, p. 1805) writes e.g. that "[t]he identifying assumption

in this class of models is that there are no other factors generating a difference in differential trends between production decisions in regulated and unregulated manufacturing firms." <sup>5</sup>

Most of the other 32 top journal articles present some intuition of what the estimator is robust against, but otherwise the information presented varies considerably. Only a few of the authors discuss a common trend or parallel trend assumption, and as the triple difference is based on a strong parallel trend assumption, it is also disturbing to see that a large part of the articles do not include unconditional plots of the outcome series they are studying. This makes it impossible to visually assess potential trends.

In Tables 2.2 and 2.3 in the appendix, we present the 50 most cited articles referencing the estimator, numbered and ordered by number of citations. There has been almost 5000 papers referencing the estimator since 1994, and it is natural to think that some of the most cited triple difference articles are methodological or represent early use of the methodology. Seven of the 50 most cited articles list Gruber as a co-author.<sup>6</sup> Six articles are covered in the review of articles in AER/QJE/JPE.<sup>7</sup> Among the rest, seven have methodological-sounding names.<sup>8</sup> A close reading of the articles with methodological-sounding names reveals that they do not give a formal ex-

---

<sup>5</sup>Some other articles in our sample have similar formulations. Hoynes et al. (2016, p. 925-926) write that "[i]n this triple-difference model, the maintained assumption is that there are no differential trends for high participation versus low participation groups within early versus late implementing counties". Deschênes et al. (2017, p. 2970) state that "[o]ur identifying assumption is that such policies did not change differentially in NBP versus non-NBP states, in winter versus summer, over this period". Finally, Kleven et al. (2013, p. 1908) write that "[i]n that case, the identifying assumption would be that there is no contemporaneous change in the differential trend between Spain and the synthetic control country".

<sup>6</sup>These are the articles 4, 9, 17, 25, 31, 34, and 39, in which 4 is Gruber (1994) and 31 is Gruber and Poterba (1994). Note also that number 30 is Yelowitz (1995).

<sup>7</sup>These are the articles 7, 11, 21, 35, 42 and 46.

<sup>8</sup>These are the articles 1, 5, 6, 10, 12, 24, and 40. Note that number 24 is Lechner (2011) which is covered previously.

position of the triple difference estimator, nor its identifying assumption. However, Ravallion (2007) cites Ravallion et al. (2005) which shows a very special case of the triple difference estimator and the identifying assumptions for that special case.<sup>9</sup>

## 2.3 The triple difference estimator

For the sake of exposition let us assume that we are talking about two American states, and that the Treatment state (T) introduces a health-care measure, while the Control state (C) does not. Further, the population of the states can be subdivided into two groups, group A and group B. The health-care measure we intend to study is only introduced to group B, i.e. group B is the group that can Benefit from the measure. Finally, there are two time periods, namely Pre- and Post-implementation of the health-care measure.

To establish a counterfactual it might seem convenient to compare group A and group B within the treatment state. This will not be valid if the health-care reform has within-state spillovers from group B to group A. Another option is to compare group B in the treatment state with group B in the control state. This will not be valid if different states have different economic conditions, so that group B in the treatment state would have trended differently from group B in the control state, regardless of the health-care measure. However, we may reasonably assume that the general economic differences will not affect the relative outcomes of group A and group B. In that case, we can use the relative difference to estimate what would have happened to

---

<sup>9</sup>This scenario does not have pre-periods, only post-periods, and two treatment groups that are treated with differential intensity. This requires a set of identifying assumptions that in general are not needed in the triple difference estimator.



the relative outcomes of group A and group B in the treatment state in the absence of treatment.

Equation 2.1 is a basic triple difference specification in accordance with the above exposition. All variables in this basic setup are dummy variables.

$$Y_{sit} = \beta_0 + \beta_1 T + \beta_2 B + \beta_3 Post + \beta_4 T * B + \beta_5 T * Post + \beta_6 B * Post + \beta_7 T * B * Post + \epsilon_{sit} \quad (2.1)$$

The conditional mean function of Equation 2.1 is  $E[Y_{sit}|T, B, Post]$ , which can take on eight values. Since the model has eight values and eight coefficients, the model is saturated (Angrist and Pischke 2008). Under standard OLS assumptions and an additive effect, we can use  $E[\epsilon_{sit}|T, B, Post] = 0$  to show the eight expected values as in Equations 2.2.

$$\begin{aligned} E[Y|T = 0, B = 0, Post = 0] &= \beta_0 \\ E[Y|T = 1, B = 0, Post = 0] &= \beta_0 + \beta_1 \\ E[Y|T = 0, B = 1, Post = 0] &= \beta_0 + \beta_2 \\ E[Y|T = 0, B = 0, Post = 1] &= \beta_0 + \beta_3 \\ E[Y|T = 1, B = 1, Post = 0] &= \beta_0 + \beta_1 + \beta_2 + \beta_4 \\ E[Y|T = 1, B = 0, Post = 1] &= \beta_0 + \beta_1 + \beta_3 + \beta_5 \\ E[Y|T = 0, B = 1, Post = 1] &= \beta_0 + \beta_2 + \beta_3 + \beta_6 \\ E[Y|T = 1, B = 1, Post = 1] &= \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7 \end{aligned} \quad (2.2)$$

Starting at the top of equation set 2.2, we can solve for the  $\beta'$ s.

$$\begin{aligned}
\beta_0 &= E[Y|T = 0, B = 0, Post = 0] \\
\beta_1 &= E[Y|T = 1, B = 0, Post = 0] - E[Y|T = 0, B = 0, Post = 0] \\
\beta_2 &= E[Y|T = 0, B = 1, Post = 0] - E[Y|T = 0, B = 0, Post = 0] \\
\beta_3 &= E[Y|T = 0, B = 0, Post = 1] - E[Y|T = 0, B = 0, Post = 0] \\
\beta_4 &= E[Y|T = 1, B = 1, Post = 0] + E[Y|T = 0, B = 0, Post = 0] - \\
&\quad E[Y|T = 1, B = 0, Post = 0] - E[Y|T = 0, B = 1, Post = 0] \\
\beta_5 &= E[Y|T = 1, B = 0, Post = 1] + E[Y|T = 0, B = 0, Post = 0] - \\
&\quad E[Y|T = 1, B = 0, Post = 0] - E[Y|T = 0, B = 0, Post = 1] \\
\beta_6 &= E[Y|T = 0, B = 1, Post = 1] + E[Y|T = 0, B = 0, Post = 0] - \\
&\quad E[Y|T = 0, B = 1, Post = 0] - E[Y|T = 0, B = 0, Post = 1] \\
\beta_7 &= (E[Y|T = 1, B = 1, Post = 1] - E[Y|T = 1, B = 1, Post = 0]) - \\
&\quad (E[Y|T = 1, B = 0, Post = 1] - E[Y|T = 1, B = 0, Post = 0]) - \\
&\quad (E[Y|T = 0, B = 1, Post = 1] - E[Y|T = 0, B = 1, Post = 0]) + \\
&\quad (E[Y|T = 0, B = 0, Post = 1] - E[Y|T = 0, B = 0, Post = 0]) \quad (2.3)
\end{aligned}$$

By rearranging the expression for  $\beta_7$  and substituting the expected values with their sample equivalents (the mean values), we get Equation 2.4. This is the triple difference estimator for the effect of the treatment for group B.

$$\begin{aligned}
\hat{\beta}_7 = & \\
& [(\bar{Y}_{T=1,B=1,Post=1} - \bar{Y}_{T=1,B=1,Post=0}) - (\bar{Y}_{T=0,B=1,Post=1} - \bar{Y}_{T=0,B=1,Post=0})] - \\
& [(\bar{Y}_{T=1,B=0,Post=1} - \bar{Y}_{T=1,B=0,Post=0}) - (\bar{Y}_{T=0,B=0,Post=1} - \bar{Y}_{T=0,B=0,Post=0})]
\end{aligned}
\tag{2.4}$$

For simplicity, we have not included control variables in the equations above. Adding control variables is common and simple when using the regression formulation of the triple difference model in Equation 2.1. The benefits are twofold. First, control variables with substantial explanatory power will reduce the residual variance, and thereby increase the precision of the causal effect of interest. Second, including control variables can account for compositional differences between groups and make the parallel trend assumption needed for identification more credible. Put differently, including control variables can mitigate selection problems if there is some selection into the treatment state and group that is based on observable characteristics. We derive the identifying assumption for the case without control variables in Section 2.5 below.

## 2.4 The difference between two difference-in-differences

The classical difference-in-differences estimator is presented in Equation 2.5.

$$\hat{\delta} = [(\bar{Y}_{T=1,Post=1} - \bar{Y}_{T=1,Post=0}) - (\bar{Y}_{T=0,Post=1} - \bar{Y}_{T=0,Post=0})] \quad (2.5)$$

Clearly, the triple difference estimator of Equation 2.4 is equivalent to the difference between two difference-in-differences. The first difference-in-differences is for group B, and is given by the first square brackets, while the second difference-in-differences is for group A, given by the second square brackets. It is also worth mentioning that due to the additive nature of the triple difference estimator of Equation 2.4, we could alternatively have presented it as a difference-in-differences for the treatment state, comparing the eligible group B and group A, minus a difference-in-differences in the control state, comparing group B and group A there. Mathematically this is equivalent, though when thinking about a specific application, one is often preferred over the other.

## 2.5 Identifying assumptions

The triple difference estimator requires a parallel trend assumption for the estimated effect to have a causal interpretation. Even though the triple difference is the difference between two difference-in-differences, it does not need two parallel trend assumptions. Rather, it requires the relative outcome of group B and group A in the treatment state to trend in the same way as the relative outcome of group B and group A in the control state, in the absence of treatment.<sup>10</sup> To see this, first take the  $\beta_7$  in Equations 2.3 and rearrange it

---

<sup>10</sup>We phrase the discussion here in terms of trends, but, as mentioned in the introduction, one can also think of triple difference as a way to remove a potential bias in an ordinary difference estimator. This requires that the two DD-estimators used have the same bias. In

to create Equation 2.6.

$$\beta_7 = \left[ \left( E[Y|T = 1, B = 1, Post = 1] - E[Y|T = 1, B = 1, Post = 0] \right) - \left( E[Y|T = 1, B = 0, Post = 1] - E[Y|T = 1, B = 0, Post = 0] \right) \right] - \left[ \left( E[Y|T = 0, B = 1, Post = 1] - E[Y|T = 0, B = 1, Post = 0] \right) - \left( E[Y|T = 0, B = 0, Post = 1] - E[Y|T = 0, B = 0, Post = 0] \right) \right] \quad (2.6)$$

Now, introduce the potential outcomes framework (see for instance Angrist and Pischke (2008)). In this framework  $E[Y_{1,sit}]$  is the expected outcome of a state, group, and time if treated, while  $E[Y_{0,sit}]$  is the expected outcome of a state, group, and time if not treated. Potential outcomes mean that we either observe  $\bar{Y}_{1,sit}$  or  $\bar{Y}_{0,sit}$ , but never both. Expressions like  $E[Y_{0,T=1,B=1,Post=1}]$  are the expectation of non-observed potential outcomes; in our case the outcome of group B in the treatment state (T), in the treatment period (Post), had it not been treated.

We can use the potential outcome framework to define  $\delta$ , the true causal effect of treatment in the treatment state (T), on the treatment group B, in the treatment period (Post) as:

$$\delta = E[Y_1 - Y_0|T = 1, B = 1, Post = 1] \quad (2.7)$$

---

fact, even the ordinary difference-in-differences estimator can in general terms be thought of as a way to remove bias rather than time trends. The parallel trend assumption is therefore sometimes referred to as a ‘bias stability’ assumption, see e.g. Frölich and Sperlich (2019, p. 230).

Equation 2.7 states that the true treatment effect is the difference between the outcome of state T, group B in period 2 as treated, and the outcome of state T, group B in period 2, had it not been treated.

Note that Equation 2.7 is the *average treatment effect on the treated*, often called ATET, ATT or TOT, see e.g. Angrist and Pischke (2008, ch. 3). Under the parallel trend assumption this is what is identified. This can be seen from the conditioning on  $T = 1$  in the definition of the true causal effect,  $\delta$ . With heterogeneous treatment effects, the population wide, unconditional, average treatment effect (ATE) is not identified. In the DD case, this has previously been pointed out by Frölich and Sperlich (2019, p. 228). They explain this by the fact that treatment effect estimation using the difference-in-differences estimator is a prediction problem where outcomes observed before the treatment started are used to predict the potential non-treatment outcome. With heterogeneous treatment effects, however, the natural experiments used for difference-in-differences estimation do not necessarily contain any information to predict the potential treatment outcome for the control group. This reasoning also applies to triple difference estimation where there are three non-predictable, counterfactual, treatment outcomes,  $E[Y_{1,T=0,B=1,Post=1}]$ ,  $E[Y_{1,T=0,B=0,Post=1}]$  and  $E[Y_{1,T=1,B=0,Post=1}]$ .

We are now ready to derive the parallel trend assumption that identifies  $\delta$ . Doing so, we rewrite Equation 2.6 using the notation from the potential outcome framework.

$$\begin{aligned}
\beta_7 = & \left[ \left( E[Y_1|T = 1, B = 1, Post = 1] - E[Y_0|T = 1, B = 1, Post = 0] \right) - \right. \\
& \left. \left( E[Y_0|T = 1, B = 0, Post = 1] - E[Y_0|T = 1, B = 0, Post = 0] \right) \right] - \\
& \left[ \left( E[Y_0|T = 0, B = 1, Post = 1] - E[Y_0|T = 0, B = 1, Post = 0] \right) - \right. \\
& \left. \left( E[Y_0|T = 0, B = 0, Post = 1] - E[Y_0|T = 0, B = 0, Post = 0] \right) \right] \quad (2.8)
\end{aligned}$$

For  $\beta_7$  to equal  $\delta$ , we need the differential in the outcomes of group A and group B in the treatment state to trend similarly to the differential in the outcomes of group A and group B in the control state, in the absence of treatment. This is the parallel trend assumption. A formal exposition of this statement is given in Equation 2.9. The first line is the change between the two periods in the outcomes of group B in the treatment state had it not been treated. The second line is the same change for group A. The difference between these two expressions is equated with an expression that is equivalent, except that it gives realized outcomes in the control state.<sup>11</sup>

---

<sup>11</sup>See Frölich and Sperlrich (2019, p. 244) for a different formulation given in the context of DDD used on a three period, two group set-up. The DD parallel trend assumption then translates into what they call a ‘parallel growth’ or ‘common acceleration’ assumption.

$$\begin{aligned}
& \left( E[Y_0|T = 1, B = 1, Post = 1] - E[Y_0|T = 1, B = 1, Post = 0] \right) - \\
& \left( E[Y_0|T = 1, B = 0, Post = 1] - E[Y_0|T = 1, B = 0, Post = 0] \right) \\
& = \\
& \left( E[Y_0|T = 0, B = 1, Post = 1] - E[Y_0|T = 0, B = 1, Post = 0] \right) - \\
& \left( E[Y_0|T = 0, B = 0, Post = 1] - E[Y_0|T = 0, B = 0, Post = 0] \right) \quad (2.9)
\end{aligned}$$

To show that this parallel trend assumption identifies  $\delta$ , the causal effect, we can substitute Equation 2.9 into Equation 2.8.

$$\begin{aligned}
\beta_7 = & \left[ \left( E[Y_1|T = 1, B = 1, Post = 1] - E[Y_0|T = 1, B = 1, Post = 0] \right) - \right. \\
& \left. \left( E[Y_0|T = 1, B = 0, Post = 1] - E[Y_0|T = 1, B = 0, Post = 0] \right) \right] - \\
& \left[ \left( E[Y_0|T = 1, B = 1, Post = 1] - E[Y_0|T = 1, B = 1, Post = 0] \right) - \right. \\
& \left. \left( E[Y_0|T = 1, B = 0, Post = 1] - E[Y_0|T = 1, B = 0, Post = 0] \right) \right] \\
& \quad (2.10)
\end{aligned}$$

Rearranging and rewriting Equation 2.10 we get



$$\begin{aligned}
\beta_7 = & E[Y_1 - Y_0|T = 1, B = 1, Post = 1] \\
& + E[Y_0|T = 1, B = 0, Post = 1] - E[Y_0|T = 1, B = 0, Post = 1] \\
& + E[Y_0|T = 1, B = 1, Post = 0] - E[Y_0|T = 1, B = 1, Post = 0] \\
& + E[Y_0|T = 1, B = 0, Post = 0] - E[Y_0|T = 1, B = 0, Post = 0] \quad (2.11)
\end{aligned}$$

By canceling out the redundant terms of Equation 2.11 we find that

$$\beta_7 = (E[Y_1 - Y_0|T = 1, B = 1, Post = 1] = \delta \quad \text{qed.} \quad (2.12)$$

## 2.6 Triple difference as difference-in-differences

Take the difference-in-differences estimator of Equation 2.5 and define the outcome variable,  $\bar{Y}$ , as:

$$\bar{Y}_{ij} = \bar{Y}_{aij} - \bar{Y}_{bij} \quad (2.13)$$

Substituting this definition into Equation 2.5 gives us

$$\begin{aligned}
\hat{\delta} &= \\
& [(\bar{Y}_{a,pre,treat} - \bar{Y}_{b,pre,treat}) - (\bar{Y}_{a,post,treat} - \bar{Y}_{b,post,treat})] - \\
& [(\bar{Y}_{a,pre,cont} - \bar{Y}_{b,pre,cont}) - (\bar{Y}_{a,post,cont} - \bar{Y}_{b,post,cont})] \\
& = \hat{\delta}_{triple} \tag{2.14}
\end{aligned}$$

This shows clearly that a basic difference-in-differences with a differential as the outcome and a symmetric structure, is a triple difference, and the other way around. This implies that all procedures for difference-in-differences can be applied to a transformed triple difference. For instance, standard robustness checks for difference-in-differences can be applied, see for instance Angrist and Pischke (2008). Also, semi-parametric versions of the difference-in-differences estimator are available, as in (Abadie, 2005), as well as non-linear models as in (Athey and Imbens, 2006), which can be directly applied to the transformed problem. Among the generalization of the simple difference-in-differences estimator, Callaway and Sant'Anna (2020) provides an appropriate estimators for caseses with many time periods, including when the parallel trend assumption holds only conditional on covariates. They also give an up to date literature review.

Finally, knowing that difference-in-differences models struggle with standard errors when there are few clusters, as documented by Bertrand et al. (2004), this will apply to the transformed triple difference, as well as to the triple difference estimator. We return to this next.

## 2.7 Inference

In the case of the difference-in-differences estimator, Bertrand et al. (2004) shows how the estimator is prone to over-rejection, i.e. finding false positives. This is due to serial correlation and intra-group correlation. This can be addressed by using cluster robust standard errors, which are based on asymptotic properties in the group dimension. However, it is common to have a limited number groups or treatment groups, violating the assumptions. For a fairly recent and extensive exposition of the issues in the difference-in-differences estimator, see Cameron and Miller (2015).

It is unclear to what extent this generalizes to the triple difference estimator, as we include additional groups, correlation structures, and explicitly try to model them. Also, we increase the number of observations and the complexity of the model. To answer these questions, we turn to a procedure from Bertrand et al. (2004), running a simulation study on data from the Current Population Survey (CPS) in which we vary the number of treated clusters.<sup>12</sup> We compare the difference-in-differences estimator with the triple difference estimator, both for individual level data and for state-year-gender aggregated data. Further, we include the triple difference estimated as difference-in-differences on a ratio, cf. Section 2.6. The full results are presented in Appendix B.

When it comes to false positives, or over-rejection, we find that the difference-in-differences and the triple difference show similar patterns of over-rejection with clustered standard errors. However, the triple difference shows greater

---

<sup>12</sup>We draw  $n$  placebo treatment states out of 51, draw a year from a uniform distribution over 1985-1995 which serves as a treatment year, estimate different models, and reiterate the process 10 000 times, considering rejection rates, i.e. how often we find a significant effect.

power to detect true (simulated) effects. Aggregation does not solve the issues of over-rejection, and comes at a cost with respect to power. Further, the triple difference as difference-in-differences and the full triple difference performs almost identical. Researchers should know that there is little to lose, and some to gain, by using the triple difference relative to difference-in-differences, but also realize that when there are few clusters, or few treated clusters, both will have severe issues of over-rejection.

## 2.8 Concluding remarks

In this paper we document the rise of the triple difference estimator. The use of the estimator has grown exponentially, yet it lacks formal derivation and is often carelessly applied in the literature, for instance by largely ignoring its parallel trend assumption, and by omitting unconditional plots, making model validation difficult.

Our main contribution is to show that the triple difference estimator does not require two parallel trend assumptions to have a causal interpretation, even though it can be computed as the difference between two difference-in-differences estimators. We also show that the triple difference parallel trend assumption is equivalent to the parallel trend assumption in a difference-in-differences model based on ratios.

When choosing between a triple difference and a difference-in-differences on a ratio-variable, there are several things to consider. The difference-in-differences estimator is much better understood, and there is a large literature that addresses the estimator and its shortcomings. However, it comes at the cost of degrees of freedom, and provides less information than the triple

difference. The triple difference will for instance provide an estimate of spillover-effects, i.e.  $\beta_5$  in Equation 2.1, which is the effect on the non-treated in the treatment state in the treatment period. This information is lost in the difference-in-differences estimator.

The triple difference estimator is often used as a heterogeneity test or as a robustness check. When comparing it with a standard difference-in-differences, Berck and Villas-Boas (2016) show conditions for when the triple difference estimator reduces bias relative to a difference-in-differences approach in the presence of omitted variable bias.

Finally, our reading of the literature points to some other key issues that demand awareness. Many of the articles spend considerable time on control variables, in which case one should be specific on whether the inclusion is to absorb variance and increase precision, or if the parallel trend assumption holds only conditional on some covariates. Note that in the case of time-invariant state-level variables, they will be differenced out, easily shown by deducting any mean from the estimator. Time-varying, state level variables, however, is a likely source of bias, and should be explicitly dealt with when evaluating the parallel trend assumption, or be dealt with in a more complex framework, as touched upon in Section 2.6.

In the literature, much less time is spent discussing functional form issues than control variables. This is unfortunate. Both the difference-in-differences and the triple difference estimator relies on a parallel trend assumption, and hence the functional form is identifying. In the triple difference estimator, we make an assumption on how the outcomes of two groups co-move relative to the co-movement in two other groups in the control state. Both a ratio and its log-transformed counterpart can be a natural choice of functional

form, depending on the situation. This requires thought, however. When the parallel trend assumption holds in logs it will not hold in levels, and vice versa, see Angrist and Pischke (2008, p. 230) and Frölich and Sperlich (2019, p. 228).<sup>13</sup>

---

<sup>13</sup>Unfortunately, what functional form to choose, seldom finds its answer in economic theory or statistics. For a discussion of these topics in the case of the difference-in-differences estimator, we recommend Kahn-Lang and Lang (2020). The recommendations of Kahn-Lang and Lang (2020) is equally applicable to the triple difference estimator, and includes addressing why there is level differences to begin with, explicitly justifying the parallel trend assumption, and noting that pre-treatment trends is indicative, but not necessary, nor sufficient for the parallel trend assumption to hold. However, in the case of the triple difference, initial level differences in the difference-in-differences might be a reason why we want to use triple difference. The general advice to reflect on level differences still stands.

## 2.A Tables

## APPENDIX A: TABLES

**Table 2.1: Use of triple difference estimation in *AER*, *JPE* and *QJE* from 2010-2017**

Cites	Authors	Title	Year	Source
829	Mian, Sufi	House prices, home equity-based borrowing, and the US household leverage crisis	2011	AER
103	Moser, Voena	Compulsory licensing: Evidence from the trading with the enemy act	2012	AER
293	Hornbeck	The enduring impact of the American Dust Bowl: Short-and long-run adjustments to environmental catastrophe	2012	AER
146	Simcoe	Standard setting committees: Consensus governance for shared technology platforms	2012	AER
243	Kleven, Landais, Saez	Taxation and international migration of superstars: Evidence from the European football market	2013	AER
320	Busso, Gregory, Kline	Assessing the incidence and efficiency of a prominent place based policy	2013	AER
57	Aaronson, Lange, Mazumder	Fertility transitions along the extensive and intensive margins	2014	AER
129	Yagan	Capital tax reform and the real economy: The effects of the 2003 dividend tax cut	2015	AER
90	Casey	Crossing party lines: The effects of information on redistributive politics	2015	AER
212	Muehlenbachs, Spiller, Timmins	The housing market impacts of shale gas development	2015	AER
291	Hoynes, Schanzenbach, Almond	Long-run impacts of childhood access to the safety net	2016	AER
440	Pierce, Schott	The surprisingly swift decline of US manufacturing employment	2016	AER
37	Duggan, Garthwaite, Goyal	The market impacts of pharmaceutical product patents in developing countries: Evidence from India	2016	AER
65	Egan, Hortáçsu, Matvos	Deposit competition and financial fragility: Evidence from the us banking sector	2017	AER
30	Deschênes, Greenstone, Shapiro	Defensive investments and the demand for air quality: Evidence from the NOx budget program	2017	AER
122	Besley, Folke, Persson, Rickne	Gender quotas and the crisis of the mediocre man: Theory and evidence from Sweden	2017	AER
79	Aaronson, Mazumder	The impact of Rosenwald schools on black achievement	2011	JPE
50	Autor, Palmer, Pathak	Housing market spillovers: Evidence from the end of rent control in Cambridge, Massachusetts	2014	JPE
163	Carneiro, Løken, Salvanes	A flying start? Maternity leave benefits and long-run outcomes of children	2015	JPE
37	Casas-Arce, Saiz	Women and power: Unpopular, unwilling, or held back?	2015	JPE
47	Nilsson	Alcohol availability, prenatal conditions, and long-term economic outcomes	2017	JPE
143	Hornbeck	Barbed wire: Property rights and agricultural development	2010	QJE
179	Shayo, Zussman	Judicial ingroup bias in the shadow of terrorism	2011	QJE
772	Ahern, Dittmar	The changing of the boards: The impact on firm valuation of mandated female board representation	2012	QJE
73	Cascio, Washington	Valuing the vote: The redistribution of voting rights and state funds following the voting rights act of 1965	2013	QJE
150	Walker	The transitional costs of sectoral reallocation: Evidence from the clean air act and the workforce	2013	QJE
155	Garthwaite, Gross, Notowidigdo	Public health insurance, labor supply, and employment lock	2014	QJE
52	Casaburi, Troiano	Ghost-house busters: The electoral response to a large anti-tax evasion program	2015	QJE
16	Agan, Starr	Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment	2017	QJE
25	Alsan, Wanamaker	Tuskegee and the health of black men	2017	QJE
44	Bandiera, Burgess, Das, Gulesci, Rasul, Sulaiman	Labor markets and poverty in village economies	2017	QJE
20	Larcom, Rauch, Willems	The benefits of forced experimentation: striking evidence from the London underground network	2017	QJE



**Table 2.2: Top 50 most cited articles referencing triple difference**

	Cites	Authors	Title	Year	Source
1	7550	Bertrand, Duflo, Mullainathan	How much should we trust differences-in-differences estimates?	2004	QJE
2	1418	Verhoogen	Trade, quality upgrading, and wage inequality in the Mexican manufacturing sector	2008	QJE
3	1306	Currie, Almond	Human capital development before age five	2011	HLE
4	1177	Gruber	The incidence of mandated maternity benefits	1994	AER
5	989	Roberts, Whited	Endogeneity in empirical corporate finance	2013	HEF
6	943	Winship, Morgan	The estimation of causal effects from observational data	1999	ARS
7	824	Mian, Sufi	House prices, home equity-based borrowing, and the US household leverage crisis	2011	AER
8	809	Ruhm	The economic consequences of parental leave mandates: Lessons from Europe	1998	QJE
9	807	Currie, Gruber	Health insurance eligibility, utilization of medical care, and child health	1996	QJE
10	774	Ravallion	Evaluating anti-poverty programs	2007	HDE
11	763	Ahem, Dittmar	The changing of the boards:	2012	QJE
12	697	Besley, Case	The impact on firm valuation of mandated female board representation	2000	TEJ
13	690	Giroud, Mueller	Unnatural experiments? Estimating the incidence of endogenous policies	2010	JFE
14	659	Zervas, Proserpio, Byers	Does corporate governance matter in competitive industries? The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry	2017	JMR
15	648	Dynarski	Hope for whom?	2000	NTJ
16	552	Costa, Kahn	Financial aid for the middle class and its impact on college attendance	2000	QJE
17	526	Gruber	Power couples: Changes in the locational choice of the college educated, 1940–1990	1997	JLE
18	512	Purnanandam	The incidence of payroll taxation: Evidence from Chile	2010	RFS
19	505	Low	Originate-to-distribute model and the subprime mortgage crisis	2009	JFE
20	500	Puri, Rocholl, Steffen	Managerial risk-taking behavior and equity-based compensation	2011	JFE
21	436	Pierce, Schott	Global retail lending in the aftermath of the US financial crisis: Distinguishing between supply and demand effects	2016	AER
22	388	Katz	The surprisingly swift decline of US manufacturing employment	1996	NBER
23	387	Sommers, Baicker, Epstein	Wage subsidies for the disadvantaged	2012	NEJM
24	384	Lechner	Mortality and access to care among adults after state Medicaid expansions	2011	FTE
25	377	Gruber	The estimation of causal effects by difference-in-difference methods	2000	JPE
26	359	Goldfarb, Tucker	Disability insurance benefits and labor supply Privacy regulation and online advertising	2011	MS

**Table 2.3: Top 50 most cited articles referencing triple difference, continued**

Cites	Authors	Title	Year	Source
27	Strauss, Thomas	Health over the life course	2007	HDE
28	Matsa, Miller	A female style in corporate leadership? Evidence from quotas	2013	AEJAE
29	Seru	Firm boundaries matter: Evidence from conglomerates and R&D activity	2014	JFE
30	Yelowitz	The Medicaid notch, labor supply, and welfare participation: Evidence from eligibility expansions	1995	QJE
31	Gruber, Poterba	Tax incentives and the decision to purchase health insurance: Evidence from the self-employed	1994	QJE
32	Milligan	Subsidizing the stork: New evidence on tax incentives and fertility	2005	RES
33	Currie	Inequality at birth: Some causes and consequences	2011	AER
34	Gruber, Madrian	Health insurance, labor supply, and job mobility: A critical review of the literature	2002	NBER
35	Busso, Gregory, Kline	Assessing the incidence and efficiency of a prominent place based policy	2013	AER
36	Eggleson, Ling, Qingyue, Lindelow, Wagstaff	Health service delivery in China: A literature review	2008	HE
37	314	The effects of Wal-Mart on local labor markets	2008	JUE
38	311	Testing, crime and punishment	2006	JPuE
39	309	Health insurance and the labor market	2000	HHE
40	296	Causal inference with observational data	2007	SJ
41	291	Do private transfers 'displace' the benefits of public transfers? Evidence from South Africa	2004	JPuE
42	290	The enduring impact of the American Dust Bowl: Short-and long-run adjustments to environmental catastrophe	2012	AER
43	287	Education in a Crisis	2004	JDE
44	286	The effect of customers' social media participation on customer visit frequency and profitability: An empirical investigation	2013	ISR
45	282	Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina	2008	JPuE
46	281	Long-run impacts of childhood access to the safety net	2016	AER
47	277	Payday lenders: Heroes or villains?	2011	JFE
48	273	Observational learning: Evidence from a randomized natural field experiment	2009	AER
49	273	Labor laws and innovation	2013	JLaE
50	267	The effects of a natural gas boom on employment and income in Colorado, Texas, and Wyoming	2012	EE

This table is produced using the software Harzinger's Publish or Perish 6. A search using each of the six most common ways to reference the triple difference estimator is conducted from 1994 until October 2018, covering almost all results for the triple difference estimator. Each search is combined with the word economics. When removing books and duplicates, this yields 3481 articles. The articles are sorted according to the number of citations, and the top 50 most cited articles are presented here. Full journal titles are found in Table 2.4.

**Table 2.4:** Title abbreviations for Tables 2.1-2.3

Abbreviaton	Full title
AEJAE	American Economic Journal: Applied Economics
AER	The American Economic Review
ARS	Annual Review of Sociology
EE	Energy Economics
FTE	Foundations and Trends® in Econometrics
HDE	Handbook of Development Economics
HE	Health Economics
HEF	Handbook of the Economics of Finance
HHE	Handbook of HE
HLE	Handbook of Labor Economics
ISR	Information Systems Research
JDE	Journal of Development Economics
JFE	Journal of Financial Economics
JLaE	Journal of Law and Economics
JLE	Journal of Labor Economics
JMR	Journal of Marketing Research
JPE	Journal of Political Economy
JPuE	Journal of Public Economics
JUE	Journal of Urban Economics
MS	Management Science
NBER	NBER Working Paper Series
NEJM	New England Journal of Medicine
NTJ	National Tax Journal
QJE	Quarterly Journal of Economics
RES	Review of Economics and Statistics
RFS	The Review of Financial Studies
SJ	Stata Journal
TEJ	The Economic Journal

## 2.B Simulations

The difference-in-differences and the triple difference estimators often have a group and a time structure, for instance individual level data in different US states over time, with some states being treated. This structure introduces issues of serial correlation and intra-group (cluster) correlation, which can

lead to biased standard errors and severe over-rejection of the null hypothesis of no effect, famously documented in the difference-in-differences case by Bertrand et al. (2004). Typically, cluster robust standard errors on the state level are used. These relies on asymptotic properties in the number of groups.<sup>14</sup> For a thorough overview on the issues and remedies, see Cameron and Miller (2015) and Angrist and Pischke (2008). It can be shown that the asymptotic properties applies to both the number of untreated and the number of treated clusters, see (Conley and Taber, 2011).

While the issue of clustered errors is well studied in the difference-in-differences estimator, it is not obvious to what extent it carries over to the triple difference estimator. One reason to expect differences is that we introduce new contrast groups that might affect correlation structures, and we explicitly try to model sub-groups within the cluster (for instance gender in state). Moreover, the number of observations typically doubles, and we increase the general complexity of the modeling approach. We will not give a full exposition of these issues in the triple difference estimator case, but we will make some points by comparing triple difference to difference-in-differences.

To aid intuition, consider the following stylized example. Some US states introduce a legal reform to affect the wage of women. Having data on wage from both before and after the reform for all states, it seems well-suited for a difference-in-differences approach. In this example, the states that introduce the legal reform are the treatment states, while the states that do not are the control states. The time period before the reform is the pre-period, while the time period after the reform is the treatment-period. However, we might be worried that the states that introduced the legislation to impact the wage

---

<sup>14</sup>As developed by White (1984) with extensions by Liang and Zeger (1986) and Arellano et al. (1987).

of women would have had higher growth rates to begin with, such that the comparison of women in the treatment state and women in the control states would be biased. While the treatment states might trend differentially from the control states regardless of treatment, we believe that this trend affects men and women similarly. Thus we consider a triple difference estimator for which we compare the relative wage of women and men in the treatment states to the relative wage of women and men in the control states, circumventing the bias from the difference-in-differences estimator.

If the assumptions hold and the right functional form is chosen, this strategy will get rid of the bias in the estimation. However, we are left with the question of standard errors. To shine some light on the issue, we use the procedure of Bertrand et al. (2004) and run simulations of placebo treatments while observing the rejection rates. The data used is the Current Population Survey in their fourth interview month, in the Merged Outgoing Rotation Group, from 1979 to 1999. The survey contains individual level data from all US states and Washington DC. The data are freely available and commonly used.<sup>15</sup> The rejection rate is defined as the proportion of times the null hypothesis is rejected on a five percent significance level, i.e. the number of times we find a significant effect for the treatment variable. When there is no effect, this should be 5 percent of the times, i.e. the significance level or probability of false positives. Note that we are “randomizing the treatment variable while keeping the set of outcomes fixed. In general, the distribution of the test statistic induced by such randomization is not a standard normal distribution and, therefore, the exact rejection rate we should expect is not known.” (Bertrand et al., 2004, p. 256). However, real data has its own

---

<sup>15</sup>Data accessed November 19 2020 from:  
<https://www.nber.org/research/data/current-population-survey-cps-data-nber>.  
Data and reproducible code is provided openly at [https://github.com/andreasolden/simulate\\_triple\\_difference](https://github.com/andreasolden/simulate_triple_difference).

advantages, and we also have the original article as a baseline. Furthermore, our comparisons are mainly to explore the relative performance of triple difference as compared to difference-in-differences, making the true rejection rate less important.

Keeping our example as close to Bertrand et al. (2004) as possible, we restrict the sample to participants between the ages of 25 and 50, with strictly positive earnings. This leaves about 1 000 000 observations. Bertrand et al. (2004) consider a difference-in-differences on women. We include men also, as they serve as a control group when adding the additional layer of the triple difference.<sup>16</sup> The procedure goes as follows:

1. Draw  $n$  states randomly. These will serve as the placebo treatment states.
2. Draw a year from a uniform distribution over 1985-1995. This year and all subsequent years will serve as the placebo treatment years.
3. Estimate different models with different standard errors.
4. Re-iterate steps 1-3 10 000 times.
5. Consider rejection rates, i.e. how often we find a significant effect.

We run five different regression models. Equation 2.15 is a difference-in-differences on females, as in Bertrand et al. (2004). Equation 2.16 is a triple difference for both sexes. Both are estimated on individual level data. Equation 2.17 is a difference-in-differences for females on data aggregated to the

---

<sup>16</sup>We deviate from Bertrand et al. (2004) by running 10 000 iterations as opposed to 200-400. We do not include individual level controls for better comparisons between the simulated models, but we always include state and year fixed effects, as well as a fixed effect of gender when applicable. Since the identification comes from group differences over time, this is unlikely to be important for our purposes.

state-year-gender level. Equation 2.18 is a triple difference performed as difference-in-differences on relative outcomes for both sexes, as shown in Section 2.6. Equation 2.19 is a full triple difference for both sexes. The latter two equations are also on data aggregated to the state-year-gender level. The motivation for the aggregation is that triple difference performed as difference-in-differences on relative outcomes is only possible for grouped data. Aggregation is sometimes also suggested as a way to circumvent intra-cluster correlation issues (Angrist and Pischke, 2008, p. 313). The outcome variable is always log-transformed weakly earnings, and  $s$  denotes state,  $i$  individual, and  $t$  time period.

$$\log Y_{sit}^{female} = \alpha_{states} + \gamma_{year} + \delta(T * post) + \epsilon_{sit} \quad (2.15)$$

$$\begin{aligned} \log Y_{sit} = & \\ & \alpha_{states} + \gamma_{year} + \beta_1(female) + \beta_2(T * female) + \beta_3(post * female) + \beta_4(T * post) + \\ & \delta(T * post * female) + \epsilon_{sit} \end{aligned} \quad (2.16)$$

$$\log \bar{Y}_{st}^{females} = \alpha_{states} + \gamma_{year} + \delta(T * post) + \epsilon_{st} \quad (2.17)$$

$$\log(\bar{Y}_{st}^{females} / \bar{Y}_{st}^{males}) = \alpha_{states} + \gamma_{year} + \delta(T * post) + \epsilon_{st} \quad (2.18)$$

$$\begin{aligned} \log \bar{Y}_{st,gender} = \\ \alpha_{states} + \gamma_{year} + \beta_1(female) + \beta_2(T * female) + \beta_3(post * female) + \\ \beta_4(T * post) + \delta(T * post * female) + \epsilon_{st,gender} \end{aligned} \quad (2.19)$$

We repeat these estimations for 25, 5, 2, and 1 treated clusters, holding the total number of clusters constant at 51.<sup>17</sup> We also show results with different ways to estimate the standard errors. We use either uncorrected standard errors assuming independent and identically distributed errors (IID), White heteroskedasticity (HC1) robust standard errors (as Stata robust), or clustered standard errors on state-level (as Stata xtreg cluster). The implementation is done by the package `Fixest` in R, Bergé (2018).<sup>18</sup> Finally, we simulate with a true effect of either two or five percent to get a sense of power, or the ability to detect true effects. The effects are simulated by adding a fixed increase of two or five percent of the pre-treatment weakly earnings for women, to the post-treatment outcomes for women in the treatment state, shifting the level,

---

<sup>17</sup>We deviate from Bertrand et al. (2004) who focus on the total number of clusters. However, as Conley and Taber (2011) point out, the asymptotics are for both treated and untreated groups, so we expect similar results as if we had just reduced the number of groups, holding the number of treated groups constant. We also expect this to be a more common issue, as even in the case of few clusters, the results will suffer from few treated clusters. For difference-in-differences, the scenario is covered in for instance Conley and Taber (2011), MacKinnon and Webb (2018), MacKinnon and Webb (2020), and Ferman and Pinto (2019).

<sup>18</sup>For more information see <https://cran.r-project.org/package=fixest> and [https://cran.r-project.org/web/packages/fixest/vignettes/standard\\_errors.html](https://cran.r-project.org/web/packages/fixest/vignettes/standard_errors.html).



but not the trend. The results are shown in Tables 2.5, 2.6, and 2.7.

## 2.B.1 Results

Table 2.5 shows rejection rates for placebo treatments for individual level data. The estimators in use are the difference-in-differences of Equation 2.15, in column 1-3, and the triple difference estimator of Equation 2.16 in column 4-6. For each estimator, the rejection rates are either based on IID standard errors, robust standard errors or clustered at the state level, in that order. Column 1 corresponds to the results from Bertrand et al. (2004), with rejection rates roughly between 60 and 70 percent, i.e. we find a significant effect 60 to 70 percent of the times in the case of IID standard errors and placebo treatment, which is severe over-rejection<sup>19</sup>. This is almost identical to the results with robust standard errors, as shown in column 2. Furthermore, the number of treated clusters, specified in the row names, has limited impact, going from rejection rates of about 70 percent to about 60 percent when reducing the number of treated clusters from 25 to 1. When clustering the standard errors, the rejection rate is 7 percent in the case of 25 treated clusters, which is close to what we would want, but it rises to 15 percent in the case of 5 treated clusters, 34 percent for 2 treated clusters, and 74 percent for 1 treated cluster, as expected when using standard errors based on cluster asymptotics.

Turning to the triple difference estimator, the results for IID and robust standard errors still over-reject, with rejection rates of about 30 percent, regardless of how many treated clusters there are, as shown in column 4

---

<sup>19</sup>Strictly speaking, column 1, row 1 is the same specification as row 1 in Table II in (Bertrand et al., 2004, p. 257), except for the choice of covariates and the number of simulations.

and 5. This is about half the rejection rate of the equivalent difference-in-differences, but still much higher than we would want. Finally, looking at column 6, the rejection rate is 6 percent for 25 treated clusters, which is close to ideal, 13 percent for 5 treated clusters, 37 percent for 2 treated clusters and 82 percent for 1 treated cluster, showing the same pattern as the difference-in-differences, however, mildly preferred for 5-25 treated clusters, but not for 1-2 treated clusters. The differences are, however, marginal. As for clustered errors, there seems to be little gain, nor lose, from moving from a difference-in-differences to a triple difference estimator, in terms of false positives.

Table 2.6 compares three scenarios, no effect (placebo treatments) as above, a simulated 2 percent true effect, and a simulated 5 percent true effect.<sup>20</sup> Each scenario contains both a difference-in-differences estimator and a triple difference estimator. All specifications are with individual level data, standard errors clustered at the state level, and either 25, 5, 2, or 1 treated clusters.

With many (25) treated clusters, the triple difference estimator shows signs of having more power than the difference-in-differences estimator, providing rejection rates of 55 percent to 21 percent in the case of a 2 percent effect, and 99.8 percent to 69 percent in the case of a 5 percent effect. When reducing the number of treated clusters to 5, the triple difference still has more power, with 36 percent to 29 percent for a 2 percent effect, and 82 percent to 57 percent for a 5 percent effect, which is better, but with smaller margins. With even fewer treated clusters the differences become slighter, and it is worth remembering that we get high rejection rates, even in the absence of treatment, when we have few treated clusters.

---

<sup>20</sup>Note that for expositional reasons Table 2.6, columns 1 and 2 are identical to Table 2.5, columns 3 and 6, respectively.

Table 2.7 are all performed on data aggregated to the state-year-gender level (equations 2.17-2.19, and clustered standard errors. Aggregation to avoid intra-cluster correlation is common and suggested for instance by Angrist and Pischke (2008, p. 313). However, in the case of the (full) triple difference, we cannot aggregate to state-year, but have to aggregate to state-year-gender, leaving two observations per state per year, not fully avoiding the intra-cluster correlation structure. However, in the special case of Equation 2.18, which is also shown (in levels) in Section 2.6, we estimate the triple difference as a difference-in-differences on a relative outcome, preserving the one observation per state per year structure.

There are some notable patterns. In the case of no effect and 25 treated clusters, the rejection rates differ only marginally, and are similar to the individual level estimations. As we decrease the number of treated clusters, the rejection rates goes up to about 12 percent for 5 treated clusters, 29-34 percent for 2 treated clusters, and 72-79 percent for 1 treated cluster. As opposed to the individual level data, the difference-in-differences is mildly preferred to the triple difference estimator (in both its forms), and there is marginal improvement by aggregating, but typically only by a few percent, for both estimators, which unfortunately is not very helpful considering the overall scale of over-rejection. Further, there is virtually no difference between the triple difference as difference-in-differences and the full triple difference estimator. This is of course partly true because they use the same cluster asymptotics. Had we used robust (for instance White HC1), it is likely that the triple difference as difference-in-differences would look more like the difference-in-differences, as it would be based on regular asymptotics and they have the same degrees of freedom.

When we introduce true, simulated effects, the same pattern as individual level data arises. The triple difference, in both its forms, has higher power. With 25 treated clusters the rejection rates are 31 percent to 10 percent for a 2 percent effect, and 95 percent to 40 percent for a 5 percent effect. However, as the number of treated clusters decrease to 5, the difference also decreases, with 21 percent to 16 percent for a 2 percent effect, and 63 percent to 32 percent for a 5 percent effect. The difference becomes even smaller for 2 and 1 treated clusters.

Overall, aggregation has only minor consequences for false positives, but major consequences for power. This is seen by comparing Tables 2.6 and 2.7. Even in the case where the asymptotics seems reasonable, i.e. 25 treated clusters, the consequences with a 2 percent effect is a reduction in rejection rates from 21 percent to 10 percent for the difference-in-differences, and 55 percent to 31 percent for the triple difference. For a 5 percent effect the reduction is from 69 percent to 40 percent for the difference-in-differences, and 99.8 percent to 95 percent for the triple difference.

To conclude, for individual level data, clustered errors, and five or more treated clusters, the triple difference typically performs slightly better than the difference-in-differences, with the reverse being true for 1-2 treated clusters, when it comes to false positives. However, the differences are marginal compared to the overall issue of over-rejection. When it comes to power, the triple difference outperforms the difference-in-differences, often by a lot, in almost all cases. Aggregation does not solve much, and comes at a large cost in terms of power. It is also noteworthy that there is close to no difference between the full triple difference and the triple difference as difference-in-differences, either in terms of false positives or power. Re-

**Table 2.5:** No effect: Rejection rates for individual level data models

	Difference-in-differences			Triple difference		
	(1)	(2)	(3)	(4)	(5)	(6)
	Uncorrected	White	Cluster	Uncorrected	White	Cluster
25	0.7058	0.7072	0.0711	0.3013	0.3021	0.0592
5	0.6757	0.6787	0.1530	0.3095	0.3106	0.1343
2	0.6187	0.6247	0.3421	0.3164	0.3192	0.3678
1	0.6015	0.6071	0.7408	0.2936	0.3096	0.8243
n	496,055	496,055	496,055	1,035,308	1,035,308	1,035,308

The table shows rejection rates for the treatment variable coefficient at a five percent significance level, over 10 000 simulations, individual level data, and a placebo treatment (no effect). Columns 1-3 is the difference-in-differences estimator, while columns 4-6 is the triple difference estimator. The columns differ in the standard errors that are used, where 1 and 4 makes no correction for correlation, 2 and 5 are White HC1 robust standard errors, while 4 and 6 are cluster robust standard errors. The row names indicate how many (placebo) treated clusters there were out of the total of 51.

searchers considering the triple difference should rest assured that in optimal cases with many (treated) clusters, the triple difference is typically at least as good as the difference-in-differences, and often much better. But beware that it suffers almost equally to the difference-in-differences estimator in the presence of few treated clusters and serially correlated errors.

**Table 2.6:** Rejection rates for individual level data models

	No effect		2 percent effect		5 percent effect	
	(1)	(2)	(3)	(4)	(5)	(6)
	DD	Triple	DD	Triple	DD	Triple
25	0.0711	0.0592	0.2148	0.5539	0.6941	0.9984
5	0.1530	0.1343	0.2910	0.3584	0.5761	0.8235
2	0.3421	0.3678	0.4213	0.4289	0.5803	0.7191
1	0.7408	0.8243	0.7830	0.7905	0.8391	0.9030
n	496,055	1,035,308	496,055	1,035,308	496,055	1,035,308

The table shows rejection rates for the treatment variable coefficient at a five percent significance level, over 10 000 simulations, individual level data, and a either a placebo treatment (columns 1 and 2), a two percent effect (columns 3 and 4), or a five percent effect (columns 5 and 6). Columns 1, 3, and 5 use the difference-in-differences estimator, while columns 2, 4, and 6 use the triple difference estimator. All standard errors are clustered at the state level. The row names indicate how many (placebo or real) treated clusters there were out of the total of 51.

Table 2.7: Rejection rates for aggregated models

No effect		2 percent effect			5 percent effect			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
DD	Triple as DD	Triple	DD	Triple as DD	Triple	DD	Triple as DD	Triple
25	0.0479	0.0515	0.0517	0.1047	0.3101	0.3131	0.3958	0.9488
5	0.1120	0.1235	0.1248	0.1562	0.2133	0.2148	0.3173	0.6330
2	0.2894	0.3371	0.3391	0.3343	0.3714	0.3730	0.4118	0.5526
1	0.7237	0.7929	0.7929	0.7490	0.8092	0.8092	0.7763	0.8495
n	1,071	1,071	2,142	1,071	1,071	2,142	1,071	2,142

The table shows rejection rates for the treatment variable coefficient at a five percent significance level, over 10 000 simulations, state-year-gender aggregated data, and a either a placebo treatment (rows 1-3), a two percent effect (rows 4-6), or a five percent effect (rows 7-9). Rows 1, 4, and 7 use the triple difference estimator, rows 2, 5, and 8 use the triple difference as difference-in-differences (i.e. a difference-in-differences on a relative outcome), and rows 3, 6, and 9 is the full triple difference estimator. All standard errors are clustered at the state level. The column names indicate how many (placebo or real) treated clusters there were out of the total of 51.

## Chapter 3

# Fraud detection by a multinomial model: Separating honesty from unobserved fraud

Jonas Andersson, Andreas Olden and Aija Polakova\*

### Abstract

In this paper we investigate the EM-estimator of the model by Caudill et al. (2005). The purpose of the model is to identify items, e.g. individuals or companies, that are wrongly classified as honest; an example of this is the detection of tax evasion. Normally, we observe two groups of items, labeled *fraudulent* and *honest*, but suspect that many of the observationally honest items are, in fact, fraudulent. The items observed as *honest* are therefore divided into two unobserved groups, *honestH*, representing the truly honest, and *honestF*, representing the items that are observed as honest, but that are actually fraudulent. By using a multinomial logit model and assuming commonality between the observed *fraudulent* and the unobserved *honestF*, Caudill et al. (2005) present a method that uses the EM-algorithm to separate them. By

---

\*Affiliation of all authors: Department of Business and Management Science, NHH Norwegian School of Economics.



means of a Monte Carlo study, we investigate how well the method performs, and under what circumstances. We also study how well bootstrapped standard errors estimates the standard deviation of the parameter estimators.

### **3.1 Introduction**

Fraud is a fact of social behaviour having increasingly important consequences including loss of revenues to businesses, government, and society. Fraud is also expensive, driving up cost for detection and fraud risk reduction. As a result, active fraud control has gradually become an integrated part of business decision-making processes. Insurance companies must deal with fraud perpetrated by consumers on the firm and spend money on fraud detection and monitoring. A lot of research has focused on the fraud detection efforts and the frequency of fraud, that is, assessing and ranking the fraud suspiciousness of individual claims (Ai et al., 2013, 2009; Artís et al., 1999; Brockett et al., 2002; Derrig and Ostaszewski, 1995; Viaene et al., 2002).

Fraud is often detected by some sort of audit process. Audits will normally reveal some information about the fraudsters, the type of situations where fraud occurs, or the products where fraud is common. This information is then used to predict which claims are more likely to be fraudulent in the future. However, for most audits, the detection probability is not one hundred percent, which skews the estimated probabilities. This occurs because there are people in the group assumed not to be fraudulent who are actually fraudulent, making the observed fraudulent and the observed honest too similar. We expect that much data has this structure, in particular insurance claims data, tax data, and medical or diagnostic data. Moreover, the larger the fraction of misclassified observations, the worse the problem

becomes.

Numerous studies develop techniques to identify or classify fraudulent claims. Predictive techniques are used to predict values for a certain target variable, such as credit scoring to predict repayment behaviour of loan applicants, and logistic regression models, both binary and multinomial logit models, are used for detecting manipulation such as dishonest insurance claims (Major and Riedinger, 2002; Olinsky et al., 1996). Artís et al. (2002) find a significant portion of the claims that were previously classified as legitimate contain omission errors, and thus are likely to be fraudulent. Further, Hausman et al. (1998) show that ignoring potential misclassification of a dependent variable can result in biased and inconsistent coefficient estimates when using standard parametric specifications.

Artís et al. (2002) present a logistic regression model accounting for misclassified claims and estimates it by the method of Hausman et al. (1998). Caudill et al. (2005) estimate this model by means of a multinomial logit model (MNL) and the EM-algorithm. They argue that identifying fraudulent claims is similar in nature to several other problems in real life including medical and epidemiological problems. They describe the methodology that can be used to produce parameter estimates with a dataset containing potentially misclassified dependent variables. Further, they estimate the proportion of fraudulent claims for car damage that are potentially erroneously classified as honest by an insurance company. The procedure is based on a transformation of the standard MNL likelihood function into a missing data formulation to which the expectation maximization (EM) algorithm can be applied (Dempster et al., 1977).

By assuming that the fraudsters that are caught have similar characteristics

with the ones that are not, a latent variable model can be specified, where the group of those that were not caught in the initial audits is divided into two groups, the uncaught fraudsters and the truly honest claims. The model can then be estimated by the Expectation Maximization (EM) algorithm for missing data and be used to identify claims that probably are fraudulent, even if the audit did not catch them. This idea was introduced by Caudill et al. (2005), who fit the model to a dataset of Spanish car insurance fraud. Since this is real data, we do not know whether the EM-algorithm actually provides an improvement over other fraud detection methods, only that it is implementable. Therefore, this paper investigates the methodology by means of a Monte Carlo study in order to evaluate its performance. We simulate data and vary the key relationships to evaluate the improvements that the EM-algorithm provides. We compare the parameter estimates obtained after running the EM-algorithm with the estimates obtained under a perfect information scenario. Additionally we compare the EM-parameters to a naive binomial logit model that does not take the misclassification into account. By doing so, we can see how much estimation accuracy we lose due to not having full information, and the improvement in performance over the naive approach.

The particular models we perform our simulation study on are guided by the empirical results of Caudill et al. (2005). The insurance claims categorized as honest, even though they are actually fraudulent, constitute the misclassified (missing) data. The data used by Caudill et al. (2005) is taken from Artís et al. (2002) and we use the standard deviations and coefficients from the paper by Artís et al. (2002) to simulate our data. In our simulated data we, of course, have full knowledge of whether a claim that is observed as being honest is really honest, or whether it is actually fraudulent.

The EM algorithm consists of two steps. In the Expectation (E) step, unobserved indicator variables associated with truly honest and honest-fraudulent claims are replaced with their conditional expectations, given the data and values of the unknown parameters. These conditional expectations or probabilities can be readily computed, given the structure of the logit model. In the Maximization (M) step, the log-likelihood function is maximized, new parameter values are obtained and then the E and M steps are repeated until the likelihood function is maximized. When the parameters are estimated, we can obtain the final estimates of the probabilities of whether a claim is fraudulent or not.

The EM-algorithm avoids the problem that the binomial logit model incurs, where all claims are assumed to be correctly classified; hence, the EM-algorithm avoids using misclassified observations for calculating probabilities, which is the cause of incorrect probabilities. Our results are aimed at revising claims initially classified as honest by reopening investigations and examining claims more closely, but might also improve prediction models. Further, this allows us to identify weaknesses in the initial classification system.

The paper is structured in the following way. Section 2 gives a literature review on theoretical and empirical studies of the detection of fraudulent claims. Section 3 presents the model by Caudill et al. (2005) and how to estimate it by means of the EM-algorithm. In Section 4 the performance of the EM-estimator is evaluated against two benchmark estimators by means of a Monte Carlo Study. A conclusion closes the paper.

## 3.2 The model

### 3.2.1 The multinomial model with missing information

The model by Caudill et al. (2005) is based on a multinomial distribution<sup>2</sup> with three categories. The first category, the honest honest (HH), are claims not caught in the audit process that are indeed not fraudulent. The second category, the honest fraudulent (HF), are fraudulent claims not caught in the audit process. The last category consists of fraudulent claims (F) caught in the audit process.

If all three categories were observed, it would simply be a trinomial logit model. Of course, this is not the case and the model was developed in order to allow for, and estimate the probability of undetected fraudulent claims. In order to do so, a similarity between the detected and undetected fraudulent claims will be assumed and reflected in a parameter restriction that will be imposed in the model. By denoting the number of HH as  $Y_1$ , the number of HF as  $Y_2$  and the number of F as  $Y_3$ , we assume that

$$(Y_1, Y_2, Y_3) \sim \text{MN}(1, (p_1, p_2, p_3)), \quad (3.1)$$

implying that

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = p_1^{y_1} \cdot p_2^{y_2} \cdot p_3^{y_3}. \quad (3.2)$$

The probability for an individual to belong to group 1, 2 or 3, respectively, is

---

<sup>2</sup>We denote this distribution  $\text{MN}(n, \mathbf{p})$ , where  $n$  is the number of trials and  $\mathbf{p}$  is a  $K$ -vector containing the probabilities for each of  $K$  categories.

assumed to be given by the following equations

$$p_1 = P(Y_1 = 1, Y_2 = 0, Y_3 = 0) = \frac{1}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_3 x}}, \quad (3.3)$$

$$p_2 = P(Y_1 = 0, Y_2 = 1, Y_3 = 0) = \frac{e^{\alpha_2 + \beta_2 x}}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_3 x}}, \quad (3.4)$$

$$p_3 = P(Y_1 = 0, Y_2 = 0, Y_3 = 1) = \frac{e^{\alpha_3 + \beta_3 x}}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_3 x}}. \quad (3.5)$$

In order to identify the parameters Caudill et al. (2005), assume that  $\beta_2 = \beta_3$ , i.e. that the probability of an individual to be fraudulent is affected by the explanatory variables in the same way, independent on whether the fraud has been detected or not. A difference in the probability for an individual to be in classes 2 or 3 is still allowed since  $\alpha_2$  and  $\alpha_3$  are free parameters.

### 3.2.2 Estimation of the model using EM algorithm

The  $\alpha$ 's and the  $\beta$ 's are parameters to be estimated, and  $x$  is a vector of exogenous variables. We can now write the log-likelihood function

$$\ln L(\alpha_2, \alpha_3, \beta_2) = \sum_{i=1}^n (Y_{1i} \ln p_1 + Y_{2i} \ln p_2 + Y_{3i} \ln p_3), \quad (3.6)$$

where  $i$  ( $i = 1, \dots, n$ ) represents an individual  $i$ , and  $n$  is the sample size. The ML-estimator is obtained by maximizing the log-likelihood with respect to the parameters  $\alpha_2$ ,  $\alpha_3$  and  $\beta_2$ .

We have only observed a binomial variable  $(Z_2, Z_3) = (Y_1 + Y_2, Y_3)$ , but model

it as a trinomial  $(Y_1, Y_2, Y_3)$ .

Since we do not observe all three categories we cannot compute  $\ln L(\alpha_2, \alpha_3, \beta_2)$ .

We therefore use the suggested by Caudill et al. (2005), which is based on Expectation Maximization (EM) algorithm. Briefly described,

1. *Expectation (E) step*

we compute the expectation of  $\ln L(\alpha_2, \alpha_3, \beta_2)$  conditional on the observed data.

$$Q(\alpha_2, \alpha_3, \beta_2) = E(\ln L(\alpha_2, \alpha_3, \beta_2) | \mathbf{Y}), \quad (3.7)$$

where  $\mathbf{Y}$  is an  $n \times 3$ -matrix where element  $(i, j)$  is equal to 1 if observation nr  $i$  belongs to category  $j$  and zero otherwise. Conditioning on the  $x$ -observations are also done but is not expressed explicitly in the formulas here. The conditional expectation of each term in (3.6) is computed by observing that only one of  $Z_2$  and  $Z_1$  is equal to one; the other is zero. We need

$$\begin{aligned} Y_{2i}^* &:= E(Y_{2i} | Z_{2i} = 1) = P(Y_{2i} = 1 | Z_{2i} = 1) = \frac{p_2}{p_2 + p_3} \\ &= \frac{e^{\alpha_2 + \beta_2 x}}{1 + e^{\alpha_2 + \beta_2 x}} \end{aligned} \quad (3.8)$$

and similarly

$$Y_{1i}^* = \frac{1}{1 + e^{\alpha_2 + \beta_2 x}} \quad (3.9)$$

2. *Maximization (M) step*

The log-likelihood function  $\ln L(\alpha_1, \alpha_2, \beta_2)$  is maximized, where  $Y_{1i}$  and  $Y_{2i}$  are substituted by  $Y_{1i}^*$  and  $Y_{2i}^*$ . New  $\alpha$  and  $\beta$  estimates are found, these are plugged in the step 1 and new  $Y_i^*$  estimates are found. Log-

likelihood function is maximized again, based on these new values, and the whole process is iterated until the log-likelihood function is at its maximum.

In R, this can be implemented by the following algorithm:

1. Select starting values for  $\alpha_2^{(1)}, \alpha_3^{(1)}, \beta_2^{(1)}$ .
2. Compute  $Y_{1i}^{*(1)}, Y_{2i}^{*(1)}$ , based on these parameter values.
3. Maximize log-likelihood function where  $Y_{1i}$  and  $Y_{2i}$  are substituted with  $Y_{1i}^{*(1)}$  and  $Y_{2i}^{*(1)}$ . After maximization, new parameter values  $\alpha_2^{(2)}, \alpha_3^{(2)}, \beta_2^{(2)}$  are obtained.
4. Steps 2 to 3 are repeated until there is a convergence to the maximum likelihood estimator.

As starting values, we use the values obtained from estimating a binomial logit model, assuming no misclassification has been done.

### 3.3 Monte Carlo Study

In order to investigate how well the method manages to estimate the parameters of the model when some observations are incorrectly classified as honest, we perform a simulation study. The results can be seen as a best case scenario since we assume that we know the data generating process (DGP) except for the parameter values, which have to be estimated. In reality, the explanatory variables, and the functional form for the probabilities, will of course, not be specified exactly in accordance with the DGP.



Though simplified, in order to study realistic situations we have chosen the parameter values of the models guided by the study of Caudill et al. (2005). For the sake of a clear exhibition, we here recapitulate the model. Each individual claim is represented by a trinomial variable  $(Y_1, Y_2, Y_3)$  with zeros in two of the three entries and 1 in the class where the claim belongs. The probabilities are given by

$$p_1 = P(Y_1 = 1, Y_2 = 0, Y_3 = 0) = \frac{1}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_2 x}}, \quad (3.10)$$

$$p_2 = P(Y_1 = 0, Y_2 = 1, Y_3 = 0) = \frac{e^{\alpha_2 + \beta_2 x}}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_2 x}}, \quad (3.11)$$

$$p_3 = P(Y_1 = 0, Y_2 = 0, Y_3 = 1) = \frac{e^{\alpha_3 + \beta_2 x}}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_2 x}}.^3 \quad (3.12)$$

In the simulation study, performed in R (R Core Team, 2020)<sup>4</sup>, we compare the results of the EM-estimators with two benchmarks. The first is to simply ignore that there is misclassification, i.e. to use a binomial logit model to estimate  $\beta_2$ ; this estimator is denoted  $\hat{\beta}_2^B$ . This would be possible to do in practice. The second benchmark is to pretend that we actually observed all three categories and estimate  $\beta_2$  by means of a trinomial logit model; this estimator is denoted  $\hat{\beta}_2^T$ . This is, obviously, not possible to do when the observations are only marked as "caught" or "not caught". It is, however, a good comparison since we, with the EM-estimator, are trying to fit exactly the same model, but with a reduced form of the data. For all models and

---

<sup>3</sup>Note that the restriction  $\beta_2 = \beta_3$  is explicitly imposed in the model.

<sup>4</sup>The code can be found at [https://github.com/andreasolden/em\\_algorithm\\_missing\\_data](https://github.com/andreasolden/em_algorithm_missing_data)

parameter combinations, 1000 replications have been performed to compute the Monte Carlo means and standard deviations. The sample size for each replication was 1000.

To compute the standard error, the estimated standard deviation, of the EM-estimator, the non-parametric bootstrap with 200 replications is used. 200 replications were chosen according to Tibshirani and Efron (1993), who show that running more than 200 replications provides very limited improvements in bootstrapped standard errors. An experiment with 1000 bootstrap replications for the standard errors was also conducted and did not indicate a noticeable improvement. For the trinomial and binomial logit model the standard errors are computed using the asymptotic distribution of the ML-estimators. Since there are no known closed form expressions for the standard deviations of the estimators, the performance of the standard errors is investigated by comparing their Monte Carlo mean with the Monte Carlo standard deviation of the estimator. In the tables, we call the latter the "True SD". Strictly speaking, this is of course only correct if the number of replications is infinitely large. For some replicates, the EM-algorithm did not converge after 100 iterations. However, we found no evidence that these estimates were systematically different from the ones that did. The tables presented below looked very similar when those replications were removed.

### 3.3.1 One explanatory variable

In this section we study the case with only one explanatory variable,  $x_1$ . The parameters of interest to estimate are therefore  $(\alpha_2, \alpha_3, \beta_2)$ . Guided by the empirical study in Caudill et al. (2005) we start by setting  $(\alpha_2, \alpha_3, \beta_2) = (-1.8, -1.5, -0.02)$ . The variable  $x_1$  is thought of as the variable *AGE* in

Caudill et al. (2005). The standard deviation of  $AGE$  is 12.3 and that is what we use as our starting case after which we also vary this quantity.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.670	0.095	0.084
$\hat{\alpha}_2^T$	-1.800	0.100	0.099
$\hat{\alpha}_2^B$	-1.664	0.087	0.087
$\alpha_3 = -1.5$			
$\hat{\alpha}_3^{EM}$	-1.486	0.101	0.090
$\hat{\alpha}_3^T$	-1.505	0.087	0.088
$\beta_2 = -0.02$			
$\hat{\beta}_2^{EM}$	-0.021	0.009	0.009
$\hat{\beta}_2^T$	-0.020	0.006	0.006
$\hat{\beta}_2^B$	-0.017	0.007	0.007

**Table 3.1:** Different estimators of  $\alpha_2$ ,  $\alpha_3$  and  $\beta_2$  when the true values are  $\alpha_2 = -1.8$ ,  $\alpha_3 = -1.5$  and  $\beta_2 = -0.02$  and  $sd(x_1) = 12.3$ . Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

From Table 3.1, we see that, with the exception of  $\alpha_2$ , the EM-estimator performs almost as well as if all three categories would have been observed when  $sd(x_1) = 12.3$ . On the other hand, the mistake of ignoring misclassification, which is done by  $\hat{\beta}_2^B$  is not that consequential. The bootstrapped standard errors are, on average, slightly underestimating the standard deviation of  $\hat{\alpha}_2^{EM}$  and  $\hat{\alpha}_3^{EM}$ .

The benefit of the EM-estimator can, however, be seen in Table 3.2, where the standard deviation of the explanatory variable is increased with a factor of 10. The Monte Carlo mean of  $\hat{\beta}_2^B$  is then  $-0.011$  compared to the true value  $-0.020$ . The EM-estimator,  $\hat{\beta}_2^{EM}$ , has a Monte Carlo mean of  $-0.021$ . The estimation uncertainty is also, as expected, much smaller for the case with a large standard deviation in the explanatory variable. The estimation uncertainty, manifested in the Monte Carlo standard deviations (True SD), is

of course larger than if we had observed all three categories.

As can be seen by comparing Table 3.1 in Table 3.2, and in all the remaining simulation experiments in the paper,  $\hat{\alpha}_2^{EM}$  is less biased when the variance of the explanatory variable(s) is larger.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.824	0.360	0.353
$\hat{\alpha}_2^T$	-1.804	0.122	0.116
$\hat{\alpha}_2^B$	-1.700	0.095	0.104
$\alpha_3 = -1.5$			
$\hat{\alpha}_3^{EM}$	-1.500	0.173	0.168
$\hat{\alpha}_3^T$	-1.503	0.108	0.108
$\beta_2 = -0.02$			
$\hat{\beta}_2^{EM}$	-0.021	0.003	0.003
$\hat{\beta}_2^T$	-0.020	0.001	0.001
$\hat{\beta}_2^B$	-0.011	0.001	0.001

**Table 3.2:** Different estimators of  $\alpha_2$ ,  $\alpha_3$  and  $\beta_2$  when the true values are  $\alpha_2 = -1.8$ ,  $\alpha_3 = -1.5$  and  $\beta_2 = -0.02$  and  $sd(x_1) = 123$ . Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

### 3.3.2 Two explanatory variables

In order to investigate how the addition of more explanatory variables affects the results we now let  $\mathbf{x} = (x_1, x_2)'$  and  $\beta_2 = (\beta_{21}, \beta_{22})'$  be 2-dimensional vectors and the terms  $\beta_2 x$  in equations (3.10)-(3.12) replaced by  $\mathbf{x}'\beta_2$ . The choice of parameter values studied is, again, guided by Caudill et al. (2005).  $x_1$  is again thought of as representing the variable *AGE* and the  $x_2$  variable *RECORDS*.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.680	0.214	0.256
$\hat{\alpha}_2^T$	-1.806	0.105	0.100
$\hat{\alpha}_2^B$	-1.671	0.088	0.089
$\alpha_3 = -1.5$			
$\hat{\alpha}_3^{EM}$	-1.483	0.107	0.110
$\hat{\alpha}_3^T$	-1.504	0.088	0.089
$\beta_2 = -0.02$			
$\hat{\beta}_2^{EM}$	-0.022	0.009	0.010
$\hat{\beta}_2^T$	-0.021	0.006	0.006
$\hat{\beta}_2^B$	-0.017	0.007	0.007
$\beta_3 = 0.2$			
$\hat{\beta}_3^{EM}$	0.207	0.065	0.071
$\hat{\beta}_3^B$	0.202	0.041	0.041
$\hat{\beta}_3^B$	0.168	0.050	0.049

**Table 3.3:** Different estimators of  $\alpha_2$ ,  $\alpha_3$ ,  $\beta_2$  and  $\beta_3$  when the true values are  $\alpha_2 = -1.8$ ,  $\alpha_3 = -1.5$ ,  $\beta_2 = -0.02$  and  $\beta_3 = 0.2$ ,  $sd(x_1) = 12.3$ ,  $sd(x_2) = 1.8$  and  $\text{Corr}(x_1, x_2) = 0$ . Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

As Table 3.3 shows, both the estimators and the standard errors are close to their respective true values. In the binomial model, the  $\beta$ -coefficients are biased towards zero due to the misspecification. Also for this case we investigate a situation with more variation in the explanatory variables. In Table 3.4, the standard deviations of  $x_1$  and  $x_2$  are both multiplied by 10.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.843	0.276	0.288
$\hat{\alpha}_2^B$	-1.808	0.140	0.136
$\hat{\alpha}_2^B$	-1.630	0.091	0.105
$\alpha_3 = -1.5$			
$\hat{\alpha}_3^{EM}$	-1.524	0.193	0.205
$\hat{\alpha}_3^T$	-1.507	0.130	0.131
$\beta_2 = -0.02$			
$\hat{\beta}_2^{EM}$	-0.020	0.003	0.004
$\hat{\beta}_2^B$	-0.020	0.002	0.002
$\hat{\beta}_2^B$	-0.007	0.001	0.001
$\beta_3 = 0.2$			
$\hat{\beta}_3^{EM}$	0.205	0.031	0.035
$\hat{\beta}_3^B$	0.202	0.014	0.014
$\hat{\beta}_3^B$	0.072	0.006	0.006

**Table 3.4:** Different estimators of  $\alpha_2$ ,  $\alpha_3$ ,  $\beta_2$  and  $\beta_3$  when the true values are  $\alpha_2 = -1.8$ ,  $\alpha_3 = -1.5$ ,  $\beta_2 = -0.02$  and  $\beta_3 = 0.2$ ,  $sd(x_1) = 123$ ,  $sd(x_2) = 18$  and  $\text{Corr}(x_1, x_2) = 0$ . Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

We now go back to the case with the original standard deviations of  $x_1$  and  $x_2$ , 12.3 and 1.8, respectively, but impose a correlation of 0.5 between them. The result is presented in Table 3.5. The difference with Table 3.3 is surprisingly small and we therefore investigated this by increasing the correlation to 0.9. This is presented in Table 3.6, which shows that the standard deviation of the estimators for  $\beta_2$  and  $\beta_3$  is larger.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.670	0.115	0.112
$\hat{\alpha}_2^T$	-1.804	0.104	0.100
$\hat{\alpha}_2^B$	-1.664	0.087	0.088
$\alpha_3 = -1.5$			
$\hat{\alpha}_3^{EM}$	-1.483	0.101	0.096
$\hat{\alpha}_3^T$	-1.503	0.087	0.088
$\beta_2 = -0.02$			
$\hat{\beta}_2^{EM1}$	-0.021	0.009	0.011
$\hat{\beta}_2^T$	-0.020	0.006	0.007
$\hat{\beta}_2^B$	-0.017	0.008	0.008
$\beta_3 = 0.2$			
$\hat{\beta}_3^{EM}$	0.207	0.071	0.075
$\hat{\beta}_3^T$	0.203	0.046	0.047
$\hat{\beta}_3^B$	0.170	0.056	0.056

**Table 3.5:** Different estimators of  $\alpha_2$ ,  $\alpha_3$ ,  $\beta_2$  and  $\beta_3$  when the true values are  $\alpha_2 = -1.8$ ,  $\alpha_3 = -1.5$ ,  $\beta_2 = -0.02$  and  $\beta_3 = 0.2$ ,  $sd(x_1) = 12.3$ ,  $sd(x_2) = 1.8$  and  $\text{Corr}(x_1, x_2) = 0.5$ . Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.667	0.085	0.064
$\hat{\alpha}_2^T$	-1.803	0.102	0.099
$\hat{\alpha}_2^B$	-1.662	0.084	0.087
$\alpha_3 = -1.5$			
$\hat{\alpha}_3^{EM}$	-1.485	0.097	0.089
$\hat{\alpha}_3^T$	-1.504	0.084	0.088
$\beta_2 = -0.02$			
$\hat{\beta}_2^{EM}$	-0.021	0.020	0.020
$\hat{\beta}_2^T$	-0.021	0.014	0.013
$\hat{\beta}_2^B$	-0.018	0.016	0.016
$\beta_3 = 0.2$			
$\hat{\beta}_3^{EM}$	0.206	0.134	0.135
$\hat{\beta}_3^T$	0.203	0.094	0.091
$\hat{\beta}_3^B$	0.173	0.113	0.111

**Table 3.6:** Different estimators of  $\alpha_2$ ,  $\alpha_3$ ,  $\beta_2$  and  $\beta_3$  when the true values are  $\alpha_2 = -1.8$ ,  $\alpha_3 = -1.5$ ,  $\beta_2 = -0.02$  and  $\beta_3 = 0.2$ ,  $sd(x_1) = 12.3$ ,  $sd(x_2) = 1.8$  and  $\text{Corr}(x_1, x_2) = 0.9$ . Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

To summarize the simulation study, for the most part, the EM-estimator estimates the parameters of the investigated data generating processes (DGP) well even though some observations are misclassified; with the exception of the EM-estimator of  $\alpha_2$ , there are no indications that the parameter estimators are biased. In one experiment, identical to Table 3.1 but for the value of  $\alpha_2$  which was  $-3.0$  instead of  $-1.8$ , the EM-estimator seem to systematically converge to value close to  $\alpha_3$  ( $-1.5$ ) with the consequence that  $\hat{\alpha}_2^{EM}$  was severely biased towards zero. This might be due to convergence to a local optimum, an hypothesis strengthened by a convergence close to the true value ( $-3.0$ ), when the true values were used as starting values.

Bootstrapped standard errors also work well for the  $\beta$ -values in the investigated DGPs. However, the standard errors for the  $\alpha$ -estimators underesti-



mate the true standard deviations of the estimators when the variances of the explanatory variables are small. When the variances are large, the standard errors seem to overestimate the standard deviations of the estimators.

### 3.4 Conclusions

In this paper we investigate, by means of a Monte Carlo study, how well the EM-estimator of the model by Caudill et al. (2005) performs. We study different levels of variation in the explanatory variables in order to evaluate what is required to estimate the parameters well. In order to investigate realistic cases, we have chosen parameter values of the models guided by the study of Caudill et al. (2005) but simplified so that we study models with one or two explanatory variables only. In addition to investigating the point estimators of the parameter values we also study bootstrapped standard errors.

For the investigated models and parameter values, most point estimators work well, on average. The exception to this is the EM-estimator of  $\alpha_2$ , which determines the difference between the correctly and incorrectly classified fraudulent observations. The estimator worked well when the variance of the explanatory variables was large. Overall, the bootstrapped standard errors also perform adequately as estimators of the standard deviations of the estimators. There is one exception also to this, though. When the variation in the explanatory variables is small, the standard deviation for  $\hat{\alpha}_2^{EM}$  is underestimated and when the variance is large, it is overestimated.

We compare the estimators with two benchmarks. The first requires more data, namely that all three categories are observed. This trinomial model

serves as an upper limit for how well the EM-estimator, which uses less information, could perform. The second benchmark is a binomial logit model where the fact that some observations are misclassified is ignored. The trinomial model, combined with observations of all three categories, as expected, captures the parameter values more precisely than the EM-estimator. The only interest in the results for the estimator of the binomial model is to show the effect of ignoring a part of the model (analogous to omitted variable bias in a linear regression). With small variance in the explanatory variables the bias is surprisingly small for the binomial estimator (for the cases when they can be seen as estimators of parameters in the trinomial model). This bias is exacerbated when the variance is increased.

## Chapter 4

# Fraud Concerns and Support for Economic Relief Programs

Ingar K. Haaland and Andreas Olden\*

### Abstract

Using a probability-based sample of the Norwegian population, we test whether an informational treatment about fewer audits by the Norwegian Tax Administration during the peak of the COVID-19 crisis affects support for an economic relief program designed to save jobs and prevent bankruptcies. The information treatment significantly reduces support for the economic relief program. The underlying mechanisms are lower trust in the tax administration and more pessimism about its ability to detect misuse of the program.

(JEL D83, H25, H26)

---

\*Haaland: Department of Economics, NHH Norwegian School of Economics; Olden: Department of Business and Management Science, NHH Norwegian School of Economics. We thank Jarle Møen and Anne-Liv Scrase for extremely valuable feedback. We also thank seminar audiences at the communication forum for the Intra-European Organisation of Tax Administrations, the OECD Behavioural Insights COI meeting, the IRS Behavioral Research Community of Practice, the Swedish Tax Administration, and several internal seminars at the Norwegian Tax Administration for valuable comments. We further thank the Norwegian Tax Administration for funding the experiment and sharing the data. The views expressed herein are those of the authors and do not necessarily reflect the views of the Norwegian Tax Administration. The study has received IRB approval from NHH. The authors report no declarations of interest.

## 4.1 Introduction

The COVID-19 pandemic has resulted in a global economic crisis. Governments around the world have responded to the crisis with extraordinary economic relief programs. For instance, the US passed a historic USD 2 trillion economic relief package in late March 2020. At the same time, tax administrations have responded to the pandemic by suspending or reducing their audit activities, especially for on-site audits (OECD, 2020). This response is in line with recommendations from the OECD, which suggests a temporary change in auditing policies to ease the burden on taxpayers and reduce disease transmission risks in the case of on-site audits.

While a temporary reduction in audit activities during an economic and public health crisis involves obvious risks related to increased fraud, a second concern is that such a reduction might negatively affect trust in the tax administration. Governments in democratic countries need public support to roll out economic relief programs. If voters are concerned about fraud and misuse of public funds, reducing audit activities when economic relief programs are most needed might unintentionally undermine public support for the programs. In this paper, we examine how information about fewer audits affects trust in the tax administration and support for economic relief programs through an online survey experiment conducted in Norway during the peak of the COVID-19 crisis.

We ran our experiment with a large, probability-based sample of the Norwegian population. We first give our respondents some background information about Norway's most important COVID-19 relief package, the , which was administered by the Norwegian Tax Administration. After Norway imple-

mented strong infection control measures in March 2020, the Norwegian Tax Administration reduced its audit activity by conducting fewer physical audits. We provide this information to a random subsample of our respondents to test how information about reduced audit activities affects support for economic relief programs. We also measure post-treatment beliefs about the detection probability for firms trying to abuse the program and trust in the tax administration.

The main results of the paper are that information about fewer audits reduces trust in the tax administration and weakens support for economic relief programs. More specifically, treated respondents are 9.6 percentage points less likely than control group respondents to express trust in the tax administration's handling of the . This effect corresponds to a 14 percent reduction in trust compared to the control group mean of 67.6 percent. We also find that treated respondents believe that the tax administration is 4.6 percentage points less likely to catch firms trying to abuse the Business Compensation Scheme compared to control group respondents. This effect corresponds to a ten percent decrease from the control group mean of 45 percent. As a result of lower trust in the tax administration and more pessimism about its ability to detect fraud, treated respondents reduce their support for the Business Compensation Scheme by 5.6 percentage points compared to control group respondents. This corresponds to a seven percent reduction compared to the baseline support of 80 percent among control group respondents. These findings highlight the importance of maintaining normal audit levels during an economic crisis to preserve public trust in the system and maintain support for economic relief packages.

Our findings contribute to the literature on how taxpayers respond to in-

formation about audit activities (Blumenthal et al., 2001; Bott et al., 2019; De Neve et al., 2021; Doerrenberg and Schmitz, 2015; Kleven et al., 2011; Perez-Truglia and Troiano, 2018). This literature has shown that information about audits can affect tax compliance. We contribute to this literature by showing that information about audits also affects trust in the tax administration and policy preferences.<sup>2</sup> More generally, we contribute to the political economy literature using information provision experiments to study beliefs and public policy preferences (Alesina et al., 2018; Cruces et al., 2013; Fehr et al., 2019; Grigorieff et al., 2020; Haaland and Roth, 2021; Karadja et al., 2017; Kuziemko et al., 2015; Lergertporer et al., 2018; Roth et al., 2021).<sup>3</sup> This literature has mostly found muted impacts of information on public policy preferences. One reason for this could be that voters are not open to persuasion on ideologically charged topics such as redistribution or affirmative action (Haaland and Roth, 2021; Kuziemko et al., 2015). We contribute to this literature by showing that policy preferences are elastic to information on a topic characterized by low political polarization.<sup>4</sup> This paper proceeds as follows. Section 4.2 describes the sample and experimental design. Section 4.3 describes the results. Section 4.4 concludes.

---

<sup>2</sup>This finding also relates to a theoretical literature on how tax evasion affects policy preferences (Borck, 2009; Roine, 2006; Traxler, 2009).

<sup>3</sup>For a recent review of information provision experiments in economics, see Haaland et al. (2021).

<sup>4</sup>For instance, in March 2020, the US Senate approved a historic \$2 trillion COVID-19 stimulus bill in a unanimous 96–0 vote.

## 4.2 Sample and experimental design

### 4.2.1 Sample

We recruited respondents using Norstat, Norway's largest market research company. Norstat administers a large online probability-based panel of the Norwegian population where all 81,000 panelists are actively recruited to join the panel, mostly via telephone. All participants need to verify their phone number and answer a questionnaire on demographics before they are allowed to join the panel. The panel is constructed to be representative of the Norwegian population. Norstat maintains several procedures to secure high-quality survey responses. First, there are clear restrictions on how often panelists can take part in surveys. Each panelist typically completes one to two surveys per month. Second, panelists who consistently speed through surveys are excluded from the panel. Third, Norstat has a system for identifying duplicate accounts, making it very unlikely that someone completes the survey twice.

The survey was fielded by Norstat between May 7 and May 20. Out of 4,840 respondents invited into the survey, 1,482 respondents started the survey. 29 respondents were screened out due to full quotas, 1 person was screened out for other reasons, and 52 respondents dropped out of the survey (of which 34 respondents were assigned a treatment). There was no differential attrition by treatment assignment. The final sample consisted of 1400 respondents, which corresponds to our pre-specified sample size. Norstat provides survey weights that makes our sample representative of the Norwegian household population on gender, age, region. As shown

in Table 2, our unweighted sample is already quite representative of the general Norwegian on these dimensions. Our sample is less representative on education as our respondents are more likely than the general population to have a college degree.

#### 4.2.2 Experimental design

We first ask pre-treatment questions about gender, age, region, and education. Thereafter, we provide all respondents with background information about a government relief program, the . Half of our respondents are then exposed to an information treatment that the tax administration is doing fewer audits during the coronavirus crisis. Finally, we measure support for the Business Compensation Scheme and elicit post-treatment beliefs about trust in the tax administration, beliefs about fraud attempts, and beliefs about the detection probability.

Section .B of the provides an English translation of the survey instruments while Section .C provides screenshots of the original survey in Norwegian.<sup>5</sup> On May 18, we submitted a pre-analysis plan to the AsPredicted registry. While the survey was already in the field when we submitted the pre-analysis plan, the data collection was fully administered by Norstat and we did not obtain access to the data before the collection ended on May 20. The pre-analysis plan is included in Section .D of the and is also available on the following link: <http://aspredicted.org/blind.php?x=qa65g2>.

---

<sup>5</sup>The survey instruments were translated into English by professional translators working in the language section of the Norwegian Tax Administration.



## **Introductory text about the Business Compensation Scheme**

The Business Compensation Scheme was the Norwegian government's leading initiative to mitigate the negative effects of the COVID-19 crisis on the economy. It was initiated in April 2020 and allowed private enterprises that experienced a revenue fall of at least 30 percent to apply for government subsidies to cover up to 90 percent of their fixed costs. The stated aim of the scheme was to prevent unnecessary bankruptcies and safeguard Norwegian jobs during the coronavirus crisis.<sup>6</sup> The scheme was approved by the Norwegian parliament for the three-month period March–April 2020. It was estimated to cost the government 20 billion NOK (or approximately 2 billion USD) per month. In late May, the Business Compensation Scheme was extended for another three-month period, but with less generous subsidies, and then discontinued in August 2020.

The Business Compensation Scheme was the most important initiative of the Norwegian government to mitigate the negative economic impact of the coronavirus crisis, and it was heavily debated in May 2020 when the survey was fielded. However, it is unclear how much the average citizen knew about the scheme. We therefore presented the following text (translated from Norwegian) to respondents in both the treatment and control group to make them familiar with the context:

Recently, the Norwegian Government launched the Business Compensation Scheme, often referred to as the cash benefit scheme for businesses. The Business Compensation Scheme was established to provide financial aid to enterprises that have been severely im-

---

<sup>6</sup>More information about the Business Compensation Scheme is available on the Tax Administration's website: <https://www.skatteetaten.no/kompensasjonsordning/>.

pacted financially by the coronavirus crisis. The aid is provided through subsidies that cover up to 90 percent of the enterprises' fixed costs, for example their rent.

The purpose of the Business Compensation Scheme is to avoid unnecessary bankruptcies and redundancies during the coronavirus crisis. **An estimate shows that the scheme will cost the state around NOK 20 billion per month.**

It has been pointed out by the media that the Business Compensation Scheme may be abused by enterprises reporting too high fixed costs to the tax authorities.

The Norwegian Tax Administration has been charged with administration of the Business Compensation Scheme on the state's behalf. **The Norwegian Tax Administration is also responsible for ensuring that the scheme is not misused.**

### **Information treatment: Reduced control activity by the tax administration**

Immediately following the short introductory text about the , we inform respondents in the treatment group that the tax administration had reduced its control activity during the coronavirus crisis. This fact received public attention in late April 2020 when Norway's largest business newspaper ran a critical story about fewer on-site audits performed by the Norwegian Tax Administration during the COVID-19 crisis.<sup>7</sup> We provide this information to our respondents with the following text (translated from Norwegian):

---

<sup>7</sup>"Færre kontroller av svindel og dagpengejuks," *Dagens Næringsliv*, April 19, 2020.

The media has also revealed that the Norwegian Tax Administration has carried out fewer on-site audits **during the coronavirus crisis because the Tax Administration’s employees were working from home and because of infection control measures.**

The information treatment was naturally embedded in the introductory text about the Business Compensation Scheme shown to all respondents (see Section .C.2 and Section .C.3 of the for screenshots of the introductory text as presented to respondents in the control and the treatment group, respectively).

The purpose of the information treatment was to give treated respondents a signal about lower control activity by the tax administration during the coronavirus crisis. A common concern about information experiments is that the information provision might induce experimenter demand effects—a bias that occurs if respondents adjust their behavior to align with perceived researcher expectations.<sup>8</sup> To mitigate concerns about demand effects, we naturally integrated the information treatment in the introductory text about the Business Compensation Scheme and framed the information treatment as additional finding revealed by the media. Importantly, respondents in both the treatment and control group were primed on the fact that there was scope to abuse the Business Compensation Scheme.

---

<sup>8</sup>Recent work suggests that experimenter demand effects are unlikely to be a concern in survey experiments, even for strongly framed treatments (de Quidt et al., 2018; Mummolo and Peterson, 2019).

## **Measuring support for the Business Compensation Scheme**

To assess how the information treatment affects policy preferences, we ask respondents the following question: “Are you in favor of or opposed to the Business Compensation Scheme?” Respondents report their answer on a five-point scale from (1) “Strongly opposed to the Business Compensation Scheme” to (5) “Strongly in favor of the Business Compensation Scheme.”

## **Mechanism questions**

After the main question about support for the , we ask respondents three additional questions to assess mechanisms and check whether the treatment successfully changed beliefs and trust in the tax administration.

**Trust in the tax administration** We first assess whether the treatment affects trust in the tax administration with the following question: “How low or how high is your trust in the Norwegian Tax Administration to manage the Business Compensation Scheme in an effective and sensible way? ” Respondents report their answer on a five-point scale from (1) “Very low trust” to (5) “Very high trust.”

**Beliefs about fraud attempts** We next assess whether the treatment affects beliefs about fraud attempts with the following question: “What percentage of the enterprises applying for a subsidy do you think will try to abuse the scheme by reporting too high fixed costs to the Tax Administration?” Respondents report their answer by moving a slider between 0 and 100 percent with intervals of ten percentage points (Section .C.4 of the provides

a screenshot of the slider).

**Beliefs about the detection probability** We finally assess whether the treatment affects beliefs about the detection probability with the following question: “What percentage of the enterprises that are trying to abuse the scheme do you think will be detected by the Tax Administration’s checks and audits?” Respondents again report their answer by moving a slider between 0 and 100 percent with intervals of ten percentage points.

## 4.3 Results

### 4.3.1 Descriptive evidence on beliefs and policy preferences

We first focus on control group respondents to provide descriptive evidence on the association between support for the Business Compensation Scheme and beliefs about fraud attempts and the detection probability. As shown in Figure 2a, control group respondents are very supportive of the : 80 percent are either in favor or strongly in favor of the scheme and only four percent are either opposed or strongly opposed to the scheme. Furthermore, beliefs about fraud attempts (Figure 2c) and beliefs about the detection probability (Figure 2d) are both quite heterogeneous.

Figure 4.1 shows that beliefs about the detection probability and about fraud attempts are both very predictive of support for the . For instance, respondents who support or strongly support the Business Compensation Scheme think the detection probability for abuse of the Business Compensation Scheme is 12.5 percentage points higher than respondents who do not sup-

port the scheme. Similarly, respondents who support or strongly support the Business Compensation Scheme think that fraud attempts are 15.9 percentage points less likely than respondents who do not support the scheme. These correlations are robust to including demographic and political party controls in a regression framework (as shown in Table 3 of the ).<sup>9</sup> But, naturally, due to concerns about omitted variable bias and reverse causality, these correlations are only suggestive and cannot be given a causal interpretation.

### 4.3.2 Treatment effects

To examine treatment effects, we estimate the following OLS equation:

$$y_i = \alpha_0 + \alpha_1 T_i + \alpha_2 \mathbf{x}_i + \varepsilon_i$$

where  $y_i$  is the outcome of interest;  $T_i$  is an indicator for whether subject  $i$  received the information treatment;  $\mathbf{x}_i$  is a vector of controls; and  $\varepsilon_i$  is an individual-specific error term. We use robust standard errors for all specifications (HC<sub>1</sub>; MacKinnon and White, 1985).

Table 4.1 presents the main regression results on our post-treatment outcomes. To assess support for the , we create an indicator variable that takes the value one for respondents who are either in favor or strongly in favor of the scheme, and zero otherwise. We also create an indicator value one for respondents who report “very high trust” or “somewhat high trust” in the Norwegian Tax Administration.<sup>10</sup> We do not transform responses to the post-treatment beliefs questions that were elicited on a 0–100 percent scale as responses to

<sup>9</sup>Figure 3 of the shows similarly strong correlations between support for the Business Compensation Scheme and trust in the tax administration.

<sup>10</sup>Table 4 presents an alternative specification where we instead z-score these two variables.

these questions already have an intuitive interpretation.

**Beliefs about the detection probability** Column 1 of Table 4.1 shows that the treatment significantly affects beliefs about the detection probability. Specifically, treated respondents think that the tax administration is 4.6 percentage points less likely to catch firms that abuses the Business Compensation Scheme (). The treatment effect corresponds to a ten percent decrease from the control group mean of 45 percent. This result is robust to inclusion of demographic and political controls (column 2) as well as population weights (column 3). These results demonstrate that treated respondents updated their beliefs from the information provided and concluded that fewer on-site audits would make it easier for firms to abuse the Business Compensation Scheme without being detected by the tax administration.

**Beliefs about fraud attempts** Since the propensity for fraud could depend on the detection probability, treated respondents might infer that firms should be more likely to abuse the Business Compensation Scheme when it is known that the tax administration conducts fewer audits. Columns 4–6 of Table 4.1 show that this is not the case: treated respondents are no more likely than control group respondents to think that firms will abuse the . Thus, our respondent’s beliefs are not in line with the standard Allingham and Sandmo (1972) model of tax evasion in which firms would respond to fewer audits with more fraud attempts.

**Trust in the tax administration** Columns 7–10 of Table 4.1 show that the treatment significantly reduced trust in the tax administration. As shown in column 7, treated respondents are 9.6 percentage points less likely than non-

treated respondents to trust the tax administration to manage the Business Compensation Scheme in an ‘effective and sensible way’ (). This effect corresponds to a 14 percent reduction in trust compared to the control group mean of 67.6 percent, or 18.5 percent of a standard deviation if we use a z-scored outcome measure (as reported in Table 4). The result is robust to the inclusion of controls and survey weights (columns 2–3). While previous studies often see control activities and trust as separate means to achieve high tax compliance (Kirchler et al., 2008), this result demonstrates that trust is directly affected by control activities.

**Support for the Business Compensation Scheme** Columns 10–12 of Table 4.1 show that the treatment significantly affects support for the . Specifically, treated respondents are 5.6 percentage points less likely to support the Business Compensation Scheme (). This corresponds to a 7 percent reduction in support compared to the control group mean of 80 percent, or 9.7 percent of a standard deviation if we use a z-scored outcome measure (as reported in Table 4).<sup>11</sup> This result is robust to the inclusion of controls and survey weights (columns 11–12).

### 4.3.3 Discussion

Our main result is that treated respondents are 5.6 percentage points less likely than non-treated respondents to support the . Is this a small or large effect size? If we assume an exclusion restriction in which the information

---

<sup>11</sup>As shown in Figure 2a, the treatment mainly affects attitudes by making some people who “favor” the scheme become “neutral.” This translates into a 0.069 change on the five-point scale ranging from (1) “strongly oppose” to (5) “strongly favor.” This effect size thus masks an important shift in attitudes from a political economy perspective, making the binary outcome our preferred specification.



treatment only affects policy preferences through changes in beliefs about the detection probability, we could conclude that the elasticity between beliefs and preferences is close to one given a “first stage” on beliefs of 4.6 percentage points. However, the exclusion restriction is unlikely to be strictly satisfied in our setting. For instance, we also see 9.6 percentage points decrease in trust toward the tax administration. This could have an independent effect on support for the , violating the exclusion restriction. The main point is that compared to how strongly the treatment affected beliefs about the detection probability and trust in the tax administration, the effect size on support for the Business Compensation Scheme is sizable. In line with the descriptive evidence from Section 4.3.1, concerns about the tax administration’s detection capacity seem important for understanding public support for economic relief programs.

The effect size of 5.6 percentage points is also rather large compared to many previous information experiments studying policy preferences, especially when taking into account that our information treatment was short and neutrally framed. For instance, in the context of policy preferences on redistribution, information experiments studying the role of mobility perceptions or beliefs about income inequality have found almost no impact of information on policy preferences despite sizable treatment effect on underlying beliefs (Alesina et al., 2018; Kuziemko et al., 2015). One explanation for why our information treatment had a sizable impact on policy views could be that support for economic relief programs is less driven by ideology than support for redistribution. In fact, we observe no differences in support for the Business Compensation Scheme between left-wing and right-wing voters, which indeed suggests that ideology is not very important in explaining support for economic relief programs.

## 4.4 Concluding remarks

This paper presents evidence that information about fewer audits to detect abuse of a large-scale economic relief program reduces public support for the program. The results are consistent with treated respondents expecting a lower detection probability for firms trying to abuse the program and displaying less trust in the tax administration.

During an economic crisis, governments often want to reduce audit activities to ease the administrative burden on businesses. In fact, during the COVID-19 pandemic, many tax administrations were publicly announcing that they were reducing audit activities to ease the administrative burden on businesses.<sup>12</sup> Furthermore, during the peak of the pandemic, tax administrations around the world were also encouraged to reduce physical audits for public health reasons. However, as our results indicate, an unintended consequence of reduced audit activities is lower trust in the tax administration. An important lesson for policy makers is thus to be aware that fewer audits, while possibly justified on economic and public health grounds, can negatively affect public trust in the system.

Furthermore, during an economic crisis, governments also want to pass economic relief packages to save jobs and prevent unnecessary bankruptcies. Our results demonstrate the importance of maintaining normal audit levels during a crisis to secure public support for economic relief programs. This finding is especially relevant for countries in which the media is likely to report about reduced control activities, as was the case in Norway during

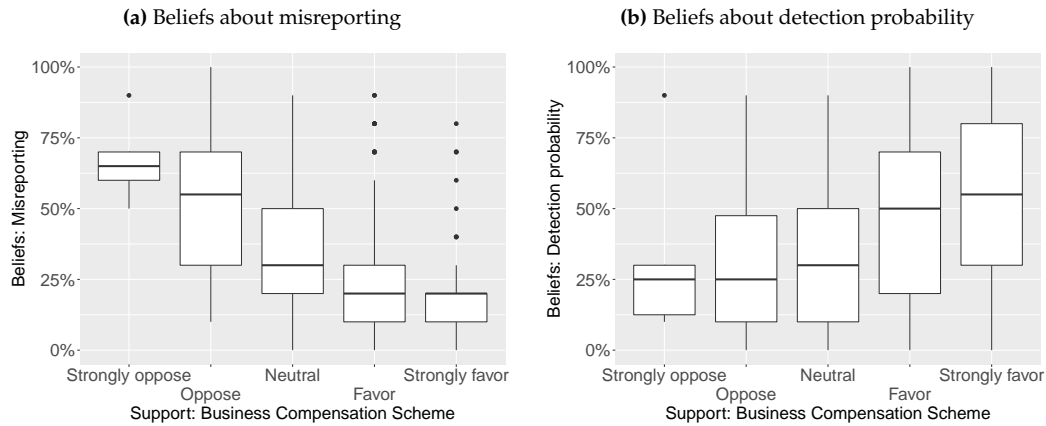
---

<sup>12</sup>See e.g. IRS's "People First" program, [www.irs.gov/newsroom/irs-unveils-new-people-first-initiative-covid-19-effort-temporarily-adjusts-suspends-key-compliance-program](https://www.irs.gov/newsroom/irs-unveils-new-people-first-initiative-covid-19-effort-temporarily-adjusts-suspends-key-compliance-program) (accessed December 2, 2020).

the peak of the COVID-19 pandemic when the media featured critical stories about fewer on-site audits performed by the Norwegian Tax Administration. An important question for future research is whether policy makers can adopt a more balanced communication strategy to preserve trust in the system despite conducting fewer on-site audits, e.g. by promising to intensify audit activities post-crisis or compensate for fewer on-site audits with less burdensome digital controls.

## Figures and tables

**Figure 4.1:** The association between beliefs and support for the



Notes: This figure uses data from control group respondents. The horizontal axis features responses to the question “Are you in favor of or opposed to the Business Compensation Scheme?”. The vertical axis of panel a) features responses to the question “What percentage of the enterprises applying for a subsidy do you think will try to abuse the scheme by reporting too high fixed costs to the Tax Administration?”. The vertical axis of panel b) features responses to the question “What percentage of the enterprises that are trying to abuse the scheme do you think will be detected by the Tax Administration’s checks and audits?”. The horizontal black lines indicate median values while the boxes display the interquartile ranges (i.e., the upper and lower part of the boxes corresponds to the 25th and 75th percentile, respectively). The upper and lower whiskers include all values up to 1.5 times the inter-quartile range. Finally, values outside the whiskers are outliers represented by individual dots.

Table 4.1: Main post-treatment outcomes

	Beliefs: Detection probability		Beliefs: Misreporting			Trust: Tax Administration			Support: Business Comp. Scheme			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment	-0.046*** (0.016)	-0.044*** (0.016)	-0.044*** (0.016)	0.014 (0.010)	0.011 (0.010)	0.013 (0.011)	-0.096*** (0.026)	-0.089*** (0.025)	-0.087*** (0.026)	-0.056** (0.022)	-0.048** (0.022)	-0.047** (0.022)
N	1341	1341	1341	1336	1336	1336	1400	1400	1400	1400	1400	1400
Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Weights	No	No	Yes	No	No	Yes	No	No	Yes	No	No	Yes
Control mean	0.45	0.45	0.45	0.27	0.27	0.27	0.68	0.68	0.68	0.80	0.80	0.80

Note: The table shows OLS regression results on our main post-treatment outcomes. In columns 1–3, the dependent variable is beliefs about the percentage of enterprises that will abuse the Business Compensation Scheme by reporting too high fixed costs (responses range from 0 to 1 in 0.1 increments). The dependent variable in columns 4–6 is beliefs about the percentage of firms trying to abuse the Business Compensation Scheme conditional on applying for a subsidy (responses range from 0 to 1 in 0.1 increments). The dependent variable in columns 7–9 is an indicator variable taking the value 1 for respondents who express “Very high trust” or “somewhat high trust” in the Norwegian Tax Administration’s handling of the Business Compensation Scheme and 0 for the remaining three response options (“Very low trust,” “Somewhat low trust,” and “Neither low nor high trust”). The dependent variable in columns 10–12 is an indicator variable taking the value 1 for respondents who are “In favor” or “Strongly in favor” of the Business Compensation Scheme and 0 for the remaining three response options (“Strongly opposed,” “Opposed,” and “Neither in favor nor opposed”). “Treatment” takes the value one for respondents who received information about fewer audits. Regressions with controls include the following variables: gender, age (in years), education (dummy for having a college degree), income (dummy for having income above NOK 500,000), employment (dummy for being full-time employed), work sector (dummy for working in the public sector), and political party preferences (a dummy for supporting one of the main left-wing parties, R, SV, or Ap, and a dummy for supporting one of the main right-wing parties, H or Frp). Regressions with probability weights make the sample fully representative of the adult Norwegian population on gender, age, and geography. “Control mean” displays the (unweighted) mean of the dependent variable for control group respondents.

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . Robust standard errors in parentheses.

For online publication only:  
**Information about Fewer Audits Reduces Support  
for Economic Relief Programs**

Ingar Haaland and Andreas Olden

Table 2 provides summary statistics , comparing the general Norwegian population with our (unweighted) Norstat sample on some key demographics. Table 3 provides descriptive evidence on the association between support for the and beliefs about fraud attempts and the detection probability. Table 4 shows treatment effects on policy preferences and trust in the tax administration using z-scored outcome measures. Figure 2 shows the distribution of responses to our main outcome questions by treatment status. Figure 3 shows correlations between the support for the and trust in the tax administration. Section .B provides instructions translated into English. Section .C provides screenshots of the original survey in Norwegian. Section .D provides a copy of the pre-analysis plan.

## .A Appendix tables and figures

**Table 2:** Summary statistics: General population versus Norstat sample

	General population	Norstat
Male	0.502	0.489
Age (in years)	47.236	46.562
Income	567480	510185.185
College degree	0.346	0.601
Observations		1400

*Note:* This table compares summary statistics of the general adult population in Norway (recovered from <https://www.ssb.no/statbank/>) and our unweighted Norstat sample (column 2). To calculate average income in our survey, we transformed the income brackets into a continuous variable using the midpoint of the answer choice given by the respondents. “College” is a dummy for having a college degree.

**Table 3:** The association between support for the and beliefs about detection probability and fraud attempts

	Detection probability		Misreporting	
	(1)	(2)	(3)	(4)
Support: Economic relief	0.125*** (0.025)	0.110*** (0.026)	-0.159*** (0.021)	-0.148*** (0.022)
N	668	668	670	670
Demographic and political controls	No	Yes	No	Yes
Mean	0.35	0.35	0.40	0.40

*Note:* The table shows OLS regression results using control group respondent only. In columns 1–2, the dependent variable is beliefs about the detection probability for abuse of the . In columns 3–4, the dependent variable is beliefs about the percentage of firms trying to misuse the . “Support: Economic relief” is an indicator taking the value one for respondents who support for the . Controls are listed in Table 4.1.

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . Robust standard errors in parentheses.

**Table 4:** Treatment effects with z-scored outcomes

	Trust: Tax Administration			Support: Business Comp. Scheme		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.185*** (0.054)	-0.171*** (0.054)	-0.164*** (0.056)	-0.097* (0.055)	-0.077 (0.053)	-0.070 (0.057)
N	1400	1400	1400	1400	1400	1400
Controls	No	Yes	Yes	No	Yes	Yes
Weights	No	No	Yes	No	No	Yes

*Note:* The table shows OLS regression results on trust in the Norwegian Tax Administration (columns 1–3) and support for the (columns 4–6). Both dependent variables were elicited using five-point Likert scales and have then been z-scored using the mean and standard deviation of control group respondents. “Treatment” takes the value one for respondents who received information about fewer audits. Controls are listed in Table 4.1. Regressions in columns 3 and 6 include probability weights for gender, age, and geography.

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ . Robust standard errors in parentheses.

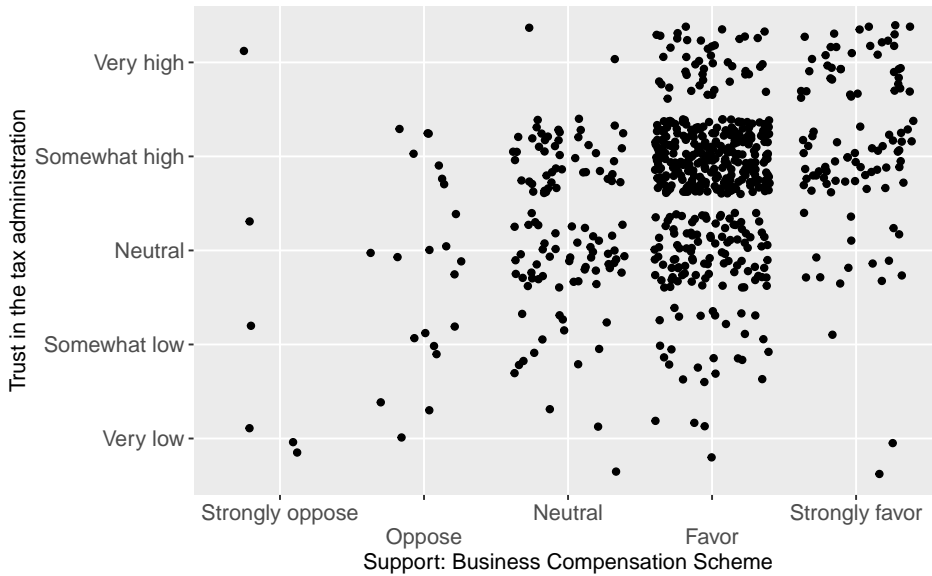


**Figure 2: Distribution of responses**



Notes: The figure shows the distribution of responses to the main outcome questions, by treatment status.

**Figure 3: The association between trust in the tax administration and support for the**



Notes: This figure uses data from control group respondents. The horizontal axis features responses to the question “Are you in favor of or opposed to the Business Compensation Scheme?” and the vertical axis features responses to the question “How low or how high is your trust in the Norwegian Tax Administration to manage the in an effective and sensible way?”

## **.B English translation of experimental instructions**

### **.B.1 Pre-treatment background questions**

- Are you ... [Male; Female]
- How old are you? [Numeric]
- Where do you live? [Numeric; Postal code]
- What is the highest degree or level of school you have completed? [Primary and lower secondary school; Upper secondary school; University/college up to and including 3 years (bachelor's degree or similar); University/college up to and including 4 years; University/college over 4 years (master's degree or similar and higher)]

### **.B.2 Introduction**

Recently, the Norwegian Government launched the Business Compensation Scheme, often referred to as the cash benefit scheme for businesses. The Business Compensation Scheme was established to provide financial aid to enterprises that have been severely impacted financially by the coronavirus crisis. The aid is provided through subsidies that cover up to 90 percent of the enterprises' fixed costs, for example their rent.

The purpose of the Business Compensation Scheme is to avoid unnecessary bankruptcies and redundancies during the coronavirus crisis. **An estimate shows that the scheme will cost the state around NOK 20 billion per month.**

It has been pointed out by the media that the Business Compensation Scheme may be abused by enterprises reporting too high fixed costs to the tax authorities.

The Norwegian Tax Administration has been charged with administration of the Business Compensation Scheme on the state's behalf. **The Norwegian Tax Administration is also responsible for ensuring that the scheme is not misused.**

*[The following paragraph was only shown to respondents in the treatment group:*  
The media has also revealed that the Norwegian Tax Administration has carried out fewer on-site audits **during the coronavirus crisis because the Tax Administration's employees were working from home and because of infection control measures.***]*

### **.B.3 Outcome questions**

#### **Support for the Business Compensation Scheme**

Are you in favor of or opposed to the Business Compensation Scheme?

- Strongly opposed to the Business Compensation Scheme
- Opposed to the Business Compensation Scheme
- Neither in favor nor opposed to the Business Compensation Scheme
- In favor of Business Compensation Scheme
- Strongly in favor of the Business Compensation Scheme

## **Trust in the Norwegian Tax Administration**

How low or how high is your trust in the Norwegian Tax Administration to manage the Business Compensation Scheme in an effective and sensible way?

- Very low trust
- Somewhat low trust
- Neither low nor high trust
- Somewhat high trust
- Very high trust

## **Beliefs about tax fraud**

What percentage of the enterprises applying for a subsidy do you think will try to abuse the scheme by reporting too high fixed costs to the Tax Administration? [Slider from 0 to 100 with intervals of ten percentage points; respondents also had a “I do not wish to answer” option]

## **Beliefs about the detection probability**

What percentage of the enterprises that are trying to abuse the scheme do you think will be detected by the Tax Administration’s checks and audits? [Slider from 0 to 100 with intervals of ten percentage points; respondents also had a “I do not wish to answer” option]

## **.B.4 Post-treatment background questions**

- What is your personal gross annual income, i.e. income before tax? [0 to 199,999; 200,000 to 399,999; 400,000 to 599,999; 600,000 to 799,999; 800,000 to 999,000; 1 million or more; I do not know / I cannot remember; I do not wish to answer]
- How would you describe your situation? If more than one option is correct, choose the one you think fits best. [I am a student; Full-time employee; Part-time employee; I have my own business; I am in military service; Maternity/paternity leave; I am a pensioner; I am looking for work; I am a homemaker; I have been laid off; I am a benefit recipient; Other; I do not wish to answer]
- Which sector do you work in? [Private; Public; Other; *only shown if respondent is in paid employment*]
- If a parliamentary election was held tomorrow, which political party would you vote for? [Ap; H; FrP; Sp; SV; V; KrF; MDG; Rødt; Folkeaksjonen nei til bompenger; Other; I would not vote; I do not wish to answer; I am not sure; I am not entitled to vote; *order randomized for all the political parties*]

## **.B.5 Comments and concluding remarks**

### **Open-ended question for comments and feedback**

If you have any comments to this survey, please write your comment in the field below. We would especially like to hear from you if anything was

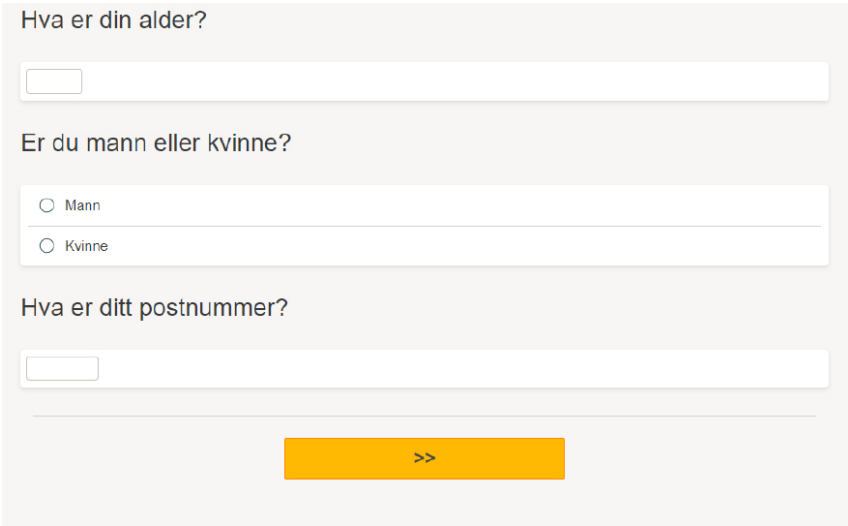
unclear or if there was anything special you reacted to during the survey.

### **Concluding remarks – sent to all participants after answering all questions**

Thank you for taking the time to participate in this survey. Your answers are important. Most of the Tax Administration's audits are performed digitally, but we did have a period of fewer on-site audits as a result of the Covid-19 outbreak. The Tax Administration will increase the number of audits and checks from now on, including for circumstances arising during the Covid-19 lockdown.

## **.C Screenshots of the experiment in Norwegian**

### **.C.1 Pre-treatment questions**



Hva er din alder?

Er du mann eller kvinne?

Mann

Kvinne

Hva er ditt postnummer?

>>

## .C.2 Main outcome screen: Control group

Regjeringen innførte nylig kompensasjonsordningen, ofte kalt kontantstøtten til næringslivet. Kompensasjonsordningen skal hjelpe bedrifter som er økonomisk hardt rammet av koronakrisen. Dette gjøres ved å dekke inntil 90 prosent av de faste utgiftene til bedriftene, for eksempel husleie.

Målet med kompensasjonsordningen er å unngå unødvendige konkurser og oppsigelser under koronakrisen. **Ordningen er anslått å koste staten cirka 20 milliarder kroner i måneden.**

Det har blitt påpekt i mediene at kompensasjonsordningen kan misbrukes av bedrifter ved å innrapportere for høye faste kostnader til skattemyndighetene.

Skatteetaten forvalter kompensasjonsordningen for staten. **Skatteetaten har også ansvar for kontrollere at ordningen ikke misbrukes.**

Er du for eller imot kompensasjonsordningen?

- Sterkt imot kompensasjonsordningen
- Imot kompensasjonsordningen
- Verken for eller imot kompensasjonsordningen
- For kompensasjonsordningen
- Sterkt for kompensasjonsordningen

>>

### .C.3 Main outcome screen: Treatment group

Regjeringen innførte nylig kompensasjonsordningen, ofte kalt kontantstøtten til næringslivet. Kompensasjonsordningen skal hjelpe bedrifter som er økonomisk hardt rammet av koronakrisen. Dette gjøres ved å dekke inntil 90 prosent av de faste utgiftene til bedriftene, for eksempel husleie.

Målet med kompensasjonsordningen er å unngå unødvendige konkurser og oppsigelser under koronakrisen. **Ordningen er anslått å koste staten cirka 20 milliarder kroner i måneden.**

Det har blitt påpekt i mediene at kompensasjonsordningen kan misbrukes av bedrifter ved å innrapportere for høye faste kostnader til skattemyndighetene.

Skatteetaten forvalter kompensasjonsordningen for staten. **Skatteetaten har også ansvar for kontrollere at ordningen ikke misbrukes.**

Det har også kommet fram i mediene at Skatteetaten har gjennomført færre fysiske kontroller av bedrifter **gjennom koronakrisen på grunn av hjemmekontor og smittevernhensyn.**

Er du for eller imot kompensasjonsordningen?

- Sterkt imot kompensasjonsordningen
- Imot kompensasjonsordningen
- Verken for eller imot kompensasjonsordningen
- For kompensasjonsordningen
- Sterkt for kompensasjonsordningen

>>



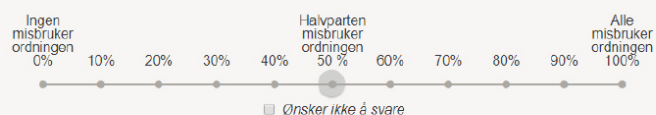
## .C.4 Post-treatment beliefs and trust in government

Hvor høy eller lav tillit har du til at Skatteetaten forvalter kompensasjonsordningen på en god og fornuftig måte?

- Svært lav tillit
- Ganske lav tillit
- Verken lav eller høy tillit
- Ganske høy tillit
- Svært høy tillit

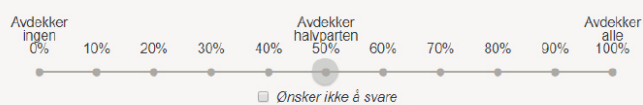
>>

Hvor mange prosent av bedriftene som søker om støtte tror du vil prøve å misbruke ordningen ved å innrapportere for høye kostnader til Skatteetaten?



>>

Hvor mange prosent av bedriftene som prøver å misbruke ordningen tror du Skatteetaten avdekker ved hjelp av sine kontroller?



>>

## .C.5 Additional demographics and political views

Hva er din høyeste fullførte utdanning?

- Grunnskole
- Videregående
- Universitet-/høyskolenivå 1 o.m. 3 år (Bachelor eller tilsvarende)
- Universitet-/høyskolenivå 1.o.m. 4 år
- Universitet-/høyskolenivå mer enn 4 år (Mastergrad eller tilsvarende og høyere grad)
- Annet

[>>](#)

Hva er din personlige inntekt (før skatt)?

- 0-100.000 NOK
- 100.001-200.000 NOK
- 200.001-300.000 NOK
- 300.001-400.000 NOK
- 400.001-500.000 NOK
- 500.001-600.000 NOK
- 600.001-700.000 NOK
- 700.001-800.000 NOK
- 800.001-900.000 NOK
- 900.001-1000.000 NOK
- 1.000.001-1.100.000 NOK
- 1.100.001-1.300.000 NOK
- 1.300.001-1.500.000 NOK
- 1.500.001 NOK eller mer
- Vil ikke svare
- Vet ikke

[>>](#)

### Hvordan vil du beskrive din daglige situasjon?

Dersom det er flere alternativ som passer, velger du det som ut fra din egen mening stemmer best.

Studier

Heltidsansatt

Deltidsansatt

Jobber i eget firma

Militærtjeneste/sivil tjeneste

Fødselspermisjon

Pensjonert

Arbeidssøker

Hjemmeværende

Permittert

Trygdet

>>

### Hvilken sektor jobber du i?

Offentlig











Privat

Er ikke i arbeid

Annet

>>

Dersom det var Stortingsvalg i morgen, hvilket parti ville du da stemme på?

-  Frp
-  Senterpartiet Sp
-  Folkeaksjonen nei til mer bompenger (FNB)
-  SV
-  Krf
-  Mdg
-  Høyre
-  Ap
-  Rødt
-  Venstre
- Andre:
- Ville ikke stemme
- Vil ikke si
- Ikke sikker
- Har ikke stemmerett

>>

## .C.6 Comments and debrief

Hvis du har noen kommentarer til denne undersøkelsen, setter vi pris på om du skriver en liten kommentar i boksen under. Vi setter spesielt pris på å høre fra deg hvis noe var uklart eller hvis det var noe spesielt du reagerte på i løpet av undersøkelsen.

>>

Takk for at du tok deg tid til å svare på spørreundersøkelsen. Dine svar er viktige for oss. De fleste av kontrollene til Skatteetaten er digitale, men vi har hatt en periode med færre fysiske kontroller som følge av Covid-19. Skatteetaten skal øke kontrollvirksomheten fremover, også på forhold som oppstod under Covid-19 nedstengningen.

---

>>

# .D Pre-analysis plan

**CONFIDENTIAL - FOR PEER-REVIEW ONLY**



## Support for Economic Relief and Beliefs about Tax Enforcement Capacity (#41206)

Created: 05/18/2020 04:59 AM (PT)

Shared: 05/22/2020 02:38 AM (PT)

---

This pre-registration is not yet public. This anonymized copy (without author names) was created by the author(s) to use during peer-review. A non-anonymized version (containing author names) will become publicly available only if an author makes it public. Until that happens the contents of this pre-registration are confidential.

---

**1) Have any data been collected for this study already?**

It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

**2) What's the main question being asked or hypothesis being tested in this study?**

We test whether beliefs about the tax administration's capacity to detect tax fraud affect public support for an economic relief bill (the "Business Compensation Scheme") for Norwegian enterprises with a significant drop in revenue due to the coronavirus situation.

**3) Describe the key dependent variable(s) specifying how they will be measured.**

The key dependent variable is support for the Business Compensation Scheme. It is measured on a 5-point scale from 1: "Strongly against the scheme" to 5: "Strongly in favor of the scheme" (translation from Norwegian). We will z-score the variable using the mean and standard deviation from control group respondents. We will also create a dummy that takes the value one for respondents who are either in favor (value 4 on the 5-point scale) or strongly in favor (value 5 on the 5-point scale) of the scheme.

We also assess treatment effects on the following three mechanism questions:

1) Trust in how well the Norwegian Tax Administration handles the Business Compensation Scheme (measured on a 5-point scale from 1: Very low trust to 5: Very high trust)

2) Beliefs about prevalence of fraud attempts among businesses applying to the scheme (measured on an 11-point scale from 0% to 100%)

3) Beliefs about how many fraud attempts the Norwegian Tax Administration is able to detect (measured on an 11-point scale from 0% of cases to 100% of cases)

**4) How many and which conditions will participants be assigned to?**

Two conditions.

The treatment group is informed that Norwegian Tax Authority has completed fewer physical controls during COVID-19 pandemic.

The control group is not informed about this.

**5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.**

The main analysis will be a linear regression of support for the Business Compensation Scheme on a treatment indicator taking the value one for respondents in the treatment group and zero otherwise. We include controls for gender, age, education, income, employment status and sector of employment, and political party preferences.

**6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.**

We will not exclude any respondents.

**7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.**

We have ordered a sample of 1400 respondents from the data collection agency (Norstat).

**8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)**

Norstat has already started to collect the data, but we do not get access to the data before the data collection is finished. This pre-registration was submitted before the data collection was finished and thus before we got access to the data.

# References

- Abadie, Alberto**, "Semiparametric difference-in-differences estimators," *The Review of Economic Studies*, 2005, 72 (1), 1–19.
- Adamopoulos, Panagiotis, Vilma Todri, and Anindya Ghose**, "Demand effects of the internet-of-things sales channel: Evidence from automating the purchase process," *Information Systems Research*, 2020, 32 (1), 238–267.
- Ai, J., P. L. Brockett, L. L. Golden, and Montserrat Guillén**, "A robust unsupervised method for fraud rate estimation," *The Journal of Risk and Insurance*, 2013, 80 (1), 121–143.
- Ai, Jing, Linda L Golden, and Patrick L Brockett**, "Assessing consumer fraud risk in insurance claims: An unsupervised learning technique using discrete and continuous predictor variables," *North American Actuarial Journal*, October 2009, 13 (4), 438–458.
- Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso**, "Intergenerational Mobility and Preferences for Redistribution," *American Economic Review*, 2018, 108 (2), 521–554.
- Allingham, Michael G and Agnar Sandmo**, "Income tax evasion: A theoretical analysis," *Journal of public economics*, 1972, 1 (3-4), 323–338.
- Alm, James**, "Testing behavioral public economics theories in the laboratory," *National Tax Journal*, 2010, 63 (4), 635–658.

- , “What motivates tax compliance?,” *Journal of Economic Surveys*, 2019, 33 (2), 353–388.
- **and Steven M Sheffrin**, “Using behavioral economics in public economics,” *Public Finance Review*, 2017, 45 (1), 4–9.
- Angrist, Joshua D and Jörn-Steffen Pischke**, *Mostly Harmless Econometrics: An Empiricist’s Companion*, Princeton University Press, 2008.
- Arellano, Manuel et al.**, “Computing robust standard errors for within-groups estimators,” *Oxford bulletin of Economics and Statistics*, 1987, 49 (4), 431–434.
- Ariely, Dan and Jonathan Levav**, “Sequential choice in group settings: Taking the road less traveled and less enjoyed,” *Journal of Consumer Research*, 2000, 27 (3), 279–290.
- Artís, Manuel, Mercedes Ayuso, and Montserrat Guillén**, “Modelling different types of automobile insurance fraud behaviour in the Spanish market,” *Insurance: Mathematics and Economics*, 1999, 24, 67–81.
- Artís, Manuel, Mercedes Ayuso, and Montserrat Guillén**, “Detection of automobile insurance fraud with discrete choice models and misclassified claims,” *Journal of Risk and Insurance*, 2002, 69 (3), 325–340.
- Athey, Susan and Guido W Imbens**, “Identification and inference in nonlinear difference-in-differences models,” *Econometrica*, 2006, 74 (2), 431–497.
- Belnap, Andrew, Jeffrey L Hoopes, Edward L Maydew, and Alex Turk**, “Real effects of tax audits: Evidence from firms randomly selected for IRS examination,” *Available at SSRN 3437137*, 2020.



- Berck, Peter and Sofia B Villas-Boas**, “A note on the triple difference in economic models,” *Applied Economics Letters*, 2016, 23 (4), 239–242.
- Bergé, Laurent**, “Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm,” *CREA Discussion Papers*, 2018, (13).
- Bernheim, B Douglas and Dmitry Taubinsky**, “Behavioral public economics,” *Handbook of Behavioral Economics: Applications and Foundations 1*, 2018, 1, 381–516.
- Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan**, “How much should we trust differences-in-differences estimates?,” *Quarterly Journal of Economics*, 2004, 119 (1), 249–275.
- Blumenthal, Marsha, Charles Christian, Joel Slemrod, and Matthew G Smith**, “Do normative appeals affect tax compliance? Evidence from a controlled experiment in Minnesota,” *National Tax Journal*, 2001, pp. 125–138.
- Borck, Rainald**, “Voting on redistribution with tax evasion,” *Social Choice and Welfare*, 2009, 32 (3), 439–454.
- Bott, Kristina M, Alexander W Cappelen, Erik Ø Sørensen, and Bertil Tungodden**, “You’ve got mail: A randomized field experiment on tax evasion,” *Management Science*, 2019, 66 (7), 2801–3294.
- Brockett, P. L., R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert**, “Fraud classification using principal component analysis of RIDITs,” *The Journal of Risk and Insurance*, 2002, 69 (3), 341–371.
- Callaway, Brantly and Pedro HC Sant’Anna**, “Difference-in-differences with multiple time periods,” *Journal of Econometrics*, 2020, p. In Press.

- Cameron, A Colin and Douglas L Miller**, "A practitioner's guide to cluster-robust inference," *Journal of Human Resources*, 2015, 50 (2), 317–372.
- Caudill, Steven B, Mercedes Ayuso, and Montserrat Guillén**, "Fraud detection using a multinomial logit model with missing information," *Journal of Risk and Insurance*, 2005, 72 (4), 539–550.
- Chandrasekhar, Arun G, Benjamin Golub, and He Yang**, "Signaling, shame, and silence in social learning," *National Bureau of Economic Research (NBER)*, 2018.
- Chetty, Raj, Adam Looney, and Kory Kroft**, "Salience and Taxation: Theory and Evidence," *American Economic Review*, 2009.
- Conley, Timothy G and Christopher R Taber**, "Inference with "difference in differences" with a small number of policy changes," *The Review of Economics and Statistics*, 2011, 93 (1), 113–125.
- Cruces, Guillermo, Ricardo Perez-Truglia, and Martin Tetaz**, "Biased perceptions of income distribution and preferences for redistribution: Evidence from a survey experiment," *Journal of Public Economics*, 2013, 98, 100–112.
- Dabholkar, Pratibha A, L Michelle Bobbitt, and Eun-Ju Lee**, "Understanding consumer motivation and behavior related to self-scanning in retailing: implications for strategy and research on technology-based self-service," *International Journal of Service Industry Management*, 2003, 14 (1), 59–95.
- Dahl, Darren W, Gerald J Gorn, and Charles B Weinberg**, "The impact of embarrassment on condom purchase behaviour," *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique*, 1998, pp. 368–370.

– , **Rajesh V Manchanda, and Jennifer J Argo**, “Embarrassment in consumer purchase: The roles of social presence and purchase familiarity,” *Journal of Consumer Research*, 2001, 28 (3), 473–481.

**De Neve, Jan-Emmanuel, Clément Imbert, Johannes Spinnewijn, Teodora Tsankova, and Maarten Luts**, “How to Improve Tax Compliance? Evidence from Population-Wide Experiments in Belgium,” *Journal of Political Economy*, 2021, 129 (5), 1425–1463.

**de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth**, “Measuring and Bounding Experimenter Demand,” *American Economic Review*, 2018, 108 (11), 3266–3302.

**Dempster, Arthur P, Nan M Laird, and Donald B Rubin**, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society. Series B (methodological)*, 1977, pp. 1–38.

**Derrig, Richard A. and Krzysztof M. Ostaszewski**, “Fuzzy techniques of pattern recognition in risk and claim classification,” *The Journal of Risk and Insurance*, 1995, 62 (3), 447–482.

**Deschênes, Olivier, Michael Greenstone, and Joseph S Shapiro**, “Defensive investments and the demand for air quality: Evidence from the NOx budget program,” *American Economic Review*, 2017, 107 (10), 2958–89.

**Doerrenberg, Philipp and Jan Schmitz**, “Tax compliance and information provision – A field experiment with small firms,” Discussion paper 15–028, ZEW-Centre for European Economic Research 2015.

**Engelbrecht, Nils, Tim-Benjamin Lembcke, Alfred Benedikt Brendel, Kilian Bizer, and Lutz M Kolbe**, “The Virtual Online Supermarket: An Open-

Source Research Platform for Experimental Consumer Research," *Sustainability*, 2021, 13 (8), 4375.

**Fehr, Dietmar, Johanna Mollerstrom, and Ricardo Perez-Truglia**, "Your Place in the World: The Demand for National and Global Redistribution," Working Paper 26555, National Bureau of Economic Research December 2019.

**Ferman, Bruno and Cristine Pinto**, "Inference in differences-in-differences with few treated groups and heteroskedasticity," *Review of Economics and Statistics*, 2019, 101 (3), 452–467.

**Frölich, Markus and Stefan Sperlich**, *Impact Evaluation Treatment Effects and Causal Analysis*, Cambridge University Press, 2019.

**Gavett, Gretchen**, "How Self-Service Kiosks Are Changing Customer Behavior," *Harvard Business Review*, last accessed 2021-12-21 at <https://hbr.org/2015/03/how-self-service-kiosks-are-changing-customer-behavior>, 2015.

**George, Alison and Anne Murcott**, "Research note: Monthly strategies for discretion: Shopping for sanitary towels and tampons," *The Sociological Review*, 1992, 40 (1), 146–162.

**Goldfarb, Avi, Ryan C McDevitt, Sampsa Samila, and Brian S Silverman**, "The Effect of Social Interaction on Economic Transactions: Evidence from Changes in Two Retail Formats," *Management Science*, 2015, 61 (12), 2963–2981.

**Grigorieff, Alexis, Christopher Roth, and Diego Ubfal**, "Does Information Change Attitudes Toward Immigrants?," *Demography*, 2020, 57 (3), 1–27.

- Gruber, Jonathan**, "The incidence of mandated maternity benefits," *American Economic Review*, 1994, 84 (3), 622–641.
- **and James Poterba**, "Tax incentives and the decision to purchase health insurance: Evidence from the self-employed," *Quarterly Journal of Economics*, 1994, 109 (3), 701–733.
- Haaland, Ingar and Christopher Roth**, "Beliefs about Racial Discrimination and Support for Pro-Black Policies," *Review of Economics and Statistics*, 2021.
- , – , **and Johannes Wohlfart**, "Designing Information Provision Experiments," *Journal of Economic Literature* (Forthcoming), 2021.
- Harris-Lagoudakis, Katherine**, "Online shopping and the healthfulness of grocery purchases," *American Journal of Agricultural Economics*, 2021.
- Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton**, "Misclassification of the dependent variable in a discrete-response setting," *Journal of Econometrics*, 1998, 87, 239–269.
- Hornbeck, Richard**, "Barbed wire: Property rights and agricultural development," *Quarterly Journal of Economics*, 2010, 125 (2), 767–810.
- Hoynes, Hilary, Diane Whitmore Schanzenbach, and Douglas Almond**, "Long-run impacts of childhood access to the safety net," *American Economic Review*, 2016, 106 (4), 903–34.
- Imbens, Guido W. and Jeffrey M Wooldridge**, "What's new in econometrics? Lecture 10 difference-in-differences estimation," *NBER Summer Institute*, available at: [www.nber.org/WNE/Slides7-31-07/slides\\_10\\_diffindiffs.pdf](http://www.nber.org/WNE/Slides7-31-07/slides_10_diffindiffs.pdf), accessed April 2018, 2007, 9, 2011.

- Kahn-Lang, Ariella and Kevin Lang**, “The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications,” *Journal of Business & Economic Statistics*, 2020, 38 (3), 613–620.
- Kahneman, Daniel and Amos Tversky**, “Prospect theory: An analysis of decision under risk,” *Econometrica*, 1979, 47 (2), 363–391.
- Karadja, Mounir, Johanna Mollerstrom, and David Seim**, “Richer (and Holier) Than Thou? The Effect of Relative Income Improvements on Demand for Redistribution,” *Review of Economics and Statistics*, 2017, 99 (2), 201–212.
- Kirchler, Erich, Erik Hoelzl, and Ingrid Wahl**, “Enforced versus voluntary tax compliance: The “slippery slope” framework,” *Journal of Economic Psychology*, 2008, 29 (2), 210–225.
- Kleven, Henrik Jacobsen, Camille Landais, and Emmanuel Saez**, “Taxation and international migration of superstars: Evidence from the European football market,” *American Economic Review*, 2013, 103 (5), 1892–1924.
- , **Martin B. Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez**, “Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark,” *Econometrica*, 2011, 79 (3), 651–692.
- Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva**, “How Elastic are Preferences for Redistribution? Evidence from Randomized Survey Experiments,” *American Economic Review*, 2015, 105 (4), 1478–1508.
- Lavik, Randi and Alexander Schjoll**, “Endring i butikkstruktur og handlemønster i norsk dagligvarehandel,” 2012.

- Lechner, Michael**, “The estimation of causal effects by difference-in-difference methods,” *Foundations and Trends® in Econometrics*, 2011, 4 (3), 165–224.
- Lergetporer, Philipp, Guido Schwerdt, Katharina Werner, Martin R West, and Ludger Woessmann**, “How information affects support for education spending: Evidence from survey experiments in Germany and the United States,” *Journal of Public Economics*, 2018, 167, 138–157.
- Liang, Kung-Yee and Scott L Zeger**, “Longitudinal data analysis using generalized linear models,” *Biometrika*, 1986, 73 (1), 13–22.
- MacKinnon, James G and Halbert White**, “Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties,” *Journal of Econometrics*, 1985, 29 (3), 305–325.
- **and Matthew D Webb**, “The wild bootstrap for few (treated) clusters,” *The Econometrics Journal*, 2018, 21 (2), 114–135.
- **and –**, “Randomization inference for difference-in-differences with few treated clusters,” *Journal of Econometrics*, 2020, 218 (2), 435–450.
- Major, John A and Dan R Riedinger**, “EFD: A hybrid knowledge/statistical-based system for the detection of fraud,” *Journal of Risk and Insurance*, September 2002, 69 (3), 309–324.
- Mickeler, Maren, Pooyan Khashabi, Marco Kleine, and Tobias Kretschmer**, “Under the Radar: User Anonymity in the Design of Organizational Platforms,” *Max Planck Institute for Innovation & Competition Research Paper*, 2021, (19-17).

- Muehlenbachs, Lucija, Elisheba Spiller, and Christopher Timmins**, “The housing market impacts of shale gas development,” *American Economic Review*, 2015, 105 (12), 3633–59.
- Mummolo, Jonathan and Erik Peterson**, “Demand Effects in Survey Experiments: An Empirical Assessment,” *American Political Science Review*, 2019, 113 (2), 517–529.
- Nilsson, J Peter**, “Alcohol availability, prenatal conditions, and long-term economic outcomes,” *Journal of Political Economy*, 2017, 125 (4), 1149–1207.
- Nunan, Daniel and MariaLaura Di Domenico**, “Older consumers, digital marketing, and public policy: A review and research agenda,” *Journal of Public Policy & Marketing*, 2019, 38 (4), 469–483.
- OECD**, “Tax administration responses to COVID-19: Measures taken to support taxpayers,” Technical Report, CIAT/IOTA/OECD, Paris April 2020.
- Olden, Andreas**, “What Do You Buy When No One’s Watching? The Effect of Self-Service Checkouts on the Composition of Sales in Retail,” 2018. NHH FOR DP 3/18, Norwegian School of Economics.
- **and Jarle Møen**, “The Triple Difference Estimator,” 2020. NHH FOR DP 1/20, Norwegian School of Economics.
- Olinsky, Alan D, Paul M Mangiameli, and Shaw K Chen**, “Statistical support of forensic auditing,” *Interfaces*, November 1996, 26 (6), 95–104.
- Perez-Truglia, Ricardo and Ugo Troiano**, “Shaming tax delinquents,” *Journal of Public Economics*, 2018, 167, 120–137.



- Povey, R, M Conner, P Sparks, R James, and R Shepherd**, “Interpretations of healthy and unhealthy eating, and implications for dietary change,” *Health Education Research*, 1998, 13 (2), 171–183.
- R Core Team**, *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing 2020.
- Ravallion, Martin**, “Evaluating anti-poverty programs,” *Handbook of Development Economics*, 2007, 4, 3787–3846.
- , **Emanuela Galasso, Teodoro Lazo, and Ernesto Philipp**, “What can ex-participants reveal about a program’s impact?,” *Journal of Human Resources*, 2005, 40 (1), 208–230.
- Roine, Jesper**, “The Political Economics of Not Paying Taxes,” *Public Choice*, 2006, 126 (1-2), 107–134.
- Roth, Christopher, Sonja Settele, and Johannes Wohlfart**, “Beliefs about public debt and the demand for government spending,” *Journal of Econometrics*, 2021.
- Shayo, Moses and Asaf Zussman**, “Judicial ingroup bias in the shadow of terrorism,” *Quarterly Journal of Economics*, 2011, 126 (3), 1447–1484.
- Slemrod, Joel**, “Tax compliance and enforcement,” *Journal of Economic Literature*, 2019, 57 (4), 904–54.
- Stein, Richard I and Carol J Nemeroff**, “Moral overtones of food: Judgments of others based on what they eat,” *Personality and Social Psychology Bulletin*, 1995, 21 (5), 480–490.
- Tibshirani, Robert J and Bradley Efron**, “An introduction to the bootstrap,” *Monographs on statistics and applied probability*, 1993, 57, 1–436.

- Traxler, Christian**, "Voting over Taxes: The Case of Tax Evasion," *Public Choice*, 2009, 140 (1-2), 43–58.
- Vartanian, Lenny R, C Peter Herman, and Janet Polivy**, "Consumption stereotypes and impression management: How you are what you eat," *Appetite*, 2007, 48 (3), 265–277.
- Viaene, S., R. A. Derrig, B. Baesens, and G. Dedene**, "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection," *The Journal of Risk and Insurance*, 2002, 69 (3), 373–421.
- Walker, W Reed**, "The transitional costs of sectoral reallocation: Evidence from the clean air act and the workforce," *Quarterly Journal of Economics*, 2013, 128 (4), 1787–1835.
- White, Halbert**, *Asymptotic theory for econometricians*, Academic press, 1984.
- Wooldridge, Jeffrey M**, *Introductory econometrics: A modern approach 7e*, Cengage, 2020.
- Yelowitz, Aaron S**, "The Medicaid notch, labor supply, and welfare participation: Evidence from eligibility expansions," *The Quarterly Journal of Economics*, 1995, 110 (4), 909–939.
- Yin, Xicheng, Hongwei Wang, Qiangwei Xia, and Qican Gu**, "How social interaction affects purchase intention in social commerce: A cultural perspective," *Sustainability*, 2019, 11 (8), 2423.