NBER WORKING PAPER SERIES

SELECTION IN SURVEYS: USING RANDOMIZED INCENTIVES TO DETECT AND ACCOUNT FOR NONRESPONSE BIAS

Deniz Dutz
Ingrid Huitfeldt
Santiago Lacouture
Magne Mogstad
Alexander Torgovitsky
Winnie van Dijk

Working Paper 29549
http://www.nber.org/papers/w29549

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2021, Revised February 2025

The authors gratefully acknowledge financial support from the Norwegian Research Council (grant no. 326391), the Becker Friedman Institute, and the National Science Foundation (grant SES-1846832). We would like to thank Bengt Oscar Lagerström and his team at Statistics Norway for implementing the survey. We would like to thank Joe Altonji, Alex Bick, Raj Chetty, Michael Greenstone, Nathan Hendren, Ali Hortaçsu, John Eric Humphries, Larry Katz, Costas Meghir, Azeem Shaikh, and seminar participants at the 2021 Cowles Foundation Conference on Labor Economics & Public Finance, the Harvard Seminar in Labor Economics, and the Arizona State University Applied Microeconomics Seminar for helpful discussion. Isabel Almazan, Marcus Lim, and Yifan Xu provided excellent research assistance. Any errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Selection in Surveys: Using Randomized Incentives to Detect and Account for Nonresponse Bias

Deniz Dutz, Ingrid Huitfeldt, Santiago Lacouture, Magne Mogstad, Alexander Torgovitsky, and Winnie van Dijk

NBER Working Paper No. 29549
December 2021, Revised February 2025
JEL No. C0, C01, C1, C36, C42, C83, H0, J0

ABSTRACT

We show how to use randomized participation incentives to test and account for nonresponse bias in surveys. We first use data from a survey about labor market conditions, linked to full-population administrative data, to provide evidence of large differences in labor market outcomes between survey participants and nonparticipants, differences which would not be observable to an analyst who only has access to the survey data. These differences persist even after correcting for observable characteristics. We then use the randomized incentives in our survey to directly test for nonresponse bias, and find evidence of substantial bias. Next, we apply a range of existing methods that account for nonresponse bias and find they produce bounds (or point estimates) that are either wide or far from the ground truth. We investigate the failure of these methods by taking a closer look at the determinants of participation, finding that the composition of participants changes in opposite directions in response to incentives and reminder emails. We develop a model of participation that allows for two dimensions of unobserved heterogeneity in the participation decision. Applying the model to our data produces bounds (or point estimates) that are narrower and closer to the ground truth than the other methods. Our results highlight the benefits of including randomized participation incentives in surveys. Both the testing procedure and the methods for bias adjustment may be attractive tools for researchers who are able to embed randomized incentives into their survey.

Deniz Dutz
University of Chicago
1126 East 59th Street
Chicago, IL 60637
deniz.dutz@gmail.com

Ingrid Huitfeldt
BI Norwegian Business School
0442 Oslo
Norway
ingrid.huitfeldt@gmail.com

Santiago Lacouture
University of Chicago
924 E 57th St
Chicago, IL 60637
slacouture@uchicago.edu

Magne Mogstad
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
magne.mogstad@gmail.com

Alexander Torgovitsky
University of Chicago
1126 E. 59th Street
Chicago, IL 60637
atorgovitsky@gmail.com

Winnie van Dijk
Department of Economics
Yale University
87 Trumbull Street, B228
New Haven, CT 06520
and NBER
winnie.vandijk@yale.edu

An online appendix is available at http://www.nber.org/data-appendix/w29549

1 Introduction

Surveys are widely used to inform policy decisions. For example, survey data collected by the U.S. Census Bureau is used to distribute around $675 billion in federal funds annually (Hotchkiss and Phelan, 2017). Survey data also plays an important role in economics research (Currie et al., 2020, and Section 2 of this paper) and social science more broadly (Sturgis and Luff, 2021). Collecting survey data requires participation on the part of those being surveyed. If participation in a survey is correlated with potential responses to the questions asked in the survey, then the results of the survey will be contaminated with nonresponse bias, making them potentially misleading descriptions of the targeted survey population.

The issue of nonresponse bias in surveys has long been appreciated in economics.[1] It is less often discussed in recent empirical research that uses survey data. We conducted a systematic review of recent empirical research in economics to document how empirical researchers in economics cope with the possibility of nonresponse bias in survey data. We find that nonresponse rates are often high, yet discussions of potential nonresponse bias are uncommon: nearly half of the reviewed papers omit any discussion of nonresponse bias. This is perhaps surprising given the discipline's increasing focus on missing data problems in causal inference.[2]

In this paper, we show how to use randomized financial incentives for survey participation to detect and account for nonresponse bias. Incentives for participation are already common features of surveys. Randomly assigning *different* incentives creates ex-ante identical groups with different participation rates whose responses should, on average, be the same if there is no nonresponse bias and if changing incentives does not directly affect participants' answers to a survey question. This observation suggests a test for the presence of nonresponse bias, which we show how to implement. We then consider ways to use the exogenous variation provided by randomly-assigned incentives to identify and estimate average responses for the entire population that the survey sample is drawn from. We develop a new model of participation that combines incentives with reminders, allowing us to better learn about population average responses even in the presence of nonresponse bias.

We apply the test for nonresponse bias, and the methods designed to account for it, to data from the Norway in Corona Times (NCT) survey. The goal of the NCT survey was to study the immediate labor market consequences of the COVID-19 lockdown that began in March 2020. In addition to randomly-assigned incentives for participation, the NCT survey has two attractive features for analyzing survey participation and nonresponse bias. First, Statistics Norway drew a random sample from the entire adult population, ensuring that the survey population is representative of the target population. Second, Statistics Norway merged the survey data with data from administrative sources, enabling us to quantify differences

[1] For example, see Hausman and Wise (1979) and the 1998 special issue of the Journal of Human Resources on attrition in longitudinal surveys.

[2] For instance, Currie et al. (2020) document that the share of economics papers containing terms related to identification strategies, (quasi-)experimental methods, or selection bias has been consistently increasing since 1980.

between participants and nonparticipants that would be unobserved to the survey analyst, and providing us with a ground truth to assess the performance of alternative methods designed to account for nonresponse bias.[3]

Our initial analysis of the NCT survey delivers three findings. First, the administrative data shows that the labor market outcomes of those who participated in the NCT survey are substantially different from those who did not participate. We find that corrections based on observable characteristics commonly used in survey research do not eliminate these differences.[4] This finding raises concerns about bias in survey responses due to selection on unobservables. Second, we directly test for nonresponse bias in survey responses by comparing responses across incentive groups. We find that there are significant and substantial differences in responses between incentive groups, and that these differences persist after adjusting for observables. These findings show that nonresponse bias can be a serious problem even in a survey implemented using best practices by a national statistical agency. Finally, we find that trying to mitigate differences between participants and nonparticipants by increasing incentives could backfire for the NCT survey: even though participation rates increase with incentives, the marginal participants induced to respond by higher incentives are even more different from nonparticipants than those who participate under lower incentives.[5]

A variety of statistical methods have been designed to account for nonresponse bias. In our review, we find that these methods are used infrequently. The most common approach is to assume that responses are missing at random after controlling for a set of observables, so that nonresponse bias can be removed by reweighting. This assumes away nonresponse bias due to unobservable differences between participants and nonparticipants. Yet our empirical results suggest that unobservable differences can play a driving role in nonresponse bias.

Because the NCT survey is linked to administrative data, it offers an opportunity to evaluate the performance of methods that allow for selection on unobservables. We apply a range of methods, including worst-case bounds, bounds that incorporate monotonicity assumptions, and approaches based on parametric and nonparametric selection models. We evaluate these methods by their fidelity to the ground truth—the population averages computed from the entire population using the administrative data—when using only administrative data on the survey participants. We find that some of the methods produce bounds that contain the population quantities, but are quite wide. Other methods produce bounds (or point estimates) that are inconsistent with the population quantities, suggesting that the underly-

[3]We use the term "ground truth" to refer to a benchmark for measuring selection bias. For example, since the administrative data includes the adult population of Norway, we can use it to calculate the population employment rate, and compare it to the employment rate among survey participants (as measured in the administrative data). This comparison isolates selection bias, because it eliminates differential measurement of employment status in the survey responses and the administrative data.

[4]This finding is in line with Bollinger et al. (2019), who use Current Population Survey individual records linked to administrative earnings data to show that nonresponse is not independent of earnings, even after controlling for observables.

[5]This finding is in line with Groves and Peytcheva (2008) and Meterko et al. (2015), who argue, based on a meta analysis, that there is no clear relationship between response rate and nonresponse bias.

ing assumptions may be violated. In some cases, even seemingly-weak assumptions lead to severely incorrect conclusions about the population quantities.

We investigate the failure of these methods in the NCT survey by taking a closer look at the determinants of participation. By considering the impacts of both incentives and reminders on response, we find evidence that there are two types of nonparticipants: "active" nonparticipants who saw the NCT survey invitation and declined to participate because the incentive was too low, and "passive" nonparticipants who never saw the invitation, but might have participated had they seen it. We also find evidence that these two types of nonparticipants have labor market outcomes different from those of the participants, but in opposite directions. We argue that such a scenario is one instance in which one might expect existing methods to perform poorly.

We develop a new method that builds on existing methods by incorporating a distinction between active and passive nonresponse. Our method uses a model of participation that accounts for both variation in randomly-assigned incentives and the timing of reminder emails. We show how to use the new method to correct for nonresponse bias and produce either bounds or point estimates on population-level average responses under different auxiliary shape restrictions. Applying the method to our data produces bounds (or point estimates) that are narrower and closer to the ground truth than existing methods.

This paper is related to literatures in statistics, economics, and survey methodology on reducing and correcting for nonresponse bias.[6] We contribute to these literatures in several ways.

First, we show how randomized financial incentives can be used to test and account for nonresponse bias due to unobserved differences between participants and nonparticipants. We document that surveys used to study aggregate statistics or treatment effects in the economics literature often include incentives, but typically do not randomly assign them.[7] In such cases, randomized incentives can often be incorporated into surveys with little to no additional costs. Our findings and methods point to additional opportunities for randomization in surveys used in economics research.

Second, our empirical results underscore that what matters for nonresponse bias is not the participation rate per se, but *who* participates. In the NCT survey, nonresponse bias actually *increases* with participation rates.[8] This suggests that guidance on survey design may benefit from more nuance. For example, the U.S. Office of Management and Budget (2006, p.60)

[6]The survey methodology literature on nonresponse is reviewed in Groves et al. (2002), Singer (2006), Bethlehem et al. (2011), and National Research Council (2013a); see also Groves et al. (2009, Section 6) for a textbook summary.

[7]Three exceptions are Moffitt (2004), Bhattacharya and Isen (2009), and Coffman et al. (2019) who use randomized incentives and interpret differences in survey participant means across incentive groups as evidence of selection. In the survey methodology literature, several studies have looked at the impact of incentives on survey participation rates conditional on demographic characteristics (see Groves et al., 2009; Singer and Ye, 2013, and references therein).

[8]These patterns may of course be different in surveys conducted outside of periods of policy and economic uncertainty. At the same time, an advantage of surveys is that they can often provide insights to policy makers on a shorter timeline than administrative data, and are thus especially useful in times of crisis. In any case, the methods we develop and apply can be used in times of crisis or non-crisis.

asserts that "response rates are an important indicator of the potential for nonresponse bias" in its guidelines of minimum methodology requirements for federally funded projects. Similarly, the Abdul Latif Jameel Poverty Action Lab (J-PAL) publishes research guidelines which state that "increasing response rates on a subsample and up-weighting the subsample will reduce bias" (J-PAL, 2021); and that the "risk of bias [is] increasing with the attrition rate" (J-PAL, 2020). The NCT survey provides an example that brings this guidance into question, consistent with previous work suggesting that encouraging survey participation tends to skew sample composition along a number of dimensions (see Juster and Suzman, 1995; Martin and Winters, 2001, and references therein).[9]

Third, there are a variety of methods that correct for nonresponse bias due to selection on observable characteristics (see, e.g. Little and Rubin, 2019). Our survey provides an example in which selection on unobservables is a main driver of nonresponse bias, and thus these methods fail to correct for nonresponse bias.[10] Moreover, we find that widely-used reweighting methods sometimes exacerbate nonresponse bias by amplifying unobservable differences.

Fourth, we evaluate the performance of existing methods that try to address selection on unobservables. Our survey provides an attractive setting for evaluating the performance of these methods against a known ground truth, in the spirit of LaLonde's (1986) evaluation of non-experimental estimators of treatment effects. Worst-case bounds and bounds that incorporate shape restrictions (such as monotonicity assumptions) are considered in a series of papers by Manski and co-authors (Manski, 1989, 1990, 1994; Horowitz and Manski, 1998; Manski and Pepper, 2000; Manski, 2016), and applied to study population parameters in the presence of sample selection by, e.g., Blundell et al. (2007). Approaches based on parametric and nonparametric selection models are based on a line of work by Gronau (1974); Heckman (1979); Heckman and Vytlacil (2001); Vytlacil (2002); Heckman and Vytlacil (2005, 2007).[11]

Fifth, we contribute to a small and mostly theoretical literature on selection models with multiple dimensions of unobserved heterogeneity. Multiple dimensions of unobserved heterogeneity arise naturally in instrumental variable models with ordered and unordered treatments (e.g. Heckman and Vytlacil, 2007; Kirkeboen et al., 2016; Heckman and Pinto, 2018; Lee and Salanié, 2018; Mountjoy, 2021; Humphries et al., 2024), as well as in settings

[9] With explicit reference to our study and findings, J-PAL has changed its advice to researchers, now emphasizing that low (high) participation rates do not necessarily indicate (small) nonresponse bias, and researchers should consider ways of testing and accounting for nonresponse bias other than adjusting for observable differences between participants and nonparticipants.

[10] Of course, the importance of observables versus unobservables depends on which variables the analyst observes. Our analysis considers a set of observables that are commonly used to assess or adjust for nonresponse bias in economics research using survey data.

[11] These techniques have been applied to survey data by, e.g., Horowitz and Manski (1998); Manski (2003); Heckman and LaFontaine (2006); Gørgens and Ryan (2008); Bollinger and Hirsch (2013); Hokayem et al. (2015); Manski (2016); McGovern et al. (2018) and Manski and Molinari (2021) to account for selection when using survey data to estimate population means, and by Bhattacharya and Isen (2009), Behaghel et al. (2015), and DiNardo et al. (2021) in a program evaluation context to identify treatment effects for local subpopulations (e.g., the individuals who respond to the survey) under additional assumptions.

with multiple instruments (Mogstad et al., 2020). While related, our multidimensional selection model is designed more specifically for modeling different forms of non-participation in surveys with randomized incentives. Our analysis of the model highlights some of the identification challenges created by multiple unobservables, and demonstrates how one can overcome these challenges with partial identification approaches.

The structure of the paper is as follows. In Section 2, we present data on the ways in which empirical research in economics addresses nonresponse bias in surveys. In Section 3, we discuss the NCT survey. In Section 4, we measure nonresponse bias in the NCT survey using administrative data, and then we show how randomized incentives can be used to directly test for nonresponse bias in survey responses. In Section 5, we evaluate the performance of existing methods that allow for selection in unobservables. In Section 6, we develop a new method based on a model that allows for both active and passive nonresponse. Section 7 concludes by distilling our results into concrete recommendations for practitioners.

2 Survey data and nonresponse in empirical economics research

In this section we present descriptive facts about the use of survey data in empirical research in economics. These facts guide our discussion in the remainder of the paper.

We draw on data from several sources. To document trends in the use of survey data since 1974, we use text data on NBER Working Papers and on the so-called "top-five publications" from multiple databases.[12] To quantify the prevalence and severity of nonresponse, we aggregate information about nonresponse rates in large-scale U.S. household surveys that are frequently used to inform policy decisions and in academic research.[13] Finally, to systematically document nonresponse in recent empirical economics research and empirical researchers' practices in coping with possible nonresponse bias, we use the results of a systematic review of survey-based research published in top-five economics journals between January 1st 2015 and August 31st 2020. We conducted this review after consulting the most recent PRISMA guidelines for systematic reviews, which are widely followed in the biomedical sciences (Page et al., 2021). A detailed description of our protocols can be found in Online Appendix D.

Descriptive Fact #1: *The collection and use of survey data in economics research has increased over the past decade.*

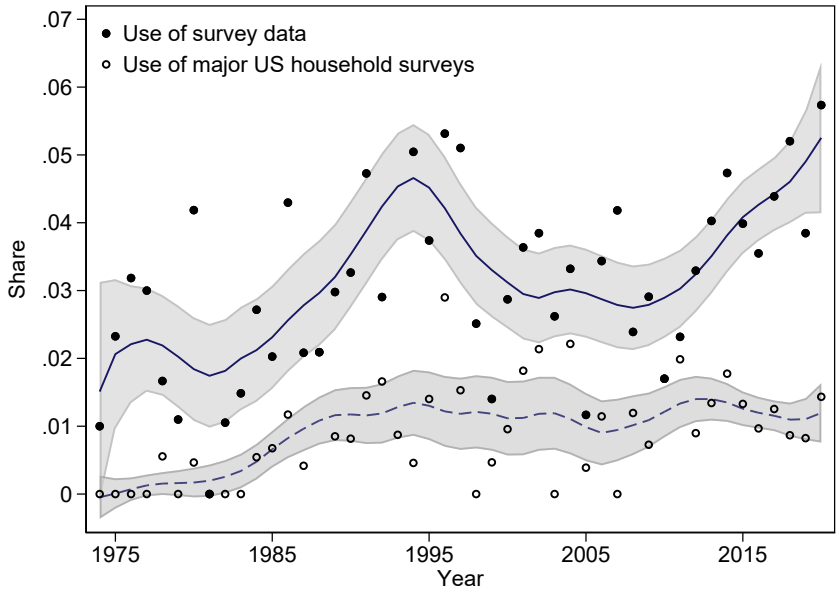
Figure 1 shows how the collection and use of survey data in economics research have evolved since 1974. The use of survey data increased during the 1980s and early 1990s, before starting to decline in the mid-1990s. The increase happened in conjunction with a rise in the use of

[12]The top-five journals referenced throughout this paper are the Journal of Political Economy, the American Economic Review, the Quarterly Journal of Economics, the Review of Economic Studies, and Econometrica. We collected data on the titles and abstracts of publications in these journals between January 1974 and August 2020 from the Web of Science database, the EconLit database, and JSTOR. Details on data collection and harmonization across the different sources are in Online Appendix A. We use the NBER Working Paper Metadata (National Bureau of Economic Research, 2020) to capture research not published in top-five journals (see Online Appendix B for details).

[13]See Online Appendix C for details on the construction of this data set.

systematically-collected household survey panels, such as the NLSY79, the HRS, and the SIPP. Since 2010, the data show a renewed upward trend despite no change in the use of these household survey panels.[14] This suggests that not only are economists using survey data more, but they have also turned to generating their own customized survey data. In principle, such a shift towards researcher-generated survey data would mean that researchers increasingly have the option to tailor their survey design and implementation to increase response rates as well as to test and correct for nonresponse bias, for example along the lines of the survey design we study in this paper.

Figure 1: Use of survey data in top-five publications



Notes: Sample consists of papers with available abstract published in top-five economics journals between January 1974 and August 2020. Records were obtained from the Web of Science, JSTOR, and EconLit. The solid line depicts the fitted values of a local linear regression of the yearly share of papers that include the word "survey", or variations thereof such as "surveyed" or "surveys", in their title or abstract. The dashed line depicts the fitted values of a local linear regression of the yearly share of papers that include the name or acronym of any of the following surveys in their abstract or title: CPS, ACS, CEX, HRS, NLSY79, NLSY97, CNLSY, SIPP, SCF, ATUS, SCE, GSS, NHIS or PSID, on year. We use a bandwidth of 2 years with an Epanechnikov kernel. 90% confidence intervals are presented in shaded areas. See Online Appendix A for more details on sample and time series construction.

Descriptive Fact #2: *Nonresponse bias is a significant possibility in most survey-based economics research: nonresponse rates are often high, and they have been increasing even for household panels that are used to validate the representativeness of other surveys.*

Our systematic literature review reveals that nonresponse rates in economics research are often high. This is especially true when the data is researcher-generated: the average nonresponse rate is 50 percent for such surveys in our review sample.[15] Among studies that

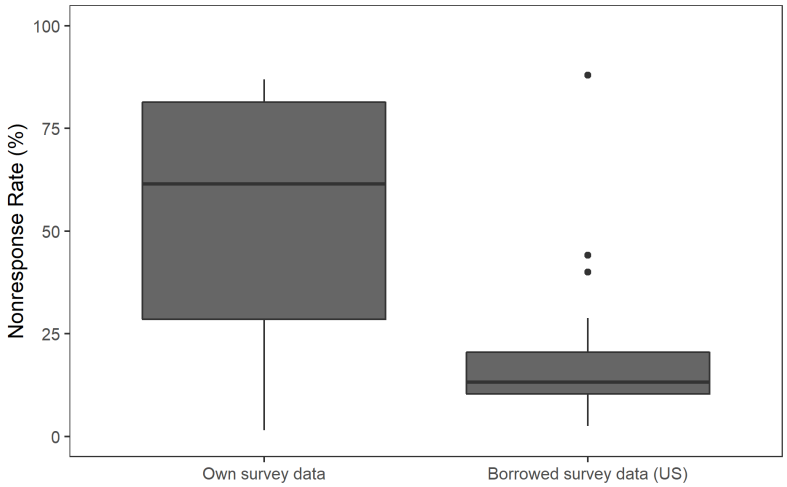
[14] The trends are similar if we restrict attention to fields classified as applied microeconomics (see Online Appendix Figure A.2), or if we instead use data on NBER Working Papers (see Online Appendix Figure B.1). Currie et al. (2020) also find similar trends using a different approach and data set (see their Online Appendix Figure A.II, Panel A).

[15] Studies that didn't use a probability sample (35 percent of papers using their own survey data) were excluded

use data borrowed from pre-existing U.S. household surveys, the average nonresponse rate is 19 percent. For studies in both categories, nonresponse rates reach as high as 87 percent. Figure 2 visualizes the nonresponse rates in our review sample.

The phenomenon of rising nonresponse rates in major household surveys has been documented repeatedly and in a wide variety of settings.[16] It is seen even in the panel surveys that are often used to validate the representativeness of other surveys, such as the Current Population Survey. This trend has not slowed over the past five years—if anything, it appears to be accelerating (see Figure 3). Although higher nonresponse rates do not necessarily imply an increase in nonresponse bias, these levels and trends suggest that nonresponse bias is a serious possibility in most survey-based economics research even when the data comes from sources widely regarded as achieving the highest possible standards of data quality.

Figure 2: Nonresponse rates in surveys used in top-five publications



Notes: This figure shows boxplots of nonresponse rates in the papers selected for our systematic review. The boxplot "Own survey data" includes papers where survey data is collected by the authors using a probability sample. The "Borrowed survey data (US)" boxplot includes papers that borrow survey data from one of the major US household surveys. See Online Appendix D for more details.

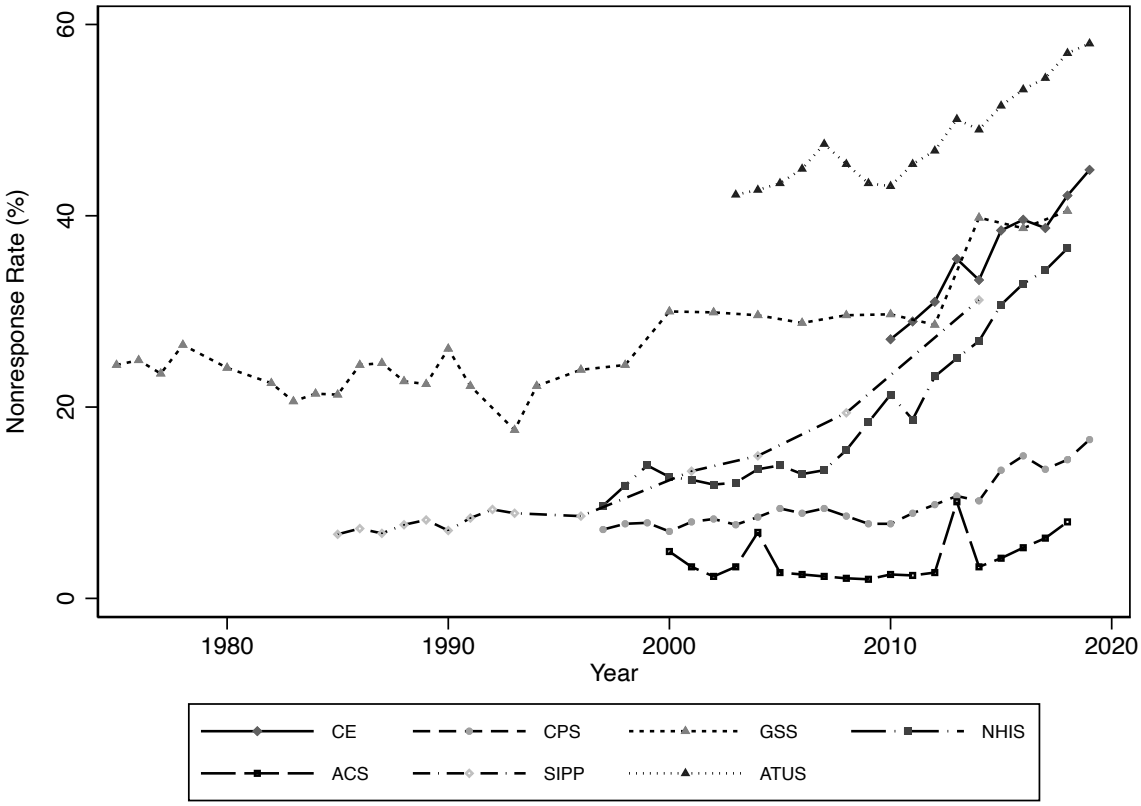
Descriptive Fact #3: *Researchers frequently omit discussion of potential nonresponse bias.*

Despite the prevalence of high nonresponse rates in economics research, we find that nearly half of the studies in our review sample do not include a discussion of potential nonresponse bias and its consequences for the study's findings. This practice stands in stark contrast to the care taken in discussing and dealing with potential selection bias when answering causal inference questions. One explanation for this practice is that researchers believe that nonresponse bias is irrelevant for the interpretation of a study's findings, which would be

from our review as it is not possible to calculate nonresponse rates for such studies that can be compared to response rates based on probability samples. Nearly half (43%) of the studies that were excluded from our response-rate analysis on this basis used data collected by marketing firms (Respondi, Qualtrics, C&T Marketing, and Growth from Knowledge).

[16]See, for example, National Research Council (2013b), Meyer et al. (2015) and Czajka and Beyler (2016) for the U.S., and de Leeuw and de Heer (2002) for other high-income countries.

Figure 3: Nonresponse rates of U.S. large household surveys over time



Notes: This figure shows time trends in the yearly nonresponse rates for seven large-scale, cross-sectional U.S. surveys: the Consumer Expenditure Surveys (CE), the Current Population Survey (CPS), the General Social Survey (GSS), the National Health Interview Survey (NHIS), the American Community Survey (ACS), the Survey of Income and Program Participation (SIPP), and the American Time Use Survey (ATUS). Details on data sources and construction of the nonresponse rates can be found in Online Appendix C.

implied by the assumption that responses are missing completely at random. The findings in our paper speak directly to whether such an assumption is warranted without further analysis and testing.

Descriptive Fact #4: *When researchers discuss potential nonresponse bias, they assume either that responses are missing completely at random, or that selection into participation is based exclusively on observables.*

In empirical research, economists largely use two strategies to explicitly address potential nonresponse bias. The first is to compare participant sample means to a reference population and (explicitly or implicitly) assert that no adjustment is necessary if little difference is found. Our systematic review shows such comparisons are found in 47 percent of papers using own survey data and in 6 percent of papers using borrowed survey data from one of the twelve prominent U.S. household surveys. The second is to apply a reweighting-on-observables procedure. This procedure is applied by 16 percent of papers using own survey data, and 53 percent of papers using borrowed data.

The current practice of assuming responses are missing completely at random or selection

is based exclusively on observables raises the question of whether nonresponse bias due to unobservables is empirically important, and how to test and correct for it. These questions motivate our paper.

Descriptive Fact #5: *Ex ante strategies for mitigating nonresponse bias—such as providing participation incentives—are common. These strategies are rarely designed to test for or address selection into survey participation based on unobservables.*

The studies in our review sample largely use two types of strategies to increase the overall response rate. The first is intensive modes of outreach, such as in-person interviews, or repeated emails or calls. The second is to offer financial or in-kind incentives for survey completion. Incentives for survey completion are typically offered uniformly across participants, or are varied in a non-random way, e.g. the type or level of incentive is determined by membership of a specific demographic group.[17] In our review of recent top-five publications, 52 percent of surveys from studies collecting their own survey data use some form of incentives, and nearly all of these (93 percent) use financial incentives.

Our findings in this paper show that such ex ante strategies could increase nonresponse bias, rather than mitigate it. Moreover, by applying these strategies uniformly across potential participants, rather than using them for a random subset of invitees, existing studies forgo the ability to test and correct for selection into survey participation based on unobserved factors. This suggests a natural direction for exploring possible improvements over current practice: data collection strategies that embed exogenous variation in participation incentives, such as the one we demonstrate in this paper.

3 The Norway in Corona Times Survey

3.1 Background

The COVID-19 (SARS-CoV-2) pandemic was confirmed to have reached Norway on February 26, 2020. The number of cases increased rapidly, prompting the government to impose severe restrictions on the behavior of individuals and firms. On March 12th, a national lockdown was announced. The majority of the workforce was told to work from home; stringent limitations were put in place banning gatherings in public and private settings; schools, daycares, and certain businesses were forced to close.

To study the consequences of this lockdown for the labor market, the national statistics agency (Statistics Norway) carried out the survey "Norway in Corona Times" (NCT). The primary motivation for carrying out the survey was that Statistics Norway's administrative data sets are updated and reported only every quarter or year, whereas surveys can provide

[17] In our review, two papers were exceptions to this rule. The first is DellaVigna et al. (2017), for whom the effect of randomly assigned incentives on survey participation is of substantive interest. The second is Coffman et al. (2019), who use survey incentives to test for selection, concluding little if any evidence of significant selection on unobservables. In Online Appendix E, we re-analyze Coffman et al. (2019)'s published data and show that, for all but one of the variables considered, their study was underpowered to detect economically meaningful differences across incentive levels.

information nearly in real time. While this presents an advantage of using survey data to inform policy, there are also drawbacks, including potential bias due to nonresponse. Our empirical analysis uses the NCT survey to study this tension.

The NCT questionnaire was designed by the authors of this paper in collaboration with Statistics Norway's unit for survey analysis. For our analysis, we focus on the questions that asked about individuals' labor market circumstances. We use these responses to construct quantities that describe the state of the Norwegian labor market before and after the lockdown.[18] The measures we consider closely resemble the labor market statistics included in, e.g., the U.S. Bureau of Labor Statistics Employment Situation Summary, which is based on the Current Population Survey.

3.2 Why we use the NCT survey to study nonresponse

The NCT survey offers three key advantages for studying participation and nonresponse bias in surveys. First, Statistics Norway has access to a census of the entire population of Norway, along with high-quality contact information, which allows them to sample randomly from the population of interest.[19] As a result, we do not have to worry that non-representativeness due to the sampling procedure confounds the assessment of nonresponse bias.

Second, Statistics Norway is able to merge the survey data with data from administrative sources through unique personal identifiers. As a result, we can observe labor market outcomes and a rich set of characteristics for each individual, independently of whether they respond to the survey. These data are reported by a third party, e.g., employers, and are inputs to the audited tax returns; consequently, they can be considered to be of high accuracy. The linked administrative data offers a ground truth that we can use both to quantify nonresponse bias in the NCT survey and to assess the performance of different methods to correct for such bias. Furthermore, some of the survey questions aim to elicit information that is also recorded in the administrative data. This allows us to examine the accuracy of the responses to the survey questions, which we do in Section 3.5.

Third, the design of the NCT survey included randomly-assigned financial incentives for participation, as well as reminder emails and text messages. We use these features to show how researchers can test for nonresponse bias and characterize selection into survey participation without requiring linked administrative data, and to correct estimates of the population mean for selection on unobservables.

[18] Appendix Table A.1 provides details on all variable definitions.

[19] The contact registry used for the survey is owned by the government and used to send official information and documents, including the tax return forms. Since individual submission of the tax return is mandatory by law and non-filers are audited and fined, coverage is almost complete and information is up-to-date. Mailing address and telephone number are available for nearly every adult individual, while email addresses are observed for 89 percent. This contact information was used to reach out to the individuals that were sampled for the NCT survey. Thus, we can be confident that the survey would give representative estimates in the absence of nonresponse bias.

3.3 Survey design and implementation

The population of interest is defined as all individuals who, as of April 1st, 2020, were Norwegian residents and at least 18 years of age. From this population, a random sample of 10,000 individuals was invited to participate in the survey. The sample was further randomized into type of survey administration. The vast majority of the sample (93 percent) was invited to complete the survey online, while the remaining individuals were invited for a phone interview. Throughout the paper, we focus on the random sample invited to the online survey. The mode of invitation for the online survey was email when available (89 percent) and regular mail otherwise. Invitations were supplemented with a notification by text message to everyone in the sample with a registered phone number (90 percent).

The initial survey invitation for the online sample was distributed on April 20, 2020. Figure 4 shows how the participation rate developed over time.[20] A total of six reminder messages were sent out before the survey was taken offline on May 22, 2020.[21] Individuals were notified of their randomized incentive for completing the survey (see below) in each contact attempt. They were also informed about the purpose of the survey and the estimated time it would take to complete it. By the end of the data collection period, 47.4 percent of those invited had completed the survey. This participation rate is similar to that of other surveys conducted by Statistics Norway,[22] and more broadly, is close to the average response rate for self-collected surveys in publications in top-five journals in economics, as described in Section 2.

Individuals in the sample were randomized into one of five incentive groups. Group assignment determined an individual's probability of receiving a prepaid credit card worth 1,000 NOK (110 USD) upon completing the survey.[23] The credit card could be spent online and in nearly all Norwegian stores. The probabilities were set to 0 percent, 1 percent, 5 percent, 7 percent and 10 percent, and individuals were assigned to the corresponding groups with probabilities 40 percent, 30 percent, 15 percent, 7.5 percent and 7.5 percent. This yields an expected payoff of 2.6 USD, ranging from 1.1 USD in the lowest incentive group to 11 USD in the highest incentive group. In comparison, the average incentive in a meta-analysis of 55 survey incentive experiments by Mercer et al. (2015) was around 10 USD. By virtue of randomization, the incentive groups are probabilistically identical. Balance tests for the administratively-linked outcomes are presented in Appendix Table A.2, and we confirm that outcomes do not differ significantly across the groups.

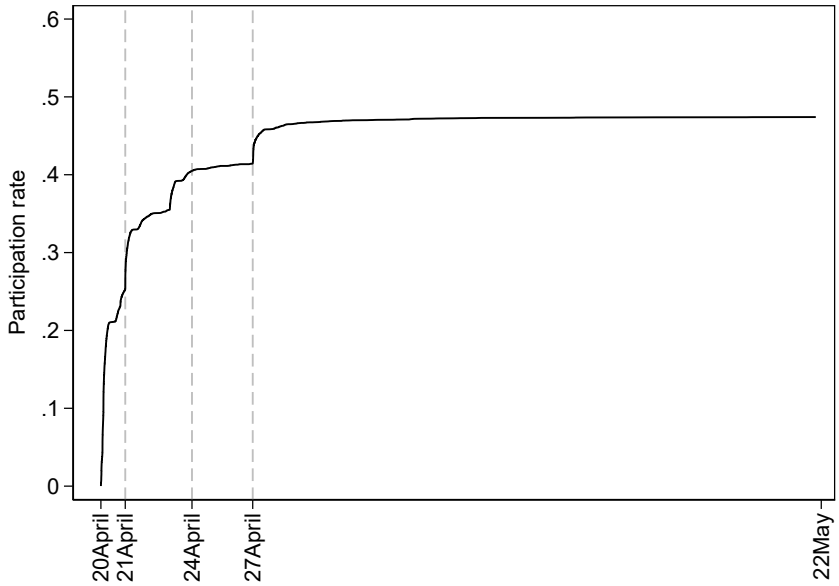
[20] Throughout the paper, we define "participation" as having completed the entire survey. Results remain unchanged if we instead define participation as having responded to all questions relating to the labor market (our main variables of interest).

[21] On April 21 (day 1), April 24 (day 4), and April 27 (day 7) text messages and emails were sent to all individuals who had not started the survey. In addition, text messages were sent on April 23 (day 3), April 29 (day 9), and May 6 (day 15) to individuals who had started but not completed the survey.

[22] For example, the Life Quality Survey, a non-recurring, voluntary survey conducted by Statistics Norway and distributed in the same period as our survey, had a participation rate of 44 percent.

[23] In a meta-analysis on the use of survey incentives in academic research, Mercer et al. (2015) point out that lotteries are the most common mechanism for providing incentives to participate in web surveys.

Figure 4: Participation rates over time



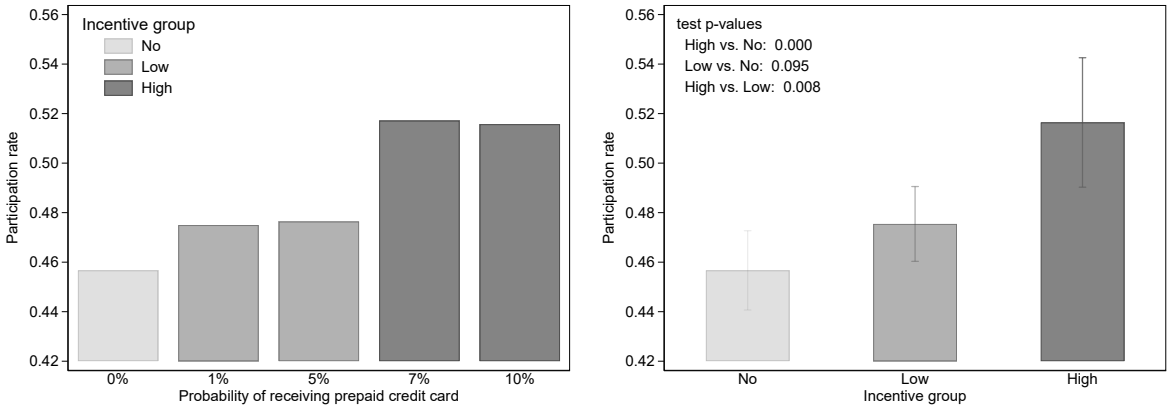
Notes: This figure shows the total share of individuals who participated in the NCT as a function of time. The vertical lines mark the dates at which reminders that were sent to all individuals who had not yet participated.

3.4 Participation rates and incentives

Figure 5a displays the proportion of individuals who participated in the survey by incentive group. Participation rates increase with the level of the incentive, with three distinct groups standing out. The participation rate is 45.7 percent in the unincentivized group, 47.5 and 47.6 percent in the two lowest incentive groups, and 51.7 and 51.6 percent in the two highest incentive groups. Given these participation rates, we chose to use three aggregated incentive groups in our analyses: "high" (7 and 10 percent probability of receiving prepaid credit card), "low" (1 and 5 percent probability of receiving prepaid credit card) and "no". This categorization, depicted in Figure 5b, helps us gain precision in the analyses. Relative to the no-incentive group, participation rates increase by around 2 percentage points for the low-incentive group, and by an additional 4 percentage points for the high-incentive group. We reject a joint test of equal participation across the three groups with p–value < 0.01.

The individuals in the NCT survey are fairly elastic to financial incentives. An expected return of 10 USD increased the participation rate by 6 percentage points, or 13 percent. By comparison, Mercer et al. (2015) found that the estimated average effect of a promised payment of the same amount was around 5 percent. Coffman et al. (2019) found that a fixed payment of 20 USD increased participation by 8.4 percentage points, while DellaVigna et al. (2017) found that a fixed payment of 10 USD increased participation by 5.4 percentage points.

Figure 5: Participation rates by incentive group



(a) **(b)**

Notes: Panel (a) shows participation rates by incentive group, where incentives are defined by the probabilities of receiving a prepaid credit card worth NOK 1,000 (USD 110) upon completing the survey. Panel (b) plots estimated coefficients and 90% CI from a regression of participation on the aggregated incentive groups (as defined in the top left corner of Panel (a)), which we use in our analyses. *P*–values for testing the pairwise equality across incentives are shown in upper left corner.

3.5 Key variables and descriptive statistics for survey participants

Table 1 lists the variables we use in our main analyses. We indicate which variables come from survey data (observed only for the survey participants) and which come from administrative data (observed for the entire population, including survey nonparticipants). In the survey data, we focus on changes in hours worked, an indicator for no longer working full-time, an indicator for becoming furloughed or unemployed, and an indicator for having applied for unemployment insurance (UI) benefits since the lockdown.[24] From the administrative data, we observe individual characteristics such as gender and age for all invited individuals, irrespective of whether they participated or not. The administrative data also collects information on monthly earnings and employment over the two months before and one month after lockdown. To further characterize how the economy responded to the lockdown, we additionally construct indicators for a large earnings loss after the lockdown (defined as earnings after lockdown being at least 20% lower than before lockdown) and for a loss of employment.

Table 1 also presents participant means and standard deviations for these variables. We find that average monthly earnings for participants was 4,030 USD before the lockdown, and dropped to 3,677 USD after the lockdown. In addition to the decrease in mean earnings, employment rate estimates for participants indicate a decrease from 68 percent before the lockdown to 58 percent after the lockdown. We also find that many individuals were severely impacted by the lockdown: 16 percent of survey participants experienced a large loss in earnings, and 11 percent experienced employment loss. Survey responses further confirm that the labor market was negatively affected by the lockdown: 28 percent of participants

[24]Inaccurate or untruthful reporting is always a concern when using surveys outcomes. Our setting allows us to examine misreporting using survey responses for which we observe the ground truth in administrative data. In Online Appendix F, we examine misreporting, and find no evidence of it.

worked fewer hours in response to the lockdown, 18 percent no longer worked full-time, and 10 percent applied for UI.

Of course, these descriptive statistics of the survey participants will only give an accurate description of the overall Norwegian economy if participants and nonparticipants had similar labor market outcomes. In the following sections, we will use our survey design as well as the linked administrative data to evaluate the accuracy of conclusions drawn based on conventional analyses of survey participant data, including the descriptive statistics provided above.

Table 1: Summary statistics

	Participant	
	Mean	SD
Panel A: Individual characteristics (administrative)		
Age	47.3	16.6
Years of Schooling	13.8	3.4
Immigrant	0.091	0.288
Female	0.524	0.5
Panel B: Outcomes		
B.1: Survey		
No longer full-time work	0.176	0.381
Reduction in work hours	0.275	0.447
Became furloughed or unemployed	0.068	0.252
Applied for UI	0.104	0.305
B.2: Administrative		
Earnings before lockdown	4,030	5,302
Earnings after lockdown	3,677	3,791
Earnings loss	0.162	0.369
Employment before lockdown	0.675	0.469
Employment after lockdown	0.577	0.494
Employment loss	0.112	0.316

Notes: This table presents participant means and standard deviations for our considered variables. We consider participants in the high incentive arm, as this arm obtained the highest participation rate. Appendix Table A.1 provides details on all variable definitions.

4 Testing for nonresponse bias and characterizing selection

In this section we introduce a framework for analyzing differences in outcomes between participants and nonparticipants, which we refer to generally as nonresponse bias. We use linked administrative data to directly measure nonresponse bias in the NCT survey. Then we show how researchers can use randomized incentives to test for nonresponse bias and

characterize selection using only survey data. In Online Appendix H, we show how the same framework can be applied when survey data is used to study treatment effects.

4.1 Defining nonresponse bias and selection

Consider a population of individuals indexed by i. Let $Y_i^\star$ denote an outcome of interest for individual i. The outcome could be measured in administrative data, or it could be a response to a question in a survey. In either case, we refer to $Y_i^\star$ as individual i's *response*. We want to measure the mean response across the population, $\mathbb{E}[Y_i^\star]$. Let $R_i \in \{0, 1\}$ denote whether individual i participates in the survey. If $Y_i^\star$ is a response to a survey question, then we observe $Y_i = Y_i^\star$ only if $R_i = 1$. If $Y_i^\star$ is an outcome measured in administrative data, then we observe $Y_i = Y_i^\star$ for all i.

It may be that an individual's decision to participate in the survey, R_i, is correlated with their response, $Y_i^\star$. It is easy to see why this could occur if $Y_i^\star$ is a labor market outcome. For example, those who are more likely to participate in the survey may be those with lower costs of time due to weaker attachment to the labor market. This would cause the unknown nonparticipant mean to differ from the participant mean, so that $\mathbb{E}[Y_i^\star] \neq \mathbb{E}[Y_i|R_i = 1]$. *Nonresponse bias* is the difference, $\mathbb{E}[Y_i|R_i = 1] - \mathbb{E}[Y_i^\star]$.

As documented in Section 2, researchers routinely assume that nonresponse bias is either absent or fully explained by observables. These assumptions are justified by assuming, respectively, that responses are missing completely at random, meaning that $Y_i^\star$ and R_i are independent, or that responses are missing at random conditional on some vector of observables, X_i, meaning that $Y_i^\star$ and R_i are independent, conditional on X_i (Little and Rubin, 2019). We will refer to the former as *no selection* and to the latter as *selection on observables*. Nonresponse bias implies that there is selection. If there is nonresponse bias after conditioning on observables, then there is *selection on unobservables*.

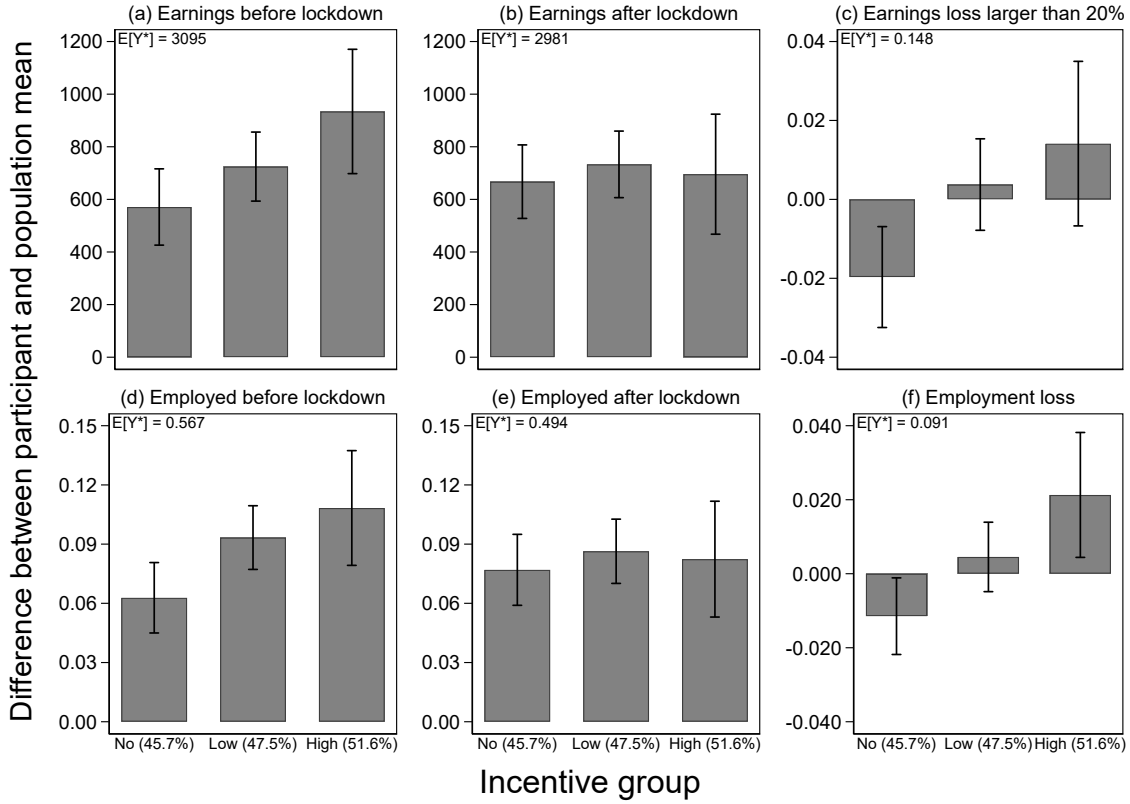
4.2 Using linked administrative data to measure nonresponse bias and characterize selection

Nonresponse bias in the NCT survey

We directly measure nonresponse bias for outcomes in the linked administrative data. Figure 6 reports the difference between the participant sample mean and the true population mean for each of the six administrative outcomes discussed in Section 3.5.[25] The results are stratified on the incentive arm (no, low, and high) as if they were distinct surveys, each with a different incentive level, but identical in every other way. Across all outcomes and incentive arms we find substantial, and statistically significant nonresponse bias; fixing either the outcome or the incentive arm, joint tests of equality always reject the null of no nonresponse bias with p–values < 0.01.

[25]Panels A and B of Appendix Table A.3 report population and participant means in table form.

Figure 6: Nonresponse bias and selection in administrative outcomes



Notes: This figure shows differences in participant means relative to population means for administrative outcomes by incentive level. Error bars represent 90% confidence intervals. Each panel presents results for one outcome. Population means are shown in upper left corners of each panel.

The magnitude of the nonresponse bias is economically important. For example, participants in the high incentive arm had on average roughly 930 USD (30 percent) higher monthly earnings before the lockdown than the full population, and they were 10.8 percentage points (19 percent) more likely to be employed. The survey estimate in the high-incentive arm that 58 percent of participants were employed after the lockdown over-estimates the true rate by 8 percentage points. A researcher or policy maker comparing this figure to the actual employment rate before the lockdown (57 percent) would conclude that the employment remained virtually unchanged over the lockdown. In fact, it dropped by 7 percentage points (see Appendix Table A.3).

Perhaps surprisingly, Figure 6 shows that nonresponse bias in the no-incentive arm is either comparable or smaller in magnitude than in the high incentive arm. For example, no-incentive participants had on average 570 USD (18 percent) higher monthly earnings before the lockdown than the full population, compared to 930 USD (30 percent) for high-incentive participants. These results show that while higher incentive surveys may have higher response rates, they do not necessarily have less nonresponse bias.

Is nonresponse bias due to selection on observables or unobservables?

In Section 2, we found that when researchers do correct for potential nonresponse bias, they typically assume that selection is fully explained by observables. A standard approach is to reweight by the propensity score, i.e. the probability of participating conditional on observable characteristics, X_i. If selection is only on observables, then the reweighted mean estimate of participant responses is a consistent estimate of the population mean (Rosenbaum and Rubin, 1983; Rubin, 1987; Little and Rubin, 2019).

We compute reweighted estimates under two specifications for the propensity score. Both specifications are logit models with characteristics that are commonly used for reweighting. We use all individual-level characteristics that are used for re-weighting in three or more papers from our systematic review—age, gender, immigration status, and years of schooling—plus municipality-level characteristics.[26] The first specification uses only the municipality-level data. The second specification uses the individual-level administrative data. In Appendix Table A.7 we show that both sets of characteristics are strong predictors of labor market outcomes and participation.

Figure 7 reports differences between the reweighted estimates and the population mean for both propensity score specifications and each of the three survey arms.[27] The effect of reweighting on the direction and magnitude of nonresponse bias varies by outcome, specification, and incentive level. However, there are two broad takeaways.

First, we continue to find substantial nonresponse bias after reweighting on observables. For each reweighting specification and incentive survey arm, a joint test rejects the hypothesis that selection for all six outcomes is fully explained by observables with p–value < 0.01. Reweighting on municipal characteristics only slightly changes estimates relative to the unweighted counterparts, and we continue to find substantial nonresponse bias.

Second, correcting for selection on observables can actually exacerbate nonresponse bias. While reweighting on individual characteristics has a larger effect than reweighting on municipal characteristics, the result is sometimes more bias, not less. For example, reweighting on individual characteristics in the high-incentive arm more than doubles the nonresponse bias for earnings loss and employment loss measures relative to the unweighted estimates.[28]

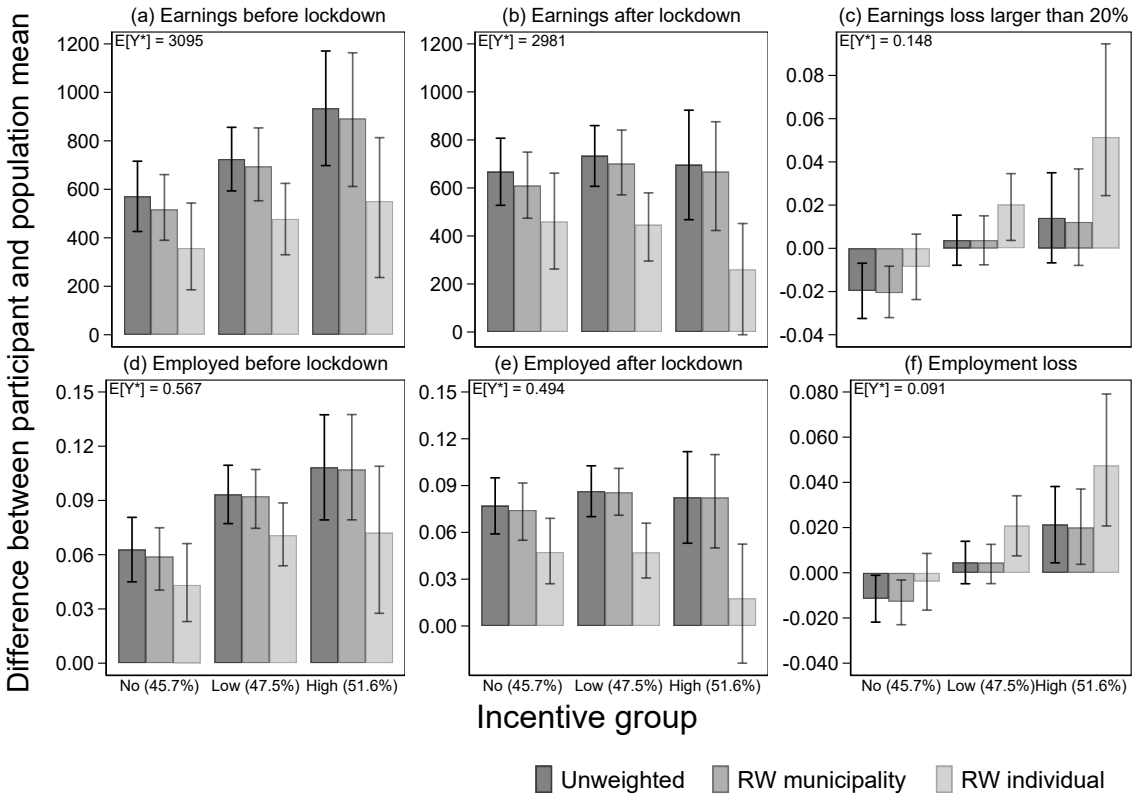
To ensure that these findings are not driven by the choice of reweighting procedure, we examine the performance of a large set of methods used to adjust for selection on observables, including machine learning algorithms, class weights, and imputation. We also examine the performance of richer sets of observable characteristics, including ones that include lagged

[26]Some papers in our review used race or ethnicity. In our context, immigration status is a more reasonable choice for re-weighting, so we use it instead. The municipality-level information we use is obtained from Fiva et al. (2020), and consists of population size, gender share, share of elderly residents, unemployment rate, and median household income. There were 356 municipalities in Norway in January 1, 2020. The average population size of a municipality is about 15,000.

[27]Panels C and D of Appendix Table A.3 report reweighted participant means in table form.

[28]Whereas the unweighted estimate for job loss is about 2.1 percentage points higher than the full population job loss rate, the reweighted estimate is 4.6 percentage points higher.

Figure 7: Nonresponse bias and selection in administrative outcomes after reweighting



Notes: This figure shows differences in participant means relative to population means by incentive level and estimation method (unweighted, reweighted by municipality characteristics, and reweighted by individual characteristics) for administrative outcomes. Dark gray bars depict unweighted estimates, as presented in Figure 6. Lighter gray bars depict estimates reweighted by municipality-level characteristics (population size, gender share, share of elderly residents, unemployment rate, and median household income (Fiva et al., 2020)); and by individual-level characteristics (gender, age, years of schooling and immigration status). We estimate a logit model of the probability of participating that is linear in these characteristics, and reweight participants by the inverse of the predicted value. Error bars represent 90% confidence intervals, for reweighting we construct these via bootstrapping. Population means are shown in upper left corners of each panel. Appendix Table A.3 presents results in table form.

outcomes. The results are reported in Online Appendix G. The findings mirror those presented in this section: regardless of the method used or the choice of characteristics, we consistently find substantial nonresponse bias after correcting for selection on observables.

If the unobservables driving nonresponse bias were constant over time, then there would be less nonresponse bias for outcomes representing changes over time than in outcomes representing levels.[29] Table A.4 reports nonresponse bias relative to the true population value for both levels and changes of earnings and employment. In our case, relative nonresponse bias is typically larger for changes than for levels. Table A.5 shows that this conclusion remains after reweighting by observable individual characteristics.

[29]Snowberg and Yariv (2021) find some evidence that differences and correlations are less susceptible to nonresponse bias. On the other hand, Heffetz and Rabin (2013) find that the gender difference in happiness changes sign when comparing easier-to-reach participants versus harder-to-reach participants.

4.3 Testing for nonresponse bias and selection using survey data

The randomized incentives in the NCT survey also allow us to test for nonresponse bias in survey responses, even though these outcomes are not observed for nonparticipants. Because incentives are randomly assigned, participants in each incentive arm should have the same distribution of (latent) responses. The tests we consider are that the average response, $\mathbb{E}[Y_i|R_i = 1, Z_i = z]$, is equal across all $z \in \mathcal{Z}$, where $\mathcal{Z}$ denotes the set of incentive levels.

For this to be a test of nonresponse bias, it is necessary to maintain an exclusion restriction that the incentives themselves do not directly affect responses. We show in Online Appendix F how the exclusion restriction can be tested with or without access to administrative data, and we find no evidence that it is violated.

Given the exclusion restriction, the random assignment of the incentives implies that average observed responses in each incentive arm should be equal to each other and to the population average. Finding different average responses across incentive arms thus implies that there is nonresponse bias in at least one of the incentive arms. Nonresponse bias in one incentive arm implies nonresponse bias in the entire survey, at least barring unusual knife-edge cases where biases of different directions offset one another when averaging across incentives.

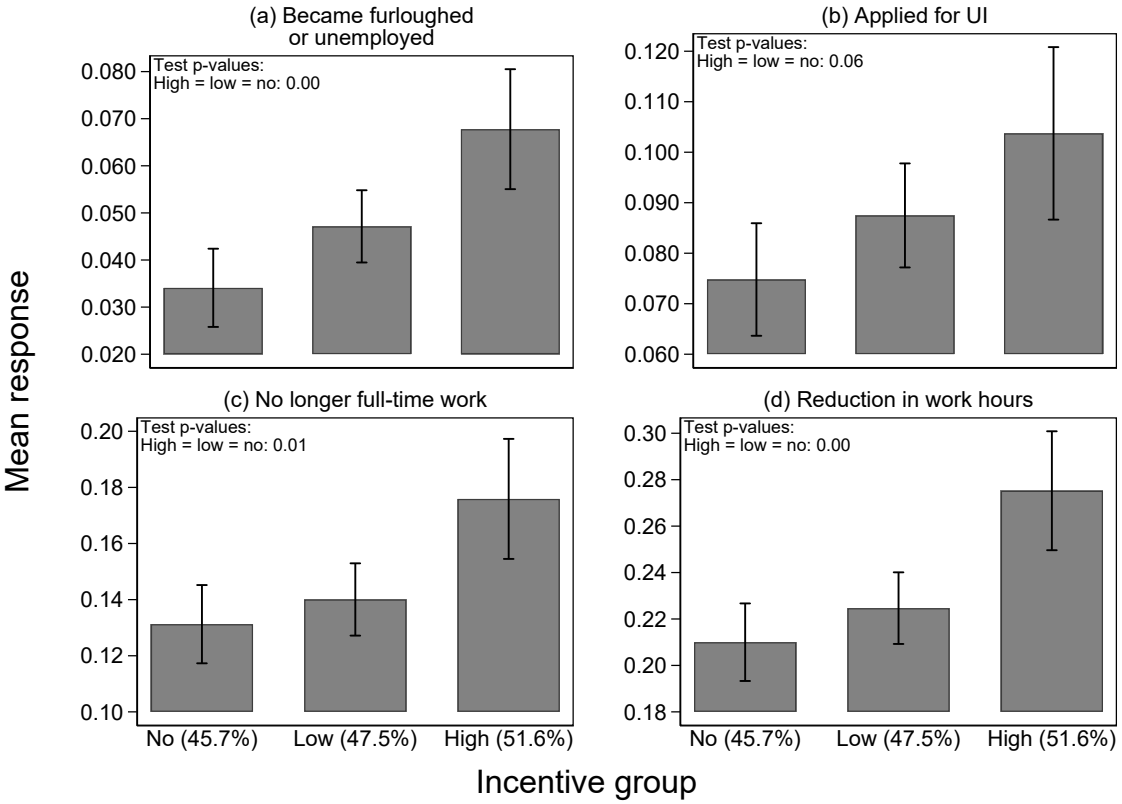
Figure 8 reports average responses by incentive arm for the survey-elicited measures discussed in Section 3.5.[30] The measures indicate that participants were negatively affected by the lockdown in all incentive arms, but the magnitudes differ substantially. For example, whereas 10.4 percent of participants in the high-incentive survey applied for UI benefits, only 7.5 percent in the no-incentive survey did. Participants in the high-incentive survey were also more likely to become furloughed or unemployed, no longer work full-time, and experience a reduction in work hours after the lockdown. For each outcome, we reject a joint test of equality in response means between the three survey arms, with all p–values under 0.1. These results show that participants differ from nonparticipants not only in their characteristics (as we found in the administrative data), but also in their responses to the survey, thus providing direct evidence of nonresponse bias.

We repeat the same analysis after reweighting to correct for selection on observables, using the same specifications as in Section 4.2. The results are reported in Panels C and D of Appendix Table A.6. Reweighting by municipality characteristics hardly affects the magnitude of the estimates. Reweighting by individual-level characteristics has a larger impact on the estimates, but the *differences* between the incentive arms typically increase rather than decrease, further highlighting the importance of selection on unobservables.[31] For each reweighting specifications and outcome, we reject the null that all selection is due to observables, with all p–values < 0.1.

[30] Appendix Table A.6 reports participant means for survey-elicited measures in table form.

[31] For example, the individually-reweighted no- and high-incentive participant estimates for becoming furloughed or unemployed differ by 7.4 percentage points, which is 4 percentage points larger than the difference in the unweighted estimates.

Figure 8: Nonresponse bias and selection in survey responses



Notes: This figure shows participant responses means by incentive level for survey-elicited outcomes. Error bars represent 90% confidence intervals. P–values for testing the joint equality across incentive groups are shown in upper left corner. Panel A of appendix Table A.6 presents estimated participant means and standard errors by incentive level and outcome.

Our findings show that the estimates a researcher would have obtained from the NCT survey are highly sensitive to the offered incentive level. These differences are large enough to have important policy implications. For example, estimated expenditures on UI benefits vary drastically depending on the considered incentive arm: while the no incentive arm would indicate that UI benefits account for 13.2 percent of total budgeted expenditures for Norwegian social insurance programs in 2020, the high incentive arm would indicate that this value is 18.4 percent.[32]

4.4 Characterizing inframarginal and marginal participants

Comparing participants from two different incentive arms involves a comparison among two types of individuals. There are the *inframarginal* individuals who participate in the higher incentive arm, but who would have participated in the lower incentive arm as well. Then there are the *marginal* individuals who participate in the arm with the higher incentive, but

[32]The Norwegian social insurance programs include old age pensions, sickness and disability insurance benefits, social benefits, health care insurance, parental leave benefits, and unemployment insurance benefits. Total budgeted expenditures on national insurance amounted to about 35 percent of the state budget in 2020 (Ministry of Finance, 2020).

would not have participated in the arm with the lower incentive. The *average* responses of the inframarginal participants may be materially different from those of the marginal participants even if responses vary smoothly with the willingness to participate in the survey.

Identification of responses of inframarginal and marginal participants

We can separate average outcomes for marginal and inframarginal participants with a simple model of the participation decision. For any $z \in \mathcal{Z}$, let $R_i(z)$ denote whether individual i *would have* participated if they had received incentive level z. If Z_i is the incentive individual i actually received, then their participation decision is

$$R_i = \sum_{z \in \mathcal{Z}} \mathbb{1}[Z_i = z] R_i(z). \tag{1}$$

We assume that any individual who would participate in the survey with one incentive would also participate with a larger incentive, or that $\mathbb{P}[R_i(z') \geq R_i(z)] = 1$ whenever $z' \geq z$. This is the well-known monotonicity condition introduced by Imbens and Angrist (1994). We continue to maintain that the incentives themselves do not directly affect responses as per the notation introduced in Section 4.1.[33]

These two assumptions allow us to estimate mean responses among the groups of individuals who are marginal or inframarginal to the incentives. If $z = 0$ denotes the smallest incentive (in our case, no incentive), then inframarginal individuals have $R_i(0) = 1$. Since they participate without incentives, they would also participate at higher incentives so that $R_i(z) = 1$ for all z. The average response for these inframarginal individuals is identified by

$$\mathbb{E}[Y_i | R_i = 1, Z_i = 0] = \mathbb{E}[Y_i^\star | R_i(0) = 1]. \tag{2}$$

We estimate the left-hand side of (2) by taking a sample mean. The marginal individuals, who do not participate at incentive level z, but would participate at incentive level $z' > z$, have $R_i(z') = 1$ but $R_i(z) = 0$. Using a similar argument to the one in Imbens and Angrist (1994), their average responses are identified by

$$\frac{\mathbb{E}[Y_i R_i | Z_i = z'] - \mathbb{E}[Y_i R_i | Z_i = z]}{\mathbb{P}[R_i = 1 | Z_i = z'] - \mathbb{P}[R_i = 1 | Z_i = z]} = \mathbb{E}[Y_i^\star | R_i(z) = 0, R_i(z') = 1]. \tag{3}$$

When contrasting two incentive levels, we estimate the left-hand side of (3) through an instrumental variables regression with $Y_i R_i$ as the outcome variable (letting $Y_i R_i = 0$ if $R_i = 0$), R_i as the endogenous variable, and Z_i as the instrument.

How do inframarginal and marginal participants differ?

Table 2 reports average labor market outcomes using both the administrative data and NCT survey data for the inframarginal group that participates without incentives, and the

[33] We verified this assumption in Section 4.3, both with and without access to administrative data.

marginal group that participates only under high incentives.[34] The estimates show that marginal participants had much stronger pre-lockdown labor market attachment.[35] For example, marginal participants earned an average of 6,806 USD per month, while inframarginal participants earned an average of 3,666 USD per month (p–value 0.08). In contrast, marginal and inframarginal participants had similar outcomes after the lockdown, with the earnings for both groups being roughly 3,600–3,800 USD per month, and statistically indistinguishable.

Consistent with these findings, the survey responses show that marginal participants were hit substantially harder by the lockdown. Table 2 shows that marginal participants were much more likely to become furloughed or unemployed, apply for UI, and experience a reduction in work hours. Marginal participants were also far more likely to experience a large loss of earnings and lose employment after the lockdown. These differences are all significant at the 5 percent level, and are large in magnitude. For example, 32.3 percent of marginal participants became furloughed or unemployed after the lockdown, compared to just 3.4 percent of inframarginal participants.

Appendix Table A.8 reports estimates of differences between marginal and inframarginal participants in their background characteristics: age, gender, immigrant status, and years of schooling. None of these differences are statistically significant at any conventional level, and a joint test of equality fails to reject with a p–value of 0.70. The fact that marginal and inframarginal participants differ so dramatically in their labor market outcomes before the lockdown, as well as in changes during the lockdown, and yet do not differ on observable background characteristics provides another strong indication of selection on unobservables.

5 Correcting for nonresponse bias due to selection on unobservables

Our findings in the previous section show substantial nonresponse bias due to selection on unobservables. In this section, we attempt to correct for selection on unobservables by applying methods from the treatment effects literature. We evaluate the methods using labor market outcomes from the administrative data. Since these outcomes are observed for both participants and nonparticipants, we can compare the different methods by their ability to reproduce the population mean, $\mathbb{E}[Y_i^\star]$, when using only data on the participants.

5.1 Worst-case bounds

In an influential paper, Manski (1989) observed that non-trivial bounds can be placed on the population mean by assuming that $\mathbb{E}[Y_i^\star|R_i = 0]$ is bounded between two known values, $\underline{y}$ and $\overline{y}$. Horowitz and Manski (1998) describe these bounds as "worst-case." The top row of each panel of Figure 9 reports worst-case bounds for our six outcomes. For the four binary

[34]The conclusions are similar, but estimates are noisier, when comparing inframarginals, marginals induced by low incentives, and marginals induced by high incentives. These results are reported in Appendix Table A.9.

[35]These results are not driven by outliers. To show this, Appendix Figure A.1 plots the estimated distributions of earnings before and after lockdown separately for the inframarginal and marginal participants. We find that the differences in the two types of participants come from differences in lower values of the distributions, and not due to outliers at the right tail.

Table 2: Instrumental variable estimates

	Inframarginal participant		Marginal participant		Inframarginal = Marginal
	Est.	(SE)	Est.	(SE)	p–value
Panel A: Administrative data					
Earnings before lockdown	3,666	(106)	6,806	(1,740)	0.08
Employed before lockdown	0.629	(0.012)	1.023	(0.199)	0.06
Earnings after lockdown	3,648	(100)	3,894	(1,471)	0.87
Employed after lockdown	0.571	(0.012)	0.618	(0.179)	0.80
Earnings loss larger than 20%	0.128	(0.009)	0.420	(0.145)	0.05
Employment loss	0.079	(0.007)	0.362	(0.125)	0.03
Panel B: NCT survey data					
Became furloughed or unemployed	0.034	(0.005)	0.323	(0.103)	0.01
Applied for UI	0.075	(0.007)	0.319	(0.115)	0.04
No longer full-time work	0.131	(0.009)	0.514	(0.159)	0.02
Reduction in work hours	0.210	(0.010)	0.767	(0.204)	0.01
Population share	45.7%		5.9%		

Notes: This table presents the estimated average labor market outcomes of individuals inframarginal and marginal to incentives. These values are estimated using an instrumental variables regression with $Y_i R_i$ as the outcome variable, survey participation R_i as the endogenous variable, and the set of indicators for incentive groups Z_i as the instrument. We reject joint tests of equality between inframarginal and marginal participants for both administrative outcomes (p–value 0.02) and survey responses (p–value <0.01). The total number of invited individuals is 9,366. Of these, we estimate that 45.7% are inframarginal participants, 5.9% are marginal participants, and 48.4% are nonparticipants.

outcomes, we take $\underline{y} = 0$ and $\overline{y} = 1$. For earnings before and after lockdown, which are continuous outcomes, we choose $\underline{y}$ and $\overline{y}$ to be the 1st and 99th percentile of the participant outcome distribution, like Lechner (1999) and Gonzalez (2005).[36] Although the bounds contain the actual population mean, they are quite wide. For example, we estimate that employment before lockdown is between 30 percent and 83 percent, while the actual value is 57 percent.[37]

5.2 Randomized incentives

Random assignment of incentives justifies assuming that $\mathbb{E}[Y_i^\star] = \mathbb{E}[Y_i^\star | Z_i = z]$ for each incentive level, z. Imposing random assignment narrows the worst-case bounds to the intersection of the worst-case bounds for $\mathbb{E}[Y_i^\star | Z_i = z]$ across each level z (Manski, 1990, 1994). The intersected bounds are necessarily narrower (weakly) than the worst-case bounds that pool participants across incentive levels. The second rows of Figure 9 shows that in our case, using incentives as instruments tightens the worst-case bounds only slightly. The resulting bounds contain the truth, but remain wide. Employment before lockdown is estimated to be between 34 and 83 percent, so that the width of the bounds is reduced by 8.5 percent relative to worst-case bounds.

[36] We use the same values of $\underline{y}$ and $\overline{y}$ in all of the subsequent results.

[37] Manski (2016) obtained much tighter worst-case bounds on employment in the March Current Population Survey. However, the nonresponse rate for the employment question in the CPS is roughly 5%, much lower than the 53% nonresponse rate in our survey or than the nonresponse rates of most surveys used in economics research (recall Figure 2).

5.3 Monotone responses

Manski and Pepper (2000) proposed adding a monotonicity assumption for outcomes with respect to a covariate, an assumption they described as monotone instrumental variables (IV). We do not have many covariates that make attractive candidates for this assumption. A potential exception is gender. Among the survey participants in our data, we find that men were more likely to be employed and had higher earnings, while being more likely to have a large earnings loss or to lose their employment during the lockdown. Using gender as a monotone IV means assuming that these relationships also hold among nonparticipants. The third rows of Figure 9 show that the assumption adds little information, and the bounds continue to be wide.

Manski and Pepper (2000) also considered a monotonicity assumption on the direction of selection bias, which they termed monotone treatment selection. For surveys, the analogous assumption can be described as monotone (positive or negative) response selection. Positive monotone response selection is the assumption that

$$\mathbb{E}[Y_i^\star | R_i = 1, Z = z] \geq \mathbb{E}[Y_i^\star | R_i = 0, Z = z] \quad \text{for all } z, \tag{4}$$

so that individuals who participate in the survey have, on average, larger outcomes than those who do not. Negative monotone response selection is the reverse inequality.

We impose monotone response selection assumptions in the directions implied by the data. For example, the evidence in Section 4 was consistent with those more reluctant to participate also being more likely to be employed before the lockdown, so we accordingly assume that nonparticipants are even more likely to be employed. The resulting bounds when adding this assumption are shown in the bottom rows of Figure 9. Monotone response selection narrows the bounds appreciably for all outcomes, and especially for employment before and after lockdown. However, the bounds remain wide and do not contain the population mean for any of the six outcomes.

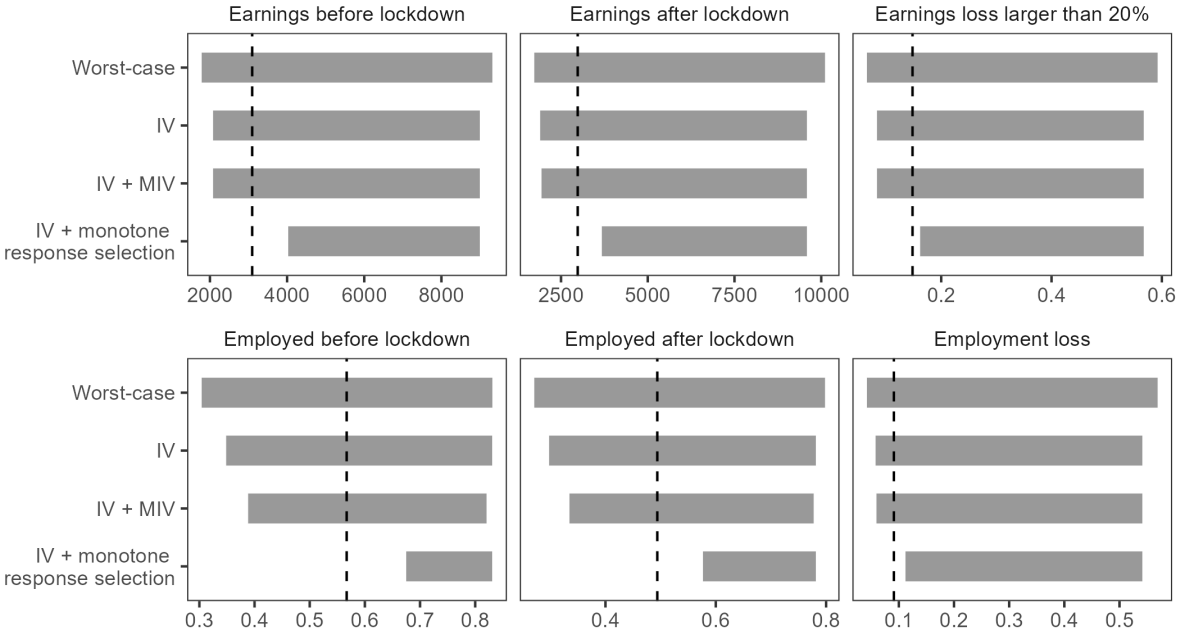
5.4 Selection model

The Imbens and Angrist (1994) monotonicity condition used in Section 4.4 provides a simple model of response behavior that can be combined with additional assumptions to correct for selection on unobservables. Vytlacil (2002) showed that the monotonicity condition is equivalent to assuming that participation follows an equation of the form

$$R_i = \mathbb{1}[U_i \leq p(Z_i)], \tag{5}$$

where U_i is an unobservable resistance to participating, and $p(z) \equiv \mathbb{P}[R_i = 1 | Z_i = z]$ is the propensity score. The unobservable U_i is independent of the assigned incentive, Z_i, due to random assignment, and normalized to have a uniform distribution on $[0, 1]$. However, it can be dependent with $Y_i^\star$, allowing for selection on unobservables. An individual's U_i

Figure 9: Estimated bounds using assumptions on the distribution of latent responses



Notes: The panels in this figure show estimated bounds under various assumptions on the distribution of the missing data. Each panel presents results for one of the six administrative outcomes. For each panel, the actual population mean is presented as a vertical dashed line. Bounds are constructed using the "no" and "high" incentive samples. In the first row (Worst-case), we assume that the mean of nonparticipants is bounded between 0 and 1 for binary variables and between the 1st and 99th percentiles of the observed distributions for continuous variables. In the second row (IV), we maintain the bounded assumption and also impose that incentives were randomly assigned. In the third row (IV + MIV), we maintain the IV assumptions and also impose the MIV gender assumption that mean responses for both participants and nonparticipants are larger for males for all outcomes. In the fourth row (IV + monotone response selection), we maintain the IV assumptions and also impose the monotone response selection assumptions in the direction implied by the data (positive for all outcomes).

characterizes their quantile of willingness to take the survey, with lower values being more willing, and higher values less willing. For example, continuing to denote $z = 0$ as no incentive and letting $z = 1$ denote high incentive, individuals with $U_i \in (p(0), p(1)]$ are the marginal participants who would participate if and only if offered the incentive, whereas individuals with $U_i \leq p(0)$ are the inframarginal participants, who would take the survey with or without an incentive.

Selection models like (5) have a long tradition, dating back to Gronau (1974) and Heckman (1974, 1979). We consider the modern nonparametric interpretation developed by Heckman and Vytlacil (2005, 2007), which is organized around the marginal survey response (MSR), $m(u) \equiv \mathbb{E}[Y_i^\star | U_i = u]$.[38] The MSR is the average response for individuals with uth quantile of willingness to participate. The population mean is the integral of the MSR over $[0, 1]$, so assumptions about the MSR can help tighten inference on the population mean. Brinch et al. (2017), Mogstad et al. (2018), and Mogstad and Torgovitsky (2018) show how to identify or bound the population mean under various types of parametric and nonparametric

[38]In treatment evaluation contexts this would be called the marginal *treatment response* with the difference between two marginal treatment responses being called the marginal treatment *effect*.

assumptions. We apply the methodology of Mogstad et al. (2018) to the survey setting in what follows; see Online Appendix I for more details.

Figure 10 contains bounds and point estimates of the population mean for our six outcomes under a variety of assumptions on the MSR. The first row requires $m(u)$ to lie between $\underline{y}$ and $\overline{y}$ for each u, with the same choices for these a priori bounds as in the previous sections. These bounds are known to be equal to the bounds that use randomized incentives from Section 5.2 (Heckman and Vytlacil, 2001). Throughout, we require $m(u)$ to lie between these values.

In the second rows of Figure 10, we assume that the MSR is a monotone function of latent willingness to participate. We set the directions of monotonicity to be the same as for the monotone response selection assumption in the previous section. The content of the assumption is similar when phrased in terms of the MSR, but stronger, since it requires the assumed direction of monotonicity to hold when comparing individuals by their propensity to participate, rather than just their participation decision. However, the result is quite similar: the bounds are substantially tighter relative to only assuming the MSR is bounded, but miss the population mean, sometimes by a wide margin.

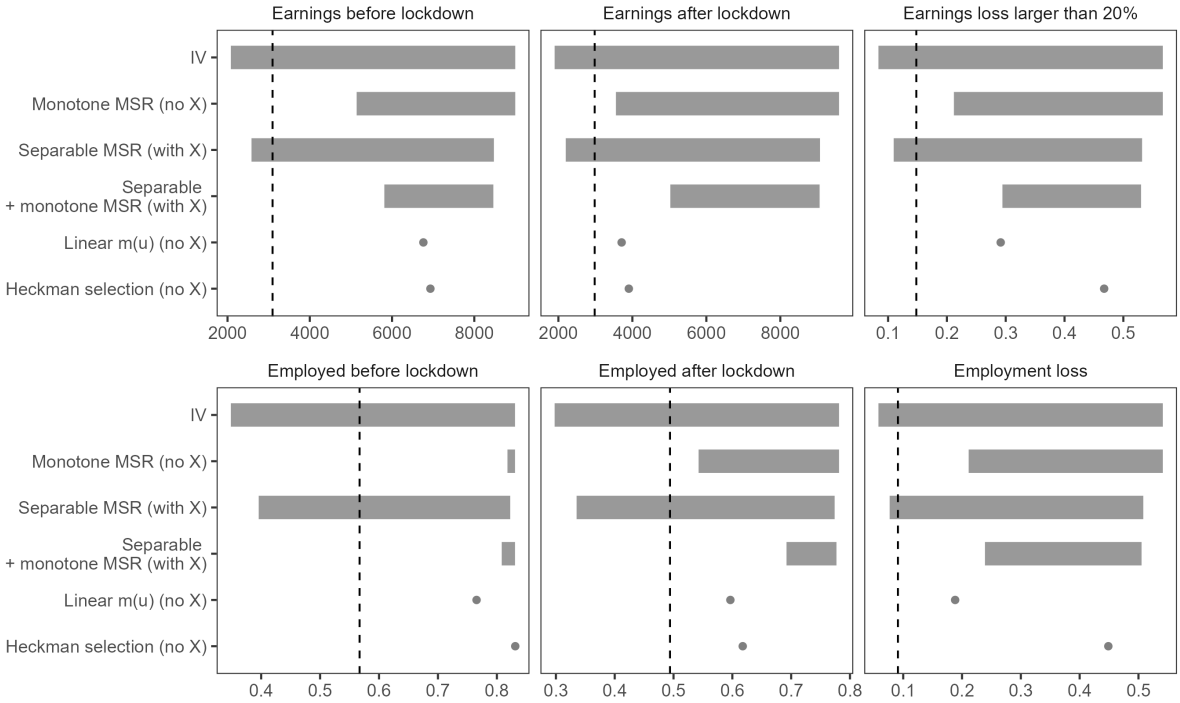
Since the bounds under monotonicity do not contain the population mean, the MSR functions cannot be monotone for any of the six outcomes. Taking pre-lockdown employment as an example, the results in Table 2 imply that the marginal participants must have larger likelihood of employment than the inframarginal participants, so that $m(u)$ is increasing in u for at least some values of u smaller than $p(1)$. But because the bounds under monotonicity do not contain the actual population mean (Figure 10), and because the population mean is the integral of m over $[0, 1]$, it must be that $m(u)$ eventually starts decreasing for values of u larger than $p(1)$.

The next two rows of Figure 10 impose the assumption that the MSR is separable in a covariate X_i, as in Carneiro et al. (2011) and Brinch et al. (2017), among others. With X_i as gender, the separability assumption is that the MSR conditional on gender has the form $m(u, x) \equiv \mathbb{E}[Y_i^\star | U_i = u, X_i = x] = m_U(u) + m_X(x)$ for functions m_U and m_X. The interpretation is that the relationship between willingness to participate and labor market outcomes (m_U) is the same for both men and women, but that men and women may differ by a constant for all values of $U = u$. In our setting, separability turns out to narrow the bounds somewhat, but they remain wide. In the fourth row of Figure 10, we combine separability with the assumption that m_U is monotone in u.[39] The resulting bounds are sometimes fairly tight, but still provide misleading estimates of the population means for all outcomes.

In the bottom two rows of Figure 10, we take a different approach and parameterize $m_U(u)$ (without covariates) to point identify the population mean. In the fifth row, we assume that $m_U(u)$ is a linear polynomial. In the sixth row, we assume that $m_U(u)$ is a linear function

[39]Despite having rejected monotonicity without separability, this does not imply that we reject it when combined with separability, because it's possible that differences in participation rates by gender, together with $m_X(x)$, are driving the observed monotone direction when we omit X_i.

Figure 10: Bounds using assumptions on the MSR



Notes: The panels in this figure show estimated bounds under the selection model in (5) and under various assumptions on the marginal survey response function (MSR). Each panel presents results for one of the six administrative outcomes. For each panel, the actual population mean is presented as a vertical dashed line. Bounds are constructed using the "no" and "high" incentive samples. For all sets of assumptions, we assume that the MSR is bounded between 0 and 1 for binary variables and between the 1st and 99th percentiles of the observed distributions for continuous variables. In the first row (IV), we make no further assumptions, and the bounds correspond to the IV bounds in Figure 9. See Online Appendix I for details on the other imposed assumptions and construction of estimated bounds.

of $\Phi^{-1}(u)$, the standard normal quantile function, which is the same parameterization used in the Heckman (1974, 1979) selection model.[40] In both cases, we obtain point estimates, but they are far from the population mean. For example, estimates of employment before lockdown are 77 percent under a linear parametrization and 83 percent under a Heckman selection model, when the true value is 57 percent. We find similarly misleading estimates for the other outcomes.

5.5 Understanding the failure of existing methods

Accounting for selection on unobservables can be viewed as an extrapolation problem, where the data on participants is used, together with some assumptions, to draw inference about the nonparticipants (e.g., Mogstad and Torgovitsky, 2018). Some of the assumptions used for extrapolation in this section produced bounds that, while containing the target population mean, are likely to be too wide to be useful for most purposes. Other assumptions produced

[40]For binary outcomes we estimate a bivariate probit by maximum likelihood. For continuous outcomes, we estimate a two-step Heckman selection model. Note that the functional form used by the Heckman selection model implies an MSR that is unbounded, and thus does not incorporate the bounded MSR restriction we maintain for the other methods.

tight bounds (or point estimates) that failed to reproduce the population mean, implying that the assumptions do not hold.

In some cases, even weak assumptions led to severely incorrect conclusions about the population mean. For example, the second row of Figure 10 for earnings before the lockdown was based on the following assumptions: incentives were randomized, individuals are more willing to participate with incentives than without (the monotonicity condition), and the relationship between their earnings and willingness to participate is monotonic along unobservables. All of these assumptions are used in many contexts in economics and in the social sciences more broadly. Yet the endpoints of the interval estimate generated based on these assumptions are 5,813 and 8,463, severely overestimating the true average pre-lockdown earnings of 3,095.

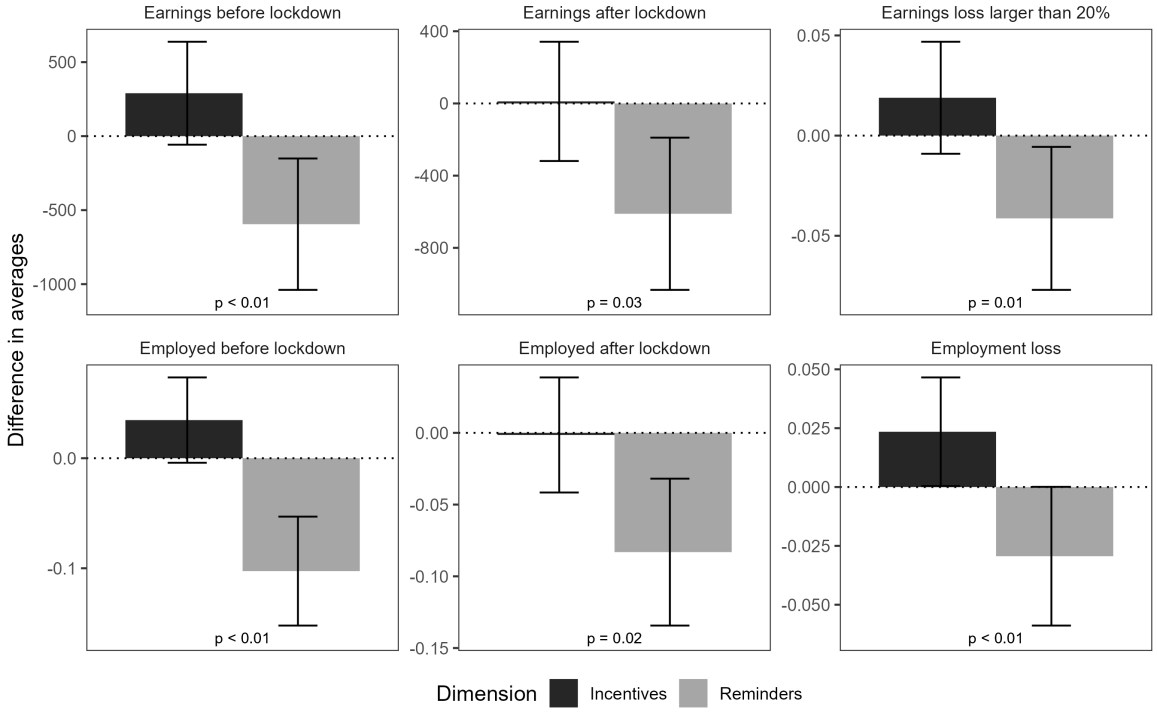
One explanation is that there are multiple types of nonparticipants who differ in fundamental ways. For instance, suppose that there are two types of nonparticipants: active nonparticipants who saw an invitation email and declined to participate because the incentive was too low, and passive nonparticipants who never saw an email, but might have participated had they seen one. These two groups might differ from the participants in opposite ways. It could be, for example, that the active nonparticipants have larger opportunity costs of time than the participants, while the passive nonparticipants have weaker labor market attachment. Imposing assumptions about the relationship between participants and nonparticipants that implicitly presume nonparticipants are of one type will fail if many of the nonparticipants are actually of the other type. Our finding that the MSR is not monotone (second row of Figure 10) is consistent with such an explanation.

We find additional evidence supporting this explanation when we split participants by when they responded. The darker bars of Figure 11 show average differences between the high and no incentive groups among participants who responded before the April 27th reminder (recall Figure 4). The lighter bars show average differences within the no incentive group between participants who responded after the reminder and participants who responded before the reminder. Across outcomes we find *positive* differences on the incentive dimension, but large *negative* differences on the pre/post-reminder dimension.[41]

These results suggest that participants differ along at least two unobservable dimensions. If the same is true of nonparticipants, then a model like (5) with a single source of unobserved heterogeneity could be badly misspecified. If nonparticipants are more similar to those who participate after the reminder than those induced by incentives, then only using variation across incentives to extrapolate could lead to the type of flawed estimates of the actual population mean seen in Figure 10.

[41] Interpreting the pre/post-reminder differences as reflecting nonresponse bias requires assuming that the underlying response $Y_i^\star$ is time-invariant. If the reminders were randomly assigned instead of being sent to all potential participants, then random assignment of the reminder could be used as an additional instrument similar to the randomly assigned incentive. We discuss this point further in Online Appendix J.

Figure 11: Selection by incentive and reminder



Notes: This figure compares mean differences between participants with and without incentives and based on whether they participate after or before the April 27th reminder. The darker bars show average differences between the high and no incentive groups among participants who responded before the reminder. The lighter bars show average differences within the no incentive group between participants who responded after the reminder and participants who responded before the reminder. 90% CIs are presented for each difference. At the bottom of each panel, we present the p-value for the test that the differences in means are equal.

6 A model of participation with financial incentives and reminders

We now develop and apply a model that incorporates a distinction between active nonparticipation (declining to participate) and passive nonparticipation (not being aware of the survey). The model allows for variation in participation both over time and due to randomly-assigned incentives. We show theoretically and empirically how to use the model to correct for nonresponse bias and produce either bounds or point estimates on population average outcomes under different auxiliary shape restrictions.

6.1 Model

Decisions and periods. Divide the survey horizon into $t = 1, \ldots, T$ time periods. Individual i sees a survey invitation email in period S_i. The email informs them of their assigned incentive level, Z_i. They then choose to take the survey if $V_i \leq \eta(Z_i)$, where V_i is a latent variable and η is an increasing function. As in Section 5, we keep the incentive binary (no or high incentive) so that $z \in \{0, 1\}$ with $\eta(0) < \eta(1)$. Let $S_i = T + 1$ if individual i never sees an invitation email and thus never makes an active participation choice. Both V_i and S_i are unobserved to the researcher.

Let $R_{it}(z)$ be an indicator for whether individual i would have participated in the survey at or before time t if they had been assigned incentive z. Then $R_{it}(z)$ is one if individual i sees an invitation email before period t and decides to participate:

$$R_{it}(z) = \underbrace{\mathbb{1}[S_i \leq t]}_{\text{sees invitation email before } t} \overbrace{\mathbb{1}[V_i \leq \eta(z)]}^{\text{would decide to participate with incentive } z}. \tag{6}$$

We do not directly observe $R_{it}(z)$. Instead, we observe $(\{R_{it}\}_{t=1}^{T}, Z_i)$, where

$$R_{it} = Z_i R_{it}(1) + (1 - Z_i) R_{it}(0) \quad \text{for } t = 1, \ldots, T. \tag{7}$$

As before, we observe an individual's response as $Y_i = Y_i^\star$ if and only if they respond during the survey horizon ($R_{it} = 1$ for some $t \leq T$, and thus $R_{iT} = 1$).

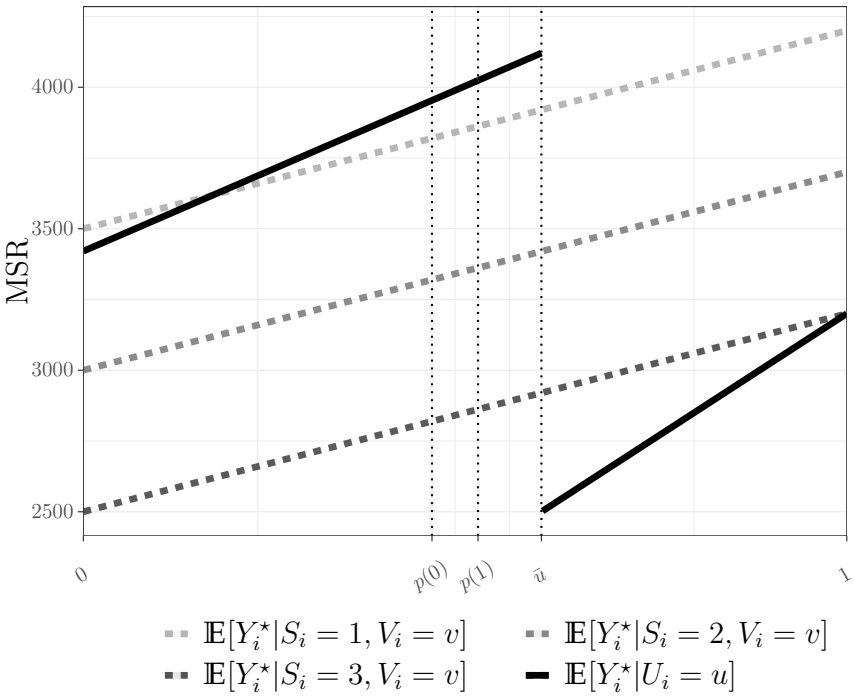
The two dimensions of heterogeneity. The model has two dimensions of unobserved heterogeneity, V_i and S_i, whereas the model in equation (5) in Section 5.4 had only one dimension, U_i. The active dimension, V_i, is, like U_i from Section 5.4, individual i's latent resistance to incentives to participate. We normalize V_i to have a uniform distribution on $[0, 1]$. The passive dimension of heterogeneity—when the individual sees an invitation email, S_i—is a categorical variable that takes on $T + 1$ values. The assumption that incentives are randomly assigned now means that Z_i is independent of $(S_i, V_i, Y_i^\star)$.

Variation in response rates due to randomized incentives, Z_i, provides information about the active dimension, V_i. Variation in response rates over time provides information about the passive dimension, S_i. To ensure sufficient variation over time, we take $T = 2$, with the first period being before April 27th and the second period being after the last major reminder email was sent on April 27th. The bump in participation after the April 27th reminder email (Figure 4) is then akin to a first stage for the passive dimension.

Benefits of two dimensions of heterogeneity. Having two dimensions of unobserved heterogeneity can help explain the difficulties with extrapolation using the one-dimensional model in Section 5.4. Individuals who didn't respond when incentivized had $U_i > p(1)$ in the one-dimensional model, while in the two-dimensional model they could have either $V_i > \eta(1)$ (incentives not high enough) or $S_i = 3$ (never saw an email), or both. They differ from individuals with $U_i \leq p(1)$ along two dimensions, since these individuals would have participated with an incentive, and so in the two-dimensional model must have both $V_i \leq \eta(1)$ and $S_i \leq 2$. Values of U_i larger than $p(1)$ initially correspond to individuals who saw an invitation email ($S_i \leq 2$) and would have participated at some larger incentive ($V_i > \eta(1)$). As U_i increases, it eventually starts to correspond to passive nonparticipants who never saw an invitation email ($S_i = 3$) and so would not have participated regardless of how large an incentive was offered.

This shift in unobservables makes reliable extrapolation difficult under the model in Section 5.4. In Section 5.5, we presented evidence that individuals who participated after the

Figure 12: Extrapolation with the one-dimensional model when heterogeneity is two-dimensional



Notes: This figure illustrates the problem of using a model with one dimension of heterogeneity to extrapolate to the population mean when the true heterogeneity is two-dimensional. The figure shows the marginal survey responses (MSR) as a function of the incentive heterogeneity (V_i in the two-dimensional model of this section and U_i in the one-dimensional model of Section 5.4). Dashed lines in grey present the true two-dimensional MSR: light grey depicts the MSR for individuals seeing an invitation email before the reminder, grey depicts the MSR for individuals seeing an invitation email after the reminder and dark grey depicts the MSR for individuals who never see an invitation email. The solid line in black shows the MSR if the two-dimensional heterogeneity is collapsed into a single dimension of heterogeneity. The map between dimensions is defined as follows. Define $\bar{u} = 1 - \mathbb{P}[S_i = 3]$, $\psi_1(x) = \mathbb{P}[V_i \leq x | S_i \leq 2]$, and $\psi_2(x) = \mathbb{P}[V_i \leq x | S_i = 3]$. Then $U_i = \mathbb{1}[S_i \leq 2][\psi_1(V_i)\bar{u}] + \mathbb{1}[S_i = 3][\bar{u} + \psi_2(V_i)(1-\bar{u})]$. For this illustration, we assume V_i is independent of S_i, $\bar{u} = 0.6$, and take $\mathbb{E}[Y_i^\star|S_i = s, V_i = v] = 700v + 3500\mathbb{1}[S_i = 1] + 3000\mathbb{1}[S_i = 2] + 2500\mathbb{1}[S_i = 3]$. The conditional expectation was chosen based on variations in earnings by incentives and reminders that we observe in administrative data.

April 27th reminder ($S_i = 2$ and $V_i \leq \eta(1)$) had substantially lower earnings and employment rates than those who participated before the reminder ($S_i = 1$ and $V_i \leq \eta(1)$). It is reasonable to expect that individuals who never saw an invitation email ($S_i = 3$) differ from both these groups, even if they would have participated ($V_i \leq \eta(1)$) had they been aware of the survey. If that's true, then the model in Section 5.4 implies an MSR function $\mathbb{E}[Y_i^\star|U_i = u]$ that is discontinuous in u. Extrapolating a discontinuous function is naturally rather difficult.

Figure 12 illustrates this argument with a numerical example. The figure plots $\mathbb{E}[Y_i^\star|V_i = v, S_i = s]$ as a function of v for each of the three values of $s \in \{1, 2, 3\}$ (dashed grey lines). The magnitude and values of the function are chosen to be roughly consistent with the variation in earnings by incentives and reminders that we observe in the administrative data. Average responses differ across individuals by a level shift based on when they see an invitation email, with those in the first period ($S_i = 1$) having the highest earnings, and those who did not see an email ($S_i = 3$) having the lowest earnings. Within these groups there is additional

heterogeneity along the incentive dimension, with individuals who are less responsive to incentives (higher V_i) having higher earnings.

The model in Section 5.4 combines these two dimensions of unobserved heterogeneity into a single dimension, shown in Figure 12 as $\mathbb{E}[Y_i^\star|U_i = u]$ (solid black line). All individuals with $S_i = 3$ must have U_i towards 1, since they wouldn't be induced by any incentive to participate. As a result, $\mathbb{E}[Y_i^\star|U_i = u]$ changes discontinuously as u crosses $\bar{u} \equiv 1 - \mathbb{P}[S_i = 3]$, making extrapolation difficult. Even if we knew that the MSR were linear up to $p(1)$, using it to extrapolate beyond $\bar{u}$ would provide misleading conclusions because of the discontinuity.

6.2 Participation groups with two-dimensional heterogeneity

The two-dimensional participation model allows for five distinct participation groups, or configurations of $(R_{i1}(0), R_{i1}(1), R_{i2}(0), R_{i2}(1))$. Table 3 lists these groups together with the realizations of (V_i, S_i) that characterize them. Always-takers participate in the first period, regardless of incentives, while never-takers don't participate in either period, even with the incentive. Incentive compliers participate in the first period if they receive an incentive, but not otherwise. Reminder compliers participate in the second period after the April 27th reminder email is sent, whether incentivized or not. Reluctant compliers only participate in the second period after the reminder email, and only if they also are incentivized.

Population shares of participation groups. The share of each participation group is identified. The share of always-takers is given by

$$\mathbb{P}[R_{i1} = 1|Z_i = 0] = \mathbb{P}[R_{i1}(0) = 1] = \mathbb{P}[S_i = 1, V_i \leq \eta(0)].$$

The share of incentive compliers is then

$$\mathbb{P}[R_{i1} = 1|Z_i = 1] - \mathbb{P}[R_{i1} = 1|Z_i = 0] = \mathbb{P}[S_i = 1, \eta(0) < V_i \leq \eta(1)].$$

Similarly, the share of reminder compliers is given by $\mathbb{P}[R_{i1} = 0, R_{i2} = 1|Z_i = 0]$, and the share of reluctant compliers by $\mathbb{P}[R_{i1} = 0, R_{i2} = 1|Z_i = 1] - \mathbb{P}[R_{i1} = 0, R_{i2} = 1|Z_i = 0]$. Because the five group shares must sum to one, the share of never-takers can be deduced from those of the other four groups.

Table 3 reports estimated group shares. The inframarginal and marginal groups under the single threshold model considered in Sections 4.4 and 5.4 are now split into four groups. Of the complier groups, the reminder and incentive compliers are the largest, with reluctant compliers comprising less than 1% of the population.

Average responses of participation groups. Average responses for the incentive compliers are given by

$$\frac{\mathbb{E}[Y_i R_{i1}|Z_i = 1] - \mathbb{E}[Y_i R_{i1}|Z_i = 0]}{\mathbb{P}[R_{i1} = 1|Z_i = 1] - \mathbb{P}[R_{i1} = 1|Z_i = 0]} = \mathbb{E}[Y_i^\star|S_i = 1, \eta(0) < V_i \leq \eta(1)],$$

Table 3: Participation group definitions and estimated shares

Group	Share (SE)	$R_{i1}(0)$	$R_{i1}(1)$	$R_{i2}(0)$	$R_{i2}(1)$	$V_i \in$		$S_i =$
Always-taker	.384 (.008)	1	1	1	1	$[0, \eta(0)]$	and	1
Incentive complier	.051 (.015)	0	1	0	1	$(\eta(0), \eta(1)]$	and	1
Reminder complier	.072 (.004)	0	0	1	1	$[0, \eta(0)]$	and	2
Reluctant complier	.009 (.008)	0	0	0	1	$(\eta(0), \eta(1)]$	and	2
Never-taker	.484 (.013)	0	0	0	0	$(\eta(1), 1]$	or	3

Notes: This table presents the estimated shares of the participation groups and their characterization based on their participation decision $R_{it}(z)$. The first column indicates the name of the participation group while the second presents the estimated population share in our survey (and its standard error in parenthesis). Columns 3 to 6 depict the groups' participation decision (1 for those who participate and 0 otherwise) under different states of $R_{it}(z)$, where $t = 1$ and $t = 2$ denote before and after April 27th, respectively, and $z = 0$ and $z = 1$ denote no incentive and high incentive, respectively. Columns 7 and 8 describe where the groups are located in the support of the two dimensions of the unobserved heterogeneity (V_i and S_i, respectively).

Table 4: Estimated average responses by group.

	Earnings			Employment		
	Before	After	Large Loss	Before	After	Loss
Always Taker (38%)	3,746	3,783	0.13	0.65	0.64	0.03
	(116)	(107)	(0.01)	(0.01)	(0.01)	(0.00)
Incentive Complier (5%)	6,766	3,944	0.31	0.92	0.70	0.13
	(1,900)	(1,546)	(0.14)	(0.20)	(0.19)	(0.08)
Reminder Complier (7%)	3,244	3,257	0.12	0.55	0.55	0.03
	(256)	(251)	(0.02)	(0.03)	(0.03)	(0.01)
Reluctant Complier (<1%)	7,030	3,920	0.84	1.35	1.38	0.27
	(5,158)	(3,903)	(0.72)	(0.85)	(0.87)	(0.27)

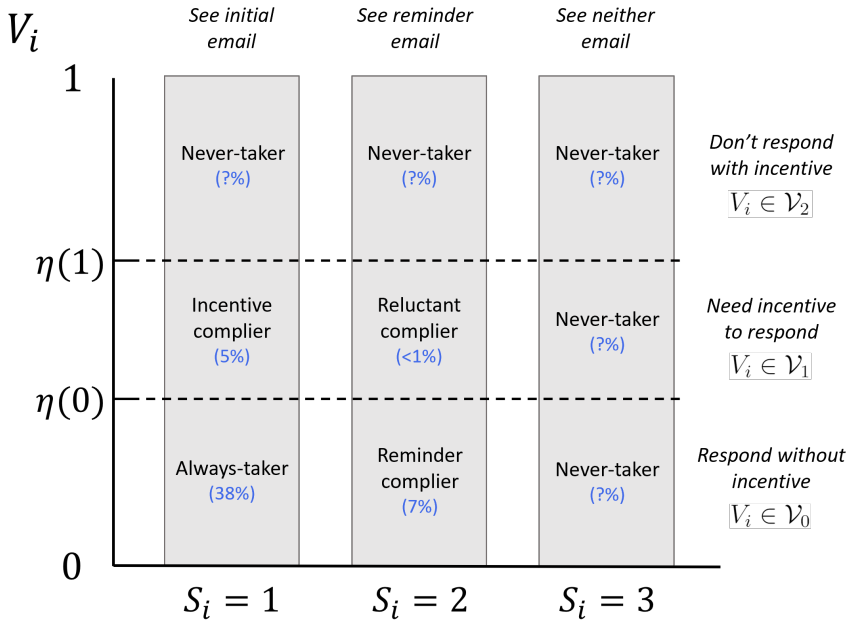
Notes: This table presents the estimated average responses for the participation groups on the six considered administrative outcomes (earnings and employment before lockdown, after lockdown, and loss). Always-taker and incentive complier group responses are estimated via an instrumental variables regression with $Y_i R_{i1}$ as the outcome variable, R_{i1} as the endogenous variable and Z_i as the instrument. Reminder complier and reluctant complier group responses are estimated via an instrumental variables regression with $Y_i(1 - R_{i1})R_{i2}$ as the outcome variable, $(1 - R_{i1})R_{i2}$ as the endogenous variable and Z_i as the instrument.

with similar arguments to identify average responses for the other groups. Average responses for the never-takers are not identified, because their responses are never observed.

Table 4 reports estimates of mean employment and earnings for the always-takers and three complier groups. Incentive compliers have higher employment and earnings than both always-takers and reminder compliers both before and after the lockdown. However, they are also more likely to have lost employment and suffered a large decline in earnings during the pandemic. In contrast, reminder compliers have lower employment and lower earnings than always-takers before and after the lockdown, and are less likely to have lost employment and suffered a large decline in earnings. The reluctant complier group is too small to draw firm conclusions about their average responses, which is reflected in the large standard errors.

The estimates suggest a situation consistent with Figure 12, with pronounced unobserved heterogeneity along both the financial incentive (V_i) and seeing (S_i) dimensions. The esti-

Figure 13: The structure of the extrapolation problem in the two-dimensional model.



Notes: This figure illustrates the nature of the problem of extrapolation under the two-dimensional heterogeneity model. The x-axis presents the seeing dimension (S_i). The y-axis depicts the heterogeneity in incentive responsiveness (V_i). Each area is labelled with the corresponding participation group and the estimated share (if identified).

mates also show that these two types of heterogeneity operate in opposing directions: for example, the incentive compliers have substantially higher monthly earnings before lockdown compared to the always-takers, while the reminder compliers have somewhat lower monthly earnings. If similar patterns occur among the group of never-takers, then using the one-dimensional model to extrapolate will run into the type of discontinuity problem illustrated in Figure 12.

6.3 Extrapolation

Figure 13 diagrams the structure of the extrapolation problem in the two-dimensional model. In terms of the latent variables, (V_i, S_i), there are nine sets representing all combinations of incentive heterogeneity (participate without incentive, only with incentive, not even with incentive) and email awareness (see in the first period, see after the reminder, never see). The always-takers and three complier groups each occupy one cell, so the masses and average responses in these cells are point identified. However, the never-takers are spread across five cells representing different combinations of V_i and S_i. The problem is to extrapolate responses from the four cells on which we have information to the five on which we don't.

Bounds on the population average response. For each $j \in \{0, 1, 2\}$ and $s \in \{1, 2, 3\}$, let $\mu_{js} \equiv \mathbb{E}[Y_i^\star | V_i \in \mathcal{V}_j, S_i = s]$, where the sets $\mathcal{V}_0, \mathcal{V}_1, \mathcal{V}_2$ are as shown in Figure 13. Similarly, let $\pi_{js} \equiv \mathbb{P}[V_i \in \mathcal{V}_j, S_i = s]$. Let $T_i = R_{i1} + 2(1 - R_{i1})R_{i2}$ denote the time period (1 or 2) in which individual i participated, if they participated, with $T_i = 0$ if they did not participate.

To be consistent with the observed data, a set of candidate values for (μ_{js}, π_{js}) must satisfy

$$\begin{aligned} \mathbb{E}[Y_i|T_i = 1, Z_i = 1] &= \mu_{01}\left(\frac{\pi_{01}}{\pi_{01}+\pi_{11}}\right) + \mu_{11}\left(\frac{\pi_{11}}{\pi_{01}+\pi_{11}}\right) \\ \text{and} \quad \mathbb{P}[T_i = 1|Z_i = 1] &= \pi_{01} + \pi_{11}, \end{aligned} \tag{8}$$

as well as similar equations for the three other combinations of $(T_i, Z_i) \in \{(1,1),(2,0),(2,1)\}$ (see Online Appendix K.1 for the equations). We let $Q(\mu, \pi) \in \mathbb{R}^8$ denote the difference in these eight equations.

Sharp bounds on the population average response, $\mathbb{E}[Y_i^\star]$, can then be found by solving the optimization problems

$$\min_{\pi \geq 0, \mu} / \max_{\pi \geq 0, \mu} \sum_{j=0}^{2}\sum_{s=1}^{3} \pi_{js}\mu_{js} \quad \text{s.t.} \quad Q(\mu, \pi) = 0 \quad \text{and} \quad \sum_{j=0}^{2}\sum_{s=1}^{3} \pi_{js} = 1. \tag{9}$$

When solving (9) we also constrain $\underline{y} \leq \mu_{js} \leq \overline{y}$ using the same a priori bounds $\underline{y}$ and $\overline{y}$ discussed in Section 5.4. While (9) is a non-convex program, it can still be solved to provable global optimality using spatial branch-and-bound algorithms (we use Gurobi Optimization (2021)). See Online Appendix K.1 for more details on implementation.

Comparing the one- and two- dimensional models. Figure 14 reports bounds and point estimates of population averages for six different outcomes in the administrative data. Each row corresponds to a different set of assumptions. Results for the two-dimensional model are shown in dark grey, while comparable results for the one-dimensional model in Section 5.4 are shown in light grey (when applicable).

As a benchmark, the first row ("IV") reports bounds that only use random assignment of the incentive together with the same a priori bounds on the outcome imposed in Section 5.4. The results for the two-dimensional model are identical to those from the one-dimensional model (and to those without any choice model), implying that the two-dimensional model by itself contains no identifying content.

In the second row, we assume that $\mu_{js} = \mu_j + \mu_s$ is separable. To match the patterns found in Table 4, we also assume that μ_j is increasing in j—so that those more reluctant to participate have higher labor market outcomes—and that μ_s is decreasing in s—so that those who see an email later have lower labor market outcomes. These assumptions have no effect on the bounds. Intuitively, the model still allows for the possibility that all nonparticipants have either high V_i and low S_i, or low V_i and high S_i. Without imposing any structure on the joint distribution of (V_i, S_i), the never-takers can be freely assigned across the five unknown cells of Figure 13. This makes it difficult to extrapolate.

We add structure in two ways. In Online Appendix K.2, we derive the strongest testable implication of V_i and S_i being independent. A bootstrap test of the implication fails to reject at all conventional significance levels with a p-value of .97. We thus assume that V_i and S_i are independent, so that the period in which an individual sees an email is unrelated to

their sensitivity to incentives. We also fix a proportion of the survey population that has $S_i = 3$, and so never sees an email invitation. For the results in the main text, we impose that $\mathbb{P}[S_i = 3] = .4$, so that 40% never see an email invitation to participate. We chose this number in consultation with the survey researchers at Statistics Norway, relying on their expertise in implementing email surveys in Norway. In Online Appendix K.4, we examine sensitivity to the 40% assumption and find that our results are largely similar if it is increased by 8 percentage points up to 48% (the largest it can be) or decreased to 32%.

Imposing these assumptions allows us to point identify the masses π_{js} in each region of Figure 13 (see Online Appendix K.3 for proof).[42] As a consequence, the unknown group means can be expressed as

$$\mu_{js} \equiv \mathbb{E}[Y_i^\star | V_i \in \mathcal{V}_j, S_i = s] = \frac{1}{\mathbb{P}[V_i \in \mathcal{V}_j]} \int_{\mathcal{V}_j} m(v, s)\, dv, \tag{10}$$

where $\mathbb{P}[V_i \in \mathcal{V}_j] = \pi_{j1} + \pi_{j2} + \pi_{j3}$ is point identified, and $m(v, s) \equiv \mathbb{E}[Y_i^\star | V_i = v, S_i = s]$ is the unknown two-dimensional MSR function. Using (10) allows us to impose assumptions directly on the MSR function rather than the higher-level group-specific means, μ_{js}. Implementation still follows (9), but now the program is linear in m, because π is identified.

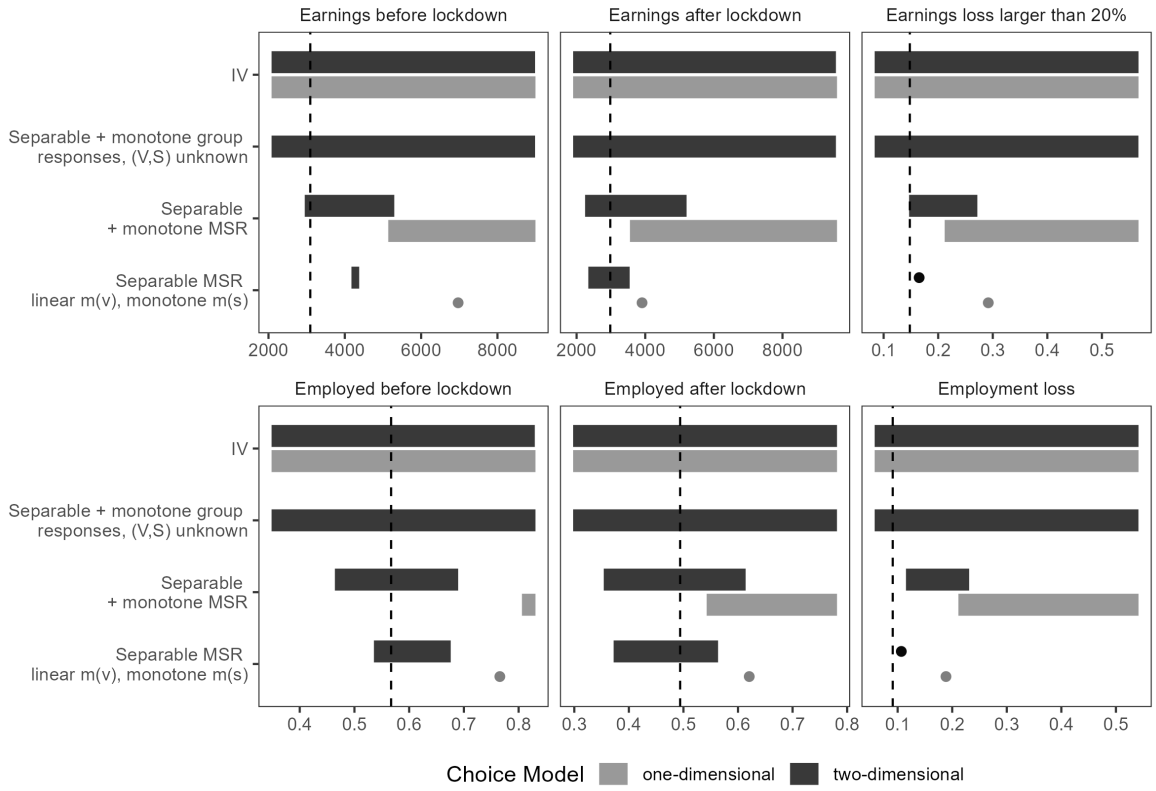
The results in the third row of Figure 14 maintain these assumptions on the joint distribution of (V_i, S_i). They also maintain separability and monotonicity assumptions similar to the second row of Figure 14, but now stated in terms of the MSR function rather than μ_{js}. The separability assumption is that $m(v, s) = m_V(v) + m_S(s)$, where m_V and m_S are unknown functions that determine unobserved heterogeneity in the active and passive dimensions. To match the patterns found in Table 4, we assume that m_V is increasing and m_S is decreasing, so that those more reluctant to participate have higher labor market outcomes and those who see an email later have lower labor market outcomes. The bounds are much narrower than in the second row. Compared to the one-dimensional model (light grey), the bounds are narrower for some outcomes, but wider for others. However, the two-dimensional model bounds contain the true population values for five out of the six outcomes, whereas the one-dimensional model bounds never do.

In the fourth row of Figure 14 we impose some parametric structure by requiring m_V to be linear, so that the relationship between the active dimension and outcomes is smooth, with a constant slope. We continue to assume that m_S is decreasing, but do not otherwise restrict its functional form. The bounds for most outcomes are tight, and in some cases points. They also get close to the truth in all cases. For example, we estimate that population average monthly earnings before the lockdown are between \$4,171 and \$4,376, against a true value of \$3,095, and that average earnings after the lockdown are between \$2,342 and \$3,546, against a true value of \$2,981. We estimate the proportion with large losses to be 16.5%, very

[42] Weaker assumptions about the joint distribution of V_i and S_i could also be considered. For example, instead of imposing full independence, we could require the distributions of $V_i | S_i = s$ to be monotone in s (that is, require V_i to be stochastically monotone in S_i). We maintain independence in the following because the empirical evidence is highly consistent with the assumption (see Online Appendix K.2).

close to the true value of 14.8%. In comparison, the one-dimensional model with a linearity assumption yields point estimates that are much farther away from the true values.

Figure 14: Bounds under double threshold model assumptions



Notes: The panels in this figure show estimated bounds under the two- (dark grey) and one- (light grey) dimensional selection models and under various assumptions on the marginal survey response function (MSR). Each panel presents results for one of the six administrative outcomes. For each panel, the actual population mean is presented as a vertical dashed line. Bounds are constructed using the "no" and "high" incentive samples. For all sets of assumptions, we assume that the MSR is bounded between 0 and 1 for binary variables and between the 1st and 99th percentiles of the observed distributions for continuous variables. For the two-dimensional model, we impose the assumptions listed on the y-axis: see Online Appendix K for details on the imposed assumptions and construction of estimated bounds. For the one-dimensional model, the first, third, and fourth rows respectively correspond to the first, second, and fifth rows of Figure 10 (for more details on these bounds, see the figure notes of Figure 10).

Including observed covariates and comparing to reweighted participant means. To incorporate covariates X_i, we modify (6) to

$$R_{it}(z) = \mathbb{1}[S_i \leq t]\mathbb{1}[V_i \leq \eta(z, X_i)], \tag{11}$$

where V_i is normalized to be uniformly distributed on $[0, 1]$, conditional on X_i. The two-dimensional MSR function is now $m(v, s, x) \equiv \mathbb{E}[Y_i^\star | V_i = v, S_i = s, X_i = x]$. We assume that $m(v, s, x) = m_{V,S}(v, s) + m_X(x)$ is separable and assume that the additional assumptions

that led to (10) hold conditional on X_i. Then

$$\mathbb{E}[Y_i^\star | V_i \in \mathcal{V}_j(X_i), S_i = s, X_i = x] = \underbrace{\frac{1}{\mathbb{P}[V_i \in \mathcal{V}_j(x)]} \int_{\mathcal{V}_j(x)} m_{V,S}(v,s)dv}_{\text{selection on unobservables}} + \overbrace{m_X(x)}^{\text{selection on observables}}, \tag{12}$$

where $\mathcal{V}_j(x)$ is as in Figure 13 but now depends on the x-specific values of $\eta(z,x)$. Equation (12) allows for selection on observables and two types of unobservables.

Panel A of Table 5 shows estimates that add individual characteristics to the specification in the final row of Figure 14. We obtain point estimates for five of the six outcomes and tight bounds for the sixth. The estimates are close to truth for all outcomes. For example, estimates of average monthly earnings before and after the lockdown are \$3,368 and \$3,232, which are close to the true values of \$3,095 and \$2,981, while the proportion that lost earnings is estimated to be 0.142, very close to the true value of 0.148.

For comparison, panel B of Table 5 shows reweighted participant means using the same set of individual characteristics. These are further from the ground truth than the estimates from the two-dimensional models for all outcomes when reweighting the no incentive participants, and for five out of six outcomes when using the high incentive participants.[43] Figure 15 compares the magnitude of the errors by plotting their absolute percentage differences from the ground truth. All points except for one lie above the 45 degree line, some by a considerable amount.

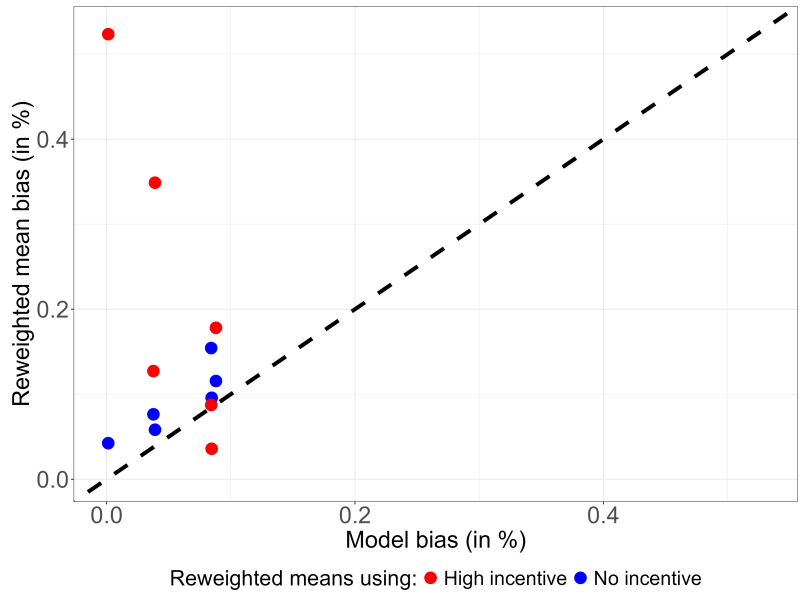
Table 5: Comparing model with characteristics and reweighted participant means

	Earnings before lockdown	Earnings after lockdown	Earnings loss larger than 20%	Employed before lockdown	Employed after lockdown	Employment loss
Ground truth	3095	2981	0.148	0.567	0.494	0.091
Panel A. Model with individual characteristics						
(Mid)point	3,368	3,232	0.142	0.588	0.536	0.091
Bounds				[0.567, 0.609]		
Panel B. Reweighted participant means using individual characteristics						
No incentive	3,453	3,441	0.139	0.610	0.542	0.087
High incentive	3,647	3,241	0.199	0.639	0.512	0.138

Notes: This table presents estimates of the population mean derived from the two-dimensional model incorporating observable individual-level characteristics (Panel A) and reweighted participant means using the same characteristics (Panel B). The observable characteristics are gender, age, years of schooling, and immigration status. The first row shows the population mean. The estimates from the two-dimensional model assume a separable MSR function in s, v, and x, which is linear in v and monotone in s; see Online Appendix K.5 for details. See footnote of Figure 7 for details on the reweighting procedure.

[43] This comparison uses the midpoint of the bounds for "employed before lockdown" (Manski, 2007).

Figure 15: Comparing model and reweighted mean bias (in %)



Notes: This figure plots the bias of estimated population means obtained from our two-dimensional model (x-axis) and from reweighting (y-axis). Each point represents the absolute percentage bias from the population mean for a specific outcome. Reweighted estimates are displayed for the high-incentive sample (red) and the no-incentive sample (blue). Corresponding estimates of the population mean are provided in Table 5.

7 Conclusion and recommendations for practice

Surveys are widely used to inform both academic research and policy decisions. We showed that nonresponse rates are often high in empirical economics, even while researchers often do not acknowledge that the validity of their conclusions may be affected by nonresponse. When researchers do acknowledge the potential role of nonresponse, they tend to either assume that responses are missing at random, or that any nonresponse bias is due to observables.

We investigated the validity of such assumptions in the Norway in Corona Time (NCT) survey by linking it to full-population administrative data. An unusual feature of the NCT survey is that it randomly assigned financial incentives for participation. We showed how to use randomized participation incentives to test for nonresponse bias. We found substantial evidence of nonresponse bias that persists after controlling for observable differences. Then, we considered various methods of correcting for nonresponse bias, finding that some more standard methods performed quite poorly, while a more elaborate new model performed better.

Our results lead to three concrete recommendations for practitioners:

Recommendation 1: Assess the potential for selection in response.
Whether nonresponse bias is a problem in any given application is an empirical question. Our results for the NCT survey provide one clear example where it is a serious problem. While this does not indicate how widespread the problem is, it does suggest that researchers cannot simply ignore the potential for nonresponse bias.

Our first recommendation is to more explicitly discuss the potential for nonresponse bias and its implications. This requires thinking about the potential determinants of response, both observed and unobserved. *Why* did some individuals respond while others did not? If these determinants are statistically related with the responses being measured, then there is scope for nonresponse bias. For example, in a labor market survey like the NCT, it is plausible that opportunity cost of time is both a determinant of response and a correlate of the types of labor market outcomes the survey is intended to measure.

Recommendation 2: Randomize incentives for participating.
Our second recommendation is that researchers incorporate randomized financial incentives into their surveys. Randomized incentives can be used to test for nonresponse bias in survey data (*without* linked administrative data) by simply comparing participant means across incentive arms, assuming that the incentives themselves do not directly affect responses (see Section 4.3). The nonresponse bias can be unpacked by comparing marginal participants with inframarginal participants through linear instrumental variable estimators (see Section 4.4). While we found strong evidence of nonresponse bias in the NCT survey, it is of course also possible to use randomized incentives to provide convincing evidence of the absence of nonresponse bias.

Incentivized surveys are more costly than unincentivized surveys. However, as our literature review in Section 2 showed, many surveys are already incentivized. *Randomizing* different incentives can be done in a cost-neutral way compared to a single deterministic incentive. While there may be implementation challenges with using randomized incentives, these challenges seem worth addressing compared to the benefits of being able to detect and measure nonresponse bias.

Recommendation 3: Consider correcting for selection (cautiously).
If evidence of nonresponse bias is found, then it may be worth trying to correct for it. While we found selection on unobservables to be more important than selection on observables in the NCT survey, this may not be the case in other settings. Correcting for observables using standard reweighting methods can certainly help, although we caution that some of our findings in Section 4.2 showed that it could actually exacerbate selection on unobservables in some cases.

Correcting for selection on unobservables is more difficult. Bounds under minimal assumptions can be useful if the response rate is already high, but response rates are low in many surveys, including ours, leading to bounds that are quite wide. Bounds under stronger assumptions can be more informative, but in our setting we found evidence that they could also be off by a considerable margin. Using a standard selection model turned out to not work any better, perhaps because of the difficulty in extrapolation.

More complex selection models can potentially help, as we demonstrated with the two-dimensional model in Section 6. That model is tailored to our particular setting and should be seen as a proof of concept rather than a widely-applicable method. Nevertheless, we think

that it provides some hope for the development of more complex methods of constructively correcting for nonresponse bias due to unobservables. Key to this development would be designing surveys that include multiple dimensions (such as financial incentives, outreach effort, and reminders) through which the propensity to participate can be varied exogenously.

References

Behaghel, L., B. Crépon, M. Gurgand, and T. Le Barbanchon (2015). Please Call Again: Correcting Nonresponse Bias in Treatment Effect Models. *Review of Economics and Statistics 97*(5), 1070–1080.

Bethlehem, J., F. Cobben, and B. Schouten (2011). *Handbook of Nonresponse in Household Surveys*, Volume 568. John Wiley & Sons.

Bhattacharya, J. and A. Isen (2009). On Inferring Demand for Health Care in the Presence of Anchoring and Selection Biases. *Forum for Health Economics & Policy 12*, 239–244.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007). Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica 75*(2), 323–363.

Bollinger, C. R. and B. T. Hirsch (2013). Is earnings nonresponse ignorable? *Review of Economics and Statistics 95*(2), 407–416.

Bollinger, C. R., B. T. Hirsch, C. M. Hokayem, and J. P. Ziliak (2019). Trouble in the Tails? What We Know about Earnings Nonresponse 30 Years after Lillard, Smith, and Welch. *Journal of Political Economy 127*(5), 2143–2185.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017). Beyond LATE with a Discrete Instrument. *Journal of Political Economy 125*(4), 985–1039.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011). Estimating Marginal Returns to Education. *American Economic Review 101*(6), 2754–81.

Coffman, L. C., J. J. Conlon, C. R. Featherstone, and J. B. Kessler (2019). Liquidity Affects Job Choice: Evidence from Teach for America. *The Quarterly Journal of Economics 134*(4), 2203–2236.

Currie, J., H. Kleven, and E. Zwiers (2020). Technology and Big Data Are Changing Economics: Mining Text to Track Methods. *American Economic Association Papers & Proceedings 110*, 42–48.

Czajka, J. L. and A. Beyler (2016). Declining Response Rates in Federal Surveys: Trends and Implications. *Mathematica policy research 1*(4), 1–86.

de Leeuw, E. and W. de Heer (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. Groves, D. Dillman, E. J.L., and R. Little (Eds.), *Survey Nonresponse*. New York: Wiley.

DellaVigna, S., J. A. List, U. Malmendier, and G. Rao (2017). Voting to Tell Others. *The Review of Economic Studies 84*(1), 143–181.

DiNardo, J., J. Matsudaira, J. McCrary, and L. Sanbonmatsu (2021). A Practical Proactive Proposal for Dealing with Attrition: Alternative Approaches and an Empirical Example. *Journal of Labor Economics 39*(S2), S507–S541.

Fiva, J. H., A. H. Halse, and G. J. Natvik (2020). Local Government Dataset. www.jon.fiva.no/data.htm. Accessed: 2021-05-23.

Gonzalez, L. (2005). Nonparametric Bounds on the Returns to Language Skills. *Journal of Applied Econometrics 20*(6), 771–795.

Gørgens, T. and C. Ryan (2008). A bounds analysis of school completion rates in australia. *Journal of Applied Econometrics 23*(3), 287–304.

Gronau, R. (1974). Wage comparisons–a selectivity bias. *Journal of political Economy 82*(6), 1119–1143.

Groves, R. M., D. A. Dillman, J. L. Eltinge, and R. J. Little (2002). *Survey Nonresponse*, Volume 23. Wiley New York.

Groves, R. M., F. J. Fowler Jr., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009). *Survey Methodology*. Wiley New York.

Groves, R. M. and E. Peytcheva (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly 72*(2), 167–189.

Gurobi Optimization, L. (2021). Gurobi Optimizer Reference Manual.

Hausman, J. A. and D. A. Wise (1979). Attrition bias in experimental and panel data: the gary income maintenance experiment. *Econometrica: Journal of the Econometric Society*, 455–473.

Heckman, J. (1974). Shadow Prices, Market Wages, and Labor Supply. *Econometrica: journal of the econometric society 42*(4), 679–694.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica: Journal of the econometric society 47*(1), 153–161.

Heckman, J. J. and P. A. LaFontaine (2006). Bias-corrected estimates of ged returns. *Journal of Labor Economics 24*(3), 661–700.

Heckman, J. J. and R. Pinto (2018). Unordered Monotonicity. *Econometrica 86*(1), 1–35.

Heckman, J. J. and E. Vytlacil (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2001). Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effect. In *Econometric Evaluation of Labour Market Policies*, pp. 1–15. Springer.

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

Heffetz, O. and M. Rabin (2013). Conclusions regarding cross-group differences in happiness depend on difficulty of reaching respondents. *American Economic Review 103*(7), 3001–3021.

Hokayem, C., C. Bollinger, and J. P. Ziliak (2015). The role of cps nonresponse in the measurement of poverty. *Journal of the American Statistical Association 110*(511), 935–945.

Horowitz, J. L. and C. F. Manski (1998). Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations. *Journal of Econometrics 84*(1), 37–58.

Hotchkiss, M. and J. Phelan (2017). Uses of Census Bureau Data in Federal Funds Distribution. Technical report, U.S. Census Bureau.

Humphries, J. E., A. Ouss, K. Stavreva, M. Stevenson, and W. van Dijk (2024). Conviction, Incarceration, and Recidivism: Understanding the Revolving Door. *Working paper*.

Imbens, G. W. and J. D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica: Journal of the Econometric Society 62*(2), 467–475.

J-PAL (2020). Data Analysis. https://www.povertyactionlab.org/resource/data-analysis#section-miscellaneous.

J-PAL (2021). Increasing Response Rates of Mail Surveys and Mailings. https://www.povertyactionlab.org/resource/increasing-response-rates-mail-surveys-and-mailings.

Juster, F. T. and R. Suzman (1995). An overview of the health and retirement study. *Journal of Human Resources*, S7–S56.

Kirkeboen, L. J., E. Leuven, and M. Mogstad (2016). Field of Study, Earnings, and Self-Selection. *Quarterly Journal of Economics 131*(3), 1057–1111.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review 76*(4), 604–620.

Lechner, M. (1999). Nonparametric Bounds on Employment and Income Effects of Continuous Vocational Training in East Germany. *The Econometrics Journal 2*(1), 1–28.

Lee, S. and B. Salanié (2018). Identifying Effects of Multivalued Treatments. *Econometrica 86*(6), 1939–1963.

Little, R. J. and D. B. Rubin (2019). *Statistical Analysis with Missing Data*, Volume 793. John Wiley & Sons.

Manski, C. F. (1989). Anatomy of the Selection Problem. *Journal of Human resources 24*(3), 343–360.

Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. *The American Economic Review 80*(2), 319–323.

Manski, C. F. (1994). The Selection Problem. In C. Sims (Ed.), *Advances in Econometrics*, Volume II of *Sixth World Congress*, pp. 143–170. Cambridge University Press.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer Science & Business Media.

Manski, C. F. (2007). Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics 139*(1), 105–115.

Manski, C. F. (2016). Credible Interval Estimates for Official Statistics with Survey Nonresponse. *Journal of Econometrics 191*(2), 293–301.

Manski, C. F. and F. Molinari (2021). Estimating the covid-19 infection rate: Anatomy of an inference problem. *Journal of Econometrics 220*(1), 181–192.

Manski, C. F. and J. V. Pepper (2000). Monotone Instrumental Variables: With an Application to the Returns to Schooling. *Econometrica 68*(4), 997–1010.

Martin, E. and F. Winters (2001). Money and motive: effects of incentives on panel attrition in the survey of income and program participation. *Journal of Official Statistics 17*(2), 267.

McGovern, M. E., D. Canning, and T. Barnighausen (2018). Accounting for Non-Response Bias Using Participation Incentives and Survey Design: An Application Using Gift Vouchers. *Economics Letters 171*, 239–244.

Mercer, A., A. Caporaso, D. Cantor, and R. Townsend (2015). How Much Gets You How Much? Monetary Incentives and Response Rates in Household Surveys. *Public Opinion Quarterly 79*(1), 105–129.

Meterko, M., J. D. Restuccia, K. Stolzmann, D. Mohr, C. Brennan, J. Glasgow, and P. Kaboli (2015). Response rates, nonresponse bias, and data quality: Results from a national survey of senior healthcare leaders. *Public Opinion Quarterly 79*(1), 130–144.

Meyer, B. D., W. K. C. Mok, and J. X. Sullivan (2015). Household Surveys in Crisis. *Journal of Economic Perspectives 29*(4), 199–226.

Ministry of Finance (2020). Prop. 1 S (2019–2020). Proposition to the Storting (draft resolution). https://www.regjeringen.no/contentassets/e5b05593a20a49a8865ef3538c7e2f1e/no/pdfs/prp2019202000001guldddpdfs.pdf.

Moffitt, R. (2004). The Three-City Study Incentive Experiment: Results from the First Two Waves. Technical report. Unpublished manuscript, https://www.researchgate.net/profile/Robert-Moffitt-2/publication/239932435_The_Three-City_Study_Incentive_Experiment_Results_from_the_First_Two_Waves/links/54ba77590cf253b50e2d019e/The-Three-City-Study-Incentive-Experiment-Results-from-the-First-Two-Waves.pdf.

Mogstad, M., A. Santos, and A. Torgovitsky (2018). Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters. *Econometrica 86*(5), 1589–1619.

Mogstad, M. and A. Torgovitsky (2018). Identification and Extrapolation of Causal Effects with Instrumental Variables. *Annual Review of Economics 10*, 577–613.

Mogstad, M., A. Torgovitsky, and C. Walters (2020). Policy Evaluation with Multiple Instrumental Variables. Technical Report w27546, National Bureau of Economic Research, Cambridge, MA.

Mountjoy, J. (2021). Community Colleges and Upward Mobility. Technical Report w29254, National Bureau of Economic Research, Cambridge, MA.

National Bureau of Economic Research (2020). Meta-data for the NBER working paper series. https://www2.nber.org/wp_metadata/.

National Research Council (2013a). *Nonresponse in Social Science Surveys: A Research Agenda*. Washington, DC: The National Academies Press.

National Research Council (2013b). The Growing Problem of Nonresponse. In *Nonresponse in Social Science Surveys: A Research Agenda*, pp. 7–39. Washington, DC: The National Academies Press.

Office of Management and Budget (2006). Questions and Answers when Designing Surveys for Information Collections. https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/pmc_survey_guidance_2006.pdf.

Page, M. J., J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al. (2021). The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *The British Medical Journal 372*.

Rosenbaum, P. R. and D. B. Rubin (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika 70*(1), 41–55.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc.

Singer, E. (2006). Introduction: Nonresponse Bias in Household Surveys. *International Journal of Public Opinion Quarterly 70*(5), 637–645.

Singer, E. and C. Ye (2013). The Use and Effects of Incentives in Surveys. *The Annals of the American Academy of Political and Social Science 645*(1).

Snowberg, E. and L. Yariv (2021). Testing the Waters: Behavior across Participant Pools. *American Economic Review 111*(2), 687–719.

Sturgis, P. and R. Luff (2021). The Demise of the Survey? A Research Note on Trends in the Use of Survey Data in the Social Sciences, 1939 to 2015. *International Journal of Social Research Methodology 24*(6).

U.S. Census Bureau (2021). SIPP Introduction & History. Technical report. https://www.census.gov/programs-surveys/sipp/about/sipp-introduction-history.html.

Vytlacil, E. (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica 70*(1), 331–341.

A Appendix figures and tables

Table A.1: Variable definitions and sources

Variable	Definition	Source
Variables from administrative sources		
Female	Indicator for female	CPR
Immigrant	Indicator for immigrant	CPR
Age	Individual's age	CPR
Years of school	Individual's years of school	NED
Live with children	Indicator for living with at least one child < 18 y.o.	CPR
Applied for UI	Indicator for application to unemployment benefits in March or April, 2020 (as_yte=DP)	ARENA
Earnings before	Average monthly earnings (USD) in Jan/Feb, 2020.	EE
Earnings after	Earnings (USD) in April, 2020 (after lockdown)	EE
Earnings loss	Indicator for 20% earnings loss after lockdown relative to before	EE
Employed before	Indicator for average earnings greater than the 'basic amount' (used to determine substantial gainful activity in Norway) and not registered as either fully or partially unemployed in Jan/Feb, 2020	EE and ARENA
Employed after	Indicator for average earnings greater than the 'basic amount' and not registered as either fully or partially unemployed in April, 2020	EE and ARENA
Employment loss	Indicator equal to 1 if employed before and not employed after, 0 otherwise	EE and ARENA
Variables from survey data		
Participation	Indicator for an individual's completion of the full survey	NCT
Became furloughed or unemployed	Indicator for reporting to be furloughed or unemployed after lockdown and not before *Do you consider yourself today primarily as ... 1. working / 2. temporary full-time laid off / 3. unemployed / 4. old age pensioner / 5. Work disabled / 6. student / 7. homeworker / 8. military service / 9. other.* (Q2a_7=2 or 3) *In the period before the lockdown, did you consider yourself primarily as ... 1. working / 2. temporary full-time laid off / 3. unemployed / 4. old age pensioner / 5. Work disabled / 6. student / 7. homeworker / 8. military service / 9. other.* (Q2a_1≠2 or 3)	NCT
Applied for UI	Indicator for reporting to have applied for unemployment benefits after lockdown *In the period after the lockdown, have you applied for any of the following governmental transfers? 1. Unemployment benefits / 2. Health-related benefits / 3. Other welfare benefits / 4. Receive no benefits / 5. Do not know.* (Q2b_2=1)	NCT
No longer full-time	Indicator for weekly work hours ≤ 37 hours after lockdown and > 37 hours before *How many hours per week do you usually work now? __ hours.* (Spm2a_10mer) *In the period before the lockdown, how many hours per week did you usually work? Include overtime and work from home. __ hours.* (Q2a_4)	NCT
Work hours reduction	Indicator for reporting to have reduced work hours after lockdown *Do you work more, less or as much as you did before the authorities implemented measures against the coronavirus? 1. I work more / 2. I work less / 3. I work just as much / 4. Do not know.* (Q2a_10=2)	NCT

Notes: This table presents definitions and data source of all variables used throughout the paper. Data sources are abbreviated as follows: CPR=Central Population Register, NED=National Education Database, EE=Employer-employee Registry, ARENA=ARENA Registry, NCT=Statistics Norway Survey "Norway in Corona Time". Individual characteristics are defined per 4/30/2020. We use the currency rate NOK/USD=9.

Table A.2: Balance test

	Female	Age	Years of school	Immigrant	Earnings before	Earnings after	Employed before	Employed after
Pr=0.1	0.500	47.61	12.69	0.153	3149.3	2865.8	0.573	0.489
	(0.0189)	(0.687)	(0.161)	(0.0142)	(152.3)	(146.9)	(0.0187)	(0.0189)
Pr=0.07	0.486	47.73	12.71	0.176	3460.4	3090.1	0.597	0.490
	(0.0189)	(0.687)	(0.161)	(0.0142)	(152.3)	(146.9)	(0.0187)	(0.0189)
Pr=0.05	0.506	47.85	12.53	0.172	3020.8	3035.7	0.564	0.498
	(0.0134)	(0.486)	(0.114)	(0.0100)	(107.7)	(104.0)	(0.0133)	(0.0134)
Pr=0.01	0.488	47.84	12.55	0.162	3120.2	2971.4	0.575	0.502
	(0.00946)	(0.344)	(0.0804)	(0.00710)	(76.19)	(73.52)	(0.00937)	(0.00946)
Pr=0	0.493	47.69	12.43	0.177	3026.2	2968.7	0.554	0.489
	(0.00819)	(0.298)	(0.0697)	(0.00615)	(65.97)	(63.66)	(0.00811)	(0.00819)
Observations	9323	9322	9323	9323	9323	9323	9323	9323
F–statistic	0.38	0.05	1.08	1.03	1.89	0.37	1.52	0.30
p–value	.82	1	.36	.39	.11	.83	.19	.88
Sample mean	0.494	47.76	12.52	0.170	3095.4	2980.9	0.567	0.494

Notes: This table reports estimates and standard errors (in parentheses) from regressions of background characteristics and outcomes from administrative data on incentive groups in the invited population. F–statistics and p–values are presented for the test of equality of means across all incentive groups. See Appendix Table A.1 for details on variable definitions.

Table A.4: Relative nonresponse bias in moments and changes

	Earnings			Employment		
	No	Low	High	No	Low	High
Levels						
Before lockdown	0.18***	0.23***	0.3***	0.11***	0.16***	0.19***
	(0.03)	(0.03)	(0.06)	(0.02)	(0.02)	(0.03)
After lockdown	0.22***	0.25***	0.23***	0.16***	0.17***	0.17***
	(0.03)	(0.03)	(0.05)	(0.02)	(0.02)	(0.04)
Changes						
Loss (as in draft)	-0.13***	0.03	0.1	-0.13**	0.05	0.23*
	(0.05)	(0.05)	(0.09)	(0.06)	(0.06)	(0.13)
Difference (after - before)	-0.84*	-0.07	2.08	-0.2**	0.1	0.36**
	(0.45)	(0.64)	(1.38)	(0.09)	(0.08)	(0.16)

Notes: This table reports estimates and standard errors (in parentheses) of relative nonresponse bias, which is nonresponse bias divided by the true population value. We consider relative nonresponse bias to better compare the severity of nonresponse bias in outcomes measured in levels and changes. The first set of columns report earnings-related outcomes, and the second set of columns report employment-related outcomes. For each set, the three columns respectively report estimates using participants in the no, low, and high incentive arms. The first row reports relative nonresponse bias for earnings and employment before lockdown, the second row for these outcomes after lockdown, the third row for earnings and employment losses, and the fourth row is for differences in earnings and employment after and before lockdown. The average absolute relative nonresponse bias across all levels estimates is 0.20, and this value is 0.36 across all changes estimates.

Table A.3: Nonresponse bias and selection in administrative outcomes

	Monthly earnings before lockdown	Monthly earnings after lockdown	Earnings loss above 20%	Employed before lockdown	Employed after lockdown	Employment loss	Joint test
Panel A: Population mean							
	3,095	2,981	0.148	0.567	0.494	0.091	
Panel B: Unweighted estimates							
No	3,666.5	3,648.3	0.128	0.629	0.571	0.079	
	(104.8)	(102.7)	(0.009)	(0.012)	(0.012)	(0.007)	0.000
Low	3,820.1	3,714.1	0.151	0.660	0.581	0.095	
	(96.9)	(94.9)	(0.008)	(0.011)	(0.011)	(0.006)	0.000
High	4,029.6	3,676.7	0.162	0.675	0.577	0.112	
	(160.9)	(157.6)	(0.013)	(0.018)	(0.018)	(0.011)	0.000
Panel C: Reweighted estimates – municipality level							
No	3,612.8	3,590.0	0.127	0.626	0.569	0.078	
	(92.4)	(98.3)	(0.008)	(0.012)	(0.012)	(0.006)	0.000
Low	3,789.9	3,681.9	0.151	0.659	0.580	0.095	
	(98.4)	(97.0)	(0.008)	(0.011)	(0.010)	(0.006)	0.000
High	3,987.3	3,648.4	0.160	0.674	0.577	0.111	
	(185.3)	(143.4)	(0.013)	(0.018)	(0.019)	(0.011)	0.000
Panel D: Reweighted estimates – individual level							
No	3,453.2	3,441.0	0.139	0.610	0.542	0.087	
	(120.0)	(132.7)	(0.010)	(0.015)	(0.014)	(0.009)	0.000
Low	3,573.6	3,427.2	0.168	0.637	0.542	0.112	
	(92.5)	(95.7)	(0.010)	(0.012)	(0.012)	(0.009)	0.000
High	3,646.9	3,241.5	0.199	0.639	0.512	0.138	
	(188.9)	(156.7)	(0.022)	(0.025)	(0.025)	(0.019)	0.000

Notes: This table shows the estimated population mean and participant mean by incentive level and estimation method for administrative outcomes. Panel A presents population means. Panels B, C and D present, respectively, unweighted, reweighted by municipality characteristics, and reweighted by individual characteristics estimated participant means and standard errors (in parentheses). The final column to the right shows p–values for a joint test of equality between the participant and population means for all six outcomes. See the figure notes of Figure 7 for more details on reweighting specifications.

Table A.5: Relative nonresponse bias in moments and changes (reweighted)

	Earnings			Employment		
	No	Low	High	No	Low	High
Levels						
Before lockdown	0.12***	0.15***	0.18***	0.08***	0.12***	0.13***
	(0.04)	(0.03)	(0.05)	(0.02)	(0.02)	(0.04)
After lockdown	0.15***	0.15***	0.09*	0.1***	0.1***	0.04
	(0.04)	(0.03)	(0.05)	(0.03)	(0.03)	(0.04)
Changes						
Loss (as in draft)	-0.06	0.14**	0.35**	-0.04	0.23**	0.52**
	(0.07)	(0.07)	(0.15)	(0.09)	(0.09)	(0.22)
Difference (after - before)	-0.89*	0.28	2.54*	-0.06	0.33***	0.75***
	(0.53)	(0.56)	(1.5)	(0.12)	(0.12)	(0.28)

Notes: This table mirrors Table A.5, but now reports estimates of relative nonresponse bias after reweighting following the approach in Section 4 and using individual characteristics. See the notes of Table A.5 for additional details. The average absolute relative nonresponse bias across all levels estimates is 0.12, and this value is 0.52 across all changes estimates.

Table A.6: Nonresponse bias and selection in survey responses

	No longer full-time	Reduction in work hours	Became furloughed or unemployed	Applied for UI
Panel A: Unweighted estimates				
No	0.131	0.210	0.034	0.075
	(0.008)	(0.010)	(0.005)	(0.007)
Low	0.140	0.225	0.047	0.087
	(0.008)	(0.009)	(0.005)	(0.006)
High	0.176	0.275	0.068	0.104
	(0.013)	(0.016)	(0.008)	(0.010)
p–value:				
High=Low=No	0.010	< 0.01	< 0.01	0.060
Panel B: Reweighted estimates – municipality level				
No	0.129	0.207	0.033	0.073
	(0.007)	(0.009)	(0.004)	(0.006)
Low	0.139	0.223	0.047	0.087
	(0.007)	(0.009)	(0.004)	(0.006)
High	0.175	0.273	0.067	0.102
	(0.014)	(0.016)	(0.009)	(0.011)
p–value:				
High=Low=No	< 0.01	< 0.01	< 0.01	0.010
Panel C: Reweighted estimates – individual level				
No	0.127	0.196	0.035	0.087
	(0.010)	(0.011)	(0.005)	(0.009)
Low	0.142	0.223	0.054	0.103
	(0.009)	(0.010)	(0.006)	(0.009)
High	0.182	0.288	0.109	0.145
	(0.020)	(0.022)	(0.021)	(0.022)
p–value:				
High=Low=No	0.020	< 0.01	< 0.01	0.010

Notes: This table shows participant responses means by incentive level and estimation method for survey-elicited outcomes. Panels A, B and C present, respectively, unweighted, reweighted by municipality characteristics, and reweighted by individual characteristics estimated participant means and standard errors (in parentheses). p-values for testing the equality of mean responses across incentive arms are shown in the lower rows of each panel. See the figure notes of Figure 7 for more details on reweighting specifications.

Table A.7: Regressions of survey participation and outcomes on background characteristics

	(1) Completed survey	(2) Earnings before	(3) Earnings after	(4) Earnings loss	(5) Employed before	(6) Employed after	(7) Employed loss
Panel A. Municipality level characteristics							
Median household income	0.313***	4304.1***	3968.2***	0.0410	0.162	0.190*	-0.0552
(in 100,000)	(0.103)	(826.9)	(797.9)	(0.0732)	(0.102)	(0.103)	(0.0593)
Inhabitants	0.00165	74.81**	63.23**	0.0000371	-0.00637	-0.00396	-0.00220
(in 100,000)	(0.00413)	(33.20)	(32.04)	(0.00294)	(0.00410)	(0.00414)	(0.00238)
Share women	2.884***	10137.5	13662.2*	0.800	0.649	-0.997	1.317**
	(0.975)	(7834.4)	(7559.6)	(0.694)	(0.967)	(0.976)	(0.562)
Unemployment rate	0.315	1065.8	-8600.9	1.286	-2.866*	-2.904*	-0.327
(benefit application)	(1.576)	(12670.1)	(12225.7)	(1.122)	(1.563)	(1.579)	(0.908)
Share aged >65 y.o.	-0.309	-4412.5*	-4554.2**	-0.0964	-1.080***	-0.731**	-0.389**
	(0.285)	(2291.2)	(2210.8)	(0.203)	(0.283)	(0.285)	(0.164)
Constant	-1.149**	-4591.6	-6001.3	-0.285	0.353	1.016**	-0.449
	(0.486)	(3908.9)	(3771.8)	(0.346)	(0.482)	(0.487)	(0.280)
F–test	9.357	23.941	23.042	1.635	11.376	8.503	3.878
p–value	0.000	0.000	0.000	0.147	0.000	0.000	0.002
Panel B. Individual level characteristics							
Female	0.0729***	-1104.5***	-1045.8***	-0.0277***	-0.0620***	-0.0436***	-0.0172***
	(0.00992)	(79.10)	(75.88)	(0.00723)	(0.00948)	(0.00967)	(0.00588)
Age	-0.000999***	-32.05***	-31.38***	-0.00331***	-0.00843***	-0.00709***	-0.00197***
	(0.000276)	(2.203)	(2.113)	(0.000201)	(0.000264)	(0.000269)	(0.000164)
Years of school	0.0267***	260.5***	269.0***	-0.00416***	0.0286***	0.0312***	-0.00305***
	(0.00126)	(10.02)	(9.609)	(0.000916)	(0.00120)	(0.00122)	(0.000745)
Immigrant	-0.131***	434.3***	388.5***	0.0118	0.0412***	-0.000882	0.0383***
	(0.0144)	(114.6)	(109.9)	(0.0105)	(0.0137)	(0.0140)	(0.00852)
Constant	0.174***	1835.5***	1561.2***	0.369***	0.635***	0.464***	0.225***
	(0.0226)	(180.0)	(172.6)	(0.0165)	(0.0216)	(0.0220)	(0.0134)
F–test	213.888	272.458	302.430	85.174	412.967	354.001	61.041
p–value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Mean outcome	0.474	3095.8	2981.3	0.148	0.567	0.494	0.0908
Observations	9322	9322	9322	9322	9322	9322	9322

Notes: This table presents the estimated coefficients and standard errors from a regression of each outcome (referenced in the top row) on a set of background characteristics. Panel A presents estimates for municipality level characteristics (Fiva et al., 2020). Panel B presents estimates for individual-level characteristics obtained from administrative data linkage. F-statistics and p–values for joint tests of significance shown in bottom rows. Standard errors in parentheses and stars denote individual statistical significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.01$.

Table A.8: Instrumental variable estimates - background variables

	Inframarginal participant (no)		Marginal participant (no-high)		p–value
	Est.	(SE)	Est.	(SE)	(no) = (no-high)
Female	0.542	(0.012)	0.388	(0.184)	0.42
Immigrant	0.103	(0.007)	0.004	(0.111)	0.39
Age	48.0	(0.4)	42.5	(6.2)	0.39
Years of school	13.7	(0.1)	14.0	(1.3)	0.85

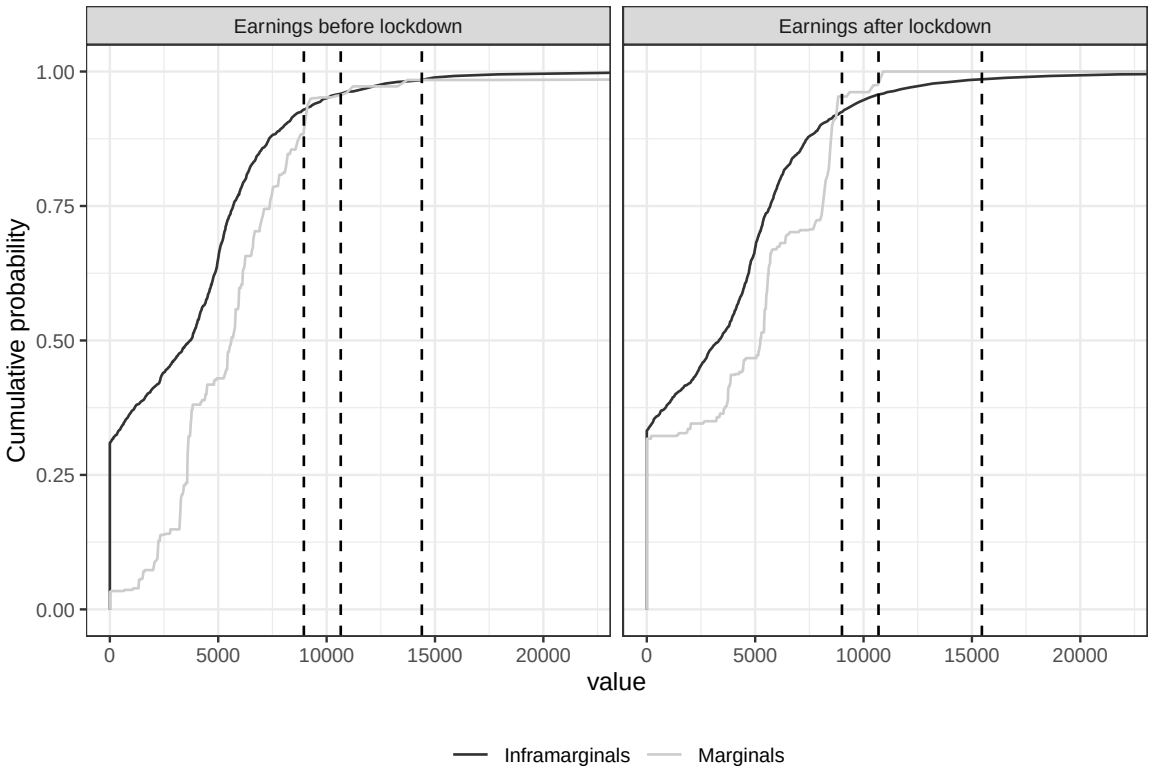
Notes: This table presents the estimated average background characteristics of individuals inframarginal and marginal to incentives. These values are estimated using an instrumental variables regression of $Y_i R_i$ as the outcome variable, survey outcome R_i as the endogenous variable and the set of indicators for incentive groups Z_i as the instrument.

Table A.9: Instrumental variable estimates using all three incentive levels

	Inframarginal participant (no)		Marginal participant (no-low)		Marginal participant (low-high)		p–value (no) = (no-low) = (low-high)
	Est.	(SE)	Est.	(SE)	Est.	(SE)	
Panel A: Administrative data							
Earnings before lockdown	3,666	(107)	7,562	(4,229)	6,460	(2,457)	0.22
Employed before lockdown	0.629	(0.012)	1.403	(0.599)	0.849	(0.258)	0.21
Earnings after lockdown	3,648	(105)	5,317	(3,636)	3,244	(2,203)	0.90
Employed after lockdown	0.571	(0.012)	0.810	(0.432)	0.531	(0.257)	0.86
Earnings loss larger than 20%	0.128	(0.009)	0.722	(0.452)	0.282	(0.189)	0.21
No longer employed	0.079	(0.007)	0.485	(0.336)	0.306	(0.170)	0.10
Panel B: NCT survey data							
Became furloughed or unemployed	0.034	(0.005)	0.356	(0.246)	0.307	(0.146)	0.02
Applied for UI	0.075	(0.007)	0.393	(0.293)	0.286	(0.159)	0.12
No longer full-time work	0.131	(0.009)	0.347	(0.304)	0.594	(0.250)	0.08
Reduction in work hours	0.210	(0.010)	0.567	(0.385)	0.862	(0.323)	0.04
Panel C: Background characteristics							
Female	0.542	(0.012)	0.418	(0.418)	0.374	(0.266)	0.72
Immigrant	0.103	(0.007)	0.040	(0.250)	-0.013	(0.161)	0.69
Age	48.0	(0.4)	33.6	(16.1)	46.5	(8.6)	0.62
Years of school	13.7	(0.1)	11.8	(3.1)	15.0	(1.9)	0.73

Notes: This table presents the estimated average labor market outcomes and background characteristics of individuals inframarginal and marginal to incentives. These values are estimated using an instrumental variables regression of $Y_i R_i$ as the outcome variable, survey outcome R_i as the endogenous variable and the set of indicators for incentive groups Z_i as the instrument.

Figure A.1: Empirical cumulative distributions by participant type



Notes: This figure presents the estimated cumulative probability for earnings before lockdown (left panel) and after lockdown (right panel) by participant group: inframarginal participants depicted in black and marginal participants in light gray. We estimate the empirical cumulative distribution for marginal participants using sample analogs of $\mathbb{P}[Y_i \leq y | R_i(1) > R_i(0)] = \frac{\mathbb{P}[R_i=1]}{\mathbb{P}[R_i(1)>R_i(0)]}[Y_i \leq y | R_i = 1, Z_i = 1] - \frac{\mathbb{P}[R_i(1)>R_i(0)]}{\mathbb{P}[R_i=1]}\mathbb{P}[Y_i \leq y | R_i = 1, Z_i = 0]$. for different values of y. The vertical dashed lines depict (from left to right) the values of the 95th, 97th and 99th percentiles, respectively.