Statistisk sentralbyrå
Statistics Norway

Selection in Surveys

TALL SOM FORTELLER
DISCUSSION PAPERS
971

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

Discussion Papers No. 971, December 2021
Statistics Norway, Research Department

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

Selection in Surveys

Abstract:
We evaluate how nonresponse affects conclusions drawn from survey data and consider how researchers can reliably test and correct for nonresponse bias. To do so, we examine a survey on labor market conditions during the COVID-19 pandemic that used randomly assigned financial incentives to encourage participation. We link the survey data to administrative data sources, allowing us to observe a ground truth for participants and nonparticipants. We find evidence of large nonresponse bias, even after correcting for observable differences between participants and nonparticipants. We apply a range of existing methods that account for nonresponse bias due to unobserved differences, including worst-case bounds, bounds that incorporate monotonicity assumptions, and approaches based on parametric and nonparametric selection models. These methods produce bounds (or point estimates) that are either too wide to be useful or far from the ground truth. We show how these shortcomings can be addressed by modeling how nonparticipation can be both active (declining to participate) and passive (not seeing the survey invitation). The model makes use of variation from the randomly assigned financial incentives, as well as the timing of reminder emails. Applying the model to our data produces bounds (or point estimates) that are narrower and closer to the ground truth than the other methods.

Keywords: survey, nonresponse, nonresponse bias

JEL classification: C01, C81, C83

Acknowledgements: The authors gratefully acknowledge nancial 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 Lagerstrm and his team at Statistics Norway for implementing the survey. We would like to thank Joe Altonji, Alex Bick, Raj Chetty, Nathan Hendren, John Eric Humphries, Larry Katz, Costas Meghir, 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.

Address: Ingrid Huitfeldt, Statistics Norway, Research Department. E-mail: ish@ssb.no

Discussion Papers comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no/en/forskning/discussion-papers
http://ideas.repec.org/s/ssb/dispap.html

ISSN 1892-753X (electronic)

Sammendrag

I denne artikkelen undersøker vi hvordan frafall påvirker konklusjoner fra spørreundersøkelser, og viser hvordan forskere kan teste og korrigere for frafallsskjevhet. For dette formålet bruker vi en spørreundersøkelse om arbeidsmarkedsforhold under COVID-19-pandemien der randomiserte finansielle insentiver ble benyttet for å motivere svardeltagelse. Ved å linke dataene fra spørreundersøkelsen til administrative registre kan vi observere den sanne arbeidsmarkedstilknytningen for både deltagere og ikke-deltagere. Vi finner at frafallsskjevheten er stor, selv etter å ha korrigert for observerbare forskjeller mellom deltagere og ikke-deltagere. Vi anvender en rekke eksisterende metoder fra økonometrisk litteratur på missing data og programevaluering som tar høyde for frafallsskjevhet som skyldes uobserverbare forskjeller. Metodene inkluderer worst-case bounds, bounds som innlemmer monotonisitetsantakelser, og tilnærminger basert på parametriske og ikke-parametriske seleksjonsmodeller. Metodene produserer bounds (eller punktestimater) som er enten for vide for å være nyttige, eller langt unna sannheten. Vi viser at disse svakhetene forbedres ved å modellere hvordan frafall kan være både aktivt (personen velger å ikke svare) og passivt (personen ser ikke invitasjonen til å svare på undersøkelsen). Modellen bruker variasjon i randomiserte finansielle insentiver, i tillegg til tidspunkt for påminnelseseposter. Ved å anvende modellen på våre data finner vi bounds (eller punktestimater) som er tettere og nærmere sannheten enn de andre metodene.

Oppsummert viser funnene våre at svarprosenter kan være en misvisende indikasjon på frafallsskjevheter i spørreundersøkelsesdata, og bør derfor ikke være viktigste målparameter når man utformer spørreundersøkelser. Isteden er det viktig at det er mulig å teste og korrigere for frafallsskjevhet. Vi viser at ved å randomisere insentiver for deltagelse kan man korrigere for uobserverbare forskjeller mellom de som deltar og ikke deltar i undersøkelsen.

1 Introduction

Surveys are widely used to inform both academic research and policy decisions. Collecting survey data requires participation on the part of those being surveyed. If participation is correlated with responses to survey questions, then estimates from a survey will be contaminated with nonresponse bias, making them potentially misleading estimates of the targeted survey population. Researchers designing surveys and analyzing survey data therefore need to be concerned with mitigating and correcting for nonresponse bias.

In Section 2, we provide a comprehensive review of how economists cope with the possibility of nonresponse bias in modern empirical research. We find that nonresponse rates are often high, yet discussion of potential nonresponse bias is uncommon. This practice stands in stark contrast to the care that is usually taken in dealing with potential selection bias when answering causal inference questions, and suggests an under-appreciation of the problems that can be caused by nonresponse. When the possibility of nonresponse bias is discussed, researchers usually either assume that responses are missing at random—so that there is no nonresponse bias—or that responses are missing at random after controlling for observable factors, so that nonresponse bias can be corrected for by reweighting.

These conventional practices raise a number of questions. Does nonresponse bias materially affect the conclusions drawn from survey data? If so, is it caused by observed differences between participants and nonparticipants, or by unobserved differences? Is it possible to reliably detect and correct for nonresponse bias? If not, can surveys be designed differently to facilitate detection and correction? The goal of our paper is to answer these questions and, in doing so, offer theoretical and empirical guidance on how researchers can address nonresponse bias when designing surveys and analyzing survey data.

In Section 3, we describe the survey we use in our empirical analysis. The Norway in Corona Times (NCT) survey was conducted by Norway's national statistical agency to study the immediate labor market consequences of the COVID-19 lockdown that began in March 2020. The survey has three 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, the survey design included randomly-assigned financial incentives for participation. Third, Statistics Norway merged the survey data with data from administrative sources, providing a ground truth that can be used to quantify selective participation in the survey, the magnitude of nonresponse bias, and the performance of methods intended to correct for it.

In Section 4, we examine nonresponse bias using the linked survey-administrative data. The analysis delivers three broad conclusions. First, in the administrative data the labor market outcomes of those who participated in the NCT survey are substantially different from those who did not participate. If these outcomes had been responses to survey questions (as they often are), there would have been large nonresponse bias in the survey. We show that correcting for differences in a rich set of observables would have

done little to reduce this bias, implying selection on unobservables. Next, we use the randomized incentives to conduct the same comparison within each incentive group in the NCT survey. We find that trying to mitigate nonresponse bias by increasing incentives to participate can actually backfire: even though participation rates increase with incentives, nonresponse bias does too. Lastly, we show that there are also large differences between incentive groups in their responses to NCT survey questions. The differences persist after adjusting for observables, consistent with the finding in the administrative data that differences between participants and nonparticipants are primarily due to unobservable factors.

In Section 5, we apply a range of methods from the econometric literature on missing data and program evaluation that account for bias due to selection on unobservables. The methods include worst-case bounds, bounds that incorporate monotonicity assumptions, and approaches based on parametric and nonparametric selection models. These methods can be viewed as alternative procedures for extrapolation, where the data on participants is used, together with some assumptions, to extrapolate to the nonparticipants (e.g., Mogstad and Torgovitsky, 2018).

We evaluate these methods by their ability to reproduce quantities in the administrative data (which are observed for the entire population) when using only data on the survey participants. We find that some of the methods produce bounds that, while containing the population quantities, are likely to be too wide to be useful for most purposes. Other methods produce tight bounds (or point estimates) that are inconsistent with the population quantities, suggesting that the underlying assumptions are suspect. In some cases, even weak assumptions lead to severely incorrect conclusions about the population quantities.

We investigate the failure of these methods for 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 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 extrapolation methods to perform poorly.

In Section 6, we develop a new framework for extrapolation that incorporates a distinction between active and passive nonparticipation. The framework makes use of variation in participation rates due to both randomly-assigned incentives and the timing of reminder emails and text messages. We show how to use the new framework to correct for nonresponse bias and produce either bounds or point estimates on population-level quantities under different auxiliary shape restrictions. Applying the framework to our data produces bounds (or point estimates) that are narrower and closer to the truth than existing methods.

This paper is related to literatures in statistics, economics, and survey methodology

on reducing and correcting for nonresponse bias.[1] We contribute to these literatures in several ways.

First, we show how financial incentives may not only be used to increase participation rates, but also to test and correct for nonresponse bias due to unobserved differences between participants and nonparticipants. The test and corrections that we propose require that incentives for participation are randomly assigned. This suggests there are missed opportunities for randomization in surveys used for economics research, where incentives for participation are typically assigned non-randomly (as we document in Section 2).

Second, our empirical results highlight that what matters for nonresponse bias is not participation *rates*, but *who* participates. Indeed, we find that nonresponse bias may well *increase* with participation rates, contrary to common guidance on survey design.[2] For instance, the U.S. Office of Management and Budget (2006, p.60) 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). Our findings suggest that participation rates could be a poor indicator of nonresponse bias and should not necessarily be the primary concern when designing surveys; instead, it is essential that it is possible to test and correct for nonresponse bias.

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 findings in the NCT survey provide an example where the majority of nonresponse bias is explained by selection on unobservables, and thus these methods fail to correct for such a bias. Moreover, we show that some widely-used reweighting methods intended to correct for selection on observables can actually exacerbate nonresponse bias by amplifying unobservable differences.

Fourth, we evaluate the performance of existing methods that acknowledge and try to address selection on unobservables. The 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 ultimately based on a line of work by Heckman (1979); Heckman and Vytlacil (2001); Vytlacil (2002); Heckman and Vytlacil (2005, 2007) and have been applied to correct for missing survey responses in a program evaluation context by DiNardo et al. (2021). The NCT 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

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

[2]This possibility is recognized by Groves (2006), who discusses several indirect methods for measuring nonresponse bias; see also the meta-analysis by Groves and Peytcheva (2008).

effects.

Fifth, we contribute to a small and mostly theoretical literature about 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), as well as in settings with multiple instruments (Mogstad et al., 2020). While related, our multidimensional selection model is distinctly tailored to survey settings. 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.

2 A survey of surveys in economics

In this section we present six descriptive facts about the use of survey data in modern economics research, the prevalence of nonresponse, and economists' practices in coping with the possibility of nonresponse bias. We use these facts to guide our discussion in the remainder of the paper.

2.1 How we collected data on surveys

Below, we provide a description of the sources and main features of the data sets we collected to study the use of survey data in practice. Further information about data sources, record screening, construction of variables, and analysis are provided in Appendices B-E.

Data on long-run trends in the use and collection of survey data. We collected and harmonized data on publications in top-five journals from three different databases: the Web of Science database, the EconLit database, and JSTOR.[3] From each database, we obtained titles and abstracts for papers published between January 1974 and August 2020. Our final merged data set includes 11,199 records. We use this data set to construct two time series. To proxy for the use of survey data, we compute the share of publications with the word "survey", or variations such as "surveyed" or "surveys", in their title or abstract. To describe trends in survey data collection, we compute the share of publications referencing one of fourteen widely-used U.S. household surveys.

Data on the use of surveys to quickly inform economic policy. To study the use of surveys in times of increased uncertainty about the state of the economy, we use NBER Working Paper Metadata (National Bureau of Economic Research, 2020). Given lags in the publication process, this data set is more suitable for documenting researchers' response to the COVID-19 pandemic than the data on published papers described above. For consistency, we also use this data set to study the use of surveys during other periods of policy and economic uncertainty, such as the 2007-08 financial crisis.

Data on nonresponse in large-scale U.S. household surveys. We use data on trends in nonresponse rates for seven large-scale cross-sectional U.S. household surveys

[3]The top-five journals referenced throughout are the Journal of Political Economy, the American Economic Review, the Quarterly Journal of Economics, the Review of Economic Studies, and Econometrica.

that are widely used to inform policy decisions and academic research: 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).

Data on economists' practices in collecting and using survey data. To describe the prevalence and severity of nonresponse in modern economics research, as well as the ways in which researchers address potential nonresponse bias in practice, we conducted a detailed systematic review of survey-based research published in top-five economics journals between January 1st 2015 and August 31st 2020. To construct the sample, we searched the Web of Science Database for top-five publications containing the word "survey", or variations thereof, in their title, abstract or keywords. In total, 83 papers matched our criteria. Applying further screening criteria led us to restrict attention to a review sample of 73 papers. For each of these, we determined (a) whether the researchers generated their "own" survey data (as opposed to using data "borrowed" from a pre-existing survey); (b) details related to sampling, survey design and implementation; (c) the nonresponse rate; and (d) ex ante and ex post strategies used to mitigate potential bias due to nonresponse.

2.2 Six descriptive facts about how economists collect and use survey data

2.2.1 The role of survey data in economics research

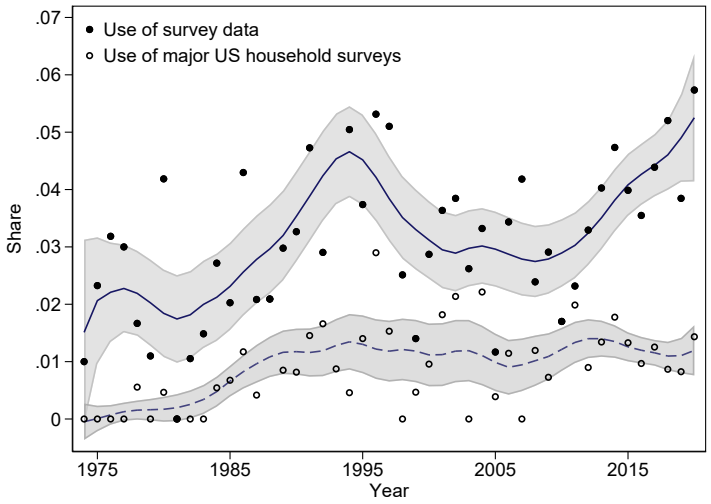
Descriptive Fact #1: *The collection and use of survey data in economics research has increased over the past decade.*

Figure 1 shows how the collection and use of survey data have evolved since 1974. The use of survey data for economics research 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 extensive, 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.[4] 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.

Descriptive Fact #2: *Survey-based research is commonly used to study rapidly-evolving changes in the economy.*

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

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 October 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, in their titles or abstracts. 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 Appendix B for more details on sample construction.

Surveys can be used to quickly generate data on the state of the economy. By contrast, government administrative data are often only available at quarterly or yearly intervals. Indeed, the primary motivation for the survey we study in this paper is to obtain timely information in the immediate aftermath of the COVID-19 lockdown in Norway. In other countries, similar survey collection efforts have emerged since March 2020. One prominent example is the Census Household Pulse survey (US Census Bureau, 2021), which aims to gather high-frequency information on the economic and health impacts of COVID-19.

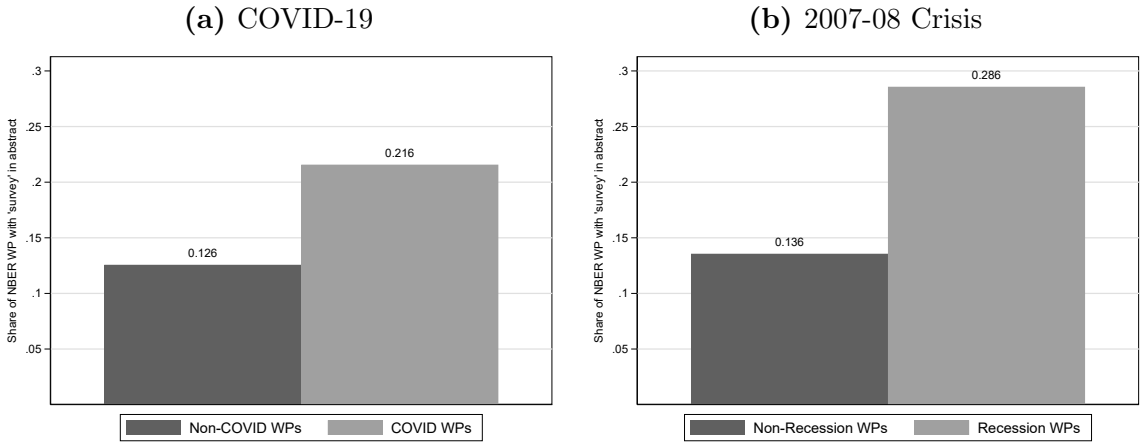
Figure 2 provides an impression of the use of survey data to track rapid economic change. As shown in panel (a), the appeal of survey data for COVID-19 research is reflected in NBER Working Papers: survey data was used in 22 percent of applied microeconomics papers studying COVID-related topics, versus only 13 percent of such papers on topics unrelated to the pandemic. Another example is economics research during the 2007-2008 financial crisis. Panel (b) shows that among applied microeconomics working papers from the two-year period surrounding the crisis, the share of recession-related papers that used surveys was twice as large as the share of other papers.

2.2.2 The prevalence of nonresponse in economics research

Descriptive Fact #3: *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

Figure 2: Use of survey data to track rapid economic change



Notes: This figure shows the frequency of survey use in applied microeconomics research during rapid periods of economic change. Working papers in applied microeconomics are identified based on the NBER program they are associated with, using the procedure in Currie et al. (2020). Panel (a): Sample consists of NBER Working Papers from March 23, 2020 (the date of publication of the earliest COVID-related NBER Working paper) until November 20, 2020. A COVID-related working paper is defined as one which includes the word "coronavirus" or "covid" in the abstract. Panel (b): Sample consists of NBER Working Papers published between October 1, 2007 and October 1, 2009. A recession-related working paper is defined as one which includes the word "crisis" or "recession" in the title or abstract. See Appendix C for more details.

nonresponse rate is 50 percent for such surveys in our review sample.[5] Among studies that 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 3 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.[6] 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 4). 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.

2.2.3 Common practices in dealing with nonresponse in economics

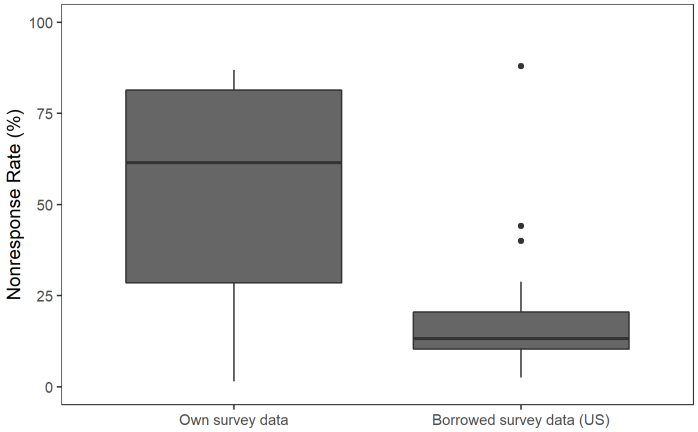
Descriptive Fact #4: *Researchers frequently omit discussion of potential nonresponse bias.*

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

[5] Studies that didn't use a probability sample (35 percent of papers using their own survey data) were excluded from our review as it is not possible to calculate comparable nonresponse rates for such studies.

[6] 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 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 Appendix E for more details.

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 is equivalent to assuming 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. Another explanation is a lack of attention to the possibility of nonresponse bias, which disregards a large literature highlighting its prevalence and consequences (see Groves and Couper, 1998; Groves, 2006; Singer, 2006, for reviews).

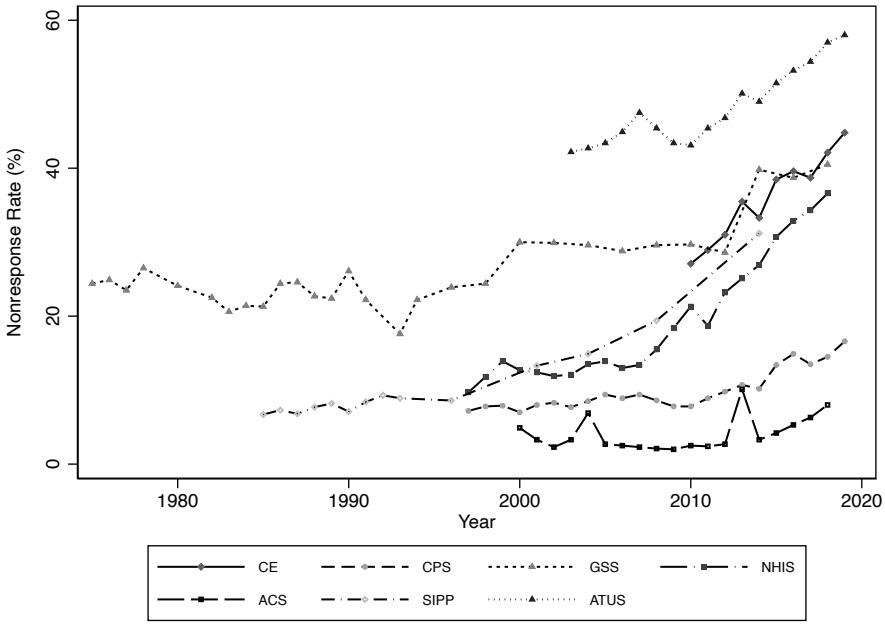
Descriptive Fact #5: *When researchers discuss potential nonresponse bias, they assume either that responses are missing completely at random, or that selection into participation is based exclusively on observables.*

Economists use two broad strategies to explicitly address potential nonresponse bias. The first is to compare respondent 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 #6: *Ex ante strategies for mitigating nonresponse bias—such as providing participation incentives—are common. These strategies are rarely designed to*

Figure 4: 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 Appendix D.

test for or address selection into survey participation based on unobservables.

The studies in our review sample 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.[7] 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 may 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.

[7] 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 Appendix F, 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.

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 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.[8] 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.[9] 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

[8] Appendix Table A.3 provides details on all variable definitions.

[9] 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 certain that the survey would give representative estimates in the absence of nonresponse bias.

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.

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 (92 percent) was invited to complete the survey online, while the remaining individuals were invited for a phone interview. The mode of contact 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 mode of contact for the phone survey was a phone call and regular mail. Throughout the paper, we focus on the random sample assigned to the online mode.

The initial survey invitation for the online sample was distributed on April 20, 2020. Figure 5 shows how the participation rate developed over time.[10] A total of six reminder messages were sent out before the survey was taken offline on May 22, 2020.[11] 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,[12] 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 2.2.2.

Individuals in the sample were randomized into one of five groups. Group assignment determined an individual's probability of receiving a prepaid credit card worth 1,000 NOK (110 USD) upon completing the survey.[13] 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

[10]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).

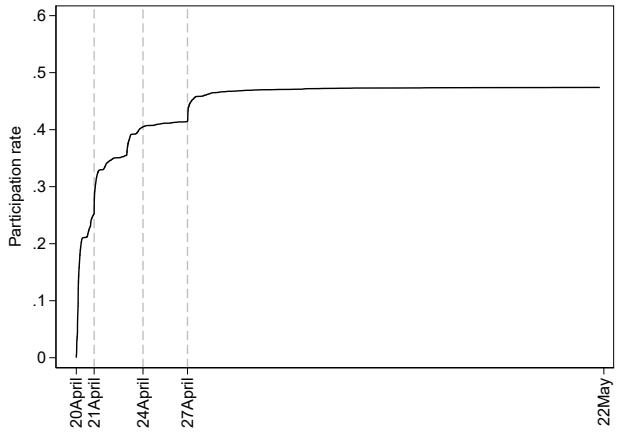
[11]On April 21 (day 1), April 24 (day 4), and April 27 (day 7) text messages and emails were sent to all individuals who had not started the survey. In addition, text messages were sent on April 23 (day 3), April 29 (day 9), and May 6 (day 15) to individuals who had started but not completed the survey.

[12]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.

[13]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.

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 groups are probabilistically identical. Balance tests for the administratively-linked outcomes are presented in Appendix Table A.1, and we confirm that outcomes do not differ significantly across the groups. Individuals were notified of the incentive in each contact attempt. They were also informed about the purpose of the survey and the estimated time it would take to complete it.

Figure 5: Participation rates over time



Notes: This figure shows the total share of individuals who participated 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.

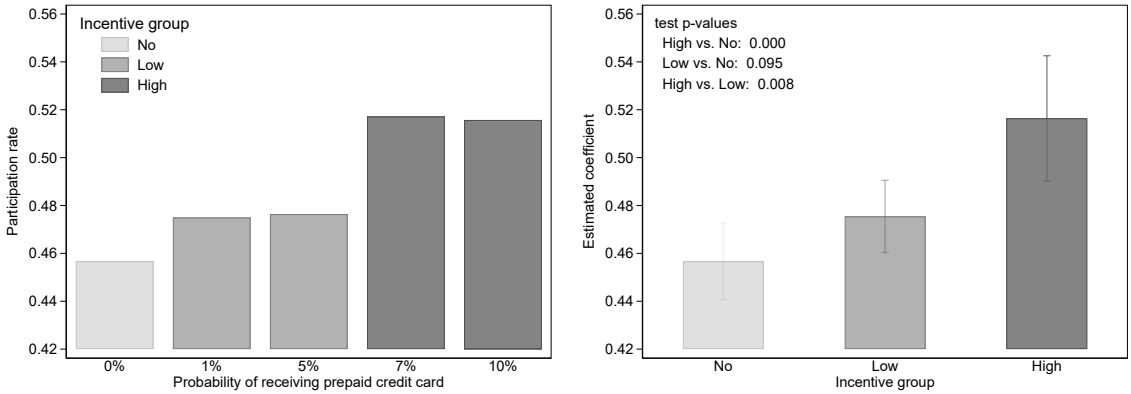
3.4 Participation rates and incentives

Figure 6a 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 gift card), "low" (1 and 5 percent probability of receiving gift card) and "no". This categorization, depicted in Figure 6b, 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 6: 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 the estimated coefficients and 90% CI from a regression of survey 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 Reliability of survey responses

Inaccurate or untruthful reporting is always a concern when using surveys. Our setting allows us to examine misreporting using survey responses for which we observe the ground truth in administrative data. Previous research suggests survey questions relating to transfer programs are particularly suited to examine the reliability of survey responses, as stigma and confidentiality concerns may lead to under-reporting.[14] To examine misreporting, we therefore focus on a question asking whether the individual applied for unemployment benefits since the lockdown. We also consider a question that is arguably less prone to misreporting: whether the individual lives with at least one child below the age of 18.

The survey responses are strikingly consistent with the administrative data: 98% of survey responses on UI applications and 95% of responses on living with children match the administrative data. The mean survey response matches that of administrative data for both variables,[15] and the consistency between survey responses and the administrative data barely varies by incentive group (see Appendix Figure A.3). This suggests that misreporting is not a concern in the NCT survey. Consistent with this finding, we find no evidence of incentives inducing different responses to the survey. In Appendix G we use the framework of Lee (2009) to show that incentives do not appear to impact responses

[14]Several factors may contribute to the under-reporting of welfare receipt, including stigma, the inclination to give socially desirable answers, concerns about confidentiality, misremembering of the timing of receipt, or confusion about program names (see, e.g. Meyer et al., 2015; Bradburn and Sudman, 1974).

[15]Participant mean for living with children is 0.34 in both survey data and administrative data, and 0.085 for applications to UI in both data sources.

directly; our estimated bounds of the effect of incentives on responses are all relatively tight around zero.

3.6 Key variables and descriptive statistics for survey respondents

We construct our variables of interest using both the survey and administrative data. From the survey, 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 benefits. From the administrative data, we use monthly earnings over the two months before and one month after lockdown, and indicators for employment two months before and one month after lockdown.

Mean outcomes for survey participants are presented in Appendix Table A.2. We consider both linked administrative and survey outcomes for participants. Taking participant means at face value, we find that average monthly earnings was 3,795 USD before the lockdown, and dropped to 3,680 USD after the lockdown. In addition to the decrease in mean earnings, employment rate estimates for participants indicate a decrease from 65 percent before the lockdown to 58 percent after the 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. The results highlight that many individuals were severely impacted by the lockdown: 14 percent of survey participants experienced a large loss in earnings, and more than 9 percent experienced employment loss. Survey outcomes further confirm that the labor market was negatively affected by the lockdown: 23 percent of participants worked fewer hours in response to the lockdown, 14 percent no longer worked full-time, and 8.5 percent applied for UI.

Of course, these descriptive statistics of the survey participants will only give an accurate description of the Norwegian economy if responses are missing at random. 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 above analysis.

4 Testing for nonresponse bias and characterizing selection

In this section we introduce a formal framework for defining and analyzing nonresponse bias. We use linked administrative data to measure nonresponse bias in the NCT survey. Then we show how researchers can use randomized incentives to test for nonresponse bias and characterize selection into survey participation without requiring linked administrative data.

4.1 Defining nonresponse bias and selection

Consider a population of individuals indexed by i. Let $Y_i^\star$ denote individual i's latent response to a survey question of interest. We want to measure the mean response across

the population, $\mathbb{E}[Y_i^\star]$, but some individuals do not participate. Let $R_i \in \{0, 1\}$ denote whether individual i participates in the survey. Then the observed response for individual i can be written as

$$Y_i \equiv \begin{cases} Y_i^\star, & \text{if } R_i = 1 \\ \text{NA}, & \text{if } R_i = 0 \end{cases}, \tag{1}$$

where NA denotes a missing observation.

It may be that an individual's decision to participate in the survey, R_i, is correlated with their latent response $Y_i^\star$. It is easy to see why this could occur in a survey like the one we study which asks questions about employment outcomes. For example, those who are more likely to participate 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 2.2, researchers routinely assume 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

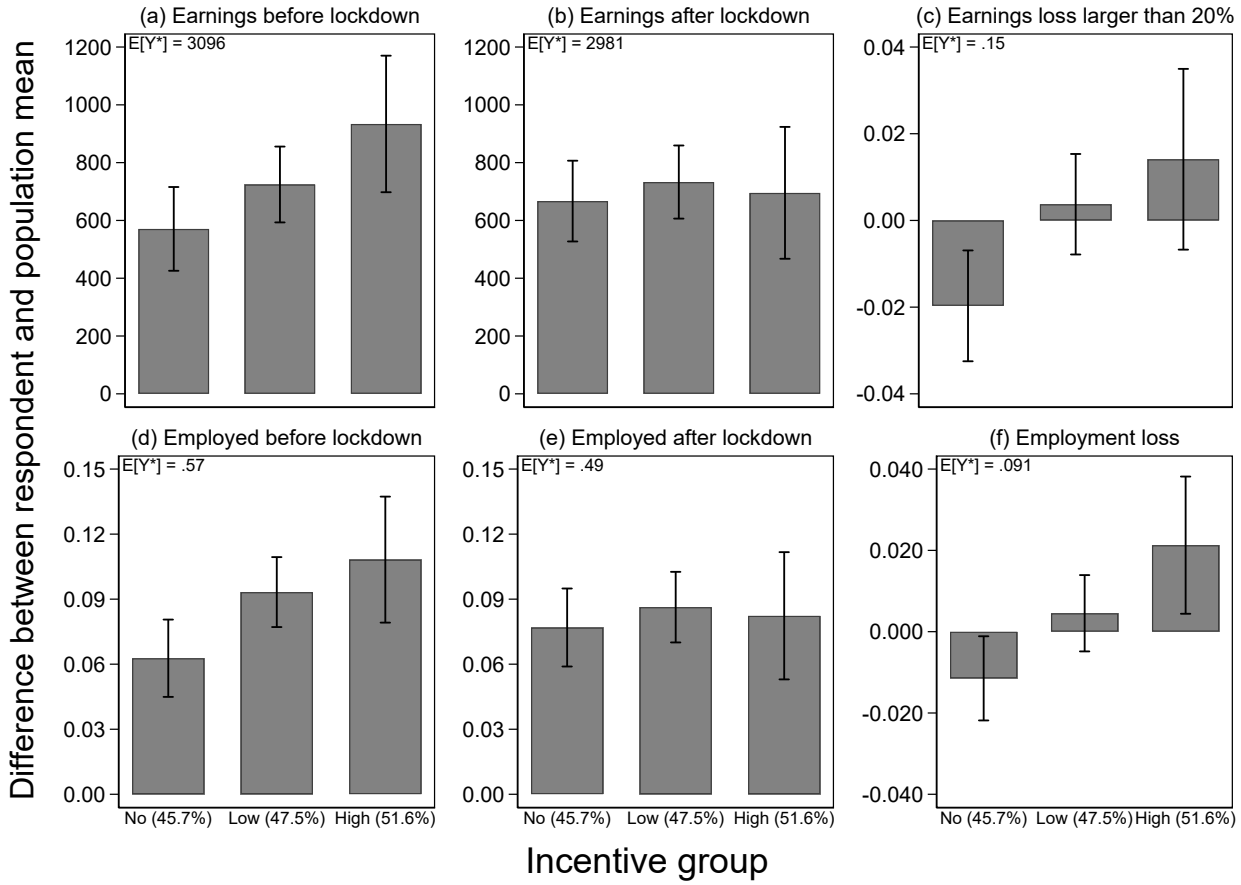
Nonresponse bias in the NCT survey

We use administrative data linked to the survey data to directly measure nonresponse bias in the NCT survey. Figure 7 reports the difference between the participant sample mean and the true population mean for each of the six administrative outcomes discussed in Section 3.6.[16] 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.

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 overestimates the true rate by 8 percentage points. A researcher or policy maker comparing

[16]Panels A and B of Appendix Table A.4 report population and participant means in table form.

Figure 7: Evidence of nonresponse bias and selection using administrative data



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. Panel A of Appendix Table A.4 presents population means by outcome, and panel B presents estimated participant means and standard errors by incentive level and outcome.

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.4).

Perhaps surprisingly, Figure 7 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. In the NCT survey, they would actually have more.

Is nonresponse bias due to selection on observables or unobservables?

In Section 2.2, 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.[17] The first specification uses municipality-level data obtained from Fiva et al. (2020): population size, gender share, elders shares, unemployment rate, and median household income.[18] The second specification uses individual-level administrative data: age, gender, immigration status, and years of schooling characteristics. In Appendix Table A.6 we show that both sets of characteristics are strong predictors of labor market outcomes and participation.

Appendix Figure A.4 reports differences between the reweighted estimates and the population mean for both propensity score specifications and each of the three survey arms.[19] 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. The majority of nonresponse bias is not explained by selection on observables.

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 often 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.[20]

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. The results are reported in Appendix H. The findings mirror those presented in this section: regardless of the method used, we consistently find substantial nonresponse bias after correcting for selection on observables. The main driver of nonresponse bias is not selection on observables, but selection on unobservables.

[17] In our systematic review of survey usage in economics, we find that researchers commonly correct for selection on observables by reweighting on individual characteristics. These characteristics are often a subset of the individual characteristics we consider in our two specifications. See Appendix Efor more details.

[18] For context, there were 356 municipalities in Norway in January 1, 2020. The average population size of a municipality is about 15,000.

[19] Panels C and D of Appendix Table A.4 report reweighted participant means in table form.

[20] 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.

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 outcomes, even though these outcomes are not observed for nonparticipants. Since the incentives are randomly assigned, each incentive arm should have the same (latent) population average response. If there is no nonresponse bias, so that participation and response are independent, then the average observed response in each incentive arm should be the same, and equal 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 3.6.[21] The measures indicate respondents 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 respondents differ from nonrespondents 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 Appendix Figure A.5. 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 surveys typically increase rather than decrease, further highlighting the importance of selection on unobservables.[22] For each reweighting specifications and outcome, we reject the null that all selection is due to observables, with all p–values < 0.1.

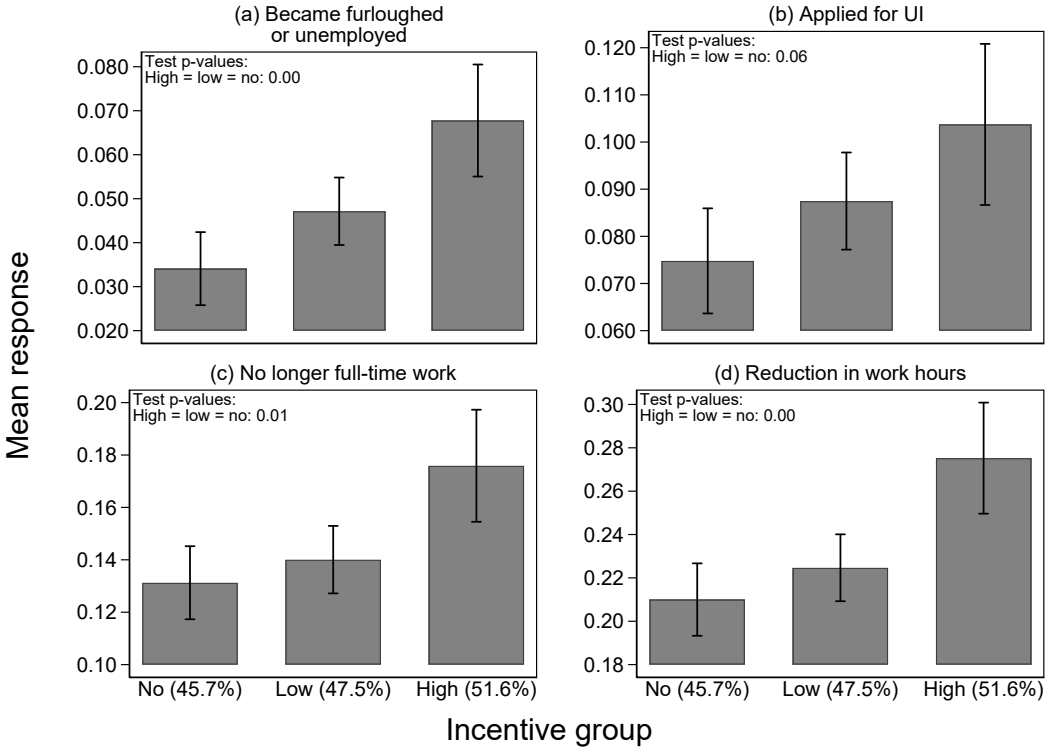
Our findings show that the estimates Statistics Norway would have obtained from the NCT survey are highly sensitive to the offered incentive level. These differences are large enough to have have important policy implications. For example, estimated expenditures on UI benefits vary drastically depending on the considered incentive arm: while the no incentive arm would indicate that UI benefits account for 13.2 percent of total budgeted experiences for Norwegian social insurance programs in 2020, the high incentive arm would indicate that this value is 18.4 percent.[23]

[21] Appendix Table A.5 reports participant means for survey-elicited measures in table form.

[22] For example, the individually-reweighted no- and high-incentive participant estimates for becoming furloughed or unemployed differ by 7.2 percentage points, which is 3.8 percentage points larger than the difference in the unweighted estimates.

[23] 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.

Figure 8: Evidence of nonresponse bias and selection using survey data



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.5 presents estimated participant means and standard errors by incentive level and outcome.

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 would not have participated in the arm with the lower incentive.

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. Let $\mathcal{Z}$ denote the set of incentives, and 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

$$R_i = \sum_{z \in \mathcal{Z}} \mathbb{1}[Z_i = z] R_i(z). \tag{2}$$

Total budgeted expenditures on national insurance amounted to about 35 percent of the state budget in 2020 (Ministry of Finance, 2020). See Appendix I for more details.

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).

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

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

We estimate the left-hand side of (3) by taking a sample mean. The marginal individuals, who comply to the incentives by participating in the survey, have $R_i(z') = 1$ but $R_i(z) = 0$, so that they do not participate at incentive level z, but would participate at $z' > z$. Using a similar argument to the one in Imbens and Angrist (1994), their average responses are identified by

$$\frac{\mathbb{E}[Y_i R_i | Z_i = z'] - \mathbb{E}[Y_i R_i | Z_i = z]}{\mathbb{P}[R_i = 1 | Z_i = z'] - \mathbb{P}[R_i = 1 | Z_i = z]} = \mathbb{E}[Y_i^\star | R_i(z) = 0, R_i(z') = 1]. \tag{4}$$

When contrasting two incentive levels, we estimate the left-hand side of (4) through an instrumental variables regression with $Y_i R_i$ as the outcome variable, R_i as the endogenous variable and Z_i as the instrument. We use the convention that $Y_i R_i = 0$ if $R_i = 0$.

How do inframarginal and marginal participants differ?

Table 1 reports average labor market outcomes using both the administrative data and NCT survey data for the inframarginal group that participates without incentives, and the marginal group that participates only under high incentives.[24] The estimates show that marginal participants had much stronger pre-lockdown labor market attachment. 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 outcomes show that marginal participants were hit substantially harder by the lockdown. Table 1 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

[24]The conclusions are similar, but estimates are noisier, when comparing inframarginals, inframarginals induced by low incentives, and inframarginals induced by high incentives. These results are reported in Appendix Table A.8.

Table 1: 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

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.

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.7 reports estimates of differences between marginal and inframarginal respondences 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 respondents 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 that nonresponse bias is driven by selection on unobservables. In this section, we attempt to correct for selection in 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 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).[25] Although the bounds contain the actual population mean, they are too wide to draw any firm conclusions. For example, we estimate that employment before lockdown is between 30 percent and 83 percent, while the actual value is 57 percent.[26]

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 (at least, weakly) than the worst-case bounds by pooling 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.

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 over 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

$$\mathbb{E}[Y_i^\star | R_i = 1, Z = z] \geq \mathbb{E}[Y_i^\star | R_i = 0, Z = z] \quad \text{for all } z, \tag{5}$$

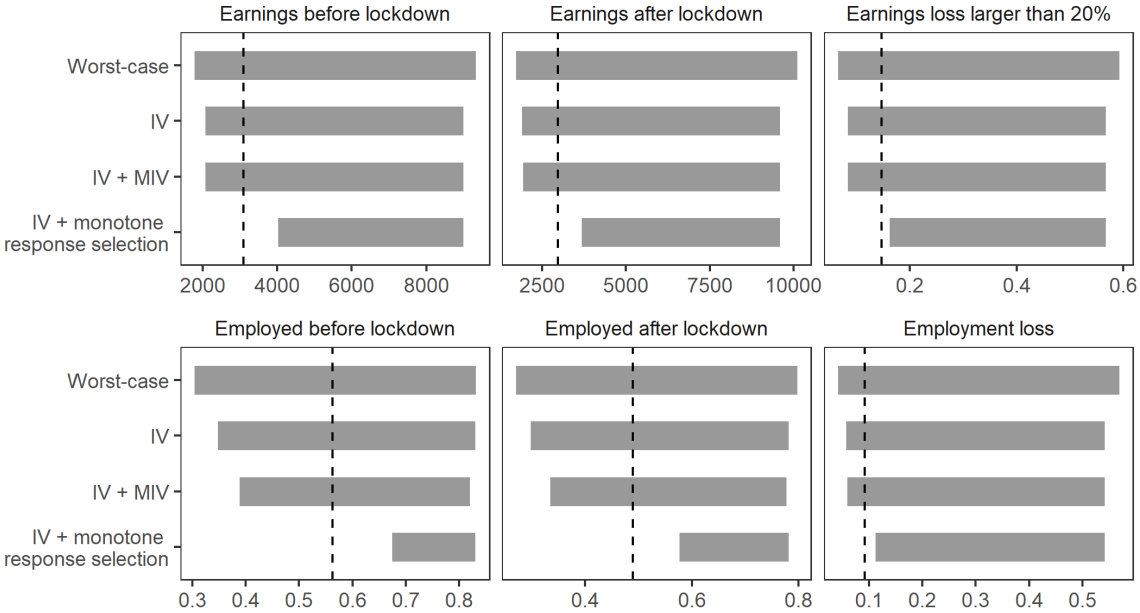
so that individuals who participate in the survey have, on average, larger outcomes than those who do not. Negative monotone response selection is the reverse inequality.

We impose monotone response selection assumptions in the directions implied by the

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

[26] 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 3).

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).

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 used 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

$$R_i = \mathbb{1}[U_i \leq p(Z_i)], \tag{6}$$

where U_i is an unobservable resistance to participating. 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 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 $U_i < p(0)$ are the inframarginal participants, who would take the survey with or without an incentive.

Selection models like (6) have a long tradition, dating back to 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), $m(u) \equiv \mathbb{E}[Y_i^\star | U_i = u]$.[27] 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 assumptions. We apply the methodology of Mogstad et al. (2018) to the survey setting in what follows; see Appendix J 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 1 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

[27] 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*.

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.[28] 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 a Heckman selection model, which assumes that $m_U(u)$ is a linear function of $\Phi^{-1}(u)$, where Φ^{-1} is the standard normal quantile function. The latter assumption is the same parameterization used in the Heckman (1974, 1979) selection model.[29] 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

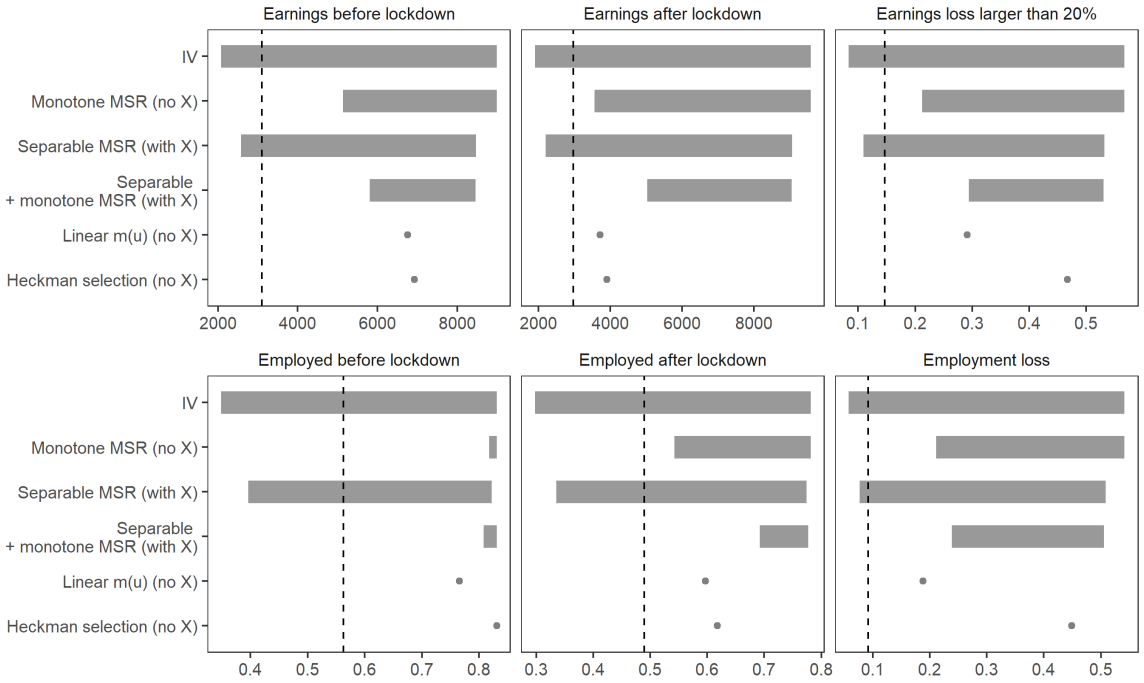
Correcting for selection on unobservables can be viewed as an extrapolation problem, where the data on participants is used, together with some assumptions, to draw inference about the nonparticipants (e.g., Mogstad and Torgovitsky, 2018). Some of the assumptions used for extrapolation in this section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

[28] 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.

[29] 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.

Figure 10: Bounds using assumptions on the MSR



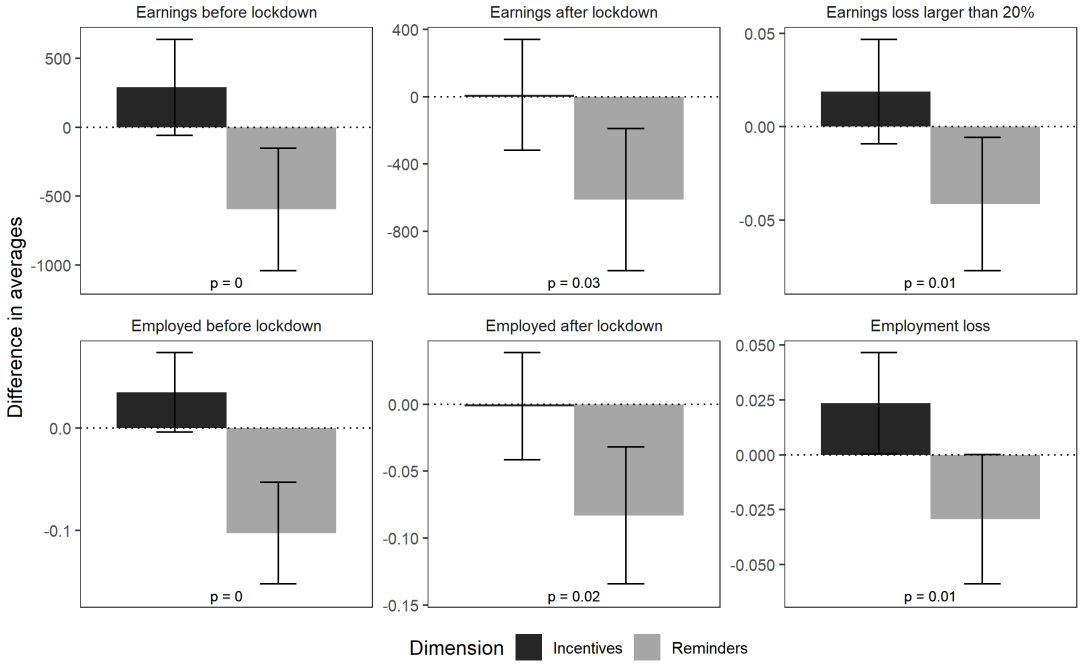
Notes: The panels in this figure show estimated bounds under the selection model in (6) and under various assumptions on the marginal survey response function (MSR). Each panel presents results for one of the six administrative outcomes. For each panel, the actual population mean is presented as a vertical 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 Appendix J for details on the other imposed assumptions and construction of estimated bounds.

based on these assumptions is 5,813 to 8,463, severely overestimating the true average pre-lockdown earnings of 3,096.

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 the email and declined to participate because the incentive was too low, and passive nonparticipants who never saw the email, but might have participated had they seen it. 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. Figure 11 compares mean differences between participants with and without incentives and after and before the reminder on April 27th (recall Figure 5). 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

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 and before the 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 and 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.

differences within the no incentive group between participants who responded after and before the reminder. Across outcomes we find weakly positive differences on the incentive dimension, but large and statistically significant negative differences on the reminder dimension.

These results suggest that participants differ along at least two unobservable dimensions. If the same is true of nonparticipants, then a model like (6) with a single source of unobserved heterogeneity could be badly misspecified. If nonparticipants are more similar to those induced by reminders than 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. In the next section, we address this problem by developing a model of survey participation that has two sources of unobserved heterogeneity.

6 A model of participation with financial incentives and reminders

In this section, we develop a model that incorporates a distinction between active (an individual declines to participate) and passive (an individual hasn't seen the invitation) nonparticipation. The model allows for variation in participation decisions due to both randomly-assigned incentives and the timing of reminder emails. We show how to use the model to correct for nonresponse bias and produce either bounds or point estimates

on the population average response under different auxiliary shape restrictions.

6.1 Model

Decisions and periods. The model has two periods. In the first period, individuals receive an email with the initial survey invitation, which includes notification of the randomized financial incentive. Individual i's first-period potential participation decision under incentive z is

$$R_{i1}(z) = \mathbb{1}[S_i = 1]\mathbb{1}[V_i \leq \eta(z)], \tag{7}$$

where $S_i = 1$ indicates that they see the email, and $V_i \leq \eta(z)$ indicates that they would have been willing to take the survey under incentive z if they had been aware of the offered financial incentive. As in Section 5, we keep the incentive binary (no and high incentives) so that $z \in \{0, 1\}$ with $\eta(0) < \eta(1)$.

We assume that the first period is long enough to collect all responses to the initial email. In our empirical analysis, we define the second period to start after the last major reminder on April 27th, which came as the response rate was flattening (Figure 5). Then, in the second period, individuals who have not yet participated receive a second email (the reminder) with notification of the original incentive level. The final participation decision for an individual assigned incentive level z is thus

$$R_{i2}(z) = R_{i1}(z) + (1 - R_{i1}(z))\mathbb{1}[S_i = 2]\mathbb{1}[V_i \leq \eta(z)] = \mathbb{1}[S_i \leq 2]\mathbb{1}[V_i \leq \eta(z)], \tag{8}$$

where $S_i = 2$ indicates seeing the reminder email but not the initial email. Let $S_i = 3$ if individual i does not see either email.

The two dimensions of heterogeneity. The model has two dimensions of unobserved heterogeneity, S_i and V_i, whereas the model in equation (6) in Section 5.4 had only one dimension, U_i. The "seeing" dimension of heterogeneity, S_i, is a categorical variable that takes on three values, which we labeled as $\{1, 2, 3\}$. The incentive dimension, V_i, is a random variable which we normalize to have support on $[0, 1]$. We do not directly observe either S_i or V_i for any individual. Instead, we observe (R_{i1}, R_{i2}, Z_i), where

$$R_{it} = Z_i R_{it}(1) + (1 - Z_i)R_{it}(0) \quad \text{for } t = 1, 2. \tag{9}$$

As before, we observe an individual's response as $Y_i = Y_i^\star$ only if they responded ($R_{i1} = 1$ or $R_{i2} = 1$).

We impose several assumptions to yield the two-dimensional model. As in Section 5.4, the separability in the incentive threshold for both equations implies that incentives do not discourage participation for any individual. We assume that the unobserved incentive dimension, V_i, does not vary over time. We have also implicitly assumed that seeing the email (S_i) is not directly affected by the incentive, which is reasonable since individuals must see the email first to learn the incentive. The assumption that incentives are randomly assigned now means that Z_i is independent of $(S_i, V_i, Y_i^\star)$.

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 current two-dimensional model they could have either $V_i > \eta(1)$ or $S_i = 3$. They differ from those with $U_i \leq p(1)$ along two dimensions, since these individuals would have participated with an incentive, and so 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 ($S_i \leq 2$) 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 ($S_i = 3$) 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 participate after the reminder ($S_i = 2$) have substantially lower earnings and employment rates than those who participate prior to the reminder ($S_i = 1$). Participation either before or after the reminder implies that $V_i \leq \eta(1)$. It is reasonable to expect that individuals who don't participate ($S_i = 3$) differ from both these groups, even if they would have responded had they seen the email, so that $V_i \leq \eta(1)$. 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.

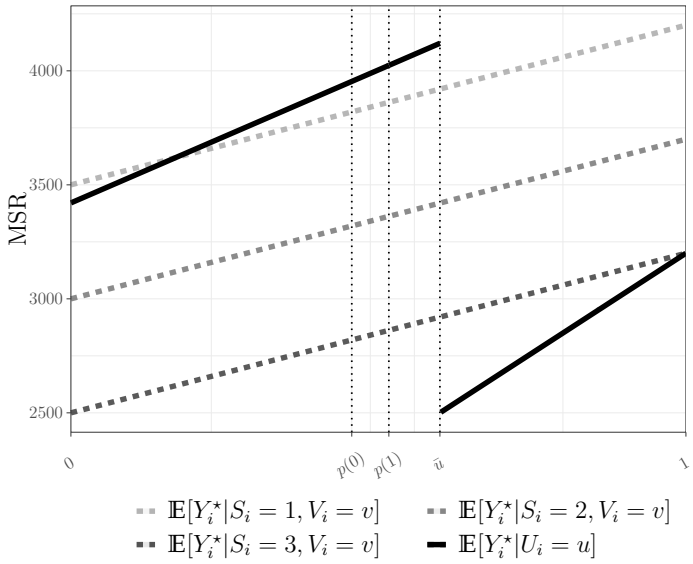
Figure 12 illustrates this argument with a numerical example. The figure plots $\mathbb{E}[Y_i^\star | V_i = v, S_i = s]$ as a function of v for each of the three values of $s \in \{1, 2, 3\}$ (dashed grey lines). The magnitude and values of the function are chosen to be roughly consistent with the variation 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 the emails, with those who saw the initial email having the highest earnings, and those who do not see either email having the lowest. Within these groups there is also systematic 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 configurations of $R_{it}(z)$, or participation groups. Table 2 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.

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 and U_i in the one-dimensional model). Dashed lines in grey present the true two-dimensional MSR: light grey depicts the MSR for individuals seeing the email before the reminder, grey depicts the MSR for individuals seeing the email after the reminder and dark grey depicts the MSR for individuals who never see the email. The solid line in black shows the implied MSR if the two-dimensional heterogeneity is collapsed into a single dimension of heterogeneity. The vertical dotted lines in black show the observed take-up propensities $p(0)$ and $p(1)$, where $p(z) = \mathbb{P}[R_i = 1|Z_i = z]$, and the value $\bar{u} \equiv 1 - \mathbb{P}[S_i = 3]$. We assume $\mathbb{E}[Y_i^\star|S_i = s, V_i = v] = 700v + 3500\mathbb{1}[S_i = 1] + 3000\mathbb{1}[S_i = 2] + 2500\mathbb{1}[S_i = 3]$. This conditional expectation was chosen based on variations in earnings by incentives and reminders that we observe in administrative data.

Incentive compliers participate in the first period if they receive an incentive, but not otherwise. Reminder compliers participate in the second period after receiving the reminder, whether incentivized or not. Reluctant compliers only participate after receiving the reminder, 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 $\mathbb{P}[R_{i1} = 0, R_{i2} = 1|Z_i = 0] - \mathbb{P}[R_{i1} = 0, R_{i2} = 1|Z_i = 1]$. 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 2 reports estimated group shares. The inframarginal and marginal groups un-

Table 2: 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 the reminder, 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 3: 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.

der the single threshold model considered in Sections 4.4 and 5.4 are now split into three complier groups of different sizes. Of the complier groups, the reminder and incentive compliers are the largest, with reluctant compliers comprising less than 1% of the population.

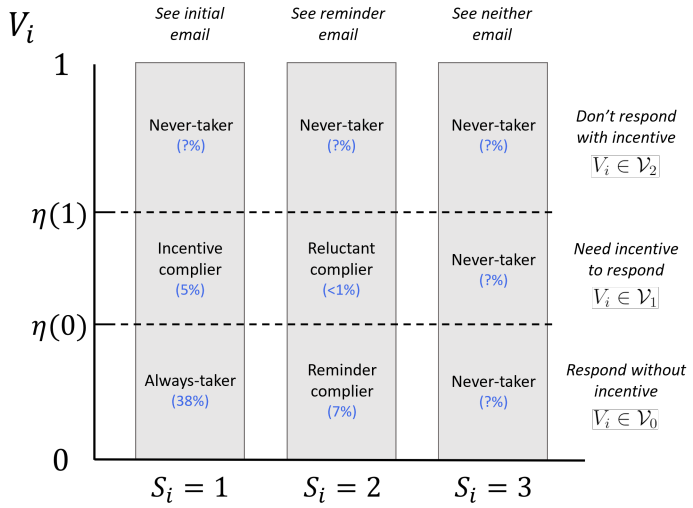
Average responses of participation groups. Average outcomes for the incentive compliers are given by

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

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

Table 3 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,

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" the invitation dimension (S_i). The y-axis depicts the heterogeneity in incentive responsiveness (V_i). Each area is named after the corresponding participation group, with the (un)known population share.

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 outcomes, 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 estimates also show that these two types of heterogeneity operate in opposing directions: for example, monthly earnings before lockdown increase for those less likely to respond to incentives, but decrease for those less likely to see the email. 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 (first email, reminder email, neither). The always-takers and three complier groups each occupy one cell, so the masses and average outcomes 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

$$\begin{aligned} \mathbb{E}[Y_i | T_i = 1, Z_i = 1] &= \mu_{01} \left(\frac{\pi_{01}}{\pi_{01} + \pi_{11}} \right) + \mu_{11} \left(\frac{\pi_{11}}{\pi_{01} + \pi_{11}} \right) \\ \text{and} \quad \mathbb{P}[T_i = 1 | Z_i = 1] &= \pi_{01} + \pi_{11}, \end{aligned} \tag{10}$$

as well as similar equations for the three other combinations of $(T_i, Z_i) \in \{(1,1), (2,0), (2,1)\}$ (see Appendix K.1 for the equations). We represent these eight equations using a vector-valued function $Q(\mu, \pi) \in \mathbb{R}^8$ whose components are zero when the equations are satisfied.

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

$$\min_{\pi \geq 0, \mu} / \max_{\pi \geq 0, \mu} \sum_{j=0}^{2} \sum_{s \in \{1,2,3\}} \pi_{js} \mu_{js} \quad \text{s.t.} \quad Q(\mu, \pi) = 0 \quad \text{and} \quad \sum_{j=0}^{2} \sum_{s \in \{1,2,3\}} \pi_{js} = 1. \tag{11}$$

When solving (11) 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 (11) is a non-convex bilinear program, it can still be solved to provable global optimality using spatial branch-and-bound algorithms (we use Gurobi Optimization (2021)). See Appendix K.1 for more details on implementation.

Results. Figure 14 reports bounds and point estimates for the 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, and 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 indeed, 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. We also assume that μ_j is increasing in j and that μ_s is decreasing in s, mimicking the patterns found in Table 3. Both assumptions are imposed by adding linear constraints to (11). These assumptions also 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. First, we assume that V_i and S_i are independent, so that an individual's email-checking behavior is unrelated to their sensitivity to incentives. In Appendix K.2, we derive the strongest testable implication of this assumption, and

we conduct a bootstrap test that fails to reject it at all conventional significance levels. Second, we choose a hypothesized proportion of the survey population that has $S_i = 3$, and so never sees either email. For the results in the main text, we impose that $\mathbb{P}[S_i = 3] = .4$, so that 40% never see the 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 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 instead decreased to 32%. Even when allowing this value drop to 10%, we show that extrapolating with the two-dimensional model outperforms the one-dimensional model.

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

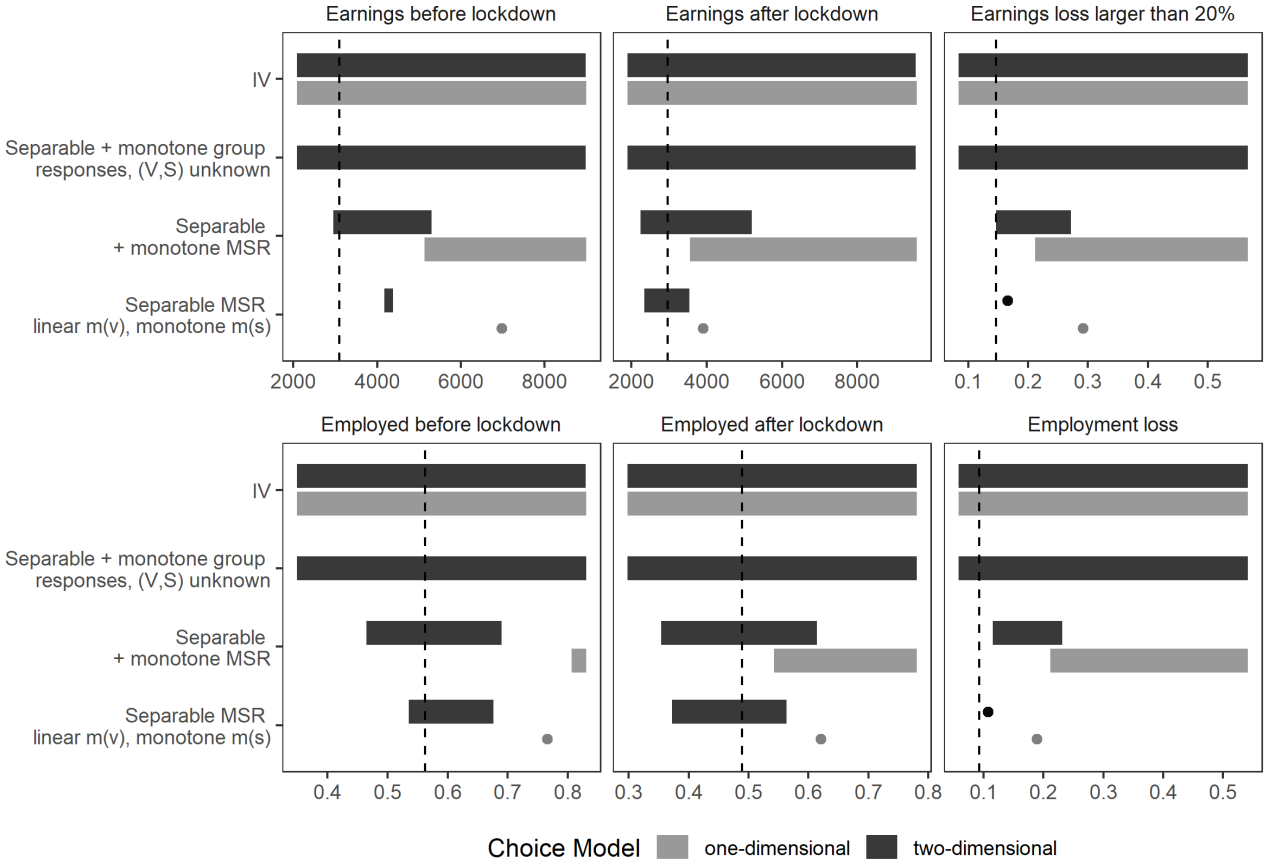
$$\mu_{js} \equiv \mathbb{E}[Y_i^\star | V_i \in \mathcal{V}_j, S_i = s] = \frac{1}{\mathbb{P}[V_i \in \mathcal{V}_j]} \int_{\mathcal{V}_j} m(v, s)\, dv, \tag{12}$$

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 (12) allows us to impose assumptions directly on the MSR function rather than the higher-level group-specific means, μ_{js}. Implementation still follows (11), but now the problem 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. The monotonicity assumptions are that m_V is increasing and m_S is decreasing. 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 never contains the truth.

In the fourth row of Figure 14 we use a semiparametric specification of the MSR in which m_V is linear. We continue to assume that m_S is decreasing. The bounds for most outcomes are tight, and in some cases points. They also get close to the truth in all cases. For example, for the three earnings outcomes, we estimate that population average monthly earnings before the lockdown are between \$4,171 and \$4,376, against a true value of \$3,096, 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%, almost the same as the true value of 16.2%. In comparison, the one-dimensional model with a linearity assumption yields point estimates that are much farther away from the true population 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 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).

7 Conclusion

Surveys are widely used to inform both academic research and policy decisions. We documented the current use of surveys in economics research. We showed that nonresponse rates are often high, but 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 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 context of the Norway in Corona Time (NCT) survey. The NCT survey randomly assigned financial incentives for participation. We showed how to use the incentives to detect nonresponse bias in both linked administrative outcomes and in the survey outcomes themselves. We found evidence of large nonresponse bias in both types of outcomes, even after correcting for

observable differences between participants and nonparticipants.

We applied a range of existing methods from the econometric literature on missing data and program evaluation that allow for unobservable differences between participants and nonparticipants. The more agnostic bounding approaches produced bounds that were too wide to be useful. Parametric and nonparametric selection models provided narrower bounds and/or point estimates, but they were often far from the administrative ground truth, suggesting that the models are misspecified.

We argued that the source of misspecification in the selection models is the reliance on a single dimension of unobserved heterogeneity, which does not allow for for non-participation to be both active (declining to participate) and passive (not seeing the survey invitation). We developed a richer choice model that addresses this shortcoming by including two dimensions of unobserved heterogeneity. Applying the model to our data produced bounds (or point estimates) that are narrow and closer to the truth than prior selection models. The results provide an example where careful attention to choice modeling can be used to successfully correct for nonresponse bias.

References

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

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.

Bradburn, N. M. and S. Sudman (1974). *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine Publishing Company.

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.

Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly 70*(1), 646–675.

Groves, R. M. and M. P. Couper (1998). *Nonresponse in Household Survey Interviews*. New York: Wiley. DOI: https://doi. org/10.1002/9781118490082.

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.

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 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.

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.

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.

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, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies 76*(3), 1071–1102.

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

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

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.

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/prp201920200001guldddpdfs.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.

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.

Shea, J. and A. Torgovitsky (2021). ivmte: An R Package for Implementing Marginal Treatment Effect Methods. *Becker Friedman Institute for Economics Working Paper* (2020-01). https://dx.doi.org/10.2139/ssrn.3516114.

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

U.S. Bureau of Labor Statistics (2018). Consumer Expenditures and Income: History. Technical report. https://www.bls.gov/opub/hom/cex/history.htm.

U.S. Census Bureau (2006). Program History. Technical report. https://www.census.gov/history/pdf/ACSHistory.pdf.

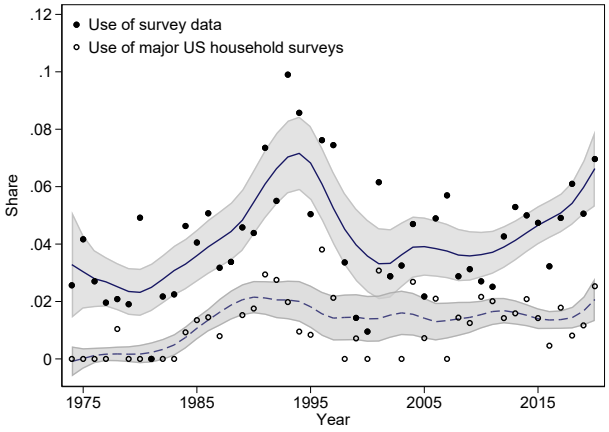
US Census Bureau (2021). Household Pulse Survey (COVID-19). https://www.census.gov/programs-surveys/household-pulse-survey.html.

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.

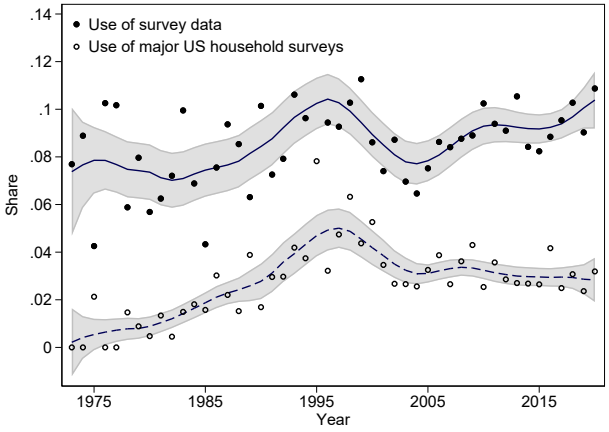
A Appendix figures and tables

Figure A.1: Use of survey data in top-five publications, applied microeconomics only



Notes: Sample consists of papers with abstract and JEL codes published in top-five economics journals between January 1974 and October 2020 that are classified as applied microeconomics. Records were obtained from the Web of Science, JSTOR, and EconLit in November 2020. The solid line represents local linear regression estimates of the share of papers that include the word "survey", or variations thereof, in their title or abstract. The dashed line represents local linear regression estimates of the 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. We use a bandwidth of 2 years with an Epanechnikov kernel. 90% confidence intervals are presented in shaded areas. See Appendix B for more details on sample construction.

Figure A.2: Use of survey data in NBER working papers



Notes: Sample consists of NBER Working Papers obtained from the NBER Metadata Website (National Bureau of Economic Research, 2020). The data starts in January 1st, 1973 and is updated through November 20, 2020. It includes 28,136 working papers. The solid line depicts the fitted values of a local linear regression of the yearly share of working papers that include the word 'survey', or variations thereof, in their titles or abstracts, on year. The dashed line represents local linear regression estimates of the share of working 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. We use a bandwidth of 2 years with an Epanechnikov kernel. 90% CIs are presented in shaded areas. See Appendix C for more details on sample construction.

Table A.1: Balance test

	(1) Female	(2) Age	(3) Years of school	(4) Immigrant	(5) Earnings before	(6) Earnings after	(7) Employed before	(8) 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.3 for details on variable definitions.

Table A.2: Descriptive statistics of participants

	Mean
Individual characteristics	
Female	0.54
Age	47.6
Years of school	13.7
Immigrant	0.10
Outcomes measured in administrative data	
Earnings before	3795.2
Earnings after	3682.6
Earnings loss	0.14
Employed before	0.65
Employed after	0.58
Employment loss	0.092
Outcomes measured in survey	
Became furloughed or unemployed	0.045
Applied for UI	0.085
No longer full-time work	0.14
Reduction in work hours	0.23

Notes: This table presents the means of background characteristics and outcomes for the participant sample. See Appendix Table A.3 for details on variable definitions.

Table A.3: Variable definitions and sources

Variable	Definition	Data source
Variables from administrative sources		
Female	Indicator for female	CPR
Immigrant	Indicator for immigrant (inv_cat = B)	CPR
Age	Individual's age	CPR
Years of school	Individual's years of school (derived from classification of education, NUS)	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 (before lockdown). Earnings are defined as the sum of the following variables: *lonn_kontant*, *lonn_natural*, *lonn_godtgjorelse*	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 > 1 'basic amount' (BA)* and not registered as either fully or partially unemployed (variable as_f ≠ 1, 2, 3 or 4) in Jan/Feb, 2020 (before lockdown)	EE and ARENA
Employed after	Indicator for average earnings > 1 'basic amount' (BA)* and not registered as either fully or partially unemployed (variable as_f ≠ 1, 2, 3 or 4) in April, 2020 (after lockdown)	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

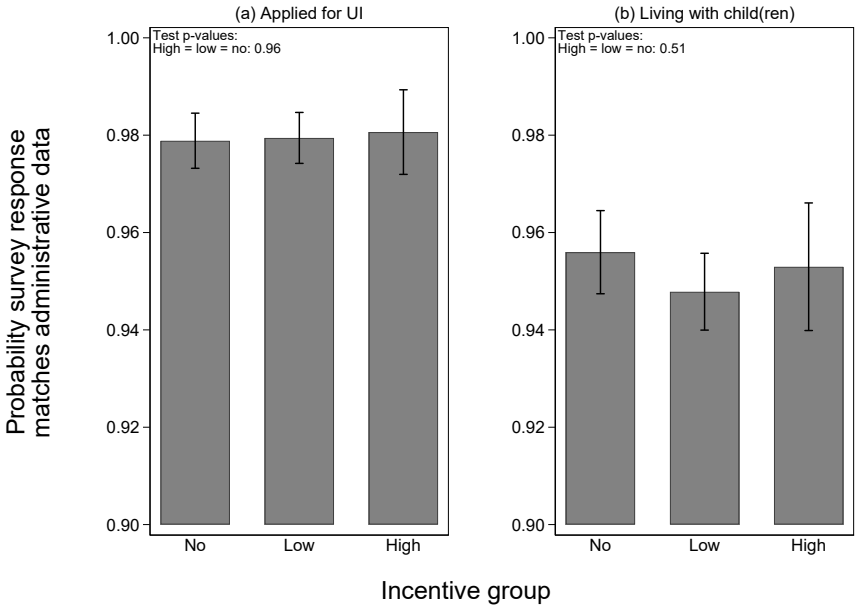
Continued on next page

Table A.3 – continued from previous page

Variable	Definition	Data source
No longer full-time	Indicator for weekly work hours $\leq$ 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
Variables from publicly available data		
Population size	Municipality number of inhabitants in 2019 (pop_kom)	Fiva et al. (2020)
Female share	Municipality female share in 2019 (women_kom)	Fiva et al. (2020)
Unemployment rate	Municipality unemployment rate in 2019 (unemployment_kom)	Fiva et al. (2020)
Share elderly	Municipality share of population aged 66 years and higher in 2019 (elderly_kom)	Fiva et al. (2020)
Median household income	Municipality median household income, in 2018 Norwegian Krone	SSB Table 06944

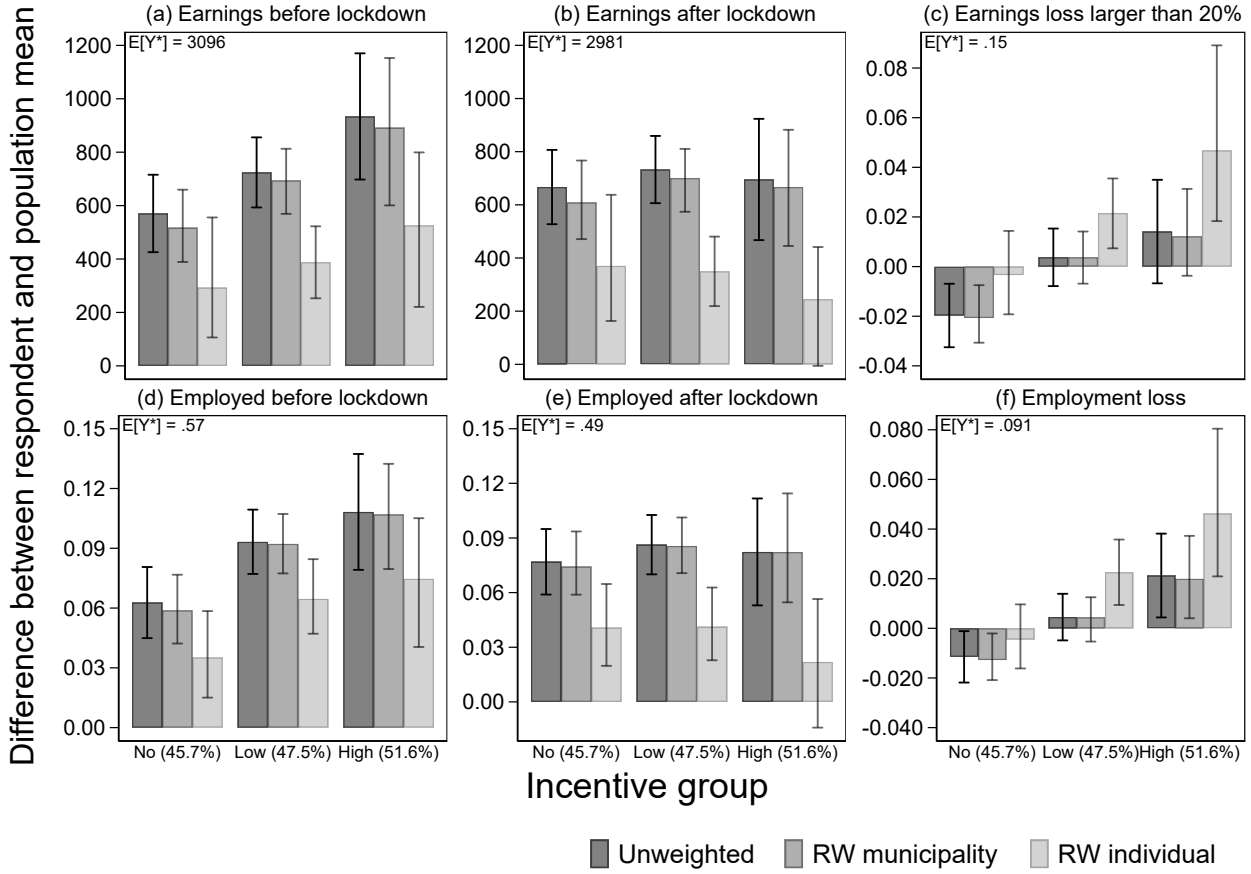
Notes: This table presents definitions and data source of all variables used throughout the paper. Data sources are abbreviated as follows: CPR=Central Population Register, NED=National Education Database, EE=Employer-employee Registry, ARENA=ARENA Registry, NCT=Statistics Norway Survey "Norway in Corona Time", SSB Table 06944=Statistics Norway, Table 06944: Household income, by type of household. Individual characteristics are defined per 4/30/2020.
*The "basic amount" (BA, substantial gainful activity level) is used by the Norwegian Social Insurance Scheme to determine eligibility for and the magnitude of benefits like old age pension, disability pension, and unemployment compensation. We use the currency rate NOK/USD=9.

Figure A.3: Match between administrative data and survey responses.



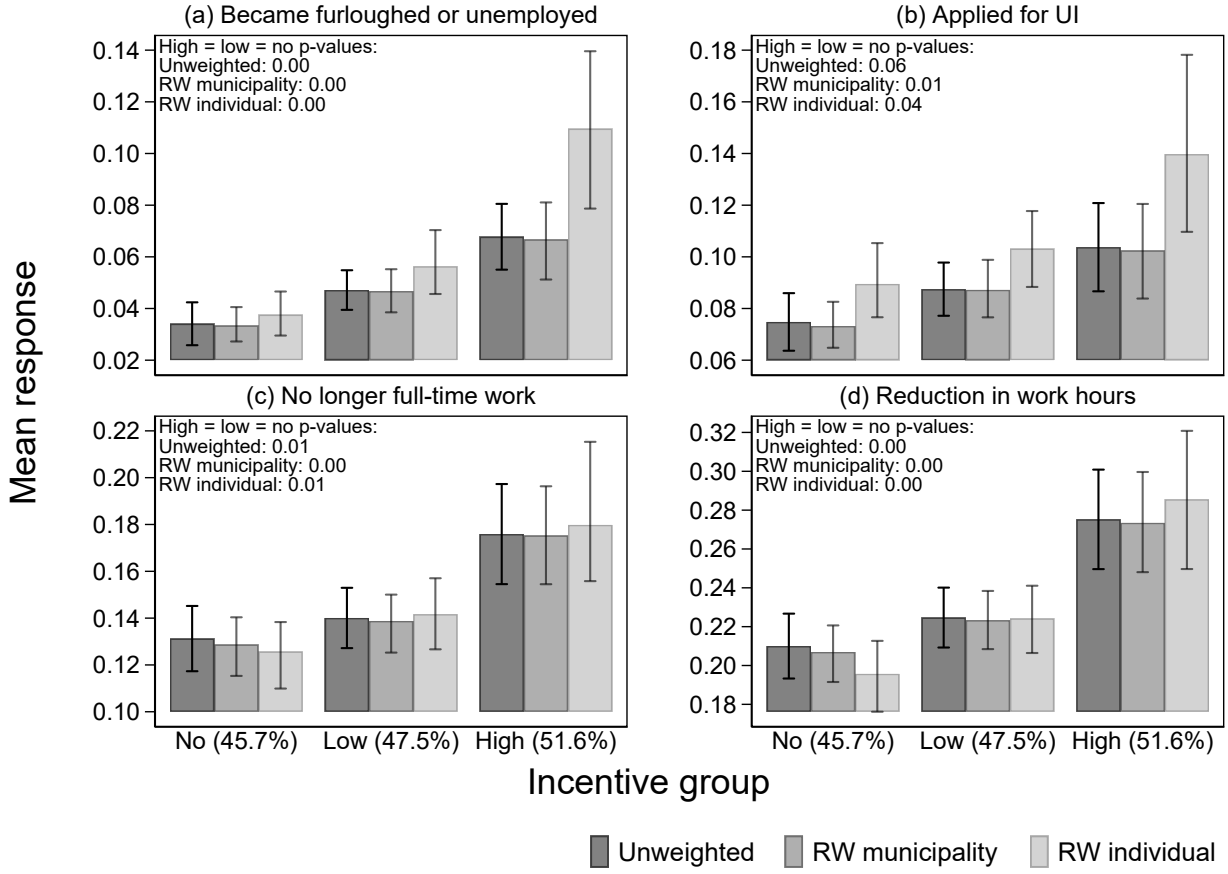
Notes: This figure plots the estimated coefficients and 90% CI from a regression of an indicator for match between survey and administrative data on the set of indicators for incentive groups. The indicator is one when the response to the survey question matches administrative data sources, and zero otherwise. Panel (a) shows the match for having applied to UI benefits after the lockdown. Panel (b) shows the match for living with a child/children below age 18. P–values for testing the equality of high vs. no incentives and high vs. low incentives are shown in upper left corner.

Figure A.4: Evidence of nonresponse bias and selection using administrative data 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 7. Lighter gray bars depict estimates reweighted by municipality-level characteristics (municipality size, gender shares, elders shares, 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 completing the survey that is linear in these characteristics, and reweight participants by the inverse of this predicted value. Error bars represent 90% confidence intervals, for reweighting we use 500 bootstrap iterations. Each panel presents results for one outcome. Population means are shown in upper left corners. Panel A of Appendix Table A.4 presents population means by outcome and panels B, C and D present estimated participant means and standard errors by incentive level and outcome, for each estimation method.

Figure A.5: Evidence of nonresponse bias and selection using survey data after reweighting



Notes: This figure shows participant responses means by incentive level and estimation method (unweighted, reweighted by municipality characteristics, and reweighted by individual characteristics) for survey-elicited outcomes. Dark gray bars depict unweighted estimates, as presented in Figure 8. Lighter gray columns bars depict reweighted by municipality-level characteristics (municipality size, gender shares, elders shares, 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 completing the survey that is linear in these characteristics, and reweight participants by the inverse of this predicted value. Error bars represent 90% confidence intervals, for reweighting we use 500 bootstrap iterations. Error bars represent 90% confidence intervals. P–values for testing the joint equality across incentive groups are shown in upper left corner. Appendix Table A.5 presents estimated participant means and standard errors by incentive level, estimation method and outcome.

Table A.4: Evidence of nonresponse bias from administrative data.

	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							
	3,095.8	2,981.3	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	
	(86.9)	(96.3)	(0.008)	(0.011)	(0.012)	(0.007)	(0.000)
Low	3,789.9	3,681.9	0.151	0.659	0.580	0.095	
	(96.8)	(100.4)	(0.009)	(0.011)	(0.011)	(0.007)	(0.000)
High	3,987.3	3,648.4	0.160	0.674	0.577	0.111	
	(192.7)	(146.8)	(0.014)	(0.018)	(0.019)	(0.012)	(0.000)
Panel D: Reweighted estimates – individual level							
No	3,388.8	3,352.0	0.144	0.602	0.535	0.086	
	(136.7)	(153.5)	(0.011)	(0.014)	(0.015)	(0.009)	(0.000)
Low	3,483.6	3,331.2	0.169	0.631	0.536	0.113	
	(89.2)	(96.4)	(0.011)	(0.013)	(0.013)	(0.009)	(0.000)
High	3,621.7	3,226.4	0.195	0.641	0.516	0.137	
	(175.7)	(140.1)	(0.023)	(0.022)	(0.023)	(0.021)	(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). We estimate a logit model of the probability of completing the survey that is linear in these characteristics, and reweight participants by the inverse of this predicted value. 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 The set of municipality-level characteristics consists of municipality size, gender shares, elders shares, unemployment rate and median household income (Fiva et al., 2020); individual-level characteristics are gender, age, years of schooling and immigration status.

Table A.5: Evidence of selection from survey data.

	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	[0.01]	[< 0.01]	[< 0.01]	[0.06]
Panel B: Reweighted estimates – municipality level				
No	0.129	0.207	0.033	0.073
	(0.008)	(0.009)	(0.004)	(0.006)
Low	0.139	0.223	0.047	0.087
	(0.008)	(0.010)	(0.005)	(0.006)
High	0.175	0.273	0.067	0.102
	(0.013)	(0.017)	(0.009)	(0.011)
p–value:				
High=Low=No	[< 0.01]	[< 0.01]	[< 0.01]	[< 0.01]
Panel C: Reweighted estimates – individual level				
No	0.126	0.196	0.038	0.090
	(0.009)	(0.010)	(0.005)	(0.009)
Low	0.142	0.224	0.056	0.103
	(0.010)	(0.011)	(0.008)	(0.009)
High	0.180	0.286	0.110	0.140
	(0.018)	(0.022)	(0.020)	(0.021)
p–value:				
High=Low=No	[< 0.01]	[< 0.01]	[< 0.01]	[0.04]

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). We estimate a logit model of the probability of completing the survey that is linear in these characteristics, and reweight participants by the inverse of this predicted value. p-values for testing the equality of mean responses across incentive arms are shown in the lower rows of each panel. The set of municipality-level characteristics consists of municipality size, gender shares, elders shares, unemployment rate and median household income (Fiva et al., 2020); individual-level characteristics are gender, age, years of schooling and immigration status.

Table A.6: 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.7: 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.8: 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.

B Construction of top-five publications data

This appendix describes the process of constructing the data sets referenced in "Data on long-run trends in the use and collection of survey data" in Section 2.1. To analyze long-run trends in top-five journal publications, we merge information from the Web of Science, JSTOR, and EconLit databases.[1] Combining sources both allows us to observe a larger number of records and reduces the possibility of false positives (records incorrectly classified as a publication in a top-five journal, or records representing other content from these journals – e.g. addenda, corrigenda) through cross-checking.[2] We subsequently construct measures of survey use and collection based on this data set.

In B.1, we define the search criteria we use to obtain records from the top-five journals from each database. In B.2, we describe how we screen these records to remove false positives and duplicates in each data set. In B.3, we explain how we merge the screened records and perform a subsequent data-quality check. In B.4, we describe how we measure survey use and collection. Finally, in B.5, we describe how we use this data set of records to construct a primary and secondary data set for our analysis. The entire process is summarized in the flow diagram in Figure B.1.

B.1 Search criteria

Below, we describe the criteria we used to query each of the three databases.

Web of Science. Accessed on 15 Nov 2020. PUBLICATION NAME: ("Journal of Political Economy" OR "American Economic Review" OR "Quarterly Journal of Economics" OR "Review of Economic Studies" OR "Econometrica"). Indexes: SCI-EXPANDED, SSCI, A&HCI, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC.

JSTOR. Accessed on 28 Nov 2020. pt:("American Economic Review") OR pt:("Journal of Political Economy") OR pt:("Quarterly Journal of Economics") OR pt:("Review of Economic Studies") OR pt:("Econometrica")

EconLit. Accessed through EBSCOhost on 30 Nov 2020. JN "American Economic Review" OR JN "Journal of Political Economy" OR JN "Quarterly Journal of Economics" OR JN "Review of Economic Studies" OR JN "Econometrica"

Using the above search criteria yielded 17,146 records from Web of Science, 23,676 records from JSTOR and 21,336 records from EconLit.

B.2 Screening

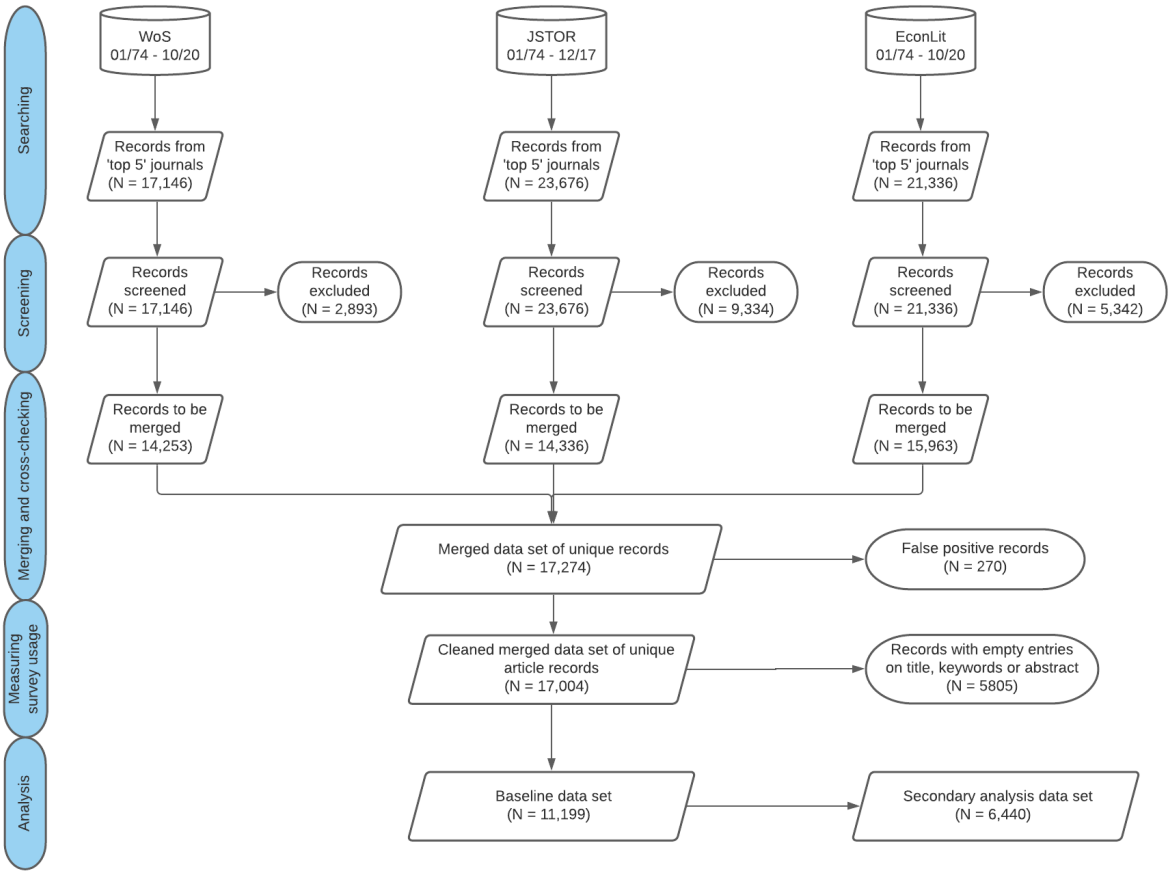
From each of the three sets of records obtained in the search stage we deleted false positives and duplicates in four steps.

First, we manually search for and delete false positives that are included in our search results because the journal name is similar to the names of the top-five journals (e.g., African Journal of Political Economy). This leads to the exclusion of 0 records from Web of Science, 750 records

[1] The "top 5" journals referenced throughout are the Journal of Political Economy, the American Economic Review, the Quarterly Journal of Economics, the Review of Economic Studies, and Econometrica.

[2] Currie et al. (2020) perform a similar analysis. Their data comes from journal websites, and as a result it only includes papers published in 2004 or later. In contrast, our data comes from the aforementioned three databases, which allow us to consider papers dating back to 1974. For the period of overlap, both data sets support our conclusions in Section 2.

Figure B.1: Record selection for data on long-run trends



Notes: This figure summarizes the sample selection process to build our database of publications in top-five journals in economics and our analysis data sets. The process is depicted in a flowchart where the top row represents the original sources and the bottom represents the final analysis data sets. The selection process consists of four steps. In the searching step, we search for all articles from top-five journals, separately for each database, using the criteria described below. In the screening step, these records are screened to exclude records not meeting the eligibility criteria described in the main text of this Appendix. In the merging and cross-checking step, we merge records from each database based on journal name, year, issue, volume, start page and end page, and cross-check to avoid duplicates and false positives. When a record appears in all three databases, we keep information about titles and abstracts from each source. Other records are further screened for eligibility by cross-checking across databases. The process produces a data set of 17,004 unique records. In the measuring survey use step, we restrict to records with abstracts to produce our baseline data set, because this allows us to proxy for survey use and collection by searching for key strings in title and abstract. In addition to the baseline data set, a secondary data set adds the restriction that a record's JEL code indicates that the record is a paper classified as being in applied microeconomics.

from JSTOR, and 2 records from EconLit. Second, we identify and remove irrelevant content like addenda, errata, corrigenda, and other notes by searching for records without page numbers or with page numbers that include Roman numerals. This excludes 0 records from Web of Science, 4,426 records from JSTOR, and 545 records from EconLit. Third, our search criteria lead to the inclusion of articles published in the May issues of the American Economic Review. These issues, known as AER Papers & Proceedings prior to 2018, contain articles that are not peer reviewed. We accordingly exclude records published in the May issue of AER before 2018. Applying these steps to each of the three databases excludes 2,893 records from Web of Science, 4,158 records from JSTOR and 4,825 records from EconLit.

Fourth, we remove duplicates within each set of records by creating a unique identifier for each record based on the following characteristics: journal name, year, issue, volume, start page and end page. We did not use authors' names or titles to eliminate the possibility that typos in those

fields would affect our deduplication process. We manually collapse duplicates identified through this process into a single record, keeping all the information from each element in the duplicate set.[3] This process removes 6 records from JSTOR and 31 records from EconLit.

After these screening steps, we are left with 14,253 records from Web of Science, 14,336 records from JSTOR and 15,963 records from EconLit.

B.3 Merging and cross-checking

We merge screened records across databases using the unique identifiers described in the previous subsection. This yields a data set of 17,274 unique records. For unique records that appear in multiple databases, we retain distinct titles and abstracts from each of the records across databases. This applies to the majority of records: 11,720 out of the 17,225 unique records (68%) appear in all three databases.

For the 5,505 (32%) unique records that do not appear in all three databases, we perform two final checks. First, we drop unique records that matched to records from another database but were previously dropped from that database during our screening process. We drop 49 records using this criterion. Second, we deduplicate based on a manual check for similar titles. This check is performed independently by two members of the research team, and leads to the elimination of 221 records.

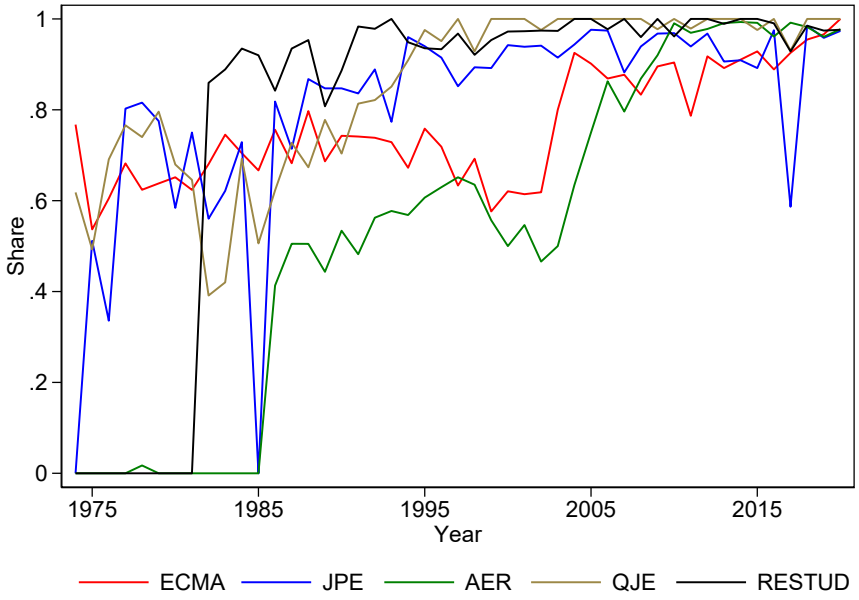
The resulting data set of merged and cross-checked records, which we refer to as the merged data set, contains 17,004 unique records. Each record represents a paper from a top-five journal between January 1974 and November 2020.

B.4 Measuring survey use and collection

We construct time series proxying for survey use and collection based on the cleaned merged data set, by searching for certain strings in the title and abstract fields for each record. We thus restrict our attention to records that have non-empty entries for these two fields. This restricted merged data set has 11,199 unique records (66% of the cleaned merged data set), and constitutes our baseline data set. Figure B.2 plots the share of records that satisfy this restriction by journal and by year. As expected, more recent years have a higher share of records with abstracts.

[3]For unique records that appear in multiple databases, we retain distinct titles and abstracts from each instance of this record across databases.

Figure B.2: Share of records with titles and abstracts



Notes: This figure plots the yearly share of records in the cleaned merged data set with non-empty titles and abstracts by journal. Sample size for this plot is 17,004 unique records.

In Section 2, to proxy for the use of survey data, we consider the share of records containing the string 'survey' (irrespective of capitalization) in either the title or abstract. Using the string allows us to capture variations including "surveyed", "surveys", etc. We identify 362 records (3.2%) satisfying this criterion.

To proxy for the type of survey data collection, we consider the share of records that include mention of large, well-known, and externally collected surveys in the United States. We identify records containing the name (or acronym) of one of the following fourteen large U.S. surveys (irrespective of capitalization): Current Population Survey (CPS), American Community Survey (ACS), Consumer Expenditure Surveys (CEX), Health and Retirement Study (HRS), National Longitudinal Survey of Youth 1979 (NLSY79), National Longitudinal Survey of Youth 1997 (NLSY97), NLSY79 Child and Young Adults (CNLSY), Survey of Income and Program Participation (SIPP), Survey of Consumer Finances (SCF), American Time Use Survey (ATUS), Survey of Consumer Expectations (SCE), General Social Survey (GSS), National Health Interview Survey (NHIS), Panel Study of Income Dynamics (PSID).[4] We identify 112 records as satisfying this criterion.

B.5 Data sets used in our analysis

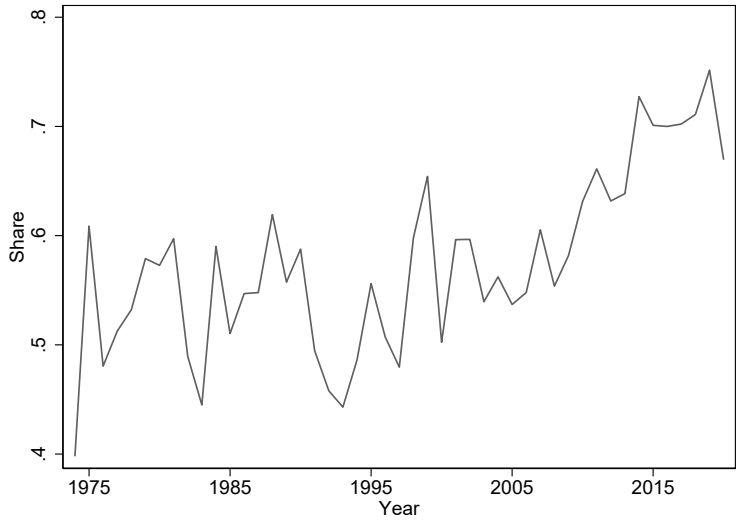
We use this data set to analyze how use of surveys and major household surveys has evolved over time, as presented in Figure 1. We also construct a secondary data set by restricting the baseline data set to records corresponding to papers classified as being in applied microeconomics. We construct this additional data set to distinguish changes in survey use from changes in the share of applied microeconomics research in the total body of published research and present the resulting trends in Appendix Figure A.1. Following Currie et al. (2020), we use JEL codes to perform this restriction.

[4]We add a space before and after the acronyms' strings to eliminate the chance of capturing other similar unrelated acronyms.

JEL codes are only available in the EconLit database.[5] Despite this limitation, 11,002 records (98% of the baseline data set) have JEL codes. As in Currie et al. (2020), we map JEL codes to fields of research following guidelines proposed by Card and DellaVigna (2013) and code records as belonging to an applied microeconomics field if all of its JEL codes belong to the following nine fields: Labor; Industrial Organization; International; Public Economics; Health and Urban Economics; Development; Lab Experiments; Welfare, Wellbeing, and Poverty; and Agriculture and Natural Resource Economics/Environmental and Ecological Economics.

Restricting to applied microeconomics papers leaves us with 6,440 records (59% of 11,002 records with abstracts and JEL codes). Figure B.3 plots the share of applied microeconomics papers in top-five journals over time, and we indeed see that such research is increasing in share over time.

Figure B.3: Fraction of applied microeconomics papers in top-five journals



Notes: This figure plots the yearly share of papers identified as belonging to an applied microeconomics field (as measured via JEL codes) among the sample of records with JEL codes.

C Construction of NBER data

This appendix describes the process of constructing the data set referenced in 'Data on the use of surveys to inform economic policy' in Section 2.1. To examine the use of surveys during periods of increased uncertainty about the state of the economy, we use publicly available NBER Working Paper Metadata (National Bureau of Economic Research, 2020), construct measures of survey use and collection for these papers, and then subset this data to two periods of economic uncertainty: the 2007-08 crisis and the COVID-19 pandemic. We also use the NBER data to construct alternatives measures of survey use and collection described in Appendix B, as a robustness check for the long-run trends.

In C.1, we describe the data, and define and construct measures of survey use and collection. In C.2, we describe how we select the subset of applied microeconomics papers. Finally, in C.3, we explain how we focus on the two periods of economic uncertainty and how we measure whether papers from these periods are related to COVID-19 or the financial crisis.

[5]Metadata of papers from EconLit contain "subject" descriptors (descriptions of JEL codes). For papers published in 1991 or later, we match descriptors to JEL codes following AEA's current JEL guide for papers published since 1991 (see American Economic Association (2021)). For papers published before 1991, we use the AEA's JEL guide from 1991 (see American Economic Association (1991)).

C.1 NBER working papers data and measuring survey use and collection

Data on NBER Working Papers is publicly available (National Bureau of Economic Research, 2020). The raw data set contains information such as titles, dates, abstracts and NBER program names for 28,136 published Working Papers between January 1st, 1973 and November 20, 2020. All NBER Working Papers contain non-empty entries for titles and abstracts.

We construct two time series proxying for survey use and collection in the exact same way as described in Appendix B.4. Based on our measure of survey use, we identify 2,518 (8.95%) NBER Working Papers as conducting research that uses surveys. Based on our measure of survey collection, we identify 891 (3.17%) NBER Working Papers as conducting research that uses major U.S. household surveys. We plot these time series in Appendix Figure A.2. We further utilize our measure of survey use to identify research using surveys to track rapid economic change. This is described in more detail in what follows.

C.2 Subsetting to applied microeconomics papers

To isolate changes in the volume of survey-based research from changes in the volume of research in applied fields, we subset to applied microeconomics papers following Currie et al. (2020), who characterize papers as applied microeconomics working papers if all of their NBER programs names belong the following list: Aging, Children, Development Economics, Education, Health Care, Health Economics, Industrial Organization, Labor Studies, Political Economy, Public Economics, International Trade, and Environment and Energy. We keep the 11,752 (42.23%) NBER Working Papers that are accordingly identified as being applied microeconomics works. In this data set, we identify 1,552 (13.21% of records in the data set) NBER Working Papers as conducting research that uses surveys.

C.3 Periods of analysis

In Section 2, we study the use of surveys during periods of increased uncertainty about the state of the economy. We consider two different time periods: the 2007-08 financial crisis and the COVID-19 pandemic.

To examine how often surveys are used in research on the 2007-08 financial crisis, we restrict our attention to the 1,950 NBER Working Papers that are identified as being applied microeconomics papers and that were uploaded between October 1, 2007 and October 1, 2009. To capture papers on the crisis, we create an indicator for each paper that is one if it contains the string "recession" or "crisis" (irrespective of capitalization). This leads to the selection of 118 NBER Working Papers (6.05% of 1,950).

To examine survey use in research on the COVID-19 pandemic, we restrict our attention to the 1,255 NBER Working Papers that are identified as being applied microeconomics works and that were uploaded between March 23, 2020 and November 20, 2020.[6] To proxy for papers about the pandemic, we create an indicator for each paper that is one if it contains the string "covid" or "coronavirus" (irrespective of capitalization). This leads to the selection of 285 NBER Working Papers (22.71% of 1,255).

D Construction of data on nonresponse in major US household surveys

This appendix describes the process of constructing the data set referenced in 'Data on nonresponse in widely-used U.S. household surveys' in Section 2.1. Throughout this appendix, we use

[6]The end date corresponds to the most recent paper available at the time we downloaded the NBER data.

the term "survey" to denote a survey data set or related data sets (CPS, NLSY79, etc). To document nonresponse rates and to analyze how these rates differ by survey and over time, we construct a longitudinal data set with yearly nonresponse rates by survey. We consider twelve large-scale U.S. household surveys.[7]

For each survey, we are interested in how the nonresponse rate evolves over time. Since survey documentations typically report response rates, we collect this information, and then take the complement (one minus the response rate) to get the nonresponse rate. The response rate is defined as "the number of interviews with reporting units divided by the number of eligible reporting units in the sample" (American Association for Public Opinion Research, 2016, p.61). One complication is that survey documentations sometimes report multiple response rates, which differ based on how "interviews with reporting unit" (the numerator of the response rate) and "eligible reporting units" (the denominator of the response rate) are defined. When this occurs, we take the highest reported response rate, thus yielding the lowest nonresponse rate. In our context of highlighting the pervasiveness of high nonresponse rates in surveys, this approach thus yields conservative estimates.

In what follows, we list and describe each of the twelve surveys. For surveys that provide information on response rates in their technical documentation, we provide the definition of the response rate we use in our analysis, and we provide details about the survey's implementation including the time frame over which response rates exist and the frequency at which the survey is conducted. For surveys that do not provide response rates, we describe the data we use to construct the response rate.

The twelve surveys can be split into two groups: cross-sectional surveys, which typically draw new individuals each time they collect point-in-time data, and longitudinal surveys, which typically consider the same group of individuals over time and collect repeated measurements for these individuals (Lavrakas, 2008). We have seven cross-sectional surveys and five panel surveys.

In what follows, Appendix D.1 considers the cross-sectional surveys, and Appendix D.2 considers the longitudinal surveys. Finally, Appendix D.3 describes the construction of the data set we use in the analysis in the main draft. In particular, we describe the (trivial) step of constructing nonresponse rates from response rates and describe how, for each survey, we aggregate nonresponse rates across waves into a yearly nonresponse rate. The resulting data set used in our analysis is a longitudinal data set of survey-by-year observations of associated nonresponse rates.

D.1 Cross-sectional surveys

We begin by considering the seven cross-sectional 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).[8]

Current Population Survey (CPS)

The CPS is administered by the Census Bureau and surveys households to produce statistics that describe the current state of the U.S. labor market.[9] It is a monthly survey that has been

[7]These surveys are the same as those used in Appendices B and C with the exception of the Survey of Consumer Expectations and the Survey of Consumer Finances, for which we were not able to find enough information on response rates.

[8]The SIPP is a collection of longitudinal surveys. We pool it with the cross-sectional surveys following the approach in Meyer et al. (2015).

[9]It is conducted using a probability sample of about 60,000 occupied households from all 50 states and the District of Columbia.

conducted since 1948.[10] We take the response rate to be the basic CPS response rate as defined by U.S. Census Bureau (2021c).[11] This response rate is defined as the number of households with completed interviews over the net housing units eligible for interviews. We obtain these response rates from Table A.7. in Czajka and Beyler (2016) for 1997-2015 and from the "Nonresponse" section of National Bureau of Economic Research (2020a) for 2015 until 2020.

American Community Survey (ACS)

The ACS is administered by the Census Bureau and surveys housing units to provide detailed population and housing information about the U.S.[12] It is a yearly survey that has been conducted since 2000.[13] We take the response rate to be the ratio of the number of units interviewed after data collection is complete to the estimate of all units that should have been interviewed (that is, an estimation of eligible units) following U.S. Census Bureau (2021a). We obtain these response rates from the table titled "Response Rates and Reasons for Noninterviews (in percent) — Housing Units" in U.S. Census Bureau (2020a), which are available yearly between 2000 and 2019.

Consumer Expenditure Surveys (CE)

The CE are administered by the Census Bureau and surveys households to find out how Americans spend their money.[14] The surveys are collected quarterly and have been conducted since the 1880s.[15] We take the response rate to be the proportion of respondents (complete and partial) over the total eligible households, and which is available from 2011 until 2020. We obtain these response rates from U.S. Bureau of Labor Statistics (2020a).

Survey of Income and Program Participation (SIPP)

The SIPP is administered by the Census Bureau and surveys households to understand income and program participation among Americans to measure the effectiveness of existing federal, state, and local programs.[16] The survey recruits a new panel every 2 to 4 years and conducts repeated surveys for each panel, and has been conducted since 1985.[17] Following the approach of Meyer et al. (2015), we take the response rate to be the number of interviewed households divided by the number of contacted and non-contacted households for the first wave of each panel. We obtain these response rates from the 'Sample Loss' column in each table of U.S. Census Bureau (2016) for 1985-2013 and the total Weighted Response Rate presented in in Table 5 in the appendix of U.S. Census Bureau (2017) for 2014.

The General Social Survey (GSS)

[10] For more details on the design and methodology of the CPS over time, see U.S. Census Bureau (2006).

[11] Alternatively, we could have considered the response rate of the Annual Social and Economic (ASEC) supplement, but as pointed out by U.S. Census Bureau (2021c), since this produces a lower response rate , we follow our method of considering the highest response rate and thus consider the basic CPS response rate.

[12] The ACS considers two different sampling units: the housing unit (HU) and the group quarters (GQ) person. We consider the HU as it is the primary unit of focus of the ACS. According to the U.S. Census Bureau (2021b), the HU sample comprises approximately 2.9 million addresses annually, while the GQ sample comprises approximately 170,000 – 200,000 individuals from GQ facilities, which include college residence halls, nursing facilities, facilities for people experiencing homelessness, etc.

[13] For more details on the design and methodology of the ACS over time, see U.S. Census Bureau (2006).

[14] The CE consist of two separate surveys, the Interview Survey and the Diary Survey. The Census Bureau selects a sample of approximately 12,000 addresses to collect data on an estimated 60 to 70 percent of total family expenditures with the Interview Survey, and a detailed daily expense record with the Diary Survey.

[15] For more details on the design and methodology of the CE over time, see U.S. Bureau of Labor Statistics (2018).

[16] The sample size ranges from approximately 14,000 to 52,000 interviewed households.

[17] The survey consists of a series of panels with short duration, as each panel ranges from 2 to 4 years. We take the first wave of each panel as if we were considering a cross-sectional survey, following the approach of Meyer et al. (2015). For more details on the design and methodology of the SIPP over time, see U.S. Census Bureau (2021).

The GSS is funded by the NORC at the University of Chicago and surveys adults to monitor and explain trends in opinions, attitudes and behaviors.[18] It is a biennial survey that has been conducted since 1972.[19] We take the response rate to be the fraction of known eligible sampled units that completed the survey. We these pull response rates from Table A.8 in NORC (2019), which includes data from 1975 until 2018.

National Health Interview Survey (NHIS)

The NHIS is administered by the National Center for Health Statistics and surveys households to monitor the health of the United States.[20] It is a yearly survey that has been conducted since 1957.[21] The survey consists of three parts: the first part asks general health-related questions about the family, and the second and third parts respectively ask specific questions for a randomly-selected adult and child (if any). We take the response rate to be the response rate from the first part of the survey, as this achieves the highest response rate. We pull response rates for the years between 1997 and 2018 from the "Family module" column of Table II of U.S. Department of Health & Human Services (2019).

American Time Use Survey (ATUS)

The ATUS is administered by the Bureau of Labor Statistics and surveys individuals to measure the amount of time people spend doing various activities, such as paid work, childcare, volunteering, and socializing.[22] It is a yearly survey that has been conducted since 2003.[23] We take the response rate to be the number of "sufficient partial interviews" over the total invited sample (regardless of eligibility) and We obtain these rates from Table 3.3. of U.S. Census Bureau (2020b).[24] Response rates are available on a yearly basis between 2003 and 2020.

D.2 Longitudinal surveys

We now turn to considering the remaining five longitudinal surveys in our set of twelve surveys: the National Longitudinal Survey of Youth 1979 (NLSY79), the NLSY79 Child and Young Adult (CNLSY), the National Longitudinal Survey of Youth 1997 (NLSY97), the Panel Study of Income Dynamics (PSID), and the Health and Retirement Study (HRS). These surveys commonly report two response rates: a cumulative response rate, measuring attrition over time; and a wave-by-wave response rate, measuring response rate at a given point of time. We focus on the latter for comparability with cross-sectional surveys, and a distinct survey is accordingly defined as a distinct wave.

One complication is that for two of the NLS surveys, response rates are not directly provided. For those surveys, we instead use provided data on number of units who participate in the survey and number of units in the sample to calculate the response rate.

National Longitudinal Survey of Youth 1979 (NLSY79)

The NLSY79 is funded by the Bureau of Labor Statistics and follows the lives of a sample of American youth born between 1957-64.[25] The sample is interviewed every 2 years and has been

[18] Total sample sizes for this survey has ranged between 2700 and 3000.

[19] For more details on the design and methodology of the GSS over time, see NORC (2019).

[20] On average, the NHIS interviews 100,000 persons in 45,000 households.

[21] For more details on the design and methodology of the NHIS over time, see IPUMS (2021).

[22] Nearly 219,000 interviews have been conducted over the past 18 years.

[23] For more details on the design and methodology of the ATUS over time, see U.S. Census Bureau (2020b).

[24] The documentation defines a "sufficient partial interview" as a case where the respondent reports at least five diary activities covering at least 21 of 24 hours. The invited sample includes noncontacted households (uncompleted callbacks, never contacted, respondent being absent, ill or hospitalized, and language barriers) and households with unknown eligibility (incorrect phone number, etc.).

[25] 12,686 young men and women aged 14 to 22 were first interviewed.

conducted since 1979.[26] We pull the response rate for each wave from Table 2 of U.S. Bureau of Labor Statistics (2020c), where reported response rates until 2018 are obtained by dividing the number of respondents interviewed over the number of individuals eligible for interview (excluding deceased).

NLSY79 Child and Young Adults (CNLSY)

The CNLSY is funded by Bureau of Labor Statistics and follows the biological children of the women in the NLSY79.[27] The survey is included as part of the NLSY79 since 1986.[28] The size of the sample depends on the number of children who reach age 15 in each survey wave. This unique feature implies that the response rate for the CNLSY depends on the mother's response rate. Thus, the BLS does not report a response rate until 2018 for the CNLSY. However, for each wave, U.S. Bureau of Labor Statistics (2020b) reports the number of interviewed children (the row *Interviewed* in Table 1) and the number of children of interviewed mothers (the row *Born* in Table 1). We calculate response rates by dividing *Interviewed* by *Born*, which, consistent with our method, yields a conservative approach for the response rate as the denominator only includes children of mothers interviewed.

National Longitudinal Survey of Youth 1997 (NLSY97)

The NLSY97 is a more recent version of the NLS79 survey and follows the lives of a sample of American youth born between 1980-84.[29] The sample is interviewed yearly and has been conducted since 1997.[30] For each wave, Table 2 of U.S. Bureau of Labor Statistics (2020d) reports the number of interviewed, non-interviewed and deceased in each wave until 2018. Consistent with NLSY79's methodology, we omit decreased when defining the total sample, and calculate response rates by dividing the number of interviewed by the number of non-interviewed (excluding decreased).

Panel Study of Income Dynamics (PSID)

The PSID is funded by the Institute for Social Research at the University of Michigan and surveys households to study the dynamics of income and poverty.[31] It is a yearly survey that has been conducted since 1968.[32] We take the response rate to be family response rate in a given wave, which is computed by Schoeni et al. (2013) by computing the proportion of families interviewed in the prior wave who completed or partially completed the interview in this wave. We pull response rates from Table 1 of Schoeni et al. (2013), which are reported between 1969 and 2009.

Health and Retirement Study (HRS)

The HRS is administered by the University of Michigan and surveys individuals to understand the challenges and opportunities of aging.[33] It is a yearly survey that has been conducted since 1992.[34] For each wave, Table 1 of Health and Retirement Study (2017) provides response rates from 1992 through 2014 and we accordingly pull them. The exact definition of the response rate varies

[26]For more details on the design and methodology of the NLSY79 over time, see National Longitudinal Surveys (2020).

[27]The number of respondent children born to NLSY79 mothers as of 2018 was 11,545.

[28]For more details on the design and methodology of the CNLSY over time, see U.S. Bureau of Labor Statistics (2020b).

[29]8,984 young men and women aged 12 to 17 were first interviewed.

[30]For more details on the design and methodology of the NLSY97 over time, see U.S. Bureau of Labor Statistics (2020d).

[31]The survey attempts to interview 18,000 individuals living in 5,000 families.

[32]For more details on the design and methodology of the PSID over time, see Institute for Social Research, University of Michigan (2021).

[33]The survey follows approx. 20,000 individuals.

[34]For more details on the design and methodology of the HRS over time, see Sonnega (2015).

slightly across time: in wave 1 (1992/1993) the response rate is the fraction of eligible individuals who completed this interview, and in follow-up waves the response rate is computed as the fraction of individuals who were attempted to be interviewed that were successfully interviewed (partial or complete) in a given wave.

D.3 Construction of data sets used in our analysis

For each survey, we thus collect response rates and start and end dates for each wave. Since the nonresponse rate is the complement of the response rate (one minus the response rate), combining this data gives us a longitudinal data set with a nonresponse rate and start and end dates for each wave. We use this data to construct a yearly data set by taking, for each survey and year, the average nonresponse rate of all waves conducted in the year. The resulting data set is one where each row is a survey×year row with the corresponding nonresponse rate.

We use the data restricted to the cross-sectional surveys to construct the series depicted in Figure 4. We use the entire data set in the systematic review described in Appendix E. Although more details can be found in that appendix, we broadly use this data to define nonresponse rates for survey-based research that employs data from these twelve surveys.

E Systematic review

This appendix describes the process of constructing the data set referenced in "Data on economists' practices in collecting and analyzing survey data" in Section 2.1. This data set is based on a systematic review of top-five economics publications that use survey data. The objective of the review is to describe the prevalence and severity of nonresponse, as well as the ways researchers attempt to deal with potential nonresponse bias. We use the Web of Science database to search for and select the records we include in our review. In Section E.1 we define our search criteria. In Section E.2 we describe how we screen these records to remove false positives and records that are outside of our scope. The process of record selection is summarized in Figure E.1 which is based on the 2020 Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) Flowchart (Page et al., 2021), the most recent version of a set of guidelines that is widely used for systematic reviews in the biomedical sciences. In Section E.3 we describe the data collected from the review sample. In Section E.4 we provide descriptive statistics for this data set.

E.1 Identification of potential studies to review

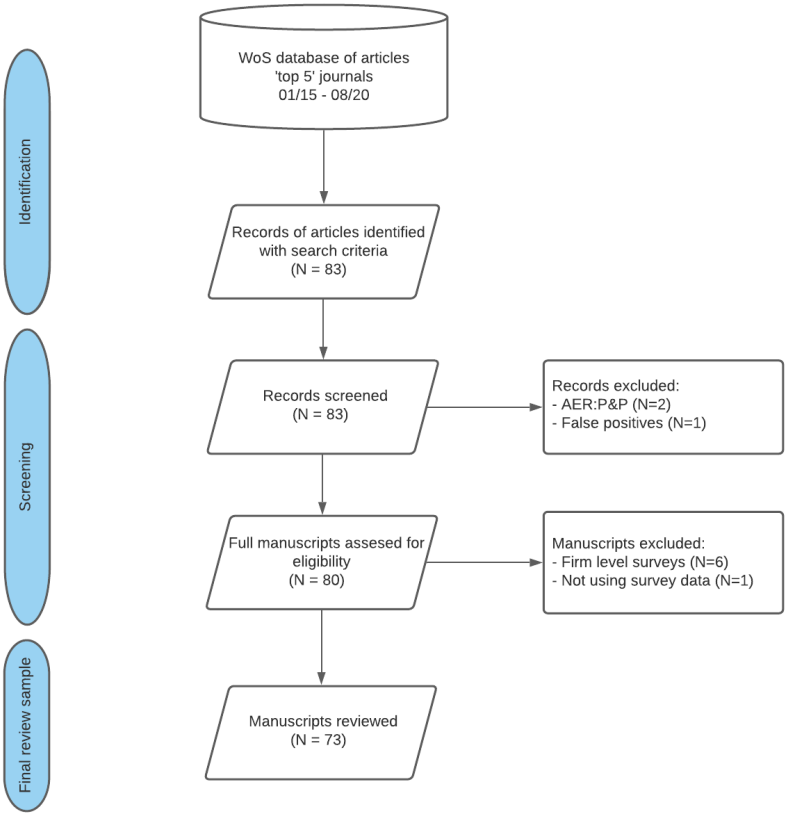
We search for papers in the Web of Science database that meet the following three eligibility criteria: (i) published in one of the top-five journals in economics, (ii) published no earlier than January 2015 and before September 2020, and (iii) used survey data collected from individuals or households. The search terms used to query the Web of Science database were:

Accessed on September 15, 2020. TOPIC: (survey) AND PUBLICATION NAME: ("Journal of Political Economy" OR "American Economic Review" OR "Quarterly Journal of Economics" OR "Review of Economic Studies" OR "Econometrica"). Indexes: SCI-EXPANDED, SSCI, A&HCI, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC.[35]

Using these search terms yielded 83 records.

[35]In Web of Science, TOPIC refers to title, abstracts and keywords. The use of the string 'survey' allows for variations like 'surveys', 'surveyed', etc

Figure E.1: Selection of studies for systematic review



Notes: This PRISMA Flow Diagram summarizes the sample selection process for our systematic review, following Page et al. (2021). *Identification* refers to the initial search using the eligibility criteria described in Appendix Section E.1. *Screening* refers to the process by which the search results are screened as described in Appendix Section E.2. *Final review sample* describes the final sample of included in the systematic review, as described in Appendix Section E.3.

E.2 Screening

We excluded ten studies because they did not meet our eligibility criteria. This was determined in two steps, each of which was conducted independently by two researchers. First, we examine the 83 records and exclude two records representing papers that were published in "AER: Papers & Proceedings", and one record of a paper without the string "survey" or its variations in its title, keywords or abstract. Second we assess the full manuscripts for the remaining papers and exclude six papers using firm-level survey data, and one paper that does not use survey data.[36] This process results in a *final review sample* of 73 studies, as depicted in Figure E.1

E.3 Collecting data from the final review sample

We now describe the data we collect from the 73 papers included in the final review sample, and the process by which this information is retrieved. Before retrieving information from each paper, we divide the sample into two categories based on the survey data source:

1. Papers that use data from a survey designed by the research team – regardless of whether the team conducted it (coded as "own survey data"), and
2. Papers that only use data from externally collected surveys so that authors have no control over the survey design (coded as "borrowed survey data").

[36]For instance, we exclude Squicciarini and Voigtländer (2015) who use a large French firm survey from the 1840s, and Ehling et al. (2018), who compared simulations to results from other papers using surveys.

Table E.1: Information collected from the review sample

	Own survey data	Borrowed survey data*
Panel A. Information collected from all papers		
Survey name		✓
Country		✓
Sampling frame	✓	
Invited sample	✓	
Probability sample	✓	
Outreach method	✓	
Response rate	✓	✓
Ex ante strategies to increase participation	✓	
Ex post strategies to mitigate potential nonresponse bias	✓	✓
Panel B. Information collected from papers using reweighting		
Exact specification	✓	✓
Method	✓	✓
Set of observables characteristics	✓	✓
Level of the characteristics	✓	✓
Characteristics from non-respondents	✓	✓

Notes: This table lists the information collected from the papers by source of the survey data – own or borrowed. For papers using borrowed survey data, the unit of observation is paper × survey, and we focus on papers using US surveys. Panel A lists the items collected for all surveys examined in the review sample. Panel B lists the items collected for the subset of papers using reweighting methods to correct for nonresponse (i.e., were identified as being in category (b) in 'Ex post strategies to mitigate potential nonresponse bias'). Data items are described in the main text of Appendix Section E.3.

Table E.1 lists the information we collect from papers in each category. For papers using borrowed survey data, we restrict ourselves to surveys conducted in the United States. Since papers may use borrowed data from more than one survey, we use paper × survey as the unit of observation.[37] We now elaborate on each piece of information collected in our review.

- *Survey name.* We collect the name of the survey(s) used in each paper.
- *Country.* We record the country where each survey was conducted.
- *Sampling frame.* We use the *Encyclopedia of Survey Research Methods* (Lavrakas, 2008) definition: "the frame represents a list (subset) of the target population from which the sample is selected".[38]
- *Invited sample.* We collect the number of individuals/households from the sampling frame invited to participate in the survey.
- *Probability sample.* We classify the invited sample as a probability sample of the sampling frame relying on the definition from Lavrakas (2008): "each member of the population has a known nonzero probability of being chosen into the sample".[39] *Outreach method.* We collect information about the mode of outreach to the invited sample. We classify methods into four categories: (a) in-person (e.g. door-to-door), (b) online (e.g. email), (c) telephone, or (d) mixed (when more than one method is used).
- *Ex-ante practices to increase participation.* We collect information on how participation into the survey was motivated beyond a single contact attempt. We classify papers into four

[37]For example, Deming (2017) use four surveys: NLSY79, NLSY97, ACS and O*NET; thus, we have four observations.

[38]This information is usually directly found in the Data section of papers and indicated by phrases such as "we sampled from...".

[39]Although it is usually clear whether the survey uses a probability sample or not, there are some cases where authors hire commercial survey or marketing companies and do not provide enough details in their paper on how the invited sample is selected. We classify these cases as non-probability sampling (e.g. Carvalho et al. (2016) and Elias et al. (2019)).

non-exclusive categories according to the strategy used: (a) surveys using intensive outreach (calling several times, re-sampling nonparticipants, etc.); (b) surveys offering monetary payments (both prepaid and postpaid) for participation; (c) surveys offering in-kind payments for participation; and (d) surveys that do not discuss the use of these practices.

- *Response rate.* We collect the response rate reported in the manuscript, and when missing, we calculate or impute it. We calculate it by dividing the number of participants over the number of invited units. Response rates from surveys with non-probability sampling were coded as *Unknown* (e.g. mTurk). We also consider cases of rates *Not reported*, which include situations where it should be possible for authors to know and report the response rate, but it is nonetheless not reported (nor is the number of invited individuals/households). We impute response rates for papers using borrowed survey data by retrieving the survey's response rate from the data collected in Appendix D. In particular, response rates are imputed perfectly if a paper used a survey for which response rates are available for every year/wave considered by the paper. In cases where the response rate is unavailable for part of the paper's time-span, we use only the overlapping period.[40] Then, we take the average response rate within the matched time range.
- *Ex post strategies to mitigate potential nonresponse bias.* We code how nonresponse to the survey is taken into account. Papers are categorized into one of the following groups: (a) papers that only present a comparison of observed characteristics between participants and invited sample, or between participants and some other external sample (e.g. Census); (b) papers that use reweighting on observable characteristics to account for nonresponse[41]; and (c) papers that do not discuss potential presence of nonresponse bias.[42]

For papers that use reweighting on observable characteristics to account for nonresponse (i.e., were identified as being in category (b) in 'Ex post strategies to mitigate potential nonresponse bias') we collect the following additional information:

- *Exact specification.* We define an exact specification of reweighting as one that would allow any researcher to reproduce the weights used. A complete exact specification consists of: (a) a detailed description of the method used; (b) the exact set of characteristics used; and (c) the way these characteristics enter the reweighting method.[43] We determine whether the complete exact specification of reweighting is described by the authors.
- *Method.* We classify reweighting methods into two broad categories: propensity weights and class weights. Papers using propensity weights are identified by statements such as "we reweight based on the inverse probability of response". Papers using class weights are identified by statements such as "we match the respondent sample in observable characteristics to ...".
- *Set of observable characteristics.* We collect the set of characteristics used by the authors to implement the reweighting correction.

[40] In total we impute 20 response rates: 16 fully matched and 4 partially matched.

[41] If a paper can be categorized into either (a) or (b), we choose category (b).

[42] Note that large U.S. surveys usually provide survey weights which are often constructed to account for the sampling design and not for nonresponse. To remain conservative we code them as if the goal is to correct for nonresponse.

[43] For example, if estimating a propensity score, researchers should specify the regression to estimate the propensity to participate in the survey, including whether characteristics enter additively or if there are interactions between them.

- *Level of the characteristics.* We code the level of observation for the characteristics into two non-exclusive groups: individual-level and geographically-aggregated level (census tract, ZIP area, county, state, etc.).[44]
- *Characteristics for nonparticipants.* We code whether the authors of the papers have access to participant-level data on background characteristics for nonparticipants. When no explicit mention on the use or availability of this information is found, we apply two heuristics: Surveys where a propensity method is used to construct weights are coded as having data on nonparticipants, since this information is needed to estimate the propensity weights. Externally collected surveys are coded as not having this data on nonparticipants. All of these surveys use class weights based on participant-level data on participants and aggregated information of the invited population.

We constructed a data extraction sheet to collect the data.[45] This sheet was pilot tested on five randomly selected articles and then refined after discussions between members of the research team. Two research assistants performed the data extraction for all included articles. Two members of the research team then reviewed data extraction, and resolved conflicts. The data is collected exclusively from the information contained in the manuscript (and its supplemental material, including appendices).

E.4 Descriptive statistics for the final review sample

Description of review sample. Figure E.2 presents characteristics of our final review sample of 73 papers. Own survey data is used in 29 papers (40 percent), while the remaining 44 papers (60 percent) borrow data from existing surveys. Among papers using own survey data, 19 papers (66 percent) use a probability sample, and 10 papers (34 percent) use a non-probability sample. Among papers borrowing survey data, 23 papers (52 percent) borrow it from at least one survey collected in the United States, while 21 papers (48 percent) borrow it exclusively from surveys collected in other countries. Finally, among the 23 papers using borrowed data from surveys collected in the U.S, we identify 32 paper×surveys.

In the remainder of this section, we elaborate on some of the descriptive facts in Section 2.

Nonresponse rates. Nonresponse rates are available in 14 of the 19 papers using own survey data and a probability sample, and they are available or imputed in 24 of the 32 paper×surveys using borrowed survey data collected in the US. Boxplots of nonresponse rates are presented in Figure 3, referenced in Descriptive Fact #3.

Ex post strategies to mitigate potential nonresponse bias. Information about ex post strategies is collected from 61 paper×surveys: 29 papers that use own survey data, and 32 paper×surveys that borrow survey data collected in the U.S. These strategies are described in Figure E.3, referenced in Descriptive Facts #4 and #5. This figure shows that reweighting on observable characteristics is used in 5 papers using own survey data, and 9 papers×survey using borrowed survey data from the U.S.

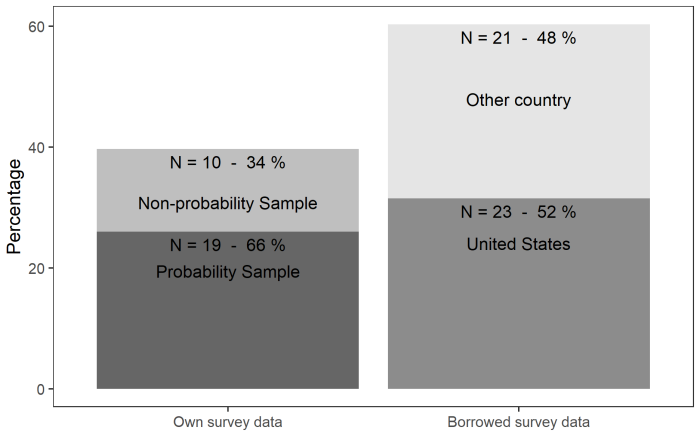
We now further describe our findings about reweighting practices.

Despite their fairly frequent use in practice, a detailed discussion of the motivation for and/or

[44] In most of the cases, we infer this information from the list of variables in the characteristics set. For instance, "age" is an individual-level characteristic but "tract age distribution" is at the geographic level.

[45] A data extraction form or data extraction sheet is a term commonly used in systematic reviews to refer to the tool used by reviewers to collect the desired information (Page et al., 2021).

Figure E.2: Composition of the final review sample



Notes: This figure describes the final review sample. Data are either "Own survey data", i.e. the survey is designed by the research team, or "Borrowed", i.e. the survey data is collected externally. "Own survey data" surveys are classified into probability sample or non-probability sample. "Borrowed" surveys are classified by whether they borrow survey data collected in the U.S. or only rely on survey data collected in other countries.

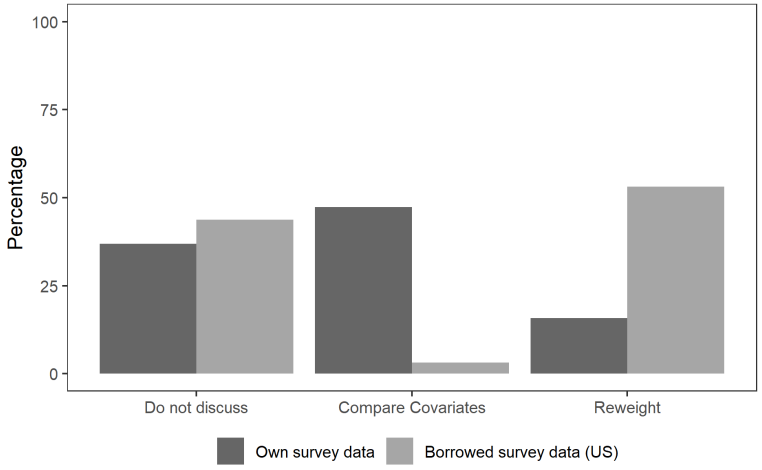
implementation of reweighting procedures is often absent.[46] For example, we find that the exact specification used to reweight survey data is described in only one out of 61 paper×surveys in our review sample. We also observe that researchers tend to use the ex post adjustment method that is best accommodated by the type of data available to them. Class weights are used as *reweighting method* in 11 paper×surveys (79 percent) and propensity weights are used in 3 paper×surveys (21 percent) (see panel A of Table E.2). None of the papers using class weights collect additional data for nonparticipants. This suggests researchers prefer to use the available information for participants only and estimate class weights instead of merging geographic-level information for each invited individual.

The set of characteristics used greatly varies across papers, although it is common to find reweighting on individual information like age, gender, race and/or geographically-aggregated information such as age distribution, racial composition, and poverty rate. Age and race are the most commonly used *characteristics* for reweighting, but others, such as gender, education and income, are also frequently used (see panel B of Table E.2). Characteristics are primarily at the individual-level, but six (more than 40 percent) of these paper×surveys use geographic-level characteristics (see panel C of Table E.2).

Ex-ante practices to increase participation. Practices to increase participation are collected from the 29 papers using their own survey data. Table E.3 presents the frequency of use of these practices by outreach method (in-person, online or telephone/mixed). We note two facts. First, strategies to increase response rates are common regardless of outreach method. Second, 52 percent of the surveys use some form of incentives, and nearly all of these (93 percent) use financial incentives. We take these facts as supporting evidence of Descriptive fact #6 where we argue that "ex ante strategies for mitigating nonresponse bias are common".

[46]This finding echoes Franco et al. (2017) who report a similar lack of 'standard operating procedure' for reweighting in political science studies using survey experiments.

Figure E.3: Treatment of nonresponse in top-five publications



Notes: This figure shows the frequency of different ex post strategies for addressing potential nonresponse bias in papers that are included in our systematic review. The sample includes all papers categorized as using "own survey data" and papers with "borrowed survey data" collected in the US. See Appendix E for details on construction of these categories. "Compare covariates" refers to papers that compare the average characteristics of their sample to some external sample averages. "Reweight" refers to papers that explicitly apply some reweighting adjustment to their estimates using observable characteristics. Papers doing both reweighting and comparing characteristics are classified as reweighting, since usually the comparison of characteristics is used as a motivation for reweighting. "Do not discuss" refers to papers that do not include any discussion about how their sample may differ from the target population and the potential effect this can have on the reported estimates.

Table E.2: Method and characteristics used for reweighting

	Frequency	**Percent**
Panel A. Method		
Class weights	11	78.6
Propensity weights	3	21.4
Panel B. Characteristics		
Age	8	57.1
Race	8	57.1
Gender	4	28.6
Education	4	28.6
Income	4	28.6
Panel C. Level of the characteristics		
Individual	10	71.4
Geographically-aggregated	6	42.9

Notes: This table describes the method, set of characteristics and level of aggregation used to reweight. Numbers are based on the papers that use reweighting adjustments – these are 5 papers using own survey data and 9 paper×surveys using borrowed survey data from the U.S. We describe the identification of the sample and coding practices in Section E.3. Panel A ('Method') presents the type of weights used (propensity weights or class weights). Panel B ('Characteristics') presents the type of characteristics used to reweight. We only include types of characteristics that are used in more than three instances. Panel C ('Level of the characteristics') shows the percentage of papers using information at each level of aggregation (non-exclusively) to reweight: individual level (e.g. individual's gender) or geographic level (e.g. gender distribution in a given area).

Table E.3: Ex ante methods to decrease nonresponse

	Own survey data	Intensive Outreach	Any incentive	Monetary incentive(s)	In-kind incentive(s)	Randomly varying incentive(s)
In-person	100%	38%	13%	13%	13%	13%
	8	3	1	1	1	1
Online	100%	13%	75%	75%	6%	6%
	16	2	12	12	1	1
Telephone/Mixed	100%	20%	40%	20%	20%	0%
	5	1	2	1	1	0
All	100%	21%	52%	48%	10%	7%
	29	6	15	14	3	2

Notes: This table presents the use of methods to increase response rates in the 29 papers using own survey data, by outreach method. The categories of ex-ante methods included in this table are described in Section E.3.

F Coffman et al. (2019) re-analysis

In this appendix, we examine the analysis of Coffman et al. (2019) (henceforth CCFK). CCFK survey applicants for Teach For America (TFA)'s transitional grants and loans (TGL) program. The authors split their sample into two groups. Within each group, individuals are randomly offered one of two incentive levels for their participation. In the first group (denoted "EC Decile 1"), individuals are told they will receive either \$20 (low) or \$40 (high) for participating. In the second group (denoted "EC Decile 2-10"), individuals are told they will receive a 0.5 percent chance (low) or a 1 percent chance (high) of receiving \$500 for participating. For each group, CCFK use the variation in incentives to test whether incentives affect participation rates and to test whether survey responses differ by incentive.

In what follows, we use CCFK's data to replicate their findings about the effect of incentives on participation rates. We then (successfully) replicate their selection test results. In particular, we replicate their finding that mean differences between the low and high incentive group participants are not statistically significant, and we fail to reject the null of no selection in their setting. However, we also observe that the CIs for these differences are wide, and do not preclude the potential for substantial selection. Accordingly, we examine whether these selection tests have sufficient power to detect selection in their setting, and conclude that they are underpowered. The data we use for this analysis is provided in supplemental materials of CCFK, which is available at Coffman et al. (2019b).[47]

Examining the effect of incentives on participation rates

CCFK find that incentives significantly increase participation rates in "EC Decile 1", but fail to find such a relationship in "EC Decile 2-10". Table F.1 presents our replication of these findings. Column (1) replicates the finding that incentives significantly increase response rates in the "EC Decile 1" group, while column (2) confirms there is no difference in response rates in the "EC Decile 2-10" group.

Table F.1: Response rates by incentive group and EC decile

	EC Decile 1	EC Decile 2-10
Low Incentive RR (%)	48.3	36.6
	(2.6)	(0.8)
High Incentive RR (%)	56.7	37.0
	(2.5)	(0.8)
Difference	8.4**	0.4
	(3.6)	(1.2)
Number of Invited Individuals	767	6,528
Number of Respondents	403	2,403

Notes: This table replicates the differences in response rates between low and high incentive group in Coffman et al. (2019). Column (1) shows the difference in response rates between low (\$20) and high (\$40) incentive for "EC Decile 1". Column (2) shows the difference in response rates between low (0.5 % chance for \$500) and high (1 % chance for \$500) incentive for "EC Decile 2 - 10". Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Testing for selection

To test for selection bias, we test for differences in participant means across the two incentive groups, as in CCFK's Appendix Table A.4. We restrict our analysis to "EC Decile 1", since it's the only group for which participation rates significantly differ by incentive. We test for significance at the .1 level. As in CCFK, we test for selection using survey-elicited responses.

[47] The data sets used are called `main_data.dta` and `survey_data.dta`. They are cleaned and merged following `main_code.do`. All of these materials can be found in Coffman et al. (2019b).

The first and second columns of Table F.2 present the respondent means in the "low" incentive and "high" incentive groups, as in CCFK. The third column, denoted 'Difference', presents mean differences between high- and low- incentive participants, and we test whether this difference is zero. Panel A presents results for mode of employment questions and Panel B presents results for credit access questions. All outcomes are binary. We replicate the authors' finding of no selection for all outcomes ($p > .1$ for each outcome). However, we note that the difference estimates are often substantial, with many estimates indicating differences of $3 - 6$ percentage points between the groups, often off of small bases.

Exploring whether selection tests are powered

The fourth column of Table F.2 presents 90% confidence intervals for the difference estimates. These confidence intervals are fairly wide, implying that the data does not preclude substantial selection.

To determine whether CCFK's tests have sufficient power to detect selection in their setting, we calculate the minimum sample size required to detect a difference of 5 percentage points for each outcome. To calculate the minimum sample size (denoted n), we use the back-to-the-envelope power calculation described by The World Bank (2020), in which $n = \left[\frac{4\sigma^2(z_{1-\alpha/2}+z_{1-\beta})^2}{D^2}\right]$. In this equation, σ is the estimated standard deviation of the considered outcome for the low-incentive group, z_x is the $x-$th quantile of a standard normal distribution, $1-\beta$ is the power of the test, α is the Type I Error, and D is the minimum detectable effect size for the mean differences between low and high incentive groups, which, as discussed above, we set to .05 to compute the corresponding minimum required sample size. We set $1-\beta \equiv 0.8$ and $\alpha \equiv 0.1$. The last column of Table F.2 depicts the required minimum sample size for each outcome under this exercise. Over all outcomes, the lowest minimum sample size is 580 while the largest is $2,484$ and nine out of the twelve outcomes require a sample size of at least $1,000$. Since the actual sample sizes in the CCFK study are 403 for mode of employment questions and 200 for credit access questions, we conclude that the study is underpowered to detect selection across participants in different incentive groups for all outcomes.

Table F.2: Comparing survey incentive groups (for "EC Decile 1")

	Low Incentive	High Incentive	Difference	90% CI of Difference	Minimum Sample Size
Mode of Employment Questions					
Teaching 0 years out (%)	77.5	81.6	4.0	[-2.7, 10.7]	1,736
	(3.1)	(2.6)	(4.1)		
Teaching 2 years out (%)	67.4	68.2	0.8	[-6.9, 8.5]	2,188
	(3.5)	(3.2)	(4.7)		
Private sector 0 years out (%)	6.2	6.9	0.7	[-3.4, 4.8]	580
	(1.8)	(1.7)	(2.5)		
Private sector 2 years out (%)	6.7	10.6	3.9	[-0.7, 8.5]	628
	(1.9)	(2.1)	(2.8)		
Graduate student 0 years out (%)	7.3	5.5	-1.8	[-5.9, 2.3]	676
	(2.0)	(1.6)	(2.5)		
Graduate student 2 years out (%)	10.7	6.9	-3.8	[-8.6, 1.0]	950
	(2.3)	(1.7)	(2.9)		
Needed additional funds (%)	52.2	50.0	-2.2	[-10.6, 6.2]	2,484
	(3.7)	(3.4)	(5.1)		
N	184	219			
Credit Access Questions					
Sought any loan (%)	86.0	87.9	1.8	[-6.1, 9.7]	1,204
	(3.6)	(3.2)	(4.8)		
Received any loan (%)	72.0	72.0	-0.1	[-10.6, 10.4]	2,016
	(4.7)	(4.4)	(6.4)		
Any denial (%)	23.7	29.0	5.3	[-4.9, 15.5]	1,808
	(4.4)	(4.4)	(6.2)		
Any discouragement (%)	30.1	29.0	-1.1	[-11.8, 9.6]	2,106
	(4.8)	(4.4)	(6.5)		
Any discouragement or denial (%)	45.2	48.6	3.4	[-8.3, 15.1]	2,478
	(5.2)	(4.9)	(7.1)		
No credit access (%)	12.9	17.8	4.9	[-3.5, 13.3]	1,126
	(3.5)	(3.7)	(5.1)		
N	93	107			

Notes: This table replicates and extends selection tests in Coffman et al. (2019) for "EC Decile 1". We only present results for "EC Decile 1" since significant differences in response rates between low-incentive and high-incentive groups is not observed for other groups (see Table F.1). To calculate the minimum sample size, we used a back-to-the-envelope power calculation, $n = \lfloor \frac{4\sigma^2(z_{1-\alpha/2}+z_{1-\beta})^2}{D^2} \rfloor$. More information on the power calculation can be found in The World Bank (2020). We set Type I Error, $\alpha = 0.1$, Power, $1-\beta = 0.8$ and population standard deviation, σ, to be equal to the low-incentive group's participant standard deviation. The difference, D, is set to 5 percentage points for all outcomes. Robust standard errors are reported in parenthesis. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

G Testing for direct effects of incentives on responses

In this appendix, we develop and implement a test for whether incentives have a direct effect on responses. As in Section 4, let Z_i denote the (randomly assigned) incentive level, let R_i be an indicator for survey participation, and let Y_i denote the observed survey response to a given question. We observe data $(Y_i R_i, R_i, Z_i)$. Furthermore, let $R_i(z)$ denote the potential participation decision under incentive level $z \in Z$. Unlike in Section 4, we also define $Y_i(z)$ to be the potential response under incentive level z. For ease of exposition, we consider the binary incentive case $Z_i \in \{0, 1\}$. The method can trivially be implemented for any pair of incentives. In our setting, $z = 0$ denotes no incentive and $z = 1$ denotes high incentive.

As in Section 4 (and the rest of the paper), we maintain the assumptions that $\{Y_i(z), R_i(z)\}_{z\in\{0,1\}}$ is independent of Z_i and that $\mathbb{P}[R_i(1) \geq R_i(0)] = 1$. The first assumption follows from random assignment of incentives. The second assumption is the monotonicity assumption from Imbens and Angrist (1994) which assumes that all individuals who participate without incentives would also participate with high incentives.

If incentives have a direct effect on responses, we would expect average treatment effects of being assigned high incentives versus no incentives on survey responses to be non-zero. Although we don't observe both $Y_i(1)$ and $Y_i(0)$ for participants, we can construct bounds on average treatment effects following the approach of Lee (2009).

Lee (2009) discusses the case when the outcome is continuous. In our setting, survey-elicited outcomes are binary, and we accordingly adapt his approach. Consider a binary outcome Y_i. We will construct treatment effect bounds on individuals who participate without incentives, i.e. on

$$\begin{aligned}\mathbb{E}[Y_i(1) - Y_i(0)|R_i(0) = 1] &= \mathbb{P}[Y_i(1) = 1|R_i(0) = 1] - \mathbb{P}[Y_i(0) = 1|R_i(0) = 1] \\ &= \mathbb{P}[Y_i(1) = 1|R_i(0) = 1] - \mathbb{P}[Y_i = 1|R_i = 1, Z_i = 0]. \qquad \text{(G.1)}\end{aligned}$$

Since the second term in (G.1) is known, it suffices to bound the first term. Let $p_k = \mathbb{P}[R_i(1) = 1|R_i(0) = k]$. Note that p_1 and p_0 are known values.[48] Then

$$\mathbb{P}[Y_i(1) = 1|R_i(1) = 1] = \mathbb{P}[Y_i(1) = 1|R_i(0) = 1]p_1 + \mathbb{P}[Y_i(1) = 1|R_i(1) = 1, R_i(0) = 0]p_0.$$

Noting that $\mathbb{P}[Y_i(1) = 1|R_i(1) = 1] = \mathbb{P}[Y_i = 1|R_i = 1, Z_i = 1]$, re-arranging the above equation yields

$$\mathbb{P}[Y_i(1) = 1|R_i(0) = 1] = \frac{1}{p_1}\mathbb{P}[Y_i = 1|R_i = 1, Z_i = 1] - \frac{p_0}{p_1}\mathbb{P}[Y_i(1) = 1|R_i(1) = 1, R_i(0) = 0].$$

Since $\mathbb{P}[Y_i(1) = 1|R_i(1) = 1, R_i(0) = 0] \in [0, 1]$, we can bound $\mathbb{P}[Y_i(1) = 1|R_i(0) = 1]$ and thus bound (G.1), which yields

$$\mathbb{E}[Y_i(1) - Y_i(0)|R_i(0) = 1] \in [\Delta_{lb}, \Delta_{ub}], \qquad \text{(G.2)}$$

with

$$\Delta_{lb} \equiv \max\left\{0, \frac{1}{p_1}\mathbb{P}[Y_i = 1|R_i = 1, Z_i = 1] - \frac{p_0}{p_1}\right\} - \mathbb{P}[Y_i = 1|R_i = 1, Z_i = 0] \qquad \text{(G.3)}$$

$$\Delta_{ub} \equiv \min\left\{1, \frac{1}{p_1}\mathbb{P}[Y_i = 1|R_i = 1, Z_i = 1]\right\} - \mathbb{P}[Y_i = 1|R_i = 1, Z_i = 0]. \qquad \text{(G.4)}$$

[48] We can write both as functions of observed quantities: $p_1 = \frac{\mathbb{P}[R_i=1|Z_i=1]}{\mathbb{P}[R_i=1|Z_i=0]}$ and $p_0 = \frac{\mathbb{P}[R_i=1|Z_i=1]-\mathbb{P}[R_i=1|Z_i=0]}{\mathbb{P}[R_i=1|Z_i=0]}$.

Given data $(Y_i R_i, R_i, Z_i)_{i=1}^n$, we can estimate treatment effect bounds $[\hat{\Delta}_{lb}, \hat{\Delta}_{ub}]$ using plug-in estimators for all probabilities in (G.3)-(G.4). If the bounds do not contain zero, we conclude that incentives have a direct effect on responses. If the bounds contain zero, we can't reject the null of no effect.

Table G.1 presents the results. For each of the four binary survey-elicited variables we consider in the paper, the first column of the Table describes the participant for each variable. The second and third columns present the estimated lower and upper bounds for the effect of being assigned the high incentive relative to no incentive on survey responses. All the estimated bounds contain zero, and we thus fail to find any evidence that incentives directly affect responses. The widths of the estimated bounds are also relatively tight. These results thus support our assumption that incentives do not affect responses.

Table G.1: Bounds on the estimated incentive effect on responses

Outcome	Mean	LB (s.e.)	UB (s.e.)
Became furloughed/unemployed	0.034	-0.034	0.043
		(0.006)	(0.012)
Applied for UI	0.075	-0.075	0.043
		(0.015)	(0.015)
No longer full-time work	0.131	-0.064	0.068
		(0.033)	(0.019)
Reduction in work hours	0.210	-0.031	0.102
		(0.033)	(0.023)

Notes: This table presents the estimated bounds on the direct effect of incentives of survey responses. Each row contains estimates for a single variable. The first column presents the mean of survey responses as a reference. The second and third columns present the estimated lower and upper bounds, respectively. Standard errors of the estimated limits are presented below the estimates and are computed using 500 bootstrap iterations.

H Correcting for nonresponse under selection on observables

In Section 4 we showed that reweighting using simple logit specifications for survey participation – which is often used in economics research – does not systematically eliminate nonresponse bias. Here, we investigate whether that conclusion changes if we use alternative methods to address selection on observables.

We rely on two sources to choose the alternative methods we consider. First, we include the methods described in chapters 3 to 5 in Little and Rubin (2019), a prominent and widely-cited book on missing data in surveys.[49] Second, we include the methods discussed in a recent review on machine learning approaches to correct for selection on observables in surveys (Buskirk et al., 2018).

This process leads us to consider thirteen different methods, which are summarized in Table H.1. As in Section 4, we consider two sets of background characteristics: municipal-level characteristics (population, share of male residents, share of elderly residents, unemployment rate, and median household income) obtained from Fiva et al. (2020); and individual-level characteristics (age, gender, immigration status, and years of schooling). Both sets of characteristics are obtained from administrative data. Our implementation closely follows the discussion in Little and Rubin (2019) and Buskirk et al. (2018). The first three methods in Table H.1 require the researcher to discretize the covariates. We dichotomize municipality-level characteristics using median values, categorize age into quartiles and transform years of schooling into education levels (high school, bachelor, or postgraduate). Methods in Panel II of Table H.1 require choices regarding estimation of tuning parameters and model specification. We describe our choices in Table H.2.

For each method, we test for selection on unobservables following the same procedure as in Section 4.2. In particular, for each method and incentive arm, we perform a joint test of no nonresponse bias across the six outcomes. The p-values for these tests are presented in Table H.3. Each row represents a method, and for each incentive arm, we perform the test using municipal background characteristics and individual background characteristics, as in Section 4.2. We find significant nonresponse bias after adjusting for observables across all adjustment methods, incentive arms, and sets of characteristics. We therefore conclude that there is indeed significant and substantial selection on unobservables.

We next examine the extent to which these methods reduce nonresponse bias, relative to the methods we considered in Section 4.2. For each adjustment method, Figure H.1 depicts box-plots of the estimated absolute standardized nonresponse bias for each adjustment method across all combinations of administrative outcomes (6), incentive levels (3), and covariate sets (2). The absolute standardized bias is defined as the absolute value of the estimated nonresponse bias over the standard deviation of the outcome in the population. This approach is commonly used to compare estimates of nonresponse bias across variables with different scales. See, for instance, Groves (2006). 'UNW' and 'BL' represent the unweighted and reweighted methods considered in Section 4. The methods we consider in this appendix perform similar to the one we considered in the Section 4.2, and thus do not change the conclusion that nonresponse bias is not primarily due to observables.

[49] We do not implement the model-based approaches discussed in the other chapters of the book for two reasons. First, these are not used in practice by applied researchers or by major household surveys (see Appendix E and Meyer et al. (2015)). Second, such methods are not easily implementable using existing software packages, as they require human judgment for making modeling decisions.

Table H.1: Correction methods for nonresponse implemented methods

Method	Abbr.	Description	Spec.
I. Methods described in Little and Rubin (2019)			
Post-stratification	CALJ	Class weights matching the joint distribution of the covariates in the population (see p. 51)	B
Raking	RAK	Class weights matching the marginal distribution of the covariates in the population (see p. 52)	B
Saturated	PS	A response propensity model with fully saturated covariates (Horvitz–Thompson estimator, see p. 46)	B
Baseline logit	BL	Logit model of the participation propensity (see p. 48)	A
Logit with complex specification	CL	Logit model including first order interactions and second order terms (see p. 48)	A
Simple regression	ImpSB	Regression imputation (see p. 62)	A
Simple regression with complex specification	ImpSC	Regression imputation including first order interactions and second order terms (see p. 62)	A
Multiple imputation	ImpMC	Imputation method that adds stochastic uncertainty iteratively (see p. 85)	A
II. Machine learning methods described in Buskirk et al. (2018)			
Adaptive LASSO	ALASSO	Least absolute shrinkage and selection operator algorithm described by Signorino and Kirchner (2018)	A
Support Vector Machine	SVML	Supervised learning model algorithm following Kirchner and Signorino (2018)	A
Neural Network classification	NN	Classification algorithm described by Eck (2018)	A
Random Forest	RF	Prediction model of participation using a random classification method following Buskirk (2018)	A
Random Forest Imputation	ImpRF	Random Forest algorithm to estimate the imputation function as described by Buskirk (2018)	A

Notes: This table presents the methods implemented to adjust for nonresponse. The column "Abbr." presents the abbreviated name of the method. The column "Spec." describes the model specification. In specification A we use the full covariate space. For methods "CL", "ALASSO" and "ImpSC" we include first order interactions and second order terms in the covariate space. In specification B all covariates are discretized. We use these to match the joint or marginal distribution of the classes in the population or to fully saturate the propensity model. Post-stratification and raking weights are implemented using Lumley (2020) R package, machine learning algorithms for propensity weights are implemented using Kuhn et al. (2020) R package, multiple imputation is implemented using van Buuren et al. (2015) R package, and Random Forest is implemented via the R package by Stekhoven and Bühlmann (2012).

Table H.2: Further implementation details of machine learning algorithms

Abbr.	Description
ALASSO	We estimate the tuning parameter for the ALASSO method along with the degree of the polynomial expansion using 10-fold cross-validation applied to the training data set, minized for a penalty of 1.14 and a polynomial degree of 3.
SVML	We employ a soft-margin SVM with a linear kernel. To determine the tuning parameter and the kernel tuning parameter, we conduct 10-fold cross-validation applied to the training data.
NN	We construct a neural network of 5 neurons in a single hidden layer using 100 training iterations
RF	We estimate the tunning parameter using 10-fold cross-validation of the training data along with a one-standard error rule and run a random forest using 500 classification trees.
ImpRF	We estimate the tunning parameter using 10-fold cross-validation of the training data along with a one-standard error rule and run a random forest to predict the outcome using 100 classification trees Buskirk (2018).

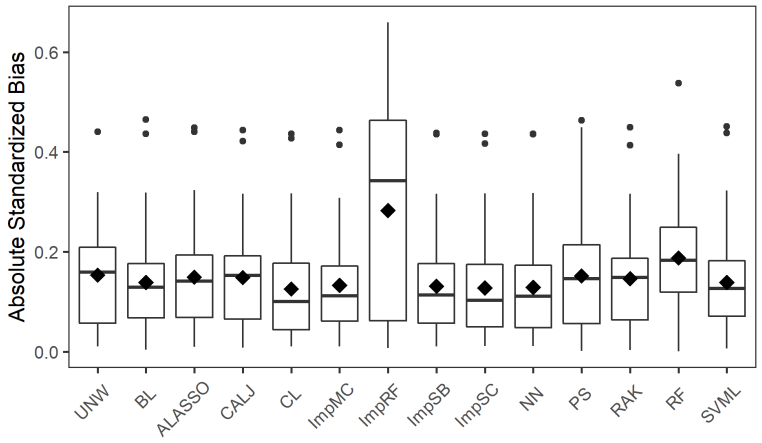
Notes: This table describes the estimation choices made to implement Machine learning algorithms (ML). ML methods are first implemented on a training data set and extrapolated on the whole data set to avoid over-fitting. We use a random subsample of 20% the unincentivized sample as our training data set.

Table H.3: Nonresponse bias test across adjustment methods

	No		Low		High	
Adjustment	**Muni**	**Ind**	**Muni**	**Ind**	**Muni**	**Ind**
ALASSO	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
BL	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
CALJ	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
CL	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
ImpMC	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
ImpRF	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
ImpSB	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
ImpSC	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
NN	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
PS	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
RAK	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
RF	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
SVML	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

Notes: This table presents the estimated joint p-values testing the null of no nonresponse bias across the 6 administrative outcomes: earnings before and after lockdown, earnings loss larger than 20%, employed before and after lockdown, and employment loss. The name of the adjustment method is the abbreviation presented in table H.1.

Figure H.1: Absolute standardized nonresponse bias



Notes: This figure presents box-plots on the estimated absolute standardized nonresponse bias. For each adjustment method, the box-plot describes the set of absolute standardized biases for all combinations of administrative outcomes (6), incentive levels (3), and covariate sets (2). 'UNW' refers to unadjusted, while the rest uses the same abbreviation described in Table H.1. The filled diamonds depict the mean of each set of estimates.

I UI expenditure projections based on survey data

Surveys can offer real-time information to inform urgent policy decisions. However, this may come at a cost in terms of lower information quality. In this section, we explore the potential magnitude of that cost by calculating survey-based projections of the government's unemployment insurance (UI) expenditures in the month following the lockdown. We consider how much projections vary across survey arms, and how far they are from projections based on the true UI application rate, which we observe in hindsight.

We start by using the survey-based variable *application to UI* to estimate the number of UI applicants in the population for the period from lockdown to survey participation (approximately six weeks). The fraction of respondents who report having applied to UI varies from 7.5 percent in the no-incentive arm to 10.4 percent in the high-incentive arm. Multiplying these rates by the size of the Norwegian population aged 18 and above yields 318,673 and 441,893 UI applicants, respectively. We then turn to estimating expenditures on UI benefits if all applicants were granted benefits. In 2019, the average benefit level among UI recipients was 1,833 USD per month (The Norwegian Labour and Welfare Administration, 2019). Multiplying the average receipt by the estimated number of UI applicants yields projections of 584 and 810 million USD for total UI expenditures. To provide a sense of magnitude, we compare these projections to average monthly budgeted expenditures for the Norwegian social insurance programs in 2020.[50]

Table I.1 summarizes the projected expenditures based on survey data and compares them to projections based on administrative data observed in hindsight. The projected expenditures on UI benefits, as a share of total expenditures on national insurance, range from 13.2 percent if based on the no-incentive survey to 18.4 percent if based on the high-incentive survey. In both surveys, projections are off by 14 to 20 percent relative to projections based on the true UI application

[50]The Norwegian social insurance programs cover 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 social insurance programs amounted to 52,946 million USD per year, or 4,412 million USD per month, in 2020 (476,518 million NOK per year, or about 35 percent of the state budget (Ministry of Finance, 2020).

rate.

Table I.1: UI expenditures by data source.

(1) Data source for UI share	(2) Estimated UI share (%)	(3) Projected UI expenditures (million USD)	(4) Projected UI expenditures as share of national insurance budget (%)
No-incentive participants	7.5	584	13.2
High-incentive participants	10.4	810	18.4
Population mean	8.7	678	15.4

Notes: Table shows projected UI expenditures for different estimates of the UI share (the share of UI applicants among the adult population). Column (1) reports the data source for UI share estimate, column (2) reports the estimate, column (3) reports projected expenditures on UI benefits based on the estimate in column (2), column (4) reports projected expenditures on UI benefits as a percentage of average monthly expenditures on the Norwegian social insurance programs in 2020. UI share estimates in the first two rows are based on an NCT survey question asking whether the participant applied for UI benefits since the lockdown; this captures a period of about six weeks. The estimate of the UI share for the full population is based on administrative data on applications to UI benefits over approximately the same period. See Appendix Table A.3 for more details on the definition of these variables.

J Extrapolation with the one-dimensional participation model

In this appendix, we describe how to (partially) identify and estimate bounds on the population mean under various types of parametric and nonparametric assumptions on the one-dimensional choice model of Section 5.4. We construct bounds by applying the methodology of Mogstad et al. (2018) and Shea and Torgovitsky (2021) to survey settings. In Appendix J.1, we discuss partial identification of the population mean. In Appendix J.2, we discuss the procedure to estimate these bounds given sampled data. In Appendix J.3, we describe how we impose the considered assumptions discussed in Section 5.4. Finally, in Appendix J.4, we describe the implementation and computation of these estimated bounds.

J.1 Partial identification of population means

Our goal is to construct bounds on the population average, which can be written as a function of the MSR $m(u,x) \equiv \mathbb{E}[Y_i^\star | U_i = u, X_i = x]$ via

$$\mathbb{E}[Y_i^\star] = \mathbb{E}[m(U_i, X_i)] = \mathbb{E}\left[\int_0^1 m(u, X_i) du\right].$$

We assume that the MSR has a finite basis representation of the form

$$m(u,x) = \sum_{k=1}^{K} \theta_k b_k(u,x), \tag{J.1}$$

where $\theta \equiv [\theta_1, \ldots, \theta_K]'$ are unknown coefficients and the b_k's are known basis functions, such as constant splines or Bernstein basis polynomials. We further assume that θ belongs to a pre-specified parameter space Θ, i.e. that $\theta \in \Theta$. Together, the choices of basis representation and Θ encode the various assumptions we consider.

Given a choice of basis representation, let $B_k(p,x) = \int_0^p b_k(u,x)du$, and

$$B(p,x) = (B_1(p,x), \ldots, B_K(p,x)).$$

We can write the population mean as

$$\begin{aligned}\mathbb{E}[Y_i^\star] &= \mathbb{E}\left[\int_0^1 m(u, X_i)du\right] \\ &= \sum_{k=1}^{K} \theta_k \, \mathbb{E}\left[\int_0^1 b_k(u, X_i)du\right] \equiv \theta^{'}\tau, \end{aligned} \tag{J.2}$$

where $\tau \equiv \mathbb{E}\left[B(1, X_i)\right]$ is known given the population distribution of X_i.[51]

Let $p(x, z) \equiv \mathbb{P}[R_i = 1 | X_i = x, Z_i = z]$ and $P_i \equiv p(X_i, Z_i)$. Given observed data $(Y_i R_i, R_i, Z_i, X_i)$, where $Y_i = Y_i^\star$ if $R_i = 1$ and is NA otherwise, we observe $\mathbb{E}[Y_i | R_i = 1, X_i = x, P_i = p]$. This known value can be written as a function of θ via

$$\begin{aligned}\mathbb{E}[Y_i | R_i = 1, X_i = x, P_i = p] = \mathbb{E}[Y_i^\star | U_i \leq p, X_i = x] &= \frac{1}{p}\int_0^p m(u, x)du, \\ &= \sum_{k=1}^{K} \theta_k \int_0^p m(u, x)du \equiv \theta^{'} B(p, x).\end{aligned}$$

Given a pre-specified parameter space Θ, the set of θ that come closest to satisfying this equation (in an L_2-sense over x, p) are given by

$$\Theta^* \equiv \arg\min_{\theta \in \Theta} \mathbb{E}\left[\left(\mathbb{E}[Y_i | R_i = 1, X_i, P_i] - \theta^{'} B(P_i, X_i)\right)^2 \big| R_i = 1\right].$$

This is a least squares problem stated in the form of best linear approximation. It can be equivalently stated in terms of best linear prediction:

$$\Theta^* \equiv \arg\min_{\theta \in \Theta} Q(\theta) \quad \text{where} \quad Q(\theta) \equiv \mathbb{E}\left[(Y_i - \theta^{'} B(P_i, X_i))^2 \big| R_i = 1\right]. \tag{J.3}$$

We define the identified set for the population mean (as defined in (J.2)) as $[\tau_{lb}^*, \tau_{ub}^*]$, where

$$\tau_{lb}^* \equiv \min_{\theta \in \Theta^*} \tau^{'}\theta \quad \text{and} \quad \tau_{ub}^* \equiv \max_{\theta \in \Theta^*} \tau^{'}\theta \tag{J.4}$$

J.2 Estimation

We estimate (J.4) using the set estimator developed in Mogstad et al. (2018) and Shea and Torgovitsky (2021). Suppose we observe an i.i.d. sample $\{Y_i R_i, R_i, Z_i, X_i\}_{i=1}^n$, with $N = \sum_{i=1}^n R_i$ respondents. First, we estimate the propensity score $\hat{p}$ via a fully saturated linear regression of R_i on Z_i, X_i, and $Z_i X_i$. Let $P_i \equiv \hat{p}(X_i, Z_i)$. Consider the sample analogue of $Q(\theta)$ (defined in (J.3))

$$\hat{Q}(\theta) \equiv \frac{1}{N} \sum_{i: R_i = 1} (Y_i - \theta^{'} B_i)^2, \tag{J.5}$$

where $B_i \equiv B(P_i, X_i)$ To find the set of θ that minimize $\hat{Q}(\theta)$, we first solve for

$$\hat{Q}^* \equiv \min_{\theta \in \Theta} \hat{Q}(\Theta). \tag{J.6}$$

An estimator of $\Theta^\star$ from (J.3) is then

$$\hat{\Theta}^\star \equiv \left\{\theta \in \Theta : \hat{Q}(\theta) \leq (1 + \kappa)\hat{Q}^\star\right\}, \tag{J.7}$$

[51] We operationalize the case of no covariates by letting X_i be a degenerate (and trivially known) distribution.

where $\kappa \geq 0$ is a small number used to improve numerical stability. We construct estimated bounds on the population mean, denoted $[\hat{\tau}^*_{lb}, \hat{\tau}^*_{ub}]$ by solving the sample analogues of (J.4), i.e.

$$\hat{\tau}^*_{lb} \equiv \min_{\theta \in \hat{\Theta}^\star} \hat{\tau}^{'} \theta \quad \text{and} \quad \hat{\tau}^*_{ub} \equiv \max_{\theta \in \hat{\Theta}^\star} \hat{\tau}^{'} \theta. \tag{J.8}$$

J.3 MSR assumptions: basis representations and parameter spaces

Together, the choice basis representation (as in (J.1)) and choice of parameter space Θ encode the various sets of assumptions we consider in Section 5.4. Depending on the considered assumptions, we either take the basis representation to be constant splines with knots on propensity scores or Bernstein basis polynomials. Mogstad et al. (2018) show that the constant spline basis exactly reproduces the nonparametric bounds.

Table J.1 presents the pair of basis representation (first column) and Θ (second and third columns) for each set of assumptions considered in Section 5.4. For example, we can impose the assumption that the MSR (without covariates) is monotone increasing by assuming the MSR has a judiciously chosen constant spline basis representation with coefficients that are weakly positive. See Table J.1 for specific details on implementation.

Table J.1: Assumptions on the MSR

	Basis representation	Parameter space restrictions
IV	$m_U(u) = \sum_{j=1}^{J} \theta_j \mathbb{1}[u \leq \zeta_j]$	Bounded: $a \leq \sum_{n=1}^{j} \theta_n \leq b$ for all j
Monotone MSR	$m_U(u) = \sum_{j=1}^{J} \theta_j \mathbb{1}[u \leq \zeta_j]$	Bounded: $a \leq \sum_{n=1}^{j} \theta_n \leq b$ for all j Inc. (Dec.): $\theta_j \geq 0 (\leq 0)$ for all j
Separable MSR	$m_U(u) = \sum_{j=1}^{J} \theta_j \mathbb{1}[u \leq \zeta_j] + \theta_{J+1} x$	Bounded: $a \leq \sum_{n=1}^{j} \theta_n + \theta_{J+1} x \leq b$ for all $j \leq J$ and $x \in X$
Separable + Monotone MSR	$m_U(u) = \sum_{j=1}^{J} \theta_j \mathbb{1}[u \leq \zeta_j] + \theta_{J+1} x$	Bounded: $a \leq \sum_{n=1}^{j} \theta_n + \theta_{J+1} x \leq b$ for all $j \leq J$ and $x \in X$ Inc. (Dec.): $\theta_j \geq 0 (\leq 0)$ for all $j \leq J$
Linear $m(u)$	$m_U(u) = \theta_0(1-u) + \theta_1 u$	Bounded: $a \leq \theta_0, \theta_1 \leq b$

Notes: This table presents the parameterization of the MSR we use for the results in the main body. Each row presents an assumption. The first column presents the parameterization used. The second column describes how boundedness is implemented. The third column presents how the assumption that the MSR is monotonic in u is estimated. We use gender as X_i. For $\zeta_j, j = 1, \ldots, J-1$, we use the $j-$th ordered estimated propensity score $\mathbb{P}[R_i = 1 | Z_i = z]$ for all realizations of z in specifications without X_i and the $j-$th ordered estimated propensity score $\mathbb{P}[R_i = 1 | Z_i = z, X_i = x]$ for all realizations of (z, x) in specifications with X_i. Letting $\mathcal{Z}$ and $\mathcal{X}$ respectively denote the (discrete) supports of Z_i and X_i and setting $\zeta_J = 1$, we have $J = |\mathcal{Z}| + 1$ for specifications without X_i and $J = |\mathcal{Z}| \times |\mathcal{X}| + 1$ for specifications with X_i.

J.4 Computation

Three optimization problems need to be solved to construct estimated bounds on the population mean: the program in (J.6) and the two in (J.8). Observing that all considered restrictions on Θ can be imposed via linear constraints on θ (see Table J.1) and that (J.5) is a quadratic function of θ, (J.6) is a quadratic program (QP). Then, since $\hat{Q}(\theta)$ is a quadratic function of θ and enters as a constraint in the programs of (J.8), these two programs are quadratically-constrained quadratic programs (QCQPs). All three programs can be solved to provable global optimality using spatial branch-and-bound algorithms (we use Gurobi Optimization (2021)). Finally, we let $\kappa > 0$ for all outcomes to avoid numerical issues during estimation and set it as $\kappa = \kappa_y \sqrt{10^{-7}}$, where $\kappa_y = (0.1, 2, 1.6, 2, 12, 0.2)$ respectively for earnings before lockdown, earnings after lockdown, earnings loss larger than 20%, employed before lockdown, employed after lockdown and employment loss.

These values were chosen so as to ensure numerical stability when solving these QCQPs with Gurobi.

K Identification and extrapolation with the two-dimensional participation model

In Appendix K.1, we discuss extrapolation using the two-dimensional model. In Appendix K.2, we prove that our test of independence between unobserved margins in Section 6 is the strongest testable implication of independence. In Appendix K.3, we show that independence between unobserved margins and knowledge of the share of individuals who never see the invitation to participate point identify group shares. Finally, in Appendix K.4, we show that our empirical findings are robust to alternative choices of the share of individuals who never see the invitation to participate.

K.1 Extrapolating with the two-dimensional participation model

We first discuss partial identification of population means. We then discuss estimation of these bounds given a finite sample. We conclude by describing the specific assumptions we impose in the main body, and discuss computation of the estimated bounds.

K.1.1 Partial identification of population means

As in Section 6, for each $z \in \{0,1,2\}$ and $s \in \{1,2,3\}$, let $\mu_{zs} = \mathbb{E}[Y_i^\star | V_i \in \mathcal{V}_z, S_i = s]$ and $\pi_{zs} = \mathbb{P}[V_i \in \mathcal{V}_z, S_i = s]$. Letting $\eta(-1) = -\infty$ and $\eta(2) = \infty$, note that $\mathcal{V}_z = [\eta(z-1), \eta(z)]$. Ordering (z,s) lexicographically, let μ be the vector that collects $\{\mu_{zs}\}$ and same for π with $\{\pi_{zs}\}$. Our goal is to construct bounds on the population average, which can be written as a function of (μ, π) via

$$\mathbb{E}[Y_i^\star] = \sum_{z,s} \mu_{zs}\pi_{zs}. \tag{K.1}$$

We assume that μ and π respectively belong to pre-specified spaces $\mathcal{M}$ and Π. These spaces encode the various assumptions we consider. Let $T_i = R_{i1} + 2(1 - R_{i1})R_{i2}$ denote the period (1 or 2) in which the individual participated, if they participated, with $T_i = 0$ if they did not participate. We observe the distribution $(Y_i \mathbb{1}[T_i \in \{1,2\}], T_i, Z_i)$, where $Y_i = Y_i^\star$ if $T_i \in \{1,2\}$, and is NA otherwise.

To be consistent with the observed data, a candidate value $(\mu, \pi) \in \mathcal{M} \times \Pi$ must satisfy two types of equality constraints. First, any such (μ, π) must satisfy

$$\mathbb{P}[T_i = t | Z_i = z] = \mathbb{P}[V_i \leq \eta(z), S_i = t] = \sum_{j \leq z} \pi_{jt} = \sum_{j,s} \pi_{js} \underbrace{\mathbb{1}[j \leq z, s = t]}_{\equiv D_{j,s}(z,t)} = \pi' D(z,t), \tag{K.2}$$

for $(z,t) \in \{0,1\}^2$ and where $D(z,t)$ is a known, vector-valued function. Second, (μ, π) must also satisfy

$$\begin{aligned}
&\mathbb{E}[Y_i | T_i = t, Z_i = z] \\
&\quad = \mathbb{E}[Y_i^\star | V_i \leq \eta(z), S_i = t] \\
&\quad = \sum_{j \leq z} \mathbb{E}[Y_i^\star | \eta(j-1) \leq V_i \leq \eta(j), S_i = t] \, \mathbb{P}[\eta(j-1) \leq V_i \leq \eta(j) | V_i \leq \eta(z), S_i = t] \\
&\quad = \sum_{j,s} \mu_{js} \underbrace{\mathbb{1}[j \leq z, s = t] \frac{\mathbb{P}[\eta(j-1) \leq V_i \leq \eta(j), S_i = s]}{\mathbb{P}[V_i \leq \eta(z), S_i = t]}}_{\equiv B_{j,s}(z,t)} = \mu' B(z,t)
\end{aligned} \tag{K.3}$$

for $(z,t) \in \{0,1\}^2$. For components $j \leq z$ and $s = t$, $B_{j,s}(z,t)$ is point identified via the probabilities point identified in (K.2); for all other values of j and s, it is 0.

We construct an identified set for the population mean through a three step procedure. Let $\mathcal{A} \equiv \{(z,s) : z \in \{0,1\}, s \in \{1,2\}\}$ be the set of (z,s) for which probabilities and moments are identified given the data. In the first step, we find $\pi \in \Pi$ that comes closest to satisfying (K.2) for all $(z,s) \in \mathcal{A}$ (in an L_2-sense) by solving

$$\arg\min_{\pi \in \Pi} \mathbb{E}\left[\left(\mathbb{1}[T_i = 1] - \pi' D(Z_i, 1)\right)^2 + \left(\mathbb{1}[T_i = 2] - \pi' D(Z_i, 2)\right)^2\right]. \tag{K.4}$$

Let $\pi^\star$ be a minimizer of (K.4). Define $\pi^\star_{\mathcal{A}} \equiv \{\pi^\star_{zs} : (z,s) \in \mathcal{A}\}$. Let $B^\star_{j,s}(z,t) \equiv \mathbb{1}[j \leq z, s = t] \frac{\pi^\star_{js}}{\sum_{j' \leq z} \pi^\star_{j't}}$, and define $B^\star(z,t)$ as the vector that collects these values, so that $B^\star(z,t)$ is equal to $B(z,t)$ in (K.3) with the probabilities replaced with those given by $\pi^\star_{\mathcal{A}}$.

In the second step, we find the set $\mu \in \mathcal{M}$ that comes closest to satisfying (K.3) for all $(z,s) \in \mathcal{A}$ by solving

$$\arg\min_{\mu \in \mathcal{M}} \mathbb{E}\left[\left(Y_i - \mu' B^\star\left(Z_i, T_i\right)\right)^2 \middle| T_i \in \{1,2\}\right]. \tag{K.5}$$

Let $\mu^\star$ denote a minimizer of (K.5). Define $\mu^\star_{\mathcal{A}} \equiv \{\mu^\star_{zs} : (z,s) \in \mathcal{A}\}$.

In the third step, we define the identified set for the population mean (defined in (K.1)) as $[\tau^\star_{lb}, \tau^\star_{ub}]$, where

$$\tau^\star_{lb} \equiv \min_{\substack{(\mu,\pi) \in \mathcal{M} \times \Pi, \\ \mu_{\mathcal{A}} = \mu^\star_{\mathcal{A}}, \pi_{\mathcal{A}} = \pi^\star_{\mathcal{A}}}} \sum_{z,s} \mu_{zs} \pi_{zs} \quad \text{and} \quad \tau^\star_{ub} \equiv \max_{\substack{(\mu,\pi) \in \mathcal{M} \times \Pi, \\ \mu_{\mathcal{A}} = \mu^\star_{\mathcal{A}}, \pi_{\mathcal{A}} = \pi^\star_{\mathcal{A}}}} \sum_{z,s} \mu_{zs} \pi_{zs}. \tag{K.6}$$

K.1.2 Estimation of bounds for population means

Given an i.i.d. sample $\left\{Y_i \mathbb{1}[T_i \in \{1,2\}], T_i, Z_i\right\}_{i=1}^n$, we estimate (K.4), (K.5), and (K.6) by taking the sample analogues. We estimate (K.4) by jointly stacking observations $(\mathbb{1}[T_i = 1], Z_i)$ and $(\mathbb{1}[T_i = 2], Z_i)$ as in pooled panel data regressions, and thus solve

$$\arg\min_{\pi \in \Pi} \frac{1}{n} \sum_i \left[\left(\mathbb{1}[T_i = 2] - \pi' D(Z_i, 2)\right)^2 + \left(\left(\mathbb{1}[T_i = 2] - \pi' D(Z_i, 2)\right)^2\right]. \tag{K.7}$$

Let $\hat{\pi}^\star$ be a minimizer of (K.7). We keep $\hat{\pi}^\star_{\mathcal{A}} \equiv \{\hat{\pi}^\star_{zs} : (z,s) \in \mathcal{A}\}$.

Next, we estimate $B^\star(z,t)$ with $\hat{B}^\star(z,t)$, where the shares correspond to $\hat{\pi}^\star_{\mathcal{A}}$. Letting $N \equiv \sum_i \mathbb{1}[T_i \in \{1,2\}]$, the estimated analogue of (K.5) is

$$\arg\min_{\mu \in \mathcal{M}} \frac{1}{N} \sum_{i : T_i \in \{1,2\}} \left[\left(Y_i - \mu' \hat{B}^\star(Z_i, T_i)\right)^2\right]. \tag{K.8}$$

Let $\hat{\mu}^\star$ be a minimizer of (K.8). We keep $\hat{\mu}^\star_{\mathcal{A}} \equiv \{\hat{\mu}^\star_{zs} : (z,s) \in \mathcal{A}\}$.

Then, estimated bounds on the population mean, $[\hat{\tau}^\star_{lb}, \hat{\tau}^\star_{ub}]$, can be constructed via the estimated analogue of (K.6), where

$$\hat{\tau}^\star_{lb} \equiv \min_{\substack{(\mu,\pi) \in \mathcal{M} \times \Pi, \\ \mu_{\mathcal{A}} = \hat{\mu}^\star_{\mathcal{A}}, \pi_{\mathcal{A}} = \hat{\pi}^\star_{\mathcal{A}}}} \sum_{j,s} \mu_{js} \pi_{js} \quad \text{and} \quad \hat{\tau}^\star_{ub} \equiv \max_{\substack{(\mu,\pi) \in \mathcal{M} \times \Pi, \\ \mu_{\mathcal{A}} = \hat{\mu}^\star_{\mathcal{A}}, \pi_{\mathcal{A}} = \hat{\pi}^\star_{\mathcal{A}}}} \sum_{j,s} \mu_{js} \pi_{js}. \tag{K.9}$$

K.1.3 Assumptions on shares, group responses, and MSR

We consider three sets of assumptions when extrapolating with the two-dimensional participation model. The first set of assumptions we consider are on group shares. These restrictions are

imposed via Π and are depicted in Table K.1. The second set of assumptions we consider are on group responses. These restrictions are imposed via $\mathcal{M}$ and are depicted in Table K.2.

Table K.1: Assumptions on shares

	Parameter space restrictions (on Π)
Valid distribution	$\pi_{zs} \in [0,1]\ \forall (z,s),\ \sum_{z,s} \pi_{zs} = 1$
Passive share equals α	$\sum_z \pi_{z3} = \alpha$
Independence	$\pi_{zs} = \pi_z \pi_s\ \forall (z,s)$ with $\pi_z \in [0,1], \pi_s \in [0,1]\ \forall (z,s)$

Notes: This table presents restrictions we consider on Π, the set of shares, where we recall $\pi_{zs} \equiv \mathbb{P}[\eta(z-1) \leq V_i \leq \eta(z), S_i = s]$. In the two period and binary incentive setting, $S_i \in \{1,2,3\}$ and $Z_i \in \{0,1\}$, where $\eta(-1) = -\infty, \eta(2) = \infty$, which fixes the dimension of Π.

Table K.2: Assumptions on group responses

	Parameter space restrictions (on $\mathcal{M}$)
Bounded grp. resp. (within $[a,b]$)	$a \leq \mu_{zs} \leq b\ \forall (z,s)$
Separable grp. resp.	$\mu_{zs} = \mu_z + \mu_s\ \forall (z,s)$
Monotone grp. resp. (incentive)	increasing: $z > z' \implies \mu_{zs} \geq \mu_{z's}$ ($\leq$ for dec.)
Monotone grp. resp. (reminder)	increasing: $s > s' \implies \mu_{zs} \geq \mu_{zs'}$ ($\leq$ for dec.)

Notes: This table presents restrictions we consider on $\mathcal{M}$, the set of group responses, where we recall $\mu_{zs} \equiv \mathbb{E}[Y_i^\star | \eta(z-1) \leq V_i \leq \eta(z), S_i = s]$. In the two period and binary incentive setting, $S_i \in \{1,2,3\}$ and $Z_i \in \{0,1\}$, where $\eta(-1) = -\infty, \eta(2) = \infty$, which fixes the dimension of $\mathcal{M}$.

The last set of assumptions we consider are on the MSR $m(v,s) \equiv \mathbb{E}[Y_i^\star | V_i = v, S_i = s]$. Since group responses also depend on the (unobserved) distribution of latent variables (V_i, S_i), we only consider assumptions on the MSR after assuming the passive share restriction and that V_i and S_i are independent (and that we have a valid distribution). Under these assumptions, all group shares $\{\pi_{zs}\}$ are point identified (see Appendix K.3 for proof). Thus, when identifying group shares via (K.4), we keep the full vector $\pi^\star$, with the estimation analogue holding for $\hat{\pi}^\star$ via (K.7).

Under these share assumptions and normalizing V_i to be uniform, group responses can be expressed as

$$\mu_{zs} = \frac{1}{\pi_{z1} + \pi_{z2} + \pi_{z3}} \int_{\eta(z-1)}^{\eta(z)} m(v,s) dv, \tag{K.10}$$

where $\eta(z) = \sum_{(z',s): z' \leq z} \pi_{z's}$. Similar to Appendix J, supposing that $m(v,s) = \sum_k \theta_k b_k(v,s)$ where b_k are known basis functions and θ belongs to some pre-specified parameter space Θ, (K.10) can be written as

$$\mu_{zs}(\theta) = \sum_k \theta_k \underbrace{\frac{1}{\pi_{z1} + \pi_{z2} + \pi_{z3}} \int_{\eta(z-1)}^{\eta(z)} b_k(v,s) dv}_{\equiv \tilde{B}_k(z,s)} \equiv \theta \tilde{B}(z,s). \tag{K.11}$$

Letting $\hat{\tilde{B}}^\star(z,s)$ denote the values of $\tilde{B}(z,s)$ given $\hat{\pi}^\star$, the analogue of (K.8) becomes

$$\arg\min_{\theta \in \Theta} \mathbb{E}\left[\left(Y_i - \theta' \hat{\tilde{B}}^\star(Z_i, T_i) \right)^2 \middle| T_i \in \{1,2\} \right]. \tag{K.12}$$

If $\hat{\theta}^\star$ is a minimizer of (K.12), define the set $\hat{\mu}_{\mathcal{A}}^\star \equiv \{\mu_{zs}(\hat{\theta}^\star) : (z,s) \in \mathcal{A}\}$. Given $(\hat{\pi}^\star, \hat{\mu}_{\mathcal{A}}^\star)$, estimated bounds for the identified set for the population mean are

$$\hat{\tau}_{lb}^\star \equiv \min_{\substack{\theta \in \Theta, \\ \mu_{\mathcal{A}} = \hat{\mu}_{\mathcal{A}}^\star}} \sum_{z,s} \mu_{zs} \hat{\pi}_{zs}^\star \quad \text{and} \quad \hat{\tau}_{ub}^\star \equiv \max_{\substack{\theta \in \Theta, \\ \mu_{\mathcal{A}} = \hat{\mu}_{\mathcal{A}}^\star}} \sum_{z,s} \mu_{zs} \hat{\pi}_{zs}^\star. \tag{K.13}$$

The considered restrictions on the MSR are imposed via choice of basis function and choice of parameter space Θ. They are depicted in Table K.3.

Table K.3: Assumptions on MSR

	Basis representation	Parameter space restrictions
Separable + monotone MSR	$m(v,s) = m_V(v) + m_S(s)$ $= \sum_{z=1}^{3} \theta_z \mathbb{1}[v \leq \eta(z)] +$ $\sum_{s'=1}^{3} \alpha_{s'} \mathbb{1}[s' = s]$	Bounded: $a \leq \sum_{z=1}^{j} \theta_z \mathbb{1}[v \leq \eta(z)] +$ $\sum_{s'=1}^{\ell} \alpha_{s'} \mathbb{1}[s' = s] \leq b$ for all (j, ℓ) Inc. (Dec.): $\theta_z \geq 0 (\leq 0)$ for all z, $\alpha_s \geq 0 (\leq 0)$ for all s
Separable MSR, linear $m(v)$, monotone $m(s)$	$m(v,s) = m_V(v) + m_S(s)$ $= \theta_0 v + \sum_{s'=1}^{3} \theta_{s'} \mathbb{1}[s' = s]$	Bounded: $a \leq \theta_0 v +$ $\sum_{s'=1}^{\ell} \alpha_{s'} \mathbb{1}[s' = s] \leq b$ for all ℓ Inc. (Dec.): $\theta_s \geq 0 (\leq 0)$ for all $s \in \{1, 2\}$

Notes: This table presents restrictions we consider on Θ. The dimension of Θ is given by the considered basis representation.

K.1.4 Implementation and computation

In our extrapolation results in Section 6.3. we consider four sets of assumptions. In the first set (IV), we restrict Π to include valid distributions and restrict $\mathcal{M}$ to be bounded group responses. In the second set (separable + monotone group responses, (V,S) unknown), we restrict Π to include valid distributions and restrict $\mathcal{M}$ to be group responses that are bounded, separable, and monotone in both margins (as determined by the data). For the third and fourth set of assumptions, we restrict Π to include shares that define valid distributions, satisfy the passive share assumption (set to .4), and satisfy the independence assumption. Under these assumptions, we impose assumptions via the MSR. In the third set, we assume that the MSR is separable and monotone in both margins (as determined by the data). The implementation is presented in the first row of Table K.3. In the fourth set, we assume that the MSR is separable, that $m(v)$ is linear, and that $m(s)$ is monotone. The implementation is presented in the second row of Table K.3.

For each set of assumptions, four optimization problems need to be solved to construct estimated bounds on the population mean: the program in (K.7), the program in (K.8) ((K.12) for the third and fourth sets of assumptions), and the two in (K.9) ((K.13) for the third and fourth sets of assumptions). All considered assumptions can be formulated as quadratic constraints, and the four programs are thus, in general, (nonconvex) quadratically-constrained quadratic programs (QCQPs). These can be solved to provable global optimality using spatial branch-and-bound algorithms (we use Gurobi Optimization (2021)).

K.2 Testing independence between unobserved margins

Let $p_{zt} \equiv \mathbb{P}[T_i = t | Z_i = z]$. Then let $\mathcal{P} \equiv \{p_{zt}\}_{(z,t) \in \{0,1\} \times \{1,2\}}$.

The two-dimensional selection model implies that for any $p_{zt} \in \mathcal{P}$,

$$p_{zt} = \mathbb{P}[V_i \leq \eta(z), S_i = t].$$

Independence of V_i and S_i implies that

$$p_{zt} = \mathbb{P}[V_i \leq \eta(z)] \, \mathbb{P}[S_i = t],$$

which immediately implies that

$$\frac{p_{01}}{p_{11}} = \frac{\mathbb{P}[V_i \leq \eta(0)]}{\mathbb{P}[V_i \leq \eta(1)]} = \frac{p_{02}}{p_{12}}.$$

We accordingly test for independence by testing whether

$$\frac{p_{01}}{p_{11}} - \frac{p_{02}}{p_{12}} = 0. \tag{K.14}$$

The estimated analogue of the LHS of (K.14) is -0.006, and we fail to reject the null that this value is zero (p-value = .97).

We next prove that (K.14) is the strongest testable implication of the independence assumption given data $\mathcal{P}$ and the two-dimensional selection model. We do this by showing that whenever (K.14) holds, there exists a distribution of latent unobservables $(\tilde{V}_i, \tilde{S}_i)$ such that

$$\mathbb{P}[\tilde{V}_i \leq \eta(z), \tilde{S}_i = t] = p_{zt} \tag{K.15}$$

for $(z,t) \in \{0,1\} \times \{1,2\}$ and such that $\tilde{V}_i$ is independent of $\tilde{S}_i$.

Let $\tilde{V}_i$ be a random variable such that $\mathbb{P}[\tilde{V}_i \leq \eta(0)] = p_{01}+p_{02}$ and $\mathbb{P}[\tilde{V}_i \leq \eta(1)] = \frac{(p_{01}+p_{02})p_{11}}{p_{01}}$. This defines a valid distribution, as $\mathbb{P}[\tilde{V}_i \leq \eta(0)] \in [0,1]$ and

$$\begin{aligned}
\mathbb{P}[\tilde{V}_i \leq \eta(1)] &= \frac{(p_{01}+p_{02})p_{11}}{p_{01}} && \text{(K.16)}\\
&= p_{11} + p_{02}\frac{p_{11}}{p_{01}} && \text{(K.17)}\\
&= p_{11} + p_{12} \in [0,1] && \text{(K.18)}
\end{aligned}$$

where the last line used that (K.14) implies $\frac{p_{11}}{p_{01}} = \frac{p_{12}}{p_{02}}$. Separately, let $\tilde{S}_i$ be a categorical random variable with support $\{1,2,3\}$ with $\mathbb{P}[\tilde{S}_i = 1] = \frac{p_{01}}{p_{01}+p_{02}}$ and $\mathbb{P}[\tilde{S}_i = 2] = \frac{p_{02}}{p_{01}+p_{02}}$. Clearly $\mathbb{P}[\tilde{S}_i = t] \in [0,1]$ for $t = 1,2$, and

$$\mathbb{P}[\tilde{S}_i = 3] = 1 - \frac{p_{01}}{p_{01}+p_{02}} - \frac{p_{02}}{p_{01}+p_{02}} = 0 \in [0,1]. \tag{K.19}$$

Clearly, $\tilde{V}_i$ can be constructed to be independent of $\tilde{S}_i$, so that

$$\mathbb{P}[\tilde{V}_i \leq \eta(z), \tilde{S}_i = t] = \mathbb{P}[\tilde{V}_i \leq \eta(z)]\,\mathbb{P}[\tilde{S}_i = t].$$

Letting $\tilde{p}_{zt} \equiv \mathbb{P}[\tilde{V}_i \leq \eta(z), \tilde{S}_i = t]$, it suffices to verify $\tilde{p}_{zt} = p_{zt}$. This holds trivially for $\tilde{p}_{01}, \tilde{p}_{11}, \tilde{p}_{02}$, so we only check $\tilde{p}_{12}$. Observe

$$\begin{aligned}
\tilde{p}_{12} &= \mathbb{P}[\tilde{V}_i \leq \eta(1)]\,\mathbb{P}[\tilde{S}_i = 2] && \text{(K.20)}\\
&= \left(\frac{(p_{01}+p_{02})\,p_{11}}{p_{01}}\right)\left(\frac{p_{02}}{p_{01}+p_{02}}\right) && \text{(K.21)}\\
&= \frac{p_{11}p_{02}}{p_{01}} && \text{(K.22)}\\
&= \frac{p_{12}p_{02}}{p_{02}} = p_{12}, && \text{(K.23)}
\end{aligned}$$

where the last line again used the fact that (K.14) implies $\frac{p_{11}}{p_{01}} = \frac{p_{12}}{p_{02}}$. This concludes the proof.

K.3 Point identification of group shares under independence and passive share restriction

Independence of V_i and S_i implies that

$$\mathbb{P}[T_i = t | Z_i = z] = \mathbb{P}[V_i \leq \eta(z)]\,\mathbb{P}[S_i = t] = \eta(z)\,\mathbb{P}[S_i = t] \tag{K.24}$$

for both $z = 0, 1$ and $t = 1, 2$, where the usual normalization on the distribution of V_i was used in the second equality. Taking the ratio of these equations—say with $z = 0$ for concreteness—identifies the ratio of $\mathbb{P}[S_i = 2]$ to $\mathbb{P}[S_i = 1]$ as

$$\frac{\mathbb{P}[S_i = 2]}{\mathbb{P}[S_i = 1]} = \frac{\mathbb{P}[T_i = 2|Z_i = 0]}{\mathbb{P}[T_i = 1|Z_i = 0]}. \tag{K.25}$$

Then, using our assumption that $\mathbb{P}[S_i = 3]$ is known (and equal to .4), we can separate out each group share by solving

$$\mathbb{P}[S_i = 1] = 1 - \mathbb{P}[S_i = 2] - \mathbb{P}[S_i = 3] = 1 - \mathbb{P}[S_i = 1]\frac{\mathbb{P}[T_i = 2|Z_i = 0]}{\mathbb{P}[T_i = 1|Z_i = 0]} - \mathbb{P}[S_i = 3] \tag{K.26}$$

and obtaining

$$\mathbb{P}[S_i = 1] = \frac{(1 - \mathbb{P}[S_i = 3])\,\mathbb{P}[T_i = 1|Z_i = 0]}{\mathbb{P}[T_i = 1|Z_i = 0] + \mathbb{P}[T_i = 2|Z_i = 0]}. \tag{K.27}$$

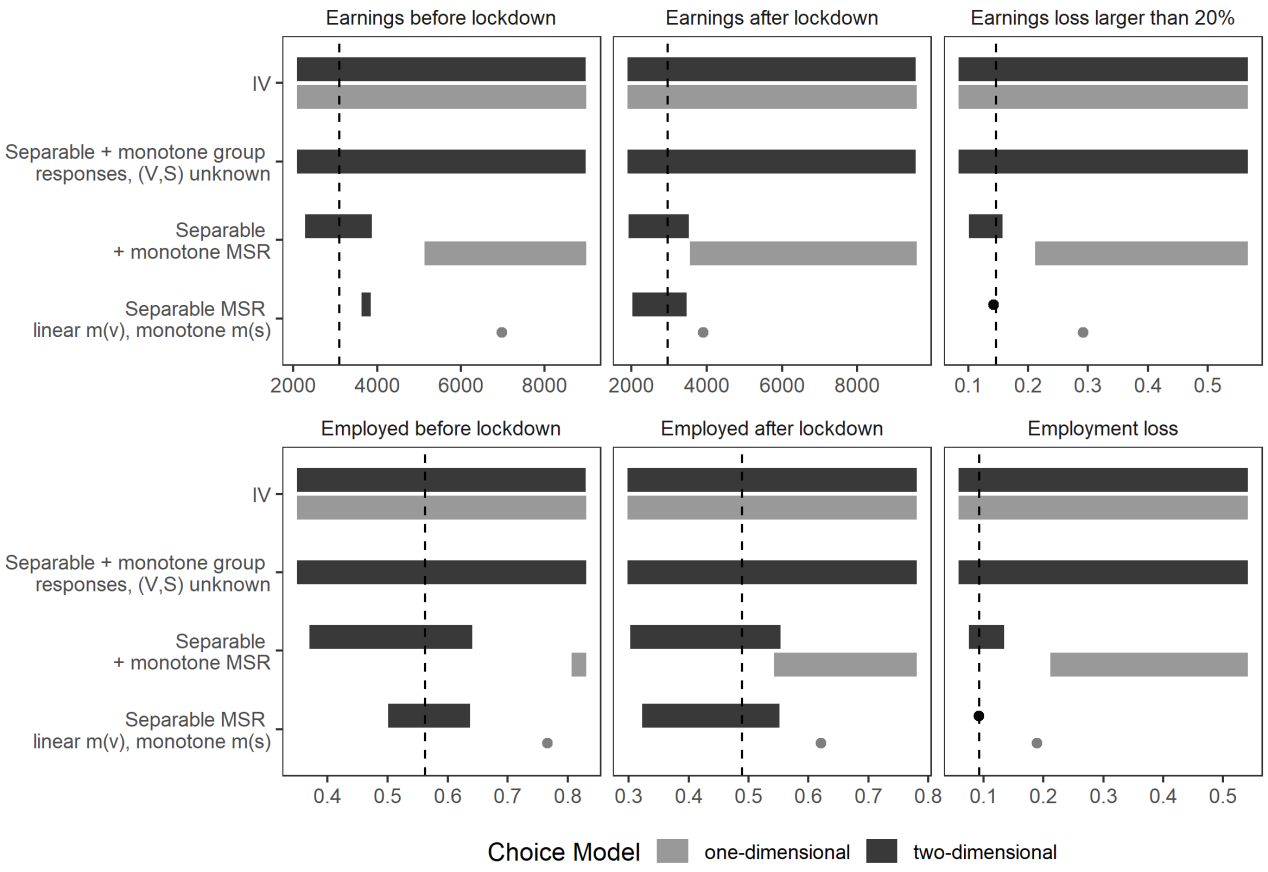
We then identify $\mathbb{P}[S_i = 2]$ by substituting (K.27) into (K.25), and then $\eta(z)$ from (K.24) with $z = 0, 1$.

K.4 Sensitivity of extrapolation estimates to choice of passive share

In the third and fourth rows of Figure 14 of Section 6, we assume that 40% never see the invitation to participate (i.e. we assume the passive share is 40%). This number was chosen in consultation with survey researchers at Statistics Norway. Figure K.1 presents results if we assumed this share was increased by 8 points to 48% (the largest is can be given the response rate of 52% in the high incentive group), and Figure K.2 presents results if we instead decrease the passive share by 8 points to 32%. In both cases, the resulting bounds are similar to those in Figure 14, and the two-dimensional model extrapolation continue to outperform the one-dimensional model extrapolation.

To further examine the robustness of our results to the choice of passive share, we compute estimated bounds under the final specification considered in Section 6.3 (fourth row of Figure 14) for all passive share choices in the range of $10\% - 48\%$ for all six outcomes. For each outcome, the second column (titled 'two-dimension') of Table K.4 presents the maximum difference between the estimated bounds and the true population mean over the considered range of passive share choices. As comparison, the third row (titled 'one-dimension') presents this difference under the one-dimensional model analogue of the final specification (see table notes of Figure 14 for more details). Even when allowing the passive share to vary from 48% to 10%, the maximum difference between the bounds and true population mean is lower under the two-dimensional model relative to the one-dimensional model for all six outcomes.

Figure K.1: Bounds under double threshold model assumptions (passive share = 48%)



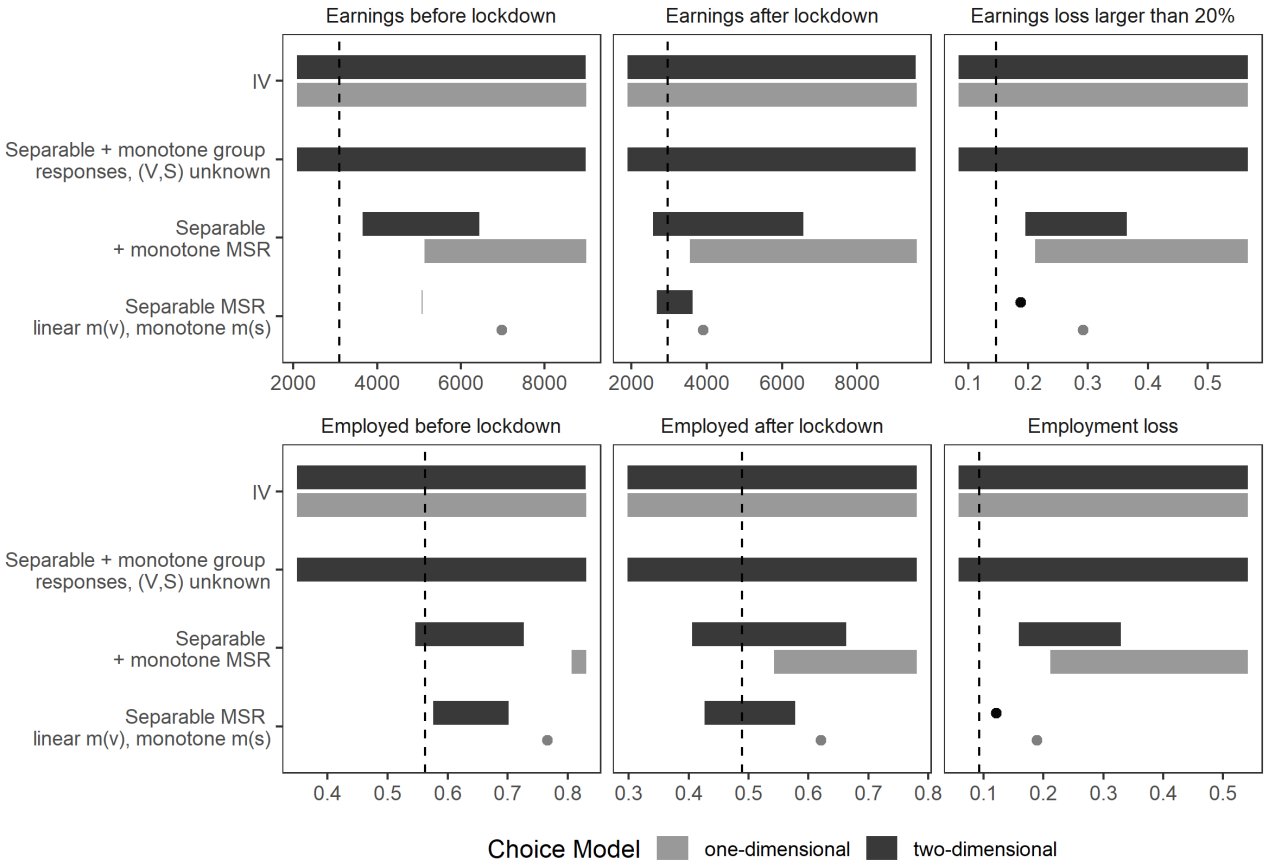
Notes: These figures report estimates of population means using both the one-dimensional (light gray) and two-dimensional (dark gray) models as in Figure 14 except that for the third and fourth rows, we assume that 48% never see the invitation under the two-dimensional model. Bars denote estimated bounds and points denote point estimates. All estimates use data from "no" and "high" incentive samples. The actual population mean is presented as a vertical dashed line. See figure notes of Figure 14 for more details.

Table K.4: Absolute difference under final extrapolation specification

	Maximum absolute difference	
	two-dimension	one-dimension
Earnings before lockdown	3194	3873
Employed before lockdown	0.13	0.20
Earnings after lockdown	526	926
Employed after lockdown	0.07	0.13
Earnings loss	0.10	0.14
Employment loss	0.07	0.10

Notes: This table presents the maximum absolute difference between estimated bounds under the final specification considered in Figure 14 (fourth row) and the true population mean for each outcome. In the two-dimensional model, given a passive share of $\alpha \in [0.10, 0.48]$, we estimate population mean bounds for an outcome under the final specification, which we denote as $[b_l(\alpha), b_u(\alpha)]$. If the true population mean is m, the absolute difference is then defined as $f(\alpha) = \min\{0, |b_u(\alpha) - m|, |b_l(\alpha) - m|\}$. The maximum absolute difference for each outcome is defined as the maximum absolute difference over all passive share choices, i.e. maximum absolute difference $\equiv \sup_{\alpha \in [0.10, 0.48]} f(\alpha)$. These values are presented in the second column (titled "two-dimension"). For comparison, the third column (titled "one-dimension") presents the absolute difference between the final specification under the one-dimensional model. Since this value is a point estimate and since there is no notion of passive share in the one-dimensional model, the maximum absolute difference is simply the absolute difference between the point estimate and the population mean.

Figure K.2: Bounds under double threshold model assumptions (passive share = 32%)



Notes: These figures report estimates of population means using both the one-dimensional (light gray) and two-dimensional (dark gray) models as in Figure 14 except that for the third and fourth rows, we assume that 32% never see the invitation under the two-dimensional model. Bars denote estimated bounds and points denote point estimates. All estimates use data from "no" and "high" incentive samples. The actual population mean is presented as a vertical dashed line. See figure notes of Figure 14 for more details.

Appendix References

American Association for Public Opinion Research (2016). Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys. Technical report, American Association for Public Opinion Research. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf.

American Economic Association (1991). Classification system: Old and new categories. *Journal of Economic Literature 29*(1), xviii–xxviii.

American Economic Association (2021). JEL Classification Codes Guide. https://www.aeaweb.org/jel/guide/jel.php.

Buskirk, T. D. (2018). Surveying the Forests and Sampling the Trees: An Overview of Classification and Regression Trees and Random Forests with Applications in Survey Research. *Survey Practice 11*(1), 1–13.

Buskirk, T. D., A. Kirchner, A. Eck, and C. S. Signorino (2018). An Introduction to Machine Learning Methods for Survey Researchers. *Survey Practice 11*(1), 1–10.

Card, D. and S. DellaVigna (2013). Nine Facts about Top Journals in Economics. *Journal of Economic Literature 51*(1), 144–61.

Carvalho, L. S., S. Meier, and S. W. Wang (2016). Poverty and Economic Decision-making: Evidence From Changes in Financial Resources at Payday. *American Economic Review 106*(2), 260–84.

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

Coffman, L. C., J. J. Conlon, C. R. Featherstone, and J. B. Kessler (2019b). Replication Data for: 'Liquidity Affects Job Choice: Evidence from Teach For America'.

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.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Eck, A. (2018). Neural Networks for Survey Researchers. *Survey Practice 11*(1), 1–11.

Ehling, P., A. Graniero, and C. Heyerdahl-Larsen (2018). Asset Prices and Portfolio Choice with Learning from Experience. *The Review of Economic Studies 85*(3), 1752–1780.

Elias, J. J., N. Lacetera, and M. Macis (2019). Paying for kidneys? A Randomized Survey and Choice Experiment. *American Economic Review 109*(8), 2855–88.

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

Franco, A., N. Malhotra, G. Simonovits, and L. Zigerell (2017). Developing Standards for Post-Hoc Weighting in Population-Based Survey Experiments. *Journal of Experimental Political Science 4*, 161–172.

Groves, R. M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly 70*(1), 646–675.

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

Health and Retirement Study (2017). Sample size and response rates. Technical report. https://hrs.isr.umich.edu/sites/default/files/biblio/ResponseRates_2017.pdf.

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.

Institute for Social Research, University of Michigan (2021). PSID Main Interview User Manual: Release 2021. Technical report, Institute for Social Research, University of Michigan. https:

//psidonline.isr.umich.edu/data/Documentation/UserGuide2019.pdf.

IPUMS (2021). NHIS Sample Design. Technical report. https://nhis.ipums.org/nhis/userNotes_sampledesign.shtml.

Kirchner, A. and C. S. Signorino (2018). Using Support Vector Machines for Survey Research. *Survey Practice 11*(1), 1–14.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, R. C. Team, et al. (2020). Package 'caret'. *The R Journal*, 223.

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods*. Sage Publications.

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies 76*(3), 1071–1102.

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

Lumley, T. (2020). Package 'survey'. *CRAN R*.

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/prp201920200001guldddpdfs.pdf.

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

National Bureau of Economic Research (2020a). Current Population Survey (CPS) Data Supplements at the NBER. http://data.nber.org/data/current-population-survey-data.html.

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

National Longitudinal Surveys (2020). NLSY79 Sample Design & Screening Process. Technical report. https://www.nlsinfo.org/content/cohorts/nlsy79/intro-to-the-sample/sample-design-screening-process.

NORC (2019). GSS Codebook Appendix A: Ssampling Design & Weighting. Technical report. http://gss.norc.org/documents/codebook/GSS_Codebook_AppendixA.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*.

Schoeni, R. F., F. Stafford, K. A. McGonagle, and P. Andreski (2013). Response rates in national panel surveys. *The Annals of the American Academy of Political and Social Science 645*(1), 60–87.

Shea, J. and A. Torgovitsky (2021). ivmte: An R Package for Implementing Marginal Treatment Effect Methods. *Becker Friedman Institute for Economics Working Paper* (2020-01). https://dx.doi.org/10.2139/ssrn.3516114.

Signorino, C. S. and A. Kirchner (2018). Using LASSO to Model Interactions and Nonlinearities in Survey Data. *Survey Practice 11*(1), 2716.

Sonnega, A. (2015). The Health and Retirement Study: An Introduction. https://hrs.isr.umich.edu/sites/default/files/Intro-to-HRS_0.pdf.

Squicciarini, M. P. and N. Voigtländer (2015). Human Capital and Industrialization: Evidence from the Age of Enlightenment. *The Quarterly Journal of Economics 130*(4), 1825–1883.

Stekhoven, D. J. and P. Bühlmann (2012). MissForest—Non-Parametric Missing Value Imputation for Mixed-type Data. *Bioinformatics 28*(1), 112–118.

The Norwegian Labour and Welfare Administration (2019). Utbetalte stønader fra NAV til bosatte i Norge 2019. Referansetall: Statsregnskap og utbetalingsstatistikk.

The World Bank (2020). Sample size and power calculations. Technical report. Available at https://dimewiki.worldbank.org/wiki/Sample_Size_and_Power_Calculations.

U.S. Bureau of Labor Statistics (2020a). Establishment Surveys Unit Response Rates. https://www.bls.gov/osmr/response-rates/establishment-survey-response-rates.htm.

U.S. Bureau of Labor Statistics (2020b). NLSY79 Child/Young Adults Sample Design. https://www.nlsinfo.org/content/cohorts/nlsy79-children/intro-to-the-sample/sample-design.

U.S. Bureau of Labor Statistics (2020c). NLSY79 Retention & Reasons for Non-interview. https://www.nlsinfo.org/content/cohorts/nlsy79/intro-to-the-sample/retention-reasons-noninterview.

U.S. Bureau of Labor Statistics (2020d). NLSY97 Retention & Reasons for Non-interview. https://www.nlsinfo.org/content/cohorts/nlsy97/intro-to-the-sample/retention-reasons-non-interview/page/0/1/#reasons.

U.S. Census Bureau (2006). History of the Current Population Survey. Technical report. https://www2.census.gov/programs-surveys/cps/methodology/Techincal%20paper%2066%20chapter%202%20history.pdf.

U.S. Census Bureau (2016). Sample Loss Rates For SIPP 1985 Through SIPP 2008 Panels. Technical report, Census Bureau. https://www2.census.gov/programs-surveys/sipp/tech-documentation/complete-documents/2008/sample_loss_reports_by_wave_for_1985-2008_panels.pdf.

U.S. Census Bureau (2017). Nonresponse Bias Analysis for Wave 1 2014 Survey of Income and Program Participation (SIPP) (ALYS-16). Technical report, Census Bureau. https://www2.census.gov/programs-surveys/sipp/tech-documentation/complete-documents/2014/2014_SIPP_Wave_1_Nonresponse_Bias_Report.pdf.

U.S. Census Bureau (2020a). American Community Survey Response Rates. https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/response-rates/.

U.S. Census Bureau (2020b). American Time Use Survey User's Guide. Technical report, Census Bureau. https://www.bls.gov/tus/atususersguide.pdf.

U.S. Census Bureau (2021a). American Community Survey (ACS) Response Rates Definitions. https://www.census.gov/programs-surveys/acs/methodology/sample-size-and-data-quality/response-rates-definitions.html.

U.S. Census Bureau (2021b). American Community Survey (ACS) Sample Size Definitions. https://www.census.gov/programs-surveys/acs/methodology/sample-size-and-data-quality/sample-size-definitions.html#:~:text=The%20full%20implementation%20of%20the%20ACS%20and%20PRCS%20Group%20Quarters,170%2C000%20persons%20starting%20in%202017.

U.S. Census Bureau (2021c). CPS Methodology: Non-Response Rates. https://www.census.gov/programs-surveys/cps/technical-documentation/methodology/non-response-rates.html.

U.S. Department of Health & Human Services (2019). 2018 National Health Interview Survey (NHIS) Survey Description. Technical report, Department of Health and Human Services. https://meps.ipums.org/meps/resources/srvydesc2018.pdf.

van Buuren, S., K. Groothuis-Oudshoorn, A. Robitzsch, G. Vink, L. Doove, and S. Jolani (2015). Package 'mice'. *Computer software*.