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

Appendix A Construction of top-five publications data

This appendix describes the process of constructing the data sets used to understand trends in the use and collection of survey data. We merge information from the Web of Science, JSTOR, and EconLit databases.[1] We subsequently construct measures of survey use and collection based on this data set. The entire process is summarized in the flow diagram in Online Appendix Figure A.1, and we now discuss each step of the process in greater detail.

A.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.

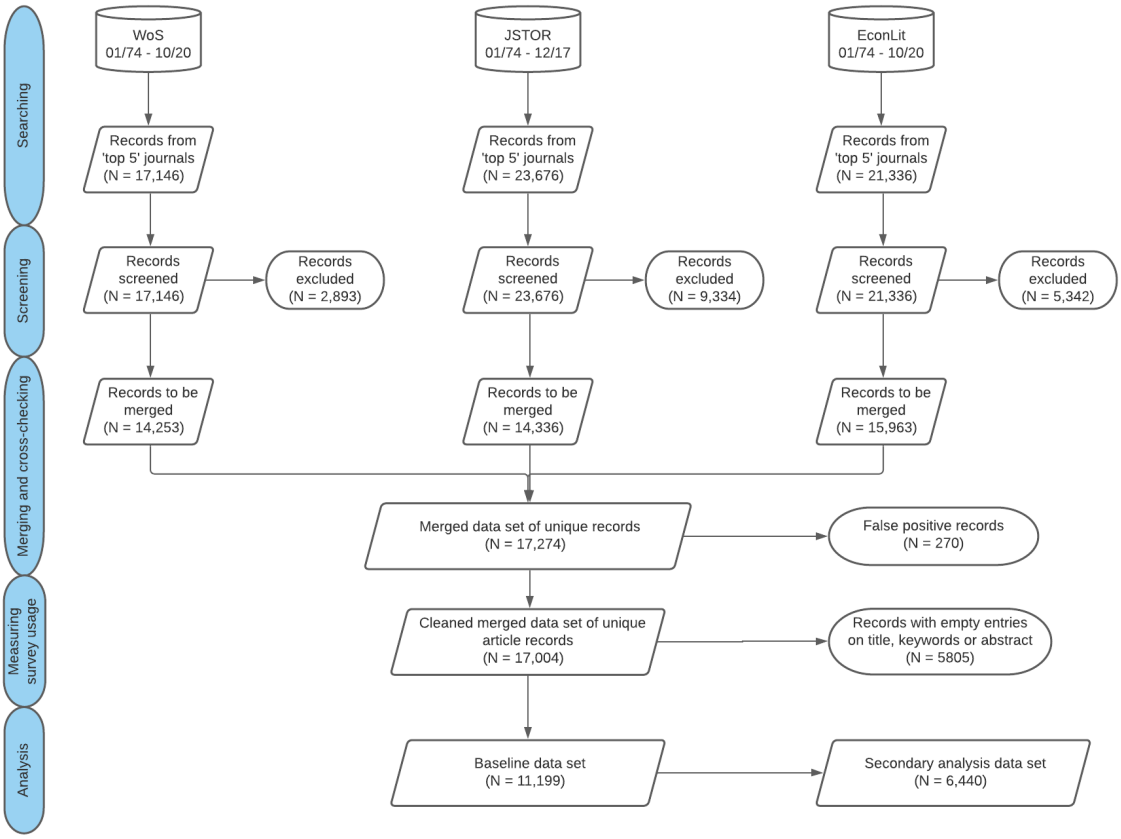
A.2 Screening

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 (for example,

[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. 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 A.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 Online 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.

African Journal of Political Economy). This leads to the exclusion of 0 records from Web of Science, 750 records 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.[2] 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.

A.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. 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.

A.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.

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. 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.

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

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).[3] We identify 112 records as satisfying this criterion.

A.5 Data sets used in our analysis

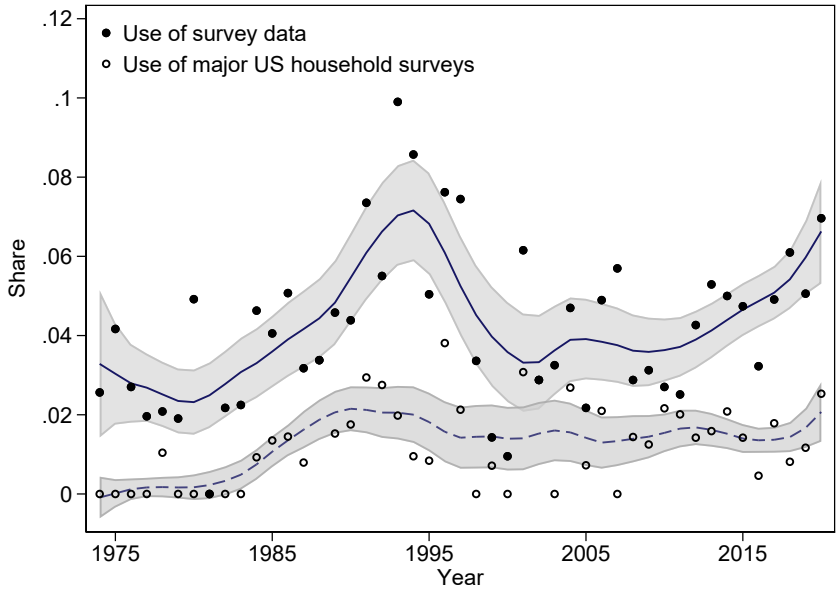
We use this data set to analyze how use of surveys and major household surveys has evolved over time, as presented in Figure 1. We also construct a secondary data set by restricting the baseline data set to records corresponding to papers classified as being in applied microeconomics. We construct this 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 Online Appendix Figure A.2. Following Currie et al. (2020), we use JEL codes to perform this restriction.

JEL codes are only available in the EconLit database.[4] 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).

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

[4] 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)).

Figure A.2: 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 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 Online Appendix A for more details on sample construction.

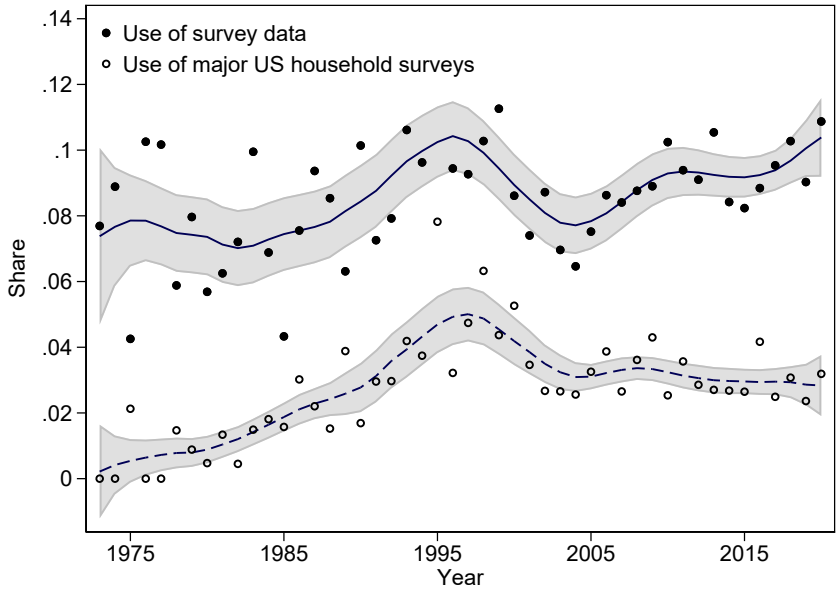
Appendix B Construction of NBER data

We use the NBER Working Paper Metadata (National Bureau of Economic Research, 2020) to capture research not published in top-five journals

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 Online Appendix A.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 Online Appendix Figure B.1.

Figure B.1: 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 Online Appendix B for more details on sample construction.

Appendix C Construction of data on nonresponse in major US household surveys

This appendix describes the process of constructing the data set used to quantify nonresponse in large-scale U.S. household surveys and used to enrich our data from our systematic review (see Online Appendix D for more on this latter purpose). Throughout this appendix, we use 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.[5]

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

[5]These surveys are the same as those used in Online Appendix A 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.

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 are conservative and use the measure that yields the lowest nonresponse rate.

The twelve surveys we consider 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, we list and describe each of the twelve surveys, and we describe how we construct nonresponse rates. The resulting data set used in our analysis is a longitudinal data set of survey-by-year observations of associated nonresponse rates.

C.1 Cross-sectional surveys

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.[6] It is a monthly survey that has been conducted since 1948.[7] We take the response rate to be the basic CPS response rate as defined by U.S. Census Bureau (2021c). 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.[8] It is a yearly survey that has been conducted since 2000.[9] 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)

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

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

[8]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.

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

The CE are administered by the Census Bureau and surveys households to find out how Americans spend their money.[10] The surveys are collected quarterly and have been conducted since the 1880s. 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.[11] The survey recruits a new panel every 2 to 4 years and conducts repeated surveys for each panel, and has been conducted since 1985.[12] 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)

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.[13] It is a biennial survey that has been conducted since 1972.[14] 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.[15] It is a yearly survey that has been conducted since 1957.[16] 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).

[10]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. See U.S. Bureau of Labor Statistics (2018) for more details.

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

[12]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).

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

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

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

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

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.[17] It is a yearly survey that has been conducted since 2003.[18] 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).[19] Response rates are available on a yearly basis between 2003 and 2020.

C.2 Longitudinal surveys

We now consider longitudinal surveys. 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.

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.[20] The sample is interviewed every 2 years and has been conducted since 1979.[21] 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.[22] The survey is included as part of the NLSY79 since 1986.[23] 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

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

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

[19] 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.).

[20] 12,686 young men and women aged 14 to 22 were first interviewed.

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

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

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

Interviewed by *Born*.

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.[24] The sample is interviewed yearly and has been conducted since 1997.[25] 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. We 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.[26] It is a yearly survey that has been conducted since 1968.[27] 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.[28] It is a yearly survey that has been conducted since 1992.[29] 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 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.

C.3 Construction of data sets used in our analysis

Combining this data across surveys 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.

[24] 8,984 young men and women aged 12 to 17 were first interviewed.

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

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

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

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

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

Appendix D Systematic review of top-five economics publications that use survey data

This appendix describes the process of constructing the data used to describe the prevalence and severity of nonresponse in empirical economics research. This data is based on a systematic review of top-five economics publications that use survey data. We conducted this review after consulting the 2020 Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines for reporting on systematic reviews that are widely followed in the biomedical sciences (Page et al., 2021). 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.

D.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.[30]

Using these search terms yielded 83 records.

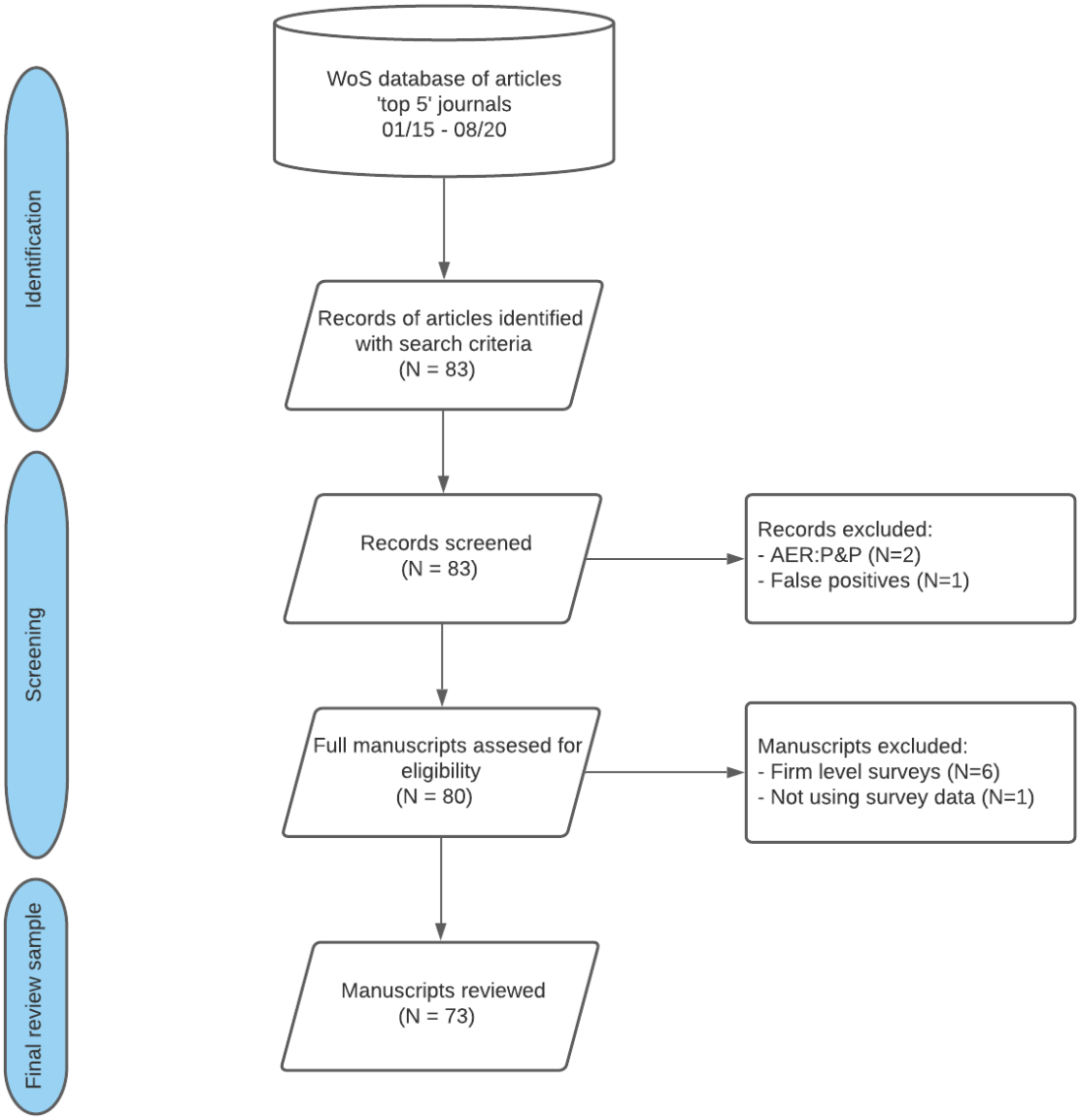
D.2 Screening

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.[31] This process is depicted in Online Figure D.1, and results in a *final review sample* of 73 studies.

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

[31] 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.

Figure D.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 Online Appendix Section D.1. *Screening* refers to the process by which the search results are screened as described in Online Appendix Section D.2. *Final review sample* describes the final sample of included in the systematic review, as described in Online Appendix Section D.3.

D.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

Table D.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 Online Appendix Section D.3.

control over the survey design (coded as "borrowed survey data").

Online Appendix Table D.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.[32] 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".[33]
- *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 pop-

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

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

ulation has a known nonzero probability of being chosen into the sample".[34] *Outreach method.* We collect information about the mode of outreach to the invited sample. We classify methods into four categories: (a) in-person (for example door-to-door), (b) online (for example 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 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* (for example 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 Online Appendix C. 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.[35] 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 (for example Census); (b) papers that use reweighting on observable characteristics to account for nonresponse[36]; and (c) papers that do not discuss potential presence of nonresponse bias.[37]

[34] 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 (for example Carvalho et al. (2016) and Elias et al. (2019)).

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

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

[37] 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.

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.[38] 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.
- *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.).[39]
- *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.[40] 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).

[38] 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.

[39] 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.

[40] 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).

Appendix E 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 confidence intervals for these differences are wide, allowing for the possibility of substantial selection. Accordingly, we examine whether these selection tests have sufficient power to detect selection in their setting, and conclude that they are underpowered due to small sample sizes and incentive levels. The analysis can only rule out implausibly large nonresponse bias, and it's unclear if the null hypothesis is not rejected due to no selection on unobservables or excessive noise.[41]

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". Online Appendix Table E.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.

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.

The first and second columns of Online Appendix Table E.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.

[41] The data we use for this analysis is provided in supplemental materials of CCFK, which is available at Coffman et al. (2019b). 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).

Table E.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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

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.

Power of CCFK's tests

The fourth column of Online Appendix Table E.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 Online Appendix Table E.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 E.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 Online Appendix Table E.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. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix F Examining misreporting and testing for direct effects of incentives on responses

In this Appendix, we first examine misreporting using survey responses for which we observe the (individual-level) ground truth in administrative data. These survey responses almost always equal the ground truth, and we thus find no evidence of misreporting. These findings also validate the exclusion restriction assumption in Section 4 that the incentives themselves do not directly affect responses. We then show how the exclusion restriction assumption can be tested without access to administrative data. We again find no evidence of a violation of this assumption.

F.1 Examining misreporting using administrative data

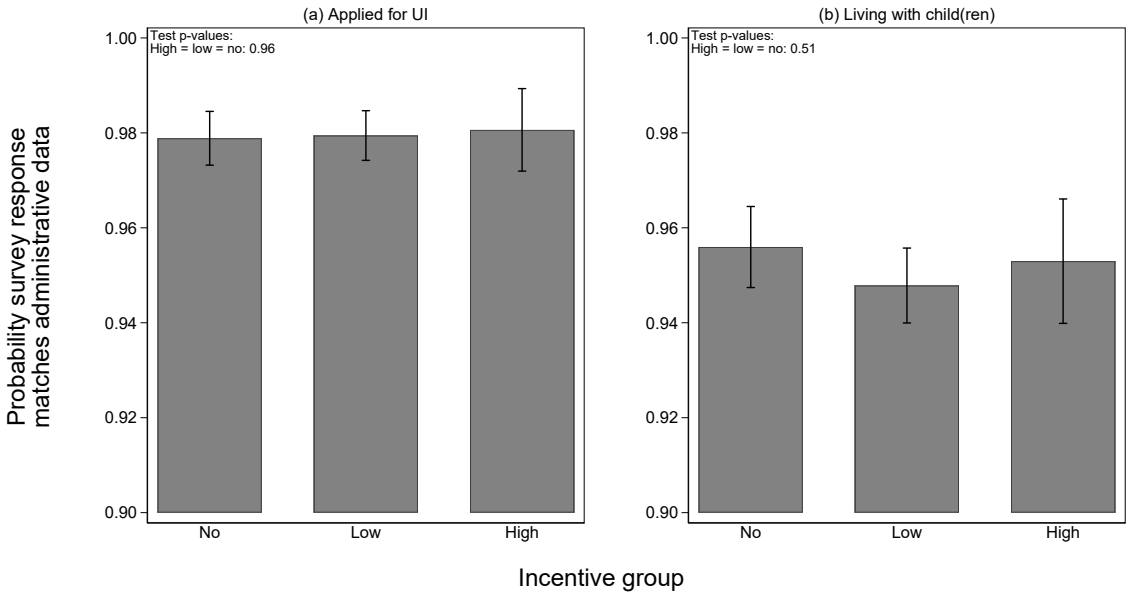
Inaccurate or untruthful reporting is always a concern when using surveys. Our setting allows us to examine misreporting using survey responses for which we observe the ground truth in administrative data. Previous research suggests survey questions relating to transfer programs are particularly suited to examine the reliability of survey responses, as stigma and confidentiality concerns may lead to under-reporting.[42] 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 also matches that of administrative data for both variables,[43] and the consistency between survey responses and the administrative data barely varies by incentive group (see Online Appendix Figure F.1). This suggests that misreporting is not a concern in the NiK survey. Consistent with this finding, we find no evidence of incentives inducing different responses to the survey.

[42] 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) and Bradburn and Sudman (1974).

[43] 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.

Figure F.1: 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.

F.2 Testing for direct effects of incentives on responses using survey data

In this appendix, we use the framework of Lee (2009) to 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_iR_i, R_i, Z_i). Let $R_i(z)$ denote the potential participation decision under incentive level 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 $Z_i \in \{0, 1\}$. In our setting, $z = 0$ denotes no incentive and $z = 1$ denotes high incentive.

As in Section 4, 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

$$\mathbb{E}[Y_i(1) - Y_i(0)|R_i(0) = 1] = \mathbb{P}[Y_i(1) = 1|R_i(0) = 1] - \mathbb{P}[Y_i(0) = 1|R_i(0) = 1]. \tag{F.1}$$

Since the second term in (F.1) is identified and equal to $\mathbb{P}[Y_i = 1|R_i = 1, Z_i = 0]$, it suffices to bound the first term. Let $p_k = \mathbb{P}[R_i(0) = k|R_i(1) = 1]$. Note that p_1 and p_0 are both identified.[44] 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 this 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 (F.1), which yields

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

with

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

$$\Delta_{ub} \equiv \min\left\{1, \frac{1}{p_1}\mathbb{P}[Y_i = 1|R_i = 1, Z_i = 1]\right\} - \mathbb{P}[Y_i = 1|R_i = 1, Z_i = 0]. \tag{F.4}$$

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

Online Appendix Table F.1 presents the results. The first column reports the participant mean for each of the four binary survey-elicited variables we consider in the main text. 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 of the estimated bounds contain zero and are relatively tight. These results thus support our assumption that incentives do not affect responses.

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

Table F.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.

Appendix G 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. This was true using two distinct sets of observable characteristics that are similar to those often used in economics research. Here, we first investigate whether that conclusion changes if we use alternative methods to address selection on observables. We then examine if the conclusion changes if we consider richer sets of observable characteristics.

G.1 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.[45] 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 Online Appendix Table G.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 obtained from our administrative data). Our implementation closely follows the

[45] 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 Online Appendix D 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.

discussion in Little and Rubin (2019) and Buskirk et al. (2018). The first three methods in Online Appendix Table G.1 require the researcher to discretize the covariates. We binarize 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 Online Appendix Table G.1 require choices regarding estimation of tuning parameters and model specification, which are described in Online Appendix Table G.1.

For each method, we test for selection on unobservables following the same procedure as in Section 4.2. In particular, for each reweighting specification, method, and incentive arm, we perform a joint test of no nonresponse bias across the six outcomes. For all joint tests, the p-value is less than 0.01, thus indicating there is significant selection on unobservables regardless of method.

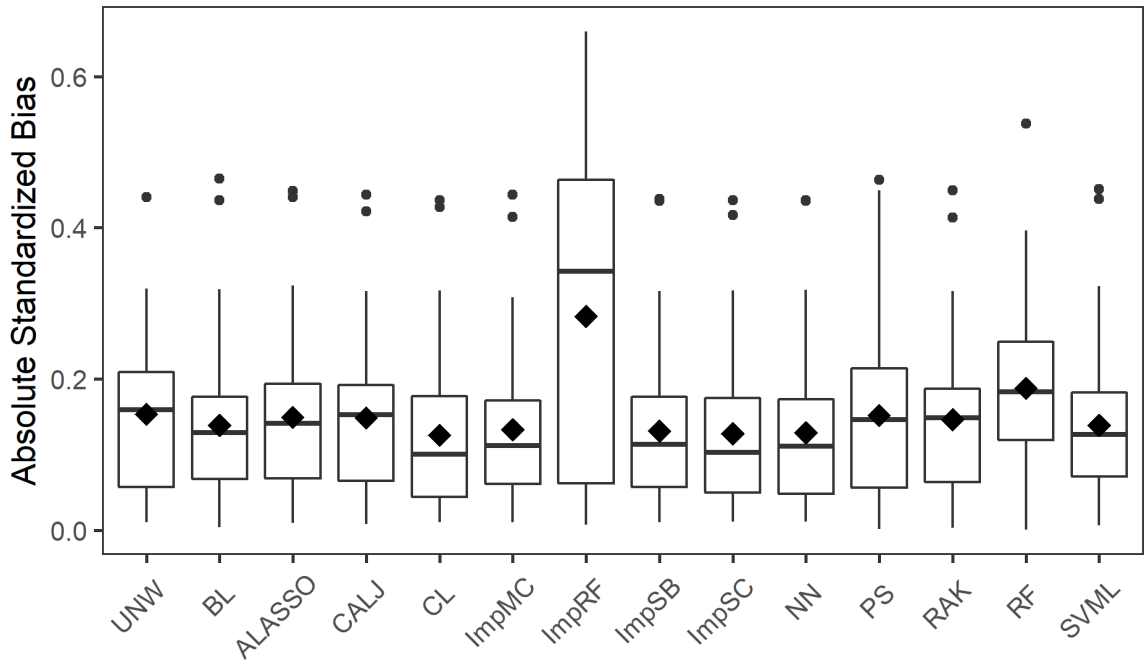
We next examine the extent to which these methods reduce nonresponse bias, relative to the methods we considered in Section 4.2. For each adjustment method, Online Appendix Figure G.1 depicts box-plots of the estimated absolute standardized nonresponse bias for each adjustment method across all combinations 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.

Table G.1: Correction methods for nonresponse implemented

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). We estimate the tuning parameter and the degree of the polynomial expansion using 10-fold cross-validation on the training data.	A
Support Vector Machine	SVML	Supervised learning model algorithm following Kirchner and Signorino (2018). 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 on the training data.	A
Neural Network classification	NN	Classification algorithm described by Eck (2018). We construct a neural network of 5 neurons in a single hidden layer using 100 training iterations.	A
Random Forest	RF	Prediction model of participation using a random classification method following Buskirk (2018). We estimate the tunning parameter using 10-fold cross-validation on the training data along with a one-standard error rule and run a random forest using 500 classification trees.	A
Random Forest Imputation	ImpRF	Random Forest algorithm to estimate the imputation function as described by Buskirk (2018). We estimate the tuning parameter using 10-fold cross-validation on the training data along with a one-standard error rule and run a random forest to predict the outcome using 100 classification trees.	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). Machine learning methods are first implemented on a training data set and extrapolated on the whole data set to avoid over-fitting. Our training dataset is obtained by randomly sampling 20% of the no incentive sample.

Figure G.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 Online Appendix Table G.1. The filled diamonds depict the mean of each set of estimates.

G.2 Alternative choices of characteristics

The two sets of characteristics we use in Section 4 are commonly used for reweighting in survey research in economics. Our ability to observe administrative data for both participants and non-participants allows us to examine whether richer sets of characteristics (which may not be observed in typical survey research) can further reduce or eliminate nonresponse bias.

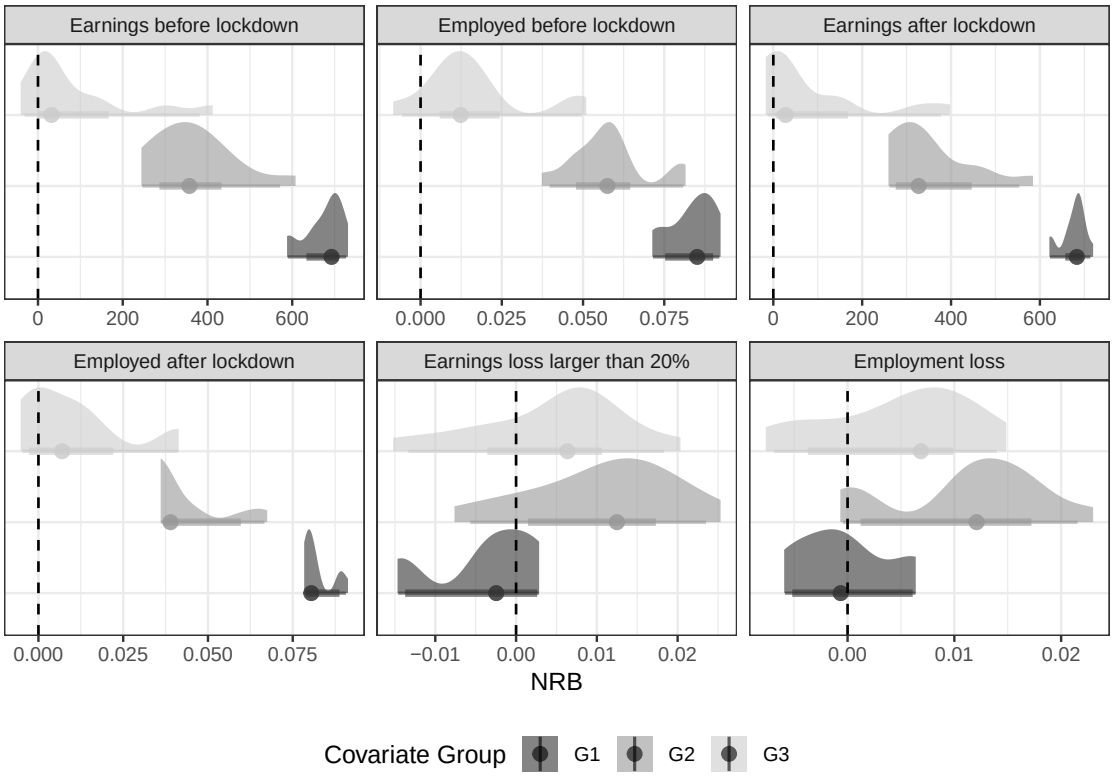
To examine this, we reweight participant means using three nested sets of characteristics. The first set, which we denote as G1, is the baseline set of municipality-level variables considered in Section 4. The second set, which we denote as G2, includes all the variables in G1 and additionally includes a richer set of individual-level variables relative to those considered in Section 4.[46] Lastly, the third set, which we denote as G3, includes all the variables in G2 and additionally includes lagged earning and employment outcomes from 2019. For each of these sets of characteristics, we produce reweighted estimates using three different specifications and for each of our participant samples (pooled and each incentive level separately).[47]

[46] In particular, we additionally include indicators for being married and having children.

[47] Specifically, the first specification includes the characteristics linearly in the logit model, the second specification adds pairwise interactions between all variables and second order terms for the continuous variables, and the third specification chooses the variables to use via Adaptive Lasso on the variables included in the second specification.

Figure G.2 presents the distribution of estimates under each set of characteristics separately by outcome. We find that there is no set of characteristics (nor specification) that eliminates nonresponse bias across all outcomes. Although reweighting using richer sets of characteristics yield estimates that tend to have lower nonresponse bias for level outcomes, the opposite is true for earnings and employment loss. Lastly, there are many choices of specifications under which using richer sets of characteristics would yield estimates with higher nonresponse bias than using less rich characteristics.

Figure G.2: Non-response bias by nested covariate groups



Notes: This figure presents density plots of the estimated nonresponse bias (NRB) by covariate group (in different colors) used for reweighting across outcomes. For each density, we present the median in a dot, and the interquantile range in a solid line.

Appendix H Treatment effects

In this section, we show how the framework used in Section 4 can be applied to analyze nonresponse bias when the object of interest is the effect of a randomly assigned treatment.

Let $D_i \in \{0,1\}$ be a binary treatment variable and let $Y_i^\star(0), Y_i^\star(1)$ be potential outcomes. The realized outcome is $Y_i^\star = (1-D_i)Y_i^\star(0) + D_i Y_i^\star(1)$. Realized outcomes are only observed for those who respond, so that $Y_i = Y_i^\star$ only if $R_i = 1$, as before. However, now the researcher's object of interest is the average treatment effect for the entire population under study:

$$\textsc{ate} \equiv \mathbb{E}[Y_i^\star(1) - Y_i^\star(0)].$$

Responses may be affected by both the incentive level and treatment status, so we now index potential responses as $R_i(d,z)$. Actual responses are then given by

$$R_i = \sum_{z,d} \mathbb{1}[Z_i = z, D_i = d] R_i(d,z),$$

Assume that (D_i, Z_i) are randomly assigned so that they are independent of $\{Y_i^\star(d), R_i(d,z)\}_{d,z}$.

Random assignment implies that the observed conditional mean for treatment arm d and incentive z is

$$\mathbb{E}[Y_i | R_i = 1, D_i = d, Z_i = z] = \mathbb{E}[Y_i^\star(d) | R_i(d,z) = 1].$$

A treatment-control contrast for incentive level z gives the difference in means

$$\begin{aligned}\textsc{dm}(z) &\equiv \mathbb{E}[Y_i | R_i = 1, D_i = 1, Z_i = z] - \mathbb{E}[Y_i | R_i = 1, D_i = 0, Z_i = z] \\ &= \mathbb{E}[Y_i^\star(1) | R_i(1,z) = 1] - \mathbb{E}[Y_i^\star(0) | R_i(0,z) = 1].\end{aligned}$$

Nonresponse bias means that $\textsc{dm}(z)$ is generally not equal to $\textsc{ate}$. The difference can be decomposed into

$$\textsc{dm}(z) - \textsc{ate} = \textsc{eb}(z) + \textsc{ib}(z),$$

where $\textsc{eb}(z)$ and $\textsc{ib}(z)$ are the "external" and "internal" nonresponse biases, defined as

$$\begin{aligned}\textsc{ib}(z) &\equiv \mathbb{E}[Y_i^\star(0) | R_i(1,z) = 1] - \mathbb{E}[Y_i^\star(0) | R_i(0,z) = 1] \\ \textsc{eb}(z) &\equiv \mathbb{E}[Y_i^\star(1) - Y_i^\star(0) | R_i(1,z) = 1] - \mathbb{E}[Y_i^\star(1) - Y_i^\star(0)].\end{aligned}$$

Internal nonresponse bias may occur if the treatment has a causal effect on responding to the survey. For example, a treatment assigned in a field experiment could affect the cost of responding to an endline survey conducted at a date after random assignment. This might be a particular consideration in a development economics context, where the treatment

assignments can have large impacts on material well-being.

In a lab experiment, it might be reasonable to assume that the researcher collects outcomes for all subjects regardless of their treatment status, so that $R_i^\star(0,z) = R_i^\star(1,z)$. In this case, there is no internal nonresponse bias. However, there could still be external nonresponse bias if the subjects who choose to participate in the lab experiment differ systematically from the population under study.

Randomized incentives enable a joint test of the null hypothesis that there is no internal or external nonresponse bias by testing that $\textsc{dm}(z)$ is constant in z. The difference with the case considered in the main text is simply that $\textsc{dm}(z)$ is a treatment-control contrast, rather than a single conditional mean. If it is assumed that internal nonresponse bias is not possible, then the test is of the null that hypothesis that there is no external nonresponse bias. To determine if nonresponse bias is due to observables or unobservables, the test can be conducted after reweighting by observable characteristics.

Appendix I 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 I.1, we discuss partial identification of the population mean. In I.2, we discuss the procedure to estimate these bounds given sampled data.

I.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

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

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 polynomials. We further assume that θ belongs to a pre-specified parameter space Θ, that is, 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

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

where $\tau \equiv \mathbb{E}\left[B(1, X_i)\right]$ is identified.[48]

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 $R_i = 1$, the quantity $\mathbb{E}[Y_i | R_i = 1, X_i = x, P_i = p]$ can be written as a function of θ via

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

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

The set of θ that satisfy this equality is equal to the set of θ that satisfy the normal equations

$$\mathbb{E}\left[B(P_i, X_i)B(P_i, X_i)|R_i = 1\right]'\theta = \mathbb{E}\left[B(P_i, X_i)Y_i|R_i = 1\right].$$

We measure the extent to which the normal equations are satisfied with the population criterion function

$$Q(\theta) \equiv \sum_{j=1}^{d_\theta} \Big| \mathbb{E}\left[(Y_i - \theta' B(P_i, X_i))' B_j(P_i, X_i)|R_i = 1\right] \Big|. \tag{I.3}$$

Then we define the identified set for θ as

$$\Theta^\star \equiv \{\theta \in \Theta : Q(\theta) = 0\}.$$

The corresponding identified set for the population mean (as defined in (I.2)) is then $[\tau_{lb}^\star, \tau_{ub}^\star]$, where

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

I.2 Estimation

We estimate (I.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 $N = \sum_{i=1}^n R_i$ participants. 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 $Q(\theta)$

$$\hat{Q}(\theta) \equiv \frac{1}{d_\theta} \sum_{j=1}^{d_\theta} \left| \frac{1}{N} \sum_{i:R_i=1} (Y_i - \theta' B_i) B_{ij} \right|, \tag{I.5}$$

where $B_i \equiv B(P_i, X_i)$ and B_{ij} denotes its j-th element. To find the set θ that minimize $\hat{Q}(\theta)$ we solve for $\hat{Q}^\star \equiv \min_{\theta \in \Theta} \hat{Q}(\theta)$. Then, we use these estimates to construct estimated bounds on the population mean, denoted $[\hat{\tau}_{lb}^\star, \hat{\tau}_{ub}^\star]$, by solving the sample analogs of (I.4), namely

$$\hat{\tau}_{lb}^\star \equiv \min_{\theta \in \hat{\Theta}^\star} \hat{\tau}'\theta \quad \text{and} \quad \hat{\tau}_{ub}^\star \equiv \max_{\theta \in \hat{\Theta}^\star} \hat{\tau}'\theta \tag{I.6}$$

where $\hat{\tau} = \frac{1}{n} \sum_i B_i(1, X_i)$ and $\hat{\Theta}^\star \equiv \{\theta \in \Theta : \hat{Q}(\theta) = \hat{Q}^\star\}$. These optimization problems can be reformulated as linear programs by introducing slack variables for the absolute value in $\hat{Q}(\theta)$. We solve these programs using Gurobi (Gurobi Optimization, 2021).

Appendix J Randomized reminders

The NiK survey sent reminders to all individuals who had not yet participated by the reminder date. In the main text, we use the reminder date as a natural division to test for nonresponse bias under the assumption that responses are time-invariant. In this appendix, we consider the value of an alternative scheme in which a reminder is sent to some randomly-assigned subset of these individuals.

Suppose that individual i is still assigned a random incentive Z_i, but is now also randomly assigned into a binary reminder ("contact") group C_i. Those with $C_i = 1$ are sent a reminder at the same time period sometime after the initial contact attempt, while those with $C_i = 0$ are not sent a reminder. Let $R_i(z, c)$ denote their potential response decisions if given (financial) incentive z and reminder status c. Actual responses are $R_i = R_i(Z_i, C_i)$. Randomization ensures that (Z_i, C_i) is independent of $(Y_i^\star, R_i(z, c))$ for every (z, c) pair. We maintain the exclusion restriction that responses are not affected by either Z_i or C_i.

Using the same arguments as in Section 4, we can test for nonresponse bias by comparing the average responses of participants assigned to different incentive and contact arms. If there is no nonresponse bias, then $\mathbb{E}[Y_i|R_i = 1, Z_i = z, C_i = c]$ should be the same for all pairs (z, c). In contrast to the case where reminders are sent to all individuals, having a randomized reminder does not require time-invariance because the reminder now operates like the incentive in producing comparisons across participants for any given time period. However, these comparisons will only have the power to detect nonresponse bias in time periods at or after the reminder is actually sent, so that the reminder has the potential to affect response decisions. The overall impact of having randomized reminders in addition to incentives then is to produce more values of (z, c), providing more equalities to test and more possible dimensions on which to detect nonresponse bias.

Appendix K Identification and extrapolation with the two-dimensional participation model

In Online Appendix K.1, we discuss extrapolation using the two-dimensional model. In Online Appendix K.2, we prove that our test of independence between unobserved margins in Section 6 is the strongest testable implication of independence. In Online 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. In Online 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. Lastly, in Online Appendix K.5, we show how to incorporate covariates.

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 sets $\mathcal{M}$ and Π. These sets 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 $T_i \in \{1,2\}$.

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

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

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

We construct an identified set for the population mean through a three step procedure that builds on the argument discussed in Appendix I. Let $\mathcal{A}\equiv\{(z,s):z\in\{0,1\},s\in\{1,2\}\}$. In the first step, we find $\pi\in\Pi$ that satisfy (K.2) for all $(z,s)\in\mathcal{A}$ by solving

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

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 $\mu\in\mathcal{M}$ that satisfy (K.3) for all $(z,s)\in\mathcal{A}$ by solving

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

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 compute an identified set for the population mean (defined in (K.1)) as $[\tau^\star_{lb},\tau^\star_{ub}]$, where

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

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 a pooled panel data regression, solving

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

Let $\hat{\pi}^\star$ be a minimizer of (K.7) and define $\hat{\pi}^\star_{\mathcal{A}}\equiv\{\hat{\pi}^\star_{zs}:(z,s)\in\mathcal{A}\}$. 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) and define $\hat{\mu}^\star_{\mathcal{A}} \equiv \{\hat{\mu}^\star_{zs} : (z,s) \in \mathcal{A}\}$. Then we estimate the population bounds $[\hat{\tau}^\star_{lb}, \hat{\tau}^\star_{ub}]$ through the estimated analogue of (K.6) by taking

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

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 Online Appendix 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 Online Appendix 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 $\pi_{zs} \equiv \mathbb{P}[\eta(z-1) < V_i \leq \eta(z), S_i = s]$. 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 $\mu_{zs} \equiv \mathbb{E}[Y_i^\star | \eta(z-1) < V_i \leq \eta(z), S_i = s]$. 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 Online 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 because V_i was normalized 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 Online Appendix I, if $m(v, s) = \sum_k \theta_k b_k(v, s)$ where b_k are known basis functions and θ belongs to some pre-specified parameter space Θ, then (K.10) can be written as

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

Let $\hat{\tilde{B}}^\star(z, s)$ denote the values of $\tilde{B}(z, s)$ given $\hat{\pi}^\star$. Then define the criterion function

$$\arg\min_{\theta \in \Theta} \hat{Q}(\theta) \quad \text{where} \quad \hat{Q}(\theta) \equiv \frac{1}{d_\theta} \sum_{j=1}^{d_\theta} \left| \frac{1}{N} \sum_{i:T_i \in \{1,2\}} (Y_i - \theta' \hat{\tilde{B}}_i^\star) \hat{\tilde{B}}_{ij}^\star \right|, \tag{K.12}$$

where $\hat{\tilde{B}}_i^\star \equiv \hat{\tilde{B}}^\star(Z_i, T_i)$ and $\hat{\tilde{B}}_{ij}^\star$ denotes its jth element. Let $Q^\star := \min_{\theta \in \Theta} \hat{Q}(\theta)$. We estimate bounds with

$$\hat{\tau}_{lb}^\star \equiv \min_{\theta \in \hat{\Theta}^\star} \hat{\tilde{\tau}}' \theta \quad \text{and} \quad \hat{\tau}_{ub}^\star \equiv \max_{\theta \in \hat{\Theta}^\star} \hat{\tilde{\tau}}' \theta \tag{K.13}$$

where $\hat{\tilde{\tau}}$ has kth component $\frac{1}{n} \sum_i \sum_s \hat{\mathbb{P}}[S_i = s] \int_0^1 b_k(v, s) dv$ and $\hat{\Theta}^\star \equiv \{\theta \in \Theta : \hat{Q}(\theta) = \hat{Q}^\star\}$.

The considered restrictions on the MSR are imposed via choice of basis function and choice of parameter space Θ. They are depicted in Online Appendix 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, we restrict Π to include valid distributions and restrict $\mathcal{M}$ to be group responses that are bounded, separable, and monotone in both margins. 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. Given these assumptions, we impose additional assumptions via the MSR. In the third set, we assume that the MSR is separable and monotone in both margins. The implementation is presented in the first row of Online Appendix 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 Online Appendix Table K.3.

Under the first two 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) and the two in (K.9). 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)).

Under the third and fourth sets of assumptions, the shares are estimated from the data using sample analogues. Thus, three programs need to be solved: the program in (K.12) and the two programs in (K.13). These optimization problems can be reformulated as linear programs by introducing slack variables for the absolute value in $\hat{Q}(\theta)$. We solve these programs using Gurobi (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

$$\mathbb{P}[\tilde{V}_i \leq \eta(1)] = \frac{(p_{01}+p_{02})p_{11}}{p_{01}} = p_{11} + p_{02}\frac{p_{11}}{p_{01}} = p_{11} + p_{12} \in [0,1], \tag{K.16}$$

where the last equality 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 $\mathbb{P}[\tilde{S}_i = 2] = \frac{p_{02}}{p_{01}+p_{02}}$. Then $\mathbb{P}[\tilde{S}_i = t] \in [0,1]$ for $t = 1,2$, and

$$\mathbb{P}[\tilde{S}_i = 3] = 1 - \frac{p_{01}}{p_{01}+p_{02}} - \frac{p_{02}}{p_{01}+p_{02}} = 0 \in [0,1]. \tag{K.17}$$

Take $\tilde{V}_i$ 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

$$\tilde{p}_{12} = \mathbb{P}[\tilde{V}_i \leq \eta(1)]\,\mathbb{P}[\tilde{S}_i = 2] = \left(\frac{(p_{01}+p_{02})\,p_{11}}{p_{01}}\right)\left(\frac{p_{02}}{p_{01}+p_{02}}\right) = \frac{p_{11}p_{02}}{p_{01}} = \frac{p_{12}p_{02}}{p_{02}} = p_{12}, \tag{K.18}$$

where the last equality 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

$$\mathbb{P}[T_i = t|Z_i = z] = \mathbb{P}[V_i \leq \eta(z)]\,\mathbb{P}[S_i = t] = \eta(z)\,\mathbb{P}[S_i = t] \tag{K.19}$$

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

$$\frac{\mathbb{P}[S_i = 2]}{\mathbb{P}[S_i = 1]} = \frac{\mathbb{P}[T_i = 2|Z_i = 0]}{\mathbb{P}[T_i = 1|Z_i = 0]}. \tag{K.20}$$

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

$$\mathbb{P}[S_i = 1] = 1 - \mathbb{P}[S_i = 2] - \mathbb{P}[S_i = 3] = 1 - \mathbb{P}[S_i = 1]\frac{\mathbb{P}[T_i = 2|Z_i = 0]}{\mathbb{P}[T_i = 1|Z_i = 0]} - \mathbb{P}[S_i = 3] \tag{K.21}$$

and obtaining

$$\mathbb{P}[S_i = 1] = \frac{(1 - \mathbb{P}[S_i = 3])\,\mathbb{P}[T_i = 1|Z_i = 0]}{\mathbb{P}[T_i = 1|Z_i = 0] + \mathbb{P}[T_i = 2|Z_i = 0]}. \tag{K.22}$$

We then identify $\mathbb{P}[S_i = 2]$ by substituting (K.22) into (K.20), and then $\eta(z)$ from (K.19) with $z = 0, 1$.

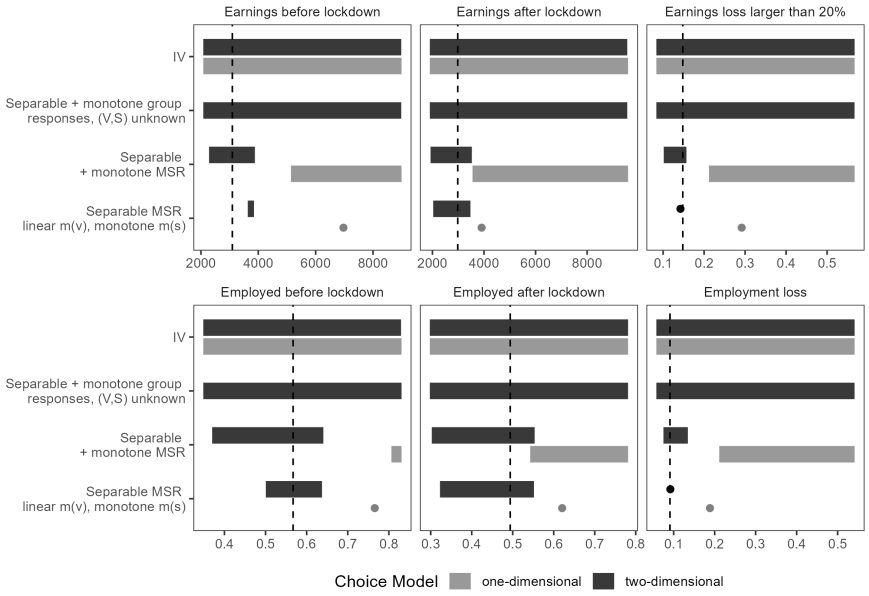
K.4 Sensitivity of extrapolation estimates to choice of passive share

In the third and fourth rows of Figure 14 of Section 6, we assume that 40% never see the invitation to participate (that is, we assume the passive share is $\mathbb{P}[S_i = 3] = .4$). This number was chosen in consultation with survey researchers at Statistics Norway. Online Appendix Figure K.1a 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 Online Appendix Figure K.1b 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.

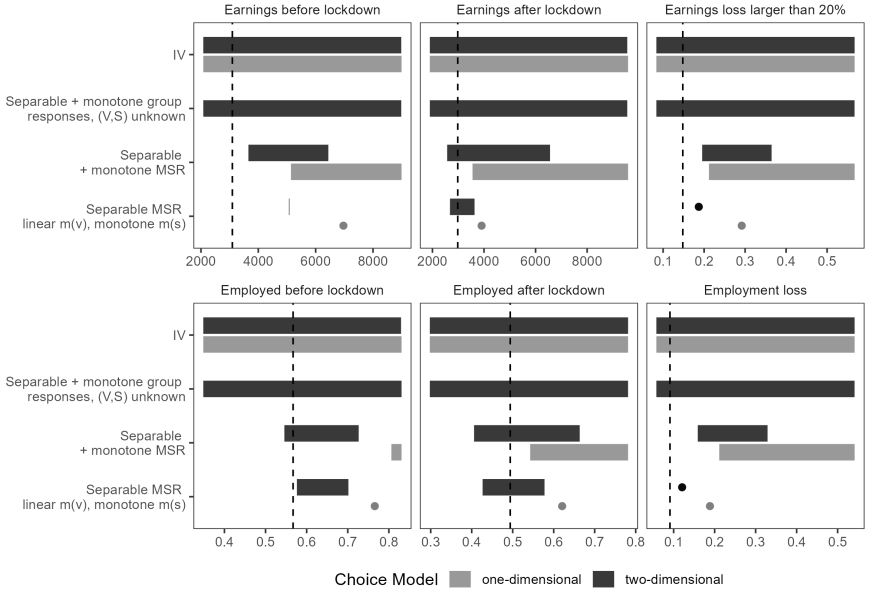
To further examine the robustness of our results to the choice of passive share, we compute estimated bounds under the final specification considered in Section 6.3 (fourth row of Figure 14) for all passive share choices in the range of $10\% - 48\%$ for all six outcomes. For each outcome, the second 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

Figure K.1: Bounds under double threshold model assumptions for different passive shares

(a) Passive share = 48%



(b) 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 48% (panel a) or 32% (panel b) 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.

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.

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 $\equiv \max_{\alpha \in [0.10, 0.48]} f(\alpha)$. 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.

K.5 Incorporating covariates

The selection model with covariates is

$$R_{it}(z) = \mathbb{1}[S_i \leq t]\mathbb{1}[V_i \leq \eta(z, X_i)], \tag{K.23}$$

where V_i is normalized to be uniform on $[0, 1]$, conditional on X_i. The MSR is $m(v, s, x) \equiv \mathbb{E}[Y_i^\star | V_i = v, S_i = s, X_i = x]$. The derivation of (12) follows from conditional-on-X_i modifications of the same arguments that lead to (K.10) under conditional-on-X_i versions of the independence and passive share assumptions. In particular, these assumptions are that V_i and S_i are independent, conditional on X_i, and that $\mathbb{P}[S_i = 3|X_i]$ is known. Under these assumptions, conditional-on-X_i versions of the same arguments that lead to (K.10) show that

$$\mathbb{E}[Y_i | T_i = s, Z_i, X_i] = \frac{1}{\eta(Z_i, X_i)} \int_0^{\eta(Z_i, X_i)} m_{V,S}(v, s) dv + m_X(X_i), \tag{K.24}$$

where $\eta(Z_i, X_i)$ is point identified given the maintained assumptions. Given estimates of η, we then proceed as in Online Appendix K.1 by solving the problems in (K.12) and (K.13) where $\hat{\tilde{B}}^\star$ is now also a function of X_i and Θ encodes boundedness restrictions for every value of X_i.

For the results reported in Table 5 and Figure 15, we use the same covariates as in Section 4: gender, age, education level, and an indicator for being an immigrant. We specify X_i to be a vector that contains an indicator for female, indicators for quartile-binned age cells, indicators for quartile-binned education levels, and an indicator for being an immigrant. We specify $m_X(X_i)$ as additive in each component of X_i.

To estimate $\eta(Z_i, X_i)$, the arguments in Online Appendix K.3 show that it suffices to

estimate $\mathbb{P}[T_i = t|Z_i = z, X_i = x]$ and $\mathbb{P}[S_i = 3|X_i = x]$. We obtain estimates of $\mathbb{P}[T_i = t|Z_i = z, X_i = x]$ using logit regressions of $\mathbb{1}[T_i = 1]$ and $\mathbb{1}[T_i = 2]$ on Z_i and additive in each component of X_i. Without covariates, our main specification imposed that $\mathbb{P}[S_i = 3] = 0.4$. To be consistent with this assumption while ensuring that $\mathbb{P}[S_i = 3|X_i = x] \leq \mathbb{P}[T_i = 3|X_i = x, Z_i = 1]$, i.e. that the share who do not see the invitation is less than the share who would not participate under the high incentive, we instead impose that $\mathbb{P}[S_i = 3|X_i = x] = \mathbb{P}[S_i = 3]\frac{\mathbb{P}[T_i=3|X_i=x,Z_i=1]}{\mathbb{P}[T_i=3|Z_i=1]}$ with $\mathbb{P}[S_i = 3] = 0.4$.

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.

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

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.

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.

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 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 (2018). Consumer Expenditures and Income: History. Technical report. https://www.bls.gov/opub/hom/cex/history.htm.

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 Noninterview. 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 Noninterview. https://www.nlsinfo.org/content/cohorts/nlsy97/intro-to-the-sample/retention-reasons-non-interview/page/0/1/#reasons.

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