Which Businesses Answer Surveys? Evidence from Dutch Administrative Data

Jack Fitzgerald*

May 5, 2025

Most Recent Version

Abstract

I leverage a unique administrative register covering the universe of establishments in the Netherlands to examine how characteristics differ between establishments that do and do not respond to business surveys. Only 20% of Dutch establishments responded to regional business surveys in 2022. Responsive establishments employed two fewer people than unresponsive establishments, and exhibited parttime employment rates 15 percentage points higher than unresponsive establishments. Sectoral and occupational response rates can vary by nearly 50 percentage points, with public-sector and white-collar occupations overrepresented amongst responsive establishments. Solo enterprises registered to residential addresses are the most common kind of establishment, but exhibit response rates 18 percentage points lower than an average office. However, controlling for contact probability reveals that most sectoral and occupational variation in response rates can be traced back to differences in contact probability rather than responsiveness. These findings highlight generalizability challenges in business surveys and opportunities to improve their design.

Keywords: Representativeness, sampling, imputation, random forest, robust regression

*Affiliation: Vrije Universiteit Amsterdam and Tinbergen Institute; School of Business and Economics, Department of Ethics, Governance, and Society; Amsterdam, Noord-Holland, The Netherlands. Email: j.f.fitzgerald@vu.nl. ORCID: 0000-0002-0322-5104. I thank Daan Breukel for providing details on the administration of LISA's regional business surveys and on the aggregation of the LISA register, as well as Peter Gerbrands, Wolter Hessink, Rutger Schilpzand, and Arjen van Witteloostuijn for organizing and providing access to the FIRMBACKBONE database. All remaining errors are my own. I am grateful to seminar participants at Vrije Universiteit Amsterdam for helpful comments and feedback.

1 Introduction

The representativeness of business surveys has implications for a considerable amount of research in management, finance, and economics. 32-41% of empirical research publications in management information systems use survey data, of which over 41% target firms as the primary unit of analysis (Karanja, Sharma, & Salama 2020). Business surveys worldwide regularly elicit information about firm expectations and financial management practices, which form a widely-used knowledge base for understanding and forecasting economic activity (Zimmermann 1999; Collins 2001; Hansson, Jansson, & Löf 2005; Baker & Mukherjee 2007; Clar, Duque, & Moreno 2007; Klein & Özmucur 2010; Baker, Singleton, & Veit 2011; Snijkers et al. 2013; Altig et al. 2022). If such surveys are conducted on unrepresentative samples of establishments, then the insights they generate may not accurately generalize to the economy at large, instead producing misleading conclusions about firm behavior, the future of financial markets, and the state and trajectory of the economy.

Like all surveys, the representativeness of businesses surveys is threatened by the fact that the sorts of firms who respond to business surveys may differ from the general population of firms in meaningful ways, but typical techniques to monitor and adjust for this selective nonresponse are unfortunately infeasible in most business surveys. In surveys with human subjects, selective nonresponse is often measured by comparing the characteristics of participants who respond to a survey with those of a sample containing people who do not respond to the survey (e.g., Schouten et al. 2012; Shlomo, Skinner, & Schouten 2012; Bianchi & Biffignandi 2017; Cornesse & Bosnjak 2018; Moore, Durrant, & Smith 2018). These differences are often observable in surveys with human subjects because nonresponsive people's characteristics can be recovered from a census or an administrative register. When such reference data is available, one can also adjust for nonresponse bias *a priori* through representative sampling or *ex post* through sample weighting. However, in most research contexts, there is no such 'business census' containing data on unresponsive businesses. Consequently, little is known about how characteristics differ between firms which do and do not respond to business surveys, nor how to adjust for selective nonresponse in business surveys.

In this paper, I leverage a unique administrative register covering the universe of establish-

ments in the Netherlands to assess how characteristics differ between establishments which do and do not respond to business surveys. I specifically use the *Landelijk Informatiesysteem van Arbeidsplaatsen (LISA)*, which combines administrative microdata from national business and real estate registers with data from regional business surveys. Only 20% of Dutch establishments responded to these regional business surveys in 2022. Because LISA draws data from other administrative sources for unresponsive establishments, I can directly examine how characteristics differ between responsive and unresponsive establishments.

I additionally exploit the random sampling process of the regional business surveys underpinning the LISA register to disentangle the determinants of *responsiveness* from the determinants of *contact probability*. Conditional on a number of variables upon which the regional business surveys stratify or deviate from their random sampling protocol, establishments are randomly contacted for surveys. Thus after controlling for these variables, estimated relationships between establishment characteristics and response rates are statistically independent from contact probability, instead solely reflecting differences in establishments' propensity to respond to surveys. These contact-conditional relationships between characteristics and response rates can generalize to other contexts, providing important insights about the sorts of establishments which are more or less likely to respond to a business survey if they are contacted for one.

I document significant compositional differences between responsive and unresponsive establishments. Responsive establishments hire two fewer employees than unresponsive establishments. The workforce of responsive establishments is also composed of fewer fulltime employees, with fulltime employment rates in responsive establishments being 15 percentage points lower than those rates in unresponsive establishments.

There are also important sectoral differences between responsive and unresponsive establishments. Response rates between the most responsive and least responsive firm types vary by around 50 percentage points. Public-sector and white-collar occupations such as public administration, real estate, finance, education, and healthcare are overrepresented amongst responsive establishments. Though the most common type of establishment in the LISA register is a solo enterprise whose address is registered at the entrepreneur's personal home, these establishments have one of the lowest response rates of all establishment types, being

over 18 percentage points less likely to respond to LISA's regional business surveys than a typical office.

However, controlling for contact probability considerably attenuates gaps in response rates between sectors and occupations. The gap in response rates between the most and least responsive sectors (facility types) attenuates from nearly 50 (53) percentage points to just eight (23) percentage points after controlling for contact probability. The response rate deficit for establishments registered to a residential address declines from over 18 percentage points to just over three percentage points after conditioning on contact determinants.

These findings demonstrate that business surveys can yield quite unrepresentative samples, but also highlight the potential for improvements in the design of business surveys. Though this paper's results concerning unconditional response rates will only generalize to business surveys with similar sampling strategies to LISA's regional business surveys, its results concerning contact-conditional response rates can generalize more broadly, revealing which sorts of establishments respond most frequently when contacted for future surveys. If one wishes to improve the representativeness of a future business survey, then one can oversample (undersample) the least (most) contact-conditionally responsive establishments. Alternatively, a resource-constrained surveyor can maximize response rates by oversampling (undersampling) the most (least) contact-conditionally responsive establishments. Given how sharply gaps in response rates decline after controlling for contact probability, these changes in survey design can likely yield strong effects on representativeness and/or response rates, ensuring better-quality survey data for more informative insights on businesses.

Section 2 introduces the LISA register and discusses the degree of nonresponse to its constitutent regional business surveys. I also discuss how LISA imputes employee headcounts for unresponsive establishments, propose an improved imputation algorithm, and show that this proposed improvement yields better predictive performance. Section 3 discusses estimation procedures and details the control strategy that allows me to isolate differences in response rates due to contact probability from differences in response rates due to responsiveness. Section 4 details the empirical results, and Section 5 concludes.

2 Data

Each year, regional bodies in the Netherlands send businesses surveys (*werkgelegenheid-senquêten*) to assess the state of work. Establishments in each of the twelve Dutch provinces can receive surveys from one of seventeen business registers, most of which represent their full province.[1] Sampling establishments from the registers of the *Kamer van Koophandel* (Chamber of Commerce; KVK), these surveys inquire about the number of workers employed by each establishment. Specifically, the surveys inquire about employee headcounts of female fulltime, female parttime, male fulltime, and male parttime workers. These surveys define a parttime worker as one who works fewer than 12 hours per week (LISA 2018).

These surveys elicit information at the establishment level, rather than the firm level. E.g., in this data, a grocery store chain will not appear as one row covering the details of the entire chain, but will instead occupy one row for each local grocery store. This information may be provided on behalf of the local establishments by representatives of the broader firm (rather than representatives of the local establishments), but the aim is to aggregate data at the establishment level.

Each year, the *Landelijk Informatiesysteem van Arbeidsplaatsen (LISA)* aggregates this regional business survey data across all regional business registers to create a national register. The LISA register combines the regional business survey data with administrative business records from the KVK's *Handelsregister* and physical establishment data from the Register of Adressess and Buildings (*Basisregistratie van Adressen en Gebouwen*), which is managed by *Kadaster*. The resulting register comprises annual panel data on the universe of establishments of all employing organizations in the Netherlands from 1996 onwards.[2]

My analysis draws on data from the 2021 and 2022 LISA registers. The data is provided through FIRMBACKBONE, a novel data infrastructure that combines register data from

[1] There are exceptions in four provinces. (1) In Groningen, the *municipality* of Groningen is covered by its own separate register. (2) North Brabant is split into five regional registers, covering Breda, Eindhoven, Helmond, Northeast Brabant, and Tilburg. (3) In North Holland, the municipality of Haarlemmermeer is covered by its own business register. (4) In South Holland, business registers split the province into two regions (MRDH, covering the metropolitan regions of Rotterdam, The Hague, and Dordrecht, and 'leftover' South Holland, which includes Leiden and Hook of Holland).

[2] 'Employing organizations' can be commercial or non-commercial. E.g., *Politie* – the Dutch national police force – is not a commercial entity, but still employs people, holds a KVK registration, and is represented in the LISA data.

LISA and the KVK. FIRMBACKBONE provides data on the 2019-2022 LISA registers. My analysis focuses on the 2022 LISA register because it is the only year of available data both (1) after the primary period of the COVID-19 pandemic and (2) with a year of lagged data. My sample is restricted to establishments that are observed in both the 2021 and 2022 LISA registers, leaving me with a total of 1,433,393 establishments.

Table 1 displays proportions of observations in the 2022 LISA register by method of data collection, showing that a relatively small proportion of LISA's data is directly provided to LISA by establishments. Just over 20% of the establishments in the 2022 LISA register have response codes ≤ 20, implying that they provided data directly to LISA in 2022. Among these responsive establishments, over 93% provide data to the regional business surveys themselves (response codes 1 and 8), rather than through a broader firm representative (response codes 11 and 20). Of the nonresponsive establishments, over 97% of these establishments' employee headcounts are imputed using a (variant of a) procedure that imputes the establishment's 2022 headcounts based on its 2021 headcounts (response codes 50, 51, and 73); I discuss this procedure and a more accurate alternative in Sections 2.1 and 2.2 respectively. The remainder of the employee headcounts for establishments in the LISA register with response codes > 20 are either imputed using data provided to the KVK or through an *ad hoc* method. Though employee headcounts are imputed from the prior year's headcounts for unresponsive establishments, other characteristics of these unresponsive establishments – such as square meterage, founding year, sector, and facility type – are typically derived from other administrative data provided by Kadaster and KVK.

Table 2 provides descriptive statistics on continuous variables. I directly compute proportions of female and fulltime employees using employee headcounts. Most establishments in the LISA register are quite small, and the median establishment is a solo enterprise. This preponderance of solo enterprises is not unique to the LISA register or the Dutch context. E.g., in the United States, 81% of small businesses were nonemployer firms in 2017, comprising 17% of the American workforce (Conway et al. 2018).

Establishment square meterage and founding year are provided by Kadaster and the KVK, respectively. The median establishment's facility is 140 m^2 in size, around the size of a small office or a large Dutch townhome. Most establishments in the data are quite young,

Response Code	Description	% Establishments in 2022 LISA
1	Data directly from company, statement per branch, obtained in writing, online, or by telephone	19.066%
8	Data directly from company, temporarily no employees	0.011%
11	Data directly from company, statement per branch, through the intervention of a third party authorized by LISA	0.001%
20	Data directly from company total statement, to be allocated to branches	1.205%
30	Data from secondary source per branch (e.g., KVK [recent], annual report, website, press release)	1.407%
40	Data from secondary source total, to be allocated to branches	0.021%
50	Data increased from previous year, VR management module	72.55%
51	Data increased from previous year, other method	3.963%
60	Data imputed, VR management module	0.182%
61	Data imputed, other method	0.103%
72	Data estimated, guesswork	0.029%
73	Data taken directly from previous year	1.462%
76	Data taken directly from following year	0%

Note: Descriptions are based on LISA (2018).

Table 1: Proportions of Establishments by Response Type, 2022 LISA

with the median firm being founded in 2018 (just four years prior to the 2022 regional business surveys).

Descriptive information on categorical variables is provided in Appendix Figures A1, A2, and A3. Appendix Figure A1 shows that the geographic distribution of establishments covered by the LISA register roughly corresponds to what one would expect from relative population differences between the LISA regions. Appendix Figure A2 shows that the registered address of establishment facilities is by far most commonly a residential address. There are nearly 850,000 establishments in the LISA register that are registered to a residential address. For comparison, there are less than 100,000 establishments registered to addresses zoned for multiple functions, the next most common type of establishment address. This preponderance of residential registrations accords with the observation that most establishments in the LISA register are solo enterprises, as solo entrepreneurs typically register their enterprise at their home address. Beyond residential and multipurpose functions, establish-

	P0.5	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5	Mean	SD	N
Employees, 2021, LISA Imputation	1	1	1	1	1	1	2	5	12	60	120	4.853	49.72	1433393
Fulltime Employees, 2021, LISA Imputation	0	0	0	0	1	1	1	4	10	51	103	4.169	46.157	1433393
Employees, 2022, LISA Imputation	1	1	1	1	1	1	2	6	12	62	121	4.954	51.129	1433393
Employees, 2022, Random Forest Imputation	1	1	1	1	3	4	5	8	14	64	123	7.399	48.938	1433393
Fulltime Employees, 2022, LISA Imputation	0	0	0	0	1	1	1	5	10	53	106	4.273	47.862	1433393
Fulltime Employees, 2022, Random Forest Imputation	0	0	1	1	3	4	5	7	12	55	109	6.786	45.429	1433393
Female Employees, 2022, LISA Imputation	0	0	0	0	0	0	1	2	6	28	55	2.23	28.214	1433393
Female Employees, 2022, Random Forest Imputation	0	0	0	0	1	1	3	5	7	29	56	3.375	27.754	1433393
Proportion Employees Fulltime, 2022, LISA Imputation	0	0	0	0	1	1	1	1	1	1	1	0.814	0.366	1433239
Proportion Employees Fulltime, 2022, Random Forest Imputation	0	0	0.615	0.778	1	1	1	1	1	1	1	0.935	0.185	1433239
Proportion Employees Female, 2022, LISA Imputation	0	0	0	0	0	0	1	1	1	1	1	0.365	0.436	1433239
Proportion Employees Female, 2022, Random Forest Imputation	0	0	0	0	0.25	0.333	0.667	0.8	1	1	1	0.415	0.28	1433239
Establishment Surface Area, m^2	17	28	55	70	100	140	233	695	1700	9160.16	16481	625.924	4784.077	1433393
Year Founded	2000	2001	2007	2010	2015	2018	2020	2022	2022	2022	2023	2016.775	4.803	1433071

Note: Data sourced from the 2021 and 2022 LISA registers, as aggregated by FIRMBACKBONE.

Table 2: Descriptive Statistics

ment locations commonly serve industrial, office, or shopping functions. Appendix Figure A3 further supports the interpretation that many establishments in the data are home-registered solo enterprises. The most frequent one-letter KVK sector code observed in the data is code M, covering consultancy, research, and specialized business services; these are services that could be reasonably provided by a solo enterprise. Other common sectors include retail, construction, and healthcare.

2.1 LISA's Imputation Method

Table 1 shows that nearly all employee headcounts for establishments who are unresponsive in 2022 are imputed by transforming the establishment's employee headcounts from 2021.[3] LISA's (2018) algorithm for imputing this data amounts to multiplying each of the four employee headcounts (i.e., female/male fulltime/parttime) of unresponsive establishments by the average growth multiplier for that headcount amongst responsive firms in the same sector-size category. Specifically, LISA stratifies firms by pre-binned groups of the KVK's one-letter 2008 SBI sector codes (A-B, C-F, G-I, H-N, O-P, Q, and S-U) and by KVK establishment size classes (specifically 2-4, 5-49, and $\geq$ 50 fulltime employees). LISA then computes the average growth multiplier for the relevant headcount between year $t-1$ and t (i.e., the average ratio of the relevant headcount in year t to that headcount in $t-1$) across all responsive firms in each stratum. Finally, the main algorithm calls for the relevant headcount of all unresponsive establishments in that stratum in year $t-1$ to be multiplied

[3]These 2021 employee headcounts may have been imputed from the 2020 employee headcounts, which may in turn have been imputed from the 2019 employee headcounts, and so forth.

by the size-sector multiplier calculated for responsive establishments in the previous step to impute the unresponsive firms' relevant headcount in year t.

LISA's (2018) imputation algorithm is *per se* unreproducible because it provides for the possibility of several manual or sequential adjustments. First, if sample sizes within geographical regions are sufficiently high, then the size-sector strata may be further stratified by COROP region codes, though no rule exists for how high sample sizes within size-sector(-region) strata must be for this additional stratification to occur. Second, to balance rounding errors across establishments, if the total sum of the rounding adjustments made to the four employee headcounts for one establishment with imputed data is greater than 0.5 (less than -0.5), then LISA subtracts (adds) an employee from (to) the headcounts of the next establishment whose headcounts are imputed. However, the order of firms imputed is not stored in the data, so it is not possible to recover which firms' headcounts have been affected by this rounding correction. Third and finally, if size-sector growth multipliers are outside the range of $[0.9, 1.1]$, then LISA manually inspects whether outliers unduly influence the growth multiplier for that stratum, idiosyncratically removing those outliers if so. Because it is impossible to know when each of these manual adjustments have happened and how (often) they impact the 2022 LISA data, I ignore their prospect when assessing the performance of LISA's imputation algorithm, instead focusing on the algorithm's performance up to systematic reproducibility.

After replicating LISA's imputation method as closely as possible, I show in Table 3 that the LISA imputation method performs quite poorly. Specifically, I restrict the LISA data to the subset of establishments who both (1) were present in the 2021 LISA data and (2) responded to the 2022 regional business surveys (i.e., exhibiting response codes ≤ 20 in 2022), randomly dividing this subset in half using 50-50 sampling to create a hold-out dataset (see also Section 2.2). I then generate imputed values of several compositional characteristics of interest by applying the LISA imputation algorithm (up to systematic reproducibility), and in Column 1 of Table 3, I regress these LISA-imputed values on the real values of those characteristics provided by establishments for the 2022 regional business surveys. A regression slope of one would indicate that a given imputation-constructed measure maps one-to-one linearly with the true value of that measure in hold-out data, whereas a regression

	Same Variable, LISA Imputation	Same Variable, 2021 LISA	Same Variable, RF Imputation
# Employees	0.575	1.003	1.211
	(0.046)	(0.042)	(0.102)
Relative MSPE	1	0.133	0.37
Relative MAD	1	0.376	0.444
Adjusted R^2	0.874	0.919	0.799
N	57679	57679	57679
% Employees Female	0.815	0.826	0.996
	(0.002)	(0.002)	(0.002)
Relative MSPE	1	0.979	0.871
Relative MAD	1	0.973	1.171
Adjusted R^2	0.72	0.722	0.724
N	57564	57675	57679
% Employees Fulltime	0.455	0.429	0.831
	(0.007)	(0.006)	(0.006)
Relative MSPE	1	1.098	0.526
Relative MAD	1	1.04	0.825
Adjusted R^2	0.266	0.276	0.437
N	57564	57675	57679

Note: Analyses are performed on a randomly-selected half of the establishments who responded to the 2022 regional business surveys and are observed in the 2021 LISA register, specifically the half which are not used to train the RF model for the RF imputation. Ordinary least squares regression slopes are displayed above HC3 heteroskedasticity-robust standard errors in parentheses (Mackinnon & White 1985). Mean squared prediction error (MSPE) and mean absolute deviation (MAD) are respectively normalized by the MSPE and MAD of each LISA-imputed variable when used to predict the real value of that variable.

Table 3: Hold-Out Performance of Different Imputation Procedures

slope above (below) one indicates that on average, the imputation procedure systematically overestimates (underestimates) the true value of the measure. Column 1 of Table 3 shows that in the hold-out data, the LISA imputation algorithm underestimates the total number of employees by over 42%, the share of female employees by over 18%, and the share of fulltime employees by over 54% on average. Column 2 of Table 3 shows that the LISA imputation algorithm even performs poorly compared to a simple imputation strategy that just carries over employee headcounts from the previous year, which exhibits similar or better regression slopes and prediction errors than the LISA imputation algorithm.

If I would simply analyze the LISA data without any further modifications, then this

problem would compromise the validity of my analyses, which are focused on examining differences in the characteristics of responsive and unresponsive establishments. Specifically, if the LISA imputation algorithm systematically underestimates my establishment-level characteristics of interest, then after using LISA-imputed data for unresponsive establishments, the differences in those characteristics between responsive and unresponsive establishments will be negatively biased. To address this issue, I offer a new imputation algorithm in the next subsection.

2.2 A Novel Imputation Strategy

My proposed improvement to the LISA imputation algorithm is a random forest (RF) imputation algorithm that leverages the panel structure of the LISA data. My RF imputation is inspired by the same insight underlying LISA's imputation: an establishment's data from year $t-1$ can be informative about that establishment's data in year t. However, rather than parametrically computing size-sector growth trends based on a series of pre-binned size-sector strata, my RF imputation algorithm flexibly allows an establishment's characteristics in year $t-1$ to be combined (potentially nonlinearly) to generate predictions of that establishment's employee headcounts in year t. Random forest imputation is seeing increasing adoption as a missing data imputation method, with recent literature showing that random forest imputation yields considerable performance improvements over more classical forms of imputation (Twala 2009; Jerez et al. 2010; Garciarena & Santana 2017; Tang & Ishwaran 2017; Luo 2022).

The RF model used for this imputation is based on seven features. First, I use the four employee headcounts requested in the regional business surveys. This allows for a given employee headcount in year t to be constructed not just from its own value in year $t-1$, but from each of the four headcounts in year $t-1$; this also yields the advantage of effectively allowing firm size to influence predictions. Second, I use COROP region codes and one-letter 2008 SBI sector codes, allowing employee headcount predictions to account for the location and sector of the establishment. These are also recognized as important predictors in the LISA algorithm, though my algorithm does not pre-bin the SBI codes into groups, instead allowing any binning of SBI codes to be data-driven. Finally, I use the establishment's survey

type in the previous year. This is because an establishment's data in year $t-1$ may be either authentically reported or imputed by LISA, and different data types may be heterogeneously informative about an establishment's true data in year t.

The RF imputation algorithm proceeds as follows. First, I randomly split rows of the responsive firms in the 2022 LISA data into training data and testing data with 50-50 sampling. Second, for each of the four employee headcounts, I train RF regression models on the relevant employee headcount in 2022. I specifically use the `ranger` package in R with the features described in the previous paragraph (see Wright & Ziegler 2017).[4] Third, I predict the relevant employee headcount for all nonresponsive firms using the RF model trained in the second step. Fourth and finally, I simply round the relevant predicted employee headcount to ensure an integer value.

My RF imputation method yields substantial performance improvements over the LISA imputation algorithm. The third column of Table 3 shows results for the RF imputation algorithm when applied to the hold-out sample of responsive firms in the 2022 LISA register. Compared to the LISA imputation procedure, my RF imputation method substantially mitigates the systematic underestimation of all three outcomes of interest to my analysis, with regression slopes between predicted and real values of the outcomes being much closer to one. The LISA algorithm also generally delivers better predictive performance, with the MSPE yielded by the RF imputation algorithm being 20-66% smaller than that yielded by the LISA imputation procedure. Due to this better performance, I use RF-imputed outcomes rather than LISA-imputed outcomes for my analyses of differences in continuous outcomes between responsive and unresponsive establishments.

3 Methods

One can investigate two different research questions concerning sampling biases in business surveys by separately analyzing *unconditional* and *contact-conditional* differences in characteristics between responsive and unresponsive firms.

[4]Default settings are applied, with the exception that I run random forest regression models with 5000 trees.

RQ1: How are characteristics expected to differ between establishments who do and do not respond to business surveys?

RQ2: *Conditional on being contacted for a business survey*, how are the characteristics of responsive and unresponsive establishments expected to differ?

The research question of interest depends on whether one is interested in (1) the representativeness of prior business surveys or (2) improving the design of future business surveys. If one only cares about (1), then it is sufficient to simply answer RQ1 and learn the expected differences in characteristics between responsive and unresponsive establishments. It is not necessary to know whether such differences emerge from the sampling process or from differences in responsiveness if all one cares about is whether they can draw generalizable conclusions from the business survey data they already have. However, the source of this sampling bias is important if one cares about (2). In this case, RQ2 can be more practically reformulated as: "If I send a business survey to a set of establishments, which of those establishments are most likely to respond?"

Insights on RQ2 can inform optimal survey design. If subsets of firms are known to respond more (less) frequently to business surveys when contacted, then one can decrease (increase) the sampling probability for these subsets to improve the chances that the final sample will be representative of the relevant population of establishments. Conversely, if one is resource-constrained and concerned about low sample sizes, then the most responsive subsets of establishments can be targeted to maximize response rates.

RQ1 and RQ2 can respectively be assessed by estimating unconditional and conditional differences in characteristics between responsive and unresponsive establishments. Let $R_i \in \{0, 1\}$ represent whether establishment i provides a response to the business survey, and let $Y_{i,j}(R)$ be a potential outcome of establishment i's j'th characteristic. Then the quantities of interest for assessing RQ1 are the *unconditional* differences in characteristics

$$\delta_{U,j} = \mathbb{E}\left[Y_{i,j}(1) - Y_{i,j}(0)\right], \tag{1}$$

which can be estimated using simple bivariate linear regression. However, letting $C_i \in \{0, 1\}$ represent whether establishment i is contacted for the business survey, the quantities of

interest for assessing RQ2 are the *contact-conditional* differences in characteristics

$$\delta_{C,j} = \mathbb{E}\left[Y_{i,j}(1) - Y_{i,j}(0) \mid C_i = c\right]. \tag{2}$$

Though $\delta_{C,j}$ is a target parameter, it is not *directly* estimable in the LISA data. Estimating each $\delta_{C,j}$ would be trivial if I had data on which establishments were contacted for the regional business surveys. However, this data is not available in the LISA register. Local registries delete the lists of contacted establishments after one month, so LISA cannot aggregate the contact status of each establishment. This prevents direct conditioning on contact status in the LISA data.

However, unbiased estimates of $\delta_{C,j}$ can be recovered by leveraging the sampling process of the regional business surveys. Conditional on a number of variables used for sampling exceptions and stratification, establishments are contacted through random sampling. After conditioning for these stratification and exception variables, remaining differences in characteristics between establishments that do and do not respond to the regional business surveys should thus have no systematic relationship with contact probability, instead solely reflecting differences in business survey responsiveness. Regression specifications that control for the full set of stratification and exception variables can therefore recover unbiased estimates of contact-conditional differences between responsive and unresponsive establishments. I.e., letting X_i represent the stratification and exception variables for establishment i, random sampling should render

$$\hat{\delta}_{C,j} = \mathbb{E}\left[Y_{i,j}(1) - Y_{i,j}(0) \mid X_i = x\right] \tag{3}$$

an unbiased estimator for $\delta_{C,j}$.

3.1 Estimating Conditional Response Rates

LISA's regional business surveys are stratified on two variables. First, LISA systematically varies contact probabilities based on establishments' KVK size classification. The KVK size classifications are based on the number of fulltime employees an establishment had in the

prior year, with bins of 1, 2-9, 10-99, and ≥ 100 employees. Establishments in the lowest two bins have a contact probability of 25%, those in the bin of establishments with 10-99 fulltime employees have a contact probability of 50%, and all establishments with ≥ 100 employees are supposed to be contacted every year (LISA 2018). Second, these contact probabilities are stratified by LISA region, indicating the business register responsible for administering each establishment's regional business survey. Each business register also idiosyncratically decides procedures and temporal/monetary investment for the regional business surveys.

There are also three exception variables I control for, upon which systematic or idiosyncratic deviations from LISA's stratified random sampling protocol can arise. First, all firms with known email addresses are contacted by email each year due to the low cost associated with sending emails. I thus control for a dummy indicating whether the establishment responded by email to the 2021 regional business surveys. Second, unresponsive establishments are systematically recontacted manually if they have five or more employees in the previous year's register. Because this count can be composed of both fulltime and parttime employees, this is distinct from the KVK size classes (which are only based on fulltime employee headcounts). I thus control for a dummy indicating whether each establishment has five or more total employees in the 2021 LISA register. Third and finally, firms may be sampled differently by regional LISA branches based on their response type in the previous year. E.g., an establishment known to have responded to the prior year's regional business survey may be contacted for the current year's survey if that establishment's branch is struggling to attain high enough response rates to meet LISA obligations. I thus control for a factor variable indicating the response code of each establishment in the 2021 LISA register.

These exception and stratification variables are important for explaining response probability. Figure 1 shows how response probabilities differ based on contact determinants.[5] Having an active email address on file substantially increases response rates, with establishments who responded by email in 2021 being 15.9 percentage points more likely to respond to the regional business surveys in 2022. As expected from LISA's sampling protocol, response rates also systematically increase with size classes, with the largest firms – those with ≥ 100 fulltime employees in 2021 – being 27.4 percentage points more likely to respond to the

[5] A table version of this figure can be found in Appendix Table A1.

2022 regional business surveys. Additionally, response rates are quite heterogeneous across regions, with the lowest-response region (Eindhoven) exhibiting response rates 17 percentage points lower than the highest-response region (Gelderland).

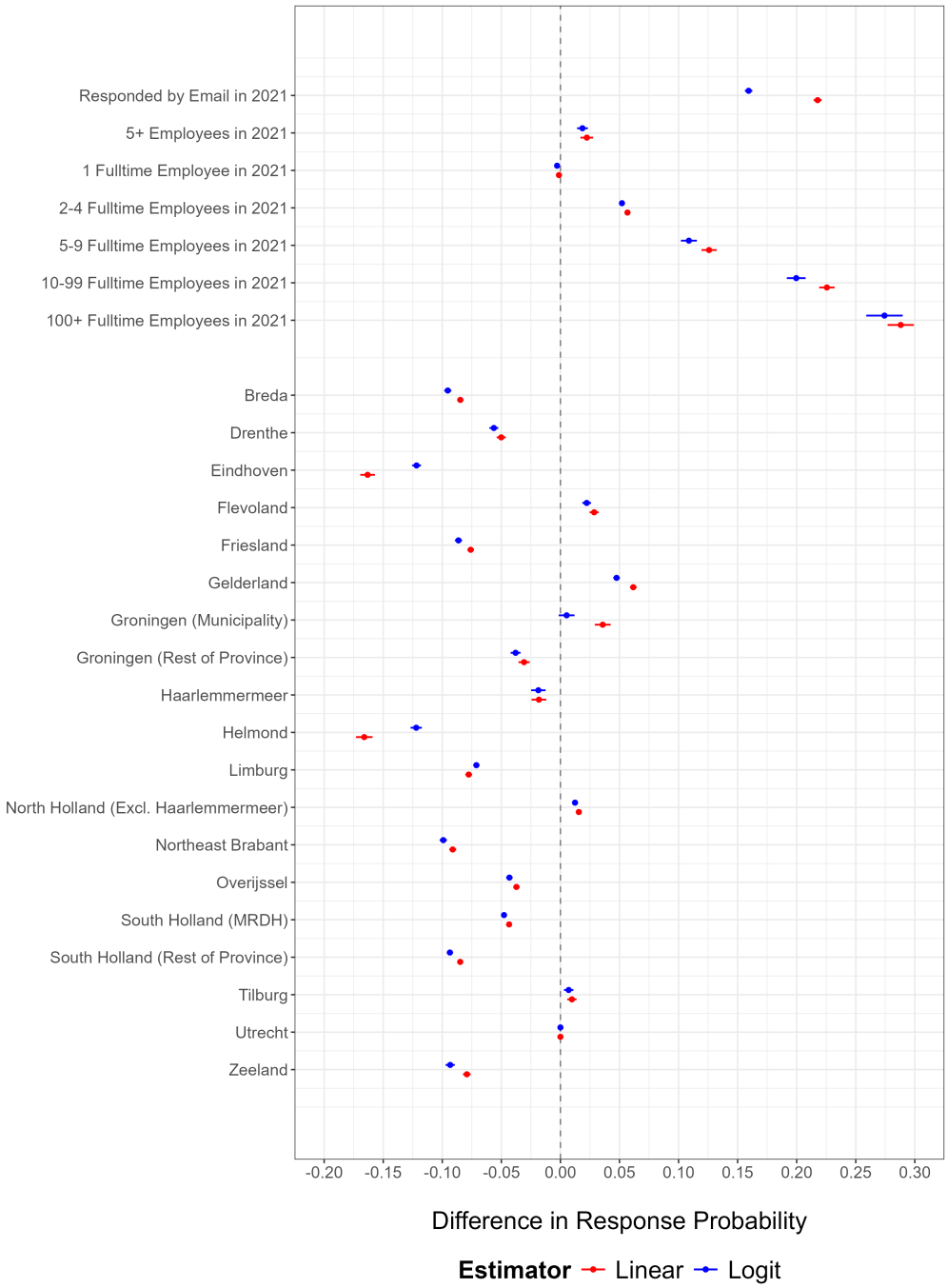
The estimates in Figure 1 arise from models that control for each establishment's responsiveness to the 2021 regional business surveys. Specifically, though not displayed in Figure 1, these models control for each establishment's response code in the 2021 LISA register. Differences in response rates observed in Figure 1 should thus primarily reflect differences in contact probability rather than differences in responsiveness. After controlling for all of the exception variables and the stratification variables discussed in this section, differences in characteristics between responsive and unresponsive firms should therefore not reflect correlations with contact probability, instead solely reflecting differences in responsiveness.

3.2 Outliers and Robust Regression

Because of extreme skew in the distributions of establishments' unbounded continuous characteristics, care must be applied to ensure that analyses on such characteristics are robust to outliers. Table 2 shows that across all upwardly-unbounded continuous characteristics in my sample (specifically employee headcount variables and square meterage), there is nearly as much variation between the 99th and 99.5th percentiles as there is between the 0.5th and 99th percentiles.[6] (Conditional) mean differences between responsive and unresponsive establishments are likely quite sensitive to these high outliers.

I thus approach analyses of unbounded continuous characteristics differently than analyses of categorical or bounded continuous characteristics. Given that all independent variables in my analysis are categorical, differences in categorical or bounded continuous characteristics can be safely examined using ordinary least squares regression. However, I examine differences in unbounded continuous characteristics using a robust regression approach. I specifically employ the M-estimator proposed by Huber (1964) and implemented by Venables & Ripley (2002) in the `MASS` R package, whose algorithms and default parameters reflect similar applications in other programming languages (e.g., Verardi & Groux 2009).

[6]The difference in variations within these quantile groupings does not exceed 10% for any upwardly unbounded continuous characteristic in my sample.



Note: Estimates are generated by regressing a dummy indicating whether the establishment responded to the 2022 regional business surveys (response code ≤ 20) on all of the indicators displayed in the figure's rows. All models control for a factor variable indicating the response code of each establishment in the 2021 LISA register. Estimates are displayed alongside 95% confidence intervals based on HC3 heteroskedasticity-robust standard errors (Mackinnon & White 1985). Linear regression estimates are based on ordinary least squares regression. Logistic regression estimates are converted to average marginal effects on linear response probability using the `marginaleffects` package in R (Arel-Bundock, Greifer, & Heiss 2024).

Figure 1: Determinants of Contact Probability and Response Rates

M-estimation is a recommended robust regression method in settings such as mine, where the independent variables of interest are categorical (Maronna & Yohai 2000; Bramati & Croux 2007; Verardi & Groux 2009).[7] When standardizing coefficients on continuous characteristics, I also follow Menkveld et al. (2024) and use interquartile ranges (IQRs) rather than standard deviations, similarly accommodating the fat-tailed distributions of the unbounded continuous characteristics.

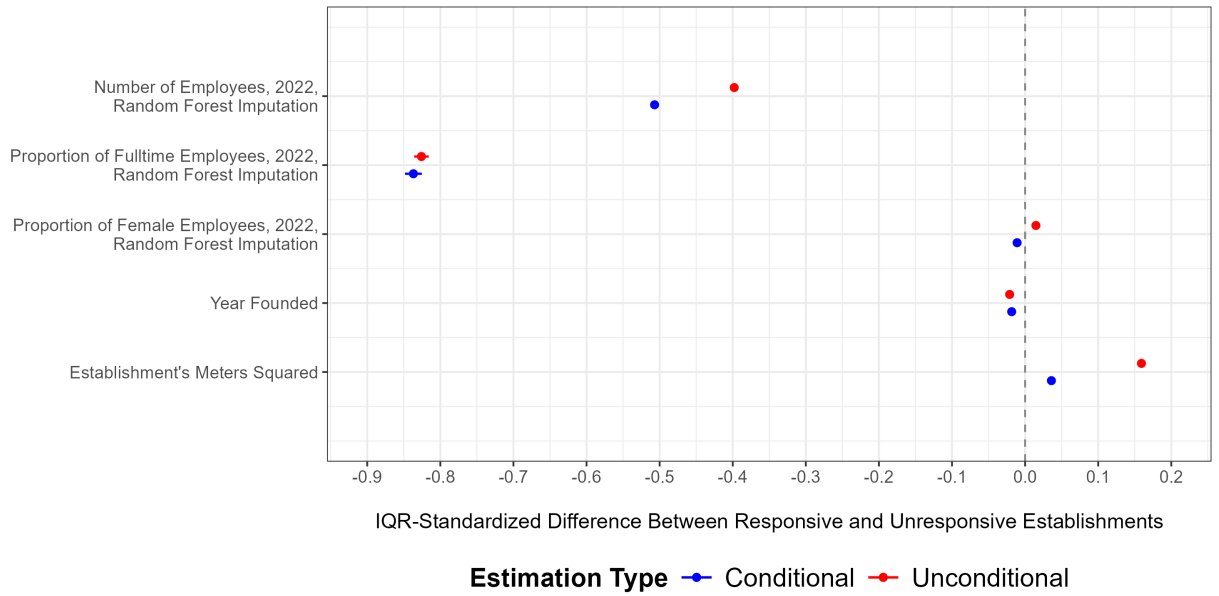
4 Results

4.1 Continuous Characteristics

Figure 2 shows the IQR-standardized (robust) regression estimates for differences in continuous characteristics between responsive and unresponsive establishments, both unconditionally and conditional on contact probability. The most striking differences are those involving employee headcounts and proportions of fulltime employees. Specifically, responsive firms appear to employ fewer people, and the people that are employed at responsive firms are less likely to work there fulltime.

Though the unconditional results concerning employee headcounts may at first appear to contradict the findings in Figure 1, which shows that larger establishments respond more to the regional business surveys because they are contacted more frequently, this discrepancy arises for two natural reasons. First, the models used to produce the employee headcount difference estimates in Figure 2 are *robust* regression models, which downweight (or even zero-weight) large outliers in the estimation procedure. The fact that the largest establishments are more likely to be contacted for the survey is perfectly compatible with the finding that within an effective subsample of representatively small establishments, those with fewer employees are more likely to respond. Second, because the KVK establishment size classes used for sampling stratification are based on *fulltime* employee headcounts, increases in *total*

[7] Though MM-estimators of the form proposed by Yohai (1987) are a common alternative, M-estimators are preferable in this setting. Because all exposure variables are categorical, there are no leverage points, so MM-estimators offer no additional robustness against leverage while simultaneously increasing the probability of estimating regressions on collinear subsamples (Maronna & Yohai 2000; Bramati & Croux 2007). Even when robust regressions include some continuous predictors, M-estimators are typically recommended for any categorical predictors (Maronna & Yohai 2000; Bramati & Croux 2007; Verardi & Groux 2009).



Note: Estimates are generated by regressing the row's outcome on a dummy indicating whether the establishment responded to the 2022 regional business surveys (response code ≤ 20). Regression coefficients are standardized by the interquartile range of the row's outcome in the 2022 LISA sample of responsive establishments and displayed alongside 99.5% confidence intervals, based on HC3 heteroskedasticity-robust standard errors (MacKinnon & White 1985). Regression results are based on ordinary least squares estimation for the proportions of fulltime and female employees, and are based on M-estimation for employee headcounts, establishment age, and establishment square meterage (Huber 1964; Venables & Ripley 2002). Conditional estimates are based on models that control for KVK size classifications, LISA regions, response by email in the 2021 LISA register, response code in the 2021 LISA register, and a dummy indicating whether the establishment is recorded as having ≥ 5 employees in the 2021 LISA register.

Figure 2: Responsive vs. Unresponsive Establishment Gaps, Continuous Characteristics

employee headcounts do not yield as much of an increase in contact probability. Appendix Table A1 shows that the difference in response rates associated with having 5-9 *fulltime* employees in the prior year is over five times larger than the difference associated with having ≥ 5 *total* employees in the prior year. It is also possible that the positive coefficient on the ≥ 5 total employees indicator in Figure 1 is driven by the largest establishments, who receive little to no weight in the robust regression estimation.

To give a sense of the practical magnitude of the differences between responsive and unresponsive firms, Table 4 displays the (robust) regression estimates without IQR standardization. Responsive establishments employ two fewer people than unresponsive establishments, and have a rate of fulltime employment 15 percentage points lower than unresponsive establishments. Though the high power afforded by the LISA register yields statistical significance for all estimates, the remainder of the estimates are quite small in practice. The proportion

	Number of Employees	Proportion of Fulltime Employees	Proportion of Female Employees	Year Founded	Establishment's Meters Squared
Panel A: Unconditional Results					
Responsive Establishment	-1.99	-0.15	0.014	-0.169	47.123
	(0.005)	(0.001)	(0.001)	(0.009)	(0.301)
N	1433393	1433239	1433239	1433071	1433393
Panel B: Conditional Results					
Responsive Establishment	-2.534	-0.152	-0.01	-0.147	10.654
	(0.003)	(0.001)	(0.001)	(0.011)	(0.291)
N	1433393	1433239	1433239	1433071	1433393
Model	M-Estimator	OLS	OLS	M-Estimator	M-Estimator

Note: Estimates are generated by regressing the column's outcome on a dummy indicating whether the establishment responded to the 2022 regional business surveys (response code ≤ 20). Regression coefficients are displayed alongside 95% confidence intervals in parentheses, based on HC3 heteroskedasticity-robust standard errors (MacKinnon & White 1985). Conditional estimates are based on models that control for KVK size classifications, LISA regions, response by email in the 2021 LISA register, response code in the 2021 LISA register, and a dummy indicating whether the establishment is recorded as having ≥ 5 employees in the 2021 LISA register.

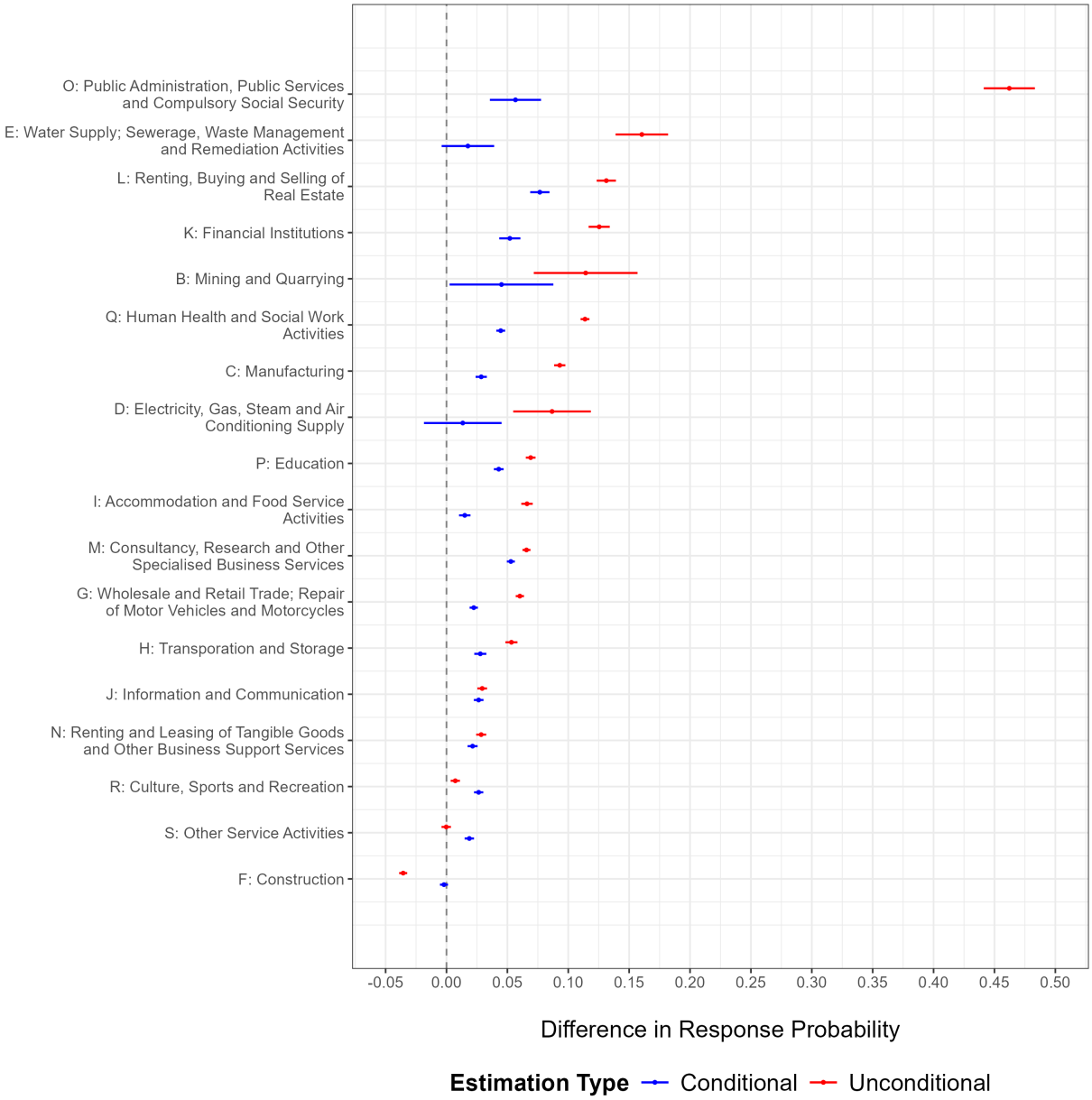
Table 4: Responsive vs. Unresponsive Establishment Gaps, Continuous Characteristics

of female employees only differs by around one percentage point between responsive and unresponsive establishments. The difference in ages between responsive and unresponsive firms can be significantly bounded beneath one year, the smallest measurable difference in firm age in this data (see Fitzgerald 2025). Though responsive establishments have 47 m^2 more physical space than unresponsive establishments before conditioning on contact probability, this difference collapses to less than 11 m^2 after such conditioning.

4.2 Sectors

Figure 3 shows average differences in response rates by one-letter 2008 SBI sector code, both unconditionally and conditional on contact probability.[8] There is large heterogeneity in unconditional response rates between sectors, with establishments in the most response-overrepresented sector (public administration) being nearly 50 percentage points more likely to respond to regional business surveys than those in the most response-underrepresented sector (construction). The most consistently and robustly overrepresented sectors amongst responsive establishments tend to be public-sector and white-collar sectors, such as public utilities, real estate, financial institutions, and healthcare.

[8] A table version of this figure is provided in Appendix Table A2.



Note: Estimates are generated by regressing a dummy indicating whether the establishment responded to the 2022 regional business surveys (response code ≤ 20) on dummy variables indicating whether the establishment belongs to each of these one-letter 2008 SBI sector codes. The omitted category is "A: Agriculture, Forestry, and Fishing." Ordinary least squares regression coefficients are displayed alongside 95% confidence intervals, based on HC3 heteroskedasticity-robust standard errors (MacKinnon & White 1985). Conditional estimates are based on models that control for KVK size classifications, LISA regions, response by email in the 2021 LISA register, response code in the 2021 LISA register, and a dummy indicating whether the establishment is recorded as having ≥ 5 employees in the 2021 LISA register.

Figure 3: Differences in Response Rates Between Sectors

Much of the differences in sectoral response rates can be traced not to the responsiveness of establishments, but to contact probability. E.g., based on unconditional response rates,

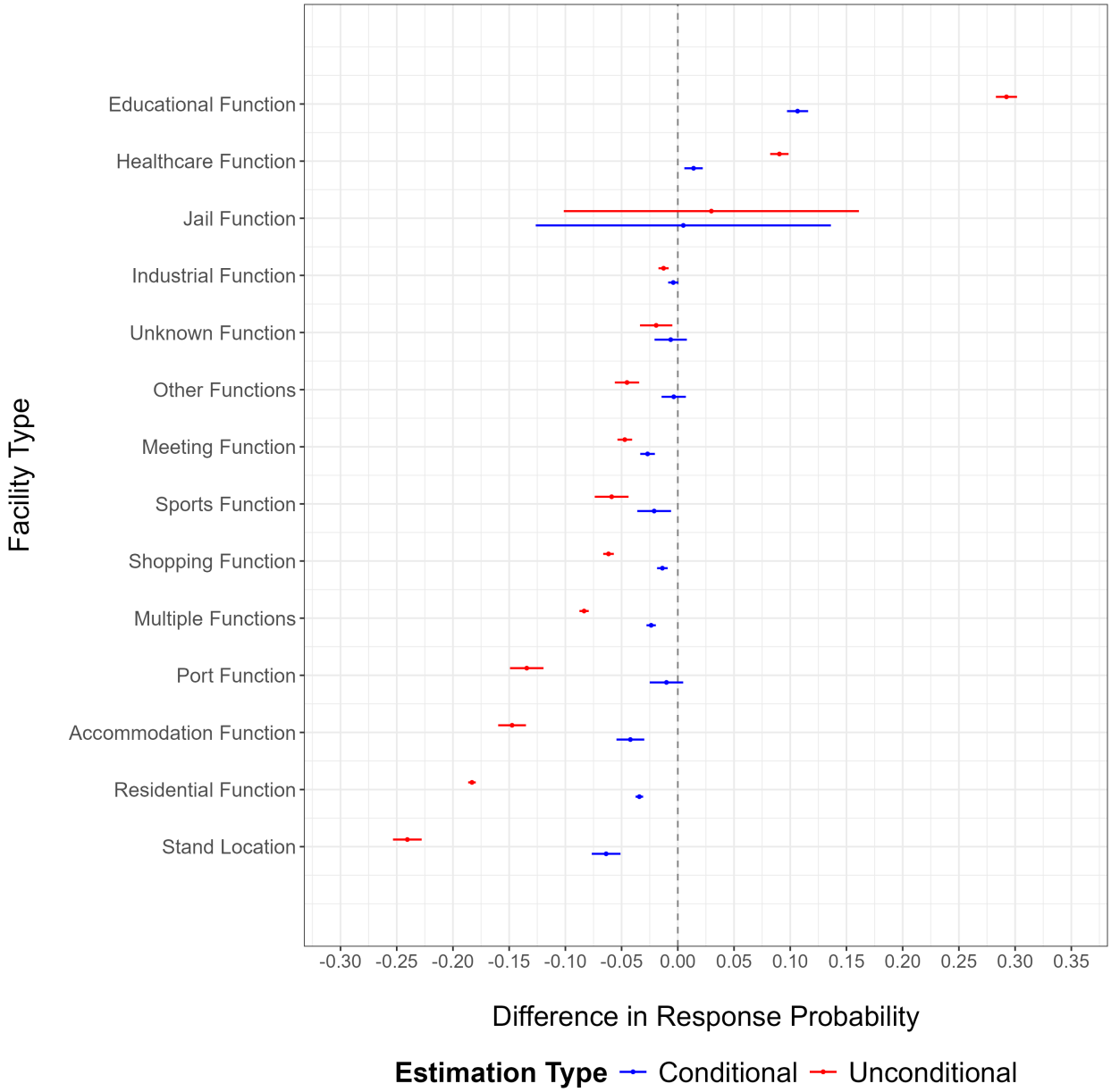
public utilities appear to be the second-most overrepresented sector among responsive firms. However, after controlling for contact probability, the estimated advantage in response rates for public utilities decreases by over 82%, and is no longer statistically significantly different from zero. In fact, Table A2 shows that 11 of the 19 sectoral response rate advantages attenuate by more than half after controlling for contact probability. The maximum difference in sectoral response rates after conditioning on contact probability — now between real estate and construction — is just eight percentage points. This implies that the most overrepresented sectors in business surveys derive much of their overrepresentation not because establishments in those sectors are more likely to respond to surveys, but instead because those establishments are more likely to be contacted for business surveys. However, much heterogeneity in response rates remains even after controlling for contact probability, suggesting that there are meaningful differences in responsiveness between sectors.

4.3 Facility Types

Figure 4 shows differences in response rates by establishment facility type, both unconditionally and conditional on contact probability.[9] There is considerable heterogeneity in response rates across establishment facility types. Both before and after conditioning on contact probability, educational institutions and healthcare facilities exhibit significantly higher response rates than other facility types. The lowest response rates are observed in establishments whose facilities serve an accommodation function (e.g., hotels), stand locations (e.g., food carts), and establishments housed in residences.

The low response rates documented amongst establishments registered at a residential address is particularly troubling for the representativeness of business surveys. As discussed in Section 2, establishments registered at facilities with a residential function are far and away the most common type of establishment in the LISA register. This reflects the fact that the median establishment is a solo enterprise, as the vast majority of solo enterprises are registered at the entrepreneur's home address. However, Appendix Table A3 shows that establishments registered at a residential address are over 18 percentage points less likely to respond to LISA's regional business surveys than the average office.

[9] A table version of this figure is provided in Appendix Table A3.



Note: Estimates are generated by regressing a dummy indicating whether the establishment responded to the 2022 regional business surveys (response code $\leq$ 20) on dummy variables indicating whether the establishment's facility is of the type in the estimate's row. The omitted category is "Office Function." Ordinary least squares regression coefficients are displayed alongside 95% confidence intervals, based on HC3 heteroskedasticity-robust standard errors (MacKinnon & White 1985). Conditional estimates are based on models that control for KVK size classifications, LISA regions, response by email in the 2021 LISA register, response code in the 2021 LISA register, and a dummy indicating whether the establishment is recorded as having $\geq$ 5 employees in the 2021 LISA register.

Figure 4: Differences in Response Rates Between Establishment Facility Types

As with differences in sectoral response rates, most of the differences in response rates between establishment facility types can be traced back to differences in contact probability

rather than differences in responsiveness. Both unconditionally and conditional on contact probability, educational institutions are the most overrepresented facility type amongst responsive establishments, whereas stand locations are the most underrepresented. Appendix Table A3 shows that the estimated difference in response rates between these facility types declines from over 53 percentage points to just over 24 percentage points after controlling for contact probability. Compared to an average office, the response rate deficit for establishments registered at a residential address declines from over 18 percentage points to just over three percentage points after controlling for contact determinants. Most other response rate gaps between establishment facility types similarly attenuate after conditioning on contact probability. This suggests that unconditionally unresponsive establishments – such as solo enterprises registered to residential addresses – are less represented among responsive firms primarily because they are less frequently contacted to take regional business surveys. However, as in Section 4.2's analysis of sectoral response rates, there is considerable heterogeneity in response rates by facility type even after controlling for contact determinants, implying meaningful heterogeneity in responsiveness between establishment types.

5 Conclusion

This paper shows that there are significant compositional, sectoral, and occupational differences between the establishments which do and do not respond to business surveys. Responsive establishments employ fewer people, and the people who *are* employed at these responsive establishments are less likely to work fulltime. If one were to examine only a dataset of responsive establishments – as occurs in analyses of most business surveys – then public-sector and white-collar occupations would be considerably overrepresented. Though solo enterprises registered to an entrepreneur's personal home are the most common kind of business in the Dutch economy, such establishments are one of the most underrepresented types of businesses amongst establishments that respond to Dutch regional business surveys. However, many of the sectoral and occupational differences in response rates lose the majority of their magnitude after controlling for contact probability. This suggests that the bulk of unrepresentativeness in business surveys arises due to survey sampling rather than selective

nonresponse, though there is still considerable selection in nonresponse even after controlling for contact probability. These results imply that though many prior business surveys may be meaningfully unrepresentative, there is substantial potential for improving business surveys by changing sampling design.

This paper's results on unconditional response rates will generalize differently from its results on contact-conditional response rates. Before conditioning on contact determinants, differences in characteristics between responsive and unresponsive establishments largely reflect differences in contact probability rather than differences in responsiveness. These differences in unconditional response rates will only generalize to business surveys that sample establishments similarly to LISA's regional business surveys. That said, many features of LISA's sampling procedure – targeting larger establishments with higher probability, contacting establishments by email when possible, etc. – are quite common in business surveys. However, after conditioning on contact probability, differences in characteristics between responsive and unresponsive establishments are statistically independent of the idiosyncratic sampling features of LISA's regional business surveys. These differences in contact-conditional response rates are informative of business' fundamental behavioral tendencies, and provide insight into which businesses respond to surveys in general.

As discussed in Section 3, a researcher's interest in this paper's results on unconditional and contact-conditional response rates will depend on whether they are interested in appraising the representativeness of existing business survey data or improving the representativeness of future business surveys. Provided that an existing business survey samples establishments similarly to LISA's regional business surveys, this paper's results on unconditional response rates can be informative about how representative that survey's sample likely is. One does not need to know whether differences in characteristics between responsive and unresponsive establishments are attributable to contact probability or responsiveness if all they care about is *whether*, and not *why*, a business survey is unrepresentative. However, differences in contact-conditional response rates can inform researchers about which firms are more or less likely to respond to a survey conditional on being contacted. Surveyors who know which firms are more likely to respond to a survey if contacted can improve response rates or representativeness by using these insights for targeted sampling. Resource-constrained sur-

veyors concerned about low sample sizes can target more contact-conditionally responsive establishments to increase response rates. Surveyors can also achieve more representative samples of businesses by oversampling establishments with low contact-conditional response rates and undersampling establishments with high contact-conditional response rates. Given the extent to which controlling for contact determinants attenuates response rate gaps between establishment types in Section 4, these adjustments in survey design will likely have a strong 'bite', yielding significant changes in response rates and sample composition. Leveraging this paper's findings, these survey design adjustments can thus meaningfully improve the power and representativeness of business surveys to generate more productive insights on businesses, markets, and the economy.

References

(2018, October). https://www.lisa.nl/include/LISA_Handboek_versie_oktober_2018.pdf

Altig, D., Barrero, J. M., Bloom, N., Davis, S. J., Meyer, B., & Parker, N. (2022). Surveying business uncertainty. *Journal of Econometrics*, *231*(1), 282–303. https://doi.org/10.1016/j.jeconom.2020.03.021

Arel-Bundock, V., Greifer, N., & Heiss, A. (2024). How to interpret statistical models using marginaleffects for R and Python. *Journal of Statistical Software*, *111*(9), 1–32. https://doi.org/10.18637/jss.v111.i09

Baker, H. K., Singleton, J. C., & Veit, E. T. (2011). *Survey research in corporate finance: Bridging the gap between theory and practice*. Oxford University Press.

Bianchi, A., & Biffignandi, S. (2017). Representativeness in panel surveys. *Mathematical Population Studies*, *24*(2), 126–143. https://doi.org/10.1080/08898480.2016.1271650

Bramati, M. C., & Croux, C. (2007). Robust estimators for the fixed effects panel data model. *The Econometrics Journal*, *10*(3), 521–540. https://doi.org/10.1111/j.1368-423x.2007.00220.x

Clar, M., Duque, J.-C., & Moreno, R. (2007). Forecasting business and consumer surveys indicators – A time-series models competition. *Applied Economics*, *39*(20), 2565–2580. https://doi.org/10.1080/00036840600690272

Collins, D. (2001). The relationship between business confidence surveys and stock market performance. *Investment Analysts Journal*, *30*(54), 9–17. https://doi.org/10.1080/10293523.2001.11082428

Conway, M., Klein, J., Robles, B., & Weindorf, M. (2018, December). 2018 report on nonemployer firms: Findings from the 2017 Small Business Credit Survey. https://doi.org/10.55350/sbcs-20181212

Cornesse, C., & Bosnjak, M. (2018). Is there an association between survey characteristics and representativeness? A meta-analysis. *Survey Research Methods*, *12*(1), 1–13. https://doi.org/10.18148/srm/2018.v12i1.7205

Fitzgerald, J. (2025). The need for equivalence testing in economics. *MetaArXiv*. https://doi.org/10.31222/osf.io/d7sqr_v1

Hansson, J., Jansson, P., & Löf, M. (2005). Business survey data: Do they help in forecasting GDP growth? *International Journal of Forecasting*, *21*(2), 377–389. https://doi.org/10.1016/j.ijforecast.2004.11.003

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2), 105–115. https://doi.org/10.1016/j.artmed.2010.05.002

Karanja, E., Sharma, A., & Salama, I. (2019). What does MIS survey research reveal about diversity and representativeness in the MIS field? A content analysis approach. *Scientometrics*, *122*(3), 1583–1628. https://doi.org/10.1007/s11192-019-03331-5

Kent Baker, H., & Mukherjee, T. K. (2007). Survey research in finance: Views from journal editors. *International Journal of Managerial Finance*, *3*(1), 11–25. https://doi.org/10.1108/17439130710721635

Klein, L. R., & Özmucur, S. (2010). The use of consumer and business surveys in forecasting. *Economic Modelling*, *27*(6), 1453–1462. https://doi.org/10.1016/j.econmod.2010.07.005

Luo, Y. (2022). Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, *23*(1). https://doi.org/10.1093/bib/bbab489

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3), 305–325. https://doi.org/10.1016/0304-4076(85)90158-7

Maronna, R. A., & Yohai, V. J. (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, *89*(1–2), 197–214. https://doi.org/10.1016/s0378-3758(99)00208-6

Menkveld, A. J., & et al. (2024). Nonstandard errors. *The Journal of Finance*, *79*(3), 2339–2390. https://doi.org/10.1111/jofi.13337

Moore, J. C., Durrant, G. B., & Smith, P. W. (2018). Data set representativeness during data collection in three UK social surveys: Generalizability and the effects of auxiliary covariate choice. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181*(1), 229–248. https://doi.org/10.1111/rssa.12256

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., & Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, *80*(3), 382–399. https://doi.org/10.1111/j.1751-5823.2012.00189.x

Shlomo, N., Skinner, C., & Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, *142*(1), 201–211. https://doi.org/10.1016/j.jspi.2011.07.008

Snijkers, G., Haraldsen, G., Jones, J., & Willimack, D. K. (2013). *Designing and conducting business surveys*. John Wiley & Sons.

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining*, *10*(6), 363–377. https://doi.org/10.1002/sam.11348

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, *23*(5), 373–405. https://doi.org/10.1080/08839510902872223

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer New York.

Verardi, V., & Croux, C. (2009). Robust regression in Stata. *The Stata Journal: Promoting communications on statistics and Stata*, *9*(3), 439–453. https://doi.org/10.1177/1536867x0900900306

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, *15*(2), 642–656. https://doi.org/10.1214/aos/1176350366

Zimmermann, K. F. (1999). Analysis of business surveys. In *Handbook of applied econometrics volume 2: Microeconomics* (pp. 369–399). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1111/b.9780631216339.1999.00010.x

Appendix

	Response Probability, Linear Model	Response Probability, Logit Model
Responded by Email in 2021	0.218	0.159
	(0.002)	(0.002)
5+ Employees in 2021	0.022	0.019
	(0.003)	(0.002)
1 Fulltime Employee in 2021	-0.001	-0.003
	(0.001)	(0.001)
2-4 Fulltime Employees in 2021	0.057	0.052
	(0.001)	(0.001)
5-9 Fulltime Employees in 2021	0.126	0.109
	(0.003)	(0.003)
10-99 Fulltime Employees in 2021	0.226	0.2
	(0.003)	(0.004)
100+ Fulltime Employees in 2021	0.288	0.274
	(0.006)	(0.008)
Breda	-0.085	-0.095
	(0.001)	(0.002)
Drenthe	-0.05	-0.056
	(0.002)	(0.002)
Eindhoven	-0.163	-0.122
	(0.003)	(0.002)
Flevoland	0.029	0.022
	(0.002)	(0.002)
Friesland	-0.076	-0.086
	(0.001)	(0.002)
Gelderland	0.062	0.048
	(0.001)	(0.001)
Groningen (Municipality)	0.036	0.005
	(0.003)	(0.003)
Groningen (Rest of Province)	-0.031	-0.038
	(0.002)	(0.002)
Haarlemmermeer	-0.018	-0.019
	(0.003)	(0.003)
Helmond	-0.166	-0.122
	(0.004)	(0.002)
Limburg	-0.078	-0.071
	(0.001)	(0.001)
North Holland (Excl. Haarlemmermeer)	0.016	0.012
	(0.001)	(0.001)
Northeast Brabant	-0.091	-0.099
	(0.002)	(0.002)
Overijssel	-0.037	-0.043
	(0.001)	(0.001)
South Holland (MRDH)	-0.043	-0.048
	(0.001)	(0.001)
South Holland (Rest of Province)	-0.085	-0.094
	(0.001)	(0.001)
Tilburg	0.01	0.007
	(0.002)	(0.002)
Utrecht	0	0
	(0)	(0)
Zeeland	-0.079	-0.093
	(0.002)	(0.002)
N	1433393	1433393

Note: Estimates are displayed alongside HC3 heteroskedasticity-robust standard errors in parentheses (Mackinnon & White 1985). Logistic regression estimates are converted to average marginal effects on linear response probability using the `marginaleffects` package in R (Arel-Bundock, Greifer, & Heiss 2024). All models control for a factor variable indicating the response code of each establishment in the 2021 LISA register.

Table A1: Determinants of Contact Probability and Response Rates

SBI Sector	Response Probability, Unconditional	Response Probability, Conditional
B: Mining and Quarrying	0.114	0.045
	(0.022)	(0.022)
C: Manufacturing	0.093	0.029
	(0.002)	(0.002)
D: Electricity, Gas, Steam and Air	0.087	0.013
Conditioning Supply	(0.016)	(0.016)
E: Water Supply; Sewerage, Waste Management	0.16	0.018
and Remediation Activities	(0.011)	(0.011)
F: Construction	-0.036	-0.002
	(0.002)	(0.002)
G: Wholesale and Retail Trade; Repair	0.06	0.022
of Motor Vehicles and Motorcycles	(0.002)	(0.002)
H: Transporation and Storage	0.053	0.028
	(0.003)	(0.003)
I: Accommodation and Food Service	0.066	0.015
Activities	(0.002)	(0.002)
J: Information and Communication	0.029	0.026
	(0.002)	(0.002)
K: Financial Institutions	0.125	0.052
	(0.004)	(0.004)
L: Renting, Buying and Selling of	0.131	0.077
Real Estate	(0.004)	(0.004)
M: Consultancy, Research and Other	0.066	0.053
Specialised Business Services	(0.002)	(0.002)
N: Renting and Leasing of Tangible Goods	0.028	0.021
and Other Business Support Services	(0.002)	(0.002)
O: Public Administration, Public Services	0.462	0.057
and Compulsory Social Security	(0.011)	(0.011)
P: Education	0.069	0.043
	(0.002)	(0.002)
Q: Human Health and Social Work	0.114	0.045
Activities	(0.002)	(0.002)
R: Culture, Sports and Recreation	0.007	0.026
	(0.002)	(0.002)
S: Other Service Activities	0	0.019
	(0.002)	(0.002)
U: Extraterritorial Organisations	0.345	0.067
and Bodies	(0.707)	(0.707)
N	1433393	1433393
Adjusted R^2	0.014	0.309

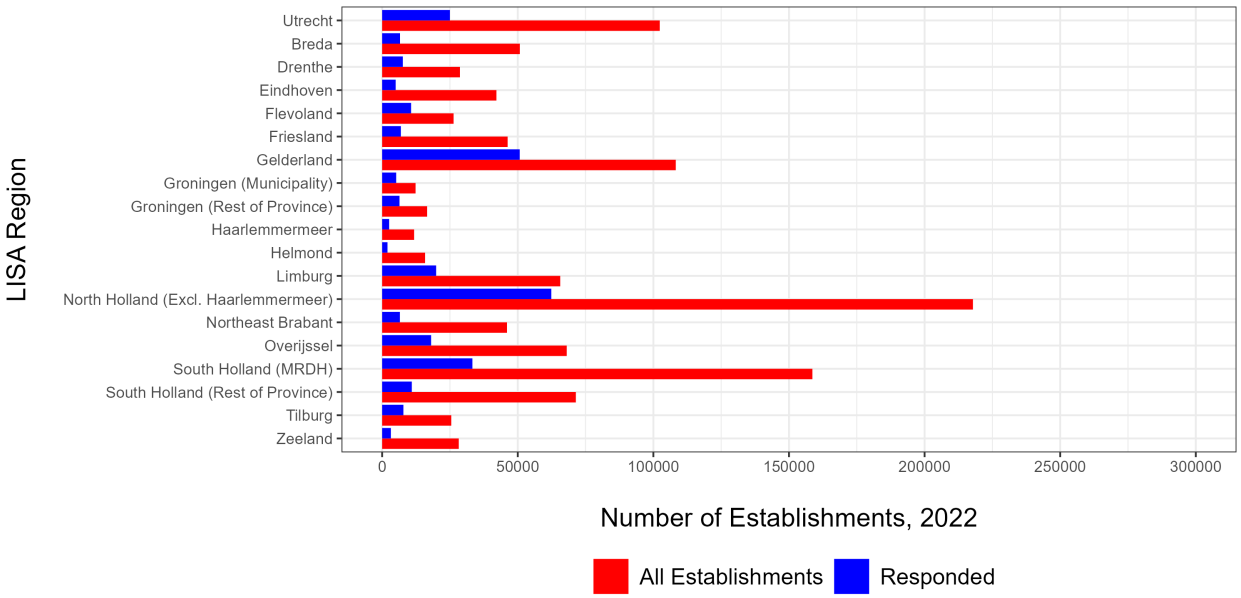
Note: Estimates are generated by regressing a dummy indicating whether the establishment responded to the 2022 regional business surveys (response code ≤ 20) on dummy variables indicating whether the establishment belongs to each of these one-letter 2008 SBI sector codes. The omitted category is "A: Agriculture, Forestry, and Fishing." Ordinary least squares regression coefficients are displayed alongside HC3 heteroskedasticity-robust standard errors in parentheses (MacKinnon & White 1985). Conditional estimates are based on models that control for KVK size classifications, LISA regions, response by email in the 2021 LISA register, response code in the 2021 LISA register, and a dummy indicating whether the establishment is recorded as having ≥ 5 employees in the 2021 LISA register.

Table A2: Differences in Response Rates Between Sectors

Facility Type	Response Probability, Unconditional	Response Probability, Conditional
Accommodation Function	-0.147	-0.042
	(0.006)	(0.006)
Educational Function	0.292	0.107
	(0.005)	(0.005)
Healthcare Function	0.09	0.014
	(0.004)	(0.004)
Industrial Function	-0.013	-0.004
	(0.002)	(0.002)
Jail Function	0.03	0.005
	(0.067)	(0.067)
Meeting Function	-0.047	-0.027
	(0.003)	(0.003)
Multiple Functions	-0.083	-0.024
	(0.002)	(0.002)
Other Functions	-0.045	-0.004
	(0.006)	(0.006)
Port Function	-0.134	-0.01
	(0.008)	(0.008)
Residential Function	-0.183	-0.034
	(0.002)	(0.002)
Shopping Function	-0.062	-0.014
	(0.002)	(0.002)
Sports Function	-0.059	-0.021
	(0.008)	(0.008)
Stand Location	-0.241	-0.064
	(0.007)	(0.007)
Unknown Function	-0.019	-0.006
	(0.007)	(0.007)
Adjusted R^2	0.039	0.308
N	1433393	1433393

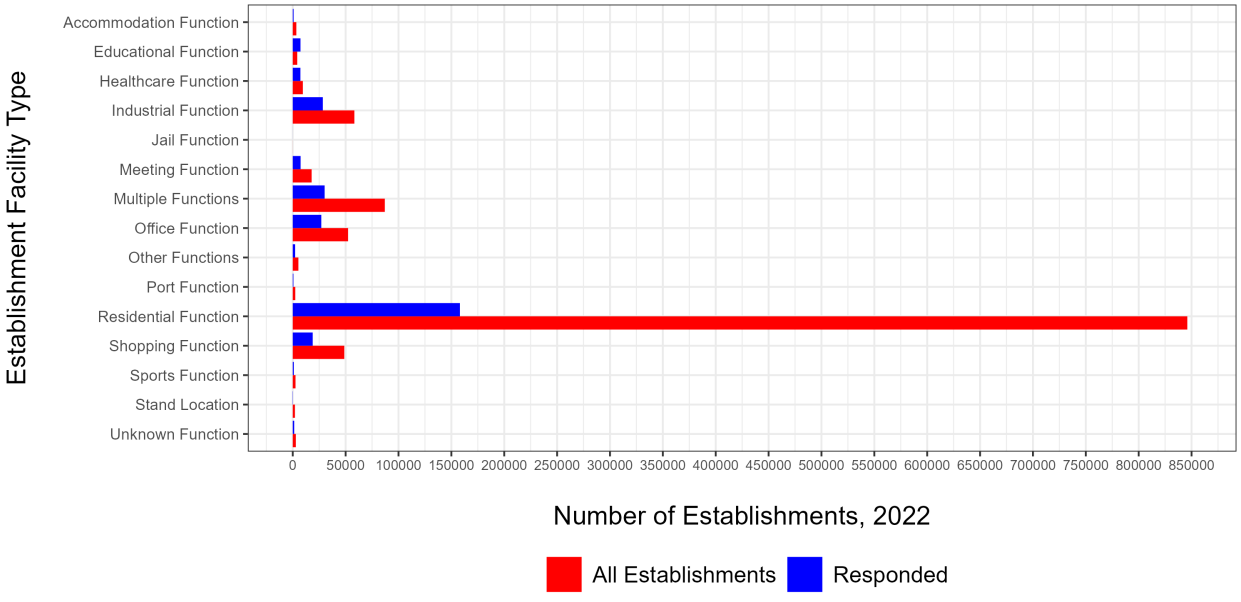
Note: Estimates are generated by regressing a dummy indicating whether the establishment responded to the 2022 regional business surveys (response code ≤ 20) on dummy variables indicating whether the establishment's facility is of the type in the estimate's row. The omitted category is "Office Function." Ordinary least squares regression coefficients are displayed alongside HC3 heteroskedasticity-robust standard errors in parentheses (MacKinnon & White 1985). Conditional estimates are based on models that control for KVK size classifications, LISA regions, response by email in the 2021 LISA register, response code in the 2021 LISA register, and a dummy indicating whether the establishment is recorded as having ≥ 5 employees in the 2021 LISA register.

Table A3: Differences in Response Rates Between Facility Types



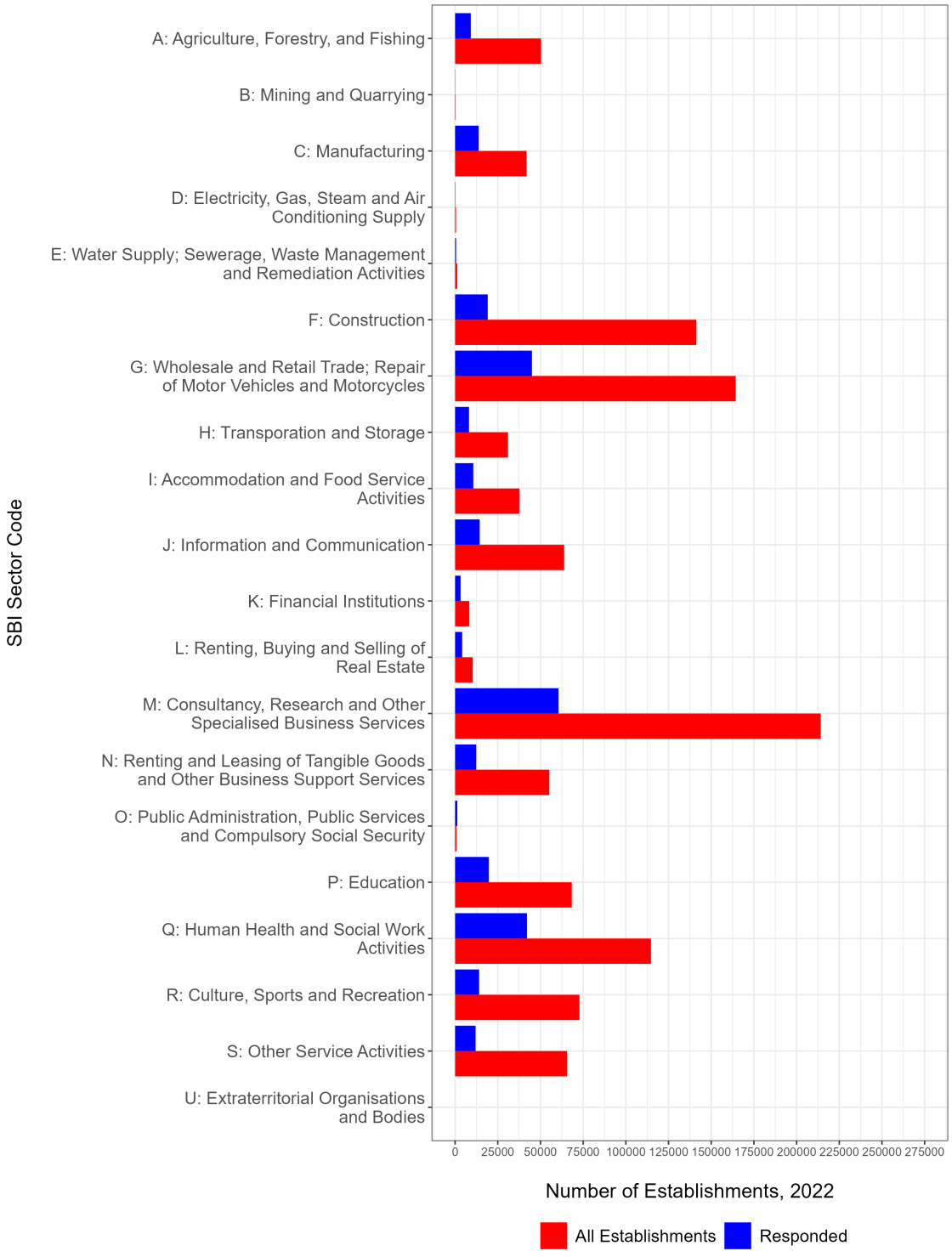
Note: Responsive establishments are those with response codes ≤ 20.

Figure A1: Counts of Establishments by LISA Region



Note: Responsive establishments are those with response codes ≤ 20.

Figure A2: Counts of Establishments by Facility Type



Note: Responsive establishments are those with response codes ≤ 20.

Figure A3: Counts of Establishments by Sector