Privacy and Survey Response: Evidence from Broadband Internet*

Scott Dallman
U.S. Census Bureau

Evan Totty
U.S. Census Bureau

September 24, 2025

Abstract

Federal statistical agencies collect survey data that are used for dispersing funds, informing policy, and aiding research. The collection of accurate data is crucial for these activities, but survey response rates have been declining for three decades. We provide new evidence on privacy and confidentiality concerns as a possible cause for declining response rates. First, we present a model in which individuals choose whether to respond to a federal agency's survey while also interacting with a firm that relies in part on the agency's published data to imperfectly price discriminate. The model demonstrates how privacy loss risk impacts survey response and provides some testable implications. Next, we empirically test the relationship between privacy and survey response using the staggered rollout of broadband internet across the United States as a technology shock that increased individuals' privacy loss risk. Refusal in the Current Population Survey increased immediately after broadband services entered a county. Broadband rollout can explain nearly all of the increase in refusal from 1995-2008. Consistent with the model, we find the impact of broadband rollout on refusal was larger for counties with greater proxied household willingness to pay for an arbitrary good or service and, among households who did respond, broadband rollout increased question refusal for topics that reveal more private and sensitive information. Finally, we present some back-of-the-envelope calculations for the implied change in privacy loss risk due to the rollout of broadband internet.

Keywords: privacy, confidentiality, survey response, survey refusal, broadband internet, high-speed internet

*Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the Census Bureau or other organizations. We thank Gary Benedetto, Justin Doty, and Jordan Stanley for their comments on the paper. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (this paper does not use any confidential data).

1 Introduction

Federal statistical agencies in the United States collect survey data on many aspects of society and the economy. These data are used to produce the official statistics on employment, poverty, health insurance coverage, inflation, and other indicators that inform public policy. They are also used in the distribution of trillions of dollars in funds annually and are a primary source of information for research in academia, industry, and government (Ross, 2023). These uses rely on high quality data that are accurate, timely, and representative of the entire population. However, survey response rates have been declining since the 1990s and the decline accelerated in the 2000s (Meyer et al., 2015; Williams & Brick, 2018). Consequently, some agencies are evolving to reduce reliance on survey data (Jarmin, 2019).

Understanding the causes of declining survey response rates is crucial for helping statistical agencies take steps to reverse the trend. Respondent burden, less leisure time for filling out surveys, decline in public spirit, rising political polarization and mistrust of government, and privacy and confidentiality concerns have been offered as possible causes, but no single factor has been identified as a key driver and the literature has called for more research into the topic (Meyer et al., 2015; National Research Council, 2013). Our focus is on privacy and confidentiality concerns, which has received little rigorous evaluation. The mechanism we have in mind is one in which individuals refuse surveys in an attempt to manage their privacy risks in a world in which data collection and personalized pricing are ubiquitous (Council of Economic Advisors, 2015).

We start with a simple model of an individual's decision to respond to a survey conducted by a federal statistical agency. The purpose of the model is to motivate how big data can influence the trade-offs associated with responding to surveys and to provide some testable predictions related to privacy and survey refusal.[1] The model includes individuals who choose whether to respond to a survey, a statistical agency that uses the survey responses to produce

[1]The meaning of big data has evolved over time and can be context dependent. We use the term to refer to the growing ability to access and combine data from multiple sources in order to generate new insights, measurements, and predictions (Council of Economic Advisors, 2015).

and disseminate public data, and a monopolistic firm who sells a product to consumers. The firm uses its own data (which it could have collected itself or purchased from a data broker) and the agency's public data in an attempt to engage in price discrimination. The firm has a prior likelihood of inferring each consumer's willingness to pay based on its own data. The firm can update this prior with intended statistical uses of the agency's data, such as learning average income for individuals of a given race and county when they already know a customer's race and county from their own data. The firm may further refine their prior by re-identifying some anonymized individuals in the agency's database with data reconstruction or data linkage attacks, thereby revealing the individuals' exact information.[2] Such attacks are not considered an intended statistical use of the agency's data and represent a privacy loss for the individual. An individual is subject to the firm updating their priors from intended statistical uses whether they participate in the survey or not, but they are only subject to the additional re-identification risk if they exist in the survey data. The decision rule for whether to respond to the survey depends upon the increased likelihood of the firm inferring their willingness to pay caused by existing in the agency's public data. Larger increases in likelihood reduce consumer surplus and thus lead to fewer responses.[3] The decision rule also depends upon the difference between an individual's willingness to pay and the market price that the firm charges to individuals whose willingness to pay it cannot infer, as individuals with greater willingness to pay have more surplus to lose by having their willingness to pay revealed.

Next, we evaluate our model of privacy and survey refusal empirically. Rather than directly measuring the risk of consumers having their willingness to pay revealed or attempting to measure willingness to pay for some product, we use the staggered rollout of broadband internet across the United States from 1995-2012 as a technology shock that reduced indi-

[2]Re-identification in our setting means a firm has confidently determined the identify of an individual in the agency's public database, allowing it to merge the agency's data on that individual with their own data on that individual.

[3]It is worth noting that the mechanism we have in mind only requires *perceived* changes in privacy loss due to responding to federal surveys, regardless of whether it is actually true.

viduals' privacy and increased firms' ability to infer willingness to pay and offer personalized pricing. The widespread applications of broadband internet have transformed the privacy landscape in the United States. The advent and diffusion of broadband internet made possible the ascent of "Web 2.0" (blogs, social media, and online networks), online marketplaces, and the Internet of Things, enabling massive amounts of individual behavior to be observed and collected, including demographic, economic, and behavioral information.[4] Combined with advances in computing power and data science, high-speed internet made possible the joint phenomena of big data and consumer surveillance, whereby firms rely on consumer data to understand, target, and influence their customers.

Firms promote personalized services and reduced search costs as benefits of big data, but these gains often come with privacy-related costs such as fraud, manipulation, stigma, and discrimination.[5] While broadband internet access was growing during the late 1990s and early 2000s, concerns regarding online privacy were garnering national attention and laws were enacted to regulate online privacy.[6] Amazon faced backlash in 2000 after it was discovered they were using personalized pricing based on consumer data collected online.[7] However, firms' data collection practices and privacy policies have remained opaque and difficult to understand, leaving privacy threats in place and making it challenging for individuals to manage their personal privacy.[8] Among those threats, there is an abundance of

[4]By the mid-2000s, user-generated content exceeded business-generated content on the internet for the first time and online advertising became the dominant form of advertising due to the ability to target ads based on individual data (Evans, 2009).

[5]The role of the internet in shaping the modern privacy landscape has been carefully described in prior work, along with associated costs and benefits that arise within the economics of privacy and digitization (Acquisti et al., 2016; Goldfarb & Que, 2023; Goldfarb & Tucker, 2019; Varian, 2010).

[6]Examples of national media coverage include the TIME Magazine cover from July 2, 2001, which was an image of a computer in the shape of a padlock with the title, "How to Protect Your Privacy Online"; and the TIME Magazine cover from February 20th, 2006, which was titled, "Can We Trust Google With Our Secrets?" Examples of laws include the Children's Online Privacy Protection Act (COPPA) of 1998, which governed the online collection of information about children, and the Gramm-Leach-Bliley Act of 1999, which allowed individuals to opt-out of financial institutions sharing non-public data with third parties.

[7]The incident was covered by many major news networks: ABC News (2000), Chicago Tribune (2000), Los Angeles Times (2000), and The Wall Street Journal (2000).

[8]E.g., Bian et al. (2024) analyze the rollout of Apple's App Tracking Transparency (ATT) policy in 2021, which significantly curtailed online data collection and sharing on the iOS platform, and showed that ATT substantially reduced consumer fraud complaints.

evidence that firms across a wide range of industries now use these data to price discriminate via personalized prices.[9]

Given this premise that broadband internet was a privacy-reducing technology shock, we use its staggered rollout across the United States in the late 1990s and 2000s as a natural experiment for evaluating the impact of privacy and confidentiality on survey refusal in the Current Population Survey (CPS). The percentage of households with access to broadband internet grew dramatically during this time frame, as did survey refusal rates. We exploit the staggered timing of when counties across the United States first gained access to broadband internet in order to test for a causal relationship between these trends.

We find that survey refusal increased after exposure to broadband internet. There were no differences in full survey refusal rates (i.e., "unit" refusal) between households in counties with versus without broadband internet in the years leading up to broadband exposure, but household refusal rates increased immediately after county-level exposure to broadband internet. The rise in refusal grew over time after initial exposure with the largest effects at the end of the event study time frame (four years after exposure). We show that broadband internet can explain nearly all of the increase in CPS refusal from 1995-2008. The actual refusal rate increased from 4.0% to 5.5%, whereas a counterfactual rate based on our results only increased from 4.0% to 4.2%. We find no evidence that broadband internet rollout was associated with changes in other forms of non-response among eligible households, which helps to validate that the estimated relationship between broadband internet rollout and survey refusal was due to changes in respondents' willingness to share their information rather than coincident factors.[10]

We further evaluate the relationship between privacy and survey refusal by testing two

[9]See Chicago Tribune (2018), Food Dive (2017), and Mohammed (2017) for a history of personalized pricing and details of recent examples. See Aparicio et al. (2024), Hannak et al. (2014), and Mikians et al. (2012) for evidence of price discrimination on the internet. See Shiller (2020) for how creating personalized pricing using consumer demographics and web-browsing histories can increase firm profits.

[10]E.g., if we found evidence of decreased non-response due to inability to locate the housing unit, we might conclude that internet use by field representatives collecting data door-to-door led to more interview attempts, some of which happened to be refused.

additional implications from our model. One implication is that increases in refusal associated with broadband internet should rise with individuals' willingness to pay for a given good or service. We find support for this using several measures of average county-level household economic well-being as proxies for individuals' willingness to pay. The second implication is that, for individuals who respond to at least some portion of the survey, question-level refusal (i.e., "item" refusal) for items that reveal more private and sensitive information should increase after exposure to broadband internet.[11] We find support for this by comparing the impact of broadband exposure on refusal of household family income (which is likely seen as a relatively private and sensitive piece of information) versus householder age (which is likely seen as less private and sensitive).[12]

Finally, we also connect the survey refusal results to the model by backing out the implied increase in a firm's ability to infer willingness to pay that can rationalize the rise in survey refusal. Using back-of-the-envelope calculations based upon inverting a latent logistic decision framework, we show that as consumer surplus increases, smaller increases in inference likelihood are needed in order to rationalize the increase in survey refusal. Using estimates from other studies on average consumer surplus for airline tickets and for household items purchased online from Amazon, a reasonable estimate of the implied average increase in inference likelihood ranges from 0.02 percentage points for relatively high surplus items such as airline tickets up to 1-2 percentage points for relatively low surplus items such as household goods purchased from Amazon.

The remainder of the paper is organized as follows. Section 2 provides more information on broadband internet and its role in the modern privacy landscape. Section 3 introduces

[11] We use *private information* to mean information that is (ideally) unknown to others, whereas we use *sensitive information* to mean information that could be used against an individual if it were known. Not all private information is sensitive (e.g., a favorite childhood toy or niche hobby), while not all sensitive information is private (e.g., criminal conviction record). The relevant information for our purposes is information that is both private and sensitive (e.g., income or medical history), although we will use the two terms interchangeably at times in the paper.

[12] The idea that income is more private and sensitive than age is supported by the existence of higher refusal rates for income (see, e.g., Figure 9). Income non-response is frequently attributed to privacy concerns (Fobia et al., 2025; Goldfarb & Tucker, 2012; Singer et al., 1997).

a model for survey response with some testable implications. Section 4 describes the data. Section 5 describes the empirical methods. Section 6 presents the results. Section 7 discusses the back-of-the-envelope estimates for the implied change in willingness to pay inference likelihood. Section 8 concludes.

2 Background

The economics of privacy pertains to the trade-offs associated with protecting or sharing personal information between individuals, organizations, and governments. Acquisti et al. (2016) highlights how the economic analysis of privacy has evolved over time. Although economists have recognized growing concerns for individuals' right to privacy since the adoption of digital technology after World War II, consumer privacy risks have historically been limited by the technological ability to digitally collect, store, and analyze information. However, the commercial success of the internet and proliferation of big data has led to a new wave of privacy research on the protection of information about an individual's preferences or type (Acquisti et al., 2016).

Initially limited by slow and costly dial-up connectivity, the construction of broadband infrastructure in the late 1990s led to the rapid growth of internet use by the United States public.[13] Figures 1 and 2 show the trend in United States internet usage and broadband internet access. Internet use grew rapidly beginning in the mid-1990s with the adoption of home computing.[14] By 2012 almost 75% of the United States population used the internet, with 30 fixed broadband subscriptions per every 100 persons.[15] Broadband connectivity

[13]The internet was first used to send and read electronic mail in 1972. However, internet access was not publicly available in the United States until the 1990s, with the privatization of infrastructure development, the creation of the World Wide Web, and the commercialization of user-friendly internet browsers (e.g., Netscape) and service providers (e.g., CompuServe and America Online) (Greenstein, 2000; Leiner et al., 2009).

[14]From 1990 to 1997, the number of United States households owning computers increased from 15 to 35 percent and spending on computers and related hardware more than tripled (BLS, 1999).

[15]Fixed broadband subscriptions refers to fixed subscriptions for high-speed access to the public internet (a TCP/IP connection), at downstream speeds equal to, or greater than, 256 kbit/s. It includes both residential subscriptions and subscriptions for organizations.

provided the high-speed bandwidth necessary to support Web 2.0 services such as keyword search engines, e-Commerce, and social media networks. While consumers benefit from these services via data-driven innovations, higher quality products, and better matched services (e.g., personalization and reduced search costs), large amounts of data on individuals are generated, stored, and exchanged on the internet as a byproduct of this activity.

Data produced and disseminated by statistical agencies can also inadvertently cause the public disclosure of sensitive information, particularly when combined with data from third parties. For example, using only the Social Security Administration's publicly available Death Master File matched to personal information from third party data brokers and social networks, Acquisti and Gross (2009) were able to statistically infer the social security number of some individuals with greater than 60% accuracy. Agencies such as the Census Bureau are modernizing disclosure avoidance methods to protect respondents' information due to rising risks of reconstruction and re-identification attacks made possible by advances in computing power, data science, and the available of auxiliary data (Abowd & Schmutte, 2019). While the Census Bureau is required by law to protect the confidentiality of its respondents, it also must deal with the fact that even *perceived* privacy loss risk may affect the public's willingness to respond and do so accurately.[16] Privacy and confidentiality concerns are often implicated as a possible cause for declining response rates in surveys such as the CPS, which is the primary source of labor force statistics for the United States.[17]

Our paper is the first we are aware of to empirically evaluate the relationship between privacy and confidentiality and survey refusal in the CPS. It is also the first we are aware of to use the rollout of high-speed broadband internet as a strategy for providing quasi-experimental variation in privacy loss risk. Other studies that have evaluated particular

[16]In a 2019 public opinion poll, 81% of the public said the potential risks they face because of data collection by companies outweigh the benefits, and 66% said the same about government data collection. Most respondents expressed concern about the way their data are used by companies (79%) and the government (64%) (PEW, 2019).

[17]References to privacy as a possible cause of declining survey response in the CPS can be found in Meyer et al. (2015) and in joint work by the Census Bureau and Bureau of Labor Statistics (BLS) on modernization efforts (Census Bureau, 2023; Linse & Johnson, 2023). More information on the CPS can be found from the Census Bureau (https://www.census.gov/programs-surveys/cps.html) and BLS (https://www.bls.gov/cps/).

explanations for rising survey refusal include Borgschulte et al. (2022), which evaluated political partisanship and CPS refusal, and Goldfarb and Tucker (2012), which inferred that privacy concerns have risen and evolved over time based on rising income refusal rates in an online marketing research survey. Other related papers have studied how the ability of firms to infer information about individuals influences financial fraud and the willingness of individuals to share their data (Acemoglu et al., 2022; Argenziano & Bonatti, 2023; Bian et al., 2024; Miklós-Thal et al., 2023). Finally, our paper is also related to work that used the staggered rollout of broadband internet as an identification strategy for studying the internet's impact on other topics such as labor supply, educational outcomes, rural connectivity, voting turnout, well-being, economic growth, and health outcomes (Atasoy, 2013; Dettling, 2017; Dettling et al., 2018; Dinterman & Renkow, 2017; Falck et al., 2014; Johnson & Persico, 2024; Kolko, 2012; Van Parys & Brown, 2023).

3 Model and Predictions

This section presents a model to illustrate how big data can influence an individual's decision to respond to a federal statistical agency's survey. We focus on price discrimination as the cost due to privacy loss, which is the ability of firms to charge different prices to different consumers based on their willingness to pay. Willingness to pay is generally an unobservable characteristic, but it could be determined with high precision based on observable characteristics such as income, sex, race, family size and structure, geographic location, and purchase history (Acquisti et al., 2016; Council of Economic Advisors, 2015; Odlyzko, 2003).

The model includes a statistical agency, a monopolistic firm, and individuals who are both potential respondents to the statistical agency and potential customers to the monopolistic

firm.[18] The agency administers a household survey and disseminates data in the form of tabulated statistics and microdata. Individuals choose whether to respond to the survey. If an individual chooses to respond, then their information is part of the tabulated statistics and microdata disseminated by the agency. Individuals also interact with the monopolistic firm. The firm has its own database on potential customers and can also use the agency's public data. For each individual $i \in N$, their valuation for the monopolist's product is equal to w_i. If the monopolist charges a price $p < w_i$, then the individual will purchase the product and receive a surplus of $w_i - p$. If the monopolist charges a price $p > w_i$ then the individual will not purchase the product. If the monopolist charges a price $p = w_i$ then the individual is indifferent between purchasing the product or not and receives zero surplus in either case.

The firm's own database allows it to infer an individual's willingness to pay with some likelihood. The firm can update this likelihood with intended statistical uses of the agency's database. For example, a firm might know a customer's county and race, and then use the agency's data to learn the average income for individuals of a given race in that county. The firm could further update this likelihood by re-identifying seemingly-anonymized individuals in the agency's public database, thus revealing the individuals' exact information. For example, the firm might attempt a linkage attack by using quasi-identifiers to merge its own database to microdata released by the agency to look for unique matches, or the firm might perform a reconstruction-abetted attack by first rebuilding the agency's unreleased microdata based on published tables and statistics and then performing a linkage attack.[19] This would not be an intended statistical use of the agency's database and represents an

[18]Part of our framework and notation are borrowed from Belleflamme and Vergote (2016) for analyzing a situation in which a monopolistic firm can use a tracking technology that allows it to imperfectly price discriminate and consumers can fully "hide" from a monopolistic firm by adopting a hiding technology to combat the firm's tracking technology. We modify this framework some as described below. We also borrow from Reiter (2005) and McClure and Reiter (2012) for modeling the risk of individual identification disclosure in microdata.

[19]See Heffetz and Ligett (2014) for examples of real world linkage attacks and Abowd et al. (2023), Acquisti and Gross (2009), Garfinkel et al. (2019), and Kosinski et al. (2013) for demonstrations of reconstruction attacks.

unintended, re-identification-based privacy loss for the individual.[20]

The key feature of the model is that individuals cannot avoid existing in the firm's database nor can they prevent the agency from releasing its own public database. However, they can control whether they exist in the agency's database, thereby controlling the additional risk of privacy loss due to re-identification risk. This feature mirrors modern society where firms collect a large amount of consumer data since the exchange of personal data is often required to receive goods and services.[21]

We assume that the individual first interacts with the agency. The agency then disseminates public microdata and/or tabulated statistics based on the survey data. The firm then uses its database and the agency's public data to set a schedule of personalized prices and a market price for individuals whose willingness to pay it could not infer. We take the firm's engagement in price discrimination and re-identification as given. However, it can only do this imperfectly due in part to data limitations. Next, we discuss the agency, the monopolist, and the individual in more detail.[22]

3.1 The statistical agency

The agency collects survey information on s sampled units of the population N, $s \leq N$. Let y_{jk} be the data collected for individual j on variable k, for $k = 0, ..., K$ and $j \in s$. The variable $k = 0$ is a unique individual identifier, such as a social security number or name and date of birth, that is not released by the agency. Let $y_j = (y_{j0}, y_{j1}, ..., y_{jK})'$ be the

[20]Technically, an individual could face a type of re-identification risk even when they do not exist in the agency's database. For instance, if some of the firm's data is aggregate (containing the individual in question) and the statistical agency fails to protect its respondents, then the firm can potentially subtract out the agency's public aggregates from their own aggregates and learn about the individual in question. We bundle this risk into the baseline risk an individual faces if they choose not to respond.

[21]Belleflamme and Vergote (2016) analyze a situation in which consumers can fully "hide" from a monopolistic firm by adopting a hiding technology to combat the firm's tracking technology. We modify this framework such that the firm has access to their own database without a tracking technology, but they also use data combination and linkage as a form of tracking technology. Consumers cannot hide from the firm's database, but they can "hide" from the statistical agency's database and thereby avoid the additional risk.

[22]Belleflamme et al. (2020) and Rhodes and Zhou (2024) generalize many of the insights from Belleflamme and Vergote (2016) to a setting with multiple sellers who still retain some market power. We leave the extension of our model to settings with multiple sellers for future work.

vector of confidential data collected for individual j. For simplicity, assume that all variables other than the unique individual identifiers are released to the public. Let $z_j = (y_{j1}, ..., y_{jK})'$ be the vector of released data for individual j and let $Z = (z_1, ..., z_s)$ be the full database disseminated to the public by the agency.[23]

3.2 The monopolist firm

The firm produces its product at a constant marginal cost that we set to zero for simplicity. A unit mass of consumers have a unit demand for the product. The distribution of consumers' valuations is given by the cumulative distribution function $F(w)$ with support $[0, \bar{w}]$, where $\bar{w} \in (0, \infty]$, and by a continuous and differentiable density $f(w) = F'(w) \geq 0$.

The firm has access to a database, A, on consumers. This may be data the firm collected itself or information purchased from other entities. A contains vectors of information on $n \leq N$ individuals from the population who may or may not correspond to an individual in database Z from the statistical agency. Let a_{ij} be the data for individual i on variable j, for $j = 0, ..., J$ and $i \in n$, where a_{i0} is a unique individual identifier that would uniquely merge to Z if y_{j0} were included in Z. Let $a_i = (a_{i0}, a_{i1}, ..., a_{iJ})'$ be the vector of data for individual i and let $A = (a_1, ..., a_n)$ be the full database.

The goal of the firm is to infer every individual's willingness to pay for its product, w_i, which we also refer to as the individual's "type." With probability ρ^{prior} the firm is able to infer the consumer's type from database A. The firm refines its inferences using the agency's database, Z, via intended statistical uses and re-identification attacks. If an individual is in the agency's database, then the firm can infer their type with probability ρ^{post}. When the individual is not in the agency's database, we denote the probability using ρ_{-i}^{post}. We assume below that $\rho^{post} \geq \rho_{-i}^{post} \geq \rho^{prior}$. The difference between ρ_{-i}^{post} and ρ^{prior} depends upon generalizable insights that can be learned from intended statistical uses of the agency's

[23] Reiter (2005) includes extensions for when the agency disseminates data for only a sub-sample of the sampled individuals and when some or all variables undergo privacy protection between collection and dissemination.

public database. The difference between ρ^{post} and ρ_{-i}^{post} depends upon the ability of the firm to re-identify individual i in the agency's database.

In terms of pricing, when an individual does not respond to the agency's survey, this means that with probability ρ_{-i}^{post} the firm knows the individual's willingness to pay and charges the individual a personalized price $p(w_i) = w_i$, whereas with probability $(1 - \rho_{-i}^{post})$ the firm does not know the individual's valuation and charges them a "regular" price p. If the individual does respond to the agency's survey then the firm charges this consumer a personalized price $p(w_i) = w_i$ with probability ρ^{post} and charges them a "regular" price p with probability $(1 - \rho^{post})$. Therefore, individuals who do respond face an increased risk of having their willingness to pay revealed to the firm (based on the assumption that $\rho^{post} \geq \rho_{-i}^{post}$) and losing out on consumer surplus if their valuation was greater than the regular price. Individuals anticipate that they will pay an equilibrium regular price p_e if they are not identified or a personalized price equal to their valuation if they are. Given this expectation, which determines the mass of individuals who decide to respond, the firm uses its database in combination with the agency's database to set prices (i.e., the regular equilibrium price p_e and a schedule of personalized prices based on individual values $p(w)$).

3.3 The individual

The individual decides whether to respond to the agency's survey. The individual also decides whether to purchase the firm's product. We assume that the second decision occurs at some point in the future after the first decision, once the agency has released database Z which the firm uses to set prices. Responding comes with the benefit of individual representation in the agency's database, which includes benefits related to accurate data for societal outcomes (government funding, informing policy, and aiding research), altruism, and any financial rewards for responding. For simplicity, we assume these benefits can be aggregated to an expected net present value equal to B.

Any individual with valuation $w_i \geq p_e$ will have utility of $B + (1 - \rho^{post})(w_i - p_e)$ if they

do respond (i.e., the benefits of representation plus any consumer surplus if their type is not identified) and utility of $(1-\rho_{-i}^{post})(w_i-p_e)$ if they do not respond (i.e., any consumer surplus if their type is not identified). Comparing these two alternatives, it is worth responding if:

$$B + (1 - \rho^{post})(w_i - p_e) \geq (1 - \rho_{-i}^{post})(w_i - p_e)$$

Re-arranging and collecting terms, it is worth responding if:

$$B \geq (\rho^{post} - \rho_{-i}^{post})(w_i - p_e) \tag{1}$$

That is, it is worth responding if the benefits of response (B) are larger than the costs. The costs of response are the surplus gains of not being identified ($w_i - p_e$) times the change in the likelihood of being identified due to existing in the agency's database ($\rho^{post} - \rho_{-i}^{post}$).

3.4 Empirical evaluation

Equation (1) motivates the empirical part of the paper. We can rewrite the decision rule as an indicator for refusal:

$$Refusal_i = \mathbb{I}\{(\rho^{post} - \rho_{-i}^{post})(w_i - p_e) \geq B\},$$

which we can also think of as a latent index model:

$$Pr(Refusal_i) = f((\rho^{post} - \rho_{-i}^{post})(w_i - p_e) - B). \tag{2}$$

We cannot directly measure willingness to pay (w_i) or re-identification risk ($\rho^{post} - \rho_{-i}^{post}$) in our data. Both could be estimated in some narrow settings, but willingness to pay is product-specific and requires purchase history data, while re-identification risk requires knowledge of data available to the firm and/or knowledge of how the agency protected the data it released.[24]

[24]See Appendix A for more discussion of the agency's role in controlling re-identification risk.

Rather than measuring re-identification risk and willingness to pay directly, we use proxies based on measures known to relate to these concepts. We propose three different tests to empirically evaluate our model of privacy and survey response:

Hypothesis 1 *We can use the rollout of broadband internet to provide quasi-experimental variation in re-identification risk. Larger amounts of data collection made possible by broadband internet increased re-identification risk. Broadband internet access should therefore be associated with an increase in survey refusal.*

As described in previous sections, increases in $\rho^{post} - \rho_{-i}^{post}$ depend upon the ability of the firm to re-identify individual i in the agency's database, and the rollout of broadband internet dramatically and rapidly increased the ability of firms to do exactly that, via expanded data availability and computational power made possible by the widespread adoption of high-speed internet and its downstream applications.[25] Equation (1) indicates that for a given individual type (w_i), an increase in the likelihood of the individual having their willingness to pay revealed due to existing in the agency's database, $\rho^{post} - \rho_{-i}^{post}$, increases the cost of response and thereby reduces the number of individuals for whom response would be worthwhile. Thus, we expect the rollout of broadband internet to be associated with an increase in survey refusal.

Hypothesis 2 *We can use measures of economic well-being as proxies for willingness to pay. The increase in refusal associated with broadband internet access should therefore be larger for those with greater economic well-being.*

Equation (1) also indicates that for a given increase in the likelihood of an individual having their willingness to pay revealed due to existing in the agency's database, $\rho^{post} - \rho_{-i}^{post}$,

[25]Many recent papers have documented the rising risk of reconstruction and re-identification attacks (e.g., Abowd & Hawes, 2024; Henriksen-Bulmer & Jeary, 2016), while others have documented specific examples of successful re-identification, reconstruction, and inference attacks (e.g., Abowd et al., 2023; Acquisti & Gross, 2009; Garfinkel et al., 2019; Heffetz & Ligett, 2014; Kosinski et al., 2013). In general, it is widely recognized that the internet and its downstream applications spawned unprecedented economic and policy issues relating to the protection of personal data (Acquisti et al., 2016).

individuals with a larger value for the product, w_i, have more to lose by having their type revealed and thus suffer larger losses due to re-identification. Consider an individual with $w_i < p_e$. This individual will never purchase the product and thus receives zero surplus whether their type is revealed or not. Similarly, an individual with $w_i = p_e$ is indifferent between purchasing the product or not and receives zero surplus in either case, regardless of whether their type is revealed. Only individuals with $w_i > p_e$ are at risk of lost consumer surplus due to having their type revealed, and this risk grows with w_i. In the empirical part of our paper, we generate measures of economic well-being that are related to (and proxy for) willingness to pay. According to the model, these proxy measures should be associated with larger increases in survey refusal following exposure to broadband internet.

Hypothesis 3 *Some survey questions reveal more private and sensitive information than others. Among individuals who respond to some portion of the survey, broadband internet access should be associated with increased refusal of questions that reveal more private and sensitive information.*

While Hypothesis 1 and Hypothesis 2 deal with overall (unit) survey refusal, some individuals partially respond to the survey by answering some questions (items) but refusing others. We can also evaluate the impact of broadband internet on item refusal. Items that reveal more private and sensitive information are likely to be the items in a database that increase re-identification risk and willingness to pay inference precision, which determines $\rho^{post} - \rho_{-i}^{post}$. We should therefore see increases in refusal of these items following exposure to broadband internet.

4 Data

4.1 Survey refusal

Our analysis is based on survey refusal in the CPS basic monthly surveys from 1995-2012.[26] The CPS spans the period of rapid broadband internet service expansion during the late 1990s and early 2000s. The CPS also identifies county of residence, which is necessary for linking the broadband data described below. The CPS is designed to interview households once per month for four months, not interview them for eight months, and then interview them again for the next four months.[27]

Importantly for our analysis, the public CPS data includes details on the reason for non-interviews, including refusal of the survey overall ("unit" non-response/refusal) and refusal of specific survey questions ("item" non-response/refusal). When a household fails to respond to the CPS survey in a certain month, it is recorded as a unit non-response and categorized as either Type A, B, or C. Type A indicates households that were eligible to be interviewed but were not because of refusal, absence from the home, language barriers, weather disruptions, illness, or inability to locate the address. The data further differentiate Type A non-response into refusal versus all other reasons. Type B indicates housing units that currently have no residents eligible for interview (e.g., vacant or occupied by people whose usual residence is elsewhere). Type C indicates housing units ineligible for interview, such as units that were demolished or converted to storage/business use.

Our outcomes of interest are two types of non-response: unit non-response and item non-response for survey questions that present varying levels of perceived privacy and sensitivity (household wage/salary income and age). Figure 4 shows Type A, Type B, and Type C unit non-response rates in the CPS from 1995-2012. Type B was the largest reason for unit

[26] We used the programs provided by the Center for Economic and Policy Research (Center for Economic and Policy Research, 2019) to prepare the Census Bureau CPS basic monthly data.

[27] As a robustness check, we merged the Census Bureau's CPS basic monthly data files with the IPUMS CPS data in order to use the IPUMS longitudinal household identifier and limit the sample to houses that do not attrite from the survey (Rivera Drew et al., 2014). The results are very similar to those reported in the paper and are available upon request.

non-response, followed by Type A, then Type C. Both Type A and Type B non-response rates increased during this time frame, whereas Type C non-response rates remained flat.

Figure 5 separates the Type A non-response rates into refusals versus other reasons. Both components of Type A non-response were higher by 2012 than they were in 1995, but from 2000 onward only the refusal rate was increasing while other Type A non-response was flat or declining. For our analysis of refusal below, we exclude Type B and Type C non-response observations and other Type A non-response observations as these observations by definition were not capable of being a refusal.[28] We also analyze the impact of broadband rollout on the rate of other Type A non-response, which we view as a falsification test. Hypothesis 1 claims that broadband internet impacted the likelihood of survey response due to increased privacy loss risk. This implies that the rollout of broadband internet should impact refusal, but not other survey-eligible non-response. Results from the falsification test help rule out the possibility that any changes in refusal associated with the staggered rollout of broadband internet were driven by other factors that influenced all components of Type A non-response (such as changes in survey implementation).[29]

4.2 Broadband internet

To identify household access to broadband internet services, we constructed county-level measures of broadband coverage using information from the Federal Communications Commission's (FCC) Form 477 and then merged them to the CPS. The FCC's Form 477 is a mandated form submitted biannually by all United States internet service providers to document their broadband infrastructure's geographic coverage. From 1999-2008, internet service providers were required to submit service coverage information at the ZIP Code level.[30]

[28] As a robustness check, we also used samples that are restricted to eight refusal or response observations per household to account for potential compositional effect biases as households transition into or out of other types of non-response. These results are very similar to those reported in the paper and are available upon request.

[29] Prior work suggests that changes in survey technology and methodology were not major factors in rising non-response rates (Brick & Williams, 2013). Moreover, our analysis period begins after the major CPS survey methodology and questionnaire redesigns in 1994.

[30] After 2008, the FCC changed the type of information collected and the geographic level of collection.

Since geographic information below the county level is not in the publicly-available CPS microdata, we harmonized and aggregated FCC 477 data from 1999-2008 to construct county-level indicators of broadband internet access. First, we converted the FCC ZIP Code data to the county level by merging county information from the ZIP Code-county crosswalk developed by the U.S. Department of Housing and Urban Development (HUD).[31] Then, we constructed summary measures of the FCC data at the county level based on the presence of at least one provider in a county, the number of providers in a county, and the fraction of a county's household addresses that have at least one provider in their Zip Code.[32]

Figure 6 summarizes the FCC broadband coverage information. The top row summarizes the proportion of ZIP Codes and counties with at least one broadband internet service provider over time. The second row summarizes the average number of broadband internet service providers at the ZIP Code and county level over time. The third row summarizes the proportion of each county's residential addresses with a broadband service provider over time (the first column of the third row shows the average share of a county's addresses with a service provider in their ZIP Code over time, while the second column of the third row shows the proportion of counties with a service provider in 100% of Zip Codes over time).

For the analysis below, we use a treatment variable based on the summary measure in the bottom-right figure of Figure 6. That is, a county is considered "treated" when every ZIP Code in the county has a broadband internet service provider. Requiring the whole county to have a broadband internet service provider before it is considered "treated" may seen likely to underestimate coverage, but the *availability* of internet service is known to over-state internet *usage*, so our definition likely mitigates some of this bias since not all

[31]The crosswalk can be downloaded from https://www.huduser.gov/portal/datasets/usps_crosswalk.html. Converting ZIP Codes to counties is an imperfect process because ZIP Codes are created by the U.S. Postal Service for mail delivery and do not always nest within county borders. When a ZIP Code in the FCC data is associated with more than one county in a single time period, we create multiple records for that ZIP Code-year observation – one for each county that it falls within. ZIP Code-county pairings are based on the existing pairings in 2010 Q1, which is the earliest version of the HUD crosswalk. Any ZIP Code-county pairings that existed from 1999-2008 but no longer existed in 2010 Q1 do not exist in the crosswalk and thus will not appear in our 1999-2008 county-level FCC data.

[32]The HUD crosswalk also has information on the fraction of each county's residential addresses that fall within each ZIP Code.

persons in a county may use the internet. This can be seen by comparing the different summary measures in Figure 6 to the household-level internet usage data according to the CPS shown in Figure B1 in Appendix B. Actual household usage data tracks best with our preferred treatment measure based on the availability of a provider in all ZIP Codes in a county: estimated internet usage at home increased from about 20% in 1998 to about 80% in 2012, whereas the proportion of counties with a service provider in every zip code increased from about 15% in 1999 to about 85% in 2008.[33]

We merge our county-year broadband availability measures to the CPS data using the county of residence associated with each household in the given CPS year. Since the FCC data cover 1999-2008, we use CPS data from 1995-2012 in order to use leading and lagging treatment effects in some of our analyses.[34] Households located in counties whose identifiers were suppressed in the CPS were dropped from the sample.[35]

Table 1 shows summary statistics for our broadband service treatment variable in the full FCC 477 data and in the subset of FCC 477 data that successfully merge to identifiable counties in the CPS. The table reports the total number of counties in each case, along with a tabulation of when each county was treated. The full FCC 477 data contains information on

[33]We considered defining treatment as occurring once there is at least one provider in the county, regardless of whether all ZIP Codes or all residences have access to a provider, but there is little variation in this treatment variable during our time frame because most counties received their first provider before the FCC data series began (see the top-right panel of Figure 6). We also considered defining treatment with a continuous variable based on the number of providers in a county, but our analysis below is based on two-way fixed effects models with staggered treatment, which have identification challenges that are more difficult to solve with continuous variables (Borusyak et al., 2024; Callaway et al., 2024; Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021). Many strands of empirical research have shifted from using continuous treatment variables to binary ones for this reason. See, e.g., the recent minimum wage literature (Cengiz et al., 2019; Clemens & Strain, 2021; Hampton & Totty, 2023).

[34]The leading and lagging years allow us to have balanced time windows for event study figures later in the paper, but they also present a challenge for classifying treatment status before the FCC 477 started in 1999 and after it ended in 2008. We feel confident leaving zero counties as fully treated before 1999, given that the rollout of broadband internet did not begin until the late 1990s. It is more challenging to determine how to handle counties in the merged FCC-CPS dataset that were still untreated as of June 2008. In our main analysis, we assume these counties were treated by the end of 2008. This seems like the most reasonable "blanket" treatment assumption, given that all identifiable counties in the CPS have large populations and therefore likely had a reasonable level of internet availability. As a robustness check, Table B1 in Appendix B shows an alternate version of our main results that drops all CPS data after June 2008 in order to avoid the issue altogether.

[35]The CPS public microdata file suppresses county identifiers for counties with a population below 100,000.

3,222 total counties, whereas the linked data contains information on 333 counties. Figure 7 shows the geographic dispersion of the staggered treatment timing in the full FCC 477 data, while Figure 8 shows the dispersion in the merged data.

4.3 Item refusal and economic well-being

As described in Section 3.4, we also evaluate heterogeneity in the impact of broadband internet on survey refusal by item sensitivity and proxies for willingness to pay. For item sensitivity, we focus on item-level refusal (rather than unit refusal) among households who responded to some portion of the survey. We study refusal of householder age and household income. The income question is likely viewed by many individuals as a private and sensitive piece of information, while age is likely viewed as less private and sensitive. Income directly impacts many economic and societal household outcomes. Income is also less likely to be publicly available and/or easily estimated. The idea that income is considered a more private and sensitive topic than age is consistent with the fact that the refusal rate for household income is much higher than that of age, as seen in Figure 9.[36]

We produce several different economic well-being measures as proxies for willingness to pay. While willingness to pay is product-specific and person-specific, it is also known to correlate with financial resources for many products. Many goods and services offered by firms that have market power and/or attempt to price discriminate are known to have positive income elasticity of demand, meaning that firms are likely to charge more to consumers with higher levels of income. Examples include colleges who may have market power due to location, prestige, or fields of study and who target financial aid based on family income; airlines who may have market power due to limited competition in origin-destination pairs and who charge dynamic prices (based on customer type, flight demand, and purchase timing) and offer tiered pricing in the form of class-specific ticket fares, airline membership benefits, and lounge access; and online retailers who can target advertising and discounts or charge

[36]For our analysis of income and age refusal, we exclude unit non-response observations, as these observations were not capable of being an item refusal by definition.

personalized prices based on user attributes including income.[37]

There are many well-being proxy variables we could construct and many different approaches to incorporate proxy variables into statistical models. For robustness, we consider four different approaches common in the proxy measurement literature (Lubotsky & Wittenberg, 2006). First, we simply use average total household income (combined from all sources) in each county-year-month. Second, we instrument for the total household income proxy using average household wage and salary income in each county-year-month. Third, we summarize five different county-year-month economic well-being measures (average household income, percent of households with a person who has employer-sponsored health insurance, percent of households not in poverty, percent of households with no members on welfare, and percent of households with no members who have a work-limiting disability) by performing principal component analysis and extracting the first component.[38] Fourth, we include all five of the well-being measures individually and use the sum of their coefficients as the estimate of the effect of willingness to pay. All of the proxy measures were standardized to have a mean of zero and standard deviation of one before use in the regression models.

5 Empirical Methods

To empirically evaluate Hypothesis 1 from Section 3.4, we use a difference-in-differences framework leveraging the staggered rollout of broadband internet services across United States counties:

$$Refused_{hcmt} = \beta Broadband_{cmt} + \alpha_c + \delta_t + \gamma_m + \epsilon_{hcmt}. \tag{3}$$

$Refused_{hcmt}$ is a binary indicator for unit refusal for household h (observed in county c and

[37] See Avery and Hoxby (2004) for discussion of the interaction between price discrimination and income in college financial aid, Aryal et al. (2024) for the airline industry, and Simonovska (2015) for online retail.

[38] The first component loads positively on all five measures, has an eigenvalue of 1.811, and explains 36.24% of the variation that is accounted for by the first five components. The next four components each have an eigenvalue between 0.64 and 0.97 and each explain between 12% and 20% of the remaining variation.

month m of year t). $Broadband_{cmt}$ is a binary indicator for whether county c had broadband availability in all ZIP Codes by month m of year t. The model includes county fixed effects (α_c), year period fixed effects (δ_t), and calendar month fixed effects (γ_m). The residual term is ϵ_{hcmt}.

We first estimate the model using ordinary least squares (OLS).[39] This type of model is commonly referred to as a "two-way fixed effects" (TWFE) model, in the sense that it accounts for fixed effects in both the cross-section dimension of the data (counties) and time dimension of the data (years). The coefficient β represents the estimated change in the probability of household survey refusal after a household's county gains broadband internet. The TWFE methodology in this setting has two key identification assumptions: (1) (conditional) parallel trends between treated and non-treated counties, meaning that the evolution of refusal in treated counties would have mirrored that of non-treated counties in the absence of treatment; and (2) homogeneous treatment effects between units and within units over time. Standard errors are always clustered at the county level.

We also estimate an event study version of (3) that estimates the evolution of refusal before and after gaining broadband internet:

$$Refused_{hcmt} = \sum_{s=-4}^{4} \beta_s Broadband_{cm[t-s]} + \alpha_c + \delta_t + \gamma_m + \epsilon_{hcmt}. \tag{4}$$

The term β_s is the estimated difference in the probability of refusal s periods before/after gaining access to broadband between households treated at time t and non-treated households. Event study results can provide evidence of dynamic effects that play out after exposure. For example, the impact of broadband internet on survey refusal may be delayed as firms improve their tracking capabilities and households become aware of the associated risks. Event studies can also provide evidence for the plausibility of the conditional parallel trends assumption associated with difference-in-differences methods. Evidence of different

[39] We also estimate equation (3) using a logit model. The results are shown in Table B2. We report the OLS results in the main text because they are more comparable to the imputation results from Borusyak et al. (2024) that we discuss below.

trends in refusal in the years before gaining broadband internet access may suggest other differences between treated and non-treated counties that are not attributable to broadband internet access.[40]

The second assumption, homogeneous treatment effects between units and within units over time, is especially unlikely to hold in many empirical settings, in which case the TWFE approach may fail to recover the average treatment effect. While the TWFE approach described in equations (3)-(4) is a common model, it is well-documented that TWFE models with staggered treatment timing estimated via OLS are subject to potential biases that arise due to the fact that treated units can serve as controls for later-treated units (Borusyak et al., 2024; Callaway & Sant'Anna, 2021; De Chaisemartin & d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Sun & Abraham, 2021). In order to evaluate the robustness of our results to this issue, we also report results using a modified TWFE estimation method from Borusyak et al. (2024). The method is based upon the construction and use of "clean controls" rather than the problematic controls described above. Borusyak et al. (2024) achieves this by first estimating the TWFE model on untreated observations only, then using the estimated parameters to impute counterfactual observations for the treated observations.[41]

To empirically evaluate Hypotheses 2 and 3 from Section 3.4, we also evaluate heterogeneity in the impact of broadband internet on survey refusal by item sensitivity and economic well-being. For Hypothesis 2, we add measures of economic well-being, which serve as a proxies for willingness to pay, as covariates and also interact them with the broadband treatment variable:

[40] Visual evidence based on event study TWFE figures estimated via OLS is generally not sufficient to formally reject or fail to reject the assumption of conditional parallel trends, although it is a common check. However, we also use an imputation estimator described below which does formally test for parallel trends in pre-treatment periods of an event study.

[41] The method from Borusyak et al. (2024) has some advantages over other recent approaches. One advantage is that it allows for more flexibility with respect to covariates and complex specifications, including the use of additional fixed effects beyond the so-called "two-way" fixed effects (e.g., calendar month fixed effects). Another advantage is that it allows for the use of not-yet treated units as controls, rather than only using never-treated or last-treated units. This is valuable in our setting since most counties are treated by the end of the time frame and the ones that are not, or are treated last, may be different than counties treated earlier. See De Chaisemartin and d'Haultfoeuille (2023) for additional discussion.

$$Refused_{hcmt} = \beta_1 Broadband_{cmt} + \beta_2 EWB_{cmt} + \beta_3 Broadband_{cmt} * EWB_{cmt} + \alpha_c + \delta_t + \gamma_m + \epsilon_{hcmt}, \quad (5)$$

where EWB_{cmt} represents one of the well-being measures described in Section 4.3. For Hypothesis 3, we switch our analysis from unit refusal to item refusal. We analyze household income refusal and householder age refusal in a pooled regression. We include an indicator for the more sensitive question (household income) and also interact it with the broadband treatment variable:

$$Refused_{qhcmt} = \beta_1 Broadband_{cmt} + \beta_2 Sensitive_q + \beta_3 Broadband_{cmt} * Sensitive_q + \alpha_c + \delta_t + \gamma_m + \epsilon_{qhcmt}, \quad (6)$$

where $Refused_{qhcmt}$ indicates refusal of question q by household h (observed in county c and month m of year t).

6 Empirical Results

6.1 Unit refusal

The difference-in-differences results for unit non-response are shown in Table 2. Columns (1) and (2) show results for the effect of staggered broadband rollout on survey refusal. Columns (3) and (4) show results for the falsification test based on other Type A non-response. Columns (1) and (3) show OLS estimates of equation (3) in the previous section, whereas columns (2) and (4) show estimates using the imputation approach from Borusyak et al. (2024).

The results show an increase in refusal after availability of broadband internet. The OLS coefficient estimate in column (1) is statistically significant at the 1% level. The magnitude of 0.00513 corresponds to approximately a 0.5 percentage point increase in survey refusal, which

is approximately a 10% increase from the sample mean refusal rate of 5.01% over the sample period. The imputation-based estimate in column (2) is also statistically significant at the 1% level and is even larger in magnitude. The coefficient estimate of 0.00853 corresponds to approximately a 0.9 percentage point increase in survey refusal, which is approximately an 18% increase from the sample mean refusal rate of 5.01% over the sample period. Thus, the evidence that exposure to broadband internet increases survey refusal is even stronger when we account for potential biases associated with OLS estimation of the TWFE specification.

The results show no relationship between broadband and other Type A non-response. The OLS coefficient estimate in column (1) is not statistically significant. The magnitude of 0.00349 corresponds to approximately a 0.3 percentage point increase in non-response, which is approximately an 8% increase from the sample mean non-response rate of 3.93% over the sample period. The imputation-based result in column (2) is also not statistically significant and is even smaller in magnitude. The coefficient estimate of 0.00143 corresponds to approximately a 0.1 percentage point increase in non-response, which is a 2.5% increase from the sample mean non-response rate of 3.93% over the sample period. Thus, while broadband exposure is associated with an increase in survey refusal, there is no association with other types of survey-eligible non-response. This suggests that the increase in refusal is due to broadband exposure itself, rather than other coincident factors that could also increase other types of non-response, such as changes in survey instruments or changes in surveyor training.

Next, we turn to the event study results for additional evidence. Figure 10 shows the event study results from equation (4) for survey refusal (left sub-figure) and other Type A non-response (right sub-figure). The results for refusal provide additional evidence supporting a relationship between broadband internet exposure and survey refusal. The leading coefficients are near zero and not statistically significant, which indicates that treated and non-treated counties have no distinguishable differences in survey refusal before broadband exposure and suggests plausibility of the conditional parallel trends assumption. After ex-

posure, the coefficients are consistently positive and statistically different from zero, indicating statistically higher refusal rates in treated counties beginning immediately after full treatment. The coefficients are generally increasing in magnitude with additional years of exposure to broadband internet, indicating that there are also time dynamics causing the likelihood of survey refusal to grow over time. This could be because it takes time for firms to enhance their tracking technologies and/or it takes time for individuals to become aware of the privacy risks associated with broadband internet.

The results for other Type A non-response in the right sub-figure of Figure 10 show no relationship with broadband internet exposure. The leading coefficients are near zero and not statistically significant (except for the OLS results, which are negative and significant for some of the leading time periods). After exposure, all the coefficients remain near zero and are not statistically significant. Collectively, the two figures support the conclusion that broadband internet exposure is causing an increase in survey refusal.

Finally, a convenient feature of the imputation estimator used above is that we can plot a counterfactual rate of refusal (or rate of other Type A non-response) in the absence of broadband internet. The average treatment effect reported earlier in the paper is the average difference between the actual refusal status observations (or other Type A non-response observations) and the imputed counterfactual refusal status observations (or other Type A non-response observations). Instead of averaging the difference between the actual and imputed observations across all individuals and years, we can plot the average rates of actual refusal and imputed refusal over time. This allows us to visualize the cumulative and dynamic effect of broadband internet on the overall refusal rate. Figure 11 plots the actual rates of refusal and other Type A non-response over time as well as their counterfactual rates, based on the sample used in the imputation regressions. For refusal, the actual and counterfactual trends begin to separate in 1999, the first year that some counties are fully exposed to broadband internet based on the FCC 477 data. The actual refusal rate increased every year from 2000 through 2007, starting at 4.2% and ending at 5.5%. The counterfactual

refusal rate was essentially flat over the same time frame, starting at 4.0% and ending at 4.2%. This suggests that nearly all of the increase in unit-refusal during this time can be explained the staggered rollout of high-speed broadband internet. For other Type A non-response, the actual and counterfactual rates are nearly identical.

The results discussed so far support Hypothesis 1 from Section 3.4 that access to broadband internet can serve as a measurable technology shock that increased re-identification risk, and that survey refusal increased after exposure to broadband internet. Next, we explore effect heterogeneity to further assess Hypotheses 2 and 3 described in Section 3.4.

6.2 Effect heterogeneity

The prior section shows evidence that exposure to broadband internet caused an increase in survey refusal, supporting Hypothesis 1. In order to further assess that this relationship is due to changes in privacy loss risk, we now evaluate Hypotheses 2 and 3 from Section 3.4, which test for heterogeneous effects of broadband internet as implied by the model from Section 3.

First, we consider heterogeneity in the effect of broadband exposure on survey refusal by measures of economic well-being. As shown in equation (1) from Section 3.3, the costs of response scale with individual consumer surplus. This is because individuals with larger willingness to pay have more potential consumer surplus to lose from lost privacy. Hypothesis 2 claims that we can use measures of economic well-being to proxy for willingness to pay, and that the impact of broadband exposure on refusal should therefore be larger for those with greater economic well-being. Columns (1) through (4) of Table 3 test this hypothesis using the regression model from equation (5) in Section 5. Column (1) uses mean county-year-month total household income as the proxy for economic well-being. Column (2) instruments for mean county-year-month total household income with mean county-year-month household wage and salary income. Column (3) uses the first principal component from a collection of five county-year-month household economic well-being indicators de-

scribed in Section 4.3. Column (4) includes each of the five well-being indicators in the regression individually and then sums the coefficients from each proxy to get the estimated impact of economic well-being (Lubotsky & Wittenberg, 2006).

The broadband exposure variable is interacted with the proxies to capture the differential impact of broadband exposure for households with greater willingness to pay as proxied for by county-level measures of economic well-being. The well-being measures are also included in the regression separately to adjust for baseline (pre-exposure) differences in refusal associated with economic well-being. The well-being measures are standardized to have a mean of zero, so the broadband variable coefficient by itself now represents the impact of exposure to broadband internet on survey refusal for counties at the mean level of economic well-being. This coefficient is positive and statistically significant at the 1% level for all four columns, indicating that exposure to broadband is associated with increases in survey refusal for counties at the mean level of well-being. The coefficient for the economic well-being variable alone is negative in all four columns and statistically significant in three of them, indicating that baseline refusal rates are somewhat lower for counties with greater well-being measures. This could indicate higher baseline levels of public spirit or trust in government for households with greater economic well-being, among other explanations.

The coefficient for the interaction between broadband exposure and economic well-being is our main variable of interest. The coefficient is positive and statistically significant in all four columns, indicating a larger increase in survey refusal for households located in counties with greater measures of economic well-being. Thus, the results in columns (1) through (4) support Hypothesis 2 that measures of economic well-being can proxy for willingness to pay and therefore increases in survey refusal associated with broadband exposure are larger for counties with greater economic well-being.

Next, we consider item refusal among individuals who responded to at least part of the survey. As described in Hypothesis 3, if increases in refusal are due to privacy concerns, then we should see increases in item refusal that are larger for (or only exist for) items that reveal

more private and sensitive information. Column (5) of Table 3 evaluates this hypothesis using the pooled interaction regression for householder age refusal and household income refusal described in equation (6) from Section 5.

The broadband exposure variable is interacted with an indicator for the household income question in order to capture the differential impact of broadband exposure on refusal of more sensitive questions. The indicator is also included in the regression separately to adjust for baseline (pre-exposure) differences in refusal between the two questions. The coefficient for the broadband variable by itself now represents the impact of exposure to broadband internet for the less sensitive item (householder age) only. This coefficient is negative and statistically significant at the 1% level, indicating that exposure to broadband internet is actually associated with a reduction in refusal of the householder age question. This could indicate that respondents do not view age as a private and sensitive topic and thus other factors besides changing privacy loss risk due to internet exposure are influencing its changing refusal rate. The coefficient for the sensitive household income question is positive and statistically significant at the 1% level, capturing the larger baseline refusal rate for the income question.

The coefficient for the interaction between broadband exposure and the sensitive income question is our main variable of interest. The coefficient is positive and statistically significant at the 1% level, indicating a larger increase in refusal associated with broadband exposure for the household income question. Panel B of Table 3 reports the full effect of broadband internet exposure on item refusal separately for householder age and household income. The effect for householder age is simply the coefficient for the broadband indicator in Panel A of Table 3 (-0.01101, statistically significant at the 1% level). The effect for household income is based on the linear combination of the three coefficients in Panel A compared to the coefficient for just the sensitive income indicator alone. The effect is 0.0143, statistically significant at the 1% level. Thus, the results for item refusal are consistent with Hypothesis 3 that household income is a more private and sensitive topic than householder age and

therefore increases in item refusal associated with broadband internet are driven by refusal of the sensitive (income) item.

7 Implied Empirical Change in Privacy Loss Risk

To further connect the empirical results for survey refusal with the model from Section 3, we perform a back of the envelope calculation to estimate an implied increase in privacy loss risk that would rationalize the observed rise in survey refusal following the rollout of broadband internet. In the model, individuals refuse to respond when the expected loss from additional inference on their willingness to pay exceeds the benefit of participation. As seen in equation (1), this loss is increasing in both the change in posterior inference, $\rho^{post} - \rho^{post}_{-i}$, and the individual's consumer surplus for the product, $w_i - p_e$. Because we cannot observe $\rho^{post} - \rho^{post}_{-i}$ directly, we use the estimated refusal effect (β) from the difference-in-differences regression (0.0085 in column 2 from Table 2) to infer what value of $\rho^{post} - \rho^{post}_{-i}$ would be necessary to explain the observed behavior, given reasonable assumptions about consumer surplus.

Recalling Section 3.4, we model an individual's decision to refuse participation as:

$$Refusal_i = \mathbb{I}\{X_i \geq 0\},$$

with latent index:

$$X_i = \Delta\rho(w_i - p_e) - B + \epsilon_i,$$

where $\Delta\rho$ represents the change in posterior inference ($\rho^{post} - \rho^{post}_{-i}$) and ϵ_i is an idiosyncratic term. Assuming a logistic distribution, $\epsilon_i \sim$ Logistic(0,1), the probability of refusal is:

$$P_i = Pr(Refusal_i = 1) = \frac{1}{1 + e^{-X_i}}, \quad X_i = \Delta\rho(w_i - p_e) - B.$$

Differentiating with respect to $\Delta\rho$ gives:

$$\frac{\partial P_i}{\partial \Delta\rho} = \frac{\partial P_i}{\partial X_i}\frac{\partial X_i}{\partial \Delta\rho} = f(X_i) * (w_i - p_e), \quad f(X_i) = \frac{e^{-X_i}}{(1+e^{-X_i})^2}.$$

For a small change in $\Delta\rho$, the linear approximation for the change in probability of refusal is:

$$\Delta P_i \approx \frac{\partial P_i}{\partial \Delta\rho} * \Delta\rho = f(X_i) * (w_i - p_e) * \Delta\rho.$$

In our difference-in-differences regression, β measures the average change in refusal probability due to the rollout of broadband internet. Interpreting β as ΔP_i averaged over individuals gives:

$$\beta \approx f(X) * (w - p_e) * \Delta\rho,$$

where $w - p_e$ is a representative value of consumer surplus and $f(X)$ is the logistic density evaluated at a representative level. Solving for $\Delta\rho$ gives:

$$\Delta\rho = \frac{\beta}{f(X) * (w - p_e)}. \tag{7}$$

To make the interpretation concrete, we consider a few representative products and services whose consumer surplus has been estimated in the literature and we evaluate the logistic density at the participation threshold, $f(X = 0) = \frac{1}{4}$. First, we consider airline tickets, which is likely among the more expensive products that many consumers purchase with some regularity. Aryal et al. (2024) estimates that the average consumer surplus per airline flight from 2009-2011, assuming pricing strategies that leave airlines unable to segment passengers based on willingness to pay, was \$145. Plugging $\beta = 0.0085$, $f(X) = 0.25$, and $(w - p_e)$=145 into equation (7), we get $\Delta\rho = 0.00023$. Next, we consider online retail shopping. Farronato et al. (2025) estimates that consumer surplus per search on Amazon for products from six common categories (health, paper products, household items, apparel, electronics, and personal care) is \$3.12, assuming no personalization of pricing or search

results. Plugging $\beta = 0.0085$, $f(X) = 0.25$, and $(w - p_e)$=3.12 into equation (7), we get $\Delta\rho = 0.01090$.

The interpretation of $\Delta\rho$ is the percentage point change in the likelihood that the firm will be able to infer a consumer's willingness to pay. Thus, our estimate of this change ranges from 0.023 percentage points for airline tickets to 1.09 percentage points for products frequently purchased on Amazon. These implied changes would increase modestly if the estimates of consumer surplus for airline tickets and Amazon products were deflated to price levels corresponding to the time range of broadband expansion. They would also increase modestly for small reductions in the assumed value of $f(X)$.

Figure 12 shows how the implied change in willingness to pay inference needed in order to rationalize the estimated increase in survey refusal changes with consumer surplus and with $f(X)$. The figure illustrates that as consumer surplus increases, a smaller value of $\Delta\rho$ is needed in order to rationalize the increase in survey refusal. For consumer surplus of \$1, the implied change in inference likelihood is nearly 3.5 percentage points, whereas at \$10 the implied change is below 0.5 percentage points and continues to asymptote toward zero for larger values of consumer surplus. The figure also illustrates that our implied values of $\Delta\rho$ are not very sensitive to modest changes in the assumed value of $f(X)$. Assuming an average consumer surplus of \$3 (similar to the estimate from Farronato et al. (2025)), the implied change only ranges from approximately 0.9 to 1.3 percentage points as $f(X)$ ranges from 0.2 to 0.3.

8 Conclusion

We evaluate the relationship between privacy and rising survey refusal rates. We first introduce a model of an individual's decision to respond to a federal statistical agency's survey. In the model, individuals have the choice of whether to respond to the survey, the statistical agency uses individuals' survey responses to disseminate public data, and a monopolistic

firm attempts to price discriminate by inferring individuals' willingness to pay for its product using its own database and the agency's public data. If the individual is in the agency's database, then they face additional inference risk due to the possibility that the firm will re-identity them in the data and learn their information directly. The decision rule for individual survey response is a function of both the marginal change in probability of having their type revealed to the firm due to existing in the agency's database and the individual's willingness to pay for the monopolistic firm's product. We then evaluate the relationship between privacy and survey refusal empirically by using the rollout of broadband internet as a privacy-reducing technology shock that enhanced the ability of firms to infer individuals' willingness to pay and offer personalized pricing. The staggered rollout of broadband internet across the United States provides a natural experiment for evaluating the evolution of survey refusal around the time of broadband rollout.

We find that full survey refusal (i.e., unit refusal) in the CPS increased when broadband internet was made available in a given county. The increase in refusal grows over time following initial exposure. The rollout of broadband internet can explain nearly all of the rise in survey refusal between 1995 and 2012. There was no increase in other types of unit non-response associated with broadband exposure, suggesting that the increase in refusal was not driven by other coincident factors.

We test two additional implications of our model to evaluate changing privacy risk as the underlying mechanism. One implication is that increases in refusal associated with broadband internet should be larger for individuals with larger willingness to pay, as these individuals have more to lose from firms inferring their willingness to pay. We find support for this using county-level measures of household economic well-being as proxies for willingness to pay. The other implication is that, among individuals who do respond to some portion of the survey, broadband exposure should be associated with an increase in item refusal that is concentrated among items that reveal more private and sensitive information. We find support for this comparing the impact of broadband exposure on refusal of household income

versus householder age.

Future work should consider other important ramifications of privacy and confidentially for survey response. Our results suggest that the rollout of broadband internet can explain a large fraction of the rise in survey refusal from 2000 through 2008, but survey refusal rates continued to rise after the time frame of our analysis. Many societal factors impacting privacy and confidentiality also evolved after the time frame of our analysis, such as the proliferation of data breaches. There is also the question of how privacy and confidentially concerns have impacted survey data *accuracy*, rather than *response rates*. Finally, many statistical agencies have begun to modernize their disclosure avoidance efforts in recent years and future work should assess the impact of these modernization efforts on refusal rates and data accuracy.

References

ABC News. (2000). Amazon error may end 'dynamic pricing'. *ABC News*. https://abcnews.go.com/Technology/story?id=119399&page=1

Abowd, J. M., Adams, T., Ashmead, R., Darais, D., Dey, S., Garfinkel, S. L., Goldschlag, N., Kifer, D., Leclerc, P., Lew, E., et al. (2023). *The 2010 Census confidentiality protections failed, here's how and why* [NBER Working Paper No. 31995].

Abowd, J. M., & Hawes, M. B. (2024). 21st century statistical disclosure limitation: Motivations and challenges. *Handbook of sharing confidential data* (pp. 24–36). Chapman; Hall/CRC.

Abowd, J. M., & Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, *109*(1), 171–202.

Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2022). Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics*, *14*(4), 218–256.

Acquisti, A., & Gross, R. (2009). Predicting social security numbers from public data. *Proceedings of the National Academy of Sciences*, *106*(27), 10975–10980.

Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature*, *54*(2), 442–92.

Aparicio, D., Metzman, Z., & Rigobon, R. (2024). The pricing strategies of online grocery retailers. *Quantitative Marketing and Economics*, *22*(1), 1–21.

Argenziano, R., & Bonatti, A. (2023). Data markets with privacy-conscious consumers. *AEA Papers and Proceedings*, *113*, 191–196.

Aryal, G., Murry, C., & Williams, J. W. (2024). Price discrimination in international airline markets. *Review of Economic Studies*, *91*(2), 641–689.

Atasoy, H. (2013). The effects of broadband internet expansion on labor market outcomes. *ILR Review*, *66*(2), 315–345.

Avery, C., & Hoxby, C. M. (2004). Do and should financial aid packages affect students' college choices? *College choices: The economics of where to go, when to go, and how to pay for it* (pp. 239–302). University of Chicago Press.

Belleflamme, P., Lam, W. M. W., & Vergote, W. (2020). Competitive imperfect price discrimination and market power. *Marketing Science*, *39*(5), 996–1015.

Belleflamme, P., & Vergote, W. (2016). Monopoly price discrimination and privacy: The hidden cost of hiding. *Economics Letters*, *149*, 141–144.

Bian, B., Pagel, M., Rava, D., & Tang, H. (2024). *Consumer surveillance and financial fraud* [NBER Working Paper 31692].

BLS. (1999). *Computer ownership up sharply in the 1990s* [U.S. Department of Labor Bureau of Labor Statistics]. https://www.bls.gov/opub/btn/archive/computer-ownership-up-sharply-in-the-1990s.pdf

Borgschulte, M., Cho, H., & Lubotsky, D. (2022). Partisanship and survey refusal. *Journal of Economic Behavior & Organization*, *200*, 332–357.

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, *91*(6), 3253–3285.

Brick, J. M., & Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 36–59.

Callaway, B., Goodman-Bacon, A., & Sant'Anna, P. H. (2024). *Difference-in-differences with a continuous treatment* [NBER Working Paper 32117].

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230.

Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics*, *134*(3), 1405–1454.

Census Bureau. (2023). *Current Population Survey (CPS) 2023 modernization efforts* [U.S. Census Bureau]. https://www.census.gov/programs-surveys/cps/about/modernization.html

Chicago Tribune. (2000). Amazon.com's variable pricing draws ire. *Chicago Tribune*. https://www.chicagotribune.com/2000/10/09/amazoncoms-variable-pricing-draws-ire/

Chicago Tribune. (2018). Are you ready for personalized pricing? *Chicago Tribune*. https://www.chicagobooth.edu/review/are-you-ready-personalized-pricing

Clemens, J., & Strain, M. R. (2021). *The heterogeneous effects of large and small minimum wage changes: Evidence over the short and medium run using a pre-analysis plan* [NBER Working Paper 29264].

Council of Economic Advisors. (2015). *Big data and differential pricing* [The White House, United States].

De Chaisemartin, C., & d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–96.

De Chaisemartin, C., & d'Haultfoeuille, X. (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal*, *26*(3), C1–C30.

Dettling, L. J. (2017). Broadband in the labor market: The impact of residential high-speed internet on married women's labor force participation. *ILR Review*, *70*(2), 451–482.

Dettling, L. J., Goodman, S., & Smith, J. (2018). Every little bit counts: The impact of high-speed internet on the transition to college. *Review of Economics and Statistics*, *100*(2), 260–273.

Dinterman, R., & Renkow, M. (2017). Evaluation of USDA's broadband loan program: Impacts on broadband provision. *Telecommunications Policy*, *41*(2), 140–153.

Evans, D. S. (2009). The online advertising industry: Economics, evolution, and privacy. *Journal of Economic Perspectives*, *23*(3), 37–60.

Falck, O., Gold, R., & Heblich, S. (2014). E-lections: Voting behavior and the internet. *American Economic Review*, *104*(7), 2238–2265.

Farronato, C., Fradkin, A., & MacKay, A. (2025). *Vertical integration and consumer choice: Evidence from a field experiment* [Working paper]. https://andreyfradkin.com/assets/amazon_exp_welfare.pdf

Fobia, A. C., Mathews, K., & Terry, R. (2025). *Privacy and confidentiality concerns in the 2020 Census* [U.S. Census Bureau report]. https://www2.census.gov/programs-surveys/decennial/2020/program-management/evaluate-docs/EAE-2020-privacy-confidentiality-concerns.pdf

Food Dive. (2017). Kroger's analytics and personalized pricing keep it a step ahead of its competitors. *Food Dive*. https://www.fooddive.com/news/krogers-analytics-and-personalized-pricing-keep-it-a-step-ahead-of-its-com/446685/

Garfinkel, S., Abowd, J. M., & Martindale, C. (2019). Understanding database reconstruction attacks on public data. *Communications of the ACM*, *62*(3), 46–53.

Goldfarb, A., & Que, V. F. (2023). The economics of digital privacy. *Annual Review of Economics*, *15*.

Goldfarb, A., & Tucker, C. (2012). Shifts in privacy concerns. *American Economic Review*, *102*(3), 349–53.

Goldfarb, A., & Tucker, C. (2019). Digital economics. *Journal of Economic Literature*, *57*(1), 3–43.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277.

Greenstein, S. (2000). Commercialization of the internet: The interaction of public policy and private choices or why introducing the market worked so well. *Innovation Policy and the Economy*, *1*, 151–186. https://doi.org/10.1086/ipe.1.25056144

Hampton, M., & Totty, E. (2023). Minimum wages, retirement timing, and labor supply. *Journal of Public Economics*, *224*(104924).

Hannak, A., Soeller, G., Lazer, D., Mislove, A., & Wilson, C. (2014). Measuring price discrimination and steering on e-commerce web sites. *Proceedings of the 2014 Conference on Internet Measurement Conference*, 305–318.

Heffetz, O., & Ligett, K. (2014). Privacy and data-based research. *Journal of Economic Perspectives*, *28*(2), 75–98.

Henriksen-Bulmer, J., & Jeary, S. (2016). Re-identification attacks—a systematic literature review. *International Journal of Information Management*, *36*(6), 1184–1192.

Jarmin, R. S. (2019). Evolving measurement for an evolving economy: Thoughts on 21st century U.S. economic statistics. *Journal of Economic Perspectives*, *33*(1), 165–184.

Johnson, K. R., & Persico, C. (2024). *Broadband internet access, economic growth, and wellbeing* [NBER Working Paper 32517].

Kolko, J. (2012). Broadband and local growth. *Journal of Urban Economics*, *71*(1), 100–113.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805.

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolff, S. (2009). A brief history of the internet. *ACM SIGCOMM Computer Communication Review*, *39*(5), 22–31.

Linse, K., & Johnson, N. (2023). *Current Population Survey (CPS) modernization* [U.S. Department of Labor Bureau of Labor Statistics]. https://apps.bea.gov/fesac/meetings/2023-12-08/Johnson-Linse.pdf

Los Angeles Times. (2000). Amazon pays a price for marketing test. *Los Angeles Times*. https://www.latimes.com/archives/la-xpm-2000-oct-02-fi-30029-story.html

Lubotsky, D., & Wittenberg, M. (2006). Interpretation of regressions with multiple proxies. *The Review of Economics and Statistics*, *88*(3), 549–562.

McClure, D., & Reiter, J. P. (2012). Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Trans. Data Priv.*, *5*(3), 535–552.

Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *29*(4), 199–226.

Mikians, J., Gyarmati, L., Erramilli, V., & Laoutaris, N. (2012). Detecting price and search discrimination on the internet. *Proceedings of the 11th ACM workshop on hot topics in networks*, 79–84.

Miklós-Thal, J., Goldfarb, A., Haviv, A. M., & Tucker, C. (2023). *Digital hermits* [NBER Working Paper 30920].

Mohammed, R. (2017). How retailers use personalized prices to test what you're willing to pay. *Harvard Business Review*. https://hbr.org/2017/10/how-retailers-use-personalized-prices-to-test-what-youre-willing-to-pay

National Research Council. (2013). *Nonresponse in social science surveys: A research agenda*. The National Academies Press.

Odlyzko, A. (2003). Privacy, economics, and price discrimination on the internet. *Proceedings of the 5th international conference on electronic commerce* (pp. 355–366).

PEW. (2019). *Americans and privacy: Concerned, confused and feeling lack of control over their personal information*. https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

Reiter, J. P. (2005). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, *100*(472), 1103–1112.

Rhodes, A., & Zhou, J. (2024). Personalized pricing and competition. *American Economic Review*, *114*(7), 2141–2170.

Rivera Drew, J. A., Flood, S., & Warren, J. R. (2014). Making full use of the longitudinal design of the Current Population Survey: Methods for linking records across 16 months. *Journal of Economic and Social measurement*, *39*(3), 121–144.

Ross, C. V. (2023). *Uses of Decennial Census programs data in federal funds distribution: Fiscal year 2021* [U.S. Census Bureau Working Paper].

Shiller, B. R. (2020). Approximating purchase propensities and reservation prices from broad consumer tracking. *International Economic Review*, *61*(2), 847–870.

Simonovska, I. (2015). Income differences and prices of tradables: Insights from an online retailer. *The Review of Economic Studies*, *82*(4), 1612–1656.

Singer, E., Schaeffer, N. C., & Raghunathan, T. (1997). Public attitudes toward data sharing by federal agencies. *International Journal of Public Opinion Research*, *9*(3), 277–285.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*, 175–199.

The Wall Street Journal. (2000). Websites vary prices, deals based on users' information. *The Wall Street Journal*. https://www.wsj.com/articles/SB10001424127887323777204578189391813881

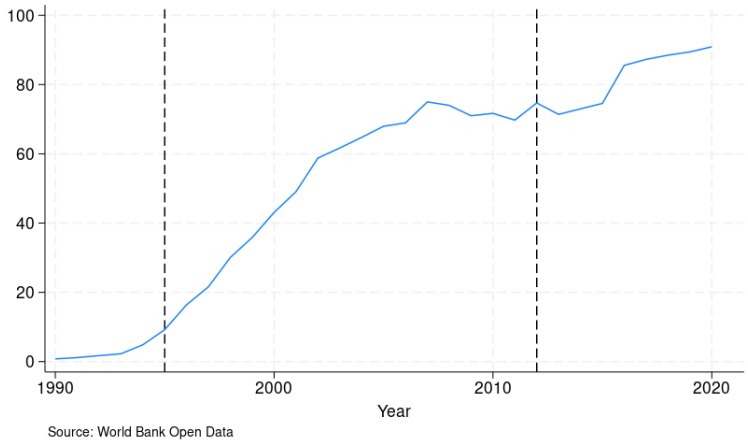
Van Parys, J., & Brown, Z. Y. (2023). *Broadband internet access and health outcomes: Patient and provider responses in medicare* [NBER Working Paper 31579].

Varian, H. R. (2010). Computer mediated transactions. *American Economic Review*, *100*(2), 1–10.

Williams, D., & Brick, J. M. (2018). Trends in U.S. face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, *6*(2), 186–211.

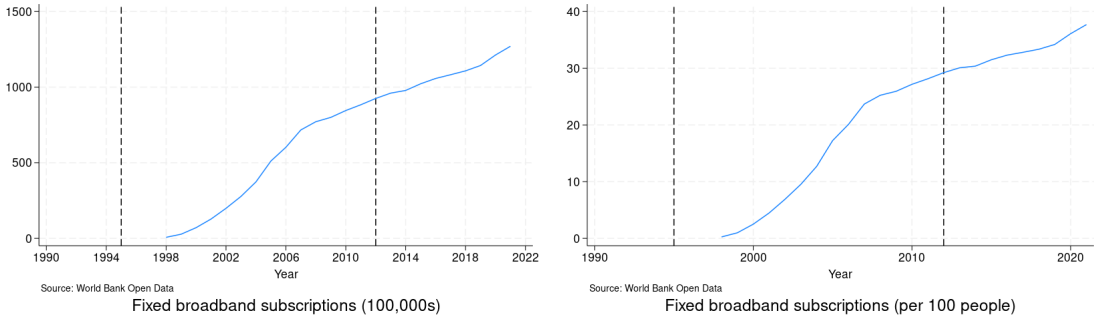
Figures and Tables

Figure 1: Internet Usage Among the United States Population



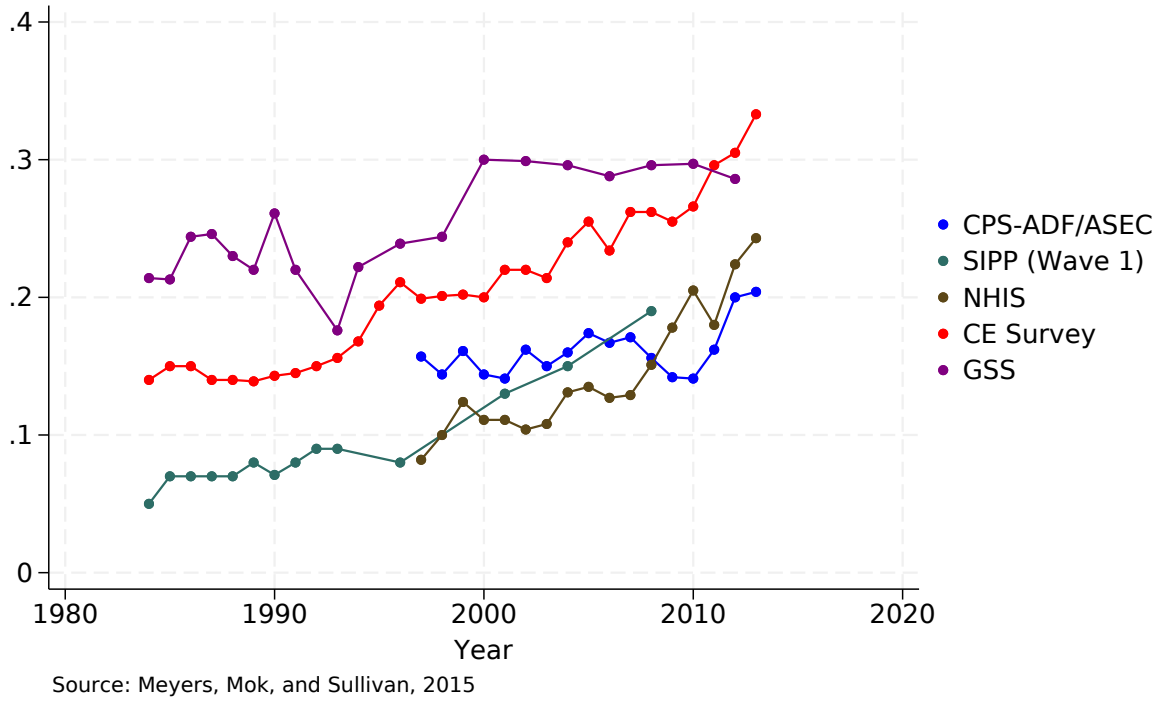
Notes: Percent of individuals using the internet in the United States over time. Vertical lines correspond to the beginning and end of our time frame of analysis.

Figure 2: Broadband Subscription Service Growth



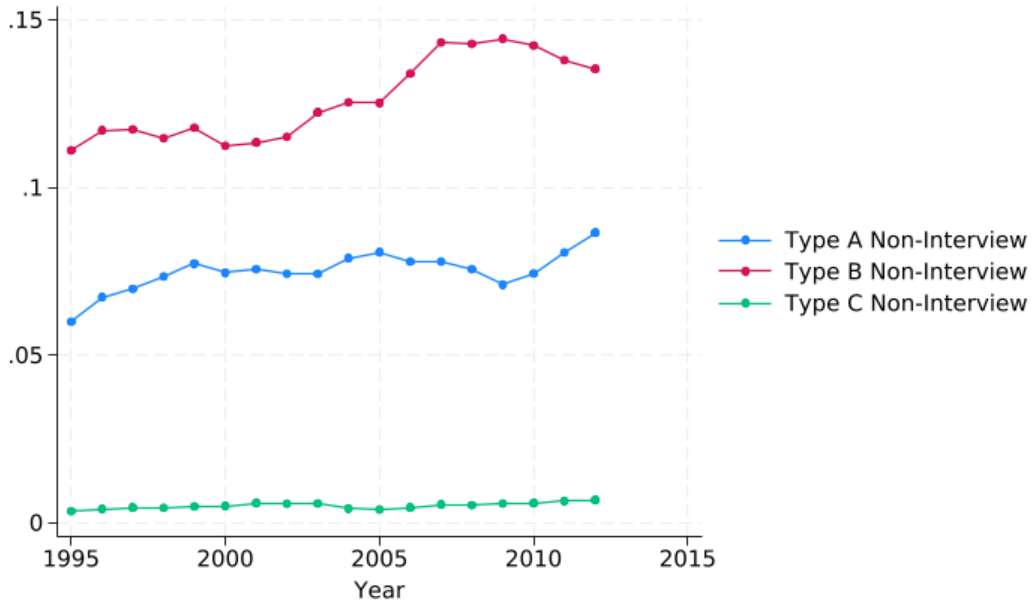
Notes: Number of fixed broadband subscriptions over time over time in the United States. Vertical lines correspond to the beginning and end of our time frame of analysis.

Figure 3: Refusal Rates Across Major Household Surveys



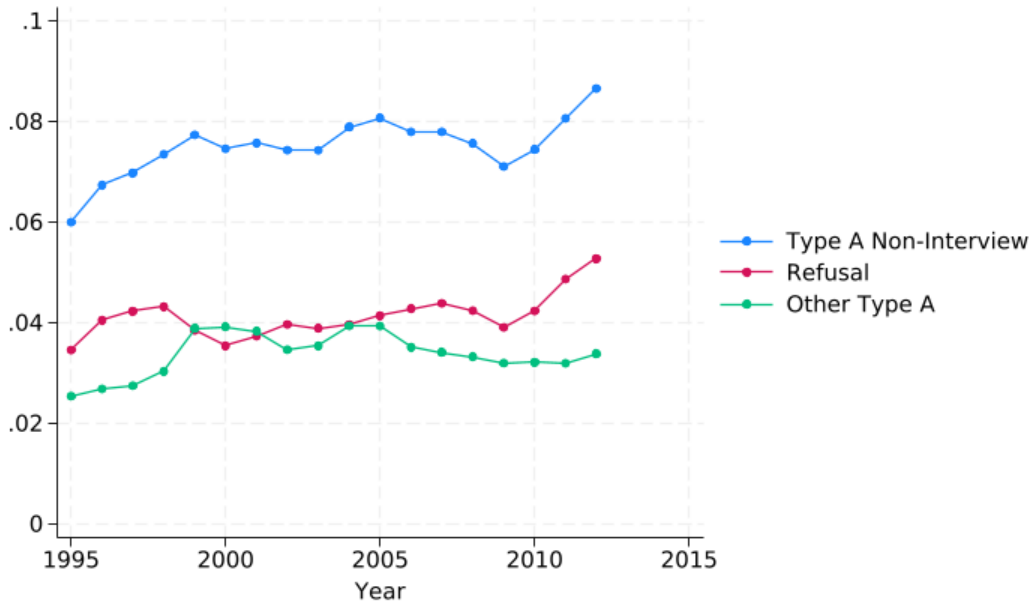
Notes: Refusal rates for the Current Population Survey Annual Demographic File/Annual Social and Economic Supplement (CPS), the Survey of Income and Program Participation (SIPP), the Consumer Expenditure (CE) Survey, the National Health Interview Survey (NHIS), and the General Social Survey (GSS). See Meyers, Mok, and Sullivan (2015) for additional details.

Figure 4: CPS Non-Response Rates by Type, 1995-2012



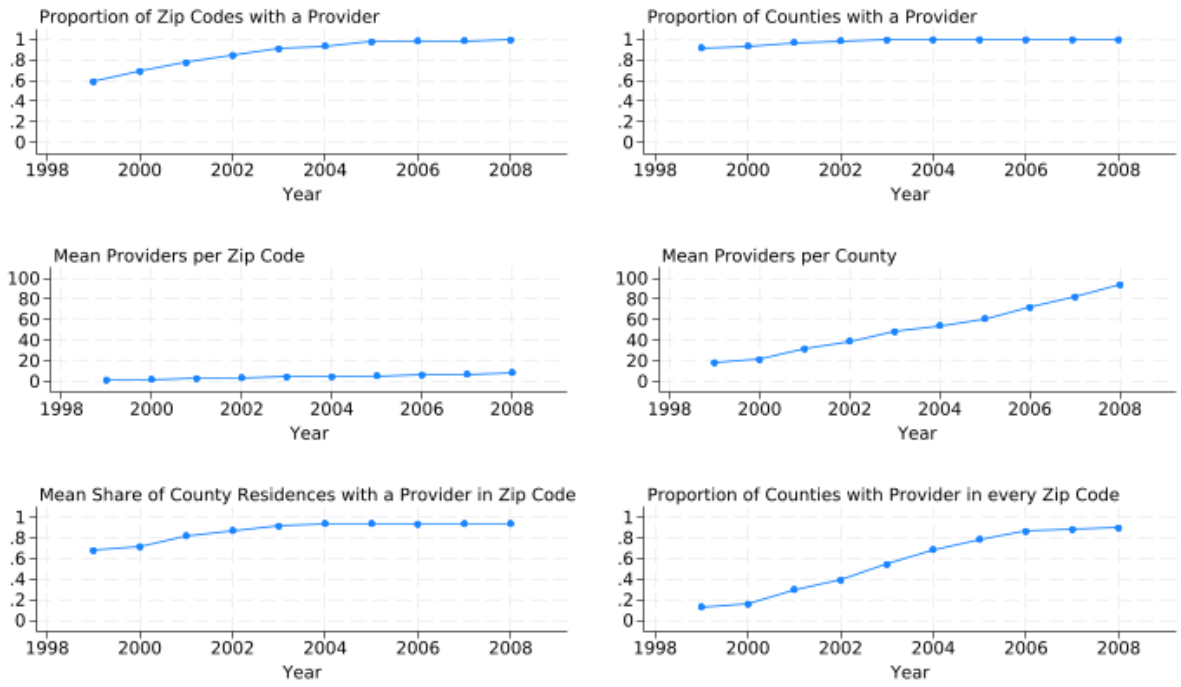
Notes: Type A non-interviews correspond to survey-eligible households that could not be interviewed due to reasons such as refusal, absence from the home, language barriers, weather, illness, or inability to locate the address. Type B non-interviews correspond to households that were ineligible for interview because they were unoccupied or occupied solely by persons not eligible for interview. Type C non-interviews correspond to households that were ineligible for interview because they have been demolished or converted into a non-residential address.

Figure 5: Type A Non-Response and its Components, 1995-2012



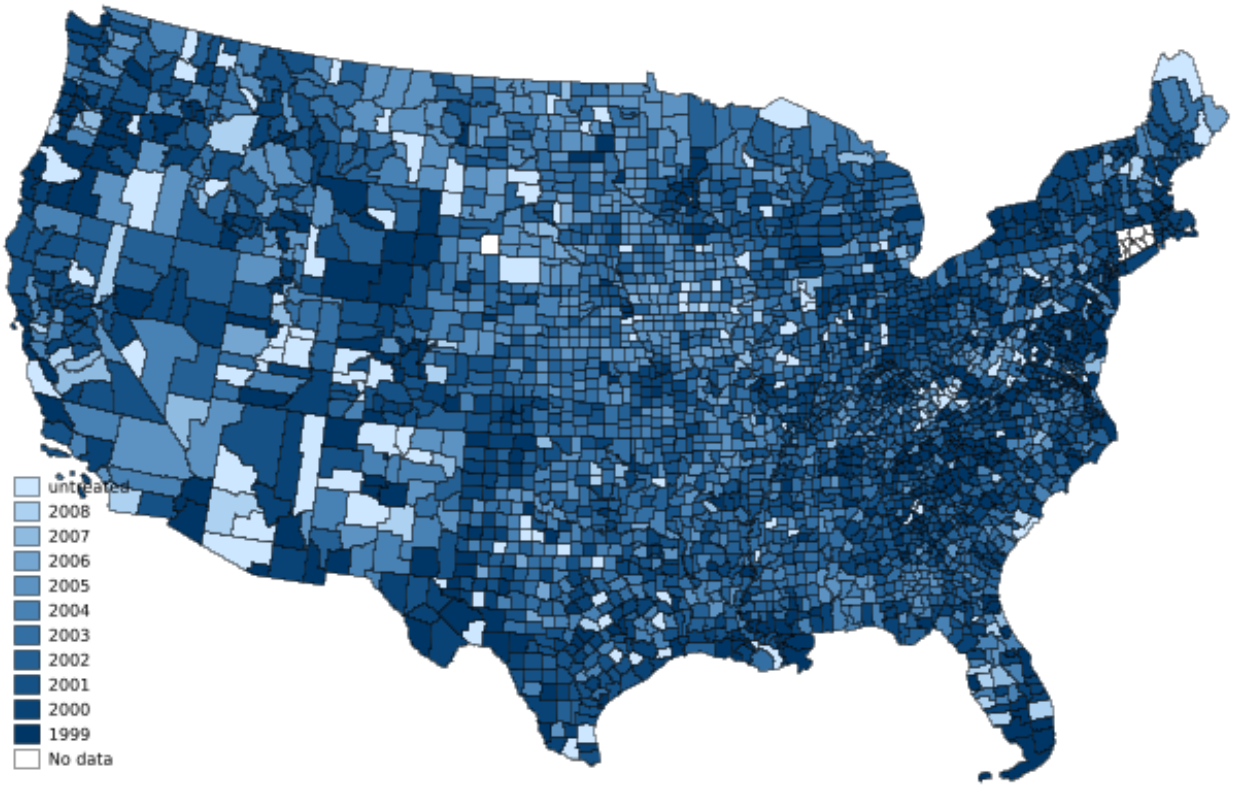
Notes: Type A non-interviews correspond to survey-eligible households that could not be interviewed due to reasons such as refusal, absence from the home, language barriers, weather, illness, or inability to locate the address. We decompose the Type A non-interviews into refusals and all other Type A non-interviews.

Figure 6: Broadband Internet Availability Over Time in FCC 477 Data



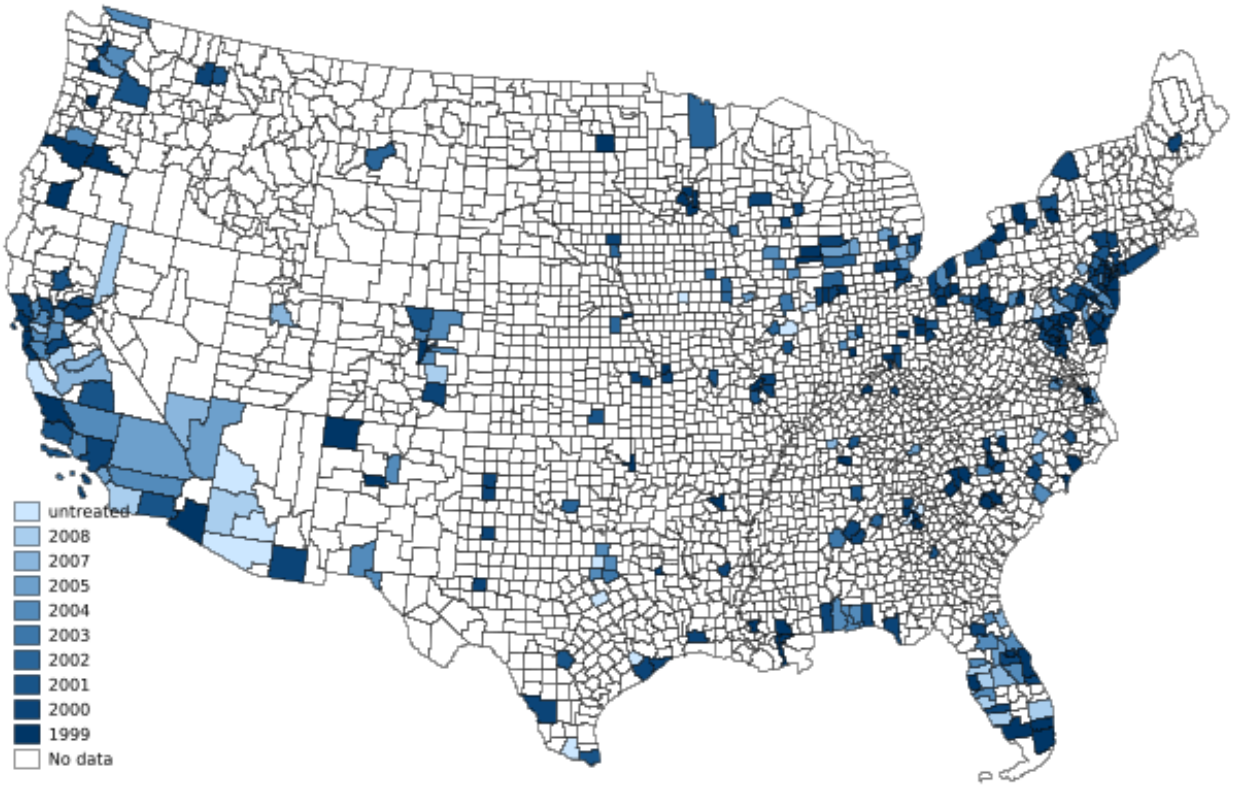
Notes: Figures are based on the Federal Communication Commission's Form 477 data (FCC 477). FCC 477 is a mandated form submitted biannually to all United States internet service providers in order to document their broadband infrastructure's geographic coverage. The public data reports the number of broadband providers at the ZIP Code level from 1999-2008. We aggregated the data to the county level using ZIP Code-county crosswalks from the U.S. Department of Housing and Urban Development (HUD).

Figure 7: First Year of County-Level Broadband Internet Coverage in FCC 477 Data



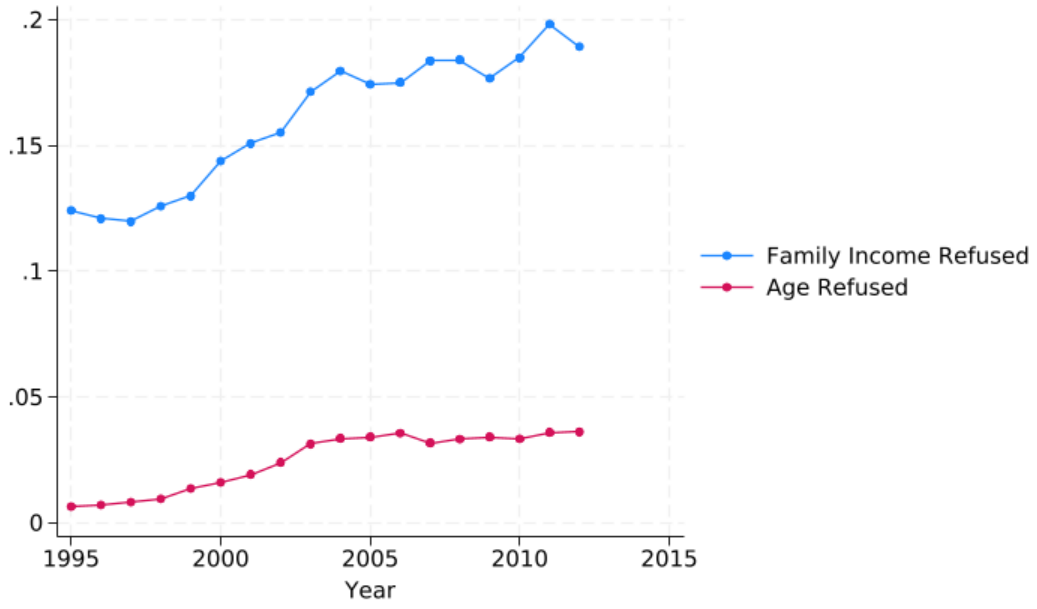
Notes: Figure is based on the Federal Communication Commission's Form 477 data (FCC 477). The year of coverage shown in the figure is first year during which all ZIP Codes in the county had a broadband internet service provider. Counties without full coverage by 2008 are labeled "untreated."

Figure 8: First Year of County-Level Broadband Internet Coverage in Merged FCC-CPS Data



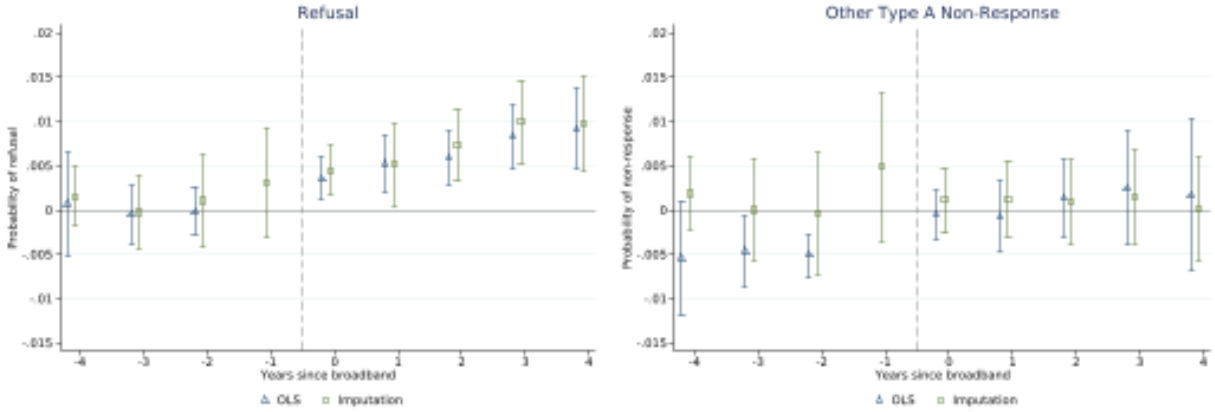
Notes: Figure is based on merged data between the Federal Communication Commission's Form 477 data (FCC 477) and the Current Population Survey. The year of coverage shown in the figure is first year during which all ZIP Codes in the county had a broadband internet service provider. Counties without full coverage by 2008 are labeled "untreated." The merged data contain fewer counties because the CPS suppresses county identifiers for counties with fewer than 100,000 people

Figure 9: Selected Item Refusal Rates, 1995-2012



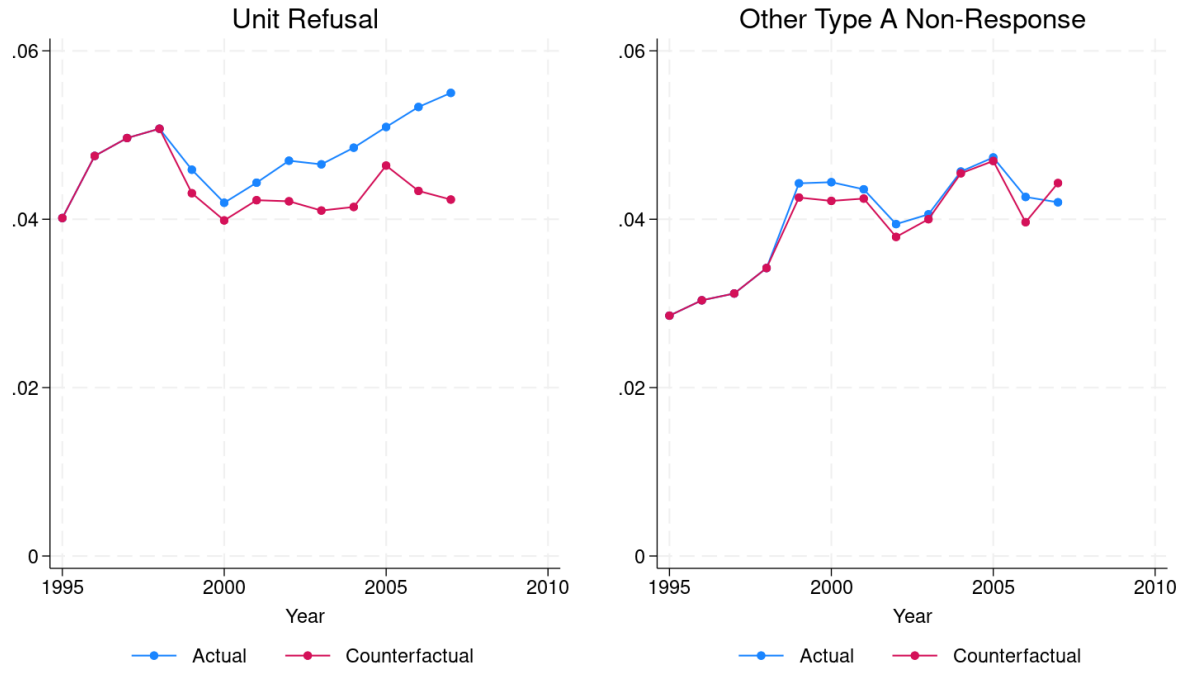
Notes: Household income refusal is based on the question that asks the household reference person to provide the total income of all family members in the household. Age refusal is based on the question that asks the reference person their age.

Figure 10: Event Study Results for Broadband Internet Exposure and Unit Non-Response



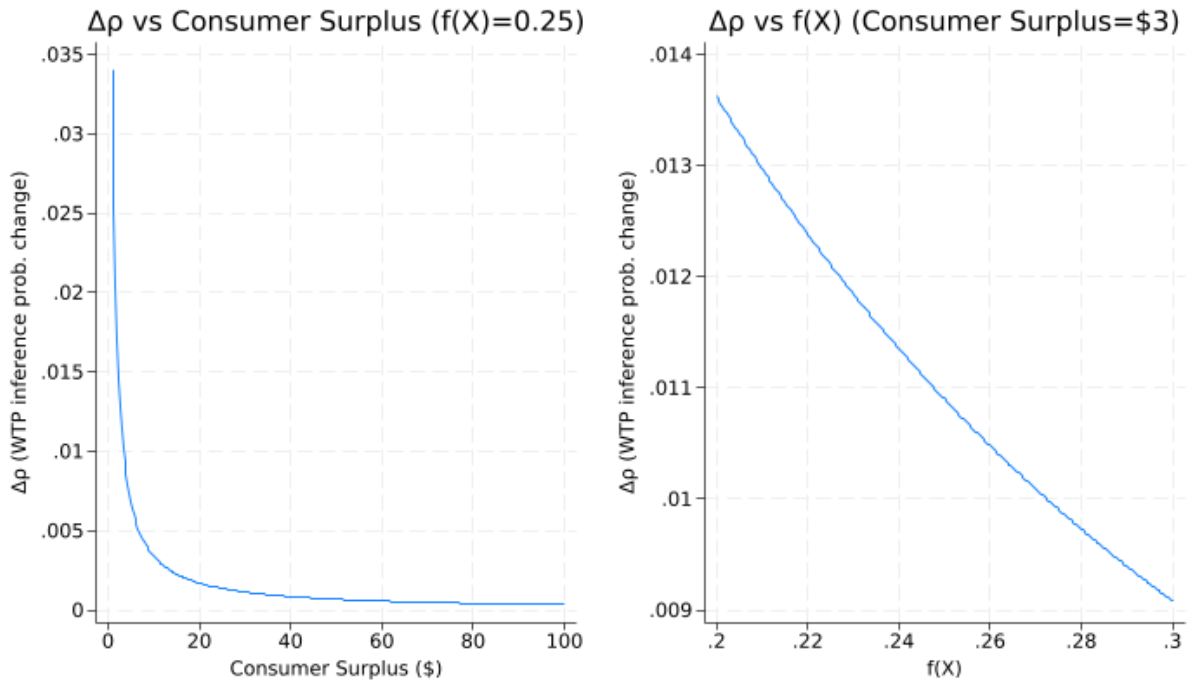
The outcome variable in the regressions is either an indicator for refusal of the entire survey (left sub-figure) or other Type A non-response (right sub-figure). "OLS" refers to ordinary least squares estimates of the regression specification from equation (4) of the main text. "Imputation" refers to estimation of equation (4) using the imputation-based estimator from Borusyak et al. (2024). Standard errors are clustered at the county level. The capped lines represent 95% confidence intervals.

Figure 11: Counterfactual Rates of Refusal and Other Type A Non-Response in the Absence of Broadband Internet



The figure shows the actual and counterfactual rates of refusal (left sub-figure) and other Type A Non-Response (right sub-figure) in the CPS basic monthly data. The counterfactual rate is based on the imputation estimator from Borusyak et al. (2024). See Section 6.1 for more details.

Figure 12: Implied Change in Inference Probability Needed to Rationalize the Increase in Refusal



The figure shows the implied change in privacy loss risk (firm inference of individuals' willingness to pay) from equation (1) caused by the rollout of broadband internet, based upon the estimated change in survey refusal caused by the rollout of broadband internet. See Section 7 for more details.

Table 1: In-Sample County Counts By Broadband Treatment Year

Treatment Year	FCC Data	Merged FCC-CPS Data
1999	419	105
2000	473	77
2001	449	24
2002	441	18
2003	406	3
2004	349	48
2005	348	17
2006	31	0
2007	49	12
2008	30	13
Untreated by June 2008	146	16
Total Counties	3,222	333

Notes: The table reports the total number of counties in the FCC 477 data and in the FCC 477 data merged to the CPS, tabulated by the first year during which all ZIP Codes in the county had at least one broadband internet provider. The FCC 477 data are based on all counties in the United States and Washington, D.C. The merged data contain fewer counties because the CPS suppresses county identifiers for counties with fewer than 100,000 people.

Table 2: Difference-in-Differences Results for Broadband Internet Exposure and Unit Non-Response

	Unit Refusal		Other Unit Non-Response	
	(1) OLS	(2) Imputation	(3) OLS	(4) Imputation
Broadband	0.00513***	0.00853***	0.00349	0.00143
	(0.00144)	(0.00205)	(0.00249)	(0.00239)
County fixed effects	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes	Yes
Observations	4,575,447	4,575,447	4,762,779	4,762,779

Notes: The outcome variable in the regressions is an indicator for either refusal of the entire survey (columns (1) and (2)) or other Type A non-response besides refusal (columns (3) and (4)). "OLS" refers to ordinary least squares estimates of the regression specification from equation (3) of the main text. "Imputation" refers to estimation of equation (3) using the imputation-based estimator from Borusyak et al. (2024). Standard errors are clustered at the county level.

Table 3: Effect Heterogeneity by Consumer Surplus and Item Sensitivity

	Unit Refusal				Item Refusal
	(1)	(2)	(3)	(4)	(5)
Panel A: Model coefficients					
Broadband	0.00548***	0.00382***	0.00469***	0.00498***	-0.0110***
	(0.00159)	(0.00134)	(0.00161)	(0.00164)	(0.00407)
Surplus	-0.00196	-0.00395**	-0.00232*	-0.00309*	
	(0.00130)	(0.00152)	(0.00119)	(0.00163)	
Broadband x Surplus	**0.00343*** **	**0.00537*** **	**0.00272** **	**0.00354** **	
	(0.00127)	**(0.00160)**	**(0.00123)**	**(0.00158)**	
Sensitive					0.1086***
					(0.00487)
Broadband x Sensitive					**0.0253*** **
					(0.00560)
Panel B: Effect of broadband exposure					
On householder age refusal					-0.0110***
					(p-val=0.007)
On household income refusal					0.0143***
					(p-val=0.006)
Consumer surplus proxy method	OLS	2SLS	PCA	Multi	
County fixed effects	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	4,467,994	4,467,994	4,467,994	4,467,994	8,573,164

Notes: Columns (1) through (4) interact exposure to broadband internet with proxies for consumer surplus based on economic well-being. Column (1) uses mean county-year-month total household income ("OLS"). Column (2) instruments for total household income with household wage and salary income ("2SLS"). Column (3) uses the first component from principal component analysis applied to a collection of five household-level economic variables averaged by county-year-month ("PCA"). Column (4) includes each of the five household economic variables in the regression individually and then sums the coefficients from each proxy to get the estimated impact of consumer surplus. ("Multi"). See Section 4.3 for more information on the construction of the proxies. Column (5) evaluates item refusal in a pooled regression for both householder age and household income refusal. We interact exposure to broadband internet with an indicator for the household income question, which is a more sensitive topic and thus more likely to induce changes in re-identification risk relative to age. Panel B shows the total marginal effect of broadband internet on refusal of householder age and household income. Numbers shown in parentheses are standard errors except where otherwise specified.

Appendix

A Relation to differential privacy

Statistical agencies cannot control how much data firms possess, but they can influence the effectiveness of re-identification technologies via privacy-preserving mechanisms applied to their database before dissemination. Mechanisms that satisfy differential privacy have become popular among data providers recently, including statistical agencies (Abowd, 2018; Drechsler, 2023). Differential privacy is a mathematical criterion for database privacy developed in the computer science literature (Dwork, 2006; Dwork et al., 2006). The appeal of differential privacy is that it provides a provable and quantifiable privacy guarantee. The guarantee is a bound on how much a database output can change based on the presence of a single individual in the database, such that it is difficult to distinguish whether a given individual was even in the database.

A common variant of differential privacy is ϵ-differential privacy, which states that a randomized privacy mechanism, M, is ϵ-differentially private if:

$$\frac{\Pr[M(D_1) \in S]}{\Pr[M(D_2) \in S]} \leq e^{\epsilon},$$

where D_1 and D_2 are two databases that differ in only one record and S is the realization of the randomized mechanism. That is, the ratio of the likelihood that realization S was based upon database D_1 rather than database D_2 is bounded by e^{ϵ}. The definition can also be expressed in terms of the ability to infer about a single individual:

$$\frac{\Pr[R = r|M(D_1)]}{\Pr[R = r|M(D_2)]} \leq e^{\epsilon}, \tag{A.1}$$

where R denotes a random variable (e.g., income) and r represents an individual's true value (Kifer et al., 2022). The interpretation is that no matter what prior an attacker uses nor what output mechanism M produces, the ability to infer about an individual in the database

is within a factor e^{ϵ} of the inference that would be made if the individual were not in the database. Therefore, although the definition of differential privacy does not directly refer to re-identification attacks, differentially private algorithms provably resist such attacks as well as other arbitrary risks (Dwork & Roth, 2014). Individuals may still experience harm from an unwanted inference based on the output from a differentially private mechanism applied to a database in which they exist, but differential privacy bounds the increase in the probability of harm caused by their existence in the database.

Now, recall the monopolistic firm described in section 3.2 whose goal is to identify every individual's willingness to pay for its product. Let i be a target individual in the firm's database (A) and let $i = j$ indicate a target individual who has been re-identified in the agency's database (Z). Recall that w_i is individual i's willingness to pay. The firm seeks to calculate the $\Pr(w_i|A, Z)$ for each $i \in n$. The $\Pr(w_i|A, Z)$ can be calculated using Bayes' rule:

$$\Pr(w_i|Z, A) = \left[\frac{\Pr(Z|i = j, w_i, A)}{\Pr(Z|A)}\right] \times \Pr(w_i|A). \tag{A.2}$$

On the right-hand size, $\Pr(w_i|A)$ is the firm's prior probability, before the agency's database is released, that the value of individual i's willingness to pay is w_i, which is equivalent to ρ^{prior} from section 3.2. On the left-hand-side, $\Pr(w_i|A, Z)$ is the firm's posterior probability that the value of individual i's willingness to pay is w_i after the agency's database is released, which is equivalent to ρ^{post} from section 3.2. The term in brackets is the Bayes factor, representing how the probability of observing database Z would change if individual j in the agency's database is identified as individual i in the firm's database and their willingness to pay is w_i. Substituting the terms from section 3.2 into (A.2), the firm's problem can be restated as:

$$\rho^{post} = \left[\frac{\Pr(Z|i = j, w_i, A)}{\Pr(Z|A)}\right] \times \rho^{prior}.$$

Recall also from section 3.2 that the firm's probability of identifying an individual's

willingness to pay when the individual is not in the agency's database is ρ_{-i}^{post}. That is, $\Pr(w_i|Z_{-i}, A) = \rho_{-i}^{post}$, where Z_{-i} represents the agency's database without data from individual i. The ratio of the probability of identifying an individual's willingness to pay when the individual is versus is not in the agency's database is:

$$\frac{\Pr(w_i|Z, A)}{\Pr(w_i|Z_{-i}, A)} = \frac{\rho^{post}}{\rho_{-i}^{post}} = \frac{\left[\frac{\Pr(Z|i=j,w_i,A)}{\Pr(Z|A)}\right] \times \rho^{pre}}{\left[\frac{\Pr(Z_{-i}|w_i,A)}{\Pr(Z-i|A)}\right] \times \rho^{pre}} \tag{A.3}$$

Note that the left-hand side of equation (A.3) is what differential privacy bounds in equation (A.1), except that it is based upon an unobservable characteristic (w_i) that is determined by observable characteristics, rather than based upon the observable characteristics themselves. If we assume that revealing observable characteristics (such as income or race) could also reveal unobservable characteristics (such as willingness to pay), then by combining equations (A.1) and (A.3) and substituting terms we are left with:

$$\frac{\Pr[R = r|M(D_1)]}{\Pr[R = r|M(D_2)]} = \frac{\Pr(w_i|Z, A)}{\Pr(w_i|Z_{-i}, A)} = \frac{\rho^{post}}{\rho_{-i}^{post}} \leq e^{\epsilon}. \tag{A.4}$$

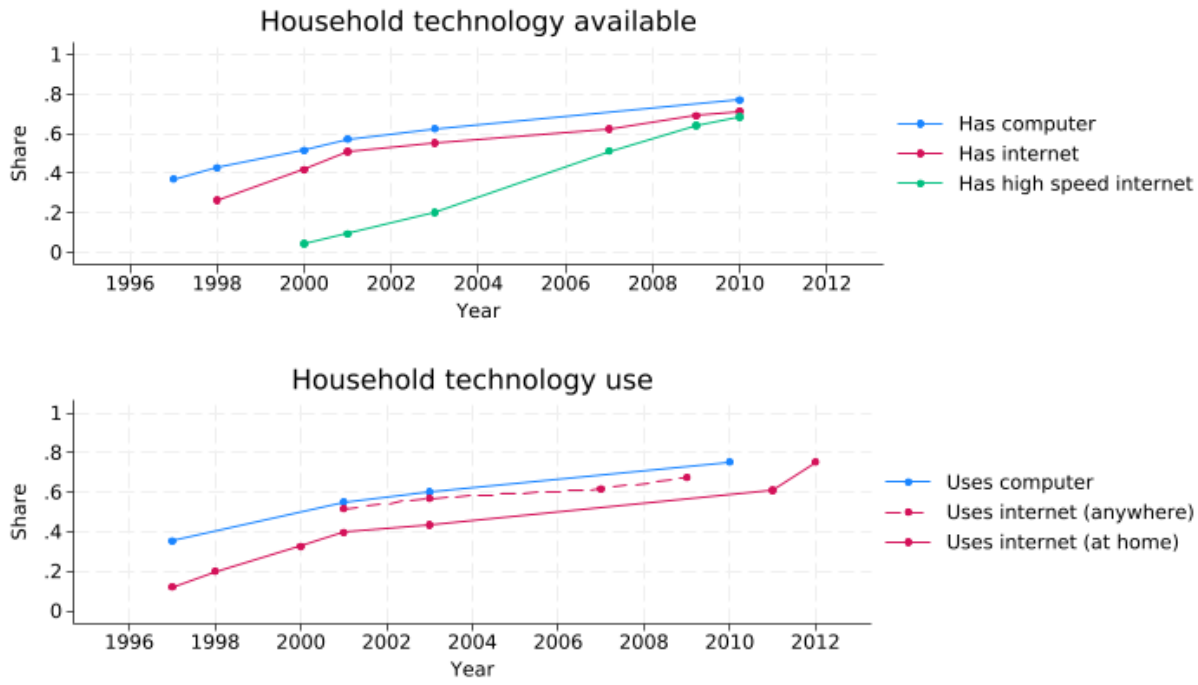
Finally, recall from equation (1) in section 3.3 that individual i will respond to the statistical agency's survey if $B \geq (\rho^{post} - \rho_{-i}^{post})(w_i - p_e)$. Re-writing equation (A.4) as $\rho^{post} \leq e^{\epsilon}\rho_{-i}^{post}$ and substituting into equation (1) from section 3.3, we see that use of an ϵ-differentially private mechanism would provide an upper bound on the right-hand side of the decision equation for a given level of w_i and ρ_{-i}^{post}:

$$(\rho^{post} - \rho_{-i}^{post})(w_i - p_e) \leq \rho_{-i}^{post}(e^{\epsilon} - 1)(w_i - p_e).$$

Thus, differential privacy can bound the likelihood of the firm determining the willingness to pay of an individual in the agency's database, relative to the likelihood the individual would face if they did not exist in the agency's database.

B Supplemental Figures and Tables

Figure B1: Household Computer and Internet Usage in the CPS



The figure shows household-level availability and usage of computers and internet, based on the Computer and Internet Use supplements in the CPS. Results are based on the whole CPS sample (rather than the merged CPS-FCC 477 sample).

Table B1: Robustness of Main Difference-in-Differences Results to Excluding Observations after June 2008

	Unit Refusal		Other Unit Non-Response	
	(1) OLS	(2) Imputation	(3) OLS	(4) Imputation
Broadband	0.00583***	0.00860***	0.00150	0.00083
	(0.00163)	(0.00224)	(0.00202)	(0.00233)
County fixed effects	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes	Yes
Observations	3,304,902	3,304,902	3,441,322	3,441,322

Notes: The results presented in this table are the same as the imputation results in Table 2, except that these results exclude CPS observations after the last date of FCC 477 broadband internet data (June 2008).

Table B2: Robustness of Main Difference-in-Differences Results to Logit Model

	Unit Refusal	Other Unit Non-Response
	(1)	(2)
	Logit	Logit
Broadband	0.0057***	0.0030
	(0.0015)	(0.0022)
County fixed effects	Yes	Yes
Year fixed effects	Yes	Yes
Month fixed effects	Yes	Yes
Observations	4,575,447	4,762,799

Notes: The results presented in this table are the same models from Table 2, except that the regression specifications are estimated via logit models rather than OLS or the imputation estimator from Borusyak et al. (2024). The table reports average marginal effects.

Appendix References

Abowd, J. M. (2018). The U.S. Census Bureau adopts differential privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2867–2867).

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, *91*(6), 3253–3285.

Drechsler, J. (2023). Differential privacy for government agencies—are we there yet? *Journal of the American Statistical Association*, *118*(541), 761–773.

Dwork, C. (2006). Differential privacy. *International Colloquium on Automata, Languages, and Programming* (pp. 1–12). Springer.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3* (pp. 265–284). Springer.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407.

Kifer, D., Abowd, J. M., Ashmead, R., Cumings-Menon, R., Leclerc, P., Machanavajjhala, A., Sexton, W., & Zhuravlev, P. (2022). Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 census. *arXiv preprint arXiv:2209.03310*.