Demand for Privacy from the U.S. Government: Evidence from the Census Income Question*

Zoë Cullen
Harvard Business School

Tom Nicholas
Harvard Business School

July 1st, 2024

Abstract

U.S. residents and representatives continue to demand limits to government data collection, restricting expansion of Census questions. Does resistance stem from demand for privacy as a right, or fear of identifiable data use? We provide evidence that individuals do not disclose their income when it highly differentiates them. Individuals at both extremes of the income distribution withhold in counties with high inequality, when the data collector's wealth is distant from theirs, and when government-collected income data has been made public in the past. These privacy patterns are consistent with fears of identifiable data disclosure, and downwardly bias measures of inequality.

JEL Classification: D83, D91, C81.
Keywords: inequality, income, privacy, government, U.S. Census.

*Cullen: zcullen@hbs.edu, Rock Center 210, Boston, MA 02163, United States. Nicholas: tnicholas@hbs.edu, Rock Center 321, Boston, MA 02163, United States. We are extremely grateful to Dan Marcin for providing access to his newspaper tax list data (ICPSR 36640), to Price Fishback for data on New Deal expenditures and to Tim Guinnane and Jesse Shapiro for detailed comments and suggestions. The Division of Research and Faculty Development at Harvard Business School provided financial support.

1. INTRODUCTION

The U.S. Census Bureau, and the data it collects, have been the subject of political debate throughout U.S. history (Bouk, 2022). Because Census data are a public good, concerns over confidentiality or surveillance can restrict the willingness of individuals to share potentially identifiable information. In 2018, the U.S. Census announced it would use differential privacy, forgoing statistical accuracy in published statistics in exchange for greater identity protection.[1] We study the reasons for resistance to government data collection of income, and discuss the implications for policies under consideration.

Individuals demand privacy for many reasons, including to protect personal data for autonomy, and for dignity. Thus, individuals express different preferences over private and public information domains (Acquisti et al., 2016a). The political debate over Census data collection has focused on the intrinsic value of privacy, and the violation of a right to privacy. In a famous 1890 *Harvard Law Review* article, Samuel Warren and Louis Brandeis asserted the right to be "left alone", was distinct from but complementary to inalienable natural rights to life, liberty and property, and that common law could defend against intrusion just as it protected against injury through slander or libel.

The perceived instrumental value of privacy, preventing ones type from becoming known through data (Stigler, 1980; Posner, 1981), has also influenced the public debate over Census data collection. Debate has centered on the rights of vulnerable populations—by race, sexual orientation or ethnicity—to be protected from discrimination, leading to assurances that the data collected will be aggregated and anonymized, used for research purposes and regionally targeted programs, and representation.

Intrinsic and instrumental motives for withholding ones data from the U.S. government have very different implications both for the patterns of data distortion we would expect to see, and the effective policies for encouraging participation. Instrumental motives for withholding data can be addressed through building trust in the continued protection of the data itself, while political and legal perspectives on the right to privacy might more effectively influence intrinsic motives. Becker (1980) rightly acknowledges that a person can have both types of privacy preferences.[2] Accordingly, our analysis considers these preferences simultaneously.

In light of current debates on Census data practices and the demand for anonymity, our analysis centers on the hotly debated issue of confidentiality in the 1940 Census, which was

[1] In the case that U.S. residents fear the identified use of these data by third parties, differential privacy may be the solution. However, if an intrinsic right to privacy matters, or there is concern the government will change its policy in the future, such policies may not be effective at garnering support for data collection.

[2] Lin (2022) shows the willingness of consumers to share data does depend on both intrinsic and instrumental components with intrinsic motives being quantitatively small.

augmented to include two questions about income—how much a person earned and whether they earned non-wage income. While Internal Revenue Services collected information about income for the purpose of taxation, fewer than 7% of the population owed income tax, and these data were not accessible to other government agencies. In support of the expansion, Harry Hopkins, the Secretary of Commerce at the time emphasized the significant research value associated with the new data as a result of labor market distortions caused by the Great Depression:

> You cannot gauge the employment and unemployment problem in terms of hours and weeks worked alone. It must be supplemented by facts concerning wages and income. Complete analysis of the unemployment situation is impossible without a tabulation of wage income in combination with age, occupation, and industry. This is a fact accepted by all concerned with the unemployment and employment problem. (United States Senate, 1940).

Opposition to the expansion of the 1940 Census to include the income questions focused on intrinsic privacy infringement. Immediately prior to enumeration, the subcommittee of the Department of Commerce discussed an opposing resolution introduced by Republican Senator Charles W. Tobey from New Hampshire, stating "no justification can exist for officials and employees of the United States to lawfully arrogate to themselves the power to make unauthorized inquiries into the private affairs of citizens" (United States Senate, 1940).

Although public discourse regarding the practical importance of income privacy from the U.S. government may have been limited, there was still potential for the data to be utilized for purposes beyond research. Income-related data collected by the federal government had been disclosed to the public in the past (Lenter et al., 2003). Noticeably, in the recent memory of most Americans, the federal government had reversed protections on tax records, allowing some local newspapers to publish individual tax returns between 1924 and 1925, before reinstating protections in 1926.[3] Partly motivated by unauthorised data use, legislators decided the income question could be answered by writing the number on a piece of paper and inserting it into an envelope when the enumerator came around.

We study concerns over the identified use of data by examining how privacy demands respond to the personal stakes associated with a data breach, and its perceived likelihood of ocurring. The personal stakes of a data breach are low when everyone has the same income, and the information does not distinguish individuals. Personal stakes rise with

[3]While not known at the time, soon after enumeration of the 1940 Census, the individual data would be used to locate, freeze and confiscate the assets of Japanese people living in the U.S. during the process of wartime relocation and internment. See further, Arellano-Bover (2022).

inequality (Luttmer, 2005; Perez-Truglia, 2020; Cullen and Perez-Truglia, 2022), due to several factors including self-image, social-image, resentment and competition. We examine how non-disclosure rates correspond to local inequality both at the county and occupational level. We also use quasi-random variation in the housing wealth-distance between the subject and the enumerator to test the causal effect of salient inequality with the data collector, and the demand for privacy. Second, we test whether perceptions about the likelihood of identified data leaks drive non-disclosure. To do this, we examine the impact of exposure to the prior tax return publicity leaks. We use the unanticipated release of tax returns in some locations, and not others, following the 1924 Revenue Act, as a shock to perceived risk that federally collected data may eventually become publicly accessible.

We construct a measure of withholding the truth from the Census enumerator regarding personal income, building on the work of historian Dan Bouk. Misreporting is a federal crime, enforced through fines and imprisonment.[4] We observe an overt form of non-disclosure in response to the new Census question. As the income question appeared at the conclusion of the survey, respondents who reported they were a wage earner with positive weeks worked in a wage-earning occupation earlier in the sequence of questions, would be forced to report their earnings as "non-wage earnings" to avoid specifying an exact amount. Bouk (2022) linked archival complaints to Census records, showing those who sent letters to their representatives protesting the income question also obscured real earnings by responding "yes" to "more than $50 as non-wage income", and entering missing or 0 wage earnings instead.[5]

Our analysis is based on millions of individual responses contained in the 1940 federal census data (Ruggles et al., 2021) along with links to multiple county-level datasets, including New Deal spending, voting patterns, urbanization, religion, and education. We find 6.2% of wage-earners aged between 16 and 80 do not disclose their earnings, equivalent to about 2 million individuals. Our main results use wage-earners between 25 and 65 years, the most active labor market group, 5.1% of whom refuse to disclose (1.2 million individuals).

Rates of withholding income data vary significantly at the level of the county, with an interquartile range of 4 percentage points or 65%. Non-disclosure rates rise monotonically across deciles of the 90/10 ratio. The top decile experienced approximately 38% higher

[4] "Uncle Sam can fine or imprison any one of us who refuses to answer or does so falsely" Willison (1940). In practice, imprisonment is rare, but has been documented: "In 1890, at least a couple dozen people were arrested in New York City for refusing to identify themselves." (Seipp (1981) p44, 49-50; Igo (2018), p46-47).

[5] "As an example, Florence Doud, who lived on rural route 2 in Michigan City, Indiana, wrote that she worried that there was "a limit" to what could or should be asked of citizens by their government. The census questions, she wrote, threatened to "breed a nation of prevaricators," a legion of liars... Doud identified her and her husband Ray's occupations as grocery clerk and barber, respectively, but claimed to have earned no wages from such work in 1939 (but to have earned more than $50 from some other source). In the manuscript data, those answers appear as "52/0/yes" for Ray (weeks worked/income from wages/more than $50 other income) and "8/0/yes" for Florence" Bouk (2022) p188.

non-disclosure rates than the bottom decile. The linear relationship between inequality and non-disclosure strengthens as the peer group narrows, for example looking within country and occupation. County level inequality in incomes is the best predictor of non-disclosure rates relative to political preferences, government spending, or religion.

The correlation between the 90/10 income ratio and non-disclosure rate is especially pronounced among the predicted top and bottom of the earnings distribution. We link a U.S. Treasury list of the highest income tax payers to the census and find that the ultra-wealthy wage earners report zero or missing incomes at a rate least four times higher than similarly aged wage earners, despite top-coding by enumerators at $5,000 for approximately the top 1% of earners.

By contrast, we cannot reject that the correlation between political preferences and non-disclosure is zero. The top decile of Republican-leaning counties experience an economically small and statistically insignificant higher non-disclosure rate relative to the bottom decile. We show robustness of our results to predicted (rather than self-reported) income measures conditional on age, race, gender, education, occupation and housing wealth, to alternative income inequality measures, and to variation in housing wealth inequality.

We use the quasi-random assignment of Census enumerators to measure how non-disclosure responds to distance in housing wealth between the subject and data collector. A subject had the opportunity to answer the income question by writing down a number and sealing it in an envelop before handing it over to the enumerator. Nevertheless, the enumerator likely represented the first person considered with potential to access the data, and as such, a salient audience. We observe a sample of 1940 Census enumerators in Census records[6] and use their home and rental values to estimate the level of capitalized and reported housing wealth inequality between them and each of their subjects. Enumerators typically enumerated in areas nearby their own residence, so enumerators and respondents were likely to have visibility into each others' housing wealth. We find a 10% increase in the housing wealth gap between the subject and the enumerator, based on reported housing values in the Census, is associated with a 3.6% increase in the non-disclosure rate in our specification with education and demographic controls. We cannot reject symmetric effects when the wealth gap favors the enumerator or subject. We find no corresponding housing-wealth effect when we estimate a 'placebo treatment' by randomly assigning another enumerator of the same gender residing in another enumeration district in a different state.

Next, we use the unanticipated release of tax returns in some locations, and not others, following the 1924 Revenue Act, as a shock to the perceived risk that federally collected data may eventually become publicly accessible. The primary source of income data collected by

[6]Enumerators self-identify as such in the occupation string from the census schedules.

the U.S. government to date was collected by the Internal Revenue Service (IRS) to levy taxes. These data included a limited portion of the population, less than 7%, because only high earners were subject to an income tax at the time. Between 1924 and 1926, tax returns, as well as the names and addresses of those filing, were publicly accessible following the surprising and controversial 1924 Revenue Act, inviting the public eye in order to curb tax evasion. Journalists around the country accessed these data and printed lists of names and individuals in local papers. Due to confusion about the legality of publication and variable preparation of tax return lists by local agencies, some papers did, and other papers did not, print these data during the brief window the 1924 Revenue Act stood.

We ask how exposure to the local publication of top-earner tax returns by name affected non-disclosure choices approximately 15 years later in the 1940 Census. Individual beliefs and choices have been show to respond significantly to personal experiences of salient economic events with effects that can persist for decades, especially when messaging is reintroduced within analogous contexts (Malmendier and Wachter, 2024).

We rely on records of the newspapers that did, and did not, print individual tax records and the reasons for these choices. The collection of these records was spearheaded by Marcin (2014), who examined the archives of newspapers publishing in the largest 50 cities. We compiled additional data on newspaper circulation from annual directories of newspaper publications to construct a measure of newsreader exposure. In our baseline analysis, we define an exposed county as one where newspapers published the tax lists in either 1923 or 1924 and assign circulation numbers to that county. We also show robustness of our results using the data in Gentzkow et al. (2014) to construct an alternative measure of exposure that allows for the circulation of newspapers across counties.

In our baseline analysis we find that newspaper circulation in a county where individual tax returns had been published is associated with an increase in the probability of non-disclosure among age groups who would have been directly exposed to the newspaper lists and in the labor market at the time. We find that 40 to 65 year olds in 1940 (i.e. 25 to 50 at the time the lists were circulated) have higher rates of non-disclosure compared to equivalently aged individuals in control counties, by around 25%. The college-educated have higher non-disclosure rates than their college-educated counterparts in control counties, consistent with demographic trends where the educated consume more through newspaper reading and with the medium though which tax list information was disseminated. Meanwhile, younger age groups for whom the messaging would have been less salient exhibit negligible differences in non-disclosure rates between counties that did and did not print individual tax returns.

Finally, we use differences in the non-disclosure of individuals in newspaper tax list versus non-tax-list counties to construct counterfactual distributions, showing how refusal to reveal

information in treated counties distorted the distribution of reported income to the 1940 U.S. Census. We provide suggestive evidence that inequality would be significantly underestimated relative to a counterfactual in which incomes were reported truthfully. For example, among those who were over 25 when exposed to highly circulated lists, the reported 90/10 income ratio is 7.0 whereas our estimated counterfactual ratio, following the methods of DiNardo et al. (1996), is 11.4. Our findings highlight the key role of instrumental privacy value in resisting data collection, and its impact on the validity of statistical inference.

The rest of the paper proceeds as follows. We next describe related literature and situate our contribution. Section 2 provides institutional context around the 1940 Census expansion, the assignment of enumerators, and the release of individual IRS income tax records 15 years earlier. Section 3 describes the details of our data construction and sources. Section 4 reports our evidence on the relationship between non-disclosure and inequality at the county-level. Section 5 examines inequality between the subject and enumerator for a sharp test of how inequality impacts non-disclosure. Section 6 examines a shock to perceptions of how likely the identified data are to be released, using exposure to publicized lists of individual tax return data 15 years earlier. Finally, we conclude.

1.1 *Contribution to literature*

Our work is connected to multiple strands of the interdisciplinary literature on privacy. Prior research has documented distortions in aggregated government survey data. Price (1947) compared the Census counts of Black residents with those of Black draft registrants, finding that the Census underestimated the Black population by about 13%. Serrato and Wingender (2016), and more recently Chi (2022), make use of the fact that Census population measures deviate from ground truth over time as people migrate and age to show that the out-of-date nature affects decisions that hinge on these numbers, such as government funding allocations and firm entry decisions. By contrast, we examine individual-level non-disclosure and the drivers of these distortions in the data.

Our paper contributes to the economics of privacy demands (Acquisti et al., 2016b; Tucker, 2023). Goldfarb and Tucker (2012) showed that, even in anonymous internet surveys, some respondents refuse to reveal information about their incomes and demographics. Athey et al. (2017) and Adjerid et al. (2013) studied the demand for privacy in the cryptocurrency market. They showed that even individuals who report that they highly value privacy are willing to give away sensitive information for small incentives. Budd and Guinnane (1991) found that incentives can cause misreporting. Some individuals in the Irish census of 1911 intentionally exaggerated their ages to meet the criteria for old-age pensions. We contribute to this literature by measuring individual preferences for privacy in a context with material consequences

for the non-discloser, e.g. potential for fines and imprisonment, and consequences for the data collector, e.g. inaccurate data as inputs to policy. In contrast to those other contexts, we find some individuals are willing to take a risk to protect their privacy. The closest study to ours looks at employees' willingness to share information about their salaries with their co-workers in return for rewards, finding the majority would pay to conceal their information but some would pay to share it. Those who choose to conceal it are, on average, those who perceive themselves as relatively high earners (Cullen and Perez-Truglia, 2023).

Our paper is related to a literature examining how transparency of tax returns affects tax evasion and income reporting. Bø et al. (2015) investigated the implications for income reporting resulting from the accessibility of online tax records in Norway in 2001 (not dissimilar to the episode we study of incomes by top earners being released in the U.S. during the 1920s). They found that business owners report higher income with greater transparency, rising by about 3% on average, when these disclosures became searchable online. By contrast, our findings show that tax evasion is not the only channel through which income transparency affects reporting behavior. The Census records we examine were disconnected from the tax collection agency. Moreover, the top earners, who prove highly responsive to the publication of tax returns, were not required by the Census to report their earnings above the top-coded level. Hence, the patterns we document are unlikely to be driven by tax evasion concerns.

Our paper also relates to the interdisciplinary literature on interactions between the surveyor (or messenger) and the respondent. Prior work has shown that people are more likely to listen to and act on information delivered by a surveyor with shared traits in common (Durantini et al., 2006; Dolan et al., 2012). We are able to examine the willingness of the respondent to share truthful information as a function of surveyor characteristics, as well as relative wealth differences between the surveyor and subject. Bouk (2022) suggested that the low share of Black enumerators could be one reason that Blacks were under-counted, particularly in those states with most White enumerators, noting that this channel is hard to test: the under-counting could be "bias of the enumerator or act of resistance." Our setting allows us to measure acts of resistance in isolation.

We also offer a channel that could contribute to the patterns of misperceptions about income inequality. We find that high income earners are less likely to disclose their income truthfully, and they are especially reluctant when they perceive inequality to be high in their environment. In face-to-face interactions with enumerators who are frequently neighbors, subjects are less likely to disclose if there are gaps in socio-economic standing. These disclosure patterns, if replicated within daily interactions with members of the community, could generate the systematic misperception that others' wages are more similar to one's own than they really are (Kreiner et al., 2022; Norton and Ariely, 2011; Hauser and Norton, 2017;

Kuziemko et al., 2015; Jaeger et al., 2021; Cullen and Perez-Truglia, 2022).

We speak to the large literature on the causes of distrust in the U.S. government. In general, citizens tend to distrust government the more it regulates (Aghion et al., 2010), and the late nineteenth and early twentieth century witnessed the rise of the U.S. regulatory state (Glaeser and Shleifer, 2003). Driven by perceptions of corruption, political bias, or fears of undue surveillance, distrust in government negates the willingness of the citizenry to comply with laws or consent to information demands (Levi and Stoker, 2000). In our context, trust in government was arguably at an all-time high, at least in areas of the country experiencing a boost from New Deal spending (Caprettini and Voth, 2022). Equally, our focus on the drivers of non-response illuminates how confidence in government can start to erode.[7]

Finally, we contribute to research on differential privacy and the practice of adding random errors to official statistics (Ruggles, 2024). Bowen (2024) argues that the exact implications of adjusting the privacy-loss budget, representing the trade-off between privacy and data utility, are not well understood. Abowd and Schmutte (2019) note that respecting privacy demands requires custodians of data to concede accuracy and propose using social welfare theory to equate marginal benefits and costs where data is a public good, like in the case of Census data. Our study shows how perceived personal stakes to share information may be large for key variables like income. These perceived stakes may interact with methods to "optimally" distort data. As such, the patterns we find could help execute differential privacy calculations.

2. BACKGROUND AND INSTITUTIONAL SETTING

In this section we outline salient aspects of the historical background, focusing on the new income question in the 1940 census, the publication of lists of top earners by newspapers during the 1920s, and privacy demands and the fear of leaks of census data.

2.1 *The Demand for Privacy*

Privacy determinations must weigh individual demands for privacy with the impact of its use by private or governmental organizations (Acquisti et al., 2016a; Becker, 1980; Lin, 2022). Warren and Brandeis' 1890 HLR article acknowledges the tradeoffs while concluding that privacy was embodied in the U.S. Constitution (it was not until 1965 that the Supreme Court certified that the Constitution implied a right of privacy). Prior to 1850 census returns

[7]This is exemplified by the macabre poem mocking the enumeration process mailed to William Lane Austin, Director of the Census Bureau: "You will wish the under-taker, Undertook the Record Maker" the writer stated, with the poem going on to read "You'll think his quizzes are all rot, You'll surely say, he should be shot" (Bouk, 2022).

were published normally in public places so that any errors could be corrected (Gatewood, 2001). Population growth and urbanization, however, increased what Warren and Brandeis described as the "intensity and complexity of life" and thereby the demand for privacy. In their view, newspapers were especially responsible for privacy infringements by "overstepping in every direction the obvious bounds of propriety and of decency."

2.2 *The Income Question*

Data protection is a core mission of census data collection efforts. On April 1st 1940, the U.S. government used more than 120,000 census enumerators to collect data on over 131 million residents in 143,000 enumeration districts. The 16th Decennial Census was taken against the backdrop of the Great Depression, so governmental agencies were particularly interested in questions around housing, labor markets and internal migration because these responses could inform welfare programs. Central to that aim was data on wage-earning.

Prior data on incomes was limited so the 1940 Census marked a turning point in the U.S. government's demand for data.[8] Enumerators were instructed to interview the most authoritative person in the household at the time of enumeration to gain the most accurate personal data. Just under half of enumerators were women with many being 'housewives' 'salespeople' or 'clerks' as temporary workers employed by the government to administer the Census (Bouk, 2022). Enumerators were organized in a hierarchy under supervisors and clerks. As an indication of the significance of what became colloquially known as "the income question", enumerators asked for this information on *all* individuals in households rather than for random samples. Sampling methods had been introduced by the Census Bureau in 1940, with a 5% sample of the population being asked about participation in social security, for example. Indeed, incomes for wage and salary earners were already being reported to the Social Security Administration (Goldfield, 1958). According to Igo (2018) few Americans voiced concerns about privacy, and were enthusiastic about social security numbers.

Questions 32 and 33 on the census form asked the wage or salary of the person and whether the person had received $50 or more in non-wage income during the prior calendar year. Annual earnings were top-coded at $5,000+ (about the top 1% of the wage distribu-

[8]Consumer expenditure surveys had asked questions about income but not at scale. Population surveys collected some data but mostly on property or other capital assets (Dray et al., 2022). Since 1840 through the Census of Agriculture farmers were asked about income generated from the use or sale of farm products and from 1850 about the value of their farms (Goldfield, 1958). A reliable guide to net incomes in agriculture can be gained from the 1915 Iowa State Census, the first to include a question on occupational wages (Goldin and Katz, 2000). Urban areas, however, experienced far less coverage. International precedents were also reasonably limited. Censuses in Britain asked if a person was a wage-earner but not the amount earned due to privacy concerns and fears that the data might be inaccurate or difficult to collect. By contrast, the 1930 Swedish Census and the 1931 Census of Canada did gather data on the earnings of wage-earners in the population, signalling significant steps towards the disclosure of personal information.

tion), though enumerators sometimes entered actual amounts for those earning above this level. The questions were intentionally asked towards the end of the interview to encourage response (Goldfield, 1958). Moreover, enumerators were reminded of the respondent's right to privacy in the Census Bureau's "Instructions to Enumerators" that they received. Point 18 of the booklet is headed "Refusal to Answer" where enumerators are told: "It is of the utmost importance that your manner should, under all circumstances, be courteous and conciliatory. In no instance should you lose your temper or indulge in disputes or threats. Much can be done by tact and persuasion." Point 19 notes "Should any person object to answering any question on the schedules, explain that the INFORMATION IS STRICTLY CONFIDENTIAL, that it will not be available to any persons except sworn census employees, that it is to be used only for statistical purposes, and that no use will be made of it that can in anyway harm the interests of individuals." Point 20 reassures the enumerator "you have a right not only to an answer, but to a truthful answer" while point 21 strongly conveys to the enumerator their "obligation to secrecy" under the law.

According to the Census Act of 1929, the penalty for non-response to the Census enumerator was a misdemeanor carrying a fine up to $100 and imprisonment of up to 60 days or both, though the Director of the Census acknowledged in 1940 "we do not use that feature of our law because the people of the United States have had confidence in the Bureau of the Census for a great many years, and they cooperate with us in these reports" (United States Senate, 1940). On the other hand, the Bureau did, in practice, impose penalties on its own staff for failure to protect the privacy of the Census data, with the law allowing for a felony charge with up to 2 years imprisonment and a fine of up to $1,000, or both.

Safeguarding personal data had been discussed widely in the press and in government (see Appendix A1). In March 1939 the *New York Herald* proclaimed 'Uncle Sam is Getting Much More Inquisitive' as the new questions had just been announced by the Census Bureau and were being tested in trial counties. Most press coverage, however, came in the months immediately prior to enumeration in response to the Congressional debate sparked by Tobey's resolution. In February 1940 the *Chicago Tribune* noted how 'Census Snooping Stirs Senate Storm' while a month later *The Christian Science Monitor* reported on a resolution to prevent 'unnecessary snooping' adopted by the New Jersey Assembly, controlled by the Republican party. That same month—March 1940—*The New York Times* covered an exchange where President Franklin D. Roosevelt emphasized the "obviously political move" on the part of Senator Tobey to disrupt collection of income data in the 1940 census. For his own part, Roosevelt declared his income as $5,000+ in the census, though provides no definitive answer to question 33 covering non-wage income despite being an active stock market speculator.

In response to the public debate, the Census Bureau conceded to a modification to how

the income data could be collected. If an individual refused to verbally disclose income information to the enumerator, they could instead write down their response on a confidential income form supplied by the enumerator, enclose it in a sealed envelope, which the enumerator would then send to the Census Bureau in Washington D.C. by mail. Goldfield (1958) notes how only 200,000 of the 15 million confidential income forms printed were used, and that subject to the relatively rudimentary statistical analysis of big data at the time, incomes in the 1940 census are considered "reasonably accurate" but "somewhat underreported."

2.3 *Publication of Top Incomes*

Unwanted publicity of tax returns by newspapers as a consequence of a provision in the 1924 Revenue Act foreshadowed the debate over the income question in the 1940 census. This episode has been documented extensively by Marcin (2014) using micro data collected from the newspapers to estimate tax response elasticities. As part of the Revenue Act, which was passed at a time when the Republican party had a majority in the House and the Senate, a private activity—tax payments to the government—became part of the public purview as newspapers across the country published long lists of tens of thousands of individual and corporate tax payers and their tax payments (see Appendix A2).

In September 1925 *The New York Times* published multiple lists of 'Downtown Manhattan's Contributions to New York's Big Share of Federal Tax' showing among others that Edward S. Harkness of 25 Broadway, a philanthropist whose wealth had derived from his father's investment in Standard Oil, paid $1,351,708 to the federal purse. Other lists showed further the tax payments of high-profile elites: for example, J.D. Rockefeller Jr. paid $6,277,669, Henry Ford paid $2,608,808, Andrew Mellon (Secretary of the Treasury at the time) paid $1,882,600 and J.P. Morgan paid $574,379, while Anna Thompson Dodge, the widow of Horace E. Dodge, one of the two Dodge brothers paid $993,028 (Marcin, 2014).[9]

Incomes could be backed out from taxes paid, thereby influencing the perceived instrumental value of privacy at a time when inequality levels were high (Piketty and Saez, 2003). In October 1924, the *Boston Post* noted the large impact these disclosures had on income revelation, with Jack Dempsey then world heavyweight boxing champion earning more than J.P. Morgan; steel magnate Charles M. Schwab less than expected; and the actress Gloria Swanson evidently earning around $120,000 a year (around $2 million today). According to the *Post* "there was no greater surprise in the whole list to Boston people than the income reported by Judge Louis B. Brandeis of the United States Supreme Court", who had warned about newspapers and privacy concerns a few decades earlier. "No one looked upon Justice

[9]The Dodge brothers' automotive parts and later automobile manufacturing company would become a division of the Chrysler Corporation after their death in 1920.

Brandeis as a rich man" the *Post* commented, "but he must be, since the tax of $9,508.22 shows he must have an income of around $55,000 a year" (almost $1 million today).

The release of these data represented the disclosure of personal information previously stored securely by the U.S. government—the essence of ones type becoming known through data. Although the publication provision in the 1924 Revenue Act was not included in the 1926 Act (henceforth only government committees could see the data) there was additional impetus during the Great Depression for public release of tax data to ensure high earners were complying with the tax rules (Lenter et al., 2003). Publication of the newspaper lists often depended on quasi-random reasons, for example local tax collection offices did not uniformly interpret the law. Some local tax collection offices prepared lists for the newspapers, others did not, and still others did not recognize the legality of publication. We exploit this variation to examine differentials in the disclosure of 1940 Census income data.

2.4 *Fear of Census Data Misuse*

Unanticipated leaks of federal data or its misuse could affect the instrumental value of privacy from the U.S. government through perceptions of harm. Recognizing the need for public trust in data confidentiality, the Reapportionment Act of 1929, encompassing both census and apportionment provisions, states: "No publication shall be made by the Census Office whereby the data furnished by any particular establishment or individual can be identified." In recent years, the implementation of differential privacy by the Census Bureau has been a response to vulnerability as reconstruction algorithms can identify individuals *indirectly* using quite disparate components of databases (Dinur and Nissim, 2003).

During the Second World War the Census Bureau *directly* provided aggregated and individual-level data to the U.S. military resulting in the incarceration of Japanese and Japanese-Americans in wartime camps as a consequence of President Roosevelt's Executive Order No. 9066 signed in February, 1942 (Seltzer and Anderson, 2001). Once notified of their relocation to camps many Japanese Americans were forced to sell belongings or property at heavily discounted prices, experiencing immediate economic costs from asset seizures and additional marginalization.[10] While the Second Wars Power Act of 1942 nullified the safeguards from the Reapportionment Act of 1929, this episode highlighted the tension between

[10]Although these disclosures of U.S. Census data were not known at the time (Pearl Harbor occurred in December 1941 after the Census returns had been completed, and the first Japanese internment camps were not established in the U.S. until mid-1942), there was some degree of preemption. Geopolitical tensions were already ongoing due to U.S. constrains on commodity flows and Japanese imperial expansion. Roxworthy (2008) documents how Japanese Americans "knew about FBI plans well in advance of Pearl Harbor", speaking of the raids of Japanese immigrant homes by FBI agents, beginning in Hawaii before spreading to ethnic enclaves throughout the U.S., in an effort to address the "Japanese problem." For studies of the labor market impact on those subjected to internment, see further Chin (2005); Saavedra (2021); Arellano-Bover (2022).

statistical agencies and the utilization of the data they generate. After it was revealed in 2004 that the Census Bureau had legally shared zip code level information about the location of Arab-Americans with the Department of Homeland Security, the Census Bureau stated that it would refrain from such actions in the future (El-Badry and Swanson, 2007).

Due to uncertainties surrounding both real and perceived breaches of anonymity, privacy has instrumental value as a protective mechanism against discrimination. Debate about invasions of privacy and the handling of government data occurred prominently after the Second World War. In 1950, Senator Joseph McCarthy (Republican, representing Wisconsin) produced his list of alleged communists employed by the U.S. government creating impetus for clearer definitions of privacy boundaries. During the McCarthy era the "Lavender Scare" resulted in the use of sexual questioning to purge "lavender lads" from careers in the State Department under the guise of a link between homosexuality and communism (Johnson, 2023). While these instances speak to the issue of public trust in the data, particularly by vulnerable or marginalized communities, we focus on a different form of privacy concerns and fears of census data misuse: those surrounding the revelation of income data. McCarthy himself is included in the 1940 census. Despite being employed as a Circuit Judge in Wisconsin and working 52 weeks of the year in 1939, his income is recorded as 0, with a 'yes' response to the question of whether he received non-wage income.

3. DATA CONSTRUCTION

Our main data source is the 1940 complete-count census data which provides information on census responses for over 131 million individuals (Ruggles et al., 2021). We incorporate various complementary datasets at the county-level. We use data from Marcin (2014) to identify the locations where newspapers did and did not publish lists of top tax payers during the 1920s, and the reason why. And we use newspaper circulation data across locations from Gentzkow et al. (2014).

3.1 *Individual-level Data from the 1940 Census*

We use data from the 1940 census restricting our analysis to individuals who were in the labor force, who self-reported being at work, and who received wages or a salary, including those who worked in government. By construction, our sample restriction will slant towards non-farm workers, since the majority of the farm population were self-employed. Our dataset includes 32.5 million individuals between 16 and 80 years of age, with our main analysis being conducted on 24.9 million 25 to 65 year olds to further restrict to a working age range.

Although these individuals were asked about incomes for the calendar year of 1939, but

their labor market status was recorded for the week of March 24-30 of 1940, the Census Bureau maintained that shifts in the nature of labor market participation were not large enough to distort estimates of income for most occupations, with movement in-and-out of public works employment under the New Deal being a notable exception (Bureau of the Census, 1943). We test for sensitivity of our results to this assumption by estimating privacy demands for full time workers only (those working 52 weeks of the year and 40 hours) and we also show non-disclosure is not being driven by high churn occupations where we might expect to see a large disconnect between employment status and earnings over time.

The census provides a vast array of additional data so we can make a granular assessment of why non-disclosure occurred. We know the share (14% in our dataset) of individuals reporting non-wage income defined as $50 or more from other sources, a level the Census Bureau set to "to identify those persons whose incomes were, for all practical purposes, limited to receipts from wages or salaries" (Bureau of the Census, 1943). Recall, Bouk (2022) notes how this category was used by individuals to "hide" their wage income as non-wage income, so it would be mechanically the inverse of wage income for those who do not disclose. The census includes data on years of education, gender, marital status, weeks and hours worked, and occupation. We know the person who responded to the enumerator on behalf of the household and we know the estimated value of the house they lived in, or otherwise the monthly rent paid, which we capitalize into a continuous series to locate each individual in the distribution of housing wealth.

3.2 *County-level Data*

We use New Deal expenditures at the county-level during the Great Depression from Fishback et al. (2003) as a proxy for the salience of federal redistribution.[11] Caprettini and Voth (2022) establish a causal link between the intensity of government-funded programs during the 1930s and patriotic sentiment during the Second World War, through individual purchases of government securities and direct participation in the war effort. This same mechanism could lead Census respondents to be more willing to share personal information with the federal government. On the other hand, since the government used its ability to finance local welfare programs as a motivation to collect Census income data it is also possible high earners would have withheld their incomes in response to the salience of local redistribution policies.

Opposition to the income question in the 1940 census ran along political party lines. Indeed, political views today are frequently hypothesized to be a determinant of privacy

[11] Our measure is the Fishback-Kantor-Wallis aggregation of New Deal spending categories. The dataset combines counties in New York, Missouri, and Virginia, which we separate by allocating spending based on 1930 population counts.

demands. We use data on voting from Clubb et al. (2006) where we take the Republican vote share by county in the 1938 election, which resulted in President Roosevelt's second term. Political beliefs can also be geographically heterogenous, most notably rural and urban voters exhibit differences in moral values and definitions of right and wrong (Enke, 2020). For this reason, we incorporate data collated by Haines (2010) measuring the share of population living in urban areas. We also include the share of population 25+ completing high school.

Empirically, the relationship between religion and trust is strong according to McCleary and Barro (2006). We use data from the 1936 Census of Religious Bodies to measure religiosity. The debate over whether questions about religion should be integrated into the formal decennial Census revolved around privacy, the desire to protect religious liberty and uphold the separation of church and state. The Census Bureau conducted the 1936 survey by sending questionnaires to religious leaders and the data are reflective of these responses. While less formal groups like the Southern Baptists will be underreported, Stark (1992) notes that the Census Bureau implemented measures to promote overall accuracy, including enlisting the aid of local personnel such as Deputy U.S. Marshals in the data collection process. He finds that the data align with qualitative evidence on church membership and with independently collected enumeration figures.

Finally, we also control for the size and affluence of counties using county-level population counts and value added in manufacturing, again from the data collated by Haines (2010).

3.3 *Inequality Measures and Enumerators*

In 1940, income inequality remained pronounced, as the top decile of income earners accounted for more than 45% of national income (Piketty and Saez, 2003). However, county-by-county, income inequality varied substantially. Empirical studies by Luttmer (2005); Perez-Truglia (2020); Kreiner et al. (2022); Hauser and Norton (2017); Cullen and Perez-Truglia (2023) underscore how salient inequality could affect day-to-day lives through social comparisons and corresponding emotions about self-worth, as well as resentment and competition. We start from the premise that the personal stakes of an identifiable income data leak rises with inequality, as income information increasingly differentiates people. To examine the relationship between disclosure rates and local inequality, we calculate the 90/10 income ratio for each county based on the income responses we do observe in the individual-level census data. Given that disclosure rates were high and top-coding accounts for about the top 1% of income earners, this measure should capture the gap between richer and poorer individuals even with income censoring. As a further step we estimate predicted income using an individuals location, occupation, housing value, age, gender, years of schooling and calculate the 90/10 income ratio using that series.

We also use Census income data to calculate inequality as the mean logarithmic deviation (MLD) of income at the county-level following the argument in Cowell and Flachaire (2023) that the MLD has favorable statistical properties. The MLD measure adheres to the principle of monotonicity in distance whereby an increase in the income of a wealthier individual results in an increase in the overall inequality measure. By contrast, in the case of other measures e.g. the Gini index, when incomes above the point of perfect equality change, both the numerator and denominator move in the same direction, lessening the impact of inequality changes when the rich get richer. We further calculate this measure using housing values to estimate inequality, bypassing the censored income data altogether.

At the individual-level, we also have a measure of the salience of local inequality. We know both who the respondent to the enumerator was in the household and we can identify the sample of enumerators who would have visited the house, consisting of neighboring individuals who reported they were a 'census taker' or a 'census enumerator' in the Census occupation string. This allows us to investigate responses to the income questions as a function of socioeconomic gaps between respondents and enumerators. Appendix A4 provides an example of a population schedule showing the information available in the IPUMS data.

3.4 *Historical Publication of Federal Data*

To examine the impact of any perceived risk of data disclosure on the preferences of individuals to maintain income privacy, we measure exposure to past publication of federally collected data. Marcin (2014) hand-searched the Library of Congress newspaper archive to locate every newspaper publishing lists of top tax payers as a consequence of disclosure-rules permitted under the 1924 Revenue Act in the 50 largest cities by their 1920 population. He identified 79 newspapers publishing these lists altogether.

We map the Marcin data to circulation at the county level using information on individual newspaper circulations from various editions of *N.W. Ayer & Son's American Newspaper Annual and Directory*, a reference publication for the newspaper industry. *Ayer* notes "Left over, unsold, returned and file copies, having never reached the hands of a possible reader, cannot be considered as part of the circulation" so we are measuring circulation that in all likelihood reached readers. We link newspapers to counties based on the location of these cities and then assign total circulation of each newspaper to their linked county. Our circulation figures encompass both the average daily circulations on weekdays and weekends.

Because newspapers publishing lists could circulate across locations we also use disaggregated circulation data from Gentzkow et al. (2014) to measure county-specific exposure. These data tell us weekday circulation of each newspaper by town for the year 1924 as compiled by the *Audit Bureau of Circulations*, an independent auditing and reporting orga-

nization to whom newspapers and publishers voluntarily submitted their circulation data. We traced 67 of the 79 newspapers from Marcin (2014) in the Gentzkow et al. data. Using both sources we test whether those individuals exposed to released, identifiable government records, demanded greater privacy from the Census.

Newspapers were able to access local tax payer information from Internal Revenue Collector offices, as the Collector was required by the new law to prepare these lists for public scrutiny. Section 257(b) of the Act stated:

> The Commissioner shall as soon as practicable in each year cause to be prepared and made available to public inspection in such manner as he may determine, in the office of the collector in each internal-revenue district and in such other places as he may determine, lists containing the name and the post-office address of each person making an income-tax return in such district, together with the amount of the income tax paid by such person.

We exploit quasi-randomness in the publication of these lists to generate variation in exposure. As Marcin notes, "[i]nterpretation and compliance with these provisions varied by local Bureau of Internal Revenue collection offices." In some cases heads of local offices instructed staff to prepare lists expeditiously, while in other cases only on request, and still in other cases not at all. Sometimes a member of the public could copy a list in its entirety, other times only partially. These differences resulted from widespread uncertainty and confusion in how the new law should be interpreted and whether or not it was legal to actually publish them. Indeed, in April 1925 the Justice Department challenged this interpretation, but a month later the Supreme Court acknowledged the legality of publicity along with the principle that privacy rules should be decided by Congress.

We identify 36 counties with a top 50 city where newspapers published these lists, with varying degrees of circulation (which we exploit), and 14 counties where lists were not published, in all covering 7.8 million wage earners between ages 25 and 65. We also exploit variation within and across counties by comparing those who were of age to be likely in the labor market and reading the newspaper at the time the lists were published, and those who were of a younger age. In Section 6, we examine the balance of attributes across counties by circulation of the lists.

3.5 *Individual-level Descriptive Statistics*

We present descriptive statistics in Appendix Table A1 for individuals reporting zero, missing and positive incomes. We consider the full U.S. population responding to the 1940 U.S. Census. Among 16 to 80 year olds, 6.19% of employees who reported earning a salary or

wage to the Census enumerator did not disclose their actual income, equivalent to about 2 million individuals. For a 5% sample of the population of wage earners the Census Bureau at the time estimated the non-disclosure rate was 4.9%, (5.4% among men and 3.5% among women) (Bureau of the Census, 1943), which is close to what we find (5.1%) for the full population of 25 to 65 year olds. Among individuals who worked 52 weeks of the year and 40 hours—the most active labor market group—3.7% did not disclose their incomes.

We find that non-disclosure is highest among those that we predict would have incomes in the top and bottom income deciles. In Figure 1, we illustrate the relationship between predicted income rank, based on occupation, housing value, age, location, gender, years of schooling, and non-disclosure. The stark U-shaped pattern reveals that non-disclosure rates among those in the top and bottom decile is nearly twice as high as the rest of the population. We observe the same U-shaped pattern when we use capitalized house values (Panel b) actual house values (Panel c) or rental values (Panel d). Moreover, the degree of inequality in one's environment, significantly magnifies the rates of non-disclosure.

3.6 *Non-Disclosure by the Super Rich*

As implied by Figure 1, we find that the super rich did exhibit a strong tendency towards non-disclosure at a rate much higher than wage earners in general. While reviewing tax policy during the early 1940s, President Roosevelt tasked the U.S. Treasury with providing a list of the largest tax payers as detailed in Brandes (1983), which we link to the 1940 Census. The list comprises 104 individuals in 98 households in areas like banking and finance, the oil industry, manufacturing and retailing. John D. Rockefeller Jr., Henry Ford and several member of the Du Pont family are included. The list (see Appendix Table A11) includes net income for each individual, averaging $1.14 million (around $30 million today).

Out of the 88 individuals we traced, 57.5% reported zero wage earnings or left the response blank when queried by the enumerator, even though the enumerator would have top-coded at $5,000+. A portion of the wealthy elite lived as rentiers, but among those who claimed to have worked in 1939, 26.5% failed to disclose their earnings, or 20.5% among those who were aged 25 to 65, compared to 5.1% observed for the population of 25 to 65 year old wage earners noted above. The ultra wealthy may gain instrumental value from privacy in contexts marked by disparity if their wealth stems from rent-seeking practices, such as imposing excessive markups, or if their income surpasses what their conspicuous consumption implies. While concealing their income to the enumerator, 93.1% responded 'yes' to the more stringent question of whether they received $50 or more in non-wage income.

3.7 *County-level Descriptive Statistics*

County-level non-disclosure rates vary: non-disclosure is 5.8% at the 25th percentile and 9.6% at the 75th percentile.[12] In Figure 2, Panel a, we display binned scatter plots that illustrate the sharp increase in non-disclosure rates as measures of inequality rise at the county level. We show this relationship is robust to a variety of inequality measures. Panel (a) shows a positive relationship between non-disclosure rates in a county, and the 90/10 *reported* income ratio. As the ratio rises from 5 to 6, the average share non-disclosing rises from 4% to 5%, and from 5% to 6% as the ratio rises to 10. Panel (b) shows a similar relationship between non-disclosure rates and the *predicted* 90/10 ratio. Panel (c) uses the mean log deviation of self-reported income, and Panel (d) displays the mean log deviation in housing wealth as an alternative measure of county-level inequality, both corroborating the strong positive relationship between non-disclosure rates and inequality in the county of the respondent. We further examine the relationship between non-disclosure by 228 narrowly defined occupations in the 1940 Census occupational classification system. Panel (e) illustrates non-disclosure rates rise even more rapidly as inequality rises in the peer group. Non-disclosure rates are around 2.5% percent when the 90/10 ratio is 3, and rise to over 10% when the ratio is 10. Meanwhile, other characteristics of the environment conjectured to impact demand for income privacy, such as political stance, religiosity, and local redistribution, have economically small and statistically insignificant relationships with the rate of non-disclosure. Panels (f-g) imply the effects of New Deal spending at the county-level, and the Republican vote share, do not co-vary with non-disclosure rates, despite the political debate at the time emphasizing the value of intrinsic privacy. Similarly, Panel (h) displays little relationship between non-disclosure rates and the share religious in a given county, despite evidence in other contexts that religion promotes positive attitudes toward honesty (McCleary and Barro, 2006).

4. NON-DISCLOSURE AND COUNTY-LEVEL INEQUALITY

In Table I, we show that these correlations are robust to including an extensive range of covariates at the individual level, county level and occupation level, as summarised in Table A1. Using millions of observations on wage earners in the Census, we estimate the following linear probability specification at the individual level i, where *Privacy* is a binary variable taking the value of 1 if an individual did not disclose their income in the Census, resulting in the enumerator recording 0 or missing to the income question:

[12]The top 5 states by mean non-disclosure are South Dakota (11.8%), Oklahoma (11.0%), Mississippi (10.9%), Tennessee (10.3%) and Missouri (9.8%). The bottom 5 states (where individuals were more responsive to the income question) are Virginia (5.7%), California (5.6%), New Hampshire (5.5%), Maine (5.2%) and Massachusetts (5.2%).

$$Privacy_i = \alpha\, Inequality_c + \beta\, \mathbf{X}_c + \delta\, \mathbf{Z}_i + \kappa_{state} + \phi_{occupation} + \epsilon_i, \tag{1}$$

We are particularly interested in α, the coefficient on $Inequality$, which we measure as the 90/10 ratio at the county-level in Panel (a), as predicted income in Panel (b), the MLD of income inequality in Panel (c), the MLD of housing wealth inequality in Appendix Table A3 and the within-occupation 90/10 income ratio in Appendix Table A4. If the value of privacy is higher in unequal environments we would expect $\alpha > 0$. $\mathbf{X}_c$ is a vector of standardized county-level variables: New Deal spending per capita, the share of individuals who voted Republican, were religious, were educated, or resided in urban areas, and we also control for county population and manufacturing value added. $\mathbf{Z}_i$ is a vector of individual characteristics from the Census. We use fixed effects for state, and occupation at three different levels—11 main categories, 22 sub-categories and 228 granular categories. Hence, our specifications control for both local area characteristics and individual characteristics to address concerns about omitted variable biases.

In Table I column 1 of Panel (a), New Deal spending per capita is a weak predictor of non-disclosure and this continues to be the case when we add county-level controls (column 2), individual demographic controls (column 3), occupation fixed effects (columns 4 to 8) and when we restrict estimation to full time workers (column 7) or to the individual who responded to the enumerator (column 8). We might expect New Deal spending to drive a lower demand for privacy—if the documented patriotism induced by the spending[13] also encouraged individuals to comply with the census requests of their government—or a higher demand for privacy by top earners—if New Deal spending was seen as a pathway to redistribution policies. Not only are most of the point estimates on the New Deal coefficients statistically insignificant, but they are small in economic magnitude. A one standard deviation increase in New Deal spending is associated with just a 0.00011 decrease in the probability of non-disclosure in column 1 or a (-0.00011/0.047)=0.23% decrease relative to the mean non-disclosure rate of 4.7%.

The Republican vote share is also a reasonably weak predictor of non-disclosure, being statistically significant in only three of the eight specifications in Panel (a). This result is important, and perhaps surprising, given how much participation in the 1940 Census was considered a political issue at the time, as well as the idea today that political preferences and partisanship motivate intrinsic privacy demands. A one standard deviation increase in the Republican vote share in a county is associated with a 3.2% increase in non-disclosure relative to the mean in column 1 with the largest effect across these specifications being a

[13] See Caprettini and Voth (2022) for an established link between New Deal spending and patriotism.

3.6% increase estimated in column 6 for full time workers.

Our results suggest a negative relationship between the religiosity of a county and non-disclosure, consistent with norms of trust increasing the response rate, but the effect sizes are not particularly large. The estimate in column 1 implies a one standard deviation increase in the share of people in a county who self-identify as being religious in 1936 is associated with a 5.3% decline in non-disclosure relative to the mean. However, the economic magnitudes are much smaller in columns 2 to 8. For full time workers in column 7, for example, a one standard deviation increase in the share religious is associated with a 2.8% decrease in the probability of non-disclosure relative to the mean non-disclosure rate of 3.5%.

We find a clear relationship between inequality and non-disclosure. A one standard deviation increase in the 90/10 income ratio is associated with between a 5.8% to 15.1% increase in non-disclosure relative to the means across columns 1 to 8. For full-time workers in column 7, the effect size is 10.5%. Figure 3 illustrates particularly large effects in a specification using dummy variables for each 90/10 decile and for each decile of the Republican vote share instead of standardized continuous measures. Privacy preferences rise linearly across deciles of the 90/10 county-level income ratio, with the top decile experiencing 37.7% higher non-disclosure rates (relative to the mean) than the bottom decile. Contrastingly, coefficients on the Republican vote share deciles hover around zero. These results are consistent with the notion that perceived personal stakes are highest when incomes are highly unequal.

While we consolidate the coefficients on the individual controls in Table I, we report these in full in Appendix Table A2. We find strong individual heterogeneity in privacy preferences and evidence of relative benchmarking. Women demand privacy more than men, though the reverse is true among full time workers (column 7).[14] Household heads (mostly men) are more likely to disclose whereas the divorced and separated less so relative to their counterparts who were married, with singles generally preferring to disclose, though less so as full time workers (column 7). Whites—about 90% of the population in 1940—exhibit a preference for privacy relative to minorities, while immigrants are more likely to disclose. The college educated exhibit a strong preference for non-disclosure, which may reflect concerns over relative income revelation or the desirability among higher earners of privacy as a luxury good. Consistent with the argument that the wealthy gain utility from instrumental privacy value, we find a positive relationship between capitalized housing values and privacy, through the effect sizes are reasonably small. In columns 3, for example, where we estimate the largest coefficient across specifications, a one standard deviation increase in housing wealth is associated with a 1.6% increase in non-disclosure relative to the mean.

[14]Following Moehling (2001), relative demand for privacy may been greater among a sub-sample of women with unemployed partners, since this may have allowed more control over how earnings were allocated towards household priorities.

Two additional sets of regressions in Table I show the robustness of the relationship between local inequality and non-disclosure. Panel (b) re-estimates all the linear probability models from Panel (a) using predicted income to measure the 90/10 ratio to avoid any mechanical bias associated with censoring in the reported income series. The effect sizes are slightly smaller, which would also be consistent with measurement error attenuating the coefficients. The effect sizes are larger in Panel (c) where we use the MLD approach to estimating income inequality. In column 7 a one standard deviation increase in local income inequality is associated with a 13.4% increase in non-disclosure relative to the mean, compared to a 10.5% increase in Panel (a) or an 8.5% increase in Panel (b).

In Appendix Table A3 we address the concern that any measure of income inequality could be distorted by the non-disclosure rate, especially in high and low income ranges, using a measure of inequality that is independent of self-reported incomes: the MLD of housing values in a county. Here, we find non-disclosure effect sizes for a standard deviation change in housing wealth inequality for full time workers (column 7) of 7.1%.

Our results are strongest when we use within occupation reported incomes to measure the 90/10 ratio, suggesting the more we narrow in on a reference group of peers, the tighter the link between the choice to disclose and the perception of inequality. In Appendix Table A4 column 7 we estimate a one standard deviation increase in the 90/10 ratio is associated with a 32.3% increase in non-disclosure for full time workers, an estimate that is stable across choices of controls, including when we control additionally for the median wage by occupation as a proxy for employment churn (see Appendix Table A5).[15]

5. NON-DISCLOSURE AND SUBJECT-ENUMERATOR INEQUALITY

As an individual-level test of how privacy demands respond to the salience of local inequality, we use the quasi-random assignment of Census Enumerators to households, and the resulting variation in the wealth gap between the respondent and the enumerator, to observe if a higher wealth gap increases the probability of non-disclosure. Enumerators were often drawn from local communities, so gaps in wealth and social standing might be noticeable to the respondent during the interview at the respondents' home. We use data on housing wealth and other socioeconomic characteristics to test whether a visit from a Census enumerator with differential wealth to the respondent increases the probability of non-disclosure.

Although their names are visible on the actual census returns at the top of the population schedules, to our knowledge no dataset of enumerators exists. We therefore identified a sample of enumerators using the occupation string in the IPUMS data. We require the

[15]If turnover is higher in lower median wage occupations individuals may be less likely to recall their earnings and therefore report missing or zero.

words 'census' and 'enumerator' or 'census' and 'taker' to be included simultaneously in this string, producing a dataset of 1,023 enumerators from 48 states plus the District of Columbia. Enumeration included individuals who went door-to-door as well as area managers and district supervisors who coordinated these activities, trained enumerators, and consolidated census returns. We cannot distinguish between these roles, but our use of specific occupation keywords means we are mostly likely to capture door-to-door enumerators.

Enumerators were not randomly selected from the population, which creates a causal challenge for estimation. Enumerators were required to be a U.S. citizen with at least a high school education; they had to have legible hand writing; and they needed to pass a formal aptitude test mimicking the completion of a schedule return (Thomson, 1940). Enumerators were often allocated to their residential district, an advantage for our purposes, in that familiarity with subjects would allow for knowledge of relative standing.

Since a wealthy individual and poor individual are both statistically more likely to have a larger inequality gap with the 'typical' enumerator than a middle income individual, we take an additional step to isolate variation stemming from the particular enumerator assignment. For comparison, we draw a 'placebo enumerator' designated as a random enumerator of the same gender from an enumeration district in a different state. We compare the effects of wealth distance between the subject and realized enumerator on non-disclosure, with the effect of differences between the subject and placebo enumerator, yielding our preferred statistical test of how salient face-to-face inequality with an enumerator impacts non-disclosure.

Specifically, we estimate the following linear probability specification at the individual level i where $Privacy$ is a binary variable coded 1 for non-disclosure, as before, and the superscripts S, E and E^* refer to subject, enumerator and placebo enumerator respectively:

$$Privacy_i = \underbrace{\gamma_1 \, log\left(\frac{House^S}{House^E}\right)_i}_{\text{Actual Enumerator}} + \underbrace{\gamma_2 \, log\left(\frac{House^S}{House^{E^*}}\right)_i}_{\text{Placebo Enumerator}} + \omega \, \mathbf{I}_i + \pi \, \mathbf{G}_i + \nu \, \mathbf{D}_d + \zeta_{county} + \epsilon_i. \quad (2)$$

Our main coefficient of interest is γ_1 capturing the wealth gap between subject and enumerator while controlling for the effect of the placebo treatment through γ_2. Our causal test of the impact of face-to-face inequality on non-disclosure is $\gamma_1 - \gamma_2 = 0$ corresponding to the null hypothesis that the difference between these coefficients is zero.

Wealth is measured as the log of the ratio of the subject to the enumerator's house value, which we also calculate for the placebo enumerator: $\left(\frac{House^S}{House^{E^*}}\right)_i$. We estimate equation 2 using both capitalized housing values and reported housing values. The vector $\mathbf{I}_i$ contains the log of the subject's own house value (to avoid conflating own wealth in the decision

to keep income private with the effect of the subject-enumerator wealth gap), as well as demographic variables including race, marital status, immigration, education, age and its square. Since homophily may also influence the willingness of subject-enumerator pairs to share information the vector $\mathbf{G}_i$ consists of age and years of education gaps between the subject and enumerator and between the subject and the placebo enumerator. We also show our results with and without restricting estimation to subject-enumerator pairs with matching genders since gender differences may also drive response rates. The vector $\mathbf{D}_d$ includes the mean of the log of housing wealth, educational attainment and age by enumeration district d to capture unobserved local characteristics that might impact privacy. We use county fixed effects and cluster the standard errors at the household level.

In column 1 of Table II we find that a 10% increase in the capitalized housing wealth gap between the subject and the enumerator is associated with an increase of $\gamma_1 \times \ln(1.10)$=0.00022 in the probability of non-disclosure or about 0.53% relative to the mean, or 0.53% in column 2 when we control for the placebo enumerator—the effect we would expect to see by chance. Under our test of the importance of face-to-face inequality, $\gamma_1 - \gamma_2 > 0$ (p-value=0.0005). With classical measurement error in the capitalized housing wealth series we would expect these coefficients to reflect lower bound estimates. Indeed, in columns 3 and 4 we find much larger effects when we use reported house values. In column 4 a 10% increase in the housing wealth gap is associated with a 2.5% increase in non-disclosure relative to the mean non-disclosure rate of 4.3%. The difference between actual and placebo coefficients again exhibits statistical significance ($\gamma_1 - \gamma_2 > 0$, p-value=0.0266).

In columns 5 to 8 we replicate the results in columns 1 to 4 using only subject-enumerator pairs of the same gender. We find consistent results and effect sizes that are slightly larger. In column 8, for example, a 10% increase in the housing wealth gap is associated with a 3.6% increase in non-disclosure relative to the mean, compared to the corresponding effect of 2.5% in column 4 when we do not gender match on subject-enumerator pairs. One explanation for the increased tendency to withhold information when we condition on gender-matching could be attributed to the importance of the reference group (Cullen and Perez-Truglia, 2022). As individuals become more closely aligned with their peers, social concerns around earnings information become more pronounced. This would be consistent with the larger magnitudes that we find in our occupation-level results, as discussed in Section 4 above.

In Appendix Table A7 we replicate columns 1, 3, 5 and 7 of Table II, splitting our sample by whether the subject housing wealth is greater than or less than the enumerator's housing wealth. We cannot reject that the sensitivity of non-disclosure to the housing wealth gap is symmetric regardless of who is wealthier, subject or enumerator.

6. NON-DISCLOSURE AND EXPOSURE TO HISTORICAL INDIVIDUAL DATA RELEASE

We now explore the effect of shifting the perception that individual income data will be leaked by the government. We look at the impact of newspaper lists of individual income tax returns published in a brief window around the 1924 Revenue Act. We test whether individuals who were most likely to observe the publication of once-private IRS tax return data during that window had a different propensity to disclose their income when the 1940 census was conducted.

We employ two complementary research designs. We carry out a triple differences estimation procedure by comparing those with high and low direct exposure to the news lists as a result of their age at the time the list is published, across counties that did and did not publish income tax returns. We predict those individuals who were alive and in the labor market (25+ years old) were more likely to have personally looked at the lists and will thus exhibit larger treatment effects (differential non-disclosure rates compared to younger individuals) in counties with published lists, compared to counties without published lists. Secondly, we introduce more granular measures of the intensity of exposure using the number of newspapers publishing tax returns in a given county and the number of copies per resident. We predict those over 25 years at the time, in counties with the top third most intensely circulated news lists, will exhibit the largest treatment effects of all relative to counterparts in counties with middle and lower third circulation intensity, or younger individuals at the time of circulation.

6.1 *Identifying Counties with Published Individual Tax Returns*

We identify counties where the lists were published or not published among states that include a top 50 city by population size due to our data source on newspaper circulation. Hence, we compare non-disclosure rates in counties like Hartford County, Connecticut where both the *Hartford Courant* and the *Hartford Times* published lists with control counties like Fairfield County, Connecticut where the lists were not published. According to Marcin, a key determinant in whether a county widely published the list of tax returns is whether or not the local tax registrar could implement the request in a timely way (while the 1924 Revenue Act stood). Appendix Tables A8 and A9 provide summary statistics and a balance table showing very few economically meaningful correlations between county characteristics and tax list circulation. New Deal spending per capita is elevated in counties where the lists

were published.[16] Counties where the lists were published also had a higher rate of home ownership, but no significant differences can be observed between these counties and those without published lists in both the level of wage and non-wage income reported in the 1940 Census.

Individual response patterns on the U.S. Census more broadly look similar between counties that did and did not publish individual tax returns. As an additional test of balance across these locations we examine a pattern in responses where individuals round their ages to specific digits, typically 0 or 5, a phenomenon referred to as age-heaping. Suppose, for example, there are different baseline patterns of obfuscation in counties that publish newspaper list, or nuclear families are more disjointed such that the person being interviewed by the enumerator may be as uncertain about a person's age as they are about their income. In such cases, we could see both age distortion and income distortion in the data that would be unrelated to the publication of the tax lists. In Appendix Figure A5 we report the distribution of Myers Index for age-county cells thereby quantifying the discrepancy between the observed age distribution and a benchmark distribution where ages are perfectly distributed across final digits from 0 to 9, with each digit having an equal likelihood of being reported (10%). We find low levels of age-heaping overall and we also fail to reject the null hypothesis that misreporting is spatially uniform. While distorting age (or lack thereof) is similar across locations, we will show distorting own income exhibits distinct spatial variation.

6.2 *Estimating Tax List Exposure Using Age at Time of Publication*

To proxy individual direct exposure to the publication of income tax returns, we categorize individuals who are 40+ years old at the time they are surveyed for the 1940 Census as more exposed than those who are younger when surveyed. The 40+ year old group would have been close to or over 25 years old at the time of the 1924 Revenue Act, and hence, recently able to vote (over 21) and recently out of college (if attended) and entering the labor market. Given age-heaping and other factors, our threshold age is not intended to be a sharp cutoff; we report results for each 5-year bucket between those aged 25 at the time of the 1940 Census, up to 65, graphically showing results are not sensitive to a specific cutoff year.

In our regression framework age is an indicator coded 0 if an individual was less than 40 at the time of the 1940 census and 1 if an individual was above 40 years old or above at the time they were surveyed for the 1940 Census.

We present results in Table III, to illustrate the relationship between non-disclosure in counties with newspaper lists and different age cohorts using linear probability models. In

[16] This is largely driven by counties in New York state receiving substantial federal relief and recovery aid during the Great Depression.

general, we find that non-disclosure is insignificantly different in news list counties (column 1), controlling for observables. However, it becomes pronounced when we consider individuals aged between 40 and 65 (column 2). While the interaction between newspaper list counties and age in column 3 is imprecisely estimated, it is significant at customary levels in column 4 when we allow covariates to differ by age, as recommended to mitigate omitted variable bias in (Feigenberg et al., 2023). The coefficient on the interaction in column 4 implies non-disclosure is (0.00276 ÷ 0.043)=6.4% higher among 40-65 year olds in newspaper list counties.

When we weight the regressions by the 1920 county population the results are strengthened. The interaction coefficient from column 3 is larger and more precisely estimated in column 5, while the interaction from column 4, estimated with weights in column 5, implies non-disclosure is (0.00579 ÷ 0.043)=13.5% higher among 40-65 year olds in newspaper list counties. Furthermore, Figure 4 shows that the effects are strongly correlated with 40+ year olds, the group most likely to remember the tax list publication on account of being of working age at the time of publication. When estimating the specifications in columns 4 and 6 of Table III with 5-year age cohort dummies, we observe a distinct shift in non-disclosure rates around this age range. The weighted and unweighted estimates are strikingly similar for 30-34 and 35-39 year olds, but diverge for the 40+ age cohorts, which is what we would expect if non-disclosure in the 1940 Census was affected by exposure to newspapers publishing tax lists 15 years prior with magnified spillovers across individuals in more populated counties.

6.3 *Estimating Tax List Exposure Using Newspaper Circulation*

To measure the impact of newspaper circulation on non-disclosure, we follow the assumption in Gentzkow et al. (2014) that news markets can be defined at the county level. Newspaper circulation was highly localized at this time (as it is today) even for national newspapers with distribution outlets across the country. Around 70% of the circulation of the *New York Times*, for example, occurred in the states of New York and New Jersey. Accordingly, our distinction between counties publishing the lists captures the local risk associated with information revelation. Newspapers largely focused on lists of local taxpayers, so circulation would reflect the degree to which this information was made public.

We now outline how we estimate exposure using direct evidence on newspaper circulation of tax lists. We perform two tests. First, suppose N^T represents the set of all newspapers publishing lists in counties and $Circulation(N_n^T, C_c^T)$ the circulation of the n-th newspaper in list county C_c^T, then the total circulation of newspapers in a single list county is $\sum_{n \in N^T} Circulation(N_n^T, C_c^T)$. We construct an exposure measure by scaling total circulation by the county population in 1920. For exposition we assign a value of zero to cases where

there is no circulation and group circulation into low, median, and high levels based on terciles. We also report our estimates weighted by the population of each county in 1920. Figure 5 shows the distribution of scaled circulation with cutoff thresholds.

Appendix Table A10 shows summary statistics across circulation terciles, revealing limited associations between attributes and tax list circulation. Of the 31 variables in total, only 2 of the county level variables and 5 individual level characteristics rise between low to medium, and medium to high circulation counties.[17] We control for all 31 observables in our regressions.

We estimate the following linear probability models:

$$Privacy_i = \sum_{j=1}^{3} \lambda_j Circulation_c + \lambda_4 Age_i^{40-65} + \sum_{j=5}^{7} \lambda_j (Circulation_c \times Age_i^{40-65}) + \beta \mathbf{X}_c + \delta \mathbf{Z}_i + \xi \mathbf{W}_i + \kappa_{state} + \phi_{occupation} + \epsilon_i, \quad (3)$$

where the indicator variables denote *Circulation* the circulation reach of the newspapers publishing the lists in county c—whether the circulation of the lists is in the low, middle or top tercile respectively. Our reference (omitted) category is a county where lists were not published (see Appendix Table A10 for summary statistics) and our reference age category is 25-39 year olds in those locations. Our key coefficients of interest are λ_5, λ_6, and λ_7, measuring variation in non-disclosure rates in counties by circulation intensity and by age cohort. Throughout our analysis we sequentially incorporate controls for individual $\mathbf{Z}_i$ and county-level characteristics $\mathbf{X}_c$, and fixed effects, as in equation 1 as well as a complete set of two-way interactions $\mathbf{W}_i$.

As a second test, we relax the assumption of exposure effects operating only though local news markets and allow for multi-way newspaper circulation across counties using data from Gentzkow et al. (2014). Newspapers published in list counties could circulate in other counties, or in control counties as well. The *Hartford Courant*, for example, circulated across all eight counties in Connecticut whereas the *Hartford Times* circulated across six counties in that state. The *New York Times* circulated across at least 245 U.S. counties and was a prominent publisher of the lists. Analysis of multi-way circulation patterns allows for extended geographic interplay between newspaper circulation of tax lists and its potential effects on income non-disclosure at the time of the 1940 Census.

As before, let C represent the set of counties in states with a top 50 city and N^T the set of all newspapers publishing the lists. $\widehat{Circulation}(N_n^T, C_c)$ then represents the circulation

[17] The county level variables are New Deal spending per capita and the share religious, and individual level characteristics are single, an immigrant, a home owner, house value and capitalized house value

of the n-th newspaper in the c-th county, such that the total circulation of all newspapers in each county is given by $\sum_{n \in N^T} \widehat{Circulation}(N_n^T, C_c)$. We scale circulation by each county's population in 1920 and again group by tercile—low, medium, and high levels—with the distribution of circulation shown in Figure 5.[18] Although the publication of taxpayer information from other locations would be less relevant to readers concerned about their own information disclosure, it may still magnify the perceived costs of income revelation. Thus, we interact our multi-way indicators of circulation exposure with our binary variable $Newslist$ denoting where the disclosures had primarily occurred and our age cohort variable denoting individuals of approximate working age or older at the time of circulation.

We estimate the following linear probability models:

$$\begin{aligned} Privacy_i = & \sum_{k=1}^{2} \psi_k \widehat{Circulation}_{c,k} + \psi_3 Age_i^{40-65} + \psi_4 Newslist_c \\ & + \psi_5 Age_i^{40-65} \times Newslist_c \\ & + \sum_{k=6}^{7} \psi_k (\widehat{Circulation}_{c,k-5} \times Age_i^{40-65}) \\ & + \sum_{k=8}^{9} \psi_k (\widehat{Circulation}_{c,k-7} \times Newslist_c) \\ & + \sum_{k=10}^{11} \psi_k (\widehat{Circulation}_{c,k-9} \times Age_i^{40-65} \times Newslist_c) \\ & + \beta \, \mathbf{X}_c + \delta \, \mathbf{Z}_i + \xi \, \mathbf{W}_i + \kappa_{state} + \phi_{occupation} + \epsilon_i \end{aligned} \tag{4}$$

where $\widehat{Circulation}$, for $k = 1$ represents medium circulation counties and $k = 2$ represents high circulation counties with low tercile circulation counties being the reference category and 25-39 olds in those counties being the reference age category. Our main coefficients of interest are ψ_{10} and ψ_{11}, the triple interaction terms which represent our estimates of the difference in non-disclosure rates among individuals with median and high circulation exposure respectively (compared to those with low exposure or none) in working age cohorts at the time of publication (compared to younger cohorts or those not yet born).

6.4 *Empirical Results and Implied Persuasion*

We report the interaction terms λ_5, λ_6, and λ_7 from Equation 3 graphically in Figure 6. Following the structure of our baseline results in Table I we begin with a specification with

[18] Note, our initial exposure measure assigns zero to no circulation and then groups positive circulation into low, median, and high levels based on terciles, giving 4 categories in total, whereas this exposure measure groups circulation solely by tercile giving 3 categories in total.

age controls and state fixed effects (Panel a), before incorporating various controls, fixed effects as well as sub-sample splits in Panels (b)-(f). Across all estimates we find consistently elevated non-disclosure rates among individuals who were most exposed to the circulation of the tax lists relative to those who were not. We find no discernable non-disclosure effects in low and medium circulation counties, implying a threshold level of newspaper circulation may have been necessary to induce a change in privacy preferences.

In Panel (a) we find non-disclosure is (0.001111 ÷ 0.043)=25.8% higher among 40-65 year olds in the highest exposure newspaper list counties using basic controls, with similar magnitudes in Panel (b) (26.0%), Panel (c) (27.0%) and Panel (d) (24.6%) when incorporating county controls, demographic controls, and occupation fixed effects respectively. Among 40-65 year olds within the most active labor market group in Panel (e) non-disclosure is (0.00657 ÷ 0.027)=24.3% higher whereas in Panel (e) non-disclosure is (0.01138 ÷ 0.043)=26.5% higher for 40-65 year olds respondents who directly interacted with enumerators in high exposure counties. If newspaper lists were genuinely shifting the demand for privacy we would expect to observe this among those with an inclination to read newspapers and with the greatest cognitive capacity to respond, namely individuals with a higher level of education. In Panel (g) we restrict estimation to only the college educated. Non-disclosure is (0.00560 ÷ 0.054)=28.0% higher among 40-65 year olds in newspaper list counties compared to their counterparts in non-list counties. Including two-way interactions in Panel (h) to account for potential omitted variable bias results in a reduction of the effect to 16.9%. In Appendix Figure A6 we show robustness of our results to individuals who remained in newspaper list counties from 1920 to 1940, since around 3% to 5% of 40-65 year olds would have moved states in the five years prior to the 1940 Census (Rosenbloom and Sundstrom, 2003).

One way to interpret the magnitude of these effects is through the lens of persuasion. According to DellaVigna and Gentzkow (2010), the persuasion rate measures how effective the persuasion treatment is in influencing behavior, while accounting for message exposure and the size of the population that still needs to be convinced—approximated empirically using the control group. In our case, the persuasion treatment occurred with the publication of the tax lists around 15 years prior to the 1940 Census, so we can only estimate persuasion using a back-of-the-envelope approach. Conditional on observables using the specification in Figure 6 Panel (h), we find that the predicted non-disclosure rate among college educated 40-65 year olds in news list circulation counties is 6.2% whereas among 40-65 year olds in control counties it is 4.6%. The implied persuasion rate at time $t + 15$ is then 1.68%. If individuals exhibit recency bias when forming beliefs, the implied persuasion rate at time $t = 1$ is 15.9% using the formula for exponential decay with an annual depreciation rate of 15%.[19] Subject

[19]Using the formula in DellaVigna and Gentzkow (2010) the implied persuasion rate P at time $t = 15$ is:

to the stipulation in Jun and Lee (2023) that persuasion rates are difficult to compare across datasets, we can gain some sense of magnitudes. Gentzkow et al. (2011) find a persuasion rate of 12.8% with respect to presidential voter turnout when introducing a newspaper to a county without one previously while switching the political slant of a newspaper from Democrat to Republican is associated with a lower persuasion rate of around 3.4%.

Finally, Figure 7 reports our estimates of equation 4 where we account for the multi-way circulation of newspapers containing the tax lists across counties, including those that published the lists and those that did not. Our focus is on ψ_{10} and ψ_{11} the coefficients on the triple interaction: $\widehat{Circulation} \times Newslist \times Age : 40-65$, as we show the effects were localized to counties where the tax lists were published. Indeed, in Panel (a) we use a preliminary specification without any interactions, regressing non-disclosure on indicators for medium and high exposure counties relative to low exposure counties as the omitted category, showing there is no effect of exposure intensity on non-disclosure. In Panel (b) we interact the exposure indicators with an age dummy for 40-65 year olds and we cannot reject the hypothesis that the effect of exposure intensity on non-disclosure is zero across counties. We then exploit additional variation by newspaper list county, plotting the triple interaction in Panels (c) and (d). In Panel (c), we find that non-disclosure is 12.6% higher among individuals aged 40-65 in high-exposure locations, and this effect increases further to 14.5% in Panel (d) when we account for a comprehensive set of two-way interactions.

Overall, our results provide suggestive evidence that demand for privacy from the U.S. Census is a function of past exposure to the release of federal income data, especially in situations of heightened exposure where individuals might be more likely to value privacy and prioritize safeguarding their personal information. Exposure can increase the perceived risk that federal data may become public at a future date, and offers first hand experience as to the consequences of that publicity.

6.5 *Counterfactual Distributions and Survey Data Distortion*

To what extent would this form of privacy preference influence the composition of survey data? We present an illustrative analysis using estimates of counterfactual distributions. We use the newspaper list data discussed above to examine what the distribution of Census incomes would have looked like without exposure to elevated concerns about the release of income data. In a counterfactual world we can remove this sensitivity and analyze the full range of reported incomes not just the zero or missing observations. We ask what would be

$100 \cdot \frac{0.062-0.046}{1} \cdot \frac{1}{1-0.046} = 1.68\%$ whereas the implied persuasion rate of 15.9% at time $t = 1$ is calculated using the following formula: $P(t) = P_0 \cdot e^{-rt}$ where $r = 0.15$ and $t = 15$. If $r = 0.10$ or $r = 0.20$ the implied persuasion rates at time $t = 1$ are 7.5% and 33.7% respectively.

the distribution of incomes for individuals in newspaper list counties if they reported their incomes like individuals with the same characteristics in non-newspaper list counties?

Of the several approaches to constructing counterfactual distributions, we use kernel re-weighting methods following DiNardo et al. (1996) (DFL) for computational feasibility given the large number of fixed effects we introduce. We first estimate a propensity score p using an indicator for individuals in newspaper list counties (coded 0) and non-newspaper list counties (coded 1) using age, state and full occupation fixed effects and all the county-level and individual-level covariates we have used in estimation so far. We then use the propensity score to both define an area of common support and re-weight the kernel density functions using DFL weights—specified as $(1-p)/p$ for treatment counties and $p/(1-p)$ for control counties. Our interest is in the gap between the distribution of reported incomes in treatment counties and the counterfactual distribution that would have otherwise prevailed.

Figure 8 Panels (a) and (b) show the distribution of reported incomes in treatment and control counties for individuals 25-65 years of age and the gap between these distributions. The remaining panels compare actual treatment and counterfactual treatment income distributions for these individuals while repeating the exercise for individuals who were 40-65 years of age, or were the direct respondent to the enumerator.

Among those above 25 when exposed to the publicized IRS tax lists, the 90/10 ratio is 7.0 when calculated using the reported distribution of incomes in newspaper list counties compared to 11.4 in the counterfactual distribution. On this measure, a policy maker might infer inequality was less pronounced than it probably was. We also note that these patterns of distortion are consistent with Figure 2, our population-level correlations between predicted income and non-disclosure.

7. CONCLUSION

Debate over the protection of privacy is fundamental to society in an age where personal data has value to governments, firms and researchers but individuals have concerns about intrusions or the potential for misuse. We have examined the origins of privacy concerns in Census data using the unprecedented circumstance associated with the 1940 Census when millions of U.S. residents, or their representatives, were asked to disclose their incomes.

We find meaningful resistance through non-disclosure where local inequality was pronounced and in places where personal data collected by the federal government had been released publicly 15 years prior. Our preferred interpretation of these results is that individuals form expectations about the private stakes of revealing their personal information. The private stakes are higher when the likelihood of a leak is higher, and when the respondent

is more highly differentiated from their reference group (e.g. on either extreme end of a dispersed income distribution). We find little evidence that privacy concerns are related to politics, or redistribution policies. Furthermore, we also find that resistance leads government statistics to underestimate the true extent of inequality. Overall, these findings highlight how the perceived instrumental value of privacy affects the quality of data collection efforts, and the long-lived consequences of the publication of individual-level data.

References

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

Abramitzky, R., L. Boustan, and M. Rashid (2020). Census linking project: Version 1.0 [dataset]. Data retrieved from, https://censuslinkingproject.org.

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

Acquisti, A., C. Taylor, and L. Wagman (2016b). The Economics of Privacy. *Journal of Economic Literature 54*(2), 442–492.

Adjerid, I., A. Acquisti, L. Brandimarte, and G. Loewenstein (2013). Sleights of Privacy: Framing, Disclosures, and the Limits of Transparency. *Proceedings of the Ninth Symposium on Usable Privacy and Security*, 1–9.

Aghion, P., Y. Algan, P. Cahuc, and A. Shleifer (2010, 08). Regulation and Distrust*. *The Quarterly Journal of Economics 125*(3), 1015–1049.

Arellano-Bover, J. (2022). Displacement, diversity, and mobility: Career impacts of japanese american internment. *The Journal of Economic History 82*(1), 126–174.

Athey, S., C. Catalini, and C. Tucker (2017). The Digital Privacy Paradox: Small Money, Small Costs, Small Talk. *NBER Working Paper No. 23488*.

Becker, G. S. (1980). Privacy and malfeasance: A comment. *The Journal of Legal Studies 9*(4), 823–826.

Bø, E. E., J. Slemrod, and T. O. Thoresen (2015). Taxes on the internet: Deterrence effects of public disclosure. *American Economic Journal: Economic Policy 7*(1), 36–62.

Bouk, D. (2022). *Democracy's Data: The Hidden Stories in the U.S. Census and How to Read Them*. MCD.

Bowen, C. M. (2024). Government data of the people, by the people, for the people: Navigating citizen privacy concerns. *Journal of Economic Perspectives 38*(2), 181–200.

Brandes, S. D. (1983). America's super rich, 1941. *The Historian 45*(3), 307–323.

Budd, J. W. and T. Guinnane (1991). Intentional age-misreporting, age-heaping, and the 1908 old age pensions act in ireland. *Population Studies 45*(3), 497–518.

Bureau of the Census (1943). Bureau of the Census: Sixteenth Census of the United States, 1940.

Caprettini, B. and H.-J. Voth (2022, 06). New Deal, New Patriots: How 1930s Government Spending Boosted Patriotism During World WarÂ II. *The Quarterly Journal of Economics 138*(1), 465–513.

Chi, F. (2022). Information waves and firm investment. *Working Paper*.

Chin, A. (2005, July). Long-Run Labor Market Effects of Japanese American Internment during World War II on Working-Age Male Internees. *Journal of Labor Economics 23*(3), 491–526.

Clubb, J. M., W. H. Flanigan, and N. H. Zingale (2006). Electoral Data for Counties in the United States: Presidential and Congressional Races, 1840-1972 [ICPSR 8611,dataset]. Technical report, ICPSR.

Cowell, F. A. and E. Flachaire (2023). Inequality measurement and the rich: Why inequality increased more than we thought. *Review of Income and Wealth 0*(0), 1–24.

Cullen, Z. and R. Perez-Truglia (2022). How Much Does Your Boss Make? The Effects of Salary Comparisons. *Journal of Political Economy 130*(3), 766–822.

Cullen, Z. and R. Perez-Truglia (2023). The Salary Taboo: Privacy Norms and the Diffusion of Information. *Journal of Public Economics 222*, 104890.

DellaVigna, S. and M. Gentzkow (2010). Persuasion: Empirical evidence. *Annual Review of Economics 2*(1), 643–669.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica 64*(5), 1001–1044.

Dinur, I. and K. Nissim (2003). Revealing information while preserving privacy. New York, NY, USA. Association for Computing Machinery.

Dolan, P., M. Hallsworth, D. Halpern, D. King, R. Metcalfe, and I. Vlaev (2012). Influencing behaviour: The mindspace way. *Journal of economic psychology 33*(1), 264–277.

Dray, S., C. Landais, and S. Stantcheva (2022). Wealth and property taxation in the united states. Working paper, Harvard University.

Durantini, M. R., D. Albarracin, A. L. Mitchell, A. N. Earl, and J. C. Gillette (2006). Conceptualizing the influence of social agents of behavior change: A meta-analysis of the effectiveness of hiv-prevention interventionists for different groups. *Psychological bulletin 132*(2), 212.

El-Badry, S. and D. A. Swanson (2007). Providing census tabulations to government security agencies in the united states: The case of arab americans. *Government Information Quarterly 24*(2), 470–487.

Enke, B. (2020). Moral values and voting. *Journal of Political Economy 128*(10), 3679–3729.

Feigenberg, B., B. Ost, and J. A. Qureshi (2023, 07). Omitted Variable Bias in Interacted Models: A Cautionary Tale. *The Review of Economics and Statistics*, 1–47.

Fishback, P. V., S. Kantor, and J. J. Wallis (2003). Can the New Deal's three Rs be rehabilitated? A program-by-program, county-by-county analysis. *Explorations in Economic History 40*(3), 278–307.

Gatewood, G. (2001). *A Monograph on Confidentiality and Privacy in the U.S. Census*. United States Government Printing Office.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2011, December). The effect of newspaper entry and exit on electoral politics. *American Economic Review 101*(7), 2980–3018.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2014, October). Competition and ideological diversity: Historical evidence from us newspapers. *American Economic Review 104*(10), 3073–3114.

Glaeser, E. L. and A. Shleifer (2003, June). The rise of the regulatory state. *Journal of Economic Literature 41*(2), 401–425.

Goldfarb, A. and C. Tucker (2012). Shifts in Privacy Concerns. *American Economic Review 102*(3), 349–353.

Goldfield, E. D. (1958). *Decennial Census and Current Population Survey Data on Income*, pp. 37–62. Princeton University Press.

Goldin, C. and L. F. Katz (2000). Education and income in the early twentieth century: Evidence from the prairies. *Journal of Economic History 60*(3), 782–818.

Haines, M. R. (2010). Historical, Demographic, Economic, and Social Data: The United States, 1790-2002 [ICPSR 2896,dataset]. Technical report, ICPSR.

Hauser, O. P. and M. I. Norton (2017). (Mis)perceptions of inequality. *Current Opinion in Psychology 18*, 21–25.

Igo, S. (2018). *The Known Citizen: A History of Privacy in Modern America*. Harvard University Press.

Jaeger, S., C. Roth, N. Roussille, and B. Schoefer (2021, December). Worker beliefs about outside options. *NBER Working Paper 29623*.

Johnson, D. K. (2023). *The Lavender Scare: The Cold War Persecution of Gays and Lesbians in the Federal Government*. University of Chicago Press.

Jun, S. J. and S. Lee (2023). Identifying the effect of persuasion. *Journal of Political Economy 131*(8), 2032–2058.

Kreiner, C. T., K. B. Hvidberg, and S. Stantcheva (2022). Social positions and fairness views on inequality. *Review of Economic Studies*.

Kuziemko, I., M. I. Norton, E. Saez, and S. Stantcheva (2015). How elastic are preferences for redistribution? Evidence from randomized survey experiments. *American Economic Review*.

Lenter, D., J. Slemrod, and D. Shackelford (2003). Public disclosure of corporate tax return information: Accounting, economics, and legal perspectives. *National Tax Journal 56*(4), 803–830.

Levi, M. and L. Stoker (2000). Political trust and trustworthiness. *Annual Review of Political Science 3*(1), 475–507.

Lin, T. (2022). Valuing intrinsic and instrumental preferences for privacy. *Marketing Science 41*(4), 663–681.

Luttmer, E. F. P. (2005, 08). Neighbors as Negatives: Relative Earnings and Well-Being. *The Quarterly Journal of Economics 120*(3), 963–1002.

Malmendier, U. and J. A. Wachter (2024). Memory of past experiences and economic decisions. In M. Kahana and A. Wagner (Eds.), *Handbook of Human Memory*. Oxford University Press.

Marcin, D. (2014). Essays on the revenue act of 1924. Phd thesis, University of Michigan.

McCleary, R. M. and R. J. Barro (2006, June). Religion and economy. *Journal of Economic Perspectives 20*(2).

Moehling, C. M. (2001). Women's work and men's unemployment. *The Journal of Economic History 61*(4), 926–949.

Norton, M. I. and D. Ariely (2011). Building a better america-one wealth quintile at a time. *Perspectives on psychological science 6*(1), 9–12.

Perez-Truglia, R. (2020, April). The effects of income transparency on well-being: Evidence from a natural experiment. *American Economic Review 110*(4), 1019–54.

Piketty, T. and E. Saez (2003, 02). Income Inequality in the United States, 1913-1998. *The Quarterly Journal of Economics 118*(1), 1–41.

Posner, R. A. (1981). The economics of privacy. *The American Economic Review 71*(2), 405–409.

Price, D. O. (1947). A check on underenumeration in the 1940 census. *American Sociological Review 12*(1), 44–49.

Rosenbloom, J. L. and W. A. Sundstrom (2003, July). The decline and rise of interstate migration in the united states: Evidence from the ipums, 1850-1990. Working Paper 9857, National Bureau of Economic Research.

Roxworthy, E. (2008). *The Spectacle of Japanese American Trauma: Racial Performativity and World War II*. University of Hawaii Press.

Ruggles, S. (2024). When privacy protection goes wrong: How and why the 2020 census confidentiality program failed. *Journal of Economic Perspectives 38*(2), 201–26.

Ruggles, S., C. A. Fitch, R. Goeken, J. D. Hacker, M. A. Nelson, E. Roberts, M. Schouweiler, and M. Sobek (2021). IPUMS Ancestry Full Count Data: Version 3.0 [dataset]. Technical report, IPUMS.

Saavedra, M. (2021). Kenji or kenneth? pearl harbor and japanese-american assimilation. *Journal of Economic Behavior & Organization 185*, 602–624.

Seipp, D. J. (1981). The right to privacy in american history. Program on information resources policy report, Harvard University.

Seltzer, W. and M. Anderson (2001). The dark side of numbers: The role of population data systems in human rights abuses. *Social Research 68*(2).

Serrato, J. C. S. and P. Wingender (2016). Estimating local fiscal multipliers. *NBER Working Paper No. 22425*.

Stark, R. (1992). The reliability of historical united states census data on religion. *Sociological Analysis 53*(1), 91–95.

Stigler, G. J. (1980). An introduction to privacy in economics and politics. *The Journal of Legal Studies 9*(4), 623–644.

Thomson, C. A. H. (1940). Public relations of the 1940 census. *The Public Opinion Quarterly 4*(2), 311–318.

Tucker, C. (2023). The economics of privacy: An agenda. Working paper, MIT.

United States Senate (1940, February-March). Hearings Before the Subcommittee of a Committee on Commerce, United States 76th Congress: A Resolution Favoring the Deletion from the 16th Census Population Schedule of Inquiries Numbered 32 and 33, Relating to Compensation Received.

Willison, G. F. (1940). World's greatest quiz session. Reader's Digest.

Table I: Privacy Demands and Inequality: 90/10 Income Ratio

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: 90/10 Income Ratio								
New Deal Spending per Capita (std)	-0.00011	0.00005	0.00011	0.00003	0.00002	0.00000	-0.00067*	0.00033
	(0.00038)	(0.00038)	(0.00037)	(0.00038)	(0.00037)	(0.00037)	(0.00036)	(0.00028)
Republican Vote Share (std)	0.00151**	0.00069	0.00059	0.00102	0.00092	0.00094	0.00126**	0.00108*
	(0.00077)	(0.00072)	(0.00070)	(0.00066)	(0.00065)	(0.00064)	(0.00058)	(0.00065)
Share Religious (std)	-0.00251***	-0.00055	-0.00070	-0.00096*	-0.00100**	-0.00100**	-0.00097**	-0.00064
	(0.00055)	(0.00054)	(0.00052)	(0.00052)	(0.00050)	(0.00050)	(0.00048)	(0.00048)
90/10 Income Ratio (std)	0.00989***	0.00708***	0.00653***	0.00432***	0.00401***	0.00367***	0.00366***	0.00271***
	(0.00063)	(0.00055)	(0.00052)	(0.00047)	(0.00047)	(0.00046)	(0.00050)	(0.00046)
Panel B: Predicted 90/10 Income Ratio								
New Deal Spending per Capita (std)	-0.00027	-0.00008	-0.00000	-0.00002	-0.00001	-0.00002	-0.00076**	0.00031
	(0.00041)	(0.00040)	(0.00039)	(0.00039)	(0.00039)	(0.00038)	(0.00039)	(0.00029)
Republican Vote Share (std)	0.00194**	0.00130*	0.00110	0.00123*	0.00109	0.00107	0.00150**	0.00113*
	(0.00081)	(0.00073)	(0.00071)	(0.00068)	(0.00067)	(0.00066)	(0.00060)	(0.00067)
Share Religious (std)	-0.00307***	-0.00089*	-0.00097*	-0.00101*	-0.00101**	-0.00098*	-0.00110**	-0.00056
	(0.00056)	(0.00054)	(0.00052)	(0.00053)	(0.00051)	(0.00051)	(0.00048)	(0.00049)
Predicted 90/10 Income Ratio (std)	0.00813***	0.00655***	0.00569***	0.00276***	0.00234***	0.00190***	0.00296***	0.00104**
	(0.00078)	(0.00064)	(0.00061)	(0.00057)	(0.00056)	(0.00054)	(0.00053)	(0.00053)
Panel C: Income Inequality (MLD)								
New Deal Spending per Capita (std)	-0.00001	0.00011	0.00016	0.00006	0.00005	0.00003	-0.00064*	0.00034
	(0.00036)	(0.00036)	(0.00036)	(0.00038)	(0.00037)	(0.00037)	(0.00036)	(0.00028)
Republican Vote Share (std)	0.00160**	0.00079	0.00066	0.00105	0.00094	0.00096	0.00128**	0.00109*
	(0.00070)	(0.00068)	(0.00067)	(0.00065)	(0.00064)	(0.00063)	(0.00056)	(0.00065)
Share Religious (std)	-0.00235***	-0.00066	-0.00080*	-0.00099*	-0.00102**	-0.00101**	-0.00100**	-0.00063
	(0.00051)	(0.00050)	(0.00048)	(0.00051)	(0.00049)	(0.00049)	(0.00046)	(0.00048)
Income Inequality (std)	0.01165***	0.00884***	0.00817***	0.00490***	0.00446***	0.00401***	0.00470***	0.00282***
	(0.00064)	(0.00059)	(0.00056)	(0.00054)	(0.00053)	(0.00052)	(0.00052)	(0.00051)
Observations	22281334	22281334	22281334	22281334	22281334	22281334	11599405	5383435
Clusters	3075	3075	3075	3075	3075	3075	3075	3075
Mean Dep Var.	0.047	0.047	0.047	0.047	0.047	0.047	0.035	0.047
Age Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic Controls	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Main Occ. FE	No	No	No	Yes	No	No	No	No
Sub Occ. FE	No	No	No	No	No	No	No	No
Full Occ. FE	No	No	No	No	Yes	Yes	Yes	Yes
Sample	Full	Full	Full	Full	Full	Full	52 wks 40 hrs	Resp.

Notes: This table reports linear probability regression coefficients where the dependent variable is 1,0 for privacy (non-disclosure). Age controls are a linear and quadratic term in age. County controls are the 1940 population, manufacturing value-added, the share urban and the share of the population 25+ completing high school. Demographic controls are indicators for gender, household head, marriage status (married, divorced/separated single), race, immigrant and college attendance and a continuous variable for capitalized house values. Occupation fixed effects at the main (11), sub (22) and full (228) levels. Standard errors clustered by county in parentheses. *p<0.1, **p<0.05, ***p<0.01.

Table II: Privacy and the Subject-Enumerator Housing Wealth Gap

	Capitalized		Reported		Capitalized		Reported	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
House Gap Ratio (log)	0.00232***	0.00246***	0.01148***	0.01143***	0.00265***	0.00292***	0.01278**	0.01630***
	(0.00064)	(0.00064)	(0.00403)	(0.00443)	(0.00081)	(0.00081)	(0.00548)	(0.00590)
Placebo House Gap Ratio (log)		-0.00041		0.00162		-0.00001		0.00214
		(0.00055)		(0.00229)		(0.00080)		(0.00304)
Difference, p-value		0.0007		0.0293		0.0111		0.0181
Mean Dep Var.	0.042	0.042	0.043	0.043	0.042	0.042	0.043	0.043
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender Matching	No	No	No	No	Yes	Yes	Yes	Yes
Clusters	181252	181252	18290	18290	124971	124971	13251	13251
Observations	276712	276712	27328	27328	163887	163887	17320	17320

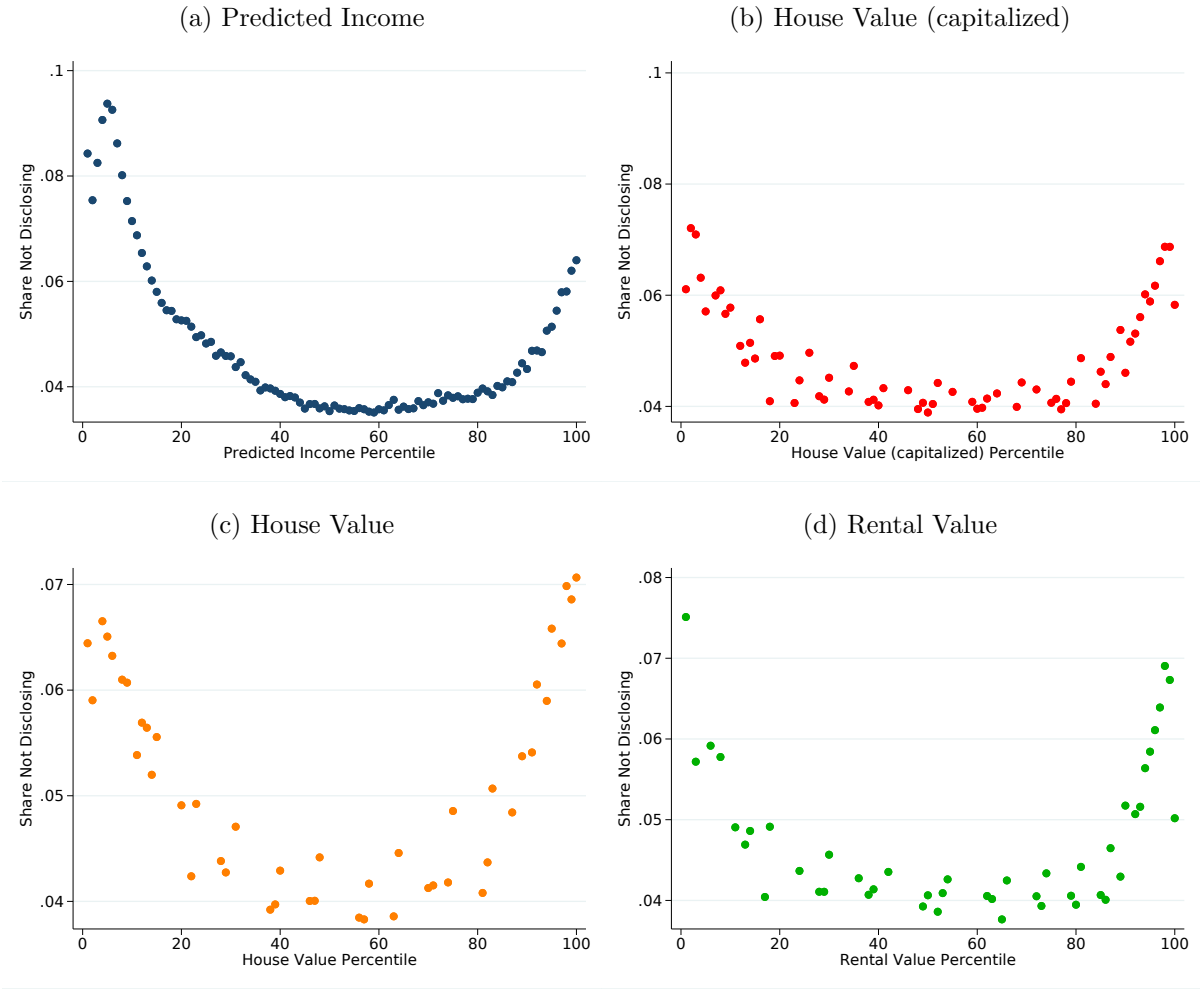
Notes: This table reports linear probability regression coefficients where the dependent variable is 1,0 for privacy (non-disclosure). The right-hand side variables specify the log of the housing wealth gap between the subject and the enumerator or the subject and a randomly selected enumerator of the same gender drawn from an enumeration district in a different state. Controls include the log of the subjects own level of housing wealth, their age and years of education, gaps in age and education between subject and (random) enumerator and mean log housing wealth, age and educational attainment in an enumeration district. Standard errors clustered by household in parentheses. *p<0.1, **p<0.05, ***p<0.01.

Table III: Privacy and Exposure to Newspaper Tax Lists

	(1)	(2)	(3)	(4)	(5)	(6)
Newslist	0.00252	0.00252	0.00163	0.00124	0.00190	0.00082
	(0.00362)	(0.00362)	(0.00369)	(0.00363)	(0.00399)	(0.00389)
Age: 40-65		0.00264***	0.00114	-0.01312***	-0.00066	-0.01823***
		(0.00041)	(0.00115)	(0.00277)	(0.00138)	(0.00402)
Newslist × Age: 40-65			0.00188	0.00275**	0.00349*	0.00579***
			(0.00147)	(0.00122)	(0.00205)	(0.00163)
Observations	7772606	7772606	7772606	7772606	7772606	7772606
Clusters	49	49	49	49	49	49
Mean Dep Var.	0.043	0.043	0.043	0.043	0.043	0.043
Age Controls	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes
County Controls	Yes	Yes	Yes	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Full Occ. FE	Yes	Yes	Yes	Yes	Yes	Yes
Two-Way Interactions			No	Yes	No	Yes
Weighted			No	No	Yes	Yes

Notes: This table reports linear probability regression coefficients where the dependent variable is 1,0 for privacy (non-disclosure). Newslist is an indicator for counties where tax lists were published. Age controls are a linear and quadratic term in age. County controls are the 1940 population, manufacturing value-added, the share urban and the share of the population 25+ completing high school. Demographic controls are indicators for gender, household head, marriage status (married, divorced/separated single), race, immigrant and college attendance and a continuous variable for capitalized house values. Occupation fixed effects at the full (228) level. In columns 5 and 6 regressions are weighted by each county's population in 1920. Standard errors clustered by county in parentheses. *p<0.1, **p<0.05, ***p<0.01.

Figure 1: Own Rank and Income Privacy Demands

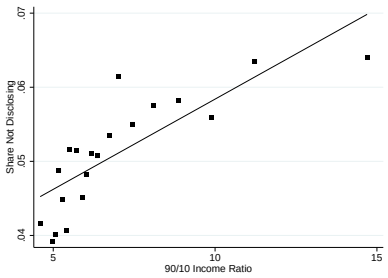


Notes: These figures show the mean income non-disclosure rate by income and wealth percentiles. The predicted income series used in Panel (a) is described in the notes to Appendix Figure A3. In Panel (b) we use capitalized house values using the actual house values and capitalized rental values whereas Panels (c) and (d) use actual house and rental values reported.

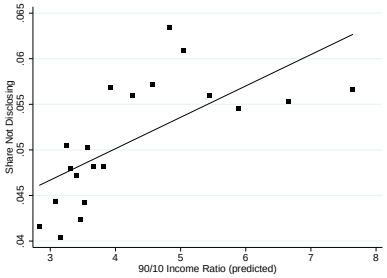
Figure 2: Population Predictors of Income Privacy Demands

Local Inequality vs. Income Privacy

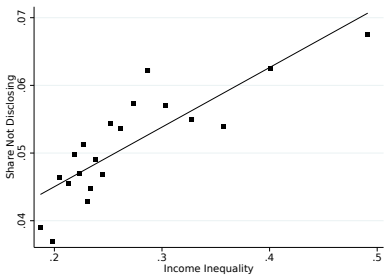
(a) 90/10 Income Ratio



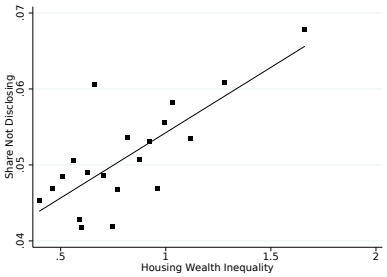
(b) Predicted 90/10



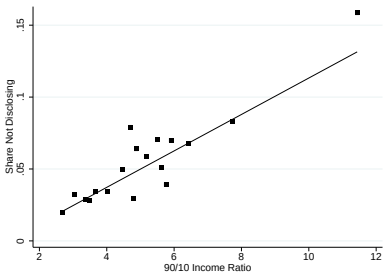
(c) Income Inequality (MLD)



(d) Housing Inequality (MLD)

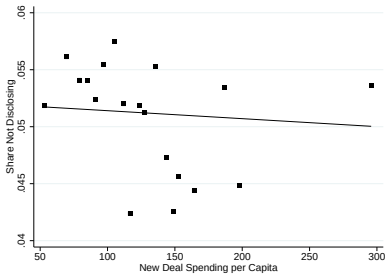


(e) Occupation 90/10

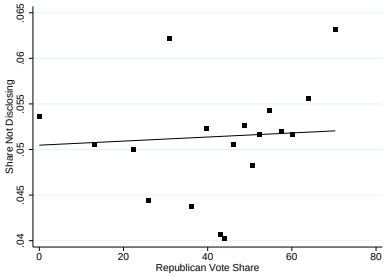


Political and Religious Views vs. Income Privacy

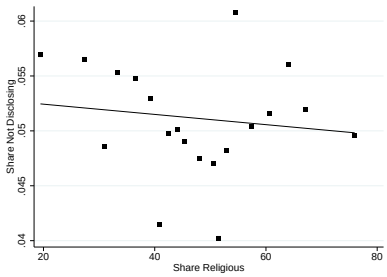
(f) New Deal per Capita



(g) Share Republican

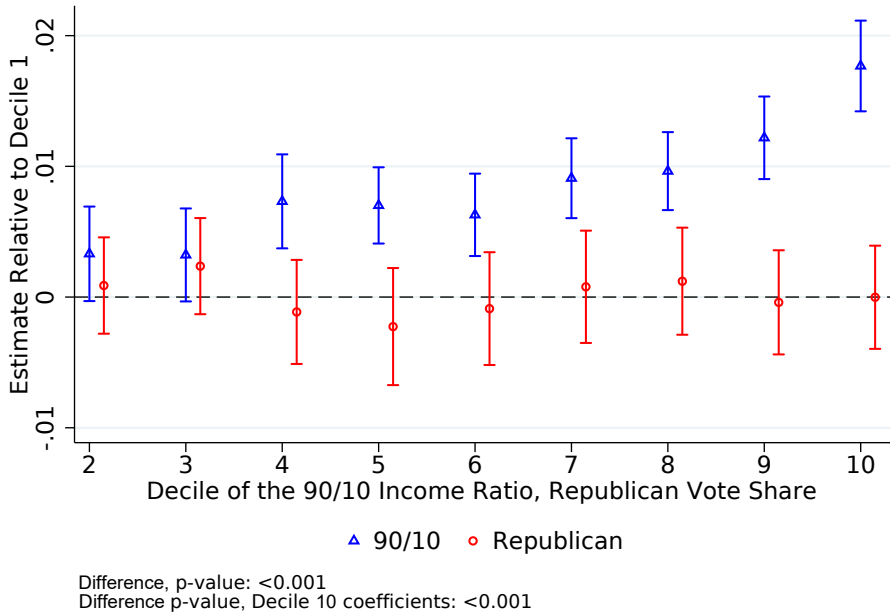


(h) Share Religious



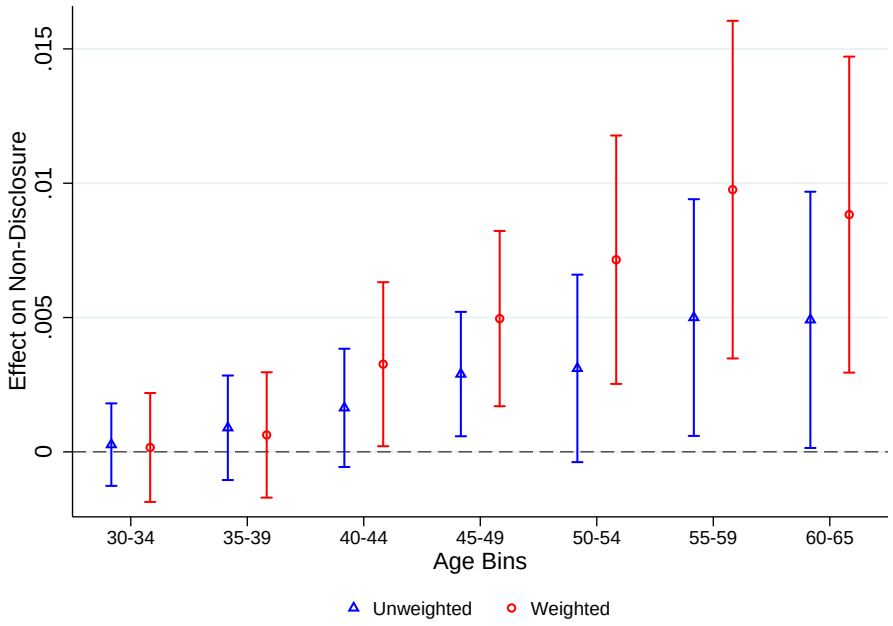
Notes: These figures show binned scatter plots. The predicted 90/10 income ratio is estimated using a predicted income series described in the notes to Appendix Figure A3. Housing and income inequality are the mean-log deviation of house values and incomes at the county-level respectively whereas the occupation 90/10 income ratio is the within occupation ratio using 228 US census occupation categories.

Figure 3: Privacy Demands by Inequality and Republican Vote Share Decile



Notes: This figure reports coefficients and 95% confidence intervals using indicators by decile of the 90/10 income ratio and the Republican vote share. Specification includes age, county and demographic controls, state fixed effects and full occupation fixed effects. The reported p-values are from tests of the null that the coefficients on the indicators by decile of the 90/10 income ratio are equal to the coefficients on the indicators for decile of the Republican vote share.

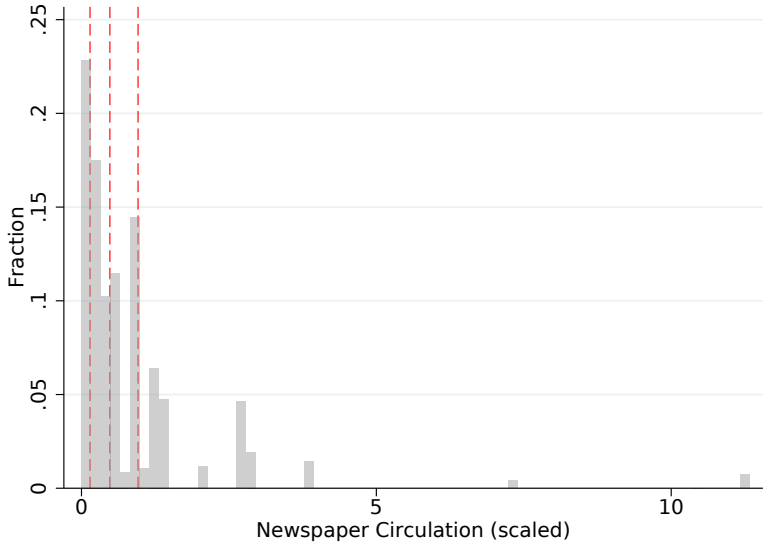
Figure 4: Privacy by Age Exposure in News List v. Non-News List Counties



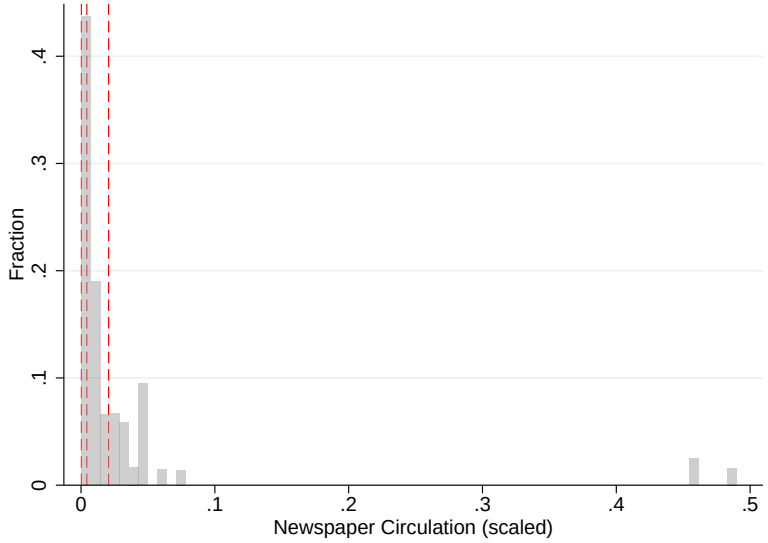
Notes: This figures reports coefficients and 95% confidence intervals of privacy demands for individuals according to their age cohorts and exposure to newspaper lists. The coefficients are interactions between a newspaper list indicator and age cohorts following the specification in column 4 (unweighted) and column 6 (weighted) of Table III. The omitted category is 25-29 year-olds.

Figure 5: Histograms of Newspaper Circulation

(a) County Circulation

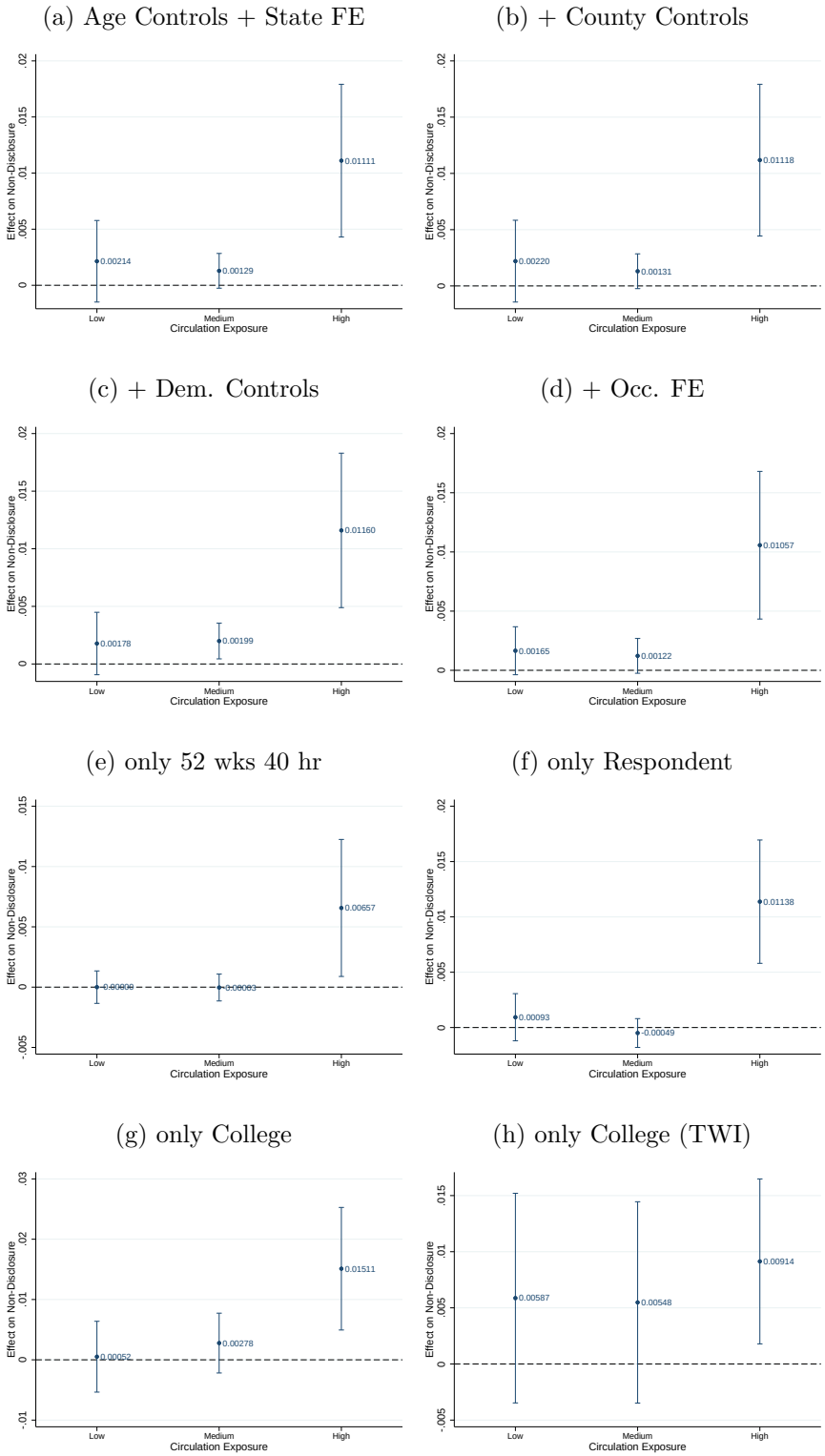


(b) County Circulation (Multi-Town)



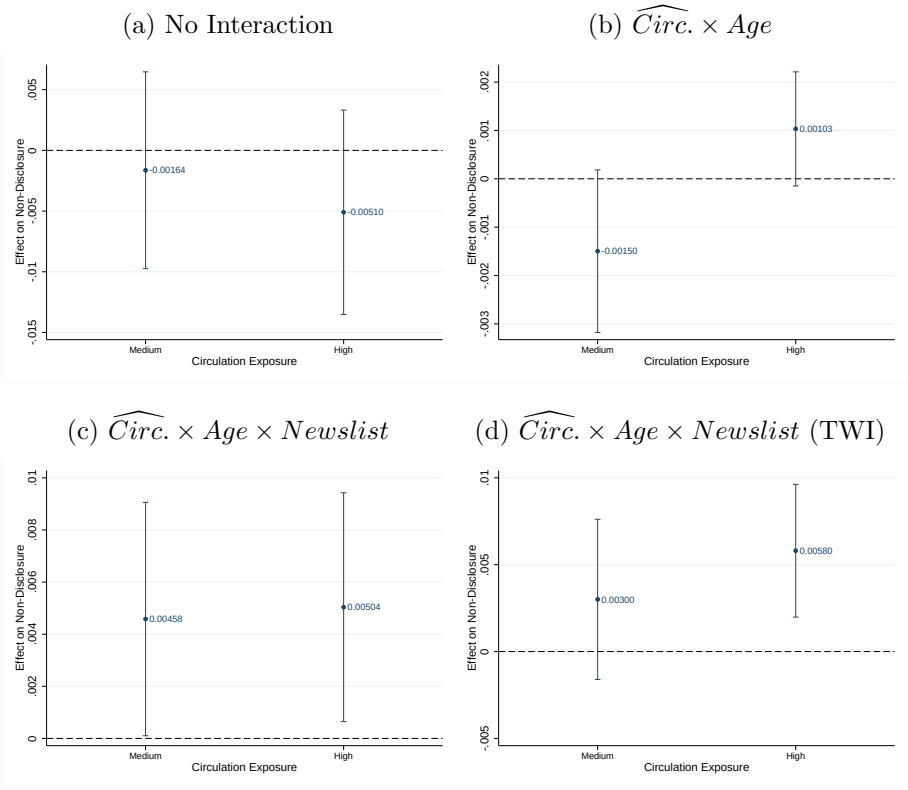
Notes: This figure shows the distribution of newspaper circulation numbers scaled by 1920 county population. Panel (a) measures circulation in counties with newspapers that published the lists. Panel (b) allows for those newspapers to circulate in other counties as well. Vertical dashed lines represent splits by tercile. Panel (a) includes weekday and weekend circulation. Panel (b) uses weekday circulation based on the data from Gentzkow et al. (2014).

Figure 6: Non-Disclosure: Newspaper List Counties and Circulation Exposure



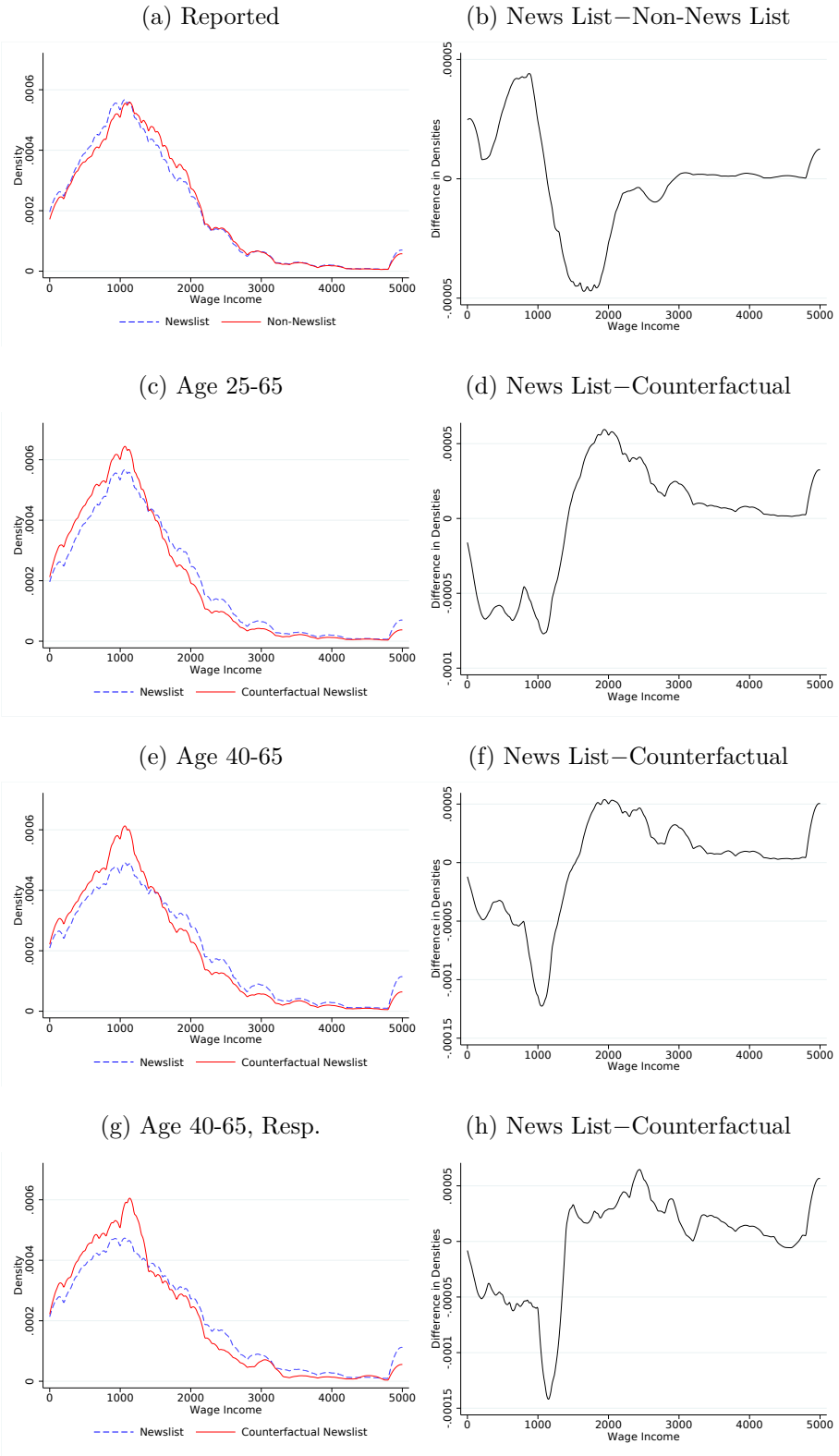
Notes: These figures reports coefficients and 95% confidence intervals of non-disclosure for individuals according to their newspaper list county status and age cohorts. These are plots of λ_5 λ_6 and λ_7 from Equation 3. Panels (a)-(d) incorporate different controls and fixed effects sequentially so Panel (d) includes age controls, state fixed effects, county controls, demographic controls and occupation fixed effects. Panels (e)-(g) include these as well and Panel (h) adds two-way interactions (TWI). Regressions are weighted by each county's population in 1920. Standard errors clustered by county.

Figure 7: Non-Disclosure: Newspaper List Counties and Multi-Way Circulation Exposure



Notes: These figures reports coefficients and 95% confidence intervals of non-disclosure for individuals according to their newspaper list county status and age cohorts. Panel (a) includes the exposure indicator variables in the original, non-interacted form whereas Panel (b) interacts these variables with the indicator for age. Panels (c) and (d) are plots of the triple interactions, ψ_{10} and ψ_7, from Equation 4 (i.e. interacting the exposure indicators with the indicator for age and the indicator for news list counties). All specifications include age controls, county controls, demographic controls, occupation and state fixed effects. TWI refers to two-way interactions. Regressions are weighted by each county's population in 1920. Standard errors clustered by county.

Figure 8: Reported and Counterfactual Income Distributions



Notes: These figures show actual and counterfactual income distributions where the counterfactual for news list counties is constructed using re-weighted kernel density functions. The right-hand side figures show the difference in distributions, news list minus non-news list or news list minus counterfactual.

Online Appendix

Demand for Privacy from the U.S. Census

Zoë Cullen
Harvard Business School

Tom Nicholas
Harvard Business School

Table A1: Descriptives

	Missing Income		Zero Income		Income Reported	
	Mean	SD	Mean	SD	Mean	SD
Panel A: County-Level						
New Deal Spending per Capita	128.600	67.505	133.734	78.929	133.173	71.467
Republican Vote Share	40.887	21.716	35.747	22.411	37.556	21.551
Share Religious	47.972	14.169	45.997	14.354	47.047	14.031
Share Urban	61.969	31.015	61.742	32.546	66.184	30.161
Share Educated	24.173	7.304	24.469	7.916	25.077	7.558
County Population 1940 (Millions)	0.543	0.877	0.664	1.030	0.700	1.049
Manufacturing Value Added (Millions)	354.687	681.030	474.344	866.078	487.991	862.583
90/10 Income Ratio	7.088	2.532	7.453	2.855	6.968	2.531
Predicted 90/10 Income Ratio	4.274	1.251	4.427	1.332	4.239	1.281
Housing Wealth Inequality	0.865	0.334	0.828	0.324	0.805	0.302
Income Inequality	0.272	0.076	0.283	0.085	0.268	0.076
Panel B: Occupation-Level						
Occupation 90/10 Income Ratio	5.375	2.067	6.089	2.411	5.021	1.690
Panel C: Individual-Level						
Privacy	1.000	0.000	1.000	0.000	0.000	0.000
Age	40.709	10.781	41.280	11.331	39.599	10.546
Male = 1	0.751	0.432	0.650	0.477	0.755	0.430
Household Head = 1	0.579	0.494	0.508	0.500	0.649	0.477
Married = 1	0.724	0.447	0.741	0.438	0.774	0.418
Divorced/Separated = 1	0.023	0.151	0.032	0.176	0.023	0.149
Single = 1	0.253	0.435	0.227	0.419	0.203	0.402
White = 1	0.918	0.275	0.883	0.321	0.905	0.294
Immigrant = 1	0.106	0.308	0.146	0.353	0.139	0.346
Years of Education	9.591	3.713	8.853	3.814	9.093	3.602
College = 1	0.168	0.374	0.133	0.340	0.130	0.336
Home Owner = 1	0.530	0.499	0.582	0.493	0.578	0.494
House Value	4423.168	7543.008	3901.968	8834.087	3846.472	6875.483
Rental Value	89.912	478.343	58.375	347.402	62.106	361.948
Capitalized House Value	7795.931	42219.042	5708.250	32350.924	5931.850	33376.577
Weeks Worked	47.503	9.828	47.279	10.374	45.076	11.320
Hours Worked	43.956	11.785	44.480	15.206	43.095	11.536
Wage Income	.	.	0.000	0.000	1224.076	891.091
Non-Wage Income = 1	0.285	0.451	0.565	0.496	0.133	0.340

Notes: This table reports descriptive statistics for observations in the 1940 census for individuals aged 25 to 65 years who were in the labor force, who self-reported being at work, and who received wages or a salary, including those who worked in government. We define income non-disclosure as missing incomes and zero incomes. We estimate capitalized house values using the actual house values and capitalized rental values at 10%. The predicted 90/10 income ratio is estimated using a predicted income series described in the notes to Appendix Figure A3. The occupation 90/10 income ratio is the within occupation ratio using 228 US census occupation categories. The variables "Housing Wealth Inequality" and "Income Inequality" are the mean-log deviation of house values and incomes at the county-level respectively.

Table A2: Privacy Demands and Inequality: Demographic Controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
New Deal Spending per Capita (std)	-0.00011	0.00005	0.00011	0.00003	0.00002	0.00000	-0.00067*	0.00033
	(0.00038)	(0.00038)	(0.00037)	(0.00038)	(0.00037)	(0.00037)	(0.00036)	(0.00028)
Republican Vote Share (std)	0.00151**	0.00069	0.00059	0.00102	0.00092	0.00094	0.00126**	0.00108*
	(0.00077)	(0.00072)	(0.00070)	(0.00066)	(0.00065)	(0.00064)	(0.00058)	(0.00065)
Share Religious (std)	-0.00251***	-0.00055	-0.00070	-0.00096*	-0.00100**	-0.00100**	-0.00097**	-0.00064
	(0.00055)	(0.00054)	(0.00052)	(0.00052)	(0.00050)	(0.00050)	(0.00048)	(0.00048)
90/10 Income Ratio (std)	0.00989***	0.00708***	0.00653***	0.00432***	0.00401***	0.00367***	0.00366***	0.00271***
	(0.00063)	(0.00055)	(0.00052)	(0.00047)	(0.00047)	(0.00046)	(0.00050)	(0.00046)
Male=1			-0.00334***	0.00456***	-0.00176**	-0.00454***	0.00412***	-0.01618***
			(0.00087)	(0.00074)	(0.00074)	(0.00076)	(0.00049)	(0.00075)
Household Head=1			-0.02769***	-0.02641***	-0.02630***	-0.02531***	-0.01215***	-0.01917***
			(0.00049)	(0.00041)	(0.00039)	(0.00035)	(0.00040)	(0.00043)
Divorced/Separated=1			0.00556***	0.00356***	0.00378***	0.00212***	0.00035	-0.00311***
			(0.00055)	(0.00052)	(0.00050)	(0.00049)	(0.00040)	(0.00059)
Single=1			-0.00403***	-0.00666***	-0.00447***	-0.00335***	0.00195***	-0.00920***
			(0.00053)	(0.00049)	(0.00048)	(0.00045)	(0.00038)	(0.00057)
White=1			0.00087	0.01118***	0.01092***	0.00706***	0.00178***	0.00949***
			(0.00088)	(0.00157)	(0.00159)	(0.00145)	(0.00066)	(0.00244)
Immigrant=1			-0.00286***	-0.00182***	-0.00207***	-0.00210***	0.00016	-0.00061
			(0.00071)	(0.00067)	(0.00061)	(0.00053)	(0.00069)	(0.00043)
College=1			0.00685***	0.00247***	0.00146***	0.00238***	0.00290***	-0.00160***
			(0.00096)	(0.00055)	(0.00042)	(0.00033)	(0.00035)	(0.00044)
Capitalized House Value (std)			0.00075***	0.00035***	0.00029***	0.00024***	0.00043***	-0.00008
			(0.00011)	(0.00010)	(0.00010)	(0.00009)	(0.00008)	(0.00014)
Observations	22281334	22281334	22281334	22281334	22281334	22281334	11599405	5383435
Clusters	3075	3075	3075	3075	3075	3075	3075	3075
Mean Dep Var.	0.047	0.047	0.047	0.047	0.047	0.047	0.035	0.047
Age Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic Controls	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Main Occ. FE	No	No	No	Yes	No	No	No	No
Sub Occ. FE	No	No	No	No	No	No	No	No
Full Occ. FE	No	No	No	No	Yes	Yes	Yes	Yes
Sample	Full	Full	Full	Full	Full	Full	52 wks 40 hrs	Resp.

Notes: This table reports linear probability regression coefficients where the dependent variable is 1,0 for privacy (non-disclosure). It shows the coefficients on the demographic controls which are otherwise condensed for reporting purposes in Table I. Age controls are a linear and quadratic term in age. County controls are the 1940 population, manufacturing value-added, the share urban and the share of the population 25+ completing high school. Occupation fixed effects at the main (11) sub (22) and full (228) levels. Standard errors clustered by county in parentheses. *p<0.1, **p<0.05, ***p<0.01.

Table A3: Privacy Demands and Inequality: MLD Housing Inequality

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
New Deal Spending per Capita (std)	-0.00012	0.00011	0.00016	0.00006	0.00005	0.00003	-0.00066*	0.00033
	(0.00043)	(0.00040)	(0.00039)	(0.00039)	(0.00038)	(0.00038)	(0.00037)	(0.00029)
Republican Vote Share (std)	0.00239**	0.00043	0.00040	0.00090	0.00081	0.00084	0.00118*	0.00102
	(0.00101)	(0.00082)	(0.00079)	(0.00071)	(0.00068)	(0.00067)	(0.00063)	(0.00067)
Share Religious (std)	-0.00272***	-0.00012	-0.00032	-0.00072	-0.00078	-0.00081	-0.00082	-0.00049
	(0.00067)	(0.00064)	(0.00061)	(0.00058)	(0.00056)	(0.00055)	(0.00053)	(0.00050)
Housing Wealth Inequality (std)	0.00507***	0.00363***	0.00347***	0.00263***	0.00266***	0.00251***	0.00247***	0.00203***
	(0.00052)	(0.00047)	(0.00045)	(0.00042)	(0.00041)	(0.00040)	(0.00038)	(0.00039)
Observations	22281334	22281334	22281334	22281334	22281334	22281334	11599405	5383435
Clusters	3075	3075	3075	3075	3075	3075	3075	3075
Mean Dep Var.	0.047	0.047	0.047	0.047	0.047	0.047	0.035	0.047
Age Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County Controls	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic Controls	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Main Occ. FE	No	No	No	Yes	No	No	No	No
Sub Occ. FE	No	No	No	No	No	No	No	No
Full Occ. FE	No	No	No	No	Yes	Yes	Yes	Yes
Sample	Full	Full	Full	Full	Full	Full	52 wks 40 hrs	Resp.

Notes: This table reports linear probability regression coefficients where the dependent variable is 1,0 for privacy (non-disclosure). Age controls are a linear and quadratic term in age. County controls are the 1940 population, manufacturing value-added, the share urban and the share of the population 25+ completing high school. Demographic controls are indicators for gender, household head, marriage status (married, divorced/separated single), race, immigrant and college attendance and a continuous variable for capitalized house values. Occupation fixed effects at the main (11), sub (22) and full (228) levels. Standard errors clustered by county in parentheses. *p<0.1, **p<0.05, ***p<0.01.

Table A4: Privacy Demands and Inequality: Occupation 90/10 Income Inequality

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
New Deal Spending per Capita (std)	-0.00002	0.00010	0.00012	0.00003	0.00001	-0.00068*	0.00029
	(0.00037)	(0.00038)	(0.00038)	(0.00039)	(0.00038)	(0.00037)	(0.00030)
Republican Vote Share (std)	0.00206***	0.00098	0.00029	0.00081	0.00083	0.00121*	0.00115*
	(0.00074)	(0.00069)	(0.00071)	(0.00069)	(0.00068)	(0.00064)	(0.00068)
Share Religious (std)	-0.00201***	-0.00032	-0.00036	-0.00065	-0.00067	-0.00077	-0.00034
	(0.00052)	(0.00050)	(0.00050)	(0.00052)	(0.00051)	(0.00050)	(0.00049)
Occupation 90/10 Income Ratio (std)	0.02197***	0.02151***	0.02122***	0.01794***	0.01300***	0.01131***	0.01571***
	(0.00049)	(0.00048)	(0.00054)	(0.00037)	(0.00037)	(0.00044)	(0.00060)
Observations	22281334	22281334	22281334	22281334	22281334	11599405	5383435
Clusters	3075	3075	3075	3075	3075	3075	3075
Mean Dep Var.	0.047	0.047	0.047	0.047	0.047	0.035	0.047
Age Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County Controls	No	Yes	Yes	Yes	Yes	Yes	Yes
Demographic Controls	No	No	Yes	Yes	Yes	Yes	Yes
Main Occ. FE	No	No	No	Yes	No	No	No
Sub Occ. FE	No	No	No	No	Yes	Yes	Yes
Sample	Full	Full	Full	Full	Full	52 wks 40 hrs	Resp.

Notes: This table reports linear probability regression coefficients where the dependent variable is 1,0 for privacy (non-disclosure). The occupation 90/10 ratio is calculated using reported incomes in 228 occupational categories. Age controls are a linear and quadratic term in age. County controls are the 1940 population, manufacturing value-added, the share urban and the share of the population 25+ completing high school. Demographic controls are indicators for gender, household head, marriage status (married, divorced/separated single), race, immigrant and college attendance and a continuous variable for capitalized house values. Occupation fixed effects at the main (11) and sub (22) levels. Standard errors clustered by county in parentheses. *p<0.1, **p<0.05, ***p<0.01.

Table A5: Privacy Demands and Inequality: Occupation 90/10 Income Inequality and Churn

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
New Deal Spending per Capita (std)	-0.00010	0.00003	0.00007	0.00002	0.00001	-0.00068*	0.00029
	(0.00038)	(0.00038)	(0.00038)	(0.00039)	(0.00038)	(0.00037)	(0.00030)
Republican Vote Share (std)	0.00190**	0.00091	0.00038	0.00080	0.00083	0.00120*	0.00115*
	(0.00075)	(0.00070)	(0.00070)	(0.00069)	(0.00068)	(0.00064)	(0.00068)
Share Religious (std)	-0.00214***	-0.00036	-0.00046	-0.00068	-0.00068	-0.00078	-0.00034
	(0.00054)	(0.00052)	(0.00051)	(0.00052)	(0.00052)	(0.00050)	(0.00049)
Occupation 90/10 Income Ratio (std)	0.02616***	0.02581***	0.02570***	0.01955***	0.01406***	0.01228***	0.01606***
	(0.00062)	(0.00062)	(0.00066)	(0.00044)	(0.00050)	(0.00052)	(0.00075)
Observations	22281334	22281334	22281334	22281334	22281334	11599405	5383435
Clusters	3075	3075	3075	3075	3075	3075	3075
Mean Dep Var.	0.047	0.047	0.047	0.047	0.047	0.035	0.047
Age Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County Controls	No	Yes	Yes	Yes	Yes	Yes	Yes
Demographic Controls	No	No	Yes	Yes	Yes	Yes	Yes
Main Occ. FE	No	No	No	Yes	No	No	No
Sub Occ. FE	No	No	No	No	Yes	Yes	Yes
Sample	Full	Full	Full	Full	Full	52 wks 40 hrs	Resp.
Churn Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports linear probability regression coefficients where the dependent variable is 1,0 for privacy (non-disclosure). The occupation 90/10 ratio is calculated using reported incomes in 228 occupational categories. The specifications control for employment churn, measured as the median reported income for each of the in 228 occupational categories. Age controls are a linear and quadratic term in age. County controls are the 1940 population, manufacturing value-added, the share urban and the share of the population 25+ completing high school. Demographic controls are indicators for gender, household head, marriage status (married, divorced/separated single), race, immigrant and college attendance and a continuous variable for capitalized house values. Occupation fixed effects at the main (11) and sub (22) levels. Standard errors clustered by county in parentheses. *p<0.1, **p<0.05, ***p<0.01.

Table A6: Subject, Enumerator Descriptives

	Subjects		Enumerators	
	Mean	SD	Mean	SD
Privacy	0.045	0.208	0.542	0.498
Age	39.840	10.583	39.442	15.919
Male = 1	0.740	0.439	0.490	0.500
Household Head = 1	0.626	0.484	0.371	0.483
Married = 1	0.751	0.432	0.706	0.456
Divorced/Separated = 1	0.030	0.171	0.019	0.136
Single = 1	0.219	0.413	0.275	0.447
White = 1	0.938	0.242	0.942	0.233
Immigrant = 1	0.154	0.361	0.144	0.351
Years of Education	9.508	3.520	9.215	3.458
College = 1	0.151	0.358	0.118	0.322
Home Owner	0.583	0.493	0.557	0.497
House Value	4049.479	6782.993	3745.144	6484.312
Rental Value	63.636	382.305	61.989	380.324
Capitalized House Value	6132.909	35283.148	5794.083	34315.518
Weeks Worked	45.421	11.166	42.601	13.824
Hours Worked	43.240	11.341	43.775	13.275
Wage Income	1260.028	892.968	524.976	804.069
Non-Wage Income = 1	0.152	0.359	0.228	0.419

Notes: This table reports descriptive statistics for a sample of individuals we identify as a 'census taker' or a 'census enumerator' in the Census occupation string in the IPUMS data.

Table A7: Privacy and the Subject-Enumerator Housing Wealth Gap: Robustness

	(1)	(2)	(3)	(4)
Panel A: Main Estimates				
House Gap Ratio (log)	0.00232***	0.01148***	0.00265***	0.01278**
	(0.00064)	(0.00403)	(0.00081)	(0.00548)
Observations	276712	27328	163887	17320
Mean Dep Var.	0.042	0.043	0.042	0.043
Clusters	181252	18290	124971	13251
Panel B: Subject Housing Wealth > than Enumerator				
House Gap Ratio (log)	0.00367***	0.00219	0.00413**	-0.01300
	(0.00139)	(0.00914)	(0.00181)	(0.01280)
Observations	124992	11330	72846	6792
Mean Dep Var.	0.042	0.046	0.042	0.049
Clusters	79787	7445	54491	5191
Panel C: Subject Housing Wealth ≤ than Enumerator				
House Gap Ratio (log)	0.00199**	0.00175	0.00268***	0.00395
	(0.00086)	(0.00593)	(0.00102)	(0.00751)
Observations	151716	15992	91030	10515
Mean Dep Var.	0.043	0.042	0.041	0.038
Clusters	101461	10839	70469	8047
County FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Gender Matching	No	No	Yes	Yes
House Values	Capitalized	Reported	Capitalized	Reported
p-value (H_0: Panel B=Panel C)	0.3043	0.9673	0.4839	0.2485

Notes: This table reports linear probability regression coefficients where the dependent variable is 1,0 for privacy (non-disclosure). The right-hand side variables specify the log of the housing wealth gap between the subject and the enumerator. Controls include the log of the subjects own level of housing wealth, their age and years of education, gaps in age and education between subject and enumerator and mean log housing wealth, age and educational attainment in an enumeration district. Panel A shows the main estimates from columns 1, 3, 5 and 7 of Table II. Panel B shows estimates where the subjects' housing wealth is greater than the enumerator's housing wealth. Panel C shows estimates where the subjects' housing wealth is less than or equal to the enumerator's housing wealth. Standard errors clustered by household in parentheses. *p<0.1, **p<0.05, ***p<0.01.

Table A8: Newspaper List Descriptives

	Non-Newslist		Newslist	
	Mean	SD	Mean	SD
Newspaper Circulation (scaled)	0.000	0.000	1.030	1.336
Panel A: County-Level				
New Deal Spending per Capita	112.252	27.509	146.591	30.094
Republican Vote Share	37.352	17.122	38.435	14.382
Share Religious	48.959	11.932	51.532	11.967
Share Urban	86.958	8.726	90.585	10.971
Share Educated	27.829	6.240	28.095	6.873
County Population 1940 (Millions)	0.929	0.752	1.626	1.318
Manufacturing Value Added (Millions)	1069.696	1138.629	1235.662	1117.527
90/10 Income Ratio	5.725	1.413	5.970	1.279
Predicted 90/10 Income Ratio	3.632	0.876	3.874	0.890
Housing Wealth Inequality	0.602	0.111	0.754	0.211
Income Inequality	0.228	0.046	0.242	0.041
Panel B: Occupation-Level				
Occupation 90/10 Income Ratio	4.766	1.575	4.849	1.683
Panel C: Individual-Level				
Privacy	0.039	0.195	0.043	0.204
Age	39.770	10.367	40.040	10.523
Male = 1	0.765	0.424	0.726	0.446
Household Head = 1	0.646	0.478	0.630	0.483
Married = 1	0.775	0.418	0.748	0.434
Divorced/Separated = 1	0.027	0.163	0.027	0.161
Single = 1	0.198	0.398	0.226	0.418
White = 1	0.930	0.256	0.913	0.281
Immigrant = 1	0.206	0.405	0.192	0.394
Years of Education	9.273	3.461	9.421	3.466
College = 1	0.124	0.329	0.137	0.344
Home Owner	0.559	0.496	0.644	0.479
House Value	4634.610	8619.573	4992.775	8459.222
Rental Value	58.625	325.969	74.111	394.560
Capitalized House Value	5977.258	29833.206	7503.927	38369.461
Weeks Worked	45.043	11.146	46.118	10.878
Hours Worked	41.604	10.486	42.682	10.455
Wage Income	1367.998	912.949	1343.914	948.200
Non-Wage Income = 1	0.134	0.341	0.135	0.342

Notes: This table shows descriptive statistics by newspaper list locations. Newspaper list counties are those with a city in the top 50 cities by population size in the US in 1920 where lists of top earners were published between 1924 and 1925. Non-list counties are those with a city in the top 50 cities by population size in the US in 1920 in which lists were not published. Newspaper circulation numbers are scaled by 1920 county population.

Table A9: Newspaper List Balance Table

Variable	(1) Non-Newslist	(2) Newslist	(3) Difference (inc. FE)
New Deal Spending per Capita	112.252	146.591	25.311***
	(27.509)	(30.094)	(7.545)
Republican Vote Share	37.352	38.435	-1.732
	(17.122)	(14.382)	(2.619)
Share Religious	48.959	51.532	0.820
	(11.932)	(11.967)	(2.309)
Share Urban	86.958	90.585	1.584
	(8.726)	(10.971)	(3.019)
Share Educated	27.829	28.095	1.346
	(6.240)	(6.873)	(1.072)
County Population 1940 (Millions)	0.929	1.626	0.490*
	(0.752)	(1.318)	(0.283)
Manufacturing Value Added (Millions)	1,069.696	1,235.662	137.469
	(1,138.629)	(1,117.527)	(225.457)
90/10 Income Ratio	5.725	5.970	0.078
	(1.413)	(1.279)	(0.175)
Age	39.770	40.040	0.046
	(10.367)	(10.523)	(0.146)
Male = 1	0.765	0.726	-0.009*
	(0.424)	(0.446)	(0.005)
Household Head = 1	0.646	0.630	0.001
	(0.478)	(0.483)	(0.007)
Married = 1	0.775	0.748	-0.013
	(0.418)	(0.434)	(0.008)
Divorced/Separated = 1	0.027	0.027	0.000
	(0.163)	(0.161)	(0.001)
Single = 1	0.198	0.226	0.012
	(0.398)	(0.418)	(0.009)
White = 1	0.930	0.913	-0.003
	(0.256)	(0.281)	(0.011)
Immigrant = 1	0.206	0.192	0.009
	(0.405)	(0.394)	(0.021)
Years of Education	9.273	9.421	0.026
	(3.461)	(3.466)	(0.098)
College = 1	0.124	0.137	0.001
	(0.329)	(0.344)	(0.003)
Home Owner	0.559	0.644	0.053***
	(0.496)	(0.479)	(0.017)
House Value	4,634.610	4,992.775	290.069
	(8,619.573)	(8,459.222)	(229.196)
Rental Value	58.625	74.111	11.225*
	(325.969)	(394.560)	(6.526)
Capitalized House Value	5,977.258	7,503.927	1,053.878*
	(29,833.207)	(38,369.461)	(525.319)
Wage Income	1,367.998	1,343.914	-19.661
	(912.948)	(948.200)	(35.064)
Non-Wage Income = 1	0.134	0.135	0.005
	(0.341)	(0.342)	(0.005)
Observations	1,552,005	6,220,601	7,772,606

Notes: This table shows a balance table of descriptive statistics by newspaper list locations. Newspaper list counties are those with a city in the top 50 cities by population size in the US in 1920 where lists of top earners were published between 1924 and 1925. Non-list counties are those with a city in the top 50 cities by population size in the US in 1920 in which lists were not published. The difference in means test includes fixed effects for state and occupation.

Table A10: Newspaper List Descriptives by Tercile of Circulation

	Non-Newslist		Low		Medium		High	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Newspaper Circulation (scaled)	0.000	0.000	0.311	0.097	0.768	0.191	2.253	1.926
Panel A: County-Level								
New Deal Spending per Capita	112.252	27.509	144.812	29.128	145.038	28.105	150.560	32.858
Republican Vote Share	37.352	17.122	40.596	14.507	42.145	11.618	31.710	14.538
Share Religious	48.959	11.932	48.731	11.922	52.718	6.437	53.991	15.263
Share Urban	86.958	8.726	87.599	8.414	94.015	7.015	90.949	15.359
Share Educated	27.829	6.240	28.455	8.452	26.115	4.116	29.686	6.333
County Population 1940 (Millions)	0.929	0.752	1.550	0.890	2.417	1.765	0.901	0.605
Manufacturing Value Added (Millions)	1069.696	1138.629	933.928	419.038	1911.393	1475.328	928.910	1003.559
90/10 Income Ratio	5.725	1.413	6.015	1.180	5.469	0.788	6.433	1.593
Predicted 90/10 Income Ratio	3.632	0.876	3.760	0.640	3.637	0.734	4.270	1.153
Housing Wealth Inequality	0.602	0.111	0.803	0.244	0.719	0.157	0.726	0.202
Income Inequality	0.228	0.046	0.239	0.032	0.230	0.030	0.257	0.054
Panel B: Occupation-Level								
Occupation 90/10 Income Ratio	4.766	1.575	4.833	1.669	4.686	1.569	5.041	1.794
Panel C: Individual-Level								
Privacy	0.039	0.195	0.044	0.205	0.038	0.191	0.048	0.214
Age	39.770	10.367	40.060	10.561	40.033	10.512	40.021	10.485
Male = 1	0.765	0.424	0.743	0.437	0.731	0.443	0.699	0.459
Household Head = 1	0.646	0.478	0.646	0.478	0.630	0.483	0.609	0.488
Married = 1	0.775	0.418	0.765	0.424	0.751	0.432	0.721	0.448
Divorced/Separated = 1	0.027	0.163	0.029	0.168	0.023	0.150	0.027	0.162
Single = 1	0.198	0.398	0.206	0.405	0.226	0.418	0.252	0.434
White = 1	0.930	0.256	0.916	0.278	0.942	0.234	0.881	0.324
Immigrant = 1	0.206	0.405	0.153	0.360	0.208	0.406	0.229	0.420
Years of Education	9.273	3.461	9.459	3.449	9.326	3.388	9.469	3.567
College = 1	0.124	0.329	0.137	0.344	0.129	0.335	0.146	0.353
Home Owner	0.559	0.496	0.584	0.493	0.647	0.478	0.720	0.449
House Value	4634.610	8619.573	4799.540	7462.205	4990.369	6757.870	5375.401	11687.538
Rental Value	58.625	325.969	75.796	409.484	72.974	392.867	73.373	379.576
Capitalized House Value	5977.258	29833.206	7307.050	37907.840	7426.552	38169.651	7844.664	39172.718
Weeks Worked	45.043	11.146	45.691	11.162	46.396	10.609	46.390	10.759
Hours Worked	41.604	10.486	42.305	10.214	42.287	9.987	43.611	11.181
Wage Income	1367.998	912.949	1324.279	921.806	1383.911	949.246	1328.002	979.737
Non-Wage Income = 1	0.134	0.341	0.138	0.345	0.138	0.344	0.129	0.335

Notes: This table shows descriptive statistics by newspaper list locations split by tercile of newspaper circulation intensity (Low, Medium or High). Newspaper list counties are those with a city in the top 50 cities by population size in the US in 1920 where lists of top earners were published between 1924 and 1925. Non-list counties are those with a city in the top 50 cities by population size in the US in 1920 in which lists were not published. Newspaper circulation numbers are scaled by 1920 county population.

Table A11: The Super Rich and Non-Disclosure

Name	Net Income 1940	Weeks Worked 1939	Census Income	Census Non-Wage
John D. Rockefeller Jr	3,789,204	0	0	Yes
Clarence Dillon	129,019	52	0	Yes
Sid W. Richardson Jr	-264,498	52	5000	Yes
Reuben H. Fleet	291,013	52	5000	Yes
Richard K. Mellon	4,069,178	52	5000	Yes
Paul Mellon	5,074,832	52	0	Yes
Sarah M. Scaife	4,021,264	0	0	Yes
George L. Hartford	3,140,642	52	5000	Yes
Ailsa M. Bruce	2,074,634	0	0	Yes
Edsel B. Ford	3,483,889	52	0	No
Charles S. Chaplin	under 100,000	52	5000	Yes
Edgar Palmer	1,883,406	52	5000	Yes
Jeremiah Milbank Sr	211,628	52	5000	Yes
Katherine S. Milbank		0	0	Yes
Arthur V. Davis	2,054,765	0	0	Yes
John A. Hartford	2,819,498	52	5000	Yes
Minnie H. Reilly	3,029,144	Missing	Missing	Yes
Alfred P. Sloan Jr	2,169,154	45	5000	Yes
Irene J. Sloan		0	0	Yes
Lammot Du Pont	1,805,381	52	5000	Yes
Jessie B. Du Pont	1,785,279	Missing	Missing	Yes
Everette L. Degolyer	under 100,000	52	0	Yes
Nell V. Degolyer		0	0	No
Edward J. Noble	209,380	52	5000	Yes
William Du Pont Jr	1,458,160	52	5000	Yes
Alwin C. Ernst	1,303,815	52	5000	Yes
Charles S. Mott	1,623,670	Missing	Missing	Yes
Ethel M. Dorrance	2,152,426	Missing	Missing	Yes
James H. Cannon	339,754			
Mary S. Harkness	1,596,543	0	0	Yes
Henry Ford	2,933,531	52	0	Yes
Irenee Du Pont	1,702,128	Missing	Missing	Yes
Felix W. Zelcer	117,247			
Ignatius J. Miranda	122,757	52	5000	Yes
Alfred J. Miranda Jr	126,632	52	5000	Yes
Mary D. Biddle	1,310,094			
Gregory Ferend	under 100,000			
Robert S. Clark	1,101,090			
Allen G. Oliphant	under 100,000	52	5000	Yes
Anita M. Blaine	1,046,439			
Walter P. Murphy	950,436	35	5000	Yes
Mills Bennett	118,582	0	0	Yes
William R. Coe	1,244,800	0	0	Yes
Lammot D. Copeland	1,109,660	0	5000	Yes
Marie H. Robertson	1,357,449			
Robert R. M. Carpenter	1,125,524	52	5000	Yes
George H. Hartford II	1,432,434			
Evelyn Mendelssohn	1,138,971			
Garfield A. Wood	under 100,000	52	0	Yes
Helen H. Whitney	1,079,321	0	0	Yes
Josephine H. McIntosh	1,301,990	0	0	Yes
Marion D. Scott	1,025,286	0	0	Yes
Joan W. Payson	357,543	0	0	Yes
Alexis F. Du Pont	875,502	0	0	Yes
William T. Grant	1,246,739	52	0	Yes
Samuel H. Kress	1,657,698	52	5000	Yes
Cartter T. Lupton	645,054	52	5000	Yes
Ella B. Kearney	246,906	0	0	Yes
Henry B. Du Pont Jr	953,829	52	5000	Yes
Jessie W Donahue	1,260,734			
Rudolf J. Schaefer	672,878	52	5000	Yes
Lucia M. Schaefer		0	0	No
Frederick M. E. Schaefer	616,211	52	5000	Yes
Eugene Du Pont Jr	697,475	52	Missing	Yes
Harry P. Bingham	930,782	52	0	Yes
Alexander Smith	163,021			
Clifford Mooers	under 100,000			

Name	Net Income 1940	Weeks Worked 1939	Census Income	Census Non-Wage
Thomas M. O'Connor Jr	826,945	52	0	Yes
Marjorie P. Davies	851,741	Missing	Missing	Missing
Katherine D. Butterworth	584,471	0	0	Yes
Carl G. Swebilius	under 100,000	52	5000	Yes
Hulda Swebilius		Missing	Missing	No
Joseph E. Widener	1,475,478	0	0	Yes
Francis N. Bard	623,735	52	5000	Yes
Philip K. Wrigley	860,257	52	5000	Yes
Edith H. Harkness	763,455			
Donaldson Brown	918,183	52	5000	Yes
Barclay Douglas	126,001			
Josephine H. Douglas				
Edwin A. Link	236,207	52	5000	Yes
George A. Adam	397,370			
Josiah K. Lilly Sr	818,883	0	0	Yes
John D. Jackson	767,878	52	5000	Yes
Raymond Pitcairn	706,051	0	0	Yes
Edward S. Moore	371,272	52	0	Yes
Robert W. Woodruff	763,187	52	5000	Yes
Mahlon D. Thatcher Jr	269,247	50	5000	Yes
William T. Rawleigh	534,490	52	Missing	Yes
Hugh R. Sharp	1,007,876	52	5000	Yes
Sarah G. Kenan	361,062	0	0	Yes
Abby A. Rockefeller	616,440	0	0	Yes
Stanley R. McCormick	464,400	0	0	Yes
E. F. Stokes	381,877			
Doris D. Cromwell	690,665	0	0	Yes
Glenn L. Martin	533,852	52	5000	Yes
Edward H. Moore	under 100,000	Missing	Missing	Missing
Walker P. Inman	772,073	0	0	Yes
Alta R. Prentice	726,204	0	0	Yes
Cora T. Burnett	687,694	0	0	Yes
Sydney M. Shoenberg	773,189	52	5000	Yes
Eli Lilly	757,626	52	5000	Yes
Leonie B. Guggenheim	719,416	0	0	Yes
Miguel J. Ossorio	789,699	52	5000	Yes
William R. Kenan Jr	381,121	52	0	Yes

Notes: We matched the list of the super rich from Brandes (1983), as compiled by the U.S. Treasury, to the 1940 Census. Names in red text are unmatched. Some individuals might have been abroad at the time of enumeration. For example, Robert S. Clark, inheritor of the Singer Sewing Machine fortune, resided in Normandy, France for part of the year, following his marriage to a French actress in 1919. Net Income is from the list itself which also includes incomes for 1941. Sid W. Richardson, an oil industry magnate, is reported as having a negative net income in 1940 but a net income of $3,948,794 in 1941. The data on weeks worked, census income (top-coded at $5,000) and census non-wage income are from the 1940 Census. Highlighted cells show our definition of non-disclosure as zero or missing responses to the income question. The list often includes individuals and their spouses because certain states allowed income sharing within the household for federal tax purposes.

Figure A1: Newspaper Stories and Senate Hearing

An Official View of Proposed Questions for 1940 Census

Director Explains Bureau's Position on Controversial Inquiries, Taking Exception to The Times Articles Which Are Defended by the Writer

Census Snooping Stirs Senate Storm

INCOME QUERIES UNLAWFUL, SAYS FOE OF SCHEME

PRESIDENT CHARGES TOBEY ASKS PUBLIC TO BREAK THE LAW

Senate Group Requests Ban On Census Income Questions

First American Senator to Do So, Says Roosevelt Defending Census Query on Income

CENSUS AIDES BACK INCOME QUESTIONS AS WOMEN PROTEST

Director Reminds Senators No One Has Ever Been Jailed Yet for Failing to Answer

REVOLT PICTURED BY FOES

Prisons Will Overflow to Halt the 'Bureaucratic Snooping,' One Woman Asserts

PREDICT ERASURE OF INCOME QUERY IN CENSUS PRYING

Senators Aroused Over Display of Burocracy.

BY CHESLY MANLY.

Hopkins Revises Census Querying To Meet Protests on Income Data

Compromise Order Permits an Objector to Fill in Blank, Unsigned, and Seal It in Franked Envelope for Mailing

Special to THE NEW YORK TIMES.

Census: Are Questions on Income Legal?

1940 Census Will Go Into Economics

Employment Will Be Studied, Along With Income, Birth Rate

Uncle Sam Is Getting Much More Inquisitive

1940 Census Will Propound Some New Questions

1940 CENSUS

HEARINGS

BEFORE A

SUBCOMMITTEE OF THE

COMMITTEE ON COMMERCE

UNITED STATES SENATE

SEVENTY-SIXTH CONGRESS

THIRD SESSION

ON

S. Res. 231

A RESOLUTION FAVORING THE DELETION FROM THE SIXTEENTH CENSUS POPULATION SCHEDULE OF INQUIRIES NUMBERED 32 AND 33, RELATING TO COMPENSATION RECEIVED

FEBRUARY 28, 29 AND MARCH 1, 1940

Notes: Newspaper headlines surrounding the 1940 income questions in the census, and the front cover of the documentary evidence summarizing the 1940 Senate Hearing on the proposed deletion of these questions.

Figure A2: Publication of Tax Lists


THE NEW YORK TIMES, WEDNESDAY, SEPTEMBER 2, 1925.

Income Tax Payments in 1925 by Individuals and Corporations on 1924 Income

Income taxes which individuals and corporations are paying this year on their incomes for last year are given in the following list, drawn from the records in the offices of the Collectors of Internal Revenue in New York City and other districts. Today's list includes all names it was possible to assemble the first day and will be followed by others on subsequent days. Cents are omitted in stating the amount of tax paid:

NEW YORK CITY

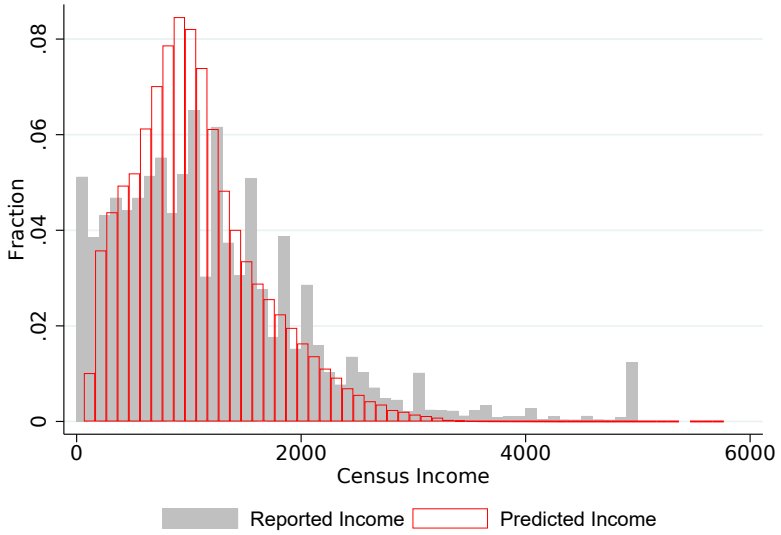
Second District

Manhattan, South of Twenty-third Street

INDIVIDUALS

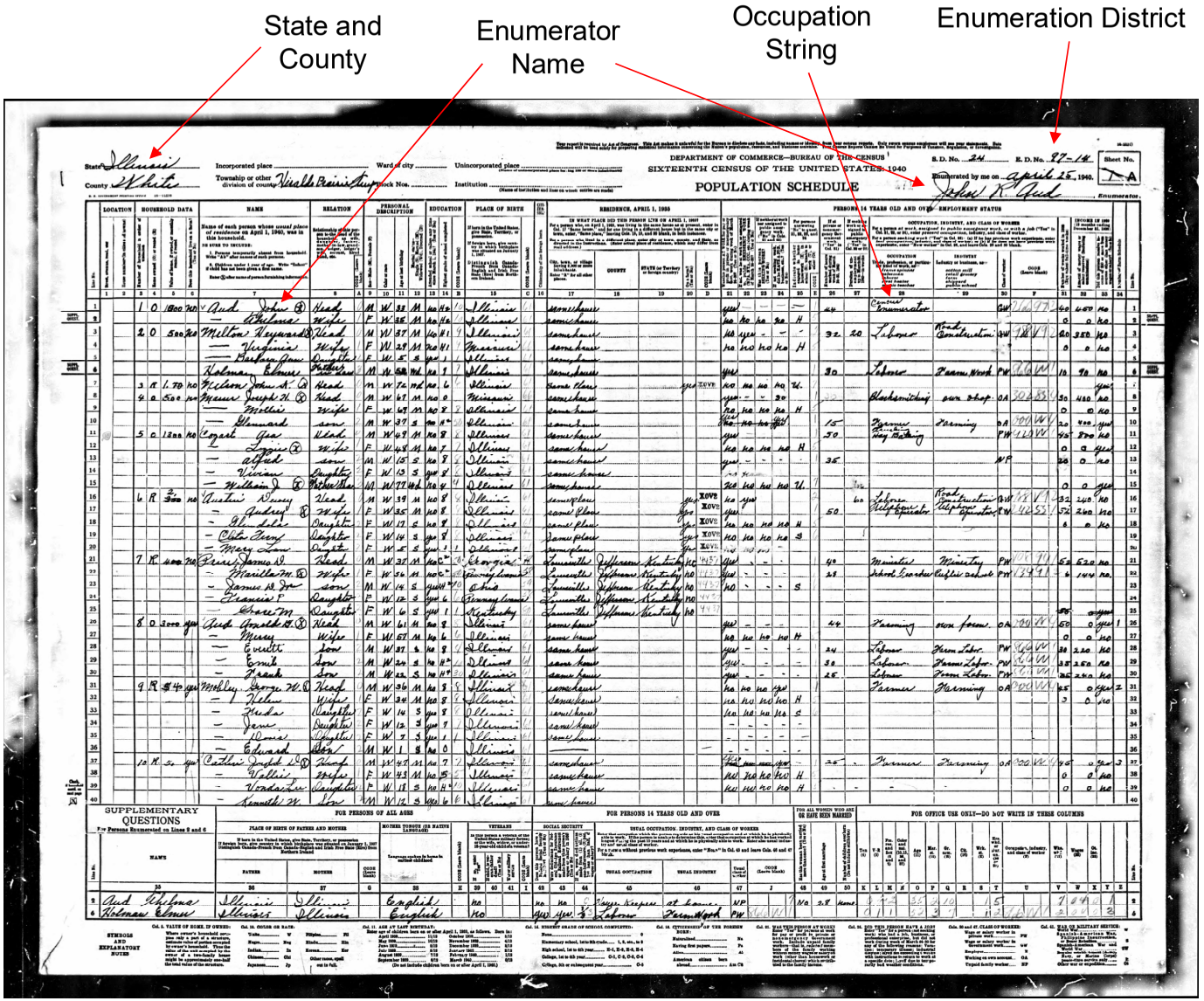
Notes: This shows an example of the tax lists published by the *New York Times*.

Figure A3: The Distribution of Reported and Predicted Income



Notes: This figure plots kernel densities of reported income from the 1940 census and predicted income based on demographic characteristics. We first regress log income for individuals with positive reported incomes on their age, a quadratic in age, their capitalized house value, years of education, indicators for gender, race, and fixed effects for state and occupation. We then use the model to predict income for all individuals in the dataset including those for whom reported income is missing or reported as zero.

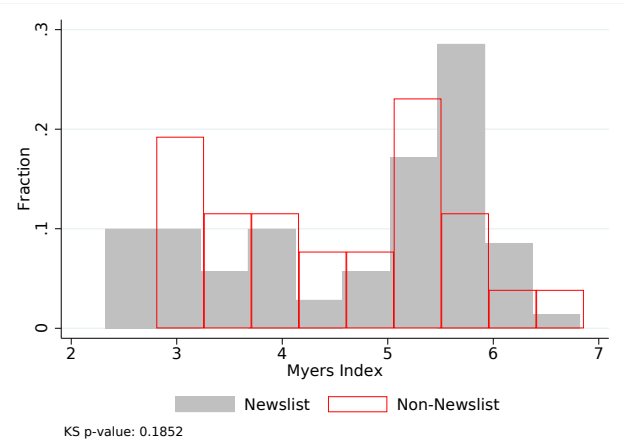
Figure A4: 1940 Census Population Schedule



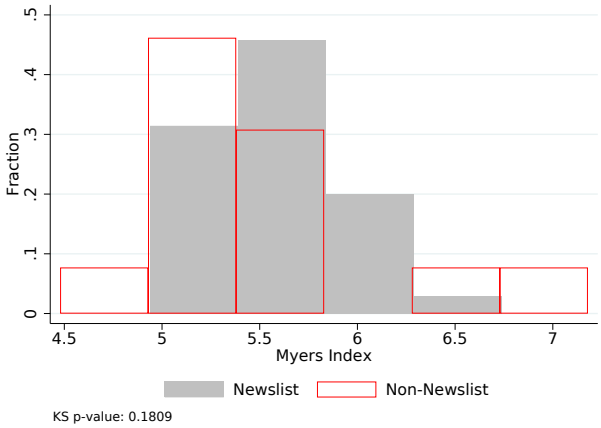
Notes: This shows an example of a Census population schedule illustrating the information we use to identify enumerators.

Figure A5: Myers Index: Newspaper List versus Non-Newspaper List Counties

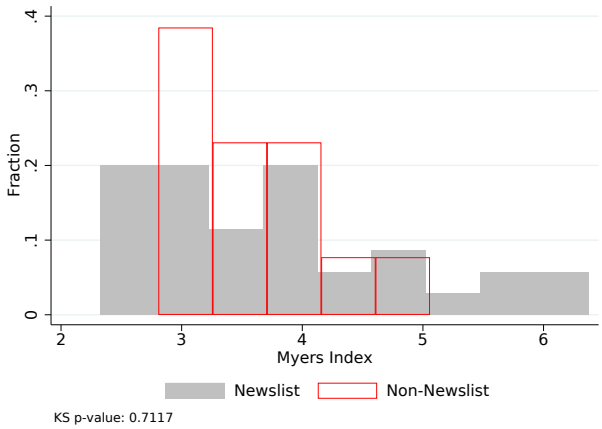
(a) Ages 25 to 65



(b) Ages 25 to 39



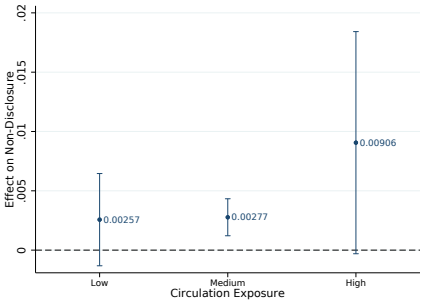
(c) Ages 40 to 65



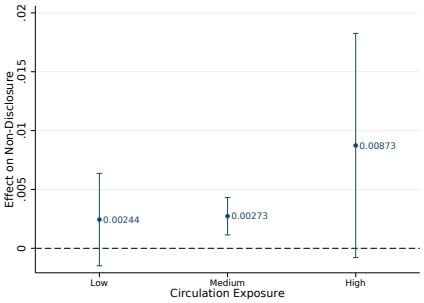
Notes: These figures show the distribution of Myers Index of age-heaping calculated for county-age cells in newspaper list and non-newspaper list counties. Age cells are 25 to 39 and 40 to 65 year olds. A Myers Index of 0 indicates no age-heaping (reported ages are evenly distributed across all final digits from 0 to 9) whereas a value of 90 indicates perfect heaping (every age is reported using only one final digit). Kolmogorov-Smirnov exact p-value reported under the null that the distributions are the same.

Figure A6: Replication of Figure 6 using only Stayers 1920-1940

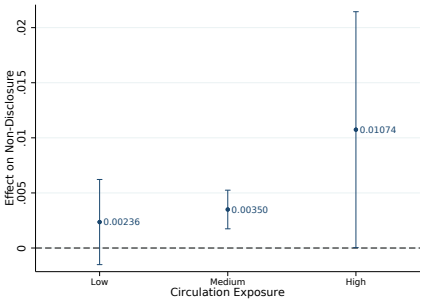
(a) Age Controls + State FE



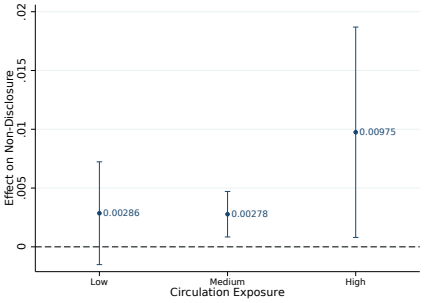
(b) + County Controls



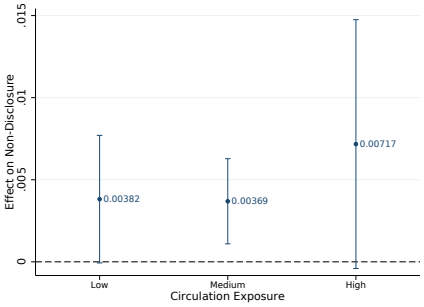
(c) + Dem. Controls



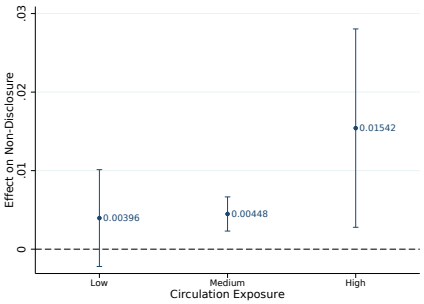
(d) + Occ. FE



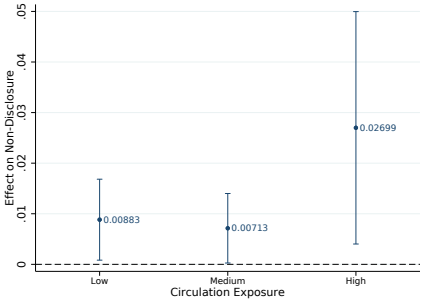
(e) only 52 wks 40 hr



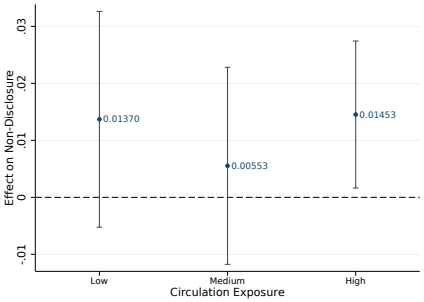
(f) only Respondent



(g) only College



(h) only College (TWI)



Notes: These figures replicate the results in Figure 6 using only individuals who remained in a newspaper list or non-newspaper list county across the 1920 to 1940 censuses. We use the male-only links provided by Abramitzky et al. (2020). A stayer is defined as an individual who remained in the same newspaper list county, moved to different newspaper list county in the same state, or a different newspaper list county in another state.