Links and legibility: Making sense of historical US Census automated linking methods *

Arkadev Ghosh[†], Sam Il Myoung Hwang[‡], Munir Squires[§]

November 3, 2022

Abstract

How does handwriting legibility affect the performance of algorithms that link individuals across census rounds? We propose a measure of legibility, which we implement at scale for the 1940 US Census, and find strikingly wide variation in enumeration-district-level legibility. Using boundary discontinuities in enumeration districts, we estimate the causal effect of low legibility on the quality of linked samples, measured by linkage rates and share of validated links. Our estimates imply that, across eight linking algorithms, perfect legibility would increase the linkage rate by 5 to 10 percentage points. Improvements in transcription could substantially increase the quality of linked samples.

*We thank Santiago Pérez and seminar participants at Midwest Economic Association conference, Western Economic Association virtual international conference, and UBC Econometrics lunch for their valuable comments. This research was undertaken thanks to funding from the Canada Excellence Research Chairs program awarded to Dr. Erik Snowberg in Data-Intensive Methods in Economics. Correspondence can be addressed to hwangii@mail.ubc.ca

[†]briq: Institute on Behavior and Inequality

[‡]University of British Columbia

[§]University of British Columbia

1 Introduction

Linking historical US Census records is crucial in the study of a range of economic outcomes such as migration and intergenerational mobility. However, data quality issues reduce the accuracy of these links, with unclear consequences for the resulting economic analyses. This paper explores one specific source of error: the difficulty in transcribing records with poor handwriting legibility. We show that this is a quantitatively important barrier to accurate linking using a novel measure of legibility that compares two independent transcriptions of the 1940 US Census schedule. We document wide variation in legibility and show that low legibility reduces the number and quality of links. We also find that the effect of legibility on these links depends on the choice of linking algorithm.

The first contribution of this paper is to propose and document a measure of legibility. For a given enumerator's handwritten entries, we use the share of recorded names where independent transcriptions by Ancestry.com and FamilySearch.org are identical.[1] Names written less legibly will be more likely to be entered differently by two individual transcribers. We find wide variation in this measure. For the lowest decile of enumerator-level legibility, fewer than half of the transcriptions agree, while for the highest decile the share is almost 90%. Figure 1 illustrates differences in legibility between enumerators: the handwriting in Figure 1a leads to fewer transcription errors than the one in Figure 1b.

Legibility is critical to link records across census rounds. We implement a variety of existing algorithms to link the 1930 and 1940 Census rounds and document how these methods perform as legibility changes.[2] The proportion of individuals who are linked across census rounds (the 'linkage rate') increases by up to two-thirds when moving from the bottom to the top decile of legibility. Further, the share of false positives declines with legibility: We find a decrease of up to 20% in the share of links that fail a validation test

[1] See Ruggles (2021) for a history of collaboration between these organizations and IPUMS, the flagship organizations for the distribution of historical and contemporary US census data.

[2] Given the scale of the data and the difficulty of determining which links are true, the literature has not yet reached a consensus on a preferred method of linking records (Abramitzky et al., 2021; Bailey et al., 2020a).

when moving from the bottom to the top decile of legibility.

One may be concerned that the quality of the linked samples and our measure of legibility are correlated for reasons that are not relevant to the underlying "true" legibility. For example, our measure of legibility may be low for enumeration districts with a high share of unusual names of foreign origin. To identify the causal effect of legibility on the quality of linked samples, we exploit discontinuities in legibility at the boundaries of enumeration districts. The discontinuity is created due to the following feature of the enumeration procedure for the US Federal Census since 1880: All households in an enumeration district are enumerated by a single census enumerator.[3] As a result, to the extent that different census enumerators have handwriting with varying degrees of legibility, our measure of legibility changes discontinuously at the boundaries of enumeration districts. Figure 2, a map of the city of Yonkers, New York, illustrates the variation in our legibility measures in neighboring enumeration districts. In fact, Figures 1a and 1b also embody our research design: these two census schedules contain information on households that live on opposite sides of the same street (South Highland Avenue, Los Angeles, California). This street happens to be on the boundary of two enumeration districts.

One challenge in implementing this research design is that one needs to know which enumeration districts share boundaries. The enumeration district maps are available for each US Federal Census in principle but they have not been digitized until recently. Shapefiles of enumeration districts in the 1940 Census have been made available for 43 cities by Logan and Zhang (2017).[4]

We find that legibility does have a causal effect on the quality of linked samples. We measure the quality of a linked sample in two ways: the *linkage rate* (i.e., the share of linkable population that is linked by a given algorithm) and the share of validated links

[3]This is at least in principle true (Jenkins (1985)). When Census Bureau officials draw the boundaries of the enumeration districts during the planning stage, they are drawn so that the enumeration of each enumeration district can be completed by a single enumerator in a reasonable time frame. See Section D for the description of how enumeration district boundaries are determined in the 1940 census.

[4]See section C for the list of these cities.

(*share validated* henceforth). The definition of a validated link may depend on the particular datasets being linked; we follow Bailey et al. (2020b) and define a link validated if the two records that are linked have matching parents' birth places. As shown in Bailey et al. (2020b), this is informative about the true-ness of a link.[5] We also present results using middle name initials as an alternative validation variable.

We find that moving from the 10th to the 90th percentile of the legibility distribution (from 53% to 88% legibility) increases the linkage rate by 16 to 41%, depending on the linking algorithm used. We also find that the share of links that are not validated drops by 12 to 23%. This implies that legibility has a large causal impact on linking performance, and the effect is much larger for some algorithms than for others. Generally, algorithms that use first and last with minimal cleaning are more sensitive to legibility than those that employ string comparators or phonetic codes.

Finally, we quantify the importance of legibility in determining the overall quality of the linked samples. We do this by estimating counterfactual values of the linkage rate and share validated as if legibility were perfect across our linkable sample. Although our boundary sample has the advantage of providing credible causal inference, it consists only of people living in large cities, and hence may not be representative of the broader population. Hence, in this section we use OLS estimates of the effects of legibility from our entire linkable sample. Using these coefficients, we estimate that perfect legibility would increase the linkage rate by 5 to 10 percentage points, depending on the linking algorithm used.[6] Observed linkage rates are approximately 75% to 88% of what they would be without legibility problems.

A concern with these estimates from the entire linkable sample is that, unlike the

[5]Another important measure of quality of linked samples is whether the linked sample is representative of the linkable population. However, we find that none of the linked samples used in this paper are representative. This is not surprising since most previous studies also find that their linked samples are not representative of their respective population.

[6]Strikingly, algorithms with higher linkage rates see somewhat *larger* increases in linkage rates from eliminating legibility errors. This suggests 'better' linking algorithms would not compensate for the problems caused by illegibility.

boundary sample, our results of the effects of legibility on the linkage rate may suffer from endogeneity. Our results suggest that this is unlikely. First, the coefficients on legibility using the entire linkable sample are very similar to the (more plausibly causal) ones from our boundary sample. Second, our results in this section include township fixed effects, which are likely to absorb much of the potentially problematic variation.[7] Third, we show that our results change very little if, as per Oster (2019), we allow for high levels of selection on unobservables.

In sum, we find that low handwriting legibility degrades transcriptions sufficiently to cause quantitatively important declines in linkage rates.

Literature review

Linked samples created from historical datasets have helped researchers answer important questions in a variety of topics, including immigration (e.g., Abramitzky et al. (2012)), internal migration (e.g., Collins and Wanamaker (2014)), intergenerational mobility (e.g., Long and Ferrie (2013)), and culture (e.g., Bazzi et al. (2020)). At the same time, considerable efforts have been made to evaluate the way linked samples are created, or in other words, evaluating the performances of linking algorithms. This is likely out of concern for the quality of datasets generated from historical sources. Relative to modern datasets, there are several quality issues with historical datasets, such as age heaping (e.g., A'Hearn et al. (2009)), reporting/enumeration errors (e.g., Ward (2021)), and transcription errors (e.g., Abramitzky et al. (2021)).

Evaluating linking algorithms would be a straightforward task if true links were observable. This is rarely the case in historical datasets.[8] In the absence of true links, some

[7] There are 25,630 townships in our data, far more than the number of counties (3,108).

[8] Two notable exceptions are the Swedish censuses of 1890 and 1900, and directories of citizens of the city of Zurich, studied in Wisselgren et al. (2014) and Favre (2019), respectively. Using these true links, they find that algorithms proposed by Ferrie (1996) or Abramitzky et al. (2012, 2014) exhibit linkage rates as high as 95 percent and type-1 error as low as 1.6 percent. However, both studies caution against applying their results to other datasets because the performance of linking algorithms may vary across datasets.

authors rely on manually constructed high-quality links (Bailey et al. (2020a), Bailey et al. (2020b), Abramitzky et al. (2021)) or crowd-sourced links available on genealogical websites (Price et al. (2021), Abramitzky et al. (2021) and Helgertz et al. (2022)).

However, these high-quality links that have been the basis of evaluation may also share some of the issues of linked samples that are being evaluated for quality. Eriksson (2017) highlights the fact that when researchers manually link a small sample to a population (because of the large costs involved in manual linking), the resulting links may contain false links. Their reasoning, which is also pointed out by Abramitzky et al. (2021), is as follows: suppose that a link is created between person A in the sample data and person B in the population data. The link between A and B may be false if there exists another person A' that is *not* in the sample data, but is actually the true match to person B. Eriksson (2017) uses Swedish census data to show that the sample-to-population linkage results in type-1 error rates of up to 24.4%, and that type-1 error rates increase as sampling rates fall.

In the case of crowd-sourced links available in genealogical websites, the links are likely to be true because the users of such websites, either genealogists or descendants of people being linked, may use additional information (e.g., birth/marriage certificates) beyond what is available in regular datasets such as the US census. However, the coverage of these user-provided links is typically not sufficient to evaluate all the links that are algorithmically created. For example, most links in Abramitzky et al. (2021) and Helgertz et al. (2022) (ranging from 63% to 95%), cannot be cross-checked with links on FamilySearch.org, a popular genealogical website. It is not clear whether these links are similar in quality to the links that can be cross-checked with links on genealogical websites, since the users of these websites may not be representative of the population.

Our approach complements existing work on the quality of linked samples. Instead of relying on high-quality linked samples to evaluate link quality, we attempt to investigate *why* the quality of linked samples is low. In other words, we attempt to measure the effect

of variation in a factor (the legibility of historical documents) that might affect the quality of linked samples. There are at least two strengths to our approach: we can quantify the role that this factor plays in degrading the quality of linked samples (see section 4); and if it is indeed identified as degrading the quality of linked samples, the research community may consider employing/developing technology that can address that issue (e.g., use advanced optical character recognition technology for digitizing the original handwritten census returns).

The remainder of this article proceeds as follows: Section 2 describes our datasets and legibility measure, and introduces the linking algorithms that we evaluate. Section 3 presents the causal effect of legibility on the quality of linked samples. In section 4, we quantify the effect of (il)legibility by simulating linkage rates and share validated under a counterfactual scenario where names are perfectly legible in all enumeration districts. Comparing simulated rates to actual rates then tells us the degree to which legibility degrades link quality. Finally, section 5 concludes. All figures and tables that contain main results are collected in a supplement file titled "Tables and Figures." Sections denoted with letters can be found in the supplemental appendices. Figures and tables for additional descriptives, analyses, and robustness checks are collected in Appendix H and prefixed with "A." Additional results, including a simple model of census data linkage and legibility, and heterogeneous effects of legibility on the quality of linked samples, can also be found in Appendices F and G, respectively.

2 Data and linking algorithms

2.1 Legibility

Our treatment variable is the legibility of census schedules. Our measure of legibility is calculated for each enumeration district (and hence each enumerator). It is defined as the share of records in the enumeration district whose transcriptions of given names

and surnames by FamilySearch.org and Ancestry.com are identical after standard cleaning procedures.[9] We obtain the transcriptions of Ancestry.com from IPUMS and that of FamilySearch.org from their website. We combine the two datasets and are successful in matching 93.1% of the 132,404,766 records in the 1940 census.[10] Among these, 71.9% have identical transcriptions. Conversely, in 28% of cases the transcriptions do not match. The mean of our enumerator-level legibility measure (across 150,156 enumeration districts) is 0.719 with a standard deviation of 0.147. For all of our analyses in the following sections, we drop enumeration districts that are too small (containing fewer than 50 people) or that do not have two transcriptions for a sufficiently large share of people (less than 90%). We also restrict analysis to White and Black males 8 years or older in 1940 (and hence could plausibly be linked to a 1930 individual). This leaves us with approximately 48 million observations. This sample will be frequently referred to as the "linkable population" in later analyses. Table A2 compares the mean of various observable characteristics of our linkable population with the overall population (all men), and with our sample of individuals who live along a relevant boundary.

We use a sample from our linkable population to identify the causal effect of legibility of census schedules on the quality of linked samples. We refer to this as the "boundary sample." This boundary sample is drawn from 43 cities for which the shapefiles of the 1940 census enumeration districts are available. It consists of individuals who live on either side of a street along the border of two neighboring enumeration districts. We obtain the exact street address of each household in the 43 selected cities from geographic reference files created by the Urban Transition Project (Logan and Zhang, 2018). We drop from our sample a) boundaries of enumeration districts that overlap with township or ward boundaries; and b) boundaries only one side of which is inhabited. We are left with

[9] To clean given names and surnames, we first change all the letters in each name to lower case; and then we remove all spaces and special characters in the names.

[10] We use National Archives and Records Administration microfilm roll number, image number within each roll, and the line number in the census schedule to merge the two transcriptions. In principle these variables uniquely identify each record in the census.

739,643 individuals living along 13,838 boundaries.

To identify the causal effect of legibility of census schedules on the quality of linked samples, we exploit discontinuities in our measure of legibility at the boundaries of enumeration districts. To test for balance across these boundaries, we present descriptive statistics for the "Less legible" and "More legible" sides of each boundary. The former group consists of all individuals who live on the side of the street where the measure of legibility is worse than the other side, and the latter group consists of the rest. Table 1 shows the mean of the legibility measure for each group: the difference in mean is 0.116, which is approximately 0.9 standard deviations of the legibility measure in the boundary sample (0.129). The difference is statistically significant at the 99.9% level. Note that in our empirical analysis, we never use this binary distinction between more and less legible sides of a boundary. Instead, we always rely on our full continuous measure of legibility, which uses the actual gap in legibility across these boundaries. Please refer to section 3 for details of the specification of the model.

To identify the effect of legibility on the quality of linked samples, it is necessary that both observable and unobservable characteristics of individuals on either side of the boundaries are balanced. We find that they are. Tables 2 and A3 compare the mean of observable characteristics between the two groups, the "Less legible" versus the "More legible" group. For most characteristics, the difference in means is not statistically significant. When it is statistically significant, the standardized difference is below 0.1, the threshold recommended in Austin (2009) to determine balance.

As for unobservable characteristics, they may not be balanced if individuals sort across enumeration district boundaries based on these characteristics. Although the existence of such sorting cannot be ruled out, it is unlikely because enumeration district boundaries are drawn only for the purpose of census enumeration and do not serve any other functions that may induce sorting. Since these boundaries may overlap with other meaningful boundaries (such as county, township or ward boundaries), we drop such overlapping

boundaries from our sample.[11] We refer interested readers to section D for a description of how enumeration district boundaries were determined for the 1940 census.

2.2 Linkage rates and the share of validated links

We measure the quality of linked samples using two outcomes: linkage rates and the share of validated links. The linkage rate is defined as the share of a given sample that is linked. The other measure of quality, the share of validated links ("share validated" henceforth), is defined as the share of linked records that is validated by an auxiliary variable, or "validation variable", that was not used as a linking variable. Our baseline validation variable is parents' birth places. That is, a link is validated if father and mother's birth places recorded in the 1930 and 1940 censuses match. We use this variable for validation because Bailey et al. (2020b) provide evidence, using ground-truth links, that links validated with parents' birth places are more likely to be true links than those that are not.[12] See section A for further discussion about using parents' birth places as a validation variable.

We also check the robustness of our results regarding share validated with an alternative validation variable: middle name initials.[13] That is, a link is validated if middle name initials match across the two censuses. Although US Federal Census questionnaires do not specifically ask about middle names, many people report them. In our main sample, 24% of records are associated with a middle name initial.[14] The validation status of a link using middle name initials is strongly correlated with that from using parents' birth places. The share of links whose validation status remains unchanged between the two validation methods is about 75% (see Table A1), suggesting that middle name initials are

[11] We are unable to drop school district boundaries from sample. This is because, as far as we are aware, the digitized maps of school districts are not available for 1940.

[12] One can infer from Table 2 of Bailey et al. (2020b) that 91% of validated links in their sample are true, whereas 82% of invalidated links are true. To determine whether a link is true, they had a group of experienced genealogical linkers at the Family History and Technology Lab at Brigham Young University verify the links.

[13] We thank an anonymous referee who suggested this alternative.

[14] We extract the middle name from the name field in the census using `abeclean` command in `STATA`.

indeed effective for validation.

2.3 Linking algorithms

This paper uses the samples created by the following eight linking algorithms.[15] In the interest of space, we henceforth use abbreviations for these algorithms:

1. Abramitzky, Boustan and Eriksson algorithm with exact names as a linking variable ("ABE-exact")

2. Apply ABE-exact algorithm and remove links with names that overlap with anyone else's in the ± 2 year band ("ABE-exact5")

3. ABE algorithm with NYSIIS-standardized names as a linking variable ("ABE-NYSIIS")

4. Apply ABE-NYSIIS algorithm and remove links with NYSIIS-standardized names that overlap with anyone elses' in the ± 2 year band ("ABE-NYSIIS5")

5. ABE algorithm where names are considered to match if they are within 0.1 Jaro-Winkler distance ("ABE-JW")[16]

6. Apply ABE-JW algorithm and remove links with names that are within 0.1 Jaro-Winkler distance from anyone else's in the ± 2 year band ("ABE-JW5")

7. Machine learning algorithm (Feigenbaum (2016), "ML")

8. The algorithm that creates the Multigenerational Longitudinal Panel dataset (Helgertz et al. (2022), "MLP")

[15] We use linked samples that are already available online where possible. Linked samples created by the ABE-exact, ABE-exact5, ABE-NYSIIS, ABE-NYSIIS5, and MLP algorithms are available online. The first four are available at `https://censuslinkingproject.org/`; the latter is available at `https://usa.ipums.org/usa/mlp_downloads.shtml`. We implement the ABE-JW, ABE-JW5, and ML algorithms using our own STATA codes (available upon request).

[16] See Winkler (1990) for a detailed description of how Jaro-Winkler distance between two strings is computed.

We refer interested readers to a review article by Abramitzky et al. (2021) (algorithms 1 to 6) or the references cited above (algorithms 7 and 8) for precise descriptions of each algorithm. We note similarities and differences between these algorithms that are important for interpreting our results in the following sections. The first seven algorithms are similar in the sense that the linking variables they use are individual characteristics that are either time invariant or evolve in a predictable way, such as given names and surnames, race,[17] birth place, and age. The main difference amongst these seven algorithms lies in how they use these linking variables (especially names) to declare links and whether they remove links with names that are common. The MLP algorithm, on the other hand, represents a departure from the other algorithms in that it expands the set of linking variables from only immutable characteristics of an individual to time-varying information about the individual (e.g., place of residence) and also to information about members in the same household (parents, spouse, siblings, etc.). This feature of the MLP algorithm likely leads to over-representation of households whose members do not change across censuses.

None of the algorithms generate linked samples that are representative of the population.[18] Table A4 compares various observable characteristics of our linkable population to each of the linked samples. All linked samples under-represent Blacks and over-represent Midwesterners relative to our linkable population. The share of Blacks in the linked samples is approximately 45% to 78% of that in the population, whereas the share of Midwesterners in the linked samples is approximately 116% to 123% of that in the population. On the other hand, for most of the other characteristics, differences in means between the population and the linked samples are statistically significant, but moderate in magnitude.

[17] There exists some evidence that recorded race for the same individual changes over time: for example, Dahis et al. (2019) argue that at least 1.4 percent of Blacks have passed for whites at some point between 1850 and 1940 censuses.

[18] As far as we know, none of the linked samples used in previous studies were representative of their respective populations.

3 The causal effect of legibility

We estimate the following model with our boundary sample to obtain estimates of the causal effect of census schedule legibility on the quality of linked samples:

$$q_i = \beta \ell_{e_i} + X_i' \gamma + \delta_{b_i} + \epsilon_i, \tag{1}$$

where the dependent variable q_i is one of our quality measures. For linkage rates, q_i is equal to 1 if person i is linked, and 0 otherwise. For share validated, q_i is equal to 1 if the link for person i is validated with his/her parents' birth places, and 0 otherwise. ℓ_{e_i} is our legibility measure for person i's enumeration district (denoted with e_i). X_i is a vector of observable characteristics of person i as well as a constant (see notes under Table 3 for the list of covariates), and δ_{b_i} is the boundary fixed effect for person i. Lastly, ϵ_i captures the effect of unobservable factors on the quality of linked samples. We estimate model (1) separately for each algorithm. We use the entire boundary sample when the outcome is linked/not linked, whereas we use only linked records with non-missing values of the validation variable when the outcome is validated/invalidated.

3.1 The effect of legibility on linkage rates

We find that the legibility of census schedules affect linkage rates for each of the eight algorithms. Figure 3 illustrates our finding. This figure, created with the boundary sample, plots mean linkage rates for each linking algorithm against different levels of legibility. There is a positive relationship between legibility and the linkage rate for each algorithm. In addition, Figure A2, which focuses on two of the algorithms (ABE-exact and ABE-JW), suggests that the effects of legibility on linkage rates are heterogeneous across these two algorithms. Specifically, the ABE-exact algorithm appears to yield higher linkage rates than ABE-JW when legibility is above 0.57 (denoted with the vertical line), yet the latter yields higher linkage rates when legibility is below 0.57. We observe a similar pattern

in Figure A3, where we focus on linkage rate-legibility profiles of three conservative algorithms (ABE-exact5, ABE-JW5, and ABE-NYSIIS5). The linkage rate of ABE-exact5 appears to be larger than that of the other two algorithms when legibility is greater than 0.61, but it falls below ABE-JW5 at 0.61 and below ABE-NYSIIS5 at 0.59 (denoted with vertical lines).

Our estimates of model (1) are consistent with the impression that Figures 3, A2, and A3 provide. Table 3 presents coefficient estimates on legibility in model (1), estimated separately for each linking algorithm. The coefficients are statistically significant for all algorithms at the 99.9% level, though the magnitude of the coefficients vary. According to our estimates, a one standard deviation increase in the legibility measure (approximately an increase of 0.129 in the boundary sample) raises linkage rates by 2.3 to 4.5 percentage points, or 6.1% to 16.1% of the mean linkage rate, depending on the linking algorithm.

Our estimates in Table 3 also indicate that the coefficient on legibility for the ABE-exact algorithm is larger than that for all of the other seven algorithms. To formally test the equality of coefficients on legibility between the ABE-exact and each of the other seven algorithms, we estimate the following model with "stacked" boundary samples:

$$\begin{aligned} q_i = &(\beta_1 + \mathbb{1}\{\text{non-ABE-exact algorithm}\}_i \cdot \beta_2) \cdot \ell_{e_i} \\ &+ \delta_{b_i} + \epsilon_i \end{aligned} \tag{2}$$

where $\mathbb{1}\{\text{non-ABE-exact-algorithm}\}_i$ is an indicator that is equal to 1 if record i is associated with one of the seven non-ABE-exact linking algorithms, and 0 otherwise. To estimate model (2), we stack two copies of boundary samples, one of which is associated with the ABE-exact algorithm and the other with one of the other seven algorithms. The null hypothesis we test is $\beta_2 = 0$, i.e., the coefficient on legibility associated with a given non-ABE-exact algorithm is equal to that associated with the ABE algorithm.

The estimates of β_1 and β_2 are presented in Table A5. $\widehat{\beta}_2$'s are negative for each of

the other seven algorithms, meaning that the coefficient on legibility is smaller for these algorithms than for the ABE-exact algorithm. We reject each of the null (equality of the coefficients) hypotheses at the 99.9% level. The magnitude of the differences (i.e., $|\widehat{\beta}_2|$) are quite large as they are (roughly) close to a half of the magnitude of $\widehat{\beta}_1$, which is the effect of legibility on the linkage rate for ABE-exact. Similarly, we also find that the coefficient on legibility for the ABE-exact5 algorithm is larger than that of each of the other seven algorithms (see Table A6 for the estimate of model (2), with dummy variables replaced appropriately).

The sensitivity of linkage rates to the legibility measure for ABE-exact and ABE-exact5 algorithms is likely because they link two records only if the names on the records exactly match. Our finding suggests that this linking strategy may yield linkage rates that are higher than algorithms that employ string comparators (e.g., Jaro-Winkler distance) or phonetic codes (e.g., NYSIIS) when the source documents are sufficiently legible. However, as the legibility of the source documents deteriorate, ABE-exact and ABE-exact5 likely yield lower linkage rates because poor legibility might induce incorrect transcription of names. Our results suggest that string comparators or phonetic codes can mitigate the effect of poor legibility on linkage rates.

Our results survive two sets of robustness checks. In our first set of robustness checks, we use two alternative measures of legibility. The first alternative is constructed in the same way as our baseline measure, except that we do not remove spaces in between letters in transcribed names. Recall that in constructing our baseline legibility measure, we remove all the spaces between letters in the names before comparing the two transcriptions. In this robustness check, we test if our results are sensitive to this particular name-cleaning procedure. For the second alternative legibility measure, we require that the two transcriptions of a person's name be sufficiently different from each other to be counted as not identical. Specifically, we require the Jaro-Winkler distance between two names to be greater than the 75th percentile value in the population, for it to be counted as not

identical. The 75th percentile is equal to 0.044, which is close to 0 because most names are transcribed identically in the two transcriptions (and therefore they have a Jaro-Winkler distance of zero).

Tables A7 and A8 present estimates of model (1) using each of the two alternative legibility measures, while Tables A9 to A12 also present estimates from model (2) for each of these two alternatives. Our baseline conclusion remains robust.

In the second set of checks, we show that our results are robust to varying the extent to which legibility changes across enumeration district boundaries. We re-run our analysis restricting our sample to boundaries where legibility changes by at least a certain threshold value. The thresholds are the 5th, 10th, 25th, and 50th percentiles of the distribution of differences in legibility (where the unit of observation of the distribution is a boundary). Table A13 presents estimates from model (1) for each threshold (as well as our baseline results for a reference). The coefficients on legibility are statistically significant and stable for all linking methods across all thresholds. Our conclusion that the linkage rate of ABE-exact(5) is more sensitive to legibility than other linking algorithms remains robust to this check as well (see Tables A14 and A15).

3.2 The effect of legibility on the share of validated links

Turning to the second measure of quality, we find that legibility also affects the share of validated links positively. Figure 4 presents the share validated-legibility profile associated with each linking algorithm. The share validated is increasing in legibility across all linked samples. We confirm this with estimates of model (1): the coefficient on legibility is positive and statistically significant at the 95% level for ABE-exact and at the 99.9% level for all the other algorithms, indicating that increases in legibility raise share validated (Table 4). To the extent that share validated is negatively correlated with type-1 error rates, our findings suggest that increases in legibility reduce type-1 error rates.

The magnitude of the effect of legibility on share validated is modest: our estimates im-

ply that a one standard deviation increase in legibility (approximately 0.129) raises share validated by 0.7% to 2.8% of the mean of share validated. Equivalently, it reduces share *invalidated* by 0.6 to 2.2 percentage points, or 4.3% to 10.1% of the mean of share invalidated, depending on the algorithm. However, we would like to emphasize that the moderate effect of legibility on share (in)validated does not necessarily mean that it has a moderate effect on type-1 error rates. A validated link can still be false.

We also find that effects of legibility on share validated are heterogeneous across algorithms. Specifically, share validated for the MLP algorithm is *less* sensitive to legibility relative to other algorithms. This pattern is visible in Figure 4: the share validated-legibility profile of the MLP algorithm appears to be flatter than others. We formally test this by estimating model (2), replacing the dependent variable and the dummy variables accordingly. Table A16 presents estimates of model (2) for the MLP and each of the other algorithms. The estimates indeed indicate that the coefficient on legibility for the MLP algorithm is positive (i.e., $\widehat{\beta}_1 > 0$) and smaller than those for other algorithms (i.e., $\widehat{\beta}_2 > 0$), and the differences are statistically significant at the 95% level or higher except when compared with ABE-exact5. Our estimates imply that share validated for the MLP algorithm is not only larger than those of other algorithms but the difference also grows as legibility deteriorates.[19]

Our results in this section are robust to various different checks. The first two are the same ones that we conducted for linkage rates in subsection 3.1, i.e., using alternative measures of legibility and restricting the sample to boundaries where differences in our baseline measure of legibility (across the boundary) is sufficiently large. Tables A21, A22, and A23 present results for these robustness checks and show that the effects of legibility on share validated are statistically significant and their magnitudes are similar to the baseline estimates.

[19]These results are robust to alternative measure of legibility, alternative validation variable, restriction of sample to boundaries across which the legibility measure is sufficiently different. See Tables A17, A18, A19 and A20.

The third robustness check uses an alternative validation variable, i.e., the initial of one's middle name, as discussed in subsection 2.2. Table A24 presents these results. The estimates of the effect of legibility are statistically significant at the 99.9% level for all of the algorithms. However, similar to the baseline estimates, once again we see that the magnitude of the effects of legibility on share validated are modest. The estimates imply a one standard deviation increase in legibility (0.129) increases share validated by 1.4%-4.9% of the mean (of share validated).

Finally, for the last robustness check, we weight each observation with the inverse of the predicted probability of being linked and having non-missing values for the baseline validation variable (i.e., parents' birth places). This check is necessary to account for the fact that our estimation uses only linked records when the dependent variable in model (1) is a validation variable. To the extent that the data linkage selects on observable/unobservable characteristics, it is possible that these characteristics of individuals are not balanced across boundaries *conditional on being linked*. Tables A25 through A32 compare the means of observable characteristics between more and less legible sides of the boundaries, similarly to Table 2, but conditional on being linked under each of the algorithms. None of the differences are large enough such that the standardized differences are greater than the threshold of 0.1.[20] Our weighting procedure, which corrects for the potential imbalance in observables created by linkage, yields estimates that are similar to our baseline estimates, although three of the eight estimates are no longer statistically significant (see Table A33). See Appendix E for details about our weighting procedure.

4 Quantifying the effect of (il)legibility

Having established that legibility positively affects linkage rates and share validated, in this section, we quantify the role that it plays in determining the overall quality of linked

[20]Lack of evidence for unbalanced observables does not necessarily mean that linkage will not cause selection on unobservables.

samples. To do so, we simulate linkage rates and share validated under a counterfactual scenario where our measure of legibility is equal to 1 in all enumeration districts. We then compare the simulated quality of linked samples to the actual quality observed in the data. The ratio of observed quality to simulated quality (or the difference between the two) is our estimate of the degree to which legibility degrades the quality of linked samples between the 1930 and 1940 censuses.

For our simulation, we estimate the effect of legibility on our quality measures using the linkable population, rather than the boundary sample. We do so because the boundary sample consists of those living in large cities, and hence is not representative of the population (see Table A2 for a comparison between our boundary sample and the linkable population). The effect of legibility in the population may therefore not be the same as in the boundary sample, whereby applying estimates from this sample to the population may lead to systematic biases.

In practice, we estimate the following model using the linkable population, with enumeration districts as the unit of observation:

$$\overline{q}_e = \beta \ell_e + \overline{X}_e' \gamma + \delta_{f(e)} + \epsilon_e, \tag{3}$$

where $\overline{q}_e$ and $\overline{X}$ are enumeration-district level averages of the corresponding variables in our baseline model (i.e., q_i and X_i in model (1), respectively), and $\delta_{f(e)}$ is a fixed effect for an administrative division that includes enumeration district e (e.g., townships, counties, or states). Finally, ϵ_e captures a random shock to the quality of linked samples in enumeration district e. This specification is similar to our baseline model (1) aggregated at the enumeration district level. The only difference is that, in model (3), we can only control for administrative divisions that are larger than an enumeration district, since our legibility measure varies at the enumeration district level. We therefore use township fixed effects in model (3), because township is the smallest unit of administrative division that is available in our dataset for the entire population.

The OLS estimates of β in model (3) for linkage rates and share validated are presented in the first column (labeled "Unadjusted") of Tables A34 and A35, respectively. We find that the signs of the coefficients on legibility are the same as those obtained from the boundary sample. That is, an increase in legibility raises both the linkage rates and share validated. The magnitude of the coefficients are also similar to our baseline estimates from model (1) obtained with the boundary sample and boundary fixed effects (see "Baseline" column in the same table).

However, we are less confident that the OLS estimate of β in model (3) represents the causal effect of legibility compared to our baseline estimates using the boundary sample. Township fixed effects may not capture all of the unobservable factors that are correlated with our measure of legibility, leading to omitted variable bias. To address this issue, we adjust for the potential bias in β by adopting the method proposed by Oster (2019) – which is devised to address selection on unobservables in linear models.

One of the key parameters in Oster (2019) is the coefficient of proportionality, denoted by δ.[21] This parameter measures the strength of selection on unobservables relative to selection on observables. Its value may vary across contexts, but assuming that δ is positive, Oster (2019) suggests that a reasonable upper bound for δ is 1 (i.e., the strength of selection on unobservables is the same as observables). Then she shows that the true treatment effect is between the unadjusted estimate (i.e., the OLS estimate) and the estimate of β under the assumption that $\delta = 1$.

We adopt this approach proposed by Oster (2019) and estimate the effect of legibility on the quality of linked samples under the assumption that $\delta = 1$. We also do it for $\delta = -1$, since we are unable to verify that δ is positive in our setting. Note that δ should to be zero

[21] There is another parameter, what Oster (2019) denotes as $R_{\max}$, which corresponds to the R squared in a hypothetical regression where all variables, observable or unobservable, that are in the true model for explaining variation in the dependent variable are included as controls. As opposed to Altonji et al. (2005), Oster (2019) allows $R_{\max}$ to be less than 1 in cases where there is measurement errors in the dependent variable. In our context, however, there is no measurement error in the dependent variable because our dependent variables are constructed with the information that is already in our dataset. Therefore, we set $R_{\max}$ equal to 1 in our implementation of the bias-adjustment procedure suggested by Oster (2019).

for our OLS estimates to be interpreted as causal. Our estimates do not vary substantially with δ. The columns labeled $\delta = 1$ and $\delta = -1$ in Tables A34 and A35 presents these results. For linkage rates, the effect of legibility is statistically significant for all linked samples regardless of the assumption about δ, and the same is true for share validated except for the three conservative algorithms (ABE-exact5, ABE-JW5, and ABE-NYSIIS5) and the MLP algorithm when $\delta = 1$. Below, we also check the sensitivity of our eventual simulation results against alternative assumptions about δ (i.e., $\delta = 1$ or $\delta = -1$).[22]

Under the counterfactual scenario where legibility is equal to 1 for all enumeration districts, we simulate the quality of linked samples for each enumeration district as follows:

$$\min \left\{ \overline{q}_e + \hat{\beta} \cdot (1 - \ell_e), 1 \right\} \tag{4}$$

where $\overline{q}_e$ and ℓ_e respectively denote the quality measure (linkage rates or share validated) observed in the data and legibility measure for enumeration district e, and $\hat{\beta}$ denotes the estimate of β in model (3). We truncate the simulated quality at 1 because that is the upper bound on these measures by construction. Note that the upper bound of 1 is rarely binding for the simulated linkage rates, because linkage rates are far lower than 1 for most enumeration districts. However, it is binding for some enumeration districts when it comes to simulating share validated – the share of enumeration districts for which the simulated share validated had to be truncated at 1 is at most 10% (see Table A36 for de-

[22] As a further robustness check, we estimate δ using our boundary sample, and then estimate β in model (3) that corresponds to the estimated δ. To estimate δ with the boundary sample, we first estimate model (1) with the boundary sample, but replacing the boundary fixed effects with township fixed effects. Then, using the formula in Proposition 3 in Oster (2019), we obtain the δ that corresponds to our baseline estimates, i.e., β in model (1) that is obtained with the boundary sample and *boundary fixed effects*. Essentially, assuming that the estimate of β obtained with boundary fixed effects is the true effect, we estimate δ that corresponds to that true effect in a model with township fixed effects. Then we estimate model (3) using the linkable population, with δ set at the estimated value. The estimates of β and δ as well as the simulation results can be found in Tables A34, A35, 5 and 6. We find that the estimated δ is close to zero, which suggests that the extent of selection on unobservables is limited, at least for the boundary sample. As a result, the estimates of β and the simulation results are similar to the one obtained without any bias-adjustments. Note that this exercise is valid if the degree of selection on unobservables relative to selection on observables is the same across the two samples. While it is difficult to test this assumption, it is comforting to find that this assumption does not make a large difference in the simulation results.

tails).

Table 5 presents the linkage rates observed in our data alongside the simulated rates for different values of δ. Using our unadjusted estimates of β, we find that observed linkage rates are approximately 74.8% to 88% of what they would be if legibility were equal to 1 in all enumeration districts. In terms of differences in levels, illegibility accounts for a 5.2 to 9.6 percentage point reduction in the linkage rate, depending on the linking algorithm used. Notably, we also do not observe smaller benefits from increasing legibility for linking algorithms with higher (baseline) linkage rates. This suggests that algorithm improvements that increase linkage rates do not compensate for low legibility.

Alternative assumptions about the value of δ do not make a considerable difference in the simulated linkage rates. Under the assumption of $\delta = 1$ ($\delta = -1$), observed linkage rates are between 76% (73%) and 88% (88%) percent of what they would have been if legibility was equal to 1 in all enumeration districts. In terms of differences in levels, illegibility accounts for a 4.4 (5.6) to 8.8 (10.1) percentage point decrease in linkage rates under the assumption that $\delta = 1$ ($\delta = -1$) (results available upon request). The lack of sensitivity of our simulation results for different values of δ is perhaps because legibility can only improve so much, and the estimated coefficients on legibility in model (3) are between 0.2 to 0.3 regardless of the value of δ. Therefore, small differences in β due to different assumptions about δ cannot make a large difference in simulated linkage rates.

While illegibility has a large effect on linkage rates, it plays a modest role in reducing share validated. Table 6 presents simulation results for share validated using parents' birth places as the validation variable. Using our unadjusted estimates of β, observed share validated is between 96.1% and 99.2% of what it would have been with perfect legibility. In terms of levels, differences between simulated and observed share validated range from 0.7 to 3.3 percentage points (results available upon request).

These results are robust under the alternative assumption that $\delta = -1$, both quantitatively and qualitatively (see column labeled "$\delta = -1$" in Table 6). However, when we

set $\delta = 1$, the differences between simulated and observed share validated are not statistically significant for three linked samples (ABE-Exact5, ABE-JW5, and ABE-NYSIIS5) – the observed share validated for these linked samples are within the 95% confidence interval of simulated share validated. These results are expected since these three linking algorithms impose more stringent conditions to declare a link than other linking algorithms (see section 2.3 or references therein for the description of the algorithm). Therefore, it is possible that there is little room for improvement in share validated for these (conservative) algorithms even if legibility improves significantly. Our results are robust to using middle name initials as the alternative validation variable (see Table A37).[23]

5 Conclusion

In this paper, we document the importance of handwriting legibility on the performance of popular linking algorithms in a case study of US Census rounds 1930-1940. We find that low enumerator handwriting legibility is associated with lower linkage rates and a smaller share of validated links, and that this holds across linking algorithms. We show that this relationship is causal by focusing on boundaries of enumeration districts.

Legibility problems are a quantitatively important source of linkage errors. We estimate that 5 to 10 percentage points more links would be found across these census rounds if all enumerators had perfect legibility. This improvement is just as large for algorithms with higher linkage rates, suggesting that improvements in linking algorithms would not substitute for improvements in legibility. Automated transcription methods may be a promising source of improvement in link quality for historical sources, if they can be programmed to outperform humans.

[23]One caveat to the results using middle name initials is that the share of enumeration districts for which the simulated share validated is greater than one (hence truncated at one) is larger than when we used parents' birth places as the validation variable. See Table A38.

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez (2021) "Automated linking of historical data," *Journal of Economic Literature*, 59 (3), 865–918.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2012) "Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration," *American Economic Review*, 102 (5), 1832–56.

——— (2014) "A nation of immigrants: Assimilation and economic outcomes in the age of mass migration," *Journal of Political Economy*, 122 (3), 467–506.

A'Hearn, Brian, Jörg Baten, and Dorothee Crayen (2009) "Quantifying quantitative literacy: Age heaping and the history of human capital," *The Journal of Economic History*, 783–808.

Altonji, Joseph G, Todd E Elder, and Christopher R Taber (2005) "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of political economy*, 113 (1), 151–184.

Austin, Peter C (2009) "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples," *Statistics in medicine*, 28 (25), 3083–3107.

Bailey, Martha, Connor Cole, and Catherine Massey (2020b) "Simple strategies for improving inference with linked data: a case study of the 1850–1930 IPUMS linked representative historical samples," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53 (2), 80–93.

Bailey, Martha J, Connor Cole, Morgan Henderson, and Catherine Massey (2020a) "How

well do automated linking methods perform? lessons from us historical data," *Journal of Economic Literature*, 58 (4), 997–1044.

Bazzi, Samuel, Martin Fiszbein, and Mesay Gebresilasse (2020) "Frontier culture: The roots and persistence of "rugged individualism" in the United States," *Econometrica*, 88 (6), 2329–2368.

Collins, William J and Marianne H Wanamaker (2014) "Selection and economic gains in the great migration of African Americans: new evidence from linked census data," *American Economic Journal: Applied Economics*, 6 (1), 220–52.

Dahis, Ricardo, Emily Nix, and Nancy Qian (2019) "Choosing racial identity in the united states, 1880-1940,"Technical report, National Bureau of Economic Research.

Eriksson, Björn (2017) "False Positves and Faulty Estimates: Linked Census Data and Bias to Estimates of Social Mobility," in *Presentation at The Systematic Linking of Historical Records Workshop, University of Guelph, Centre for Economic Demography-Lund University*.

Favre, Giacomin (2019) "Bias in social mobility estimates with historical data: evidence from Swiss microdata," *University of Zurich, Department of Economics, Working Paper* (329).

Feigenbaum, James J (2016) "Automated census record linking: A machine learning approach."

Ferrie, Joseph P (1996) "A new sample of males linked from the public use microdata sample of the 1850 US federal census of population to the 1860 US federal census manuscript schedules," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29 (4), 141–156.

Helgertz, Jonas, Joseph Price, Jacob Wellington, Kelly J Thompson, Steven Ruggles, and Catherine A Fitch (2022) "A new strategy for linking US historical censuses: A case

study for the IPUMS multigenerational longitudinal panel," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 55 (1), 12–29.

Jenkins, Robert M (1985) *Procedural History of the 1940 Census of Population and Housing*, 1940: University of Wisconsin Press.

Logan, John R and Weiwei Zhang (2017) "Developing GIS Maps for US Cities in 1930 and 1940," *The Routledge Handbook of Spatial History. Routledge: UK*.

——— (2018) "Developing GIS Maps for US Cities in 1930 and 1940," in *The Routledge Companion to Spatial History*, 229–249: Routledge.

Long, Jason and Joseph Ferrie (2013) "Intergenerational occupational mobility in Great Britain and the United States since 1850," *American Economic Review*, 103 (4), 1109–37.

Oster, Emily (2019) "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, 37 (2), 187–204.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley (2021) "Combining family history and machine learning to link historical records: The Census Tree data set," *Explorations in Economic History*, 80, 101391.

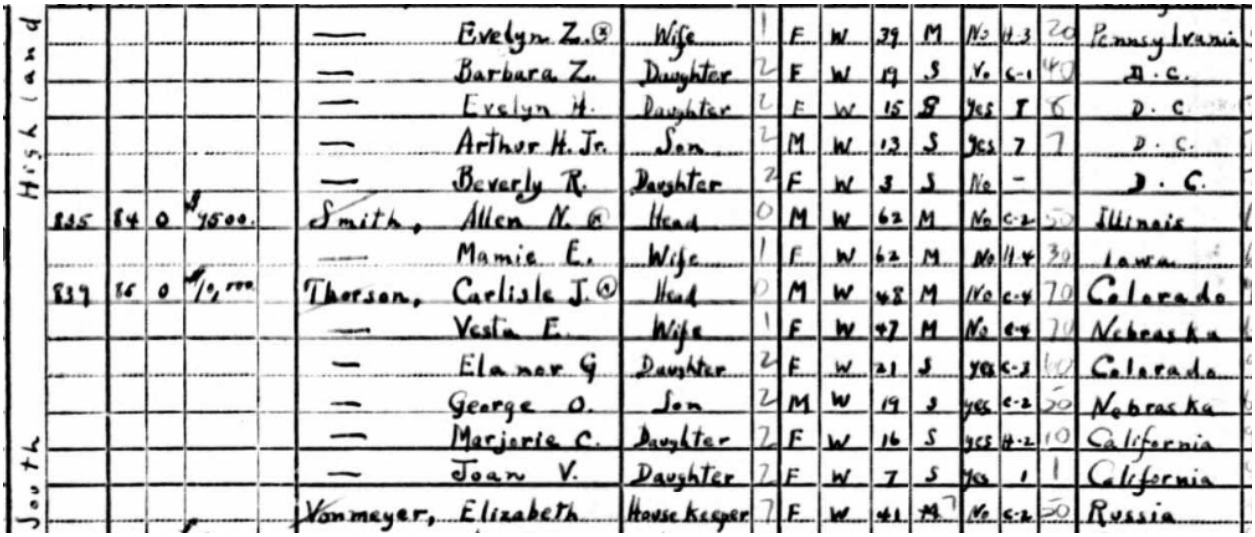
Ruggles, Steven (2021) "Collaborations between IPUMS and Genealogical Organizations."

Ward, Zachary (2021) "Intergenerational mobility in American history: Accounting for race and measurement error,"Technical report, National Bureau of Economic Research.

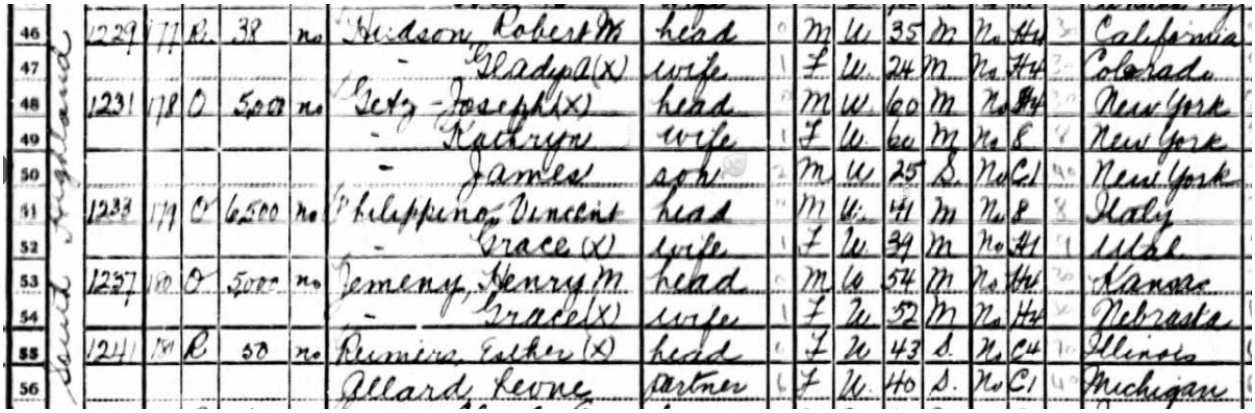
Winkler, William E (1990) "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.."

Wisselgren, Maria J, Sören Edvinsson, Mats Berggren, and Maria Larsson (2014) "Testing methods of record linkage on Swedish censuses," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 47 (3), 138–151.

1 Figures referenced in the main text



(a) South Highland Ave., ED 60-863, Los Angeles, CA



(b) South Highland Ave., ED 60-869, Los Angeles, CA

Figure 1: Differences in handwriting legibility across enumerators

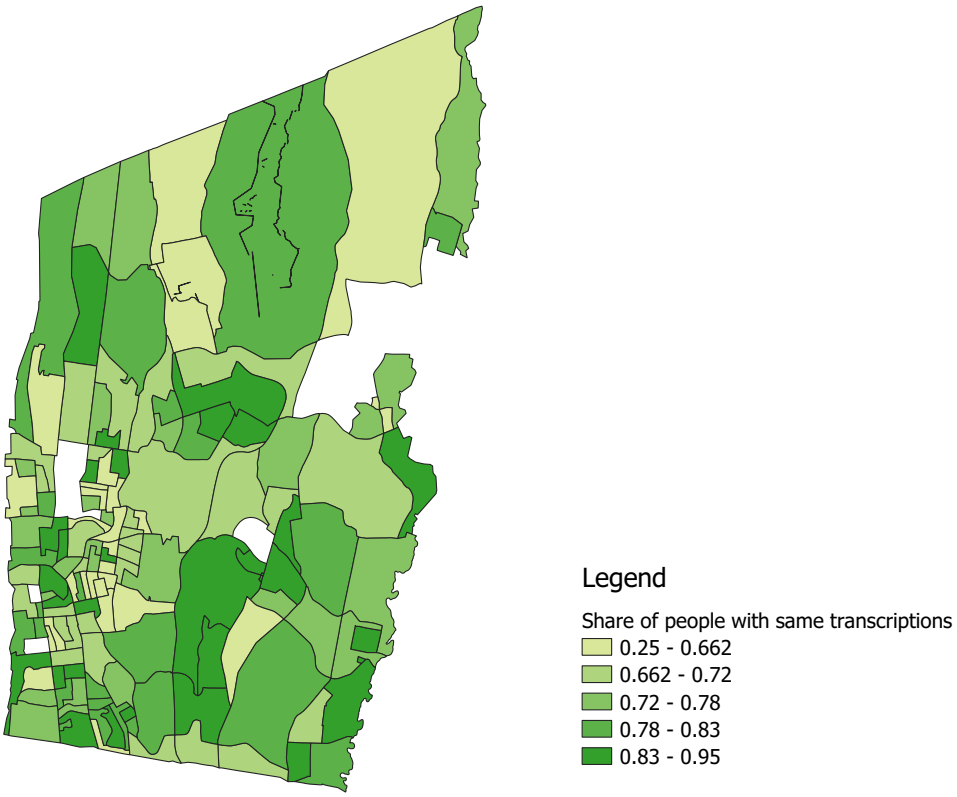


Figure 2: Legibility by enumeration districts in Yonkers, NY

Note: "Share of people with same transcriptions" is equal to the number of people in each enumeration district for whom two transcriptions (one by Ancestry.com and another by FamilySearch.org) agree, divided by the number of people in that enumeration district for whom the two transcriptions exist. A few enumeration districts are missing because none of the people in those districts have two transcriptions of their names.

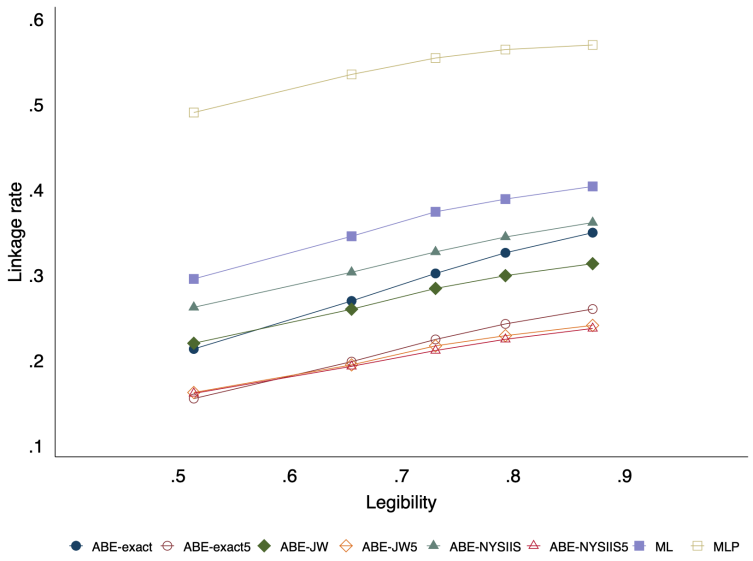


Figure 3: The effect of legibility on linkage rates

Note: We use the boundary sample to create this figure (N=739,634). For each linking algorithm (see legend), the symbol corresponds to the linkage rate for that particular bin. The bins are of equal size. Confidence intervals are omitted for clarity of presentation.

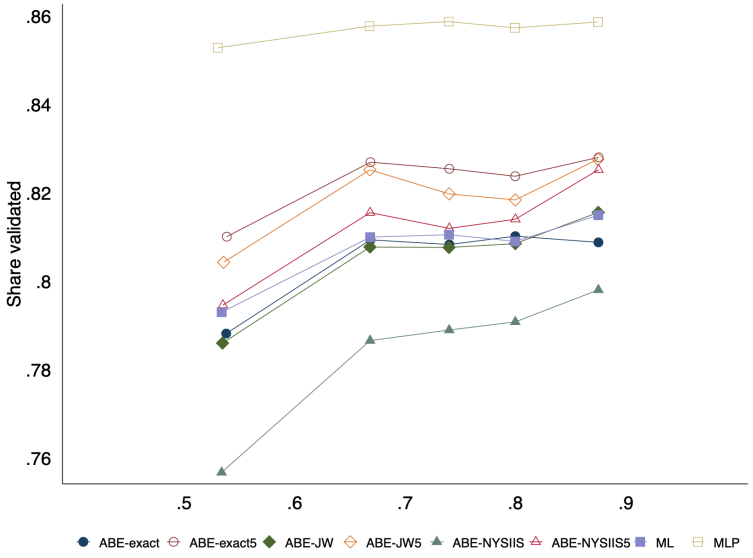


Figure 4: The effect of legibility on share validated

Note: This figure is created only with linked observations, because the validation variable (whether linked observations have matching parents' birth places) is missing for unlinked observations. The number of observations for each algorithm is as follows: 92,225 (ABE-exact, 12.5% of the boundary sample), 69,480 (ABE-exact5, 9.4%), 90,016 (ABE-JW, 12.2%), 67,003 (ABE-JW5, 9.1%), 100,096 (ABE-NYSIIS, 13.5%), 64,975 (ABE-NYSIIS5, 8.8%), 116,947 (ML, 15.811%), 201,261 (MLP, 27.211%).

2 Tables referenced in the main text

Table 1: Legibility and the number of people on each side of the boundaries

Variable	Less legible	More legible	Diff.	Std. diff.
Legibility	0.656	0.773	0.116***	0.712
# people on the boundary	26.007	27.442	1.435***	0.033
Observations	13,838	13,838	27,676	

Note: The unit of observations for this table is boundary $\times$ enumeration district. The standardized difference for a continuous covariate x is equal to:

$$\frac{\overline{x}_{\text{more legible}} - \overline{x}_{\text{less legible}}}{\sqrt{\frac{s^2_{\text{more legible}} + s^2_{\text{less legible}}}{2}}}$$

where $\overline{x}_{\text{more legible}}$ and $\overline{x}_{\text{less legible}}$ are sample means of the covariate x for more legible group and less legible group, respectively, and $s^2_{\text{more legible}}$ and $s^2_{\text{less legible}}$ are sample variances for the two groups, respectively. The standardized difference for a binary covariate is defined analogously (see Austin (2009) for a reference). $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Balance of observables at the boundary of enumeration districts

Variable	Less legible	More legible	Diff.	Std. diff.
Own a house	0.411	0.412	0.001	0.001
Multi-gen. household	0.813	0.814	0.000	0.000
Age	35.055	35.127	0.072*	0.003
Black	0.090	0.090	0.000	0.000
Married	0.528	0.526	-0.002	-0.002
In school	0.210	0.209	-0.001	-0.001
# characters in full name	12.820	12.887	0.067***	0.021
Main provider of info.	0.150	0.155	0.005***	0.009
Absent when enumerated	0.005	0.006	0.001***	0.009
Head of household	0.500	0.499	-0.002	-0.002
Years of schooling	10.244	10.232	-0.012	-0.002
Foreign born mother	0.367	0.368	0.001	0.001
Foreign born father	0.417	0.417	-0.000	-0.000
Non-institutional residence	0.989	0.988	-0.001***	-0.008
Employed	0.628	0.629	0.001	0.002
In labor force	0.799	0.800	0.001	0.003
# weeks worked	44.544	44.350	-0.194***	-0.011
# hours per week worked	43.576	43.592	0.015	0.001
Labor income	891.885	889.591	-2.294	-0.002
Nonlabor income $\geq$ \$50	0.204	0.213	0.008***	0.014
Size of household	4.723	4.736	0.013**	0.004
Observations	359,890	379,744	739,634	

Note: This table presents the means of various observable characteristics of those who live on either side of an enumeration district boundary — the side very legibility is relatively lower ('Less legible") and the side where it is relatively higher ("More legible"). The unit of observations in this table is a person. The sample used to create this table is the boundary sample, i.e., those who live on streets that serve as the border between two neighboring enumeration districts. The variables are obtained from the 1940 census. Note that the share of observations for which "Foreign born mother" and "Foreign born father" are non-missing is 39% and 34% in the boundary sample. These are greater than 5%, which is the sampling rate for the census "long form" in which parents' birth places are surveyed. This is because IPUMS assigned the birth places of parents to those who did not take the "long form" survey, but were living with one of their parents at the time of the 1940 census. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: The effect of legibility on linkage rates

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.269***	0.190***	0.180***	0.348***	0.230***	0.239***	0.263***	0.255***
	(0.00696)	(0.00691)	(0.00689)	(0.00774)	(0.00765)	(0.00806)	(0.00824)	(0.00836)
BDFE	Y	Y	Y	Y	Y	Y	Y	Y
N	725790	725790	725790	725790	725790	725790	725790	725790
adj. R^2	0.051	0.052	0.047	0.052	0.055	0.044	0.067	0.142

Note: We use the boundary sample to estimate model (1). The dependent variable is linked/not linked (1/0). Controls include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided most or all of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); one's birth state; and highest grade of schooling. All controls are directly obtained from 1940 census. BDFE refers to boundary fixed effects. Robust standard errors are in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: The effect of legibility on share validated

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.0673*	0.116***	0.124***	0.0841***	0.136***	0.169***	0.116***	0.0480***
	(0.0277)	(0.0278)	(0.0284)	(0.0238)	(0.0234)	(0.0227)	(0.0197)	(0.0125)
BDFE	Y	Y	Y	Y	Y	Y	Y	Y
N	68529	66057	64040	90885	88703	98660	115204	198087
adj. R^2	0.103	0.102	0.096	0.090	0.088	0.084	0.080	0.085

Note: We use only the linked observations in the boundary sample to estimate model (1). The dependent variable is validated/not validated (1/0), where the validation variable is parents' birth places. Covariates include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided the most (or all) of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); birth state; highest grade of schooling. All controls are directly obtained from 1940 census. BDFE refers to boundary fixed effects. Robust standard errors are in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5: Comparison between the observed and simulated linkage rates of the linked samples

Link. alg.	Obs. qual.	Simulated quality			
		Unadjusted	$\delta = 1$	$\delta = -1$	Estimated δ
ABE-exact5	0.218	0.293	0.287	0.297	0.293
			(.285, .287)	(.296, .297)	(.292, .293)
ABE-JW5	0.218	0.270	0.262	0.275	0.270
			(.26, .262)	(.274, .275)	(.269, .27)
ABE-NYSIIS5	0.215	0.267	0.262	0.271	0.268
			(.26, .262)	(.27, .272)	(.266, .268)
ABE-exact	0.296	0.392	0.384	0.397	0.392
			(.383, .386)	(.396, .398)	(.391, .393)
ABE-JW	0.284	0.346	0.337	0.352	0.347
			(.335, .337)	(.351, .352)	(.345, .347)
ABE-NYSIIS	0.327	0.399	0.392	0.403	0.399
			(.39, .392)	(.402, .404)	(.397, .399)
ML	0.373	0.444	0.435	0.449	0.444
			(.433, .436)	(.448, .45)	(.442, .444)
MLP	0.534	0.609	0.607	0.611	0.609
			(.604, .607)	(.61, .612)	(.608, .61)

Note: This table presents the observed linkage rates for each linked sample (in column labeled "Obs. qual.") as well as the simulated linkage rates under different assumptions about δ. The column labeled "Unadjusted" contains the simulated linkage rate obtained with the OLS estimate of β in model (3) (note that its confidence intervals are omitted, given how tight the confidence intervals for the OLS estimates of β is). The columns labeled "$\delta = 1$" and "$\delta = -1$" contain simulated linkage rates under the corresponding assumption about δ, and includes the 95 percent bootstrap confidence for the simulated linkage rates. The column labeled "Estimated δ" contains simulated linkage rates when δ is set at the estimated value. See footnote 22 in the main text for details about how δ is estimated, and see Table A34 for the estimates of δ.

Table 6: Comparison between the observed and simulated share validated of the linked samples

Link. alg.	Obs. qual.	Simulated quality			
		Unadjusted	$\delta = 1$	$\delta = -1$	Estimated δ
ABE-exact5	0.841	0.860	0.842	0.860	0.860
			(.785, .891)	(.858, .861)	(.859, .862)
ABE-JW5	0.840	0.861	0.844	0.861	0.862
			(.819, .86)	(.859, .862)	(.86, .863)
ABE-NYSIIS5	0.835	0.860	0.818	0.860	0.860
			(.756, .858)	(.858, .861)	(.859, .862)
ABE-exact	0.823	0.845	0.853	0.844	0.845
			(.843, .866)	(.843, .845)	(.843, .847)
ABE-JW	0.826	0.852	0.855	0.852	0.852
			(.845, .876)	(.851, .853)	(.85, .854)
ABE-NYSIIS	0.809	0.842	0.847	0.841	0.842
			(.842, .857)	(.84, .842)	(.841, .844)
ML	0.827	0.848	0.853	0.848	0.848
			(.845, .864)	(.846, .848)	(.847, .85)
MLP	0.875	0.882	0.878	0.882	0.882
			(.861, .889)	(.881, .883)	(.881, .883)

Note: This table presents the observed share validated for each linked sample (in column labeled "Obs. qual.") as well as the simulated share validated under different assumptions about δ. The validation variable in this table is parents' birth places. The column labeled "Unadjusted" contains the simulated share validated obtained with the OLS estimate of β in model (3) (note that its confidence intervals are omitted, given how tight the confidence intervals for the OLS estimates of β is). The columns labeled "$\delta = 1$" and "$\delta = -1$" contain simulated share validated values under the corresponding assumption about δ, and includes the 95% bootstrap confidence for the simulated estimates. The column labeled "Estimated δ" contains share validated when δ is set at the estimated value. See footnote 22 in the main text for details about how δ is estimated, and see Table A35 for the estimates of δ.

Supplemental Appendices (not for publication)

October 18, 2022

A Parents' birth places as a validation variable

There is one challenge with using parents' birth places as a validation variable. It originates from the fact that information about parents' birth places are available only for 5% of the population in the 1940 census. In the 1940 census, parents' birth places were surveyed in a "long-form" questionnaire. The long-form questionnaire was administered only to a 5% random sample of the population. However, in the 1930 census parents' birth places were surveyed for the entire population. As a result, the size of the sample with which we can estimate the effect of legibility on share validated is *less* than 5% of the boundary sample, if we use only those who answered questions in the long-form survey. This is because we can only validate linked records, and the linkage rate is at most 54%.[1] With such a small sample, the effect of legibility on share validated is likely to be imprecisely estimated, as can be seen in Table A39.

To address this issue, we increase the size of the sample with non-missing values of parents' birth places by including those who were living with at least one of their parents in both censuses. Including this demographic group increases the size of the sample (with information about parents' birth places) by approximately 9 to 10 times, depending

[1] See Table A4 for the linkage rates for each linking algorithm.

on the linking algorithm.[2] This augmented sample is more selected than those who took the long-form survey, because individuals living with their parents at the time of both censuses are likely to be younger than the average population. This is indeed the case as confirmed in Table A41. Including those residing with their parents in both censuses makes the sample younger, as well as include a higher share of individuals who are single and more likely to be the children of the household head.

Note that the non-representativeness of our sample does not affect the validity of our research design (i.e. exploiting discontinuities in the legibility measure at the boundary of enumeration districts) as long as our sample is balanced across enumeration boundaries. Table A42 indicates that this is indeed the case. This table compares the mean of various observable characteristics of those living on more or less legible sides of the boundaries – restricting to those with information about parents' birth places. The differences in means are either not statistically significant or the standardized differences are smaller than the threshold of 0.1, recommended by Austin (2009) to determine balance. Therefore, our research design applied to this sample would still allow us to identify the causal effect of legibility on share validated. However, the magnitude of the effects we find may differ relative to a more representative sample.

We have weak evidence that our sample construction leads to under-estimation of the effect of legibility on share validated. Tables A39 and A40 present estimates of the effect of legibility on share validated for those who took the long-form survey and those who did not (but whom we include in the sample because we have information about their parents' birth places) respectively. We find that the magnitude of the effect for the former group is larger than that for the latter. Given that the size of the latter group is much larger, the estimated effect for the augmented sample is closer to that for the latter group

[2]To compare how much the sample size changes by adding this demographic group, compare the sample sizes in Table 4 with that in Table A39).

(see Table 4). However, this is not conclusive since the estimates for the long-form survey takers are statistically imprecise (though the effect sizes are large).

B Abbreviations in Tables A41 and A4

- Link. rate: linkage rate
- Sh. val. (PB): share validated, where the validation variable is parents' birth places
- Sh. val. (MI): share validated, where the validation variable is the middle name initial
- BPL: NE: birth state in New England census region
- MA: Mid-Atlantic census region
- ENC: East North Central census region
- WNC: West North Central census region
- SA: South Atlantic census region
- ESC: East South Central census region
- WSC: West South Central census region
- MTN: Mountain census region
- In BPL: living in birth state
- 5-yr mig.: 5 year interstate migration
- NEast: Northeast
- MW: Midwest

C The list of selected cities

The Urban Transition Project team digitized enumeration district maps for 43 selected cities. We use these maps to identify households that are at borders of two neighboring enumeration districts. The list of these 43 cities are as follows: Akron, OH; Baltimore, MD; Birmingham, AL; Boston, MA; Bridgeport, CT; Chattanooga, TN; Chicago, IL; Columbus, OH; Dallas, TX; Denver, CO; Des Moines, IA; Flint, MI; Fort Worth, TX; Grand Rapids, MI; Hartford, CT; Houston, TX; Indianapolis, IN; Jacksonville, FL; Jersey City, NJ; Kansas City, KS; Los Angeles, CA; Miami, FL; Milwaukee, WI; Minneapolis, MN; Nashville, TN; Newark, NJ; New Haven, CT; New Orleans, LA; Oakland, CA; Oklahoma City, OK; Omaha, NE; Philadelphia, PA; Pittsburgh, PA; San Antonio, TX; San Francisco, CA; Seattle, WA; St. Paul, MN; Syracuse, NY; Trenton, NJ; Tulsa, OK; Washingtonn DC; Worchester, MA; and Yonkers, NY.

D Enumeration district boundaries in 1940 Census

This section summarizes geographical planning of the 1940 U.S. Federal Census, described in Jenkins (1985).

The map of enumeration districts was created by the Census Bureau's Division of Geography. The task of preparing the maps began by dividing the states of the U.S. into supervisors' districts. One or more counties were allotted to each supervisor's district. A "plan of division by enumeration district (ED)" was then prepared for each county.

EDs were designed to be clearly defined areas. Their boundaries were to follow either the boundaries of municipalities, wards, or minor civil divisions; or roads, streets, railways, public survey lines, and other well-known lines. Note that we drop those enumerations districts from our (boundary) sample whose boundaries overlap with that of townships or wards.

The size of each ED was determined so that it can be canvassed by a single enumerator in a desirable time frame: about two weeks in urban areas or a month in rural areas. In order to achieve this goal, the Division of Geography had to take into consideration the number of inhabitants, the number of farms and access to each residential area. In practice, ED boundaries in rural areas for 1940 census were mostly based on those for 1930 census, except in cases where the Field Division had recommended that the ED be divided, where changes had occurred in the minor civil divisions, or where the description of the ED in the 1930 census was incorrect.

Similarly, ED boundaries used in the 1930 census in urban areas were to be used in the 1940 census, except when changes had occured in minor civil divisions, assembly districts, or ward areas that resulted in a fragmented ED; where there was a revision of census tracts; where the ED description was incorrect; where the ED had impractical boundaries; or where available information indicated that the population within the ED

was too large or too small.

E Weight estimation

We closely follow the inverse probability weighting specification described in Bailey et al. (2020b). Their's is a useful benchmark for linked census data because they show that their inverse probability weighting restores the representativeness of a widely used linked census sample, the 1860-1880 IPUMS Linked Representative Sample (see column 7 of Table 5 in that paper).

To compute the weight for each observation, first we estimate the following probit model:

$$\begin{aligned} &\mathbb{1}\{\text{Observation i is linked and has non-missing values for parents' birth places}\} \\ =&\mathbb{1}\{X_i\beta + \epsilon \geq 0\}, \epsilon_i \sim N(0,1) \end{aligned} \tag{1}$$

The control variables are: dummy variables for size of local city (under 1,000 or unincorporated; 1,000 to 2,499; 2,500 to 3,999; 4,000 to 4,999; 5,000 to 9,999; 10,000 to 24,999; 25,000 to 49,999; 50,000 to 74,999; 75,000 to 99,999, 100,000 to 199,999; 200,000 to 299,999; 300,000 to 499,999; 500,000 to 599,999; 600,000 to 749,999; 750,000 to 999,999; 1 million to 1.99 million and 2 million and up); dummy variables for census region of birth location (Northeast, Mid-Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, and Mountain); dummy variables for relationship of individual to household head (head/householder, spouse, child, child-in-law, parent, parent-in-law, sibling, sibling-in-law, grandchild, other relatives, parent friend or visitor); dummy variables for occupation categories (professional-technical, farmers, managers, officials, proprietors, clerical and kindred, sales workers, craftsmen, operatives, service workers, farm Laborers, laborers, non-occupational response); dummy variables for five-year categories of ages; dummy variables for region of residence (Northeast, Midwest, West); dummy variables for whether an individual lives with his mother, lives with his fa-

ther, or lives with both parents; dummy variables for married/currently separated, widowed, or divorced/never married; dummy variable for living on a farm; whether they were living in the same state as birth; and number of siblings in the household.

After estimating the probit model (equation 1 above), we predict the probability of being linked and having non-missing values for parents' birth places using the estimates of the model. We use the inverse of the predicted probability as the weight for each linked observation.

F A simple model of legibility and the quality of linked sample

In this section, we present a simple model of census data linkage. The purpose of this model is to motivate our empirical investigation in the main text and in section G. Our framework shows that the effect of legibility on the quality of linked samples depend on certain unobservable parameters, and, without knowing the magnitude of those parameters, the sign of the effect can be ambiguous for some quality measures.

We posit a simple model where two people (persons 1 and 2) are enumerated in two consecutive censuses (e.g., 1930 and 1940 censuses). To focus on how legibility affects the quality of linked samples through its impact on transcription errors, we assume that linking relies only on names. One way to interpret this assumption is that each of these two people have already satisfied the other conditions to be declared a link. For example, the difference in recorded ages between the two censuses is within an acceptable range. We also assume that the names of persons 1 and 2 are sufficiently different that, if the two names are transcribed without error, then any linking algorithm can create correct links for each of them.[3]

For tractability, we assume that the names are transcribed without errors in the first census. However, the names may be transcribed with errors in the second census. Transcription errors may occur due to poor legibility of the census schedule.[4] Factors that affect the degree of transcription error are treated as random. We do not directly model how the realization of these random factors translate to transcription errors, but instead assume that there are three possible events, each of which occurs with a certain probability. For each person $i \in \{1, 2\}$, the three events are as follows:

[3]This model abstracts from enumeration errors.

[4]Names may also be mis-transcribed due to mistakes by the transcriber.

Event A In the second census, i's names are transcribed sufficiently similarly to i's names in the first census, so i is (correctly) a candidate to be linked to himself in the first census;

Event B In the second census, i's names are transcribed sufficiently similarly to $-i$'s names in the first census, so i is (incorrectly) a candidate to be linked to $-i$ in the first census; and

Event C In the second census, i's names are not sufficiently similar to either i's or $-i$'s names in the first census, so i is not a candidate to be linked to either himself or $-i$ in the first census,

where $-i$ refers to person 2 if $i = 1$, and person 1 if $i = 2$.[5]

Note that the likelihood that these events occur depends on the extent of errors in the transcribed names. If the extent of error is sufficiently small, then Event A occurs. Event B may occur if the extent of error is severe; and Event C may occur if the extent of error is intermediate. We denote the probability of Events A, B, and C with a, b and c respectively, and assume that these probabilities are the same for person 1 and 2. We also assume that the two names are sufficiently different so that Events A and B cannot occur simultaneously for one person. That is, it is not possible that a person i in the second census is a candidate for both persons 1 and 2 in the first census. Therefore, the sum of the three probabilities a, b, and c is equal to 1. Finally, we assume for tractability that these

[5]A person in the second census, say i, could be a *candidate* to be linked to another person in the first census, say j. But i's candidacy does not guarantee that she is *linked* to j because $-i$ (also in the second census) can also be a candidate for j. In case of such ties, all existing linking algorithms link neither i nor $-i$ (in the second census) to j (in the first census). In other words, all linking algorithms link i to j if and only if i and j are unique candidates for each other. In our model, we follow this common practice.

events are independent across the two persons, i.e., for all $(k, l) \in \{A, B, C\} \times \{A, B, C\}$

$$\begin{aligned}&\Pr(\text{Event } k \text{ occurs for person 1} \cap \text{Event } l \text{ occurs for person 2})\\ =&\Pr(\text{Event } k \text{ occurs for person 1}) \cdot \Pr(\text{Event } l \text{ occurs for person 2})\end{aligned}$$

Figure A1 illustrates all configurations of the candidacies for links, the probability with which they occur, and the linkage rates and the share of true links in each configuration. In this model, the expected linkage rate, denoted with LR, is equal to the following:

$$a^2 + b^2 + ac + bc = a + b - 2ab \tag{2}$$

where the equality is obtained by substituting $1 - a - b$ for c. Similarly, the expected share of true links, denoted by ST, is equal to the following:

$$\frac{a^2 + 2ac}{a^2 + b^2 + 2ac + 2bc} = \frac{-(a+b)^2 + 2(a+b) + b^2 - 2b}{-(a+b)^2 + 2(a+b) - 2ab} \tag{3}$$

where the equality once again derives from the same substitution as above. Note that the denominator of (3) is equal to the probability that the linkage rate is not zero. Since the share of true links is well defined only if the linkage rate is non-zero, the expected share of true links must condition on the events in which linkage rates are greater than zero.

We assume that greater legibility increases the probability that each person in the second census is a candidate to herself in the first census (i.e., it increases a), while the probabilities for the other two events decrease (i.e., it decreases b and c). That is, we assume the following:

$$\frac{\partial a}{\partial \ell} \geq 0, \quad \frac{\partial b}{\partial \ell} \leq 0, \quad \frac{\partial c}{\partial \ell} \leq 0$$

where ℓ denotes the legibility of the census schedule. Under these assumptions, we can derive the following result:

Proposition 1. *The change in the expected linkage rate in response to an increase in legibility is equal to the following:*

$$\frac{dLR}{d\ell} = \frac{\partial a}{\partial \ell}(1-2b) + \frac{\partial b}{\partial \ell}(1-2a) \tag{4}$$

The sign of this derivative is positive if $a \geq b$. The change in the expected share of true links in response to an increase in legibility is equal to the following:

$$\frac{dST}{d\ell} = \left(\frac{1}{-(a+b)^2 + 2(a+b) - 2ab}\right) \cdot \left[2(1-ST)\left(1-(a+b)\right)\left(\frac{\partial a}{\partial \ell} + \frac{\partial b}{\partial \ell}\right) - 2 \cdot \frac{\partial b}{\partial \ell}\left(1 - \left(ST \cdot a + \left(1 + ST \cdot \frac{\frac{\partial a}{\partial \ell}}{\frac{\partial b}{\partial \ell}}\right) b\right)\right)\right]$$

The sign of this derivative is positive.

Proof. Differentiating the expected linkage rate with respect to legibility ℓ, we get

$$\begin{aligned}
\frac{dLR}{d\ell} =& \frac{d(a+b-2ab)}{d\ell} \\
=& \frac{\partial a}{\partial \ell} + \frac{\partial b}{\partial \ell} - 2b\frac{\partial a}{\partial \ell} - 2a\frac{\partial b}{\partial \ell} \\
=& \frac{\partial a}{\partial \ell}(1-2b) + \frac{\partial b}{\partial \ell}(1-2a) \\
=& -\frac{\partial b}{\partial \ell}\left(\underbrace{\frac{\frac{\partial a}{\partial \ell}}{-\frac{\partial b}{\partial \ell}}}_{\geq 1} \cdot (1-2b) - (1-2a)\right) \\
\geq& -\frac{\partial b}{\partial \ell}\left\{(1-2b) - (1-2a)\right\} = -\frac{\partial b}{\partial \ell}(2a-2b) \geq 0
\end{aligned}$$

where the first inequality holds under the assumption that $(1-2b) \geq 0$.

On the other hand, differentiating the expected share of true links with respect to legibility ℓ, we get

$$\begin{aligned}\frac{dST}{d\ell} =& \frac{d\left(\frac{-(a+b)^2+2(a+b)+b^2-2b}{-(a+b)^2+2(a+b)-2ab}\right)}{d\ell} \\ =& \underbrace{\frac{1}{-(a+b)^2+2(a+b)-2ab}}_{>0} \cdot \\ & \left((-(a+b)^2+2(a+b)+b^2-2b)' - \underbrace{ST}_{\in(0,1]} \times (-(a+b)^2+2(a+b)-2ab)'\right) \quad (5)\end{aligned}$$

where the operator $(\cdot)'$ denotes differentiation with respect to legibility ℓ. The derivative $(-(a+b)^2+2(a+b)+b^2-2b)'$ (i.e., the first term in the parentheses) is equal to:

$$\begin{aligned}& -2(a+b)\left(\frac{\partial a}{\partial \ell}+\frac{\partial b}{\partial \ell}\right)+2\left(\frac{\partial a}{\partial \ell}+\frac{\partial b}{\partial \ell}\right)+2b\frac{\partial b}{\partial \ell}-2\frac{\partial b}{\partial \ell} \\ =& 2\underbrace{(1-(a+b))}_{\geq 0}\cdot\underbrace{\left(\frac{\partial a}{\partial \ell}+\frac{\partial b}{\partial \ell}\right)}_{\geq 0}\underbrace{-2\cdot\frac{\partial b}{\partial \ell}}_{\geq 0}\underbrace{(1-b)}_{\geq 0} \quad (6) \\ \geq& 0\end{aligned}$$

The derivative $(-(a+b)^2+2(a+b)-2ab)'$, the second term in the parentheses without the multiplier ST, is equal to:

$$\begin{aligned}& -2(a+b)\left(\frac{\partial a}{\partial \ell}+\frac{\partial b}{\partial \ell}\right)+2\left(\frac{\partial a}{\partial \ell}+\frac{\partial b}{\partial \ell}\right)-2b\frac{\partial a}{\partial \ell}-2a\frac{\partial b}{\partial \ell} \\ =& 2\underbrace{(1-(a+b))}_{\geq 0}\cdot\underbrace{\left(\frac{\partial a}{\partial \ell}+\frac{\partial b}{\partial \ell}\right)}_{\geq 0}\underbrace{-2\cdot\frac{\partial b}{\partial \ell}}_{\geq 0}\left(a+b\frac{\frac{\partial a}{\partial \ell}}{\frac{\partial b}{\partial \ell}}\right) \quad (7)\end{aligned}$$

Note that (6) is greater than equal to (7) because

$$(1-b) \geq \left(a + b \frac{\frac{\partial a}{\partial \ell}}{\frac{\partial b}{\partial \ell}} \right) \tag{8}$$

Therefore, it follows that:

$$(-(a+b)^2 + 2(a+b) + b^2 - 2b)' - ST \times (-(a+b)^2 + 2(a+b) - 2ab)' \geq 0$$

which implies that (5) is greater than or equal to zero, as desired. □

Our sufficient condition for the change in the expected linkage rate to be positive (i.e., $a \geq b$) cannot be tested with our data because neither a nor b are directly observable. We can come up with non-pathological examples where violations of our sufficient condition causes expected linkage rates to *decrease* with legibility.[6] Intuitively, this can occur because decreases in b can reduce linkage rates. While we believe that this condition is mild and likely to be true in most historical linkages, whether this assumption holds is ultimately an empirical question, which we turn to in the next section.

On the other hand, our results for share of true links ("share validated") do not depend on the relative magnitude of these parameters. Intuitively, if the probability of true candidacy (a) increases and the probability of false candidacy (b) decreases due to an increase in legibilty, any candidacy (hence links) should be more likely to be true relative to when legibility was lower. However, our results are silent about the magnitude of this effect or about how legibility may affect the share of true links differentially across sociodemographic groups. This motivates our analyses of heterogeneous effects in section G.

[6]For example, if $a = 0.26$, $b = 0.29$, $\frac{\partial a}{\partial \ell} = 0.29$, and $\frac{\partial b}{\partial \ell} = -0.28$, then $\frac{dLR}{d\ell} = -0.0126$.

G Heterogeneity in the effect of legibility on the quality of linkage

In this section, we present findings on the heterogeneous effects of legibility on the quality of the linked samples. Recall that in our theoretical model in Section F, the effect of legibility on the quality of linkage was a function of several parameters, including the probability of correct/incorrect candidacy (a and b) and how sensitive these probabilities are to legibility ($\frac{\partial a}{\partial \ell}$ and $\frac{\partial b}{\partial \ell}$). Because these parameters are typically not observable and since they may be different across socio-demographic groups, it is difficult to predict a priori whether the effect of legibility would be heterogeneous. In this section, we empirically show that heterogeneous effects exist in certain dimensions, but not in others.

To estimate heterogeneous effects, we modify our baseline specification (1) by interacting our measure of legibility with different socio-demographic variables Z_i. That is, we estimate the following model:

$$q_i = \beta_1 \ell_{e_i} + \beta_2 (\ell_{e_i} \cdot Z_i) + X_i' \gamma_1 + \gamma_2 Z_i + \delta_{b_i} + \epsilon_i \tag{9}$$

The parameter of interest is β_2, which shows how increases in legibility differentially affect the sociodemographic group of interest. A positive β_2 would indicate that the quality of linkage for that particular group is comparatively more sensitive to legibility than the average effect on other groups (recall from subsections 3.1 and 3.2 that β_1 is positive). For continuous variables, a positive β_2 would imply higher sensitivity for larger values of Z_i.

Tables A43 and A44 present these results. We test if heterogeneity exists along the following important demographic dimensions: race (black vs. white), age, foreign vs. U.S.-born, children of immigrant fathers (or "second generation") vs. children of U.S.-born fathers, years of schooling, occupational score, wage/salary income, and occupational

group (white-collar occupations vs. blue-collar occupations). All of these variables are obtained from the 1940 census. Our results for the linkage rates are summarized below (see also Table A43):

1. The linkage rates for blacks are less sensitive to legibility;

2. The linkage rates for foreign-borns are less sensitive to legibility than U.S.-borns; but those for the children of foreign-born fathers are more sensitive to legibility than those for the children of U.S.-born fathers;

3. The linkage rates for more educated people and those that have white-collar occupations are more sensitive to legibility; and

4. For other socio-demographic characteristics, we either do not have evidence that legibility has heterogeneous effects along those dimensions (the effects are not statistically significant), or the magnitude of the effects ($\hat{\beta}_2$) are negligible.

These results are robust under alternative legibility measures (Tables A46 and A45) and also when restricting samples to boundaries with sufficiently large differences in legibility between the two sides (results available upon request).

On the other hand, we do not find evidence that the effect of legibility on share validated is heterogeneous, except for foreign-borns and children of foreign-born fathers (see Tables A47 and A48). Both for foreign-born individuals and the children of immigrant fathers, share validated is less sensitive to legibility than native borns or for children of U.S.-born fathers. However, we note that this finding is not robust under the alternative validation variable, i.e., using middle name initials (see Table A49).

Given the relationship between the unobservable parameters ($a, b, \frac{\partial a}{\partial \ell}$, and $\frac{\partial b}{\partial \ell}$) and the effect of legibility on the quality of linked samples shown through our theoretical model

(section F), it is challenging to pin down why we observe this heterogeneity[7]. Ideally, one would estimate these parameters for different socio-demographic groups with ground-truth links. Because of the large costs involved with constructing ground-truth links, we leave it for future research. We also caution that the findings from this section may not generalize to other contexts where these parameters may have different values.

[7]Of course, we could make a reasonable guess for certain categories. For example, individuals with more education and those in white-collar occupations may have more complicated names and hence their names are more difficult to transcribe if illegible, and hence more sensitive. But these would simply be conjectures and since we do not have a systematic way of doing this currently, understanding why these heterogeneous effects exist, is best left for future work

H Additional figures and tables

H.1 Figures

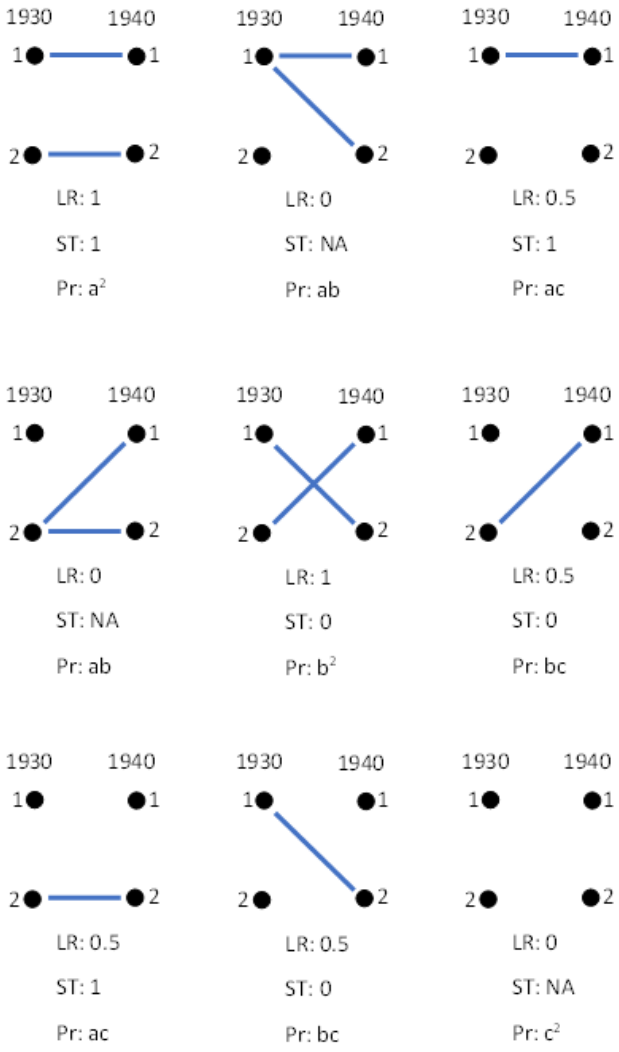


Figure A1: All configurations of link candidacies in the model in section F, and associated linkage rates, share of true links, and probabilities

Note: "1930" and "1940" in the figure corresponds to the census years, and "1"'s and "2"'s next to the nodes are the index of the person. For each configuration, "LR" denotes the linkage rate, "ST" denotes the share of true links, and "Pr" denotes the probability that the given configuration occurs. Note that share of true links is undefined if the linkage rate is equal to zero. A link between two people is valid only if each person associated with the link is a unique candidate to the other. For example, the configuration in the center of the top line has a linkage rate of zero because both people in 1940 are candidates for a link to person 1 in 1930.

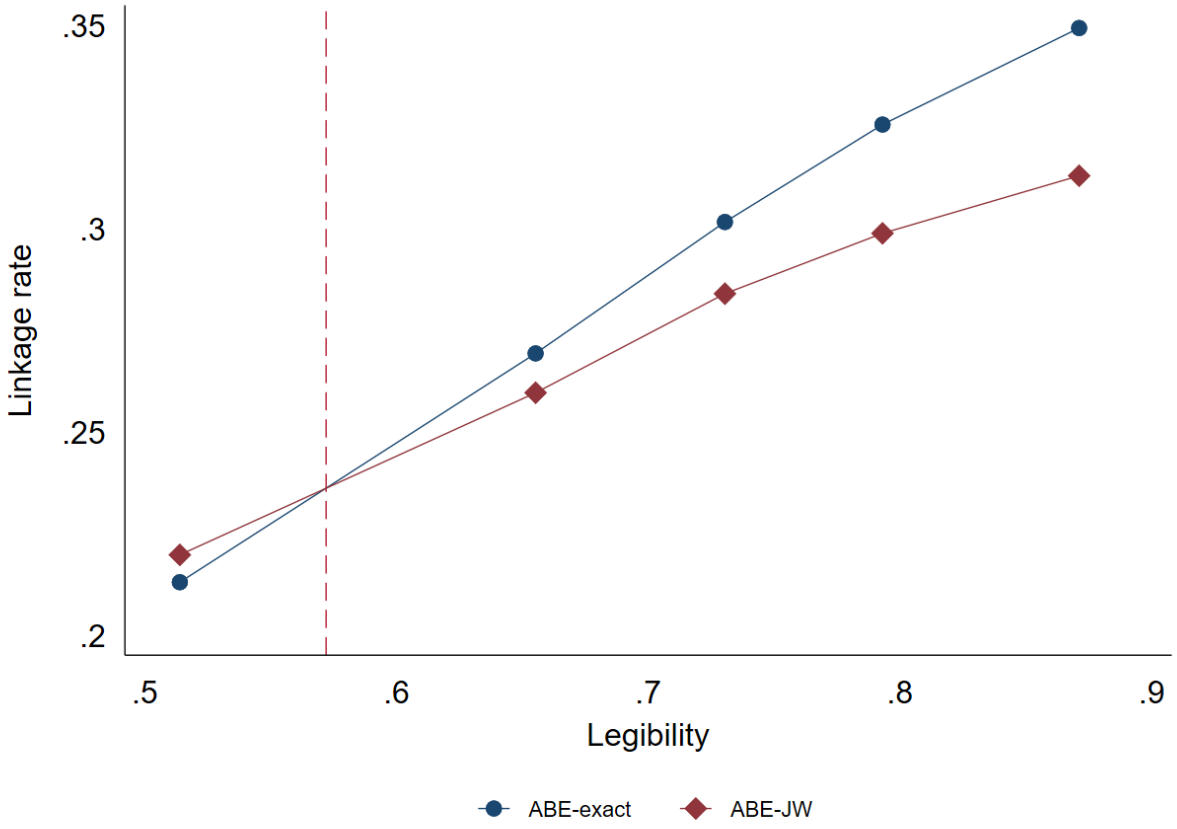


Figure A2: The effect of legibility on the linkage rate (ABE-exact and ABE-JW only)

Note: The vertical line indicates the level of legibility (0.57) at which the linkage rates of ABE-exact and ABE-JW coincide.

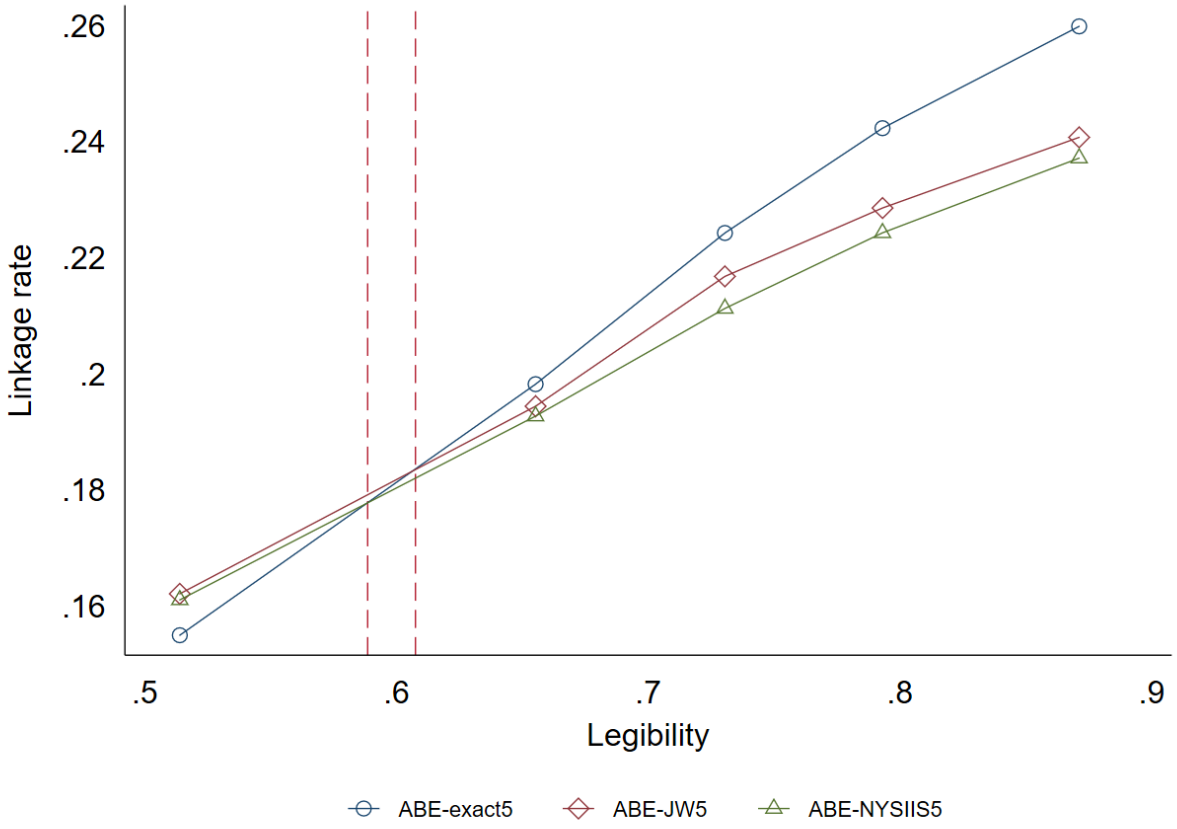


Figure A3: The effect of legibility on the linkage rate (ABE-exact5, ABE-JW5, and ABE-NYSIIS5 only)

Note: The vertical line to the left indicates the level of legibility (0.59) at which the linkage rates of ABE-exact5 and ABE-NYSIIS5 coincide; the vertical line to the right indicates the level of legibility (0.61) at which the linkage rates of ABE-exact5 and ABE-JW5 coincide.

H.2 Tables

Table A1: Comparison between two validation variables

Linking algorithms	# of obs. not missing			Validation status coincides
	PBPL	MI	Either	
ABE-exact	92,225	36,771	12,506	0.757
ABE-JW	90,016	31,541	10,957	0.754
ABE-NYSIIS	100,096	36,113	12,197	0.755
ABE-exact5	69,480	26,854	9,080	0.762
ABE-JW5	67,003	23,531	7,925	0.759
ABE-NYSIIS5	64,975	23,122	7,676	0.755
ML	116,947	45,801	15,862	0.759
MLP	201,261	69,048	25,599	0.758

Note: This table presents the number of linked observations in our boundary sample for which we have information about each of the validation variables. The column labeled "PBPL" contains the number of linked observations, in each linked sample, for which we have information about an individual's parents' birth places. The column labeled "MI" is the corresponding column for the alternative validation variable – middle name initials. The column labeled "Either" presents the number of linked observations for which we have information about both validation variables. Lastly, the column labeled "Validation status coincides" contains the share of linked observations for whom we have information about both validation variables and for whom the validation status coincides across the two validation variables.

Table A2: Comparison of observable characteristics across samples

	U.S.-born white or black male	Linkable population	Boundary sample
# obs.	65,815,534	48,097,569	739,634
Legibility	0.72	0.72***	0.71***
Age	31.02	34.54***	35.09***
Black	0.09	0.09***	0.09***
Married	0.46	0.53***	0.53***
BPL: NE	0.05	0.06***	0.07***
MA	0.17	0.17***	0.21***
ENC	0.18	0.18***	0.20***
WNC	0.11	0.10***	0.07***
SA	0.14	0.14***	0.10***
ESC	0.10	0.10***	0.04***
WSC	0.10	0.10***	0.09***
MTN	0.02	0.02***	0.01***
In BPL	0.70	0.69***	0.62***
5-yr mig.	0.05	0.04***	0.03***
Head	0.45	0.52***	0.50***
Child	0.41	0.36***	0.36***
-in-law	0.01	0.01***	0.02***
Parent	0.01	0.01***	0.01***
-in-law	0.00	0.01***	0.01***
Sibling	0.01	0.01***	0.02***
-in-law	0.01	0.01***	0.01***
Lives with both parents	0.35	0.30***	0.28***
w/mother	0.06	0.07***	0.08***
w/father	0.02	0.02***	0.02***
# of siblings	1.05	0.97***	0.89***
In NEast	0.27	0.28***	0.36***
In MW	0.31	0.30***	0.33***
In West	0.11	0.08***	0.09***

Note: The linkable population consists of U.S.-born white or black males who are 8 years old or older. We also exclude those who live in enumeration districts that contain fewer than 50 people, or where less than 90% of people have both transcriptions of their names. The boundary sample consists of those who live in one of the 43 selected cities (see section C for the list of these cities) and at the boundary of two enumeration districts. We exclude boundaries that overlap with township or ward boundaries or only one side of which are inhabited.

Table A3: Balance of observables at the boundary of enumeration districts 2

Variable	Less legible	More legible	Diff.	Std. diff.
Live in: Northeast region	0.356	0.366	0.011***	0.016
Midwest	0.335	0.332	-0.003**	-0.004
South	0.218	0.211	-0.007***	-0.012
West	0.091	0.090	-0.001	-0.003
Occupation: professional	0.043	0.043	0.000	0.001
Farmer	0.003	0.002	-0.000***	-0.004
Manager	0.064	0.065	0.001	0.002
Clerical	0.075	0.076	0.001	0.002
Sales	0.056	0.056	0.000	0.001
Craftsman	0.145	0.146	0.001	0.001
Operatives	0.156	0.155	-0.000	-0.001
Service	0.067	0.067	-0.000	-0.000
Farm labor	0.011	0.010	-0.000	-0.002
Laborer	0.097	0.096	-0.000	-0.000
Non-occupational response	0.284	0.283	-0.002*	-0.003
Lived in: same house 5 years ago	0.441	0.447	0.006***	0.008
same county	0.488	0.483	-0.004***	-0.006
different county	0.022	0.022	0.000	0.000
abroad	0.003	0.002	-0.000***	-0.005
Unknown	0.018	0.016	-0.003***	-0.015
Observations	359,890	379,744	739,634	

Note: This table presents the means of various observable characteristics of those who live on either side of an enumeration district boundary — the side where legibility is relatively lower ('Less legible") and the side where it is relatively higher ("More legible"). The unit of observations in this table is a person. The sample used to create this table is the boundary sample, i.e., those who live on the streets that serve as the border of the two neighboring enumeration districts. The variables are obtained from the 1940 Census. We follow IPUMS's classification to code occupations. See `https://usa.ipums.org/usa-action/variables/OCC1950#codes_section` for the detailed coding scheme for occupations in the U.S. census.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A4: Comparison of mean characteristics between linkable population and linked samples

	Pop.	ABE-						ML	MLP
		exact5	JW5	NYSIIS5	exact	JW	NYSIIS		
Link. rate	1.00	0.22	0.22	0.22	0.30	0.29	0.33	0.37	0.54
Sh. val. (PB)	1.00	0.85	0.85	0.84	0.83	0.83	0.82	0.84	0.88
Sh. val. (MI)	1.00	0.86	0.86	0.86	0.82	0.83	0.81	0.84	0.88
Legibility	0.72	0.74***	0.74***	0.74***	0.74***	0.74***	0.74***	0.73***	0.73***
Age	34.54	35.14***	34.85***	35.16***	35.23***	34.53	35.23***	34.80***	34.65***
Black	0.09	0.04***	0.07***	0.05***	0.06***	0.07***	0.06***	0.07***	0.06***
Married	0.53	0.55***	0.55***	0.56***	0.55***	0.54***	0.55***	0.55***	0.51***
BPL: NE	0.06	0.07***	0.07***	0.07***	0.07***	0.07***	0.07***	0.07***	0.06***
MA	0.17	0.17***	0.14***	0.15***	0.18***	0.16***	0.16***	0.16***	0.18***
ENC	0.18	0.22***	0.20***	0.20***	0.21***	0.21***	0.20***	0.20***	0.20***
WNC	0.10	0.14***	0.14***	0.14***	0.13***	0.14***	0.13***	0.13***	0.12***
SA	0.14	0.11***	0.14***	0.12***	0.12***	0.13***	0.13***	0.14***	0.12***
ESC	0.10	0.07***	0.09***	0.08***	0.08***	0.09***	0.08***	0.09***	0.09***
WSC	0.10	0.08***	0.09***	0.09***	0.08***	0.09***	0.09***	0.08***	0.09***
MTN	0.02	0.03***	0.03***	0.03***	0.02***	0.02***	0.02***	0.02***	0.02***
In BPL	0.69	0.72***	0.72***	0.71***	0.72***	0.73***	0.71***	0.72***	0.73***
5-yr mig.	0.04	0.04***	0.04***	0.04***	0.04***	0.04***	0.04***	0.04***	0.03***
Head	0.52	0.55***	0.54***	0.55***	0.55***	0.54***	0.55***	0.54***	0.50***
Child	0.36	0.36***	0.36***	0.36***	0.36***	0.37***	0.35***	0.36***	0.44***
-in-law	0.01	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***
Parent	0.01	0.01***	0.01***	0.01***	0.01***	0.01***	0.01^{+}	0.01***	0.01***
-in-law	0.01	0.00***	0.00***	0.01***	0.01***	0.00***	0.01	0.00***	0.00***
Sibling	0.01	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***
-in-law	0.01	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***	0.01***
Live with both parents	0.30	0.30***	0.30***	0.30***	0.30***	0.31***	0.29***	0.30***	0.37***
w/mother	0.07	0.06***	0.07***	0.06***	0.06***	0.07***	0.06***	0.07	0.07***
w/father	0.02	0.02***	0.02***	0.02***	0.02***	0.02***	0.02***	0.02***	0.03***
# of siblings	0.97	0.95***	0.97***	0.95***	0.94***	0.98***	0.95***	0.97***	1.21***
In NEast	0.28	0.27***	0.24***	0.26***	0.28***	0.26***	0.27***	0.26***	0.29***
In MW	0.30	0.37***	0.36***	0.36***	0.35***	0.35***	0.35***	0.35***	0.35***
In West	0.08	0.10***	0.10***	0.10***	0.09***	0.09***	0.10***	0.09***	0.08***

Note: The column labeled "pop" is created using our linkable population, i.e., white or black males who are 8 years or older in 1940. We exclude enumeration districts with fewer than 50 people or where less than 90% of the people have both transcriptions of their names. After dropping 9,220,999 such observations, we are left with 48,097,569 observations. The null hypothesis that the mean of the linkable population is equal to that of each of the linked samples is tested. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. For the meaning of abbreviations in the left-most column, see section B.

Table A5: Comparing the effect of legibility on linkage rates: ABE-exact vs others

	ABE-exact vs ABE-JW	ABE-exact vs ABE-NYSIIS	ABE-exact vs ABE-exact5	ABE-exact vs ABE-JW5	ABE-exact vs ABE-NYSIIS5	ABE-exact vs ML	ABE-exact vs MLP
$\widehat{\beta}_1$	0.360***	0.347***	0.353***	0.360***	0.353***	0.359***	0.352***
	(0.00611)	(0.00620)	(0.00594)	(0.00595)	(0.00594)	(0.00624)	(0.00629)
$\widehat{\beta}_2$	-0.141***	-0.108***	-0.0887***	-0.182***	-0.177***	-0.106***	-0.101***
	(0.00553)	(0.00570)	(0.00527)	(0.00527)	(0.00527)	(0.00576)	(0.00585)
BDFE	Y	Y	Y	Y	Y	Y	Y
N	1451580	1451580	1451580	1451580	1451580	1451580	1451580
adj. R^2	0.056	0.053	0.066	0.062	0.062	0.067	0.157

Note: We use stacked boundary samples to estimate model (2). In each column the reference linking algorithm is the ABE-exact algorithm. The dependent variable is linked/not linked (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on linkage rate for the ABE-exact algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of the other algorithms and ABE-exact. For each column in this table, the dummy variables in model (2) are appropriately modified. Robust standard errors are reported in parentheses. BDFE refers to boundary fixed effects.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A6: Comparing the effect of legibility on linkage rates: ABE-exact5 vs others

	ABE-exact5 vs ABE-exact	ABE-exact5 vs ABE-JW	ABE-exact5 vs ABE-NYSIIS	ABE-exact5 vs ABE-JW5	ABE-exact5 vs ABE-NYSIIS5	ABE-exact5 vs ML	ABE-exact5 vs MLP
$\widehat{\beta}_1$	0.264***	0.276***	0.263***	0.276***	0.269***	0.275***	0.268***
	(0.00568)	(0.00567)	(0.00577)	(0.00549)	(0.00549)	(0.00582)	(0.00586)
$\widehat{\beta}_2$	0.0887***	-0.0521***	-0.0192***	-0.0931***	-0.0887***	-0.0172**	-0.0118*
	(0.00527)	(0.00525)	(0.00543)	(0.00497)	(0.00497)	(0.00549)	(0.00559)
BDFE	Y	Y	Y	Y	Y	Y	Y
N	1451580	1451580	1451580	1451580	1451580	1451580	1451580
adj. R^2	0.066	0.061	0.064	0.055	0.053	0.087	0.203

Note: We use stacked boundary samples to estimate model (2). In each column the reference linking algorithm is the ABE-exact5 algorithm. The dependent variable is linked/not linked (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on linkage rate for the ABE-exact5 algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and ABE-exact5. For each column in this table, the dummy variables in model (2) are appropriately modified. Robust standard errors are reported in parentheses. BDFE refers to boundary fixed effects.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A7: The effect of legibility on linkage rates (alternative legibility measure 1)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.274***	0.193***	0.184***	0.354***	0.235***	0.244***	0.269***	0.259***
	(0.00707)	(0.00703)	(0.00701)	(0.00786)	(0.00778)	(0.00819)	(0.00837)	(0.00849)
BDFE	Y	Y	Y	Y	Y	Y	Y	Y
N	725790	725790	725790	725790	725790	725790	725790	725790
adj. R^2	0.051	0.052	0.047	0.052	0.055	0.044	0.067	0.142

Note: We use the boundary sample to estimate model (1). The alternative legibility measure used for this table is based on the name-cleaning procedure in which we do not remove spaces between the letters of a name. The dependent variable is linked/not linked (1/0). Controls include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided most or all of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); one's birth state; and highest grade of schooling. All controls are directly obtained from 1940 census. BDFE refers to boundary fixed effects. Robust standard errors are reported in parentheses.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A8: The effect of legibility on linkage rates (alternative legibility measure 2)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.280***	0.197***	0.187***	0.362***	0.240***	0.248***	0.276***	0.267***
	(0.00726)	(0.00721)	(0.00720)	(0.00808)	(0.00799)	(0.00843)	(0.00861)	(0.00875)
BDFE	Y	Y	Y	Y	Y	Y	Y	Y
N	725790	725790	725790	725790	725790	725790	725790	725790
adj. R^2	0.051	0.052	0.047	0.052	0.055	0.044	0.067	0.142

Note: We use the boundary sample to estimate model (1). The alternative legibility measure used in this table is based on a different criteria for declaring two transcriptions of the same person's name as not identical. Specifically, we declare two transcriptions *not* to be identical only if the Jaro-Winkler distance between the two transcriptions is greater than 0.044 (which is the 75th percentile of the population distribution of Jaro-Winkler distance among the transcriptions). The dependent variable is linked/not linked (1/0). Controls include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided most or all of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); one's birth state; and highest grade of schooling. All controls are directly obtained from 1940 census. BDFE refers to boundary fixed effects. Robust standard errors are reported in parentheses.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A9: Comparing the effect of legibility on linkage rates: ABE-exact5 vs Others (alternative legibility 1)

	ABE-exact5 vs ABE-exact	ABE-exact5 vs ABE-JW	ABE-exact5 vs ABE-NYSIIS	ABE-exact5 vs ABE-JW5	ABE-exact5 vs ABE-NYSIIS5	ABE-exact5 vs ML	ABE-exact5 vs MLP
$\widehat{\beta}_1$	0.270***	0.281***	0.268***	0.281***	0.273***	0.279***	0.273***
	(0.00577)	(0.00575)	(0.00586)	(0.00558)	(0.00557)	(0.00591)	(0.00595)
$\widehat{\beta}_2$	0.0889***	-0.0525***	-0.0188***	-0.0940***	-0.0885***	-0.0156**	-0.0135*
	(0.00533)	(0.00531)	(0.00549)	(0.00503)	(0.00503)	(0.00555)	(0.00565)
BDFE	Y	Y	Y	Y	Y	Y	Y
N	1451580	1451580	1451580	1451580	1451580	1451580	1451580
adj. R^2	0.066	0.061	0.064	0.055	0.053	0.087	0.203

Note: We use stacked boundary samples to estimate model (2). The alternative legibility measure used for this table is based on the name-cleaning procedure in which we do not remove spaces between letters of the name. In each column the reference linking algorithm is the ABE-exact5 algorithm. The dependent variable is linked/not linked (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on the linkage rate of the ABE-exact5 algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and ABE-exact5. For each column in this table, the dummy variables in model (2) are appropriately modified. Robust standard errors are in parentheses. BDFE refers to boundary fixed effects.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A10: Comparing the effect of legibility on linkage rates: ABE-exact vs others (alternative legibility measure 1)

	ABE-exact vs ABE-JW	ABE-exact vs ABE-NYSIIS	ABE-exact vs ABE-exact5	ABE-exact vs ABE-JW5	ABE-exact vs ABE-NYSIIS5	ABE-exact vs ML	ABE-exact vs MLP
$\widehat{\beta}_1$	0.365***	0.353***	0.358***	0.365***	0.358***	0.364***	0.358***
	(0.00620)	(0.00629)	(0.00603)	(0.00604)	(0.00603)	(0.00634)	(0.00638)
$\widehat{\beta}_2$	-0.141***	-0.108***	-0.0889***	-0.183***	-0.177***	-0.104***	-0.102***
	(0.00560)	(0.00576)	(0.00533)	(0.00533)	(0.00533)	(0.00583)	(0.00592)
BDFE	Y	Y	Y	Y	Y	Y	Y
N	1451580	1451580	1451580	1451580	1451580	1451580	1451580
adj. R^2	0.056	0.053	0.066	0.062	0.062	0.067	0.157

Note: We use stacked boundary samples to estimate model (2). The alternative legibility measure used for this table is based on the name-cleaning procedure in which we do not remove spaces between letters of the name. In each column the reference linking algorithm is the ABE-exact algorithm. The dependent variable is linked/not linked (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on the linkage rate for the ABE-exact algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and ABE-exact. For each column in this table, the dummy variables in model (2) are appropriately modified. Robust standard errors are reported in parentheses. BDFE refers to boundary fixed effects.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A11: Comparing the effect of legibility on linkage rates: ABE-exact5 vs others (alternative legibility measure 2)

	ABE-exact5 vs ABE-exact	ABE-exact5 vs ABE-JW	ABE-exact5 vs ABE-NYSIIS	ABE-exact5 vs ABE-JW5	ABE-exact5 vs ABE-NYSIIS5	ABE-exact5 vs ML	ABE-exact5 vs MLP
$\widehat{\beta}_1$	0.275***	0.287***	0.274***	0.287***	0.280***	0.286***	0.278***
	(0.00593)	(0.00592)	(0.00602)	(0.00573)	(0.00573)	(0.00607)	(0.00613)
$\widehat{\beta}_2$	0.0928***	-0.0535***	-0.0205***	-0.0964***	-0.0928***	-0.0162**	-0.00961
	(0.00550)	(0.00549)	(0.00567)	(0.00519)	(0.00519)	(0.00574)	(0.00585)
BDFE	Y	Y	Y	Y	Y	Y	Y
N	1451580	1451580	1451580	1451580	1451580	1451580	1451580
adj. R^2	0.066	0.061	0.064	0.055	0.053	0.087	0.203

Note: We use stacked boundary samples to estimate model (2). The alternative legibility measure used for this table is based on an alternative criterion for declaring two transcriptions of the same person's name as not identical. Specifically, we declar two transcriptions not to be identical only if they are sufficiently different in the sense the Jaro-Winkler distance between the two transcriptions is greater than 0.044, which is the 75th percentile of the population distribution of Jaro-Winkler distance between the two transcriptions. In each column the reference linking algorithm is the ABE-exact5 algorithm. The dependent variable is linked/not linked (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on the linkage rate for the ABE-exact5 algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and ABE-exact5. For each column in this table, the dummy variables in model (2) are appropriately modified. Robust standard errors are reported in parentheses. BDFE refers to boundary fixed effects.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A12: Comparing the effect of legibility on linkage rates: ABE-exact vs others (alternative legibility measure 2)

	ABE-exact vs ABE-JW	ABE-exact vs ABE-NYSIIS	ABE-exact vs ABE-exact5	ABE-exact vs ABE-JW5	ABE-exact vs ABE-NYSIIS5	ABE-exact vs ML	ABE-exact vs MLP
$\widehat{\beta}_1$	0.374***	0.362***	0.368***	0.374***	0.367***	0.374***	0.366***
	(0.00638)	(0.00647)	(0.00621)	(0.00621)	(0.00621)	(0.00652)	(0.00657)
$\widehat{\beta}_2$	-0.146***	-0.113***	-0.0928***	-0.189***	-0.186***	-0.109***	-0.102***
	(0.00578)	(0.00595)	(0.00550)	(0.00551)	(0.00551)	(0.00602)	(0.00612)
BDFE	Y	Y	Y	Y	Y	Y	Y
N	1451580	1451580	1451580	1451580	1451580	1451580	1451580
adj. R^2	0.056	0.053	0.066	0.062	0.062	0.067	0.157

Note: We use the boundary sample to estimate model (1). The alternative legibility measure used in this table is based on a different criteria for declaring two transcriptions of the same person's name as not identical. Specifically, we declare two transcriptions *not* to be identical only if the Jaro-Winkler distance between the two transcriptions is greater than 0.044 (which is the 75th percentile of the population distribution of Jaro-Winkler distance between the two transcriptions). In each column the reference linking algorithm is the ABE-exact algorithm. The dependent variable is linked/not linked (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on the linkage rate for the ABE-exact algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and ABE-exact. For each column in this table, the dummy variables in model (2) are appropriately modified. Robust standard errors are in parentheses. BDFE refers to boundary fixed effects.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A13: The effect of legibility on linkage rates (sample restricted to boundaries across which the difference in legibility is sufficiently large)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Baseline	0.269***	0.190***	0.180***	0.348***	0.230***	0.239***	0.263***	0.255***
	(0.00696)	(0.00691)	(0.00689)	(0.00774)	(0.00765)	(0.00806)	(0.00824)	(0.00836)
5th percentile	0.269***	0.190***	0.180***	0.349***	0.231***	0.239***	0.264***	0.256***
	(0.00696)	(0.00691)	(0.00689)	(0.00774)	(0.00765)	(0.00806)	(0.00824)	(0.00836)
10th percentile	0.269***	0.190***	0.180***	0.349***	0.230***	0.239***	0.263***	0.255***
	(0.00696)	(0.00691)	(0.00689)	(0.00774)	(0.00765)	(0.00806)	(0.00824)	(0.00836)
25th percentile	0.270***	0.190***	0.181***	0.350***	0.231***	0.240***	0.265***	0.258***
	(0.00697)	(0.00693)	(0.00691)	(0.00776)	(0.00767)	(0.00809)	(0.00826)	(0.00838)
50th percentile	0.269***	0.190***	0.182***	0.349***	0.230***	0.240***	0.265***	0.258***
	(0.00715)	(0.00710)	(0.00708)	(0.00795)	(0.00787)	(0.00829)	(0.00848)	(0.00862)

Note: This table presents estimates of model (1) restricting the sample only to boundaries where legibility changes by at least a certain threshold value. We use different quantiles (i.e., 5th, 10th, 25th, or 50th) of the distribution of differences in legibility as the thresholds for sample restrictions. The reference estimates obtained with no sample restrictions are shown in the row labeled "Baseline." The dependent variable is linked/not linked (1/0). Controls include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided most or all of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); one's birth state; and highest grade of schooling. All controls are directly obtained from 1940 census. Robust standard errors are reported in parentheses.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A14: Comparing the effect of legibility on linkage rates: ABE-exact vs Others (sample restricted to boundaries across which the difference in legibility is sufficiently large)

Thresholds	Parameters	ABE-exact vs ABE-JW	ABE-exact vs ABE-NYSIIS	ABE-exact vs ABE-exact5	ABE-exact vs ABE-JW5	ABE-exact vs ABE-NYSIIS5	ABE-exact vs ML	ABE-exact vs MLP
Baseline	$\widehat{\beta}_1$	0.360***	0.347***	0.353***	0.360***	0.353***	0.359***	0.352***
		(0.00611)	(0.00620)	(0.00594)	(0.00595)	(0.00594)	(0.00624)	(0.00629)
	$\widehat{\beta}_2$	-0.141***	-0.108***	-0.0887***	-0.182***	-0.177***	-0.106***	-0.101***
		(0.00553)	(0.00570)	(0.00527)	(0.00527)	(0.00527)	(0.00576)	(0.00585)
5th percentile	$\widehat{\beta}_1$	0.359***	0.347***	0.353***	0.359***	0.353***	0.358***	0.351***
		(0.00612)	(0.00621)	(0.00597)	(0.00597)	(0.00596)	(0.00626)	(0.00630)
	$\widehat{\beta}_2$	-0.139***	-0.107***	-0.0886***	-0.180***	-0.177***	-0.104***	-0.0973***
		(0.00563)	(0.00579)	(0.00535)	(0.00536)	(0.00536)	(0.00586)	(0.00595)
10th percentile	$\widehat{\beta}_1$	0.359***	0.347***	0.352***	0.359***	0.352***	0.357***	0.351***
		(0.00615)	(0.00624)	(0.00599)	(0.00600)	(0.00599)	(0.00628)	(0.00632)
	$\widehat{\beta}_2$	-0.138***	-0.107***	-0.0874***	-0.179***	-0.176***	-0.102***	-0.0988***
		(0.00572)	(0.00589)	(0.00544)	(0.00545)	(0.00545)	(0.00596)	(0.00605)
25th percentile	$\widehat{\beta}_1$	0.358***	0.348***	0.353***	0.359***	0.352***	0.357***	0.352***
		(0.00624)	(0.00632)	(0.00609)	(0.00610)	(0.00609)	(0.00637)	(0.00641)
	$\widehat{\beta}_2$	-0.136***	-0.107***	-0.0866***	-0.178***	-0.173***	-0.100***	-0.0964***
		(0.00605)	(0.00623)	(0.00575)	(0.00576)	(0.00576)	(0.00630)	(0.00640)
50th percentile	$\widehat{\beta}_1$	0.357***	0.348***	0.352***	0.358***	0.352***	0.354***	0.348***
		(0.00656)	(0.00663)	(0.00642)	(0.00643)	(0.00642)	(0.00667)	(0.00672)
	$\widehat{\beta}_2$	-0.134***	-0.106***	-0.0865***	-0.177***	-0.173***	-0.0938***	-0.0901***
		(0.00684)	(0.00703)	(0.00649)	(0.00650)	(0.00650)	(0.00712)	(0.00724)

Note: This table presents estimates of model (1) restricting the sample only to boundaries where legibility changes by at least a certain threshold value. We use different quantiles (i.e., 5th, 10th, 25th, or 50th) of the distribution of differences in legibility as the thresholds for sample restrictions. The reference estimates obtained with no sample restrictions are shown in the row labeled "Baseline." In each column the reference linking algorithm is the ABE-exact algorithm. The dependent variable is linked/not linked (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on linkage rate for the ABE-exact algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and ABE-exact. For each column in this table, the dummy variables in model (2) are appropriately modified. Robust standard errors are reported in parentheses. $^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A15: Comparing the effect of legibility on linkage rates: ABE-exact5 vs Others (sample restricted to boundaries across which the difference in legibility is sufficiently large)

Thresholds	Parameters	ABE-exact5 vs ABE-exact	ABE-exact5 vs ABE-JW	ABE-exact5 vs ABE-NYSIIS	ABE-exact5 vs ABE-JW5	ABE-exact5 vs ABE-NYSIIS5	ABE-exact5 vs ML	ABE-exact5 vs MLP
Baseline	$\widehat{\beta}_1$	0.264***	0.276***	0.263***	0.276***	0.269***	0.275***	0.268***
		(0.00568)	(0.00567)	(0.00577)	(0.00549)	(0.00549)	(0.00582)	(0.00586)
	$\widehat{\beta}_2$	0.0887***	-0.0521***	-0.0192***	-0.0931***	-0.0887***	-0.0172**	-0.0118*
		(0.00527)	(0.00525)	(0.00543)	(0.00497)	(0.00497)	(0.00549)	(0.00559)
5th percentile	$\widehat{\beta}_1$	0.264***	0.275***	0.263***	0.275***	0.269***	0.274***	0.267***
		(0.00570)	(0.00568)	(0.00578)	(0.00551)	(0.00550)	(0.00583)	(0.00588)
	$\widehat{\beta}_2$	0.0886***	-0.0505***	-0.0182***	-0.0917***	-0.0879***	-0.0155**	-0.00864
		(0.00535)	(0.00534)	(0.00552)	(0.00505)	(0.00505)	(0.00559)	(0.00568)
10th percentile	$\widehat{\beta}_1$	0.265***	0.275***	0.264***	0.275***	0.269***	0.273***	0.268***
		(0.00571)	(0.00570)	(0.00580)	(0.00553)	(0.00552)	(0.00584)	(0.00589)
	$\widehat{\beta}_2$	0.0874***	-0.0510***	-0.0195***	-0.0919***	-0.0883***	-0.0149**	-0.0114*
		(0.00544)	(0.00543)	(0.00561)	(0.00514)	(0.00514)	(0.00568)	(0.00578)
25th percentile	$\widehat{\beta}_1$	0.267***	0.275***	0.265***	0.276***	0.269***	0.274***	0.269***
		(0.00578)	(0.00577)	(0.00587)	(0.00561)	(0.00561)	(0.00591)	(0.00596)
	$\widehat{\beta}_2$	0.0866***	-0.0493***	-0.0203***	-0.0912***	-0.0862***	-0.0134*	-0.00974
		(0.00575)	(0.00574)	(0.00593)	(0.00543)	(0.00543)	(0.00600)	(0.00611)
50th percentile	$\widehat{\beta}_1$	0.266***	0.274***	0.264***	0.275***	0.269***	0.271***	0.265***
		(0.00605)	(0.00604)	(0.00612)	(0.00589)	(0.00589)	(0.00616)	(0.00622)
	$\widehat{\beta}_2$	0.0865***	-0.0477***	-0.0195**	-0.0909***	-0.0867***	-0.00731	-0.00363
		(0.00649)	(0.00648)	(0.00669)	(0.00612)	(0.00612)	(0.00678)	(0.00691)

Note: This table presents estimates of model (1) restricting the sample only to boundaries where legibility changes by at least a certain threshold value. We use different quantiles (i.e., 5th, 10th, 25th, or 50th) of the distribution of differences in legibility as the thresholds for sample restrictions. The reference estimates obtained with no sample restrictions are shown in the row labeled "Baseline." In each column the reference linking algorithm is the ABE-exact5 algorithm. The dependent variable is linked/not linked (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on linkage rate for the ABE-exact5 algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and ABE-exact5. For each column in this table, the dummy variables in model (2) are appropriately modified. Robust standard errors are reported in parentheses. $^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A16: Comparing the effect of legibility on share validated: MLP vs others

	MLP vs ABE-exact5	MLP vs ABE-JW5	MLP vs ABE-NYSIIS5	MLP vs ABE-exact	MLP vs ABE-JW	MLP vs ABE-NYSIIS	MLP vs ML
$\widehat{\beta}_1$	0.0490***	0.0581***	0.0545***	0.0511***	0.0586***	0.0612***	0.0592***
	(0.0115)	(0.0115)	(0.0115)	(0.0113)	(0.0113)	(0.0113)	(0.0110)
$\widehat{\beta}_2$	0.0252$^+$	0.0320*	0.0479***	0.0310*	0.0495***	0.0799***	0.0336**
	(0.0140)	(0.0141)	(0.0144)	(0.0129)	(0.0127)	(0.0126)	(0.0115)
BDFE	Y	Y	Y	Y	Y	Y	Y
N	266616	264144	262127	288972	286790	296747	313291
adj. R^2	0.104	0.106	0.104	0.104	0.106	0.106	0.106

Note: We use stacked linked samples to estimate model (2). In each column the reference linking algorithm is the MLP algorithm. The dependent variable is validated/not validated (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on share validated for the MLP algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and MLP. BDFE refers to boundary fixed effects. Robust standard errors are reported in parentheses.$^+$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A17: Comparing the effect of legibility on share validated: MLP vs others (alternative legibility measure 1)

	MLP vs ABE-exact5	MLP vs ABE-JW5	MLP vs ABE-NYSIIS5	MLP vs ABE-exact	MLP vs ABE-JW	MLP vs ABE-NYSIIS	MLP vs ML
$\widehat{\beta}_1$	0.0489***	0.0587***	0.0550***	0.0509***	0.0594***	0.0613***	0.0601***
	(0.0117)	(0.0117)	(0.0117)	(0.0115)	(0.0115)	(0.0115)	(0.0111)
$\widehat{\beta}_2$	0.0267$^+$	0.0322*	0.0467**	0.0331*	0.0500***	0.0804***	0.0337**
	(0.0142)	(0.0143)	(0.0146)	(0.0130)	(0.0129)	(0.0127)	(0.0116)
BDFE	Y	Y	Y	Y	Y	Y	Y
N	266616	264144	262127	288972	286790	296747	313291
adj. R^2	0.104	0.106	0.104	0.104	0.106	0.106	0.106

Note: We use stacked linked samples to estimate model (2). The alternative legibility measure used for this table is based on the name-cleaning procedure in which we do not remove spaces between letters of the name. In each column the reference linking algorithm is the MLP algorithm. The dependent variable is validated/not validated (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on share validated for the MLP algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and MLP. BDFE refers to boundary fixed effects. Robust standard errors are reported in parentheses.$^+$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A18: Comparing the effect of legibility on share validated: MLP vs others (alternative legibility measure 2)

	MLP vs ABE-exact5	MLP vs ABE-JW5	MLP vs ABE-NYSIIS5	MLP vs ABE-exact	MLP vs ABE-JW	MLP vs ABE-NYSIIS	MLP vs ML
$\widehat{\beta}_1$	0.0529***	0.0623***	0.0585***	0.0550***	0.0629***	0.0657***	0.0640***
	(0.0120)	(0.0120)	(0.0121)	(0.0118)	(0.0118)	(0.0118)	(0.0115)
$\widehat{\beta}_2$	0.0284$^+$	0.0358*	0.0505***	0.0341*	0.0528***	0.0847***	0.0371**
	(0.0148)	(0.0148)	(0.0151)	(0.0135)	(0.0134)	(0.0132)	(0.0121)
BDFE	Y	Y	Y	Y	Y	Y	Y
N	266616	264144	262127	288972	286790	296747	313291
adj. R^2	0.104	0.106	0.104	0.104	0.106	0.106	0.106

We use stacked linked samples to estimate model (2). The alternative legibility measure used for this table is based on an alternative criterion for declaring two transcriptions of the same person's name as not identical. Specifically, we declar two transcriptions not to be identical only if they are sufficiently different in the sense the Jaro-Winkler distance between the two transcriptions is greater than 0.044, which is the 75th percentile of the population distribution of Jaro-Winkler distance between the two transcriptions. In each column the reference linking algorithm is the MLP algorithm. The dependent variable is validated/not validated (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on share validated for the MLP algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and MLP. BDFE refers to boundary fixed effects. Robust standard errors are reported in parentheses.$^+$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A19: Comparing the effect of legibility on share validated: MLP vs others (alternative validation variable)

	MLP vs ABE-exact5	MLP vs ABE-JW5	MLP vs ABE-NYSIIS5	MLP vs ABE-exact	MLP vs ABE-JW	MLP vs ABE-NYSIIS	MLP vs ML
$\widehat{\beta}_1$	0.0959***	0.118***	0.101***	0.107***	0.117***	0.120***	0.126***
	(0.0231)	(0.0232)	(0.0234)	(0.0229)	(0.0230)	(0.0230)	(0.0216)
$\widehat{\beta}_2$	0.0476$^+$	0.0729**	0.101***	0.0805***	0.102***	0.105***	0.0472*
	(0.0245)	(0.0256)	(0.0261)	(0.0228)	(0.0237)	(0.0232)	(0.0195)
BDFE	Y	Y	Y	Y	Y	Y	Y
N	94167	90884	90475	103893	98744	103232	112758
adj. R^2	0.068	0.064	0.064	0.077	0.074	0.080	0.072

Note: We use stacked linked samples to estimate model (2). The validation variable for this table is middle name initials. In each column the reference linking algorithm is the MLP algorithm. The dependent variable is validated/not validated (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on share validated for the MLP algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and MLP. BDFE refers to boundary fixed effects. Robust standard errors are reported in parentheses.$^+$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A20: Comparing the effect of legibility on share validated: MLP vs others (sample restricted to boundaries across which the difference in legibility is sufficiently large)

Thresholds	Parameters	MLP vs ABE-exact5	MLP vs ABE-JW5	MLP vs ABE-NYSIIS5	MLP vs ABE-exact	MLP vs ABE-JW	MLP vs ABE-NYSIIS	MLP vs ML
Baseline	$\widehat{\beta}_1$	0.0490***	0.0581***	0.0545***	0.0511***	0.0586***	0.0612***	0.0592***
		(0.0115)	(0.0115)	(0.0115)	(0.0113)	(0.0113)	(0.0113)	(0.0110)
	$\widehat{\beta}_2$	0.0252^{+}	0.0320*	0.0479***	0.0310*	0.0495***	0.0799***	0.0336**
		(0.0140)	(0.0141)	(0.0144)	(0.0129)	(0.0127)	(0.0126)	(0.0115)
5th percentile	$\widehat{\beta}_1$	0.0488***	0.0570***	0.0546***	0.0506***	0.0579***	0.0604***	0.0584***
		(0.0115)	(0.0115)	(0.0116)	(0.0113)	(0.0113)	(0.0113)	(0.0110)
	$\widehat{\beta}_2$	0.0260^{+}	0.0364*	0.0479**	0.0328*	0.0515***	0.0824***	0.0358**
		(0.0143)	(0.0144)	(0.0146)	(0.0131)	(0.0130)	(0.0128)	(0.0117)
10th percentile	$\widehat{\beta}_1$	0.0488***	0.0570***	0.0549***	0.0507***	0.0582***	0.0603***	0.0581***
		(0.0115)	(0.0115)	(0.0116)	(0.0113)	(0.0113)	(0.0113)	(0.0110)
	$\widehat{\beta}_2$	0.0265^{+}	0.0369*	0.0476**	0.0328*	0.0507***	0.0832***	0.0364**
		(0.0146)	(0.0147)	(0.0150)	(0.0134)	(0.0133)	(0.0131)	(0.0120)
25th percentile	$\widehat{\beta}_1$	0.0487***	0.0562***	0.0536***	0.0505***	0.0590***	0.0599***	0.0576***
		(0.0116)	(0.0116)	(0.0116)	(0.0114)	(0.0114)	(0.0114)	(0.0111)
	$\widehat{\beta}_2$	0.0308*	0.0456**	0.0553***	0.0351*	0.0538***	0.0866***	0.0402**
		(0.0155)	(0.0156)	(0.0159)	(0.0142)	(0.0140)	(0.0139)	(0.0127)
50th percentile	$\widehat{\beta}_1$	0.0551***	0.0618***	0.0593***	0.0568***	0.0646***	0.0651***	0.0649***
		(0.0120)	(0.0120)	(0.0121)	(0.0119)	(0.0118)	(0.0118)	(0.0116)
	$\widehat{\beta}_2$	0.0261	0.0393*	0.0533**	0.0317^{+}	0.0442**	0.0812***	0.0311*
		(0.0178)	(0.0178)	(0.0182)	(0.0163)	(0.0160)	(0.0158)	(0.0145)

Note: This table presents estimates of model (1) restricting the sample only to boundaries where legibility changes by at least a certain threshold value. We use different quantiles (i.e., 5th, 10th, 25th, or 50th) of the distribution of differences in legibility as the thresholds for sample restrictions. The reference estimates obtained with no sample restrictions are shown in the row labeled "Baseline." In each column the reference linking algorithm is the MLP algorithm. The dependent variable is validated/not validated (1/0). $\widehat{\beta}_1$ is to be interpreted as the effect of an increase in legibility on share validated of the MLP algorithm, and $\widehat{\beta}_2$ as the difference in the effect between each of other algorithms and MLP. Robust standard errors are in reported parentheses.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A21: The effect of legibility on share validated (alternative legibility measure 1)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.0674*	0.117***	0.123***	0.0844***	0.138***	0.169***	0.117***	0.0483***
	(0.0281)	(0.0281)	(0.0288)	(0.0241)	(0.0237)	(0.0230)	(0.0200)	(0.0127)
BDFE	Y	Y	Y	Y	Y	Y	Y	Y
N	68529	66057	64040	90885	88703	98660	115204	198087
adj. R^2	0.103	0.102	0.096	0.090	0.088	0.084	0.080	0.085

Note: We only use linked observations in the boundary sample to estimate model (1). The alternative legibility measure used for this table is based on the name-cleaning procedure where we do not remove spaces between the letters of a name. The dependent variable is validated/not validated (1/0). Covariates include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided the most (or all) of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); birth state; highest grade of schooling. All controls are directly obtained from 1940 census. BDFE refers to boundary fixed effects. Robust standard errors are reported in parentheses.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A22: The effect of legibility on share validated (alternative legibility measure 2)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.0766**	0.127***	0.133***	0.0923***	0.146***	0.181***	0.127***	0.0512***
	(0.0291)	(0.0291)	(0.0298)	(0.0250)	(0.0246)	(0.0238)	(0.0207)	(0.0131)
BDFE	Y	Y	Y	Y	Y	Y	Y	Y
N	68529	66057	64040	90885	88703	98660	115204	198087
adj. R^2	0.103	0.102	0.096	0.090	0.089	0.084	0.080	0.085

Note: We only use linked observations in the boundary sample to estimate model (1). The alternative legibility measure used for this table is based on a different criteria for declaring two transcriptions of the same person's name as not identical. Specifically, we declare two transcriptions *not* to be identical only if the Jaro-Winkler distance between the two transcriptions is greater than 0.044 (which is the 75th percentile of the population distribution of Jaro-Winkler distance between the two transcriptions). The dependent variable is validated/not validated (1/0). Covariates include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided the most (or all) of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); birth state; highest grade of schooling. All controls are directly obtained from 1940 census. BDFE refers to boundary fixed effects. Robust standard errors are reported in parentheses.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A23: The effect of legibility on share validated (sample restricted to boundaries across which the difference in legibility is sufficiently large)

Thresholds	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Baseline	0.0673*	0.116***	0.124***	0.0841***	0.136***	0.169***	0.116***	0.0480***
	(0.0277)	(0.0278)	(0.0284)	(0.0238)	(0.0234)	(0.0227)	(0.0197)	(0.0125)
5th percentile	0.0680*	0.116***	0.125***	0.0845***	0.136***	0.169***	0.116***	0.0479***
	(0.0277)	(0.0278)	(0.0284)	(0.0238)	(0.0234)	(0.0227)	(0.0197)	(0.0125)
10th percentile	0.0673*	0.115***	0.125***	0.0838***	0.136***	0.169***	0.115***	0.0481***
	(0.0277)	(0.0278)	(0.0284)	(0.0238)	(0.0234)	(0.0227)	(0.0197)	(0.0125)
25th percentile	0.0693*	0.118***	0.124***	0.0835***	0.140***	0.169***	0.116***	0.0490***
	(0.0278)	(0.0279)	(0.0285)	(0.0239)	(0.0235)	(0.0228)	(0.0198)	(0.0126)
50th percentile	0.0773**	0.128***	0.134***	0.0920***	0.140***	0.171***	0.121***	0.0530***
	(0.0288)	(0.0287)	(0.0294)	(0.0247)	(0.0242)	(0.0235)	(0.0203)	(0.0129)

Note: This table presents estimates of model (1) restricting the sample only to boundaries where legibility changes by at least a certain threshold value. We use different quantiles (i.e., 5th, 10th, 25th, or 50th) of the distribution of differences in legibility as the thresholds for sample restrictions. The reference estimates obtained with no sample restrictions are shown in the row labeled "Baseline." The dependent variable is validated/not validated (1/0). Controls include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided most or all of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); one's birth state; and highest grade of schooling. All controls are directly obtained from 1940 census. Robust standard errors are in reported parentheses.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A24: The effect of legibility on share validated (alternative validation variable)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.206***	0.288***	0.256***	0.221***	0.291***	0.300***	0.224***	0.0935***
	(0.0574)	(0.0597)	(0.0607)	(0.0490)	(0.0510)	(0.0490)	(0.0374)	(0.0256)
BDFE	Y	Y	Y	Y	Y	Y	Y	Y
N	26433	23150	22741	36159	31010	35498	45024	67734
adj. R^2	0.044	0.038	0.044	0.045	0.047	0.050	0.032	0.024

Note: We only use linked observations in the boundary sample to estimate model (1). The validation variable for this table is middle name initials. The dependent variable is validated/not validated (1/0). Covariates include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided the most (or all) of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); birth state; highest grade of schooling. All controls are directly obtained from 1940 census. BDFE refers to boundary fixed effects. Robust standard errors are included in parentheses.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A25: Balance of observables at the boundary of enumeration districts (conditional on being linked under ABE-Exact5)

Variable	Less legible	More legible	Diff.	Std. diff.
Own a house	0.484	0.485	0.001	0.001
Multi-gen. household	0.980	0.981	0.001	0.005
Age	22.134	22.218	0.085	0.005
Black	0.047	0.047	-0.001	-0.002
Married	0.126	0.125	-0.001	-0.002
In school	0.449	0.444	-0.004	-0.006
# characters in full name	12.947	13.024	0.077***	0.024
Main provider of info.	0.056	0.056	0.001	0.002
Absent when enumerated	0.008	0.010	0.002**	0.012
Head of household	0.096	0.095	-0.001	-0.002
Years of schooling	11.078	11.098	0.019	0.004
Foreign born mother	0.337	0.341	0.004	0.006
Foreign born father	0.392	0.394	0.002	0.003
Non-institutional residence	0.999	0.998	-0.001*	-0.009
Employed	0.430	0.439	0.010**	0.014
In labor force	0.655	0.664	0.009**	0.014
# weeks worked	40.979	40.768	-0.211	-0.010
# hours per week worked	41.723	42.086	0.364***	0.024
Labor income	542.607	553.765	11.158*	0.011
Nonlabor income ≥ 50	0.092	0.095	0.003	0.008
Size of household	5.366	5.346	-0.020	-0.007
Observations	31,362	38,118	69,480	

Note: This table presents the means of various observable characteristics of those who live on either side of an enumeration district boundary — the side where legibility is relatively lower ('Less legible") and the side where it is relatively higher ("More legible"). The unit of observations in this table is a person. The sample used to create this table is the boundary sample, i.e., those who live on streets that serve as the border between two neighboring enumeration districts, *conditional on being linked by the ABE-exact5 algorithm*. $^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A26: Balance of observables at the boundary of enumeration districts (conditional on being linked under ABE-JW5)

Variable	Less legible	More legible	Diff.	Std. diff.
Own a house	0.479	0.477	-0.002	-0.003
Multi-gen. household	0.979	0.980	0.001	0.007
Age	22.094	22.032	-0.062	-0.004
Black	0.075	0.078	0.003	0.008
Married	0.126	0.123	-0.003	-0.007
In school	0.448	0.448	-0.000	-0.000
# characters in full name	13.189	13.285	0.096***	0.030
Main provider of info.	0.055	0.056	0.001	0.004
Absent when enumerated	0.009	0.010	0.001*	0.010
Head of household	0.094	0.092	-0.002	-0.005
Years of schooling	11.034	11.017	-0.016	-0.004
Foreign born mother	0.345	0.359	0.014***	0.021
Foreign born father	0.406	0.417	0.011***	0.016
Non-institutional residence	0.998	0.998	0.000	0.004
Employed	0.429	0.433	0.003	0.005
In labor force	0.654	0.660	0.006	0.009
# weeks worked	41.113	40.660	-0.453**	-0.021
# hours per week worked	41.864	42.136	0.272**	0.017
Labor income	542.195	537.194	-5.001	-0.005
Nonlabor income ≥ 50	0.092	0.095	0.002	0.006
Size of household	5.386	5.414	0.028	0.009
Observations	30,875	36,128	67,003	

Note: This table presents the means of various observable characteristics of those who live on either side of an enumeration district boundary — the side where legibility is relatively lower ('Less legible") and the side where it is relatively higher ("More legible"). The unit of observations in this table is a person. The sample used to create this table is the boundary sample, i.e., those who live on streets that serve as the border between two neighboring enumeration districts, *conditional on being linked by the ABE-JW5 algorithm.* $^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A27: Balance of observables at the boundary of enumeration districts (conditional on being linked under ABE-NYSIIS5)

Variable	Less legible	More legible	Diff.	Std. diff.
Own a house	0.480	0.477	-0.003	-0.005
Multi-gen. household	0.978	0.980	0.002*	0.010
Age	22.254	22.234	-0.021	-0.001
Black	0.050	0.054	0.004**	0.012
Married	0.129	0.125	-0.004	-0.008
In school	0.447	0.442	-0.004	-0.006
# characters in full name	13.304	13.413	0.109***	0.034
Main provider of info.	0.057	0.059	0.002	0.005
Absent when enumerated	0.008	0.010	0.001*	0.010
Head of household	0.099	0.095	-0.004*	-0.009
Years of schooling	11.031	11.071	0.039	0.009
Foreign born mother	0.365	0.372	0.007*	0.010
Foreign born father	0.421	0.428	0.007	0.010
Non-institutional residence	0.998	0.998	-0.000	-0.003
Employed	0.432	0.438	0.006	0.009
In labor force	0.657	0.665	0.008*	0.012
# weeks worked	41.084	40.741	-0.343*	-0.016
# hours per week worked	41.834	42.240	0.405***	0.026
Labor income	541.840	545.534	3.693	0.004
Nonlabor income ≥ 50	0.093	0.097	0.004	0.010
Size of household	5.352	5.359	0.007	0.002
Observations	30,101	34,874	64,975	

Note: This table presents the means of various observable characteristics of those who live on either side of an enumeration district boundary — the side where legibility is relatively lower ('Less legible") and the side where it is relatively higher ("More legible"). The unit of observations in this table is a person. The sample used to create this table is the boundary sample, i.e., those who live on streets that serve as the border between two neighboring enumeration districts, *conditional on being linked by the ABE-NYSIIS5 algorithm*. $^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A28: Balance of observables at the boundary of enumeration districts (conditional on being linked under ABE-Exact)

Variable	Less legible	More legible	Diff.	Std. diff.
Own a house	0.473	0.475	0.002	0.003
Multi-gen. household	0.979	0.979	0.001	0.003
Age	22.238	22.306	0.068	0.004
Black	0.055	0.055	-0.000	-0.001
Married	0.129	0.127	-0.001	-0.003
In school	0.446	0.444	-0.002	-0.003
# characters in full name	12.869	12.930	0.061***	0.019
Main provider of info.	0.056	0.057	0.001	0.003
Absent when enumerated	0.008	0.010	0.002***	0.013
Head of household	0.098	0.098	-0.001	-0.001
Years of schooling	11.041	11.062	0.022	0.005
Foreign born mother	0.326	0.331	0.005	0.008
Foreign born father	0.373	0.377	0.004	0.006
Non-institutional residence	0.998	0.998	-0.000	-0.006
Employed	0.429	0.438	0.008**	0.012
In labor force	0.656	0.661	0.006	0.008
# weeks worked	41.011	40.830	-0.181	-0.008
# hours per week worked	41.802	42.059	0.257**	0.017
Labor income	542.854	552.467	9.614*	0.009
Nonlabor income ≥ 50	0.090	0.095	0.005**	0.011
Size of household	5.361	5.365	0.005	0.002
Observations	41,631	50,594	92,225	

Note: This table presents the means of various observable characteristics of those who live on either side of an enumeration district boundary — the side where legibility is relatively lower ('Less legible") and the side where it is relatively higher ("More legible"). The unit of observations in this table is a person. The sample used to create this table is the boundary sample, i.e., those who live on streets that serve as the border between two neighboring enumeration districts, *conditional on being linked by the ABE-exact algorithm*. $^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A29: Balance of observables at the boundary of enumeration districts (conditional on being linked under ABE-JW)

Variable	Less legible	More legible	Diff.	Std. diff.
Own a house	0.474	0.476	0.003	0.004
Multi-gen. household	0.980	0.981	0.001	0.006
Age	21.902	21.961	0.059	0.004
Black	0.067	0.069	0.002	0.006
Married	0.120	0.120	0.000	0.000
In school	0.450	0.449	-0.001	-0.001
# characters in full name	13.124	13.196	0.072***	0.023
Main provider of info.	0.054	0.055	0.001	0.003
Absent when enumerated	0.008	0.010	0.001**	0.011
Head of household	0.090	0.090	0.001	0.001
Years of schooling	11.041	11.040	-0.001	-0.000
Foreign born mother	0.342	0.354	0.012***	0.018
Foreign born father	0.397	0.408	0.011***	0.016
Non-institutional residence	0.998	0.998	0.000	0.001
Employed	0.424	0.429	0.005	0.007
In labor force	0.651	0.655	0.004	0.006
# weeks worked	41.002	40.530	-0.472***	-0.022
# hours per week worked	41.768	41.973	0.204*	0.013
Labor income	533.234	532.065	-1.168	-0.001
Nonlabor income ≥ 50	0.089	0.092	0.004	0.009
Size of household	5.377	5.403	0.026*	0.008
Observations	41,617	48,399	90,016	

Note: This table presents the means of various observable characteristics of those who live on either side of an enumeration district boundary — the side where legibility is relatively lower ('Less legible") and the side where it is relatively higher ("More legible"). The unit of observations in this table is a person. The sample used to create this table is the boundary sample, i.e., those who live on streets that serve as the border between two neighboring enumeration districts, *conditional on being linked by the ABE-JW algorithm*. $^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A30: Balance of observables at the boundary of enumeration districts (conditional on being linked under ABE-NYSIIS)

Variable	Less legible	More legible	Diff.	Std. diff.
Own a house	0.471	0.471	-0.000	-0.000
Multi-gen. household	0.978	0.979	0.001	0.006
Age	22.317	22.351	0.034	0.002
Black	0.055	0.056	0.002	0.006
Married	0.130	0.128	-0.002	-0.004
In school	0.441	0.439	-0.002	-0.003
# characters in full name	13.079	13.178	0.098***	0.031
Main provider of info.	0.057	0.058	0.001	0.003
Absent when enumerated	0.008	0.010	0.002***	0.013
Head of household	0.100	0.098	-0.002	-0.005
Years of schooling	10.980	11.007	0.028	0.006
Foreign born mother	0.357	0.364	0.007**	0.010
Foreign born father	0.409	0.416	0.007**	0.011
Non-institutional residence	0.998	0.998	0.000	0.001
Employed	0.431	0.438	0.007**	0.010
In labor force	0.660	0.665	0.005	0.008
# weeks worked	41.023	40.797	-0.226	-0.010
# hours per week worked	41.855	42.129	0.275**	0.018
Labor income	542.038	545.751	3.713	0.004
Nonlabor income ≥ 50	0.092	0.097	0.005**	0.012
Size of household	5.369	5.379	0.010	0.003
Observations	46,732	53,364	100,096	

Note: This table presents the means of various observable characteristics of those who live on either side of an enumeration district boundary — the side where legibility is relatively lower ('Less legible") and the side where it is relatively higher ("More legible"). The unit of observations in this table is a person. The sample used to create this table is the boundary sample, i.e., those who live on streets that serve as the border between two neighboring enumeration districts, *conditional on being linked by the ABE-NYSIIS algorithm*. $^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A31: Balance of observables at the boundary of enumeration districts (conditional on being linked under ML)

Variable	Less legible	More legible	Diff.	Std. diff.
Own a house	0.473	0.475	0.001	0.002
Multi-gen. household	0.979	0.980	0.001	0.005
Age	22.069	22.042	-0.028	-0.002
Black	0.067	0.068	0.001	0.002
Married	0.125	0.123	-0.002	-0.005
In school	0.450	0.448	-0.003	-0.004
# characters in full name	13.093	13.170	0.077***	0.024
Main provider of info.	0.055	0.056	0.001	0.002
Absent when enumerated	0.008	0.010	0.001**	0.009
Head of household	0.095	0.092	-0.004**	-0.009
Years of schooling	11.014	11.016	0.001	0.000
Foreign born mother	0.354	0.362	0.009***	0.013
Foreign born father	0.408	0.416	0.008**	0.011
Non-institutional residence	0.998	0.998	0.000	0.001
Employed	0.424	0.431	0.006**	0.009
In labor force	0.652	0.660	0.008***	0.012
# weeks worked	41.079	40.622	-0.457***	-0.021
# hours per week worked	41.832	42.039	0.207**	0.013
Labor income	538.050	540.411	2.360	0.002
Nonlabor income ≥ 50	0.089	0.093	0.003*	0.008
Size of household	5.376	5.389	0.013	0.004
Observations	54,484	62,463	116,947	

Note: This table presents the means of various observable characteristics of those who live on either side of an enumeration district boundary — the side where legibility is relatively lower ('Less legible") and the side where it is relatively higher ("More legible"). The unit of observations in this table is a person. The sample used to create this table is the boundary sample, i.e., those who live on streets that serve as the border between two neighboring enumeration districts, *conditional on being linked by the ML algorithm.* $^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A32: Balance of observables at the boundary of enumeration districts (conditional on being linked under MLP)

Variable	Less legible	More legible	Diff.	Std. diff.
Own a house	0.485	0.485	0.000	0.001
Multi-gen. household	0.986	0.986	-0.000	-0.003
Age	21.823	21.889	0.067	0.004
Black	0.050	0.053	0.003***	0.009
Married	0.110	0.109	-0.001	-0.002
In school	0.451	0.449	-0.002	-0.003
# characters in full name	12.977	13.038	0.061***	0.019
Main provider of info.	0.050	0.051	0.001	0.004
Absent when enumerated	0.009	0.011	0.002***	0.013
Head of household	0.082	0.081	-0.001	-0.002
Years of schooling	11.004	11.014	0.010	0.002
Foreign born mother	0.375	0.377	0.002	0.003
Foreign born father	0.425	0.426	0.001	0.002
Non-institutional residence	0.999	0.999	-0.000	-0.003
Employed	0.416	0.423	0.007***	0.010
In labor force	0.646	0.653	0.007***	0.011
# weeks worked	40.672	40.507	-0.165	-0.008
# hours per week worked	41.621	41.774	0.153*	0.010
Labor income	515.153	521.611	6.459*	0.006
Nonlabor income ≥ 50	0.084	0.088	0.003**	0.009
Size of household	5.550	5.556	0.006	0.002
Observations	95,682	105,579	201,261	

Note: This table presents the means of various observable characteristics of those who live on either side of an enumeration district boundary — the side where legibility is relatively lower ('Less legible") and the side where it is relatively higher ("More legible"). The unit of observations in this table is a person. The sample used to create this table is the boundary sample, i.e., those who live on streets that serve as the border between two neighboring enumeration districts, *conditional on being linked by the MLP algorithm*. $^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A33: The effect of legibility on share validated (weighted)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.0669	0.153**	0.110*	0.0655	0.123**	0.199***	0.147***	0.0538
	(0.0514)	(0.0511)	(0.0512)	(0.0482)	(0.0469)	(0.0453)	(0.0410)	(0.0331)
BDFE	Y	Y	Y	Y	Y	Y	Y	Y
N	68529	66057	64040	90885	88703	98660	115204	198087
adj. R^2	0.445	0.462	0.453	0.405	0.414	0.383	0.374	0.331

Note: We only use linked observations in the boundary sample to estimate model (1). Each observation is weighted by the predicted probability of being linked and having non-missing values of parents' birth places. See section E for details about how we estimate weights for each observation. The dependent variable is validated/not validated (1/0). Covariates include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided the most (or all) of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); birth state; highest grade of schooling. All controls are directly obtained from 1940 census. BDFE refers to boundary fixed effects. Robust standard errors are reported in parentheses.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A34: The effect of legibility on linkage rates (linkable population)

Linking algorithm	Estimates of β					
	Unadjusted	Baseline	$\delta = 1$	$\delta = -1$	Estimated δ	
					$\hat{\beta}$	δ
ABE-exact5	0.266	0.269	0.244	0.281	0.266	-0.007
	(.2659,.2661)	(.2554,.2826)	(.239, .246)	(.277, .282)	(.262, .267)	
ABE-JW5	0.187	0.190	0.158	0.205	0.187	-0.012
	(.1869,.1871)	(.1765,.2035)	(.151, .159)	(.2, .204)	(.182, .187)	
ABE-NYSIIS5	0.186	0.180	0.166	0.200	0.186	-0.007
	(.1859,.1861)	(.1665,.1935)	(.161, .167)	(.196, .2)	(.182, .187)	
ABE-exact	0.345	0.348	0.317	0.363	0.345	0.000
	(.3449,.3451)	(.3328,.3632)	(.311, .321)	(.358, .364)	(.341, .347)	
ABE-JW	0.223	0.230	0.190	0.244	0.224	-0.013
	(.2229,.2231)	(.215,.245)	(.182, .191)	(.239, .244)	(.219, .224)	
ABE-NYSIIS	0.255	0.239	0.230	0.271	0.255	0.001
	(.2549,.2551)	(.2232,.2548)	(.224, .232)	(.267, .272)	(.25, .256)	
ML	0.250	0.263	0.220	0.271	0.251	-0.014
	(.2499,.2501)	(.2468,.2792)	(.214, .223)	(.266, .272)	(.246, .252)	
MLP	0.268	0.255	0.258	0.275	0.268	-0.016
	(.2679,.2681)	(.2386,.2714)	(.252, .261)	(.27, .277)	(.263, .27)	

Note: This table presents estimates of β in model (3) when the dependent variable is the linkage rate for a given linked sample. The column labeled "Unadjusted" contains the OLS estimates of β. Our baseline estimates from model (1), obtained using the boundary sample with boundary fixed effects, are presented as a reference in the column labeled "Baseline". The columns labeled $\delta = 1$ and $\delta = -1$ contain estimates of β that correspond to the respective value of δ. Lastly, the columns labeled "Estimated δ" contain the estimates of β when we set the value of δ equal to the one estimated using the boundary sample (see footnote 22 in the main text for details about how δ is estimated). Note that we set the maximum R squared, denoted by $R_{\max}$ in Oster (2019), equal to 1 when estimating β with various δ's. Numbers in the parentheses are 95% bootstrap confidence intervals, except for that for estimates contained in the column labeled "unadjusted" and "baseline", for which we use standard error to calculate confidence intervals.

Table A35: The effect of legibility on share validated (linkable population)

Linking algorithm	Estimates of β					
	Unadjusted	Baseline	$\delta = 1$	$\delta = -1$	Estimated δ	
					$\hat{\beta}$	δ
ABE-exact5	0.074	0.067	0.005	0.072	0.073	0.003
	(.0738,.0742)	(.0127,.1213)	(-.198, .2)	(.067, .078)	(.068, .08)	
ABE-JW5	0.080	0.116	0.012	0.077	0.081	0.045
	(.0798,.0802)	(.0616,.1704)	(-.075, .073)	(.073, .082)	(.077, .088)	
ABE-NYSIIS5	0.098	0.124	-0.060	0.095	0.098	0.035
	(.0978,.0982)	(.0683,.1797)	(-.278, .087)	(.091, .1)	(.094, .105)	
ABE-exact	0.080	0.084	0.109	0.077	0.080	0.020
	(.0798,.0802)	(.0374,.1306)	(.074, .16)	(.073, .082)	(.075, .086)	
ABE-JW	0.097	0.136	0.110	0.096	0.097	0.073
	(.0968,.0972)	(.0901,.1819)	(.074, .192)	(.093, .1)	(.093, .105)	
ABE-NYSIIS	0.118	0.169	0.138	0.115	0.119	0.095
	(.1178,.1182)	(.1245,.2135)	(.118, .177)	(.111, .12)	(.115, .126)	
ML	0.077	0.116	0.097	0.075	0.077	0.044
	(.0768,.0772)	(.0773,.1547)	(.065, .136)	(.071, .079)	(.073, .083)	
MLP	0.028	0.048	0.013	0.029	0.028	0.008
	(.0279,.0281)	(.0234,.0726)	(-.047, .055)	(.026, .032)	(.024, .031)	

Note: This table presents estimates of β in model (3) when the dependent variable is share validated for a given linked sample. The column labeled "Unadjusted" contains the OLS estimates of β. Our baseline estimates from model (1), obtained using the boundary sample with boundary fixed effects, are presented as a reference in the column labeled "Baseline". The columns labeled $\delta = 1$ and $\delta = -1$ contain estimates of β that correspond to the respective value of δ. Lastly, the columns labeled "Estimated δ" contain estimates of β when we set the value of δ equal to the one estimated using the boundary sample (see footnote 22 in the main text for details about how δ is estimated). Note that we set the maximum R squared, denoted by $R_{\max}$ in Oster (2019), equal to 1 when estimating β with various δ's. Numbers in parentheses are 95% bootstrap confidence intervals, except for estimates contained in the columns labeled "unadjusted" and "baseline", for which we use standard errors to calculate the confidence intervals.

Table A36: Share of enumeration districts where the simulated quality measures are greater than 1 (hence truncated at 1)

Quality measure	Linking algorithm	Unadjusted	Assumptions about δ		
			$\delta = 1$	$\delta = -1$	Estimated δ
Linkage rate	ABE-exact5	0.0001	0.0001	0.0001	0.0001
			(0, 0)	(0, 0)	(0, 0)
	ABE-JW5	0.0001	0.0001	0.0001	0.0001
			(0, 0)	(0, 0)	(0, 0)
	ABE-NYSIIS5	0.0001	0.0001	0.0001	0.0001
			(0, 0)	(0, 0)	(0, 0)
	ABE-exact	0.0002	0.0002	0.0002	0.0002
			(0, 0)	(0, 0)	(0, 0)
	ABE-JW	0.0002	0.0002	0.0002	0.0002
			(0, 0)	(0, 0)	(0, 0)
	ABE-NYSIIS	0.0001	0.0001	0.0001	0.0001
			(0, 0)	(0, 0)	(0, 0)
	ML	0.0002	0.0002	0.0002	0.0002
			(0, 0)	(0, 0)	(0, 0)
	MLP	0.0003	0.0003	0.0003	0.0003
			(0, 0)	(0, 0)	(0, 0)
Share validated	ABE-exact5	0.1037	0.0966	0.1032	0.1037
			(0, .181)	(.102, .105)	(.102, .106)
	ABE-JW5	0.1025	0.0912	0.1014	0.1027
			(0, .1)	(.099, .103)	(.1, .106)
	ABE-NYSIIS5	0.1023	0.0000	0.1010	0.1025
			(0, .097)	(.098, .103)	(.1, .106)
	ABE-exact	0.0638	0.0753	0.0628	0.0638
			(.061, .101)	(.061, .065)	(.062, .066)
	ABE-JW	0.0727	0.0782	0.0723	0.0728
			(.063, .125)	(.07, .074)	(.07, .076)
	ABE-NYSIIS	0.0587	0.0668	0.0577	0.0588
			(.058, .085)	(.056, .059)	(.057, .062)
	ML	0.0500	0.0576	0.0491	0.0500
			(.046, .077)	(.047, .051)	(.048, .052)
	MLP	0.0492	0.0454	0.0496	0.0492
			(0, .062)	(.048, .051)	(.047, .051)

Note: This table presents the share of enumeration districts in our sample for which the simulated quality was truncated at 1 (see (4) in the main text for the formula for simulation). The number of enumeration districts used to create this table is 120,862 for linkage rates, and between 119,937 and 120,320 for share validated, depending on the linked sample. The validation variable for this table is parents' birth places. The column labeled "Unadjusted" corresponds to simulation results using OLS estimate of β in model (3) (note that its confidence intervals are omitted, given how tight the confidence intervals for the OLS estimates of β is). The columns labeled $\delta = 1$ and $\delta = -1$ correspond to simulation results using the estimates of β assuming $\delta = 1$ or $\delta = -1$. Lastly, the column labeled "Estimated δ" corresponds to simulation results using the estimates of β when we set the value of δ equal to the one estimated using the boundary sample (see footnote 22 in the main text for details about how δ is estimated). Numbers in the parentheses are 95% bootstrap confidence intervals.

Table A37: Comparison between observed and simulated share validated of linked samples (alternative validation variable)

Link. alg.	Obs. qual.	Simulated quality			
		Unadjusted	$\delta = 1$	$\delta = -1$	Estimated δ
ABE-exact5	0.843	0.880	0.877	0.880	0.880
			(.872, .881)	(.877, .882)	(.877, .882)
ABE-JW5	0.845	0.884	0.879	0.884	0.884
			(.875, .882)	(.881, .886)	(.881, .886)
ABE-NYSIIS5	0.839	0.881	0.876	0.882	0.882
			(.873, .88)	(.878, .885)	(.878, .885)
ABE-exact	0.794	0.838	0.844	0.838	0.838
			(.839, .848)	(.835, .84)	(.835, .84)
ABE-JW	0.805	0.854	0.855	0.853	0.853
			(.851, .86)	(.848, .857)	(.848, .857)
ABE-NYSIIS	0.779	0.836	0.836	0.835	0.835
			(.831, .841)	(.832, .839)	(.832, .839)
ML	0.816	0.855	0.866	0.854	0.854
			(.861, .87)	(.85, .857)	(.85, .857)
MLP	0.876	0.905	0.909	0.905	0.905
			(.906, .911)	(.903, .906)	(.903, .906)

Note: This table presents observed share validated for each linked sample as well as simulated share validated for various assumptions about δ. The validation variable for this table is middle name initials. The column labeled "Unadjusted" presents simulated share validated obtained with OLS estimates of β in model (3) (note that its confidence intervals are omitted, given how tight the confidence intervals for the OLS estimates of β is). The columns labeled "$\delta = 1$" and "$\delta = -1$" show simulated share validated under the corresponding assumption about δ, and includes their 95% bootstrap confidence intervals. The column labeled "Estimated δ" shows share validated when δ is set at the estimated value. See footnote 22 in the main text for details about how δ is estimated.

Table A38: Share of enumeration districts where simulated quality measures are greater than 1 (hence truncated at 1) (alternative validation variable)

Quality measure	Linking algorithm	Unadjusted	Assumptions about δ		
			$\delta = 1$	$\delta = -1$	Estimated δ
Share validated	ABE-exact5	0.2879	0.2838	0.2879	0.2879
			(.279, .29)	(.284, .294)	(.284, .294)
	ABE-JW5	0.3072	0.2985	0.3079	0.3079
			(.294, .305)	(.302, .313)	(.302, .313)
	ABE-NYSIIS5	0.3058	0.2959	0.3064	0.3064
			(.291, .303)	(.301, .313)	(.301, .313)
	ABE-exact	0.1833	0.1899	0.1831	0.1831
			(.184, .196)	(.181, .186)	(.181, .186)
	ABE-JW	0.2169	0.2191	0.2164	0.2164
			(.214, .226)	(.21, .223)	(.21, .223)
	ABE-NYSIIS	0.1778	0.1787	0.1775	0.1775
			(.174, .185)	(.174, .183)	(.174, .183)
	ML	0.1618	0.1755	0.1611	0.1611
			(.169, .185)	(.158, .166)	(.158, .166)
	MLP	0.1734	0.1804	0.1732	0.1732
			(.176, .185)	(.17, .177)	(.17, .177)

Note: This table presents the share of enumeration districts in our sample for which the simulated quality was truncated at 1 (see (4) in the main text for the formula for simulation). The number of enumeration districts used to create this table is between 107,471 and 115,170, depending on the linked sample. The results for linkage rates are omitted because they are the same as presented in Table A36. The validation variable for this table is middle name initials. The column labeled "Unadjusted" corresponds to simulation results using OLS estimates of β in model (3) (note that its confidence intervals are omitted, given how tight the confidence intervals for the OLS estimates of β is). The columns labeled $\delta = 1$ and $\delta = -1$ correspond to simulation results using the estimates of β assuming $\delta = 1$ and $\delta = -1$ respectively. Lastly, the column labeled "Estimated δ" shows simulation results using estimates of β when we set the value of δ equal to the one estimated using the boundary sample (see the main text for details). Numbers in the parentheses are 95% bootstrap confidence intervals.

Table A39: The effect of legibility on share validated (long-form questionnaire respondents only)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.174	0.269	0.198	0.0503	0.283*	0.208+	0.292**	0.0902
	(0.180)	(0.186)	(0.181)	(0.136)	(0.138)	(0.119)	(0.0985)	(0.0628)
BDFE	Y	Y	Y	Y	Y	Y	Y	Y
N	7902	7564	7488	10705	9976	11800	13244	20548
adj. R^2	0.093	0.123	0.124	0.091	0.118	0.105	0.104	0.105

Note: We use only the linked observations in the boundary sample to estimate model (1). The dependent variable is validated/not validated (1/0) with the validation variable being parents' birth places. For this table, we restrict the sample to those who were administered the long-form questionnaire in the 1940 census. Covariates include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided the most (or all) of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); birth state; highest grade of schooling. All controls are directly obtained from 1940 census. BDFE refers to boundary fixed effects. Robust standard errors are in parentheses.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A40: The effect of legibility on share validated (those who did not take the long-form survey but lived with a parent in both censuses)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.0747*	0.131***	0.126***	0.0890***	0.143***	0.165***	0.106***	0.0448***
	(0.0294)	(0.0293)	(0.0300)	(0.0253)	(0.0246)	(0.0241)	(0.0208)	(0.0129)
BDFE	Y	Y	Y	Y	Y	Y	Y	Y
N	60627	58493	56552	80180	78727	86860	101960	177539
adj. R^2	0.101	0.098	0.095	0.089	0.085	0.082	0.076	0.084

Note: We use only the linked observations in the boundary sample to estimate model (1). The dependent variable is validated/not validated (1/0) with the validation variable being parents' birth places. For this table, we restrict the sample to those who were not administered the long-form questionnaire in the 1940 census but lived with a parent in both censuses, so that we have information about their parents' birth places. Covariates include: age; black (1/0); married (1/0); attended school on March 1st, 1940 (1/0); head of the household (1/0); whether you provided the most (or all) of the information about the household (1/0); whether you were temporarily absent from the household (1/0); number of alphabet letters in one's name; whether your household owns the house (1/0); whether more than one generation live in your household (1/0); birth state; highest grade of schooling. All controls are directly obtained from 1940 census. BDFE refers to boundary fixed effects. Robust standard errors are in parentheses.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A41: Comparison of observable characteristics across samples

	White or black male "Linkable pop."	Long form respondents	Obs. with parents' BPL	Obs. with middle name initials
Share non-missing	1.00	0.05	0.42	0.24
Legibility	0.71	0.71	0.71***	0.73***
Age	35.09	35.09	21.50***	37.70***
Black	0.09	0.09	0.08***	0.05***
Married	0.53	0.53	0.12***	0.61***
BPL: NE	0.07	0.07	0.10***	0.08***
MA	0.21	0.22	0.27***	0.17***
ENC	0.20	0.20	0.25***	0.18***
WNC	0.07	0.06	0.06***	0.10***
SA	0.10	0.10	0.10***	0.14***
ESC	0.04	0.04	0.03***	0.06***
WSC	0.09	0.09	0.09	0.12***
MTN	0.01	0.01	0.01	0.01***
In BPL	0.62	0.62	0.82***	0.64***
5-yr mig.	0.03	0.03	0.02***	0.04***
Head	0.50	0.50	0.09***	0.59***
Child	0.36	0.36	0.85***	0.30***
-in-law	0.02	0.02	0.00***	0.01***
Parent	0.01	0.01	0.00***	0.01
-in-law	0.01	0.01	0.00***	0.01***
Sibling	0.02	0.02	0.01***	0.01***
-in-law	0.01	0.01	0.01***	0.01***
Lives with both parents	0.28	0.28	0.68***	0.24***
w/mother	0.08	0.08	0.19***	0.07***
w/father	0.02	0.02	0.05***	0.02***
# of siblings	0.89	0.89	2.03***	0.61***
In NEast	0.36	0.36	0.39***	0.26***
In MW	0.33	0.33	0.33**	0.28***
In West	0.09	0.09	0.08***	0.13***

Note: The linkable population consists of white or black males who are 8 years old or older. We also exclude those who live in enumeration districts that contain fewer than 50 people, or where less than 90% of people have both transcriptions of their names. "Long form respondents" are those in our linkable population who took the long-form survey in the 1940 census. "Obs. with parents' BPL" are those who either took the long-form survey in the 1940 or those who lived with a parent during both 1930 and 1940 censuses. Lastly, "Obs. with middle name initials" are those in our linkable population that have non-missing values of middle name initials in the 1940 census.

Table A42: Balance of observables at the boundary of enumeration districts (observations with non-missing information about parents' birth places)

Variable	Less legible	More legible	Diff.	Std. diff.
Own a house	0.434	0.436	0.001	0.002
Multi-gen. household	0.976	0.977	0.000	0.001
Age	21.484	21.509	0.025	0.002
Black	0.075	0.077	0.002*	0.004
Married	0.125	0.123	-0.002*	-0.004
In school	0.471	0.472	0.001	0.002
# characters in full name	12.898	12.958	0.059***	0.018
Main provider of info.	0.055	0.056	0.001	0.003
Absent when enumerated	0.008	0.009	0.001***	0.010
Head of household	0.096	0.094	-0.002**	-0.005
Years of schooling	10.311	10.315	0.004	0.001
Foreign born mother	0.367	0.368	0.001	0.002
Foreign born father	0.417	0.417	-0.000	-0.000
Non-institutional residence	0.997	0.997	-0.000	-0.002
Employed	0.396	0.399	0.004**	0.005
In labor force	0.665	0.668	0.003	0.005
# weeks worked	41.064	40.816	-0.248***	-0.012
# hours per week worked	41.908	42.002	0.093	0.006
Labor income	535.914	537.101	1.187	0.001
Nonlabor income $\geq$ \$50	0.092	0.095	0.003**	0.008
Size of household	5.413	5.433	0.020**	0.006
Observations	150,409	158,558	308,967	

Note: The unit of observations in this table is a person. The observations used to create this table satisfy the following two conditions: a) they are in the boundary sample, i.e., those who live on the streets that serve as the border of the two neighboring enumeration districts; and b) they have non-missing values of parents' birth places in the 1940 census, either because they took the long-form survey, or because they lived with at least one parent in both censuses. The variables are obtained from the 1940 Census. We follow the IPUMS's classification of the occupations. See `https://usa.ipums.org/usa-action/variables/OCC1950#codes_section` for the detailed coding scheme for occupations in the U.S. census.$^{+}$ $p < 0.1$,* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$

Table A43: Heterogeneous effect of legibility on linkage rates 1

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.278***	0.189***	0.184***	0.357***	0.232***	0.245***	0.267***	0.260***
	(0.00722)	(0.00711)	(0.00715)	(0.00799)	(0.00788)	(0.00833)	(0.00848)	(0.00863)
× Black	-0.105***	0.00745	-0.0456**	-0.110***	-0.0197	-0.0708***	-0.0459*	-0.0589**
	(0.0160)	(0.0183)	(0.0165)	(0.0194)	(0.0202)	(0.0206)	(0.0222)	(0.0219)
Legibility	0.279***	0.207***	0.181***	0.365***	0.256***	0.244***	0.284***	0.259***
	(0.00723)	(0.00718)	(0.00714)	(0.00802)	(0.00797)	(0.00834)	(0.00854)	(0.00861)
× Foreign-born	-0.0563***	-0.0929***	-0.00700	-0.0895***	-0.139***	-0.0275*	-0.112***	-0.0255*
	(0.00880)	(0.00860)	(0.00903)	(0.00991)	(0.00940)	(0.0107)	(0.0106)	(0.0115)
Legibility	0.276***	0.166***	0.147***	0.381***	0.219***	0.218***	0.253***	0.249***
	(0.0143)	(0.0141)	(0.0139)	(0.0158)	(0.0158)	(0.0162)	(0.0167)	(0.0160)
× 2nd gen.	0.00475	0.0424**	0.0424**	-0.0128	0.0304^{+}	0.0409*	0.0442*	0.000929
	(0.0150)	(0.0148)	(0.0148)	(0.0163)	(0.0165)	(0.0171)	(0.0175)	(0.0169)
Legibility	0.270***	0.191***	0.181***	0.350***	0.232***	0.240***	0.265***	0.256***
	(0.00696)	(0.00692)	(0.00690)	(0.00774)	(0.00766)	(0.00807)	(0.00825)	(0.00836)
× Yrs. of sch.	0.0436***	0.0371***	0.0336***	0.0537***	0.0509***	0.0418***	0.0531***	0.0312***
	(0.00342)	(0.00347)	(0.00347)	(0.00384)	(0.00385)	(0.00411)	(0.00420)	(0.00448)
Legibility	0.263***	0.187***	0.184***	0.342***	0.230***	0.246***	0.260***	0.258***
	(0.00879)	(0.00872)	(0.00869)	(0.00981)	(0.00968)	(0.0102)	(0.0104)	(0.0105)
× White collar occ.	0.0385***	0.0256**	0.0274**	0.0384***	0.0342**	0.0252*	0.0371**	0.0259*
	(0.00976)	(0.00964)	(0.00968)	(0.0108)	(0.0106)	(0.0112)	(0.0114)	(0.0115)

Note: Each panel in this table corresponds to separate estimates of model (9) (for different socio-demographic variables). Black, foreign-born, second generation (i.e., children of immigrant fathers), and white-collar occupations are binary variables, and years of schooling is standardized (i.e., its sample average is subtracted from the the raw values and then divided by the sample standard deviation). The sample size for each panel is as follows: black (725,790), foreign-borns (725,790), second generation (234,930), years of schooling (725,790), white-collar occupations (524,024). The sample size for second generation is smaller because we use only those whose parents' birth places are non-missing. The sample size for white-collar occupations is smaller because we use only those whose occupations are not housewives, students, or unemployed. The dependent variable is linked/not linked (1/0). We use the same set of controls as model (1) (see notes under Table A33 for the list of controls). In addition, we control for the socio-economic variable that is being interacted with legibility as well. Robust standard errors are reported in the parentheses.

Table A44: Heterogeneous effect of legibility on linkage rates 2

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.269***	0.190***	0.180***	0.348***	0.230***	0.239***	0.263***	0.255***
	(0.00696)	(0.00691)	(0.00689)	(0.00774)	(0.00765)	(0.00806)	(0.00824)	(0.00836)
× Age	-0.00140	-0.00562	0.00627^{+}	-0.0120**	-0.0180***	-0.00106	-0.0166***	0.0269***
	(0.00349)	(0.00348)	(0.00350)	(0.00390)	(0.00384)	(0.00411)	(0.00417)	(0.00440)
Legibility	0.276***	0.196***	0.193***	0.355***	0.241***	0.254***	0.272***	0.267***
	(0.00831)	(0.00822)	(0.00821)	(0.00924)	(0.00910)	(0.00962)	(0.00979)	(0.00987)
× Occ. score	0.00994*	0.00572	0.00570	0.0111*	0.00652	0.00750	0.00610	0.0133*
	(0.00445)	(0.00442)	(0.00444)	(0.00493)	(0.00486)	(0.00516)	(0.00525)	(0.00532)
Legibility	0.279***	0.200***	0.193***	0.356***	0.245***	0.258***	0.276***	0.253***
	(0.00975)	(0.00969)	(0.00965)	(0.0108)	(0.0107)	(0.0112)	(0.0115)	(0.0115)
× Wage/salary inc.	0.0131*	0.00346	0.0111*	0.0124^{+}	0.00174	0.0115	0.0162*	0.0349***
	(0.00598)	(0.00488)	(0.00533)	(0.00686)	(0.00549)	(0.00710)	(0.00795)	(0.00761)

Note: Each panel in this table corresponds to separate estimates of model (9) (for different socio-demographic variables). Age, occupational scores, and wage/salary income are standardized (i.e., its sample average is subtracted from the raw values and then divided the by the sample standard deviation). The sample size for each panel is as follows: age (726,261), occupational scores (524,024) and wage/salary income (389,230). The sample size for occupational scores is smaller because we use only those whose occupational scores are not coded as "N/A". The sample size for wage/salary income is smaller because we use only those whose wage/salary income is not missing and who are between the ages of 20 and 50. The dependent variable is linked/not linked (1/0). We use the same set of controls as model (1) (see notes under Table A33 for the list of controls). In addition, we control for the socio-economic variable that is being interacted with legibility as well. Robust standard errors are reported in the parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, ***

Table A45: Heterogeneous effects of legibility on linkage rates (alternative legibility measure 1)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.283***	0.193***	0.188***	0.364***	0.237***	0.250***	0.274***	0.265***
	(0.00734)	(0.00723)	(0.00726)	(0.00812)	(0.00800)	(0.00846)	(0.00861)	(0.00876)
× Black	-0.106***	0.00676	-0.0478**	-0.112***	-0.0201	-0.0726***	-0.0480*	-0.0576**
	(0.0161)	(0.0185)	(0.0166)	(0.0196)	(0.0203)	(0.0207)	(0.0224)	(0.0221)
Legibility	0.285***	0.211***	0.186***	0.372***	0.262***	0.250***	0.291***	0.264***
	(0.00734)	(0.00730)	(0.00725)	(0.00814)	(0.00809)	(0.00846)	(0.00867)	(0.00874)
× Foreign-born	-0.0575***	-0.0949***	-0.00881	-0.0912***	-0.141***	-0.0298**	-0.114***	-0.0254*
	(0.00890)	(0.00869)	(0.00913)	(0.0100)	(0.00950)	(0.0108)	(0.0107)	(0.0116)
Legibility	0.280***	0.169***	0.149***	0.387***	0.225***	0.222***	0.259***	0.252***
	(0.0146)	(0.0144)	(0.0142)	(0.0160)	(0.0160)	(0.0165)	(0.0169)	(0.0162)
× 2nd gen.	0.00817	0.0442**	0.0436**	-0.00971	0.0319+	0.0410*	0.0466**	0.00249
	(0.0152)	(0.0150)	(0.0150)	(0.0165)	(0.0167)	(0.0173)	(0.0177)	(0.0171)
Legibility	0.275***	0.195***	0.185***	0.356***	0.237***	0.245***	0.271***	0.261***
	(0.00707)	(0.00703)	(0.00701)	(0.00786)	(0.00778)	(0.00819)	(0.00837)	(0.00849)
× Yrs. of sch.	0.0435***	0.0376***	0.0342***	0.0536***	0.0516***	0.0423***	0.0540***	0.0308***
	(0.00345)	(0.00351)	(0.00350)	(0.00387)	(0.00389)	(0.00415)	(0.00424)	(0.00452)
Legibility	0.269***	0.191***	0.188***	0.348***	0.235***	0.251***	0.266***	0.263***
	(0.00892)	(0.00885)	(0.00882)	(0.00995)	(0.00983)	(0.0104)	(0.0106)	(0.0107)
× White collar occ.	0.0386***	0.0263**	0.0276**	0.0394***	0.0349**	0.0259*	0.0388***	0.0271*
	(0.00984)	(0.00973)	(0.00976)	(0.0109)	(0.0107)	(0.0113)	(0.0115)	(0.0116)

Note: The alternative legibility measure used for this table is based on the name-cleaning procedure where we do not remove spaces between the letters of a name. Each panel in this table corresponds to separate estimates of model (9) (for different socio-demographic variables). Black, foreign-born, second generation (i.e., children of immigrant fathers), and white-collar occupations are binary variables, and years of schooling is standardized (i.e., its sample average is subtracted from raw values and then divided by the sample standard deviation). The sample size for each panel is as follows: black (725,790), foreign-borns (725,790), second generation (234,930), years of schooling (725,790), white-collar occupations (524,024). The sample size for second generation is smaller because we use only those whose parents' birth places are non-missing. The sample size for white-collar occupations is smaller because we use those whose occupations are not housewives, students, or unemployed. The dependent variable is linked/not linked (1/0). We use the same set of controls as model (1) (see notes under Table A33 for the list of controls). In addition, we control for the socio-economic variable that is being interacted with legibility as well. Robust standard errors are reported in the parentheses.

Table A46: Heterogeneous effect of legibility on linkage rates (alternative legibility measure 2)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.289***	0.197***	0.191***	0.372***	0.242***	0.254***	0.280***	0.273***
	(0.00753)	(0.00741)	(0.00745)	(0.00834)	(0.00822)	(0.00869)	(0.00884)	(0.00902)
× Black	-0.109***	0.00860	-0.0473**	-0.115***	-0.0205	-0.0743***	-0.0478*	-0.0693**
	(0.0168)	(0.0192)	(0.0173)	(0.0203)	(0.0212)	(0.0216)	(0.0233)	(0.0230)
Legibility	0.292***	0.216***	0.188***	0.381***	0.268***	0.253***	0.299***	0.273***
	(0.00754)	(0.00750)	(0.00746)	(0.00838)	(0.00832)	(0.00871)	(0.00893)	(0.00901)
× Foreign-born	-0.0597***	-0.0977***	-0.00665	-0.0939***	-0.146***	-0.0290**	-0.119***	-0.0276*
	(0.00915)	(0.00895)	(0.00939)	(0.0103)	(0.00979)	(0.0112)	(0.0110)	(0.0120)
Legibility	0.290***	0.177***	0.155***	0.400***	0.234***	0.229***	0.269***	0.264***
	(0.0150)	(0.0148)	(0.0146)	(0.0165)	(0.0165)	(0.0170)	(0.0175)	(0.0168)
× 2nd gen.	0.00457	0.0434**	0.0426**	-0.0144	0.0296+	0.0371*	0.0435*	-0.00233
	(0.0156)	(0.0154)	(0.0154)	(0.0170)	(0.0172)	(0.0179)	(0.0183)	(0.0177)
Legibility	0.282***	0.199***	0.188***	0.365***	0.242***	0.249***	0.278***	0.269***
	(0.00727)	(0.00722)	(0.00720)	(0.00808)	(0.00800)	(0.00843)	(0.00861)	(0.00874)
× Yrs. of sch.	0.0461***	0.0390***	0.0352***	0.0569***	0.0536***	0.0442***	0.0561***	0.0332***
	(0.00357)	(0.00362)	(0.00361)	(0.00401)	(0.00402)	(0.00429)	(0.00439)	(0.00468)
Legibility	0.274***	0.194***	0.190***	0.356***	0.239***	0.255***	0.273***	0.270***
	(0.00917)	(0.00909)	(0.00906)	(0.0102)	(0.0101)	(0.0107)	(0.0109)	(0.0110)
× White collar occ.	0.0416***	0.0270**	0.0288**	0.0417***	0.0358**	0.0267*	0.0389**	0.0294*
	(0.0102)	(0.0101)	(0.0101)	(0.0113)	(0.0111)	(0.0117)	(0.0119)	(0.0121)

Note: The alternative legibility measure used for this table is based on a different criteria for declaring two transcriptions of the same person's name as not identical. Specifically, we declare two transcriptions to *not* be identical only if the Jaro-Winkler distance between the two transcriptions is greater than 0.044 (which is the 75th percentile of the population distribution of Jaro-Winkler distance between the two transcriptions). Each panel in this table corresponds to separate estimates of model (9) for each socio-demographic variables. Black, foreign-born, second generation (i.e., children of immigrant fathers), and white-collar occupations are binary variables, and years of schooling is standardized (i.e., its sample average is subtracted from the raw variable and then we divide the difference by the sample standard deviation). The sample size for each panel is as follows: black (725,790), foreign-borns (725,790), second generation (234,930), years of schooling (725,790), white-collar occupations (524,024). The sample size for second generation is smaller because we used for estimation only those whose parents' birth places are non-missing. The sample size for white-colloar occupations is smaller because we used for estimation only those whose occupations are not housewives, students, or unemployed. The dependent variable is linked/not linked (1/0). We use the same set of controls as model (1) (see notes under Table A33 for the list of controls). In addition, we control for the socio-economic variable that is being interacted with legibility as well. Robust standard errors are reported in the parentheses.

Table A47: Heterogeneous effects of legibility on share validated 1

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.0598*	0.125***	0.123***	0.0833***	0.141***	0.168***	0.113***	0.0454***
	(0.0279)	(0.0282)	(0.0286)	(0.0240)	(0.0238)	(0.0230)	(0.0200)	(0.0127)
× Black	0.191	-0.130	0.0259	0.0162	-0.0895	0.0109	0.0425	0.0579
	(0.116)	(0.0939)	(0.116)	(0.0954)	(0.0815)	(0.0871)	(0.0703)	(0.0518)
Legibility	0.0705*	0.121***	0.132***	0.0886***	0.140***	0.175***	0.121***	0.0499***
	(0.0279)	(0.0279)	(0.0286)	(0.0240)	(0.0235)	(0.0229)	(0.0199)	(0.0126)
× Foreign-born	-0.0763	-0.146*	-0.174**	-0.111*	-0.125*	-0.142**	-0.128**	-0.0444
	(0.0637)	(0.0612)	(0.0588)	(0.0528)	(0.0547)	(0.0463)	(0.0435)	(0.0283)
Legibility	0.115**	0.190***	0.176***	0.0997**	0.183***	0.188***	0.168***	0.0825***
	(0.0358)	(0.0368)	(0.0380)	(0.0308)	(0.0306)	(0.0299)	(0.0260)	(0.0166)
× 2nd gen.	-0.108**	-0.130**	-0.131**	-0.0657+	-0.0950**	-0.0675*	-0.113***	-0.0917***
	(0.0408)	(0.0412)	(0.0423)	(0.0353)	(0.0343)	(0.0332)	(0.0289)	(0.0182)
Legibility	0.0601*	0.111***	0.123***	0.0771**	0.130***	0.168***	0.116***	0.0483***
	(0.0279)	(0.0279)	(0.0286)	(0.0240)	(0.0236)	(0.0229)	(0.0198)	(0.0126)
× Yrs. of sch.	0.0313+	0.0241	0.00212	0.0315*	0.0276+	0.00480	0.000715	-0.00152
	(0.0170)	(0.0170)	(0.0175)	(0.0149)	(0.0148)	(0.0141)	(0.0125)	(0.00791)
Legibility	-0.0399	0.0609	0.0371	0.00585	0.0954*	0.0944**	0.0348	0.00927
	(0.0435)	(0.0456)	(0.0459)	(0.0371)	(0.0373)	(0.0354)	(0.0309)	(0.0192)
× White collar occ.	0.0913*	0.0428	0.0424	0.0534	0.0602	-0.0222	0.0573+	0.0354+
	(0.0440)	(0.0463)	(0.0470)	(0.0375)	(0.0387)	(0.0363)	(0.0322)	(0.0201)

Note: Each panel in this table corresponds to separate estimates of model (9) (for different socio-demographic variables). Black, foreign-born, second generation (i.e., children of immigrant fathers), and white-collar occupations are binary variables, and years of schooling is standardized (i.e., its sample average is subtracted from raw values and then divided by the sample standard deviation). The sample size varies across columns (i.e., across linking algorithms) because only linked observations are used for estimation. The dependent variable is validated/not validated (1/0). We use the same set of controls as model (1) (see notes under Table A33 for the list of controls). In addition, we control for the socio-economic variable that is being interacted with legibility as well. Robust standard errors are reported in the parentheses.

Table A48: Heterogeneous effect of legibility on share validated 2

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.0600^{+}	0.0930**	0.0887**	0.0713*	0.112***	0.140***	0.0931***	0.0322*
	(0.0330)	(0.0333)	(0.0338)	(0.0282)	(0.0285)	(0.0269)	(0.0239)	(0.0157)
× Age	-0.0100	-0.0307	-0.0486*	-0.0178	-0.0324	-0.0401*	-0.0308^{+}	-0.0211^{+}
	(0.0239)	(0.0239)	(0.0244)	(0.0203)	(0.0209)	(0.0194)	(0.0176)	(0.0119)
Legibility	-0.000347	0.0761^{+}	0.0514	0.0319	0.119***	0.0889**	0.0536^{+}	0.0217
	(0.0410)	(0.0427)	(0.0433)	(0.0351)	(0.0353)	(0.0333)	(0.0292)	(0.0184)
× Occ. score	0.0307	-0.000318	-0.00671	0.0327^{+}	0.00894	0.0126	-0.00642	0.0000722
	(0.0227)	(0.0234)	(0.0235)	(0.0188)	(0.0200)	(0.0182)	(0.0164)	(0.0101)
Legibility	0.00520	0.0722	0.0303	0.0593	0.134**	0.113**	0.0662^{+}	0.0310
	(0.0492)	(0.0519)	(0.0508)	(0.0413)	(0.0418)	(0.0391)	(0.0341)	(0.0213)
× Wage/salary inc.	0.0496	0.0139	0.0341	0.0536^{+}	0.00121	0.0400	0.00967	0.00996
	(0.0344)	(0.0357)	(0.0365)	(0.0293)	(0.0293)	(0.0278)	(0.0243)	(0.0151)

Each panel in this table corresponds to separate estimates of model (9) (for different socio-demographic variables). Age, occupational scores, and wage/salary income are standardized (i.e., its sample average is subtracted from raw values and then divided by the sample standard deviation). The sample size varies across columns (i.e., across linking algorithms) because only linked observations are used for estimation. The dependent variable is validated/not validated (1/0). We use the same set of controls as model (1) (see notes under Table A33 for the list of controls). In addition, we control for the socio-economic variable that is being interacted with legibility as well. Robust standard errors are reported in the parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, ***

Table A49: Heterogeneous effect of legibility on share validated (alternative validation variable)

	ABE-exact5	ABE-JW5	ABE-NYSIIS5	ABE-exact	ABE-JW	ABE-NYSIIS	ML	MLP
Legibility	0.207***	0.287***	0.255***	0.220***	0.288***	0.298***	0.221***	0.0994***
	(0.0577)	(0.0598)	(0.0610)	(0.0494)	(0.0511)	(0.0493)	(0.0375)	(0.0259)
× Foreign-born	-0.0202	0.0747	0.0233	0.00965	0.129	0.0306	0.0720	-0.0822+
	(0.117)	(0.168)	(0.132)	(0.0984)	(0.155)	(0.103)	(0.0982)	(0.0467)
Legibility	0.0886	0.256	0.350+	0.0783	0.287*	0.314*	-0.0115	0.0646
	(0.168)	(0.178)	(0.185)	(0.133)	(0.136)	(0.127)	(0.0909)	(0.0652)
× 2nd gen.	0.0573	-0.0931	-0.0249	-0.0290	-0.0686	0.208	-0.0124	-0.00750
	(0.240)	(0.287)	(0.280)	(0.185)	(0.215)	(0.187)	(0.137)	(0.0908)

Note: Each panel in this table corresponds to separate estimates of model (9) (for different socio-demographic variables). In this table we only include the results for foreign-born and second generation (i.e., children of immigrant fathers) individuals because legibility appears to have heterogeneous effect for these two groups (see Table A47). The sample size varies across columns (i.e., across linking algorithms) because only linked observations are used for estimation. The validation variable for this table is middle name initials. The dependent variable is validated/not validated (1/0). We use the same set of controls as model (1) (see notes under Table A33 for the list of controls). In addition, we control for the socio-economic variable that is being interacted with legibility as well. Robust standard errors are reported in the parentheses.