An Assessment of CenSoc Match Quality[*]

Casey Breen [†]   Maria Osborne [‡]

Draft Version: June 10, 2022

Abstract

The CenSoc datasets link individual-level 1940 Census records to Social Security death records using deterministic record linkage algorithms. In this technical report, we describe our record linkage methodology and assess the accuracy and representativeness of the CenSoc Version 2.1 matches. The main takeaways of this report are:

1. The CenSoc-DMF and CenSoc-Numident datasets are comprised of individuals that are broadly representative of the general population but slightly skewed towards higher socioeconomic status individuals (e.g., 35.2% of individuals in CenSoc-DMF vs. 32.5% of individuals in the general population completed high school). Black people are underrepresented in both datasets, comprising 9.6% of the general population but only 4.8% of CenSoc-DMF and 6.5% of CenSoc-Numident. However, the Black samples are broadly representative of the general Black population. Non-representativeness has the potential to bias estimates if the outcome of interest is heterogeneous across the under or over-represented population subgroups. To account for this, researchers can stratify for covariates such as race and education in their analysis.
2. The overall mortality-adjusted match rate for the CenSoc-DMF is 30% (18% for our set of conservative matches), while the overall mortality-adjusted match rate for CenSoc-Numident is approximately 30% for men (22% conservative) and 32% for women (24% conservative). The match rate for Censoc-Numident is lower for earlier birth cohorts (1895-1915) because of the higher rates of missingness of birthplace, a required matching field.
3. For both datasets, restricting to conservative matches reduces sample size but increases the quality of the matches. The conservative matches are comparably representative of the general population but contain fewer false matches than the standard matches. False matches introduce measurement error resulting in attenuated estimates within a regression framework. We generally recommend researchers restrict to conservative matches to avoid this attenuation bias.
4. For analyses of multiple birth cohorts, we recommend including birth cohort fixed effects. Birth cohort fixed-effects control for each birth cohort being observed for a different window of ages of death and the potential sample composition bias introduced by differential match rates across birth cohorts in CenSoc-Numident.

[*] For helpful discussions and feedback the authors thank Joshua R. Goldstein, Leora Lawton, Ryan Edwards and Carl Boe. We thank Kathy Perez for research assistance. All replication code is available on https://github.com/caseybreen/censocdev.

[†] Department of Demography, University of California, Berkeley. caseybreen@berkeley.edu.

[‡] Department of Demography, University of California, Berkeley mariaosborne@berkeley.edu.

Contents

1 Overview

The CenSoc datasets – so termed because they link the full-count 1940 Census ("Cen") with Social Security Administration mortality records ("Soc") – are a publicly available administrative data resource for researchers studying mortality. These individual-level datasets provide researchers access to millions of mortality records with rich sociodemographic covariates. In this technical report, we assess the accuracy and representativeness of the CenSoc matches.

This report proceeds as follows. In Section 2, we provide background on the ABE record linkage algorithm we used to link the 1940 Census to mortality records. Section 3 presents the raw and mortality-adjusted match rate of the CenSoc datasets, and Section 4 assesses the accuracy and representativeness of the CenSoc matches. We conclude in Section 5 with a discussion of considerations and best practices for researchers using the CenSoc dataset.

2 Background

The CenSoc project disseminates two different datasets linking the 1940 Census to Social Security mortality records (Goldstein et al., 2021). The first is the CenSoc-DMF dataset, which links the 1940 census to the Death Master File (DMF), a collection of over 83 million death records reported to the Social Security Administration. This file includes only men, as surname changes during marriage preclude the accurate linkage of women. The second is the CenSoc-Numident dataset, which links the 1940 Census to the Social Security Numident records publicly available from the National Archives and Records Administration. Table 1 shows the key features of both datasets.

	CenSoc-DMF	**CenSoc-Numident**
Sex	Men-Only	Men and Women
1940 Census Covariates	Yes	Yes
High Coverage of Deaths	1975-2005	1988-2005
Size (Standard)	7.8 Million	9.4 Million
Size (Conservative)	4.7 Million	7.0 Million

Table 1: Summary of key features of CenSoc datasets

2.1 ABE Linking Algorithm

CenSoc Version 2.1 links the 1940 Census to Social Security mortality records using the ABE exact record linkage algorithm (Abramitzky, Boustan and Eriksson, 2012, 2014; Abramitzky et al., 2019). This linking strategy requires an exact match on first name, last name, and place of birth, while allowing ±2 years flexibility on year of birth. The specific steps of our implementation of this algorithm are:

1. Perform a series of steps to clean names, including removing common titles (e.g., Dr.), name standardization (e.g., Billy to William), and removing non-alphabetic characters such as dashes.

2. Restrict the 1940 Census to people unique by first name, last name, and year of birth (and place of birth in CenSoc-Numident).

3. For each record in the 1940 Census, try to find a Social Security death record that agrees on (1) first name, (2) last name, and (3) exact birth year (and exact match on state of birth in CenSoc-Numident).

    (a) If there is one and only one match, declare this pair of records to be a match.

    (b) If there are several potential matches that match exactly on year of birth, the match is discarded.

    (c) If there are no matches, the algorithm expands its search to allow flexibility on birth year. Specifically, it look for matches ± 1 year of reported birth. If there is one and only one match, declare this pair of records to be a match. If there is more than one match, discard this record. If there are no matches, then repeat this process a final time for ± 2 years of reported birth.

Table 2 shows a stylized illustration of the ABE record linkage algorithm.

2.2 Conservative Matches

After establishing the standard matches, we establish a set of "conservative" matches. The conservative matches are a subset of the standard matches; every conservative match is also a standard match. The conservative variant requires first and last name to be unique within ± 2 years around year of birth (a 5-year band) within a given state (or for CenSoc-DMF, at the national level).

2.3 Matching methods for women

For women, surname changes during marriage present a challenge for record linkage. To address this, we first identify marital status in the 1940 Census. For ever-married women, we link using last name in the Numident, exactly as we would men. For never-married women, we use father's last name from the Numident as a proxy for the surname a woman was assigned at birth (and reported in the 1940 Census), allowing for the linkage of women never-married in 1940. We are not able to link women in the CenSoc-DMF because parents' last names are not available in the DMF.

Datasets

1940 Census

ID	Raw Name	Cleaned Name	Birth Year
1A	Stewie Smith	Stewart Smith	1910
2A	Ben Lawson	Benjamin Lawson	1914
3A	James Johnson	James Johnson	1917

Death Master File (DMF)

ID	Raw Name	Cleaned Name	Birth Year
1B	Stew Smith	Stewart Smith	1911
2B	Benjamin Lawson	Benjamin Lawson	1915
3B	Ben Lawson	Benjamin Lawson	1915
4B	James Johnson	James Johnson	1917
5B	Jimmy Johnson	James Johnson	1919

ABE Matches

Standard Variant

Cleaned Name	Established Match
Stewart Smith	1A ↔ 1B
James Johnson	3A ↔ 4B

Conservative Variant

Cleaned Name	Established Match
Stewart Smith	1A ↔ 1B

Table 2: **Stylized illustration of ABE record linkage algorithm.** The ABE linkage algorithm established a match for "Stewart Smith" because there was an exact match on first name, last name, and a ± 1 difference on year of birth. Additionally, this was deemed a conservative match because the name is unique within a 5-year band (± 2 years) in both the 1940 Census and DMF. A match was established for "James Johnson" because there was one and only one exact match on first name, last name, and exact year of birth. However, this was not deemed a conservative match because "James Johnson" is not a unique name within a 5-year window in the DMF. No match was established for "Benjamin Lawson" because there were two potential matches in the DMF.

3 Match Rate

We define the raw match rate M_{raw} as the proportion of individuals observed in the 1940 Census[1] successfully linked to mortality records:

$$M_{raw} = \frac{\text{Number Established Matches}}{\text{Number of Records in 1940 Census}} \tag{1}$$

The raw match rate does not take into account mortality. Adjusting for mortality gives a better sense of the match rate conditional on someone dying during our doubly-truncated mortality observation window (1975-2005 for CenSoc-DMF and 1988-2005 for CenSoc-Numident). We define the mortality-adjusted match rate $M_{adjusted}$ to be

$$M_{adjusted} = \underbrace{\left(\frac{\text{Number Established Matches}}{\text{Number of Records in 1940 Census}}\right)}_{\text{Raw mortality rate}} \times \underbrace{\left(\frac{1}{P(\text{Dying in window})}\right)}_{\text{Adjustment factor for mortality}}, \tag{2}$$

where $P(\text{Dying in window})$ is the probability that someone dies in the mortality observation window conditional on living until 1940. Formally, this can be expressed as:

$$P(\text{Dying in window}) = P\left(\theta_l \leq D_{year} \leq \theta_r | D_{year} > 1940\right) \tag{3}$$

where θ_l is the year of left truncation, θ_r is the year of right truncation, and D_{year} is the year of death. We calculate the probability that someone dies in our observation window separately for each birth cohort using data from the Human Mortality Database (HMD, 2021). These probabilities are shown in Figure 1; see Section A for full mathematical details.

While the mortality-adjusted match rate gives a better sense of match rate, it doesn't fully account for (1) emigration or (2) people enumerated in the 1940 Census who are not captured in mortality records (e.g., people who never received a Social Security number).

[1] We also estimate match rates in Section B.1 using an alternative denominator, the mortality data files (DMF or Numident). The match rates calculated using the alternative denominator are highly comparable.

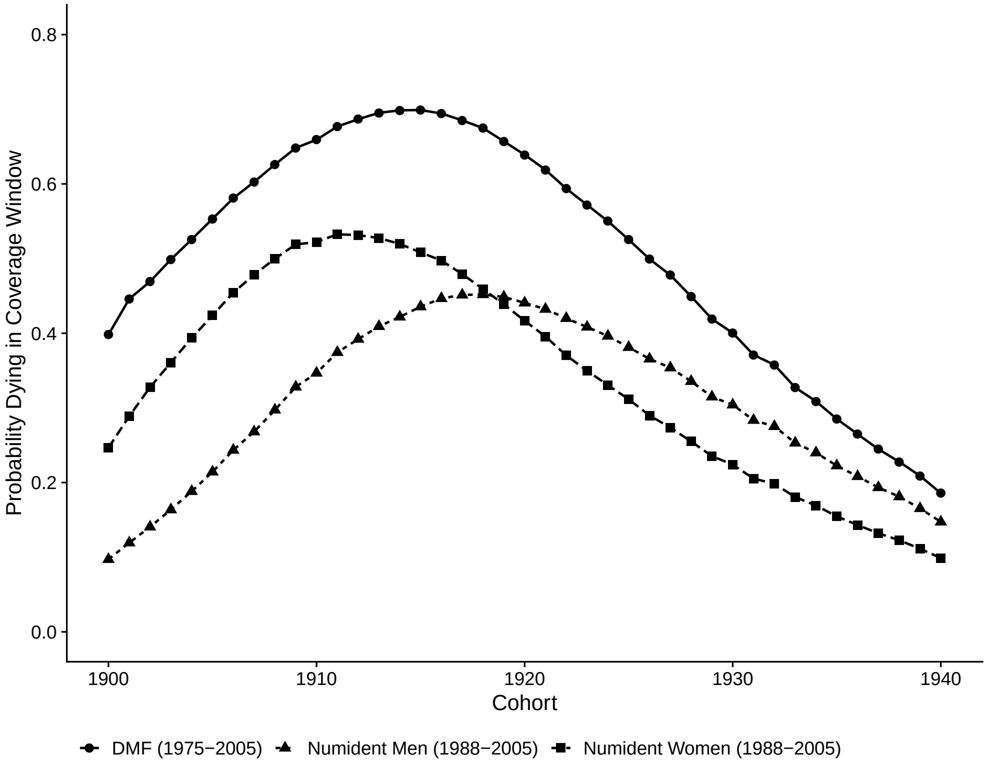

Figure 1: **Probability of dying in the mortality coverage window.** Circles show the probability that a man observed in the 1940 Census dies during the mortality coverage window for the CenSoc-DMF (1975-2005). Triangles show the probability that a man observed in the 1940 Census dies in the mortality coverage window for the CenSoc-Numident (1988-2005). Squares show the probability that a woman observed in the 1940 Census dies in the mortality coverage winow for the CenSoc-Numident (1988-2005). The probability of dying in the DMF mortality coverage window is higher than the probability of dying in the Numident coverage window for all birth cohorts because the DMF includes a wider window of deaths.

3.1 CenSoc-DMF Match Rate

Figure 2, panel (a) shows the raw match rate for the CenSoc-DMF, calculated separately for each birth cohort and linkage variant (standard and conservative). The raw match rate peaks at 19.7% for the birth cohort of 1913, and declines below 10% for birth cohorts after 1930. The mortality-adjusted match rate for CenSoc-DMF is shown in Figure 2, panel (b). The mortality-adjusted match rate for the standard variant is relatively stable around 30%, while the conservative mortality-adjusted match rate is relatively stable around 20%.

3.2 CenSoc-Numident Match Rate

We calculate match rates separately for men and women in the CenSoc-Numident. For both genders, there is a sharp uptick in match rate beginning in 1910 due to the increased availability of birthplace information in the Numident, which is a required matching field; we do not attempt to link individuals with a missing birthplace. Figure 5 shows birthplace was available for less than 25% of men born prior to 1910.

For later birth cohorts, the mortality-adjusted birth rate for the standard sample is over 40%, and the mortality-adjusted birth rate for the conservative sample is over 30%. These match rates are approximately 10% higher than the CenSoc-DMF match rates. The higher match rate is achieved because birthplace is used as an additional matching field in CenSoc-Numident, reducing number of records discarded because they have several different potential matches.

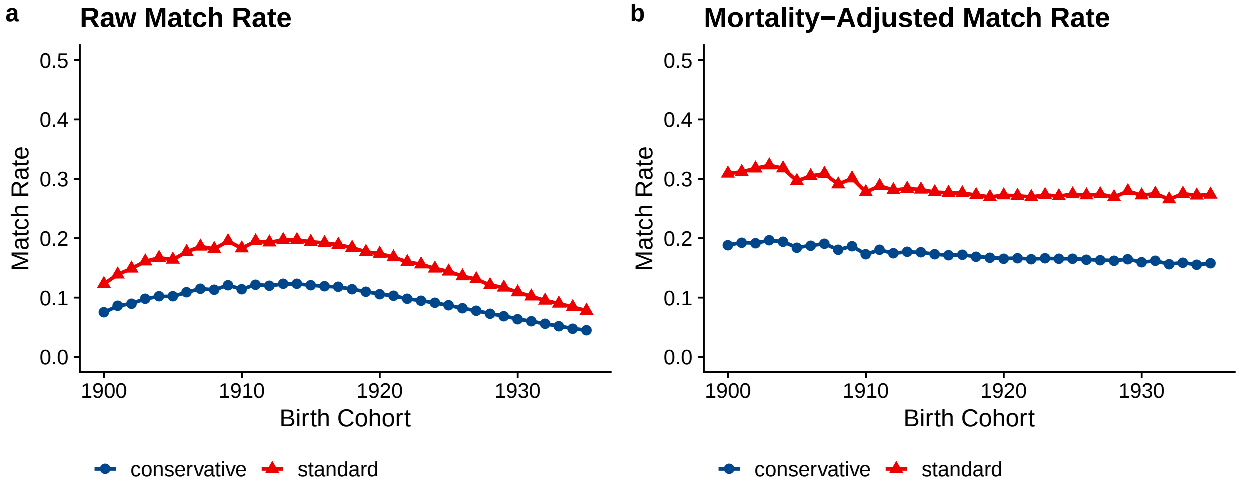

Figure 2: **CenSoc-DMF Match Rates.** Panel (a) shows the raw match rate and Panel (b) shows the mortality-adjusted match rate.

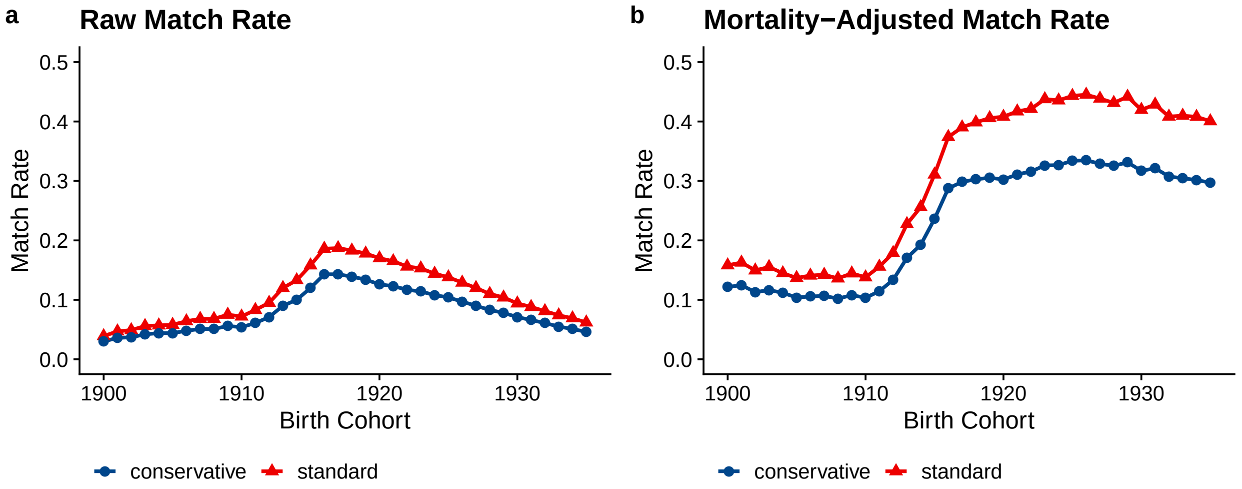

Figure 3: **CenSoc-Numident Match Rates for women.** Panel (a) shows the raw match rate and panel (b) shows the mortality-adjusted match rate.

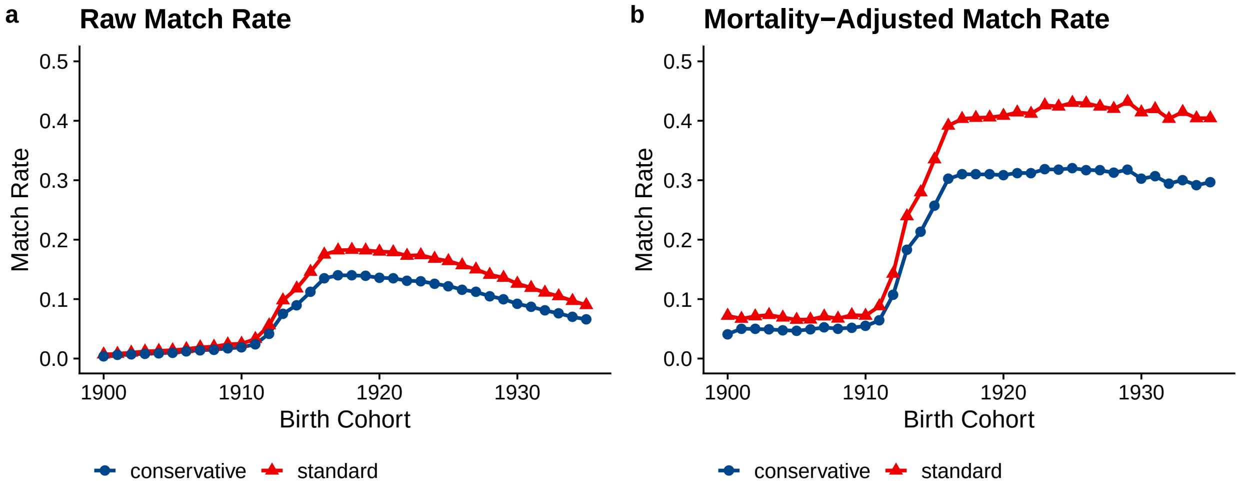

Figure 4: **CenSoc-Numident Match Rates for men.** Panel (a) shows the raw match rate and panel (b) shows the mortality-adjusted match rate.

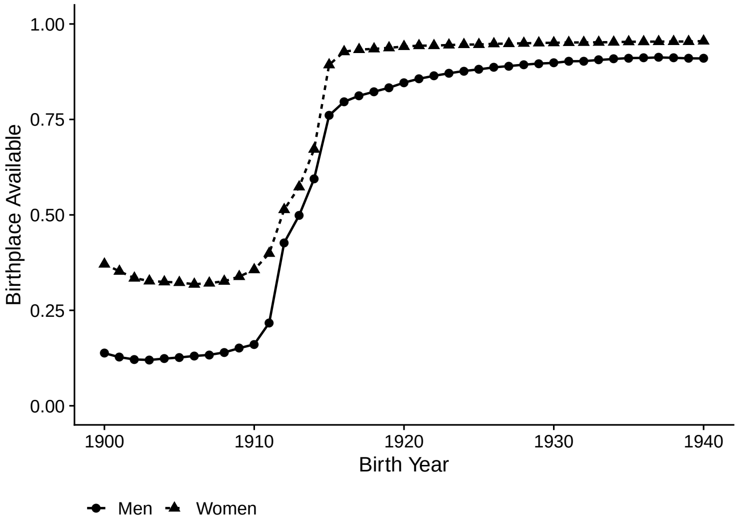

Figure 5: **CenSoc-Numident birthplace availability**.

4 Match Quality

Match quality is generally characterized by the false matches (Type I error) and missed matches (Type II error). Missed matches can lead to a selection bias – that is, the characteristics of the matched population differ systematically from the unmatched population. This lack of representativeness can often be measured and largely corrected for using weighting methods (Ruggles, Fitch and Roberts, 2018). However, false matches present more pressing challenges, introducing systematic error into inference. For instance, false matches will dramatically upwardly bias estimates of migration rates and socioeconomic mobility. The general recommendation is to prioritize minimizing the number of false matches over maximizing the overall match rate (Ruggles, Fitch and Roberts, 2018; Bailey et al., 2020).

4.1 Representativeness of matches

To assess how representative our CenSoc matches are of the general population, we compare the socioeconomic characteristics of individuals enumerated in the 1940 Census who were matched and unmatched. Figure 6 shows that the socioeconomic characteristics of individuals in the CenSoc-DMF align closely with the general population, albeit having slightly higher socioeconomic status. However, Black Americans are significantly underrepresented.[2] Similarly, Figure 7 and Figure 8 show that individuals in the CenSoc-Numident tend to be higher socioeconomic status than the general population, and Black Americans are underrepresented.

Table 3, Table 4, and Table 5 show the representativeness of the CenSoc matches for the pooled birth cohorts of 1900-1920. Similar to the age-specific analysis, these tables show that the matched population is similar in composition to the 1940 population, but tends to be Whiter and higher socioeconomic status. For the CenSoc-Numident, the differential match rates by birth cohort can lead to large relative differences in the composition of the pooled sample. For example, the relative proportion of married men is much higher in the 1940 Census than in the CenSoc-Numident, reflecting the lower match rate for older cohorts

[2] However, despite the lower match rate, the sociodemographic characteristics of the Black people successfully matched align closely with the general Black population. See Section B.3 for details.

that would be more likely to be married. To address this issue, researchers can use birth year fixed effects in a regression models to help address compositional differences related to differential match rate for different birth cohorts.

4.2 Implications for Differential Linkage for Inference

The representativeness of the matches has implications for inference. Specifically, if the under or over-represented population subgroups are also heterogeneous on the outcome of interest, this may lead to biased estimates of population-level parameters in linked samples (Bailey et al., 2020). To address this, researchers can conduct stratified analyses (e.g., fit separate models for Black and White subgroups). However, the errors introduced by sample non-representativeness are generally modest compared to errors introduced by false matches (Bailey et al., 2020).

One limitation to this approach of comparing the socioeconomic characteristics of matched and unmatched individuals enumerated in the 1940 Census is differential mortality: some subgroups may be more or less likely to die within our mortality observation window. This is a larger consideration for the CenSoc-Numident, with its narrower mortality coverage window, than with the CenSoc-DMF. The extent to which these compositional differences are driven by differential mortality is an open area of investigation.

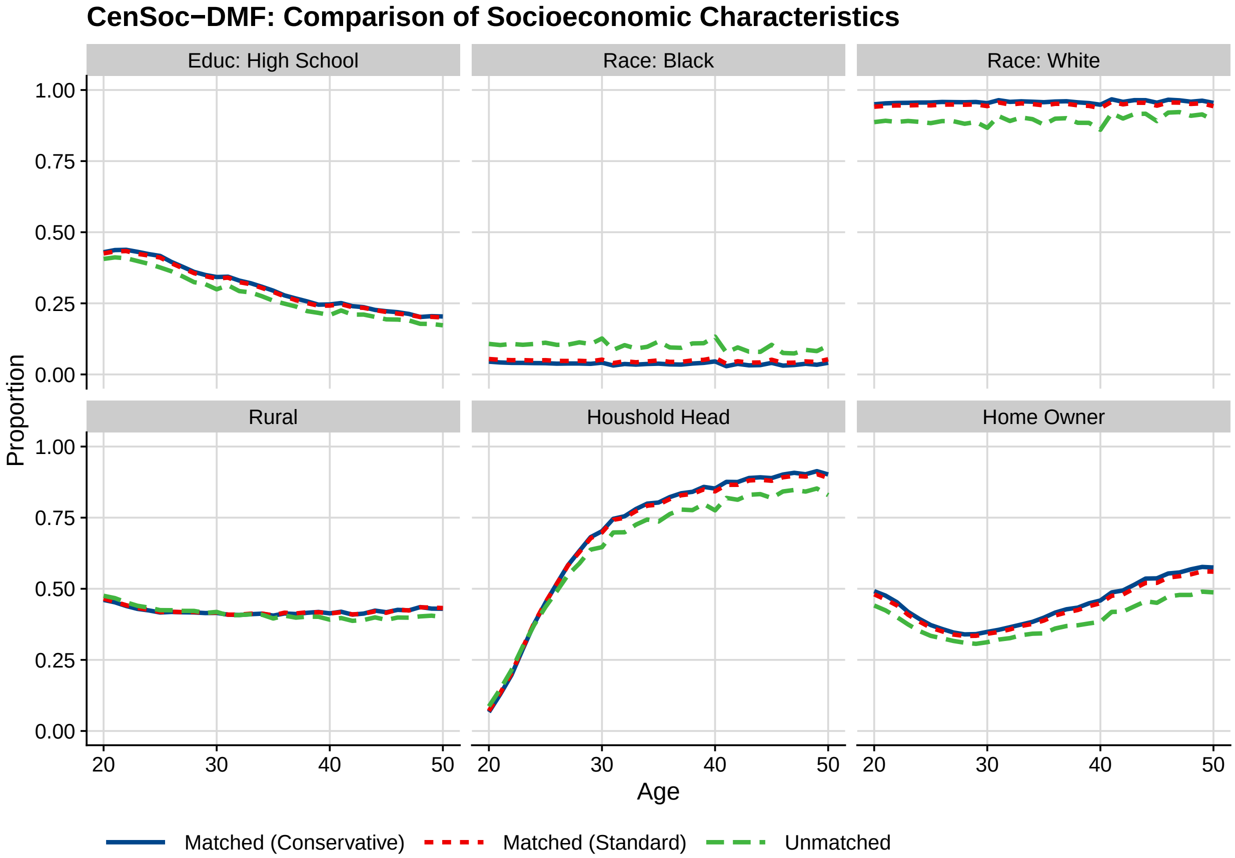

Figure 6: For each panel, lines show the proportion of men with a given socioeconomic characteristics by census age who were not matched to the DMF (green line), matched with the standard algorithm (red line), and matched with the conservative algorithm (blue line).

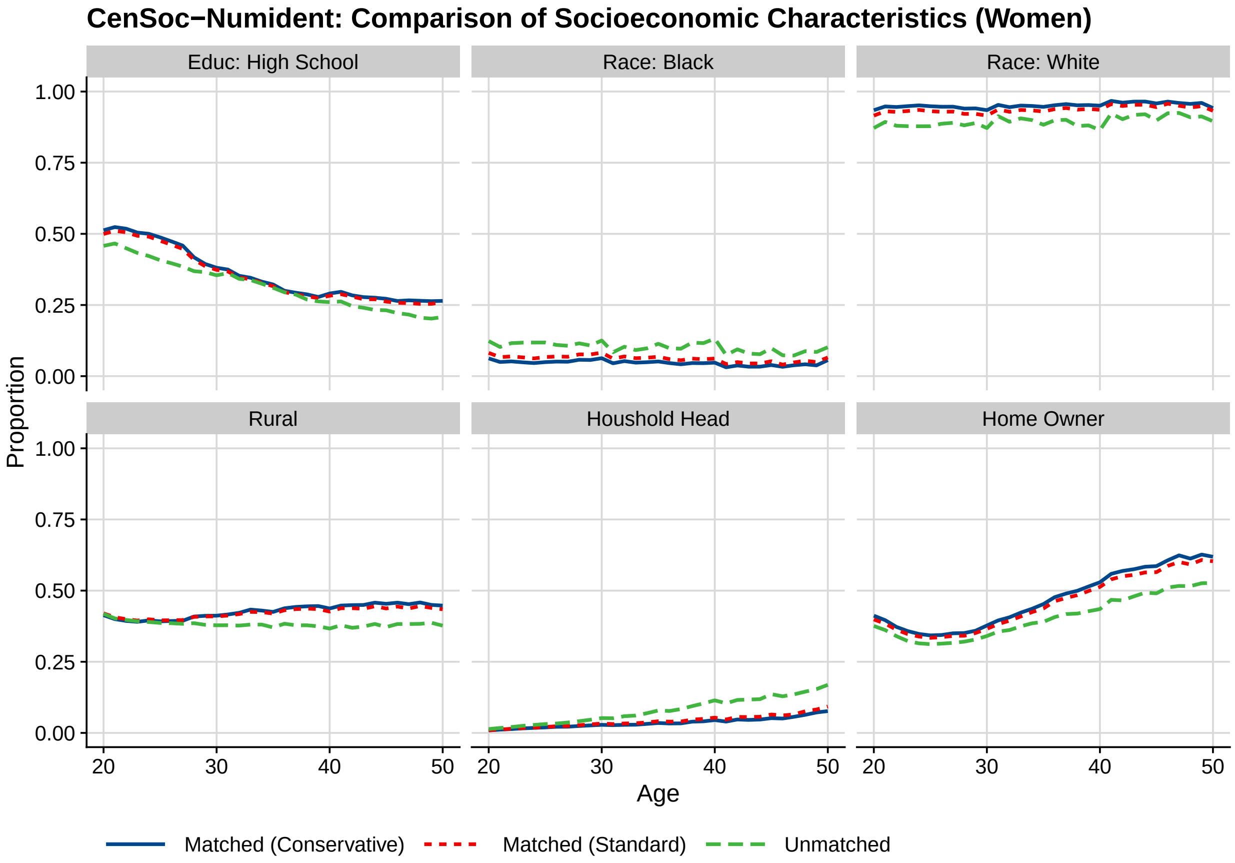

Figure 7: For each panel, lines show the proportion of women with a given socioeconomic characteristics by census age who were not matched to the Numident (green line), matched with the standard algorithm (red line), and matched with the conservative algorithm (blue line).

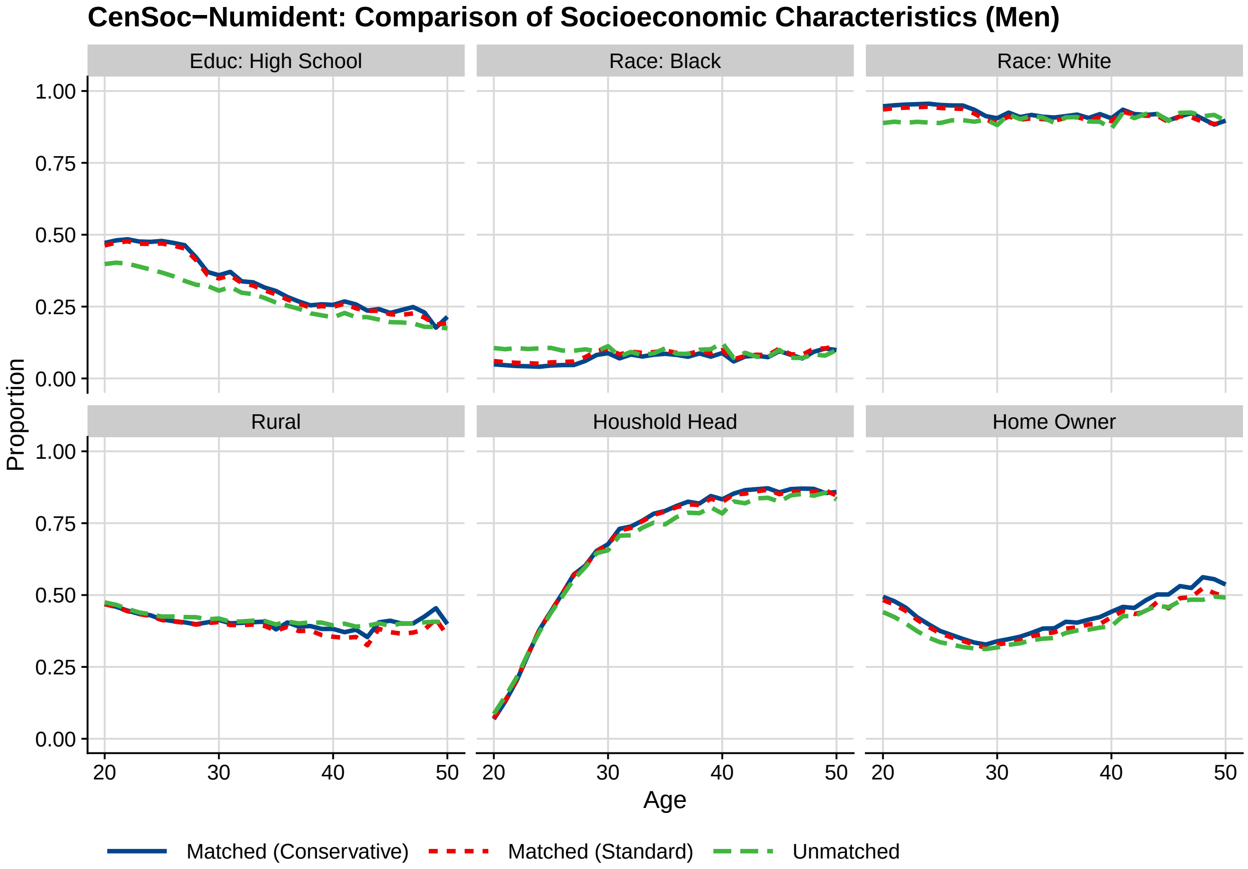

Figure 8: For each panel, lines show the proportion of men with a given socioeconomic characteristics by census age who were not matched to the Numident (green line), matched with the standard algorithm (red line), and matched with the conservative algorithm (blue line).

	1940 Census		Censoc-DMF		CenSoc-DMF (Conservative)	
	No.	%	No.	%	No.	%
Education						
< High School	14486193	65.7	2595341	63.2	1587055	62.8
High School or some college	5836199	26.5	1179529	28.7	732880	29.0
Bachelors Degree	896975	4.1	181246	4.4	113425	4.5
Advanced Degree	410021	1.9	86674	2.1	55,402	2.2
NA	430420	2.0	62588	1.5	36604	1.4
Race						
Black	2114903	9.6	198963	4.8	97886	3.9
White	19828277	89.9	3889049	94.7	2415505	95.6
Other	116628	0.5	17366	0.4	11975	0.5
Marital Status						
Married	13628978	61.8	2565849	62.5	1582257	62.7
Not married	8430830	38.2	1539529	37.5	943109	37.3
Home Ownership						
Home Owner	7964879	36.1	1594140	38.8	1002039	39.7
Not Home Owner	14094929	63.9	2511238	61.2	1523327	60.3
Socioeconomic Index						
1-9	4255639	19.3	713319	17.4	426153	16.9
10-14	2802663	12.7	529972	12.9	330740	13.1
15-25	5626737	25.5	1095464	26.7	679261	26.9
26+	7377168	33.4	1453855	35.4	905154	35.8
N/A	1997601	9.1	312768	7.6	184058	7.3
Rural						
Rural	9298119	42.1	1734573	42.3	1063007	42.1
Urban	12761689	57.9	2370805	57.7	1462359	57.9
Region						
East North Central	4458267	20.2	968308	23.6	626028	24.8
East South Central	1730090	7.8	234986	5.7	127529	5.0
Middle Atlantic	4729114	21.4	903518	22.0	563648	22.3
Mountain	695808	3.2	132592	3.2	81358	3.2
New England	1349283	6.1	267554	6.5	162163	6.4
Pacific	1747202	7.9	352633	8.6	220193	8.7
South Atlantic	3022025	13.7	416072	10.1	226237	9.0
West North Central	2136265	9.7	479397	11.7	313880	12.4
West South Central	2191754	9.9	350318	8.5	204330	8.1

Table 3: Representativeness, by match method for CenSoc-DMF file for pooled birth cohorts of 1900-1920.

	1940 Census		CenSoc-Numident		CenSoc-Numident Conservative	
	No.	%	No.	%	No.	%
Educ						
< High School	14486193	65.7	909049	53.7	677883	52.8
High School or some college	5836199	26.5	653873	38.6	506673	39.4
Bachelors Degree	896975	4.1	73691	4.4	56784	4.4
Advanced Degree	410021	1.9	31245	1.8	24281	1.9
NA	430420	2.0	25740	1.5	18946	1.5
Race						
Black	2114903	9.6	103106	6.1	63596	5.0
Other	116628	0.5	6266	0.4	5183	0.4
White	19828277	89.9	1584226	93.5	1215788	94.6
Marital Status						
Married	13628978	61.8	698102	41.2	522558	40.7
Not married	8430830	38.2	995496	58.8	762009	59.3
Home Ownership						
Home Owner	7964879	36.1	686930	40.6	534365	41.6
Not Home Owner	14094929	63.9	1006668	59.4	750202	58.4
Socioeconomic Indicator						
1-9	4255639	19.3	329962	19.5	246365	19.2
10-14	2802663	12.7	172520	10.2	131446	10.2
15-25	5626737	25.5	457542	27.0	349245	27.2
26+	7377168	33.4	535468	31.6	408982	31.8
NA	1997601	9.1	198106	11.7	148529	11.6
Rural						
Rural	9298119	42.1	727400	42.9	554823	43.2
Urban	12761689	57.9	966198	57.1	729744	56.8
Region						
East North Central	4458267	20.2	371854	22.0	291486	22.7
East South Central	1730090	7.8	105520	6.2	72774	5.7
Middle Atlantic	4729114	21.4	352103	20.8	256780	20.0
Mountain	695808	3.2	59710	3.5	49644	3.9
New England	1349283	6.1	129089	7.6	99755	7.8
Pacific	1747202	7.9	144517	8.5	115768	9.0
South Atlantic	3022025	13.7	196749	11.6	136443	10.6
West North Central	2136265	9.7	187910	11.1	154761	12.0
West South Central	2191754	9.9	146146	8.6	107156	8.3

Table 4: Representativeness, by match method for CenSoc-Numident men for pooled birth cohorts of 1900-1920.

	1940 Census		CenSoc-Numident		CenSoc-Numident Conservative	
	No.	%	No.	%	No.	%
Education						
< High School	13915933	61.3	1319380	54.9	977979	53.9
Advanced Degree	175235	0.8	17222	0.7	13340	0.7
Bachelors Degree	819819	3.6	92117	3.8	70640	3.9
High School or some college	7376597	32.5	939352	39.1	725170	40.0
NA	406200	1.8	36014	1.5	26607	1.5
Race						
Black	2407467	10.6	164580	6.8	92902	5.1
Other	72535	0.3	5197	0.2	4336	0.2
White	20213782	89.1	2234308	92.9	1716498	94.6
Marital status						
Married	16208239	71.4	1602579	66.7	1202079	66.3
Not married	6485545	28.6	801506	33.3	611657	33.7
Home Ownership						
Home Owner	8247623	36.3	904365	37.6	702493	38.7
Not Home Owner	14446161	63.7	1499720	62.4	1111243	61.3
Socioeconomic Indicator						
1-9	1173816	5.2	100050	4.2	67653	3.7
10-14	333347	1.5	25539	1.1	17627	1.0
15-25	2441448	10.8	250058	10.4	188616	10.4
26+	3909290	17.2	454267	18.9	348724	19.2
NA	14835883	65.4	1574171	65.5	1191116	65.7
Rural						
Rural	8776272	38.7	980765	40.8	737783	40.7
Urban	13917512	61.3	1423320	59.2	1075953	59.3
Region						
East North Central	4539245	20.0	505590	21.0	398716	22.0
East South Central	1830078	8.1	183550	7.6	124551	6.9
Middle Atlantic	4965782	21.9	508062	21.1	370875	20.4
Mountain	670434	3.0	70605	2.9	59498	3.3
New England	1423164	6.3	167467	7.0	131051	7.2
Pacific	1659092	7.3	167820	7.0	136261	7.5
South Atlantic	3137101	13.8	316708	13.2	217393	12.0
West North Central	2180243	9.6	251995	10.5	207593	11.4
West South Central	2288645	10.1	232288	9.7	167798	9.3

Table 5: Representativeness, by match method for CenSoc-Numident women for pooled birth cohorts of 1900-1920.

4.3 Middle Initial Analysis

To assess the accuracy of matches in the absence of ground-truth data, we check agreement between the middle initial reported in the Census and the mortality record. As middle initial was not used as a matching field, we assume that disagreement on middle initials likely corresponds to false match. We use middle initials rather than full middle names because full middle names are rarely available in both Census and mortality records, and restrict this analysis to men to avoid complications with middle name changes during marriage for women.

Middle initials are available for 78% of records in the Numident, 30% of records in the 1940 Census, and 27% of records in both datasets. Therefore, our analysis is restricted to the 27% of records that have a middle initial in both datasets. In the Numident, middle initials agree for 87% for the conservative matches, 78% of standard matches, and only 53% of standard matches that were not deemed conservative matches. Figure 9 shows middle initial agreement by birth cohort for CenSoc-Numident.

Middle initials are available for 43% of records in the DMF, 30% of records in the 1940 Census, and 15% of records in both datasets. Middle initials agree in 85% of conservative matches, 72% of standard matches, and 51% of standard matches that were not deemed conservative matches. Figure 10 shows middle initial agreement by birth cohort for CenSoc-DMF.

In the CenSoc-Numident, we assess middle initial agreement rates when birth year in the 1940 Census and Numident records disagree (the ABE algorithm allows flexibility ± 2 years). The motivation behind this analysis is to assess whether the additional matches gained by allowing flexibility on birth year are as accurate as the matches established with an exact agreement on birth year. Figure 11 shows matches in the CenSoc-Numident that disagree on birth year are significantly more likely to have a mismatch on middle initial.

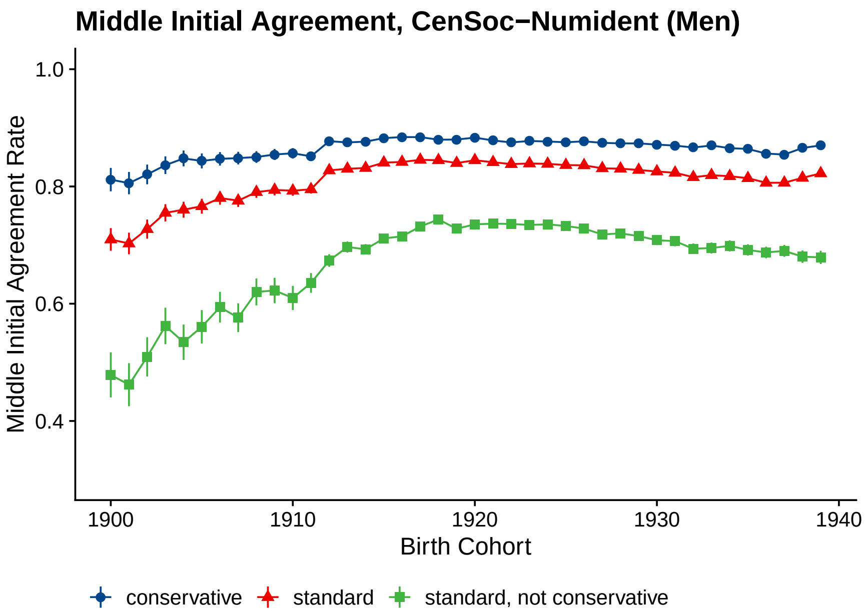

Figure 9: Middle initial agreement by birth cohort for men in the CenSoc-Numident. Middle initial agreement is highest in the conservative matches (blue) and lowest for the standard matches that not deemed "conservative" matches (green).

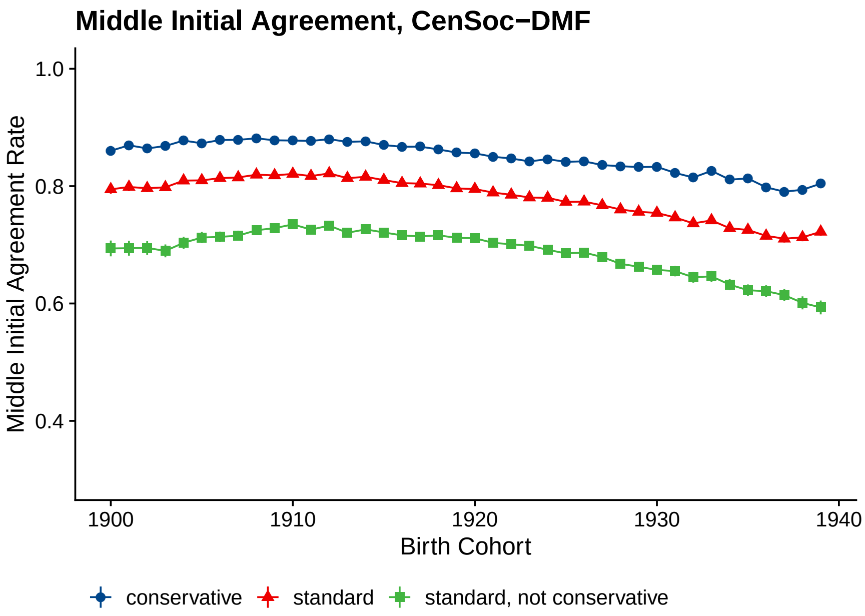

Figure 10: Middle initial agreement by birth cohort for men in the CenSoc-DMF. Middle initial agreement is highest in the conservative matches (blue) and lowest for the standard matches that not deemed "conservative" matches (green).

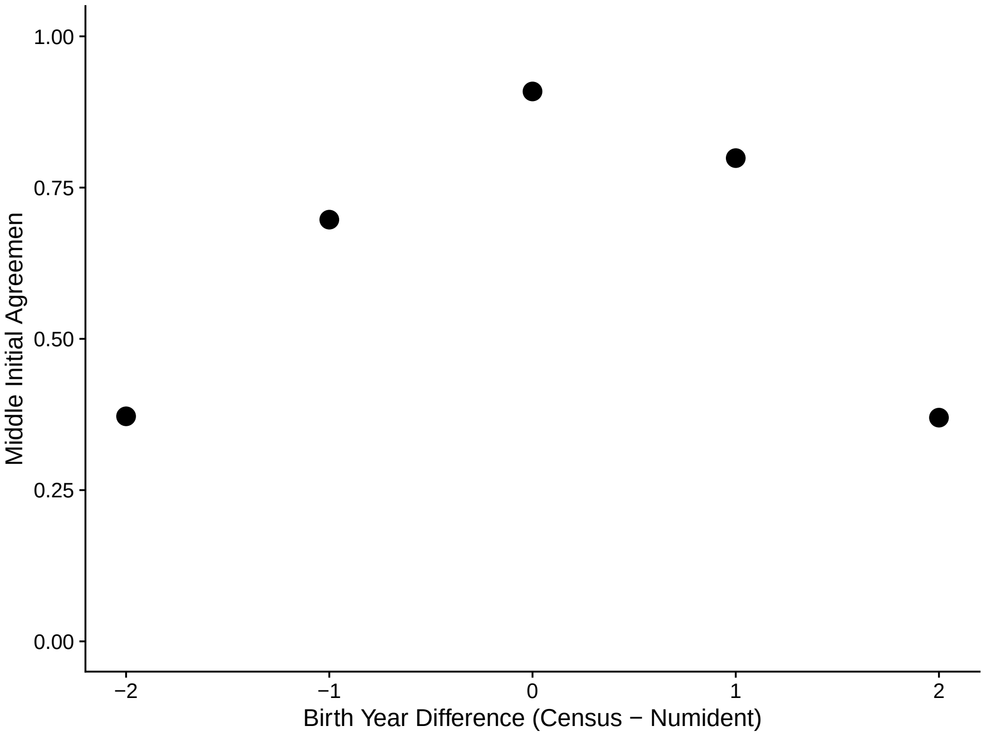

Figure 11: Middle initial agreement for matches established with discrepant birth years.

4.4 Implications of False Matches for Inference

To investigate the effect of false matches on research results, we estimate the association between years of education and longevity from OLS regression on age of death using different samples. Specifically, we first defined three samples from the Numident cohorts of 1900-1920: standard matches, conservative matches, and standard matches that were not deemed conservative matches. Fore each sample, we defined three subsamples based on middle initial agreement – agree, disagree, or both agree and disagree ("pooled"). In total, this gives nine different samples. On each of the nine samples, we ran separate regression estimating the association between years of education and longevity.

Figure 12 plots each of the estimated regression coefficients. Several insights emerge from this figure. First, the regression coefficient for the full "pooled" sample was largest for the conservative matches, very slightly attenuated for the standard matches, and substantially attenuated for the standard matches not deemed conservative. Second, when middle initials agree, regression coefficient point estimates are identical across all three samples (conservative, standard, standard not conservative). Third, when middle initials disagree, the estimated regression coefficient is highly attenuated, and is most attenuated for the standard, matches deemed not conservative sample. Finally, for conservative matches, the estimated coefficient is nearly identical for the "pooled" sample and "agree" sample, suggesting that false matches have minimal impact on inference for this sample.

This analysis demonstrates that false matches systematically introduce measurement error, downwardly biasing the magnitude of estimated regression coefficients (Bailey et al., 2020). While the attenuation bias in this example is modest, we generally recommend researchers restrict to conservatives matches to limit the number of false matches and attenuation bias.

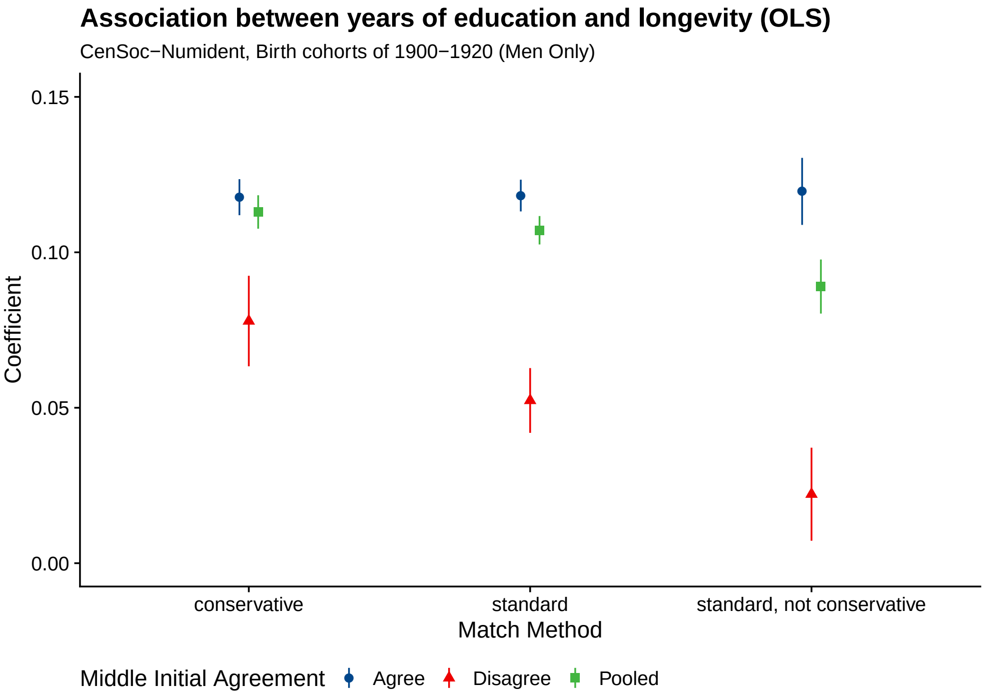

Figure 12: The estimated education gradient using regression on age of death from nine different CenSoc-Numident samples. Blue estimates, where middle initials matched, are nearly identical across samples. Green estimates from the full sample ("pooled") include records where middle initials agree and disagree. Red estimates, where middle initials didn't match are attenuated (biased towards 0).

5 Considerations for Researchers

In summary, there are several caveats and considerations of the CenSoc datasets that warrant discussion. The overall mortality-adjusted match rate for the CenSoc-DMF is 30% (18% conservative), while the overall mortality-adjusted match rate for CenSoc-Numident is approximately 30% for men (22% conservative) and 32% for women (24% conservative). The match rate for Censoc-Numident is lower for earlier birth cohorts (1895-1915) because of the higher rates of birthplace missingness, a required matching field. For analyses of pooled birth cohorts (e.g, individuals born between 1910 and 1920), we recommend including birth cohort fixed effects in regressions for two reasons. First, this helps account for each birth cohort being observed for a different window of ages of death (Breen and Goldstein, 2022). Second, it helps address potential sample composition bias introduced by differential match rates across birth cohorts in CenSoc-Numident.

The CenSoc datasets are not perfectly representative of the general population. While the socioeconomic characteristics of the matched and unmatched samples align closely for the both datasets, the matched sample is slightly more advantaged across a range of socioeconomic dimensions. For instance, 35.2% of individuals in in CenSoc-DMF had completed high school, while 32.5% of individuals in the general population had completed high school. Black people are underrepresented in both datasets, comprising only 4.8% of CenSoc-DMF and 6.5% of CenSoc-Numident, but 9.6% of the general population. However, the population of Black people successfully matched is broadly representative of the general Black population. Non-representativeness has the potential to bias estimates of population-level parameters in linked samples if the under or over-represented population subgroups are also heterogeneous on the outcome of interest (Bailey et al., 2020). Researchers can address this bias by conducting stratified analyses (e.g., fit separate models for Black and White subgroups).

Our middle initial analysis demonstrated that the conservative CenSoc datasets contain fewer false matches than the standard CenSoc datasets. False matches introduce measurement error resulting in attenuated estimates (within a regression framework). For most analyses, we recommend researchers restrict to conservative matches to avoid this attenuation bias. The trade-off of restricting to conservative matches is a decrease in sample size. Researchers must weigh these consideration when working with the CenSoc datasets.

References

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):1832–1856.

Abramitzky, Ran, Leah Platt Boustan and Katherine Eriksson. 2014. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." *Journal of Political Economy* 122(3):467–506.

Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson, James Feigenbaum and Santiago Pérez. 2019. Automated Linking of Historical Data. Technical Report w25825 National Bureau of Economic Research Cambridge, MA: .

Bailey, Martha J., Connor Cole, Morgan Henderson and Catherine Massey. 2020. "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data." *Journal of Economic Literature* 58(4):997–1044.

Breen, Casey F and Joshua R Goldstein. 2022. "Berkeley Unified Numident Mortality Database: Public Administrative Records for Individual-Level Mortality Research." p. 40.

Goldstein, Joshua R., Monica Alexander, Casey F. Breen, Andrea Miranda-González, Felipe Menares, Maria Osborne and Ugur Yildrim. 2021. "CenSoc Mortality File: Version 2.0.".

HMD. 2021. "Human Mortality Database.".

Ruggles, Steven, Catherine A. Fitch and Evan Roberts. 2018. "Historical Census Record Linkage." *Annual Review of Sociology* 44(1):19–37.

Wachter, Kenneth. 2014. *Essential Demographic Methods*. Harvard University Press.

Supplemental Information

A Estimating Probability of Dying in Coverage Window

To estimate the mortality-adjusted match rate, we calculate the probability that an individual dies in our mortality observation window conditional on living until 1940:

$$P(\text{Dying in window}) = P\left(\theta_l \leq D_{year} \leq \theta_r | D_{year} > 1940\right) \tag{4}$$

where θ_l is the year of left truncation, θ_r is the year of right truncation, and D_{year} is the age of death. Because full cohort lifetables are not available for the U.S., we use 1x1 mortality rates from Human Mortality Database (HMD). We convert the mortality rates using the following conversion formula (Wachter, 2014):

$$_nq_x = \frac{(n)(_nM_x)}{1 + (n -_n a_x)(_nM_x)} \tag{5}$$

which, assuming $_1a_x = 0.5$ and $n = 1$, simplifies to

$$_1q_x = \frac{_1M_x}{1 - 0.5(_1M_x)}. \tag{6}$$

We define the probability of survival to be $_1p_x = 1 -_1 q_x$. We estimate two quantities, the probability of surviving from 1940 until the observation window ($_np_x$) and the probability of dying during the observation window ($_\lambda p_{n+x}$). The produce of dying in the mortality observation window is the product of these two quantities:

$$\begin{aligned} P(\text{Dying in window}) &= P\left(\theta_l \leq D_{year} \leq \theta_r | D_{year} > 1940\right) \\ &= \underbrace{_np_x}_{\text{Living until observation window}} \times \underbrace{(1 -_{30} p_{n+x})}_{\text{Dying during observation window}} \end{aligned} \tag{7}$$

B Alternative Denominators

In our main analysis, we used the 1940 Census as our reference baseline for calculating match rates. In this section, we assess match rate using an alternative denominator, the mortality data files (DMF or Numident).

B.1 CenSoc-DMF

Here, we define the match rate as the proportion of individuals observed in the DMF file successfully linked to the 1940 Census:

$$M = \left(\frac{\text{Number established matches}}{\text{Number of deaths 1975-2005 in the DMF}} \right) \times \left(\frac{1}{\text{Sex ratio of deaths 1975-2005}} \right) \quad (8)$$

Because there is no information on sex in the DMF, we approximated the number of male deaths based on a cohort-specific sex ratio from the Human Mortality Database HMD (2021). Figure 13 shows the CenSoc-DMF match rate by birth cohort and linkage method. Match rates for the standard variant remain stable around 30% across birth cohorts and the conservative rate remains stable at slightly below 20%. This agrees with the mortality-adjusted CenSoc-DMF match rate reported in Figure 2.

These calculations do not account for immigrants arriving to the US after 1940 and dying within our observation window. The denominators for these rates therefore include individuals whose deaths are recorded in the DMF but cannot be observed in or linked to the 1940 Census.

B.2 CenSoc-Numident

B.2.1 Match Rates

We calculate match rates for the CenSoc-Numident using the Numident records as the universe of potential matches. We restrict to deaths occurring in the observation window of 1988-2005.

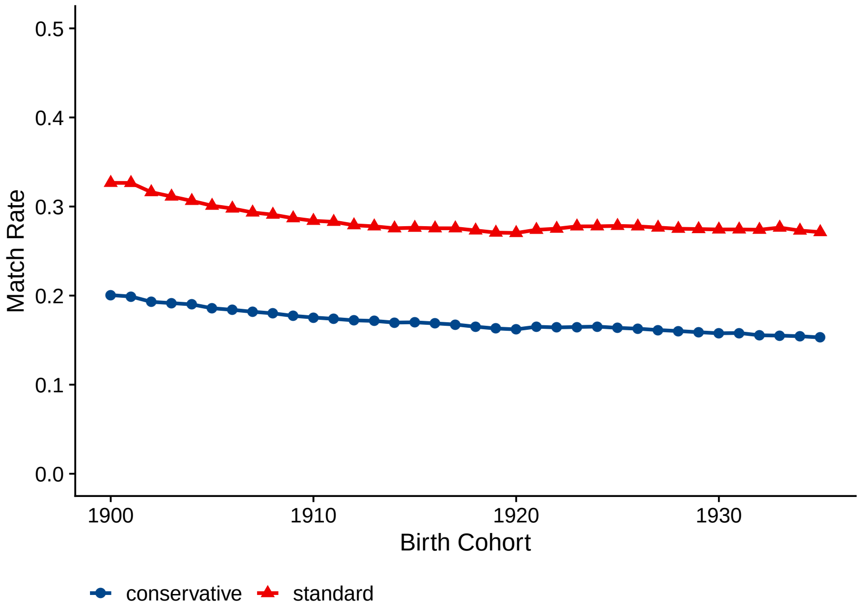

Figure 13: CenSoc-DMF match rate. Only deaths within the window of 1975-2005 are included in calculations. The cohort sex ratios of deaths in the DMF are assumed to be identical to those in the HMD.

Match rates for men and women are shown in Figure 14 and Figure 15, respectively. The "raw" match rates reflect the proportion of all death records 1988-2005 successfully linked to the 1940 census. The birthplace-available match rate reflects the proportion of death records 1988-2005 with non-missing birthplace information successfully linked to the census. We do not attempt to match records with missing information on birthplace.

The raw match rates for both men and women improve significantly after 1910 because of the increased availability of birthplace data shown in Figure 5. Match rates conditional on the availability of birthplace, however, are substantially more consistent across cohorts.

B.2.2 Socioeconomic characteristics

The BUNMD includes a limited number of covariates in addition to birth and death dates, allowing us to consider the representativeness of the CenSoc-Numident from a different perspective. Here, we compare place of birth and race in the linked and unlinked BUNMD. For purposes of comparison, the universe of records will consist only of those with a valid birthplace, as we do not attempt to link BUNMD records without birthplace.

Figure 16 and Figure 17 show the racial composition (proportion of Black and White

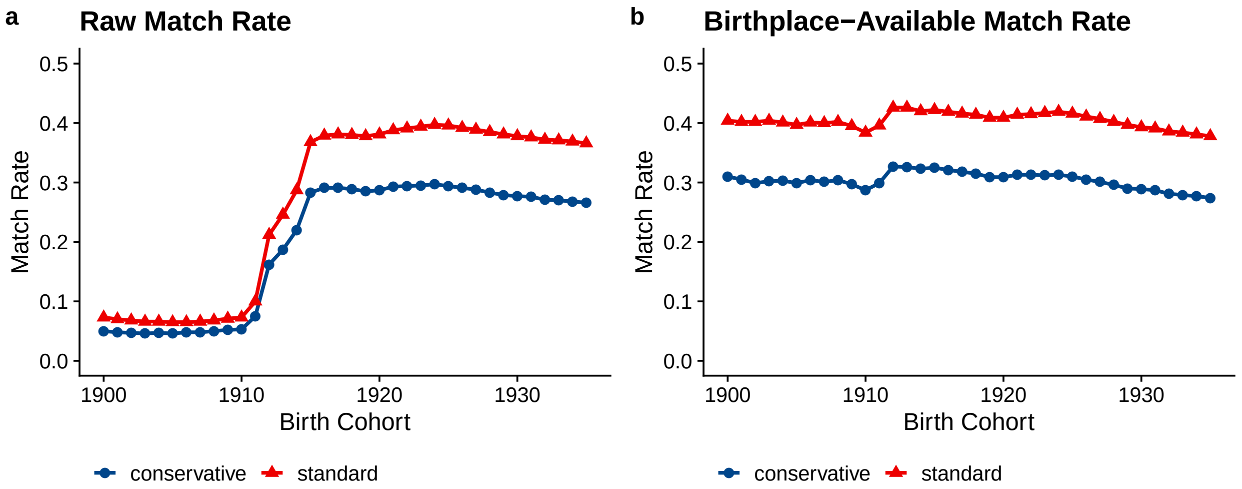

Figure 14: **CenSoc-Numident Match Rates for men.** Panel (a) shows the raw CenSoc-Numident match rate for men. Panel (b) shows the match rate for women with non-missing birthplace.

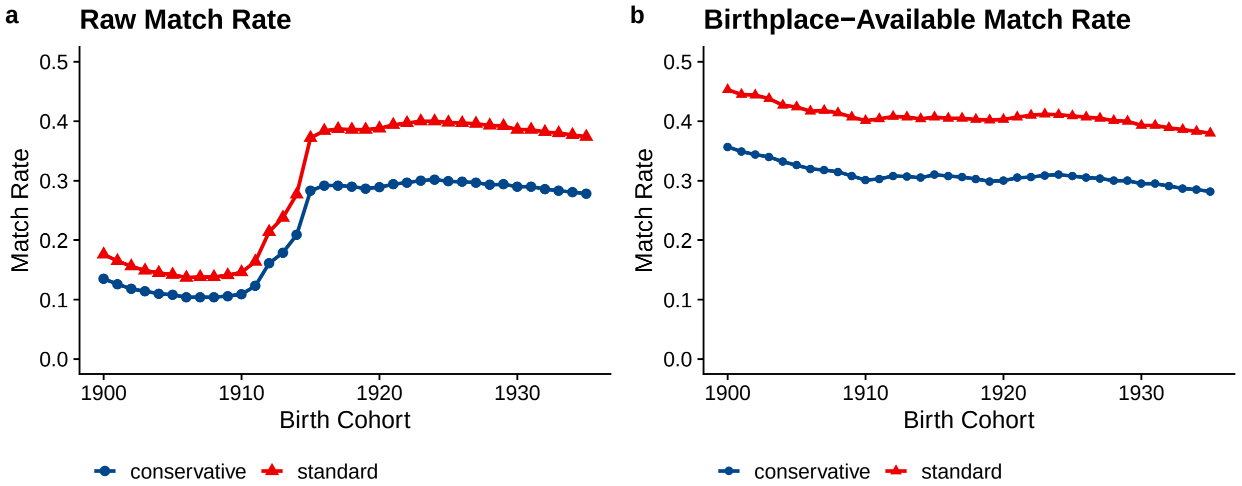

Figure 15: **CenSoc-Numident Match Rates for women.** Panel (a) shows the raw CenSoc-Numident match rate for women. Panel (b) shows the match rate for women with non-missing birthplace.

people) of matched and unmatched records in the BUNMD.[3] These figures include only the cohorts of 1915-1940, as race information is missing in a high proportion of records for earlier cohorts.

In all cases, the linked sets contains a smaller proportion of Black individuals and larger proportion of White individuals than the unlinked set of records. The conservative ABE

[3] The BUNMD contains data on race for each individual's first and last social security application. These are highly consistent; for these calculations, we use first reported race.

matching variant links relatively more White individuals and fewer Black individuals than the standard matching variant.

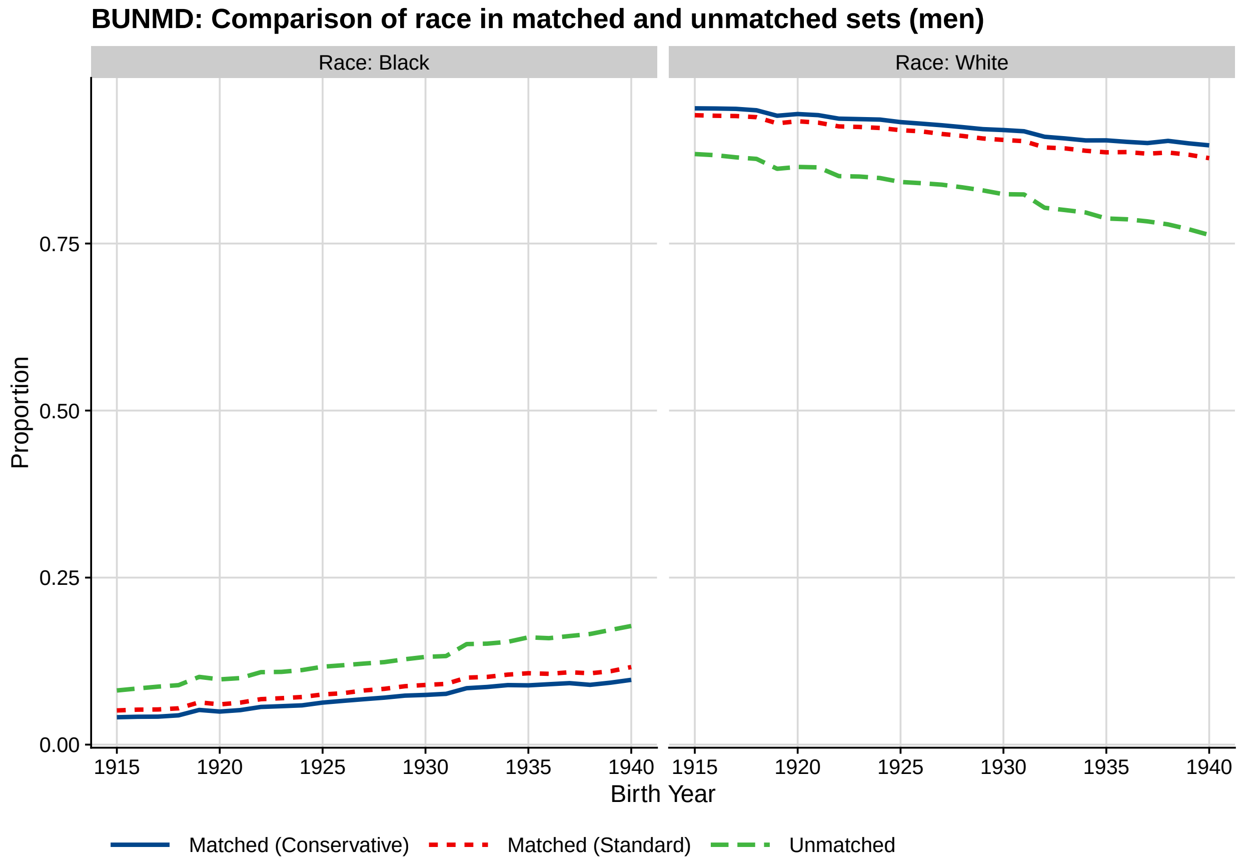

Figure 16: Race of men in matched and unmatched BUNMD/Numident

Tables 7 and 9 compare the racial and geographic composition of the full BUNMD to the BUNMD records successfully linked to the 1940 Census. About 9.1% of individuals in the BUNMD with valid birth dates, death dates, and birth place data are foreign born, some proportion of which are not matchable because they immigrated to the US after the 1940 census enumeration.

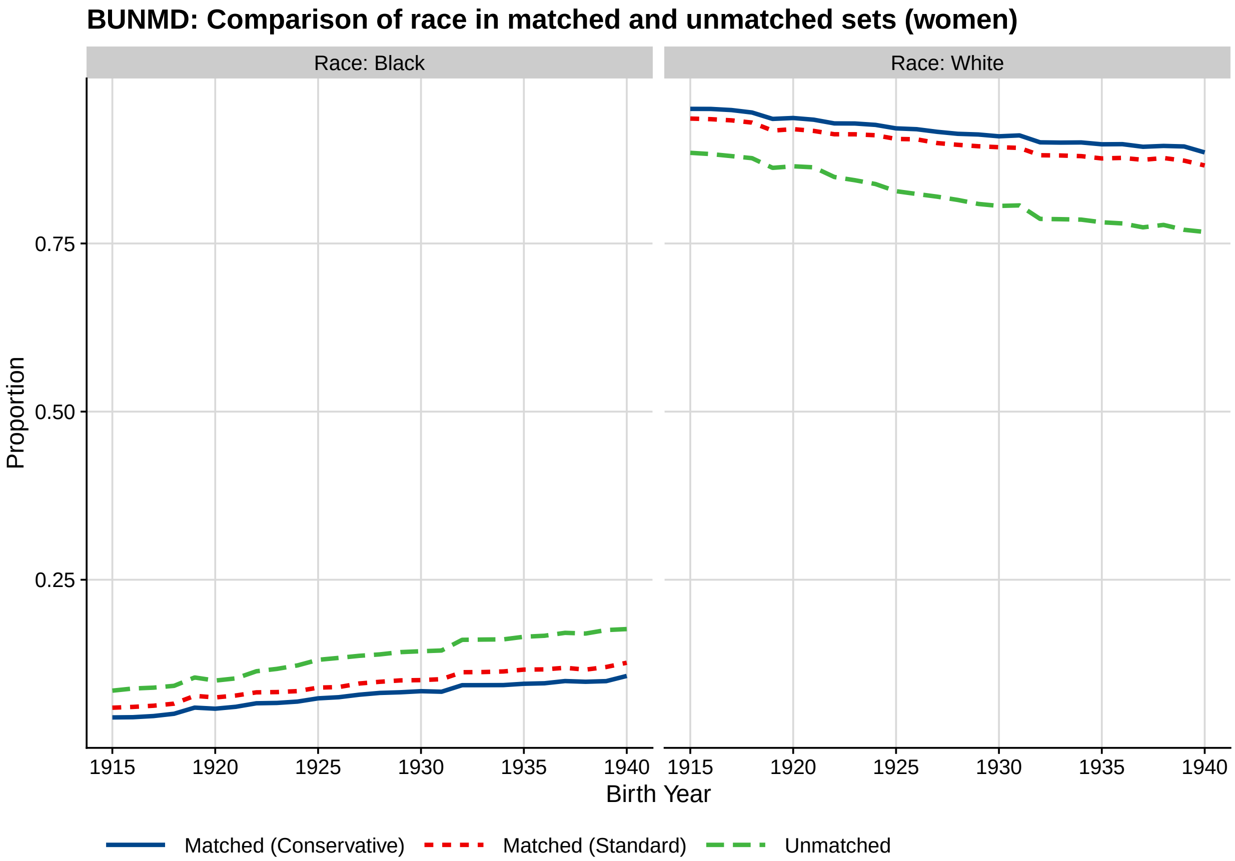

Figure 17: Race of women in matched and unmatched BUNMD/Numident

	Full BUNMD		CenSoc-Numident		CenSoc-Numident Conservative	
	No.	%	No.	%	No.	%
Race						
Black	972675	10.0	295100	7.5	182595	6.2
White	8449562	86.6	3604872	91.4	2732835	92.7
Other	264015	2.7	24227	0.6	18224	0.6
N/A	70673	0.7	20885	0.5	15256	0.5
Region of Birth						
East North Central	1683907	17.3	820542	20.8	638079	21.6
East South Central	908534	9.3	334416	8.5	222279	7.5
Foreign Born	812904	8.3	25181	0.6	21026	0.7
Middle Atlantic	1716206	17.6	741294	18.8	538248	18.3
Mountain	286768	2.9	139207	3.5	118401	4.0
New England	550360	5.6	269800	6.8	210012	7.1
Pacific	418788	4.3	191956	4.9	153227	5.2
South Atlantic	1363932	14.0	523233	13.3	354175	12.0
West North Central	967023	9.9	496764	12.6	406495	13.8
West South Central	1048503	10.7	402691	10.2	286968	9.7

Table 7: **Representativity by match method for CenSoc-Numident Men. Pooled Cohorts of 1915-1940. Only records with available birthplace are included.**

	Full BUNMD		CenSoc-Numident		CenSoc-Numident Conservative	
	No.	%	No.	%	No.	%
Race						
Black	878007	10.5	281496	8.4	170043	6.7
White	7182392	85.8	3027747	90.2	2318416	91.8
Other	212901	2.5	16838	0.5	12911	0.5
N/A	101735	1.2	32450	1.0	24176	1.0
Region of Birth						
East North Central	1429614	17.1	669616	19.9	529264	21.0
East South Central	798075	9.5	310553	9.2	204889	8.1
Foreign Born	685409	8.2	21414	0.6	18169	0.7
Middle Atlantic	1472574	17.6	636727	19.0	471329	18.7
Mountain	249953	3.0	111813	3.3	97462	3.9
New England	463685	5.5	220809	6.6	177276	7.0
Pacific	347540	4.1	152296	4.5	124778	4.9
South Atlantic	1203721	14.4	486425	14.5	326693	12.9
West North Central	805627	9.6	389975	11.6	321517	12.7
West South Central	918837	11.0	358903	10.7	254169	10.1

Table 9: **Representativity by match method for CenSoc-Numident Women. Pooled Cohorts of 1915-1940. Only records with available birthplace are included.**

B.3 Additional Tables

Representativeness tables for Black Americans, pooled to birth cohorts of 1900-1920.

	1940 Census		CenSoc-DMF		CenSoc-DMF Conservative	
	No.	%	No.	%	No.	%
Education						
< High School	1883899	89.1	174005	87.5	84836	86.7
High School or some college	155922	7.4	18104	9.1	9585	9.8
Bachelors Degree	20591	1.0	2383	1.2	1299	1.3
Advanced Degree	6250	0.3	783	0.4	456	0.5
NA	48241	2.3	3688	1.9	1710	1.7
Marital Status						
married	1397386	66.1	133258	67.0	66291	67.7
not married	717517	33.9	65705	33.0	31595	32.3
Home Ownership						
Home Owner	394106	18.6	41637	20.9	21331	21.8
Not Home Owner	1720797	81.4	157326	79.1	76555	78.2
Socieconomic Indicator						
1-9	975464	46.1	89376	44.9	43314	44.2
10-14	389789	18.4	38560	19.4	19558	20.0
15-25	388463	18.4	37638	18.9	18816	19.2
26+	149136	7.1	15487	7.8	7951	8.1
NA	212051	10.0	17902	9.0	8247	8.4
Rural						
Rural	1005019	47.5	97067	48.8	47957	49.0
Urban	1109884	52.5	101896	51.2	49929	51.0
Region						
East North Central	189108	8.9	19142	9.6	9760	10.0
East South Central	434240	20.5	39414	19.8	18979	19.4
Middle Atlantic	223565	10.6	21034	10.6	10636	10.9
Mountain	6876	0.3	678	0.3	358	0.4
New England	14931	0.7	1700	0.9	880	0.9
Pacific	25848	1.2	2926	1.5	1555	1.6
South Atlantic	773558	36.6	67248	33.8	31137	31.8
West North Central	56110	2.7	5837	2.9	2947	3.0
West South Central	390667	18.5	40984	20.6	21634	22.1

Table 10: Representativeness, by match method for CenSoc-DMF Black Men.

	1940 Census		CenSoc-Numident		CenSoc-Numident Conservative	
	No.	%	No.	%	No.	%
Education						
< High School	1883899	89.1	87016	84.4	53150	83.6
Advanced Degree	6250	0.3	343	0.3	231	0.4
Bachelors Degree	20591	1.0	1262	1.2	844	1.3
High School or some college	155922	7.4	12643	12.3	8292	13.0
NA	48241	2.3	1842	1.8	1079	1.7
Marital Status						
Married	1397386	66.1	56012	54.3	34723	54.6
Not married	717517	33.9	47094	45.7	28873	45.4
Home Ownership						
Home Owner	394106	18.6	22686	22.0	14541	22.9
Not Home Owner	1720797	81.4	80420	78.0	49055	77.1
Socioeconomic Indicator						
1-9	975464	46.1	46323	44.9	28227	44.4
10-14	389789	18.4	17992	17.5	11366	17.9
15-25	388463	18.4	19905	19.3	12340	19.4
26+	149136	7.1	7754	7.5	5008	7.9
NA	212051	10.0	11132	10.8	6655	10.5
Rural						
Rural	1005019	47.5	52777	51.2	32750	51.5
Urban	1109884	52.5	50329	48.8	30846	48.5
Region						
East North Central	189108	8.9	9318	9.0	5950	9.4
East South Central	434240	20.5	20726	20.1	12519	19.7
Middle Atlantic	223565	10.6	10361	10.0	6439	10.1
Mountain	6876	0.3	352	0.3	236	0.4
New England	14931	0.7	1091	1.1	722	1.1
Pacific	25848	1.2	1432	1.4	988	1.6
South Atlantic	773558	36.6	36499	35.4	21367	33.6
West North Central	56110	2.7	2825	2.7	1867	2.9
West South Central	390667	18.5	20502	19.9	13508	21.2

Table 11: Representativeness, by match method for CenSoc-Numident Black men.

	1940 Census		CenSoc-Numident		CenSoc-Numident Conservative	
	No.	%	No.	%	No.	%
Education						
< High School	2074427	86.2	135259	82.2	74715	80.4
Advanced Degree	5245	0.2	429	0.3	266	0.3
Bachelors Degree	30639	1.3	2472	1.5	1578	1.7
High School or some college	253236	10.5	23712	14.4	14856	16.0
N/A	43920	1.8	2708	1.6	1487	1.6
Marital Status						
Married	1716520	71.3	106457	64.7	59364	63.9
Not married	690947	28.7	58123	35.3	33538	36.1
Home Ownership						
Home Owner	470117	19.5	34299	20.8	20622	22.2
Not Home Owner	1937350	80.5	130281	79.2	72280	77.8
Socioeconomic Indicator						
1-9	618635	25.7	40377	24.5	21608	23.3
10-14	143976	6.0	8086	4.9	4353	4.7
15-25	255792	10.6	17004	10.3	9534	10.3
26+	98371	4.1	7689	4.7	4767	5.1
N/A	1290693	53.6	91424	55.5	52640	56.7
Rural						
Rural	1022353	42.5	72724	44.2	41036	44.2
Urban	1385114	57.5	91856	55.8	51866	55.8
Region						
East North Central	213053	8.8	14338	8.7	8460	9.1
East South Central	500012	20.8	32747	19.9	17775	19.1
Middle Atlantic	278106	11.6	19245	11.7	11140	12.0
Mountain	6346	0.3	383	0.2	254	0.3
New England	16540	0.7	1447	0.9	953	1.0
Pacific	27102	1.1	1690	1.0	1075	1.2
South Atlantic	854483	35.5	60305	36.6	33241	35.8
West North Central	63780	2.6	4405	2.7	2766	3.0
West South Central	448045	18.6	30020	18.2	17238	18.6

Table 13: Representativeness, by match method for CenSoc-Numident Black women.