Name-Based Estimators of Intergenerational Mobility

– Working Paper –

Torsten Santavirta* and Jan Stuhler†‡

August 20, 2020

Abstract

A recent development in intergenerational research is the use of *names* – first names and surnames – to overcome data limitations. Yet, while the use of name-based estimators has produced innovative evidence on mobility across multiple generations, historical periods, or regions, it remains unclear how different methods compare and how reliable they are. This paper reviews and validates name-based methods, based on newly digitized data from Finland and other sources. We show that the different methods are closely related, but that their interpretation crucially depends on the sampling properties of the data, which differ across studies. To demonstrate their reliability, we compare the intergenerational mobility of the two combatant groups in the Finnish Civil War of 1918. Both conventional and name-based methods indicate substantially higher downward mobility among members of the socialist "Red Guard" as compared to the conservative "White Guard."

JEL classification: J62.

*Stockholm University, Swedish Institute for Social Research (SOFI) and Institute for Housing and Urban Research (IBF) (email: torsten.santavirta@ibf.uu.se)

†Universidad Carlos III de Madrid and Swedish Institute for Social Research (email: jan.stuhler@uc3m.es)

‡We thank Ilkka Jokipii and Virva Liski for excellent data collection and data preparation for this project. We are very grateful to James Feigenbaum, Claudia Olivetti and Daniele Paserman for the provision of samples and replication files to link records in the U.S. Census. We also thank Maia Güell, Sevi Rodríguez Mora and seminar participants at SOFI at Stockholm University, Uppsala University, Purdue University, Universidad Autonoma de Barcelona, Statistics Norway, and the 2019 SOLE conference for comments. Support from the Ministerio de Ciencia, Innovación y Universidades (Spain, MDM 2014-0431, ECO2017-87908-R and IJCI-2016-30011) and Comunidad de Madrid (MadEco-CM S2015/HUM-3444) is gratefully acknowledged.

1 Introduction

A recent development in research on intergenerational mobility is the use of *names* to overcome data limitations. Conventional measures require linked data on two generations. For example, a common measure is based on the regression of the child's socioeconomic status y_i in family i on the parent's corresponding status x_i,

$$y_i = \alpha + \beta x_i + \epsilon_i. \tag{1}$$

with a steeper slope β indicating a greater dependence of child status on parental status. When family links are not available, this "*direct*" regression becomes infeasible. However, names—first names and surnames—can serve as a proxy for these links. Based on this insight, researchers have developed different types of *name-based* estimators of intergenerational mobility, which have become instrumental in several strands of the literature. Examples include recent work on the long-run persistence of inequality across multiple generations (e.g., Clark and Cummins 2014, Barone and Mocetti 2020) and on trends in intergenerational mobility (e.g, Clark 2014, Olivetti and Paserman 2015, Güell, Rodríguez Mora and Telmer 2015) and its pattern across regions (Güell et al. 2018).

Table 1 provides a partial list of recent contributions. While all studies are motivated by the observation that names contain socioeconomic information, they exploit that information in different ways. Most authors focus on the innovative features of their respective method or its application in a particular setting. The conceptual similarities that link the various methods have, however, received less attention. This methodological diversity not only complicates the interpretation of name-based estimators and their further development, but also masks the degree to which insights and criticisms regarding one method extend to another—or the general approach as such.

In this paper, we therefore present a systematic review of name-based estimators of intergenerational mobility.[1] Specifically, we (i) provide an overview of the proposed methods, (ii) evaluate their properties, strengths, and weaknesses, and (iii) describe how the various methods are linked. Our conceptual arguments are empirically supported by evidence from U.S. Census data and newly digitized historical data from Finland that contain all the required elements (including direct family links) necessary for comparing name-based and conventional estimators. We conclude our study by evaluating their performance in a typical application, estimating mobility rates of the two antagonistic parties in the Finnish Civil War.

[1] A lively debate has ensued on the validity and interpretation of specific name-based studies. Recent contributions include Chetty et al. (2014), Vosters and Nybom (2017), Torche and Corvalan (2018), Braun and Stuhler (2018), Güell et al. (2018), Solon (2018), Adermon, Lindahl and Palme (2019), Vosters (2018), Clark (2018), and Choi, Gu and Shen (2018). However, no systematic review has thus far been conducted. Feigenbaum (2018) comes closest in spirit, showing that in a U.S. sample the grouping and direct estimators qualitatively arrive at the same conclusion.

Table 1: Name-Based Intergenerational Studies

Authors	*Year*	*Publication*	*Method*	*Data*	*Main Application*
Clark	2012	Working Paper	Surnames, Name Frequencies	Repeated cross-section of surname frequencies	Multigenerational mobility in Sweden
Clark and Cummins	2012	Working Paper	Surnames, Grouping	Repeated cross-section of rare surnames	Multigenerational mobility in England
Collado, Ortuño and Romeu	2012	Reg. Science and Urban Econ.	Surnames, Grouping (by region)	Single cross-section across areas	Intergenerational consumption mobility in Spain
Collado, Ortuño and Romeu	2013	Working Paper	Surnames, Grouping	Repeated cross-section of surname averages	Multigenerational mobility in Spanish provinces
Clark	2014	Princeton University Press	Surnames, Grouping	Repeated cross-section of rare surnames	Inter- and multi-generational mobility in various
Clark and Cummins	2014	Economic Journal	Direct and Surnames, Grouping	Repeated cross-section of rare surnames	Multigenerational wealth mobility in England
Güell, Rodríguez and Telmer	2015	Review of Economic Studies	Surnames, R2	Single cross-section	Intergenerational mobility level and trends in Catalonia
Clark and Diaz-Vidal	2015	Working Paper	Surnames, Grouping	Repeated cross-section of surname averages	Multigenerational and assortative mobility in Chile
Olivetti and Paserman	2015	American Economic Review	First names, Two-sample Two-stage IV	Repeated cross-section	Historical mobility trends in the United States
Nye, Mason, Bryukhanov, Poly-achenko, Rusanov	2016	Working Paper	Surnames, Name Frequencies	Repeated cross-section of name frequencies	Intergenerational mobility in Russia
Durante, Labartino and Perotti	2016	Working Paper (R&R AEJ:Policy)	Surnames, Name Frequencies	Single cross-section of surname frequencies	Family connections at Italian universities
Feigenbaum	2018	Economic Journal	Direct, First and Surnames, R2, Grouping		Historical mobility level in Iowa, United States
Güell, Pellizzari, Pica, and Rodríguez	2018	Economic Journal	Surnames, R2	Single cross-section across areas	Regional variation in mobility in Italy
Olivetti, Paserman and Salisbury	2018	Explorations in Economic History	First names, Two-sample Two-stage IV	Repeated cross-section	Multigenerational mobility in the United States
Barone and Mocetti	2020	Review of Economic Studies (Forthcoming)	Surnames, Two-sample Two-stage IV	Repeated cross-section of surname averages	Multigenerational mobility in Florence, Italy

Note: The table lists selected intergenerational mobility research that use first or surnames to overcome the lack of direct parent-child links.

Table 2: A Classification of Name-Based Methods

Method	*First names*	*Last names*
R-squared Estimators	-	Güell, Rodríguez and Telmer (2015), Güell, Pellizzari, Pica and Rodríguez (2018)
Grouping Estimators	Olivetti and Paserman (2015), Olivetti, Paserman and Salisbury (2018), Feigenbaum (2018)	Clark (2012), Collado Ortuño and Romeu (2012), Collado Ortuño and Romeu (2013), Clark (2014), Clark and Cummins (2014), Feigenbaum (2018), Barone and Moretti (2020)

We first provide an overview of the various methods that have, to date, been developed. We argue that the vast majority of name-based studies can be categorized within the simple two-by-two diagram shown in Table 2, with name type (*first names* vs. *surnames*) on the horizontal axis and the type of estimator on the vertical axis (R^2 vs. *grouping* estimators).[2]

Starting from the top-right cell, the R^2 *estimator* developed by Güell, Rodríguez Mora and Telmer (2015) considers the joint distribution of names and socioeconomic status in a given generation, thereby entirely circumventing the need to link generations. If both surnames and status are transmitted from one generation to the next, then rare last names should explain status variation in the cross-section. The R^2 of a regression of individual-level outcomes on a set of surname dummies summarizes this *informational content of surnames*. Because a high R^2 implies strong status inheritance (and vice versa), the estimator can be used to rank groups or regions by their level of intergenerational mobility.

Despite being labeled differently by different authors, most studies use what is fundamentally the same type of *grouping estimator* (bottom-row, left and right cells), which can be described as a two-step estimator. In the first step, the average socioeconomic status within each name group and generation is computed. In the second step, a variant of the intergenerational regression (1) is estimated in which the parent's socioeconomic status x_i is imputed by the group-level means for their generation. Prominent examples include Clark (2014) and related studies, such as Clark et al. (2015). Using historical sources that span several countries and centuries, they show that socioeconomic status only slowly regresses at the surname level. Using medieval census data from the Italian city of Florence,

[2]Names have been used in a number of other creative ways, which we do not review here. For example, Collado, Ortuño-Ortín and Romeu (2012) circumvent the lack of intergenerational data on consumption choices by comparing the spatial distributions of consumption behavior and surnames. Clark et al. (2015) study the relative frequency of names on the admissions lists of Oxford and Cambridge Universities, as far back as 1170. Paik (2014) show that in Korea, the average prestige of historical clan lineages in a region is predictive of its contemporary educational level. And in some settings, names can be used to impute direct links between parents and their offspring, see Abramitzky, Boustan and Eriksson (2012), Long and Ferrie (2013), Johnson, Massey and O'Hara (2015), Feigenbaum (2016), Modalsli (2017) or Abramitzky, Mill and Pérez (2018).

Barone and Mocetti (2020) show that intergenerational correlations can persist across *six centuries* (i.e., over the very long run). While small in absolute value, these long-run correlations are much higher than what extrapolations of conventional parent-child estimates would suggest.

Olivetti and Paserman (2015) develop a grouping estimator based on *first names*, in which the individual's given name serves as a proxy for family background. One advantage of this approach is that first names do not change upon marriage, and therefore remain informative in parental and maternal lineages, for both sons and daughters. Olivetti and Paserman note that their empirical strategy can be interpreted as a two-sample two-stage least squares (TS2SLS) estimator, in which the first stage groups parental status by first name and the second stage regresses child socioeconomic outcomes against their parental group mean. This approach allows them to provide evidence on U.S. mobility trends in a previously unexplored period, the late 19th and early 20th century. Olivetti, Paserman and Salisbury (2018) further extend this approach to track paternal and maternal lineages in a multigenerational context.

Finally, we note that the R^2 estimator proposed by Güell, Rodríguez Mora and Telmer (2015) can be adapted to measure the informational content of *first names*. The conceptual motivation given by Güell et al. is based on the surname naming process, and does not necessarily extend to first names. However, the informational content of the latter, and their potential value for mobility research has been demonstrated by Olivetti and Paserman (2015). Furthermore, the R^2 estimator based on first names performs well in our application. We also show that the various estimators are closely related. The R^2 estimator proposed by Güell, Rodríguez Mora and Telmer (2015) is approximately the (adjusted) R^2 from the first stage of the two-stage estimators proposed by Olivetti and Paserman (2015) for first names or Barone and Mocetti (2020) for surnames. Moreover, these two-stage estimators belong to the same class of *grouping estimators* used in earlier studies that directly relate surname averages across generations, such as Clark (2014).

To interpret the results of the different studies, and to understand how they compare, we develop a simple regression framework that highlights a number of implications and dependencies that have not been made explicit before. In particular, we show that the *same* grouping estimator actually estimates *different* statistical objects depending on the sampling properties of the underlying data. A key property is the conditional probability that a parent is sampled when his or her child is included in the child sample (i.e., the "*overlap*" between the parent and child samples). As this probability differs widely across studies, the existing estimates are not directly comparable—even among studies that use the same type of estimator.

The argument that the grouping estimator is highly dependent on properties of the underlying data also links the grouping and R^2 estimators. Specifically, a low informational

content of names corresponds to a "weak" first stage in the grouping estimator. Interestingly, this does not pose much of an issue if the parent and child samples overlap. In such settings, the grouping is a standard 2SLS estimator and is therefore biased towards the OLS estimator; yet such bias is desirable if the (feasible) grouping estimator is meant to approximate the (infeasible) OLS estimator. However, if the parent and child samples do not overlap, the grouping estimator instead corresponds to a split-sample IV estimator, and is biased towards zero (Choi, Gu and Shen 2018, Khawand and Lin 2015). We show that the resulting bias can be large in typical applications, and propose a simple bias correction procedure that accounts for (i) the degree of overlap between the parent and child samples, and (ii) the extent to which the name group means in the parent sample predict the corresponding means in the child sample. The correction procedure performs well even in small samples, and its application could improve the comparability of estimates in the literature.

In addition, the grouping and R^2 estimators are subject to similar conceptual issues. First, both are identified from rare names; the approach chosen by Clark (2014) has been criticized on these grounds, but this observation applies to all name-based estimators, including those using first names (although, as we show, to a lesser extent). However, while the R^2 estimator decreases in name frequency, the grouping estimates tend to increase. Second, name frequency decreases with socioeconomic status. While this suggests that name-based estimators capture mobility within a selective portion of the population, we argue that this concern does not invalidate the approach. Third, the grouping estimator tends to be more susceptible to sample size than the R^2 estimator. Fourth, name-based estimators weight the underlying transmission mechanisms differently than individual-level estimators. Intergenerational correlations reflect various transmission processes within the family or more aggregate groups (e.g., defined by ethnicity or region), and name-based estimators weight aggregate processes more heavily. We address these concerns and as well as several additional issues.

We conclude our study with an application based on historical data on veterans of the Finnish 1918 Civil War, which was fought between the land-owning and educated conservative elite and the working class and agricultural proletariat. Our objective is to compare the prewar mobility of the members of the socialist *"Red Guard"*with that of the conservative *"White Guard"*. All name-based methods estimate lower relative mobility for the Red compared to the White Guard. These findings are consistent with the direct estimates, which show that downward intergenerational mobility was more prevalent among the Red Guard. Hence all name-based methods "pass the test".

The remainder of this paper is organized as follows. Section 2 reviews the main insights from recent name-based studies. Section 3 introduces the data. Section 4 explores the informational content of first names and surnames and reviews the R^2 estimators. Section

5 assesses the grouping estimators, while Section 6 examines the properties and various conceptual caveats affecting both estimators, highlighting the different methods' stance and evidence relative to name mutations. Section 7 presents our applications on cross-group and cross-regional mobility pattern during the time of the Finnish Civil War. Section 8 concludes.

2 Recent Applications

By enabling the exploitation of historical and cross-sectional data, name-based estimators have opened up promising new research areas. They have been instrumental in three particularly active strands of the literature, and are starting to change our understanding of intergenerational processes in a number of key aspects:

First, they are informative about the extent of intergenerational mobility in the very long run. Studies such as Clark (2014), Clark and Cummins (2014) or Barone and Mocetti (2020) show that the average socioeconomic status of surnames can be highly persistent across many generations, in fact much more so than the socioeconomic status of individual families as captured by conventional estimators and direct parent-child links. Clark (2014) notes that this observation is consistent with the idea that conventional measures understate the degree to which economic inequalities persist, because they do not capture the transmission of unobserved characteristics that affect the socioeconomic prospects of future generations. If correct, this interpretation would drastically change our understanding of intergenerational processes. It has however triggered a lively debate, with some scholars remaining decidedly critical as to the validity of the surname-based grouping estimator itself (see footnote 1). However, recent studies that directly link distant family members largely confirm that conventional parent-child measures understate the multigenerational transmission of economic advantages (e.g. Lindahl et al., 2015; Braun and Stuhler, 2018; Neidhöfer and Stockhausen, 2019; Colagrossi, d'Hombres and Schnepf, 2019; Collado, Ortuño-Ortín and Stuhler, 2019).

Second, name-based studies can shed light on the extent of mobility for countries and historical periods for which intergenerational panels with direct family links are not available. For example, using cross-sectional Census data, Long and Ferrie (2013) and Olivetti and Paserman (2015) find that while the U.S. may have been characterized by high intergenerational mobility in the 19th century, mobility was lower in the early 20th century. Clark (2014) and others provide evidence on the extent of intergenerational mobility for a number of countries and time periods for which few if any other estimates are available. And Barone and Mocetti (2020) use the R^2 estimator to show that intergenerational mobility in the Italian city of Florence may have been much lower during the 15th century than in modern times. Name-based studies can therefore greatly expand our knowledge

about how intergenerational processes vary across time and countries.

Third, name-based methods can be used to characterize the geography of intergenerational processes in greater detail. Following the influential work by Chetty et al. (2014), a number of studies compare how mobility processes vary across regions within countries, based on large-scale administrative data. This work is interesting from a descriptive perspective, but also opens the door for causal research designs to estimate how regional characteristics affect social mobility. Unfortunately, the type of administrative data used in these studies is not available for most countries. Name-based estimators however can be used to compare mobility rates across regions based on more standard data sources, such as Census data. For example, Güell et al. (2018) use the R^2 estimator in cross-sectional data to study how mobility rates differ between Italian provinces.

3 Data

To compare name-based estimators in the type of setting for which they were designed for (in which linkages based on individual identifiers are uncommon), we use historical data sources from Finland and the United States. Our main source are longitudinal records from the turn of the 19th century and the 20th century in Finland, which for a number of reasons are well suited for studying the performance of name-based methods. First, they include various socioeconomic outcomes for individuals of two adjacent generations, complete names, and direct father-son links for estimation of a benchmark mobility measure. Second, the first decade of the 20th century was a particularly active period of surname changes in Finland. Such name changes are recorded in our data, allowing us to explore how they affect each method. Third, previous studies suggest a large decline in social mobility around the turn of the century in the U.S. and the UK (Long and Ferrie, 2013; Olivetti and Paserman, 2015; and Feigenbaum, 2018), making this period an interesting window for mobility research also in other countries. In particular, we will compare the social mobility of the two antagonistic sides that fought against each other in the Finnish Civil War of 1918.

To illustrate that our key results are generalizable, we replicate them in linked records from the U.S. Census. First, we use the *IPUMS Linked Representative Sample 1880-1900*, which links records from the 1880 complete-count to 1% samples of the 1900 U.S. Census. The data contain complete names, as well as the occupational mean income of both father's and sons. Olivetti and Paserman (2015) report results from these data in Table 3 of their study. They also provide replication files to reconstruct their samples, which we use here. As in their study, our analyses are restricted to white father-son pairs in which the son was aged 0-15 in 1880. These restrictions and the requirement of non-missing values for log occupational income (based on 1950 median occupational earnings calculated by

IPUMS from a 1950 Census Report) for both generations renders a sample size of 9,076 observations. Finally, we use the digitized Iowa State Census 1915 Sample (Goldin and Katz, 2000) linked to the 1940 U.S. Federal Census by Feigenbaum (2018). Feigenbaum restricts the analysis to father-son pairs in which the son was aged 3-17 in 1915, resulting in 3,204 father-son pairs with non-missing values for log occupational income.[3]

3.1 Finnish Longitudinal Veteran Database

We assemble our main sample by linking individual-level data on veterans of the Finnish Civil War 1918 from various sources of archives of the National Archives of Finland. The data base, named the *Finnish Longitudinal Veteran Database*, contains 16,318 individuals born between 1865 and 1904 who survived the Civil War 1918. It includes information on first names, surnames, schooling, occupation, parental occupation, demographic characteristics, and the side on which the individual fought in the Civil war. After dropping all females and males with missing occupation our analytic sample contains 14,811 individuals, of which 6,507 fought in the Red Guard and 8,304 fought in the White Guard. We observe father's occupation for 7,051 father-son pairs through the son's self-report of his father's occupation. These self-reported links are complemented with matched links from digitized genealogy records, matching the individuals in our data to their own birth certificates as stored at www.ancestry.fi based on complete name, date of birth and place of birth.[4] These birth certificates also contain father's occupation. In total, 1,864 successful matches were made. Table 3 reports the sample and name characteristics. Appendix A describes in detail the individual registries from which the variables were acquired.

Coding of Names. We used the first of up to three given names, henceforth, first name. Surnames were cleaned from obvious spelling mistakes. We further harmonized the first name so as to account for different spelling forms of one and the same phonetic name. We differentiated between Finnish and Swedish spelling forms in order not to forego the socioeconomic content that the language may convey.

Measuring Socioeconomic Status. We use two quantitative measures of socioeconomic status: occupational status and years of schooling. We observe occupation as of 1918, referring to occupation at the time of enrollment in the troops for the civil war,

[3] Olivetti and Paserman (2015) and Feigenbaum (2018) both use the same 1950-based occupational income measure from IPUMS.

[4] Our study sample was matched to digitized birth certificates obtained from the www.ancestry.fi maintained by the Genealogical Society of Finland (http://hiski.genealogia.fi/hiski/93id4x?en) using a matching algoritm developed specifically for this purpose by Eric Malmi (http://ancestryai.cs.hut.fi/). The universe of birth certificates for the years 1850-1900 are digitized for 41 parishes out of 194 parishes in total.

for everyone in the study sample. Members of the White Guard also reported their occupation in midlife (as of the mid-1930s). Our preferred measure of occupational status is HISCAM, a one-dimensional social stratification scale adapted from Cambridge Social Interaction and Stratification (CAMSIS) that is based on the Historical International Standard Classification of Occupations (HISCO) developed by Miles, Leeuwen and Maas (2002). The CAMSIS approach uses patterns of social interaction to determine the position of an occupation in the overall hierarchy, mainly using information on marriage and partner selection (Lambert et al., 2013).[5] In the absence of a country-specific version for Finland we use the universal scale of HISCAM, which is standardized to have a mean of 50 and a standard deviation (s.d.) of 15 in a nationally representative sample of individuals. In our full sample (n=14,754), the HISCAM score based on occupation in 1918 has a mean of 51.3 (std. dev. 10.80). The HISCAM score as of the 1930s was only recorded for members of the White Guard (n=8,680) and has a mean of 59.6 (std. dev. 15.8). Table 3 reports the summary statistics for our socioeconomic outcomes and background covariates. Years of schooling is coded as number of completed years of schooling based on the self-reported highest completed level of education. Each reported category of education, e.g. compulsory schooling, was coded according its default duration during the period of study. Moreover, individuals who did not complete the reported highest level of education were asked to report the number of years completed.

The Distribution of Names. Table 3 reports summary statistics, separately for veterans of the White Guard and Red Guard. In both samples, the share of individuals with singleton first names is roughly 2.5 percent. The first name distribution is more compressed among the Red Guard, with roughly 76 percent of the individuals having a first name that ranks within the 50 most popular names. Rare surnames are more common than rare first names, and roughly 30 percent of the individuals in our data have a unique surname. As for first names, the surname distribution is more compressed among the Red Guard veterans, with 26.1 percent of all individuals having a top-50 ranked surname as compared to 11.6 percent among the White Guard veterans. The difference in the first name and surname distributions is illustrated further in Figure 1, which shows that surnames have a right-skewed distribution while first names do not.

[5] Individuals who are socially close to one another are more likely to interact and form marriages than individuals who are socially far apart. The CAMSIS project website (http://www.camsis.stir.ac.uk/index.html) describes one methodological approach to deriving a social interaction distance scale based on occupational data.

Table 3: Summary Statistics

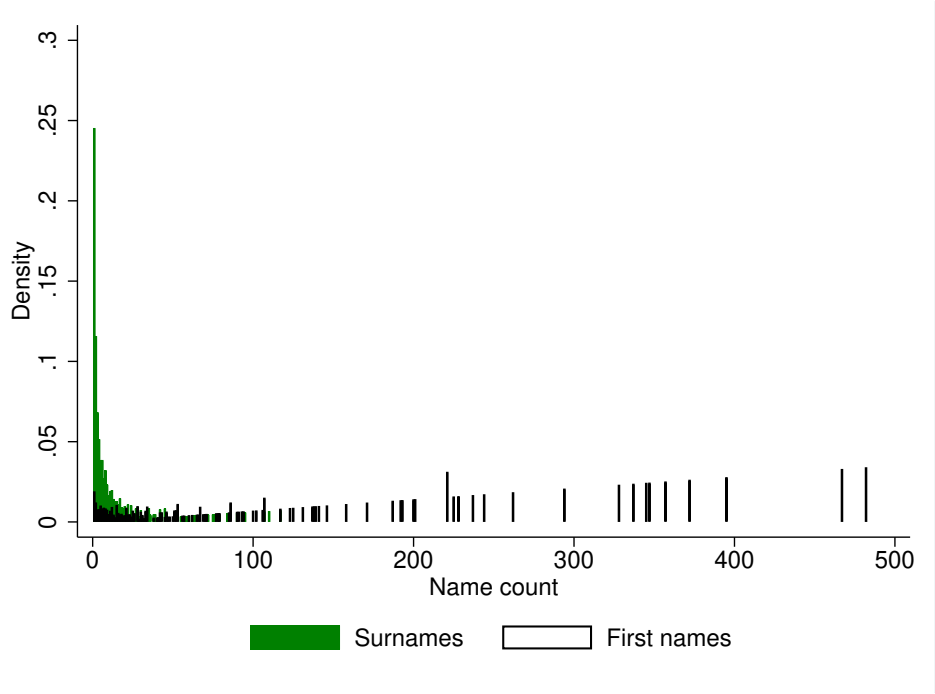
	Red Guards	White Guards
Number of sons	6,666	9,652
Linked fathers (self-reported)	–	7,051
Linked fathers (birth records)	1,117	1,389
First names		
Number of distinct names	426	585
Mean frequency per name	15.6	16.5
Sons with singleton first name	2.5%	2.2%
Top-50 names	76.1%	66.5%
Surnames		
Number of distinct names	2,861	4,404
Mean frequency per name	2.3	2.2
Sons with singleton surname	30.2%	29.6%
Top-50 surnames	26.1%	11.6%
Socioeconomic Outcomes		
Son's years of schooling	5,851	7,065
mean (std. dev.)	3.24 (1.56)	6.79 (4.85)
Son's occupational status (1918)	6,472	8,282
mean (std. dev.)	47.77 (7.58)	54.01 (12.07)
Son's occupational status (1930)		8,680
mean (std. dev.)		59.64 (15.84)
Father's occupational status		7,012
mean (std. dev.)		55.01 (12.14)
Father's occupational status (BR)	1,117	1,389
mean (std. dev.)	47.88 (5.36)	52.99 (10.18)

Note: Father's occupational status for the members of the Red Guard is only available from birth records (BR).

4 The Informational Content of Names

Name-based mobility studies start from the observation that names predict socioeconomic status. Both first names and surnames are informative though for different reasons. The informational content of surnames stems from a mechanical process—children *inherit* their surname, along with other factors that influence their socioeconomic status. In contrast, the informational content of first names results from deliberate action. In other words, parents *choose* a name for their child, and that choice correlates with status. However, first names do not simply capture parental socioeconomic status, they also capture name preferences *conditional* on that status. Because these preferences may be intertwined with the mobility process itself, more detailed arguments are necessary to justify the use of first names in mobility research (Olivetti and Paserman 2015).

Figure 1: The Sample Frequency of First and Surnames



As surnames are more straightforward to use, they have been the more popular choice in such work (see Table 2).[6] Yet, the contrast between first names and surnames is not as sharp as it may initially seem. First, surnames too are ultimately a matter of individual choice. In fact, Güell, Rodríguez Mora and Telmer (2015) note that name *mutations*—deliberate or accidental—are essential for surnames to retain their informational content. In the absence of such mutations, the surname distribution would eventually collapse into a small number of frequent and uninformative last names such as "Smith" and "Jones". The informational content of surnames thus also depends on choices, albeit less directly. Second, while individual choice creates conceptual difficulties, it is attractive from a predictive perspective. Indeed, while first names may remain informative irrespective of their frequency, the predictive power of surnames becomes negligible for frequent last names, due to their mechanical transmission. It is therefore not a priori obvious whether first names or surnames are more useful for mobility research.

4.1 The R^2 Estimator

Most name-based studies compute the average socioeconomic status of each name, to then estimate a traditional intergenerational regression on the name-level (see Section 5). The informational content of names plays only an implicit role in this two-step *group-*

[6] Surnames have also been used in other fields, such as population genetics and health sciences, to link current and previous generations (see Collado, Ortín and Romeu, 2008).

ing estimator. However, Güell, Rodríguez Mora and Telmer (2015) have developed an alternative method to estimate social mobility in the absence of direct child-parent links. They show that researchers can make inference about intergenerational mobility without running a single intergenerational regression, by quantifying the informational content of surnames in cross-sectional data. Intuitively, if socioeconomic status is strongly transmitted, then surnames should explain a large share of its variance – the R^2 in a regression of socioeconomic status on name dummies is increasing in the degree of intergenerational transmission. This *"R^2 estimator"* has been also applied in Barone and Mocetti (2020) and Güell et al. (2018), and we show below that a corresponding estimator can also be constructed from *first* names.

Specifically, Güell et al. (2015) estimate a linear regression of the economic status y_{is} of individual i with surname s on a set of surname dummies,

$$y_{is} = \beta' Surname_s + \gamma' X_{is} + \varepsilon_{is}. \tag{2}$$

The vector X_{is} may include sociodemographic characteristic such as region of birth, year of birth, ethnicity and – in our application – the side on which the individual fought in the civil war. In order to confirm the true *incremental* information that surnames carry, the R^2 obtained from this regression is contrasted against a placebo R^2_P from an otherwise identical regression as (1), in which surnames are reshuffled across individuals (while maintaining their marginal distribution). Güell et al. define the *informational content of surnames (ICS)* as the difference between the true and the placebo R^2

$$\text{ICS} \equiv R^2 - R^2_P. \tag{3}$$

While not directly comparable, Güell et al argue that the ICS is monotonically increasing in the degree of intergenerational status persistence on the individual level.[7] The ICS can therefore be used to compare intergenerational mobility across time, regions or groups.

The R^2 estimator has both intuitive and counterintuitive aspects. The idea that surnames contain socioeconomic information is straightforward: surnames are passed on from parents to their children, along with other characteristics that affect economic status. While surnames can be subject to choice (Collado, Ortín and Romeu, 2008), familial linkages account for much of the partition into surnames. The insight that disruptions in this process are imperative for surnames to retain their socioeconomic content is perhaps less intuitive. A stationary process would dilute the informativeness of surnames over time. But the surname distribution is almost universally skewed, with a large share of the surnames being held by few individuals and a small share of the surnames held by a

[7] In support of this result, Acciari, Polo and Violante (2016) show that direct estimates of intergenerational mobility from Italian tax data correlate with name-based estimates presented in Güell et al. (2018).

large share of individuals. This skewness is generated by a birth-death process through which some surnames become extinct (e.g., because family members fail to reproduce on the male lineage) and new names are being created by migration or name mutations (see Section 6.6).

The key advantages of the R^2 estimator are its modest data requirements and its broad applicability. The estimator requires only cross-sectional data, and does not require intergenerational panel or the observation of family links. Moreover, a single cross-section may be sufficient for the estimation of mobility trends. For example, Güell et al (2015) study the evolution of intergenerational mobility and assortative mating in Catalonia, exploiting that a single Census contains information on a wide range of birth cohorts. While the R^2 is less closely linked to the direct estimator, we illustrate that it has a number of technical advantages compared to alternative approaches. In particular, it is less sensitive to sampling properties (see Section 5) and is not sensitive to sample size (see Section 6.3).

4.2 The Informational Content of Surnames

Panel A in Table 4 reports estimates from our full sample. Column (1) reports an OLS regression of occupational status against surname dummies. Column (2) adds indicator variables for individuals belonging to the White guard (vs. the Red guard), for ethnicity as proxied by an indicator for Finnish sounding surname (Swedish being the predominant ethnicity in the reference category), and year of birth to the model. Column (3) further controls for place of birth by including county fixed effects. We aggregate the individuals' geocoded parishes of birth into 10 synthetic counties by k-medoid clustering.[8] We also estimated identical models with years of schooling as the socioeconomic outcome instead of occupational status, with similar results.

The implied ICS reported is the difference between the adjusted R^2 from these regressions (reported) and the adjusted R^2 from the corresponding placebo regression in which the surname dummies are randomly reshuffled.[9] Because their analysis is based on a full-count Census, Güell, Rodríguez Mora and Telmer (2015) reshuffle names only once and do not address sampling uncertainty. The method therefore needs to adapted for applicability in settings with more limited sample size, and we discuss such refinements in Section 6.3. First, we reshuffle the surname dummies 1,000 times and report the *mean* ICS across these repetitions in Table 4. Second, we report 95% confidence intervals that

[8] Birth places were linked to geocodes acquired from the Linked Data Finland portal. Geocoded birth place information was clustered using PAM (Partitioning Around Medoids) algorithm.

[9] Güell, Rodríguez Mora and Telmer (2015) consider the difference in the *adjusted* R^2 in their applications, as we do here. The outcome of individuals with a unique surname (*singletons*) can be predicted perfectly irrespectively of whether the name dummies represent actual or placebo names. Because singletons are not a random subgroup of the population, they nevertheless contribute to the ICS (see Section 6.2). We retain singletons in our analysis but our results are not sensitive to their inclusion.

Table 4: The Informational Content of Surnames and First Names

	Dependent variable: Son's occupational status					
	ICS			ICF		
	(1)	(2)	(3)	(4)	(5)	(6)
Surname dummies	Yes	Yes	Yes	Yes	Yes	Yes
Demographic controls		Yes	Yes		Yes	Yes
Region of birth (county/state)			Yes			Yes
Panel A: Finnish Longitudinal Veteran Database (n=14,754)						
AR2	0.206	0.293	0.308	0.065	0.197	0.215
Implied informational content	0.206	0.165	0.147	0.065	0.069	0.056
Bootstrapped 95% CI	[0.163, 0.246]	[0.131, 0.199]	[0.116, 0.181]	[0.045, 0.088]	[0.053, 0.085]	[0.042, 0.069]
Panel B: IPUMS Linked Representative Sample 1880-1900 (n=9,076)						
AR2	0.125	0.165	0.223	0.045	0.081	0.132
Implied informational content	0.082	0.082	0.081	0.035	0.034	0.021
Bootstrapped 95% CI	[0.034, 0.131]	[0.040, 0.133]	[0.046, 0.136]	[0.015, 0.057]	[0.012, 0.051]	[0.0002, 0.041]
Panel C: Linked 1915 Iowa State Census Sample (n=3,841)						
AR2	0.171	0.175	0.180	0.013	0.016	0.022
Implied informational content	0.172	0.171	0.171	0.013	0.012	0.012
Bootstrapped 95% CI	[0.098, 0.244]	[0.102, 0.243]	[0.102, 0.245]	[-0.023, 0.050]	[-0.023, 0.048]	[-0.021, 0.047]

Note: The table reports statistics from a regression of son's occupational status (HISCAM) (Panel A) or log occupational income (Panels B and C) on a set of name dummies and control variables. The implied informational content is the difference between the adjusted R-squared (AR2) reported in the column and the AR2 of an otherwise identical regression in which the surname dummies are randomly reshuffled. Panel A reports estimates from the Finnish Longitudinal Veteran Database. Demographic controls include a dummy for ethnicity (Finnish-sounding name), a dummy for White Guard (the reference cagtegory being the Red Guard), and year and region of birth. Panel B reports estimates from the IPUMS Linked Representative Sample 1880-1900 of the U.S. Census (Olivetti and Paserman, 2015). Demographic controls include an indicator for immigrant status, dummies for country of birth, and year of birth. Panel C reports estimates from the Linked 1915 Iowa State Census Sample, which is a digitized sample of the 1915 Iowa State Census (Goldin and Katz, 2000) linked to the 1940 US Federal Census (Feigenbaum, 2018). Demographic controls include an indicator for immigrant status, dummies for country of birth and year of birth. 95% confidence interval across 1,000 bootstrap samples in brackets.

are based on 1,000 bootstrap samples. In each round, we draw cluster of observations on the surname level with replacement. We then report the 2.5 and 97.5 percentile of the resulting distribution.

The ICS is somewhat higher for educational than for occupational status, and varies substantially with the choice of control variables. Region of birth and ethnicity, as proxied by a dummy for a Finnish sounding surname, are particularly important. The ICS in the

regression without controls is 20.6% but falls to 14.7% when including demographic controls and county of birth. Ethnicity does not seem to affect the ICS much once region is accounted for in our context. This is partly because region and ethnicity are overlapping, as the ethnic Swedish minority mostly lived along the coastline. The informational content of surnames in our Finnish sample is therefore partially due to their covariance with region of birth and ethnicity. As a qualitative result this is not a concern, as intergenerational persistence on the individual-level likewise reflects variation across regions and ethnic groups. The concern is however that the R^2 estimator weights these factors more heavily than the direct estimator. Practitioners should therefore check whether comparative findings are robust to the inclusion of group-level control variables, in particular those related to regional and cultural differences in naming conventions (we return to these considerations in Subsection 6.5). Surnames still retain substantial explanatory power when abstracting from such factors.

Panels B and C in Table 4 report the corresponding estimates from two other samples, the IPUMS Linked Representative Sample 1880-1900 of the U.S. Census (Olivetti and Paserman, 2015) and the Linked 1915 Iowa State Census Sample (Feigenbaum, 2018). The ICS is smaller in these samples compared to our main sample, and less sensitive to the inclusion of demographic and regional controls. As in the Finnish sample, surnames explain a substantial share of the variation in occupational status.

4.3 The Informational Content of First Names

The R^2 method proposed by Güell, Rodríguez Mora and Telmer (2015) is based on the informational content of surnames. However, first names also carry informational content, as parent's active name choice correlates with parental socioeconomic characteristics. Similarly to our analysis of surnames, we compare a linear regression of the socioeconomic status y_{in} of individual i with first name n on a set of first name dummies,

$$y_{in} = \beta' Firstname_n + \gamma' X_{in} + \varepsilon_{in}. \tag{4}$$

to a placebo regression in which those dummies are reshuffled across individuals. The *informational content of first names (ICF)* is again defined as the difference in the (adjusted) R^2 between the two regressions,

$$\text{ICF} \equiv R^2 - R^2_P. \tag{5}$$

It is apriori not clear whether first names or surnames have higher informational content. On the one hand, first names are more selective, and may therefore encode more information. As noted by Clark et al. (2015): "*First names carry much more information typically about family status at the time of birth than do surnames. This is because the*

surname links someone to the status of some distant ancestor, while the first name gives information about the status of parents at the time of birth." On the other hand, first names are less dispersed, with the average group size being ten times larger for first than for surnames in our sample.

We report estimates of the ICF in Columns (4) to (6) of Table 4. They confirm that a non-negligible share of the variation in socioeconomic status across individuals can be explained by their first names. The structure of the table follows the corresponding table for surnames. Column (1) reports an OLS regression of occupational status against first name dummies. Column (2) adds demographic controls, and Column (3) further controls for county of birth. The implied ICF is the difference between the adjusted R^2 reported in the column and the mean adjusted R^2 of 1,000 placebo regression (not reported).

The informational content of first names is lower than for surnames in our sample, but the two estimators otherwise follow a similar pattern – both decrease substantially when place of birth fixed effects are included, and are larger for years of schooling (not shown) than for occupational status. When including the full set of controls, the ICF is estimated to be 5.6% for occupational status and 7.0% for years of schooling. The ICF is smaller in the other two samples from the U.S. Census (Panels B and C), but remains positive and statistically significant.

While the theoretical motivation for the ICS does not generalize to the ICF, both are informative in practice. Moreover, we show below that both can capture mobility differences between groups in a typical application (see Section 7). The ICS and ICF-variants of the R^2 estimator are therefore promising measures for intergenerational mobility research. However, as other name-based estimators, they are subject to several conceptual caveats that warrant attention. In particular, the ICS and ICF are sensitive to the frequency of names, and names have *added informational content* over and above their role as proxies for observable socioeconomic status. We review those caveats in Section 6.

5 The Grouping Estimator

A more commonly used approach estimates a variant of the standard intergenerational regression, in which name-group averages are used to replace the unavailable parental outcomes. The informational content of names motivates the grouping principle of this two-step estimator, and its first step corresponds to the name dummy regression (2) underlying the R^2 method. While differently framed in the literature, we observe that all studies use the same type of estimator – a Wald or *grouping* estimator, in which groups are defined by first names or surnames. But despite using the same estimator, prior work has produced very different estimates, an observation that we aim to rationalize here.

We first link the grouping estimator to the conventional *direct* estimator in equation

(1), and show that their relation crucially depends on the sampling properties of the underlying data. In particular, the grouping estimator tends to be larger than the direct estimator if the families in the parent and child samples *overlap*, such that an offspring is sampled whenever his or her parents are sampled as well. In contrast, the grouping estimator can be much smaller than the direct estimator when the two samples do not overlap fully, as in cross-sectional data with partial coverage of the population.

The grouping estimator therefore behaves very differently depending on whether or not the parent and child samples overlap. Moreover, in overlapping samples, the grouping estimator is remarkably insensitive to other properties of the data, such as sample size, name frequencies, or the informational content of names. Meanwhile, these properties do matter in non-overlapping samples. As such, existing estimates are not directly comparable – which helps to explain why grouping estimates are larger than direct estimates in some studies while others find the reverse pattern.

We also link the grouping estimator to the R^2 estimator presented in the previous section. A low informational content of names corresponds to a "weak" first stage in the grouping estimator. Interestingly, this is not so much of an issue if the parent and child samples overlap. In such settings, the grouping is a standard 2SLS estimator, and is biased towards the OLS estimator; yet such bias is desirable if the (feasible) grouping estimator is meant to approximate the (infeasible) direct OLS estimator. However, if the parent and child samples do not overlap the grouping estimator instead corresponds to a split-sample IV estimator, and is biased towards *zero* (Choi, Gu and Shen, 2018). We show that the resulting bias can be large in typical applications, as well as propose a simple bias correction procedure that accounts for the overlap between the parent and child samples.

5.1 The Grouping Estimator

The grouping estimator has been presented in different ways in the literature. Clark (2014) and related studies (such as Clark and Cummins, 2014) consider regression to the mean on the *surname* level. In a first step, the average socioeconomic status across individuals within each name and generation is computed. In a second step, the mean status in one generation is regressed on the mean in the previous generation. Others consider instead a two-sample two-stage least squares (TS2SLS) estimator, instrumenting parent's status in equation (1) with a set of first name (Olivetti and Paserman, 2015) or surname dummies (Barone and Mocetti, 2020). However, a two-stage least squares estimator based on a set of dummy variables is tantamount to running a weighted linear regression on a set of group means, and is therefore also called the "grouping" estimator.[10] The approach by

[10] This equivalence is underscored by the standard Wald estimator based on a binary instrument, which scales the bivariate regression with binary explanatory variable by a simple difference of two group means. Indeed, a weighted regression on group means can be understood as a linear combination of all Wald estimators that can be constructed from pairs of means (Angrist and Pischke 2008).

Clark and co-authors is therefore equivalent to the instrumental variable approach used in more recent studies, as long as group means are appropriately weighted.[11] Accordingly, we adopt the label *grouping estimator* for either approach. The TS2SLS perspective remains useful, and we return to it below.[12]

We compare estimates from the "direct" regression of the child's socioeconomic status y_{ij} in family i with first or surname j on the parent's status x_{ij},

$$y_{ij} = \beta x_{ij} + \epsilon_{ij}, \tag{6}$$

with the corresponding grouping estimator, in which x_{ij} is replaced by a group mean $\bar{x}_j$ defined by a child's first name or surname,

$$y_{ij} = \delta \bar{x}_j + u_{ij}. \tag{7}$$

We argue that the properties of the group coefficient δ depend crucially on if the group mean is defined over the parents of the sampled children, or over *other* individuals who merely share the same name.[13] Specifically, its level and interpretation depends on the *overlap* between the parent and child samples, defined as the probability p that a parent is sampled if his or her parent is contained in the child sample, i.e.

$$p \equiv P(i \in \text{parent sample} \,|\, i \in \text{child sample}).$$

We start by considering the two polar cases. Consider first the "*short*" group-level regression

$$y_{ij} = \pi \bar{x}_{ij} + v_{ij}. \tag{8}$$

where $\bar{x}_{ij}$ with subscript i represents the "*inclusive*" mean that averages over the parents of sampled children (including i). Equation (8) is the relevant object if families in the parent and child samples *overlap* exactly ($p = 1$), as in Chetty et al. (2014). Such exact

[11] Since the two approaches produce numerically identical coefficients, it might seem easier to construct the group means manually, which is computationally much simpler. Considering this question in the context of quasi-experimental "examiner" designs, Hull (2017) however notes that the manual approach may lead to inflated first-stage F-statistics, as it understates the true dimensionality of the underlying instruments – a severe bias from weak instruments may therefore go undetected. This argument is particularly relevant in the intergenerational context as names are only weak predictors of socioeconomic status.

[12] The TS2SLS estimator (i) takes uncertainty from the first stage into account and (ii) automatically weights name groups by their frequency, while other implementations may not. However, the TS2SLS perspective also has its pitfalls. As Olivetti, Paserman and Salisbury (2018) note in response to a critique by Choi, Gu and Shen (2018), names are unlikely to be a valid instrument in the sense of satisfying the exclusion restriction. Moreover, we show that the properties of the grouping estimator depend critically on the extent to which the parent and child samples overlap, on which the TS2SLS perspective imposes a polar assumption.

[13] We abstract from the intercept by expressing all variables as deviations from their mean. Note that the coefficient δ does not depend on whether one also takes means of the left-hand side variable.

overlap also occurs if the grouping estimator is applied in complete-count census data, or data that track families according to some fixed criteria (abstracting from international migration, variation in fertility rates, survey attrition, etc.).

In other settings, the parent and child samples might not overlap fully – for a family i in name group j we might observe an ancestor or a descendant but not both. To illustrate such settings, consider the group-level regression in which the group mean is constructed over individuals who are *not* parents of the sampled children. We can approximate such settings by

$$y_{ij} = \kappa \bar{x}_{(i)j} + w_{ij} \tag{9}$$

in which $\bar{x}_{(i)j} = \frac{N_j \bar{x}_j - x_{ij}}{N_j - 1}$ represents the "*leave-out*" mean in a name group of size N_j, in which each descendant's own ancestor is excluded. It corresponds to the predicted value of x_{ij} underlying the jackknife instrumental variables (JIVE) estimator (Kolesár et al., 2015). Equation (9) represents the grouping estimator in settings in which there is zero or only negligible overlap between the parent and child samples ($p = 0$), for example because each sample is a small and independent draw from the population.[14]

Table 5 reports estimates from equations (6), (8) and (9), separately for our *main sample* (Panel A), the *IPUMS Linked Representative Sample 1880-1900* (Panel B) and the *linked 1915 Iowa State Census Sample* (Panel C). Column (1) reports the direct estimates based on equation (6), which are $\hat{\beta} = 0.600$ in the Finnish sample, and $\hat{\beta} = 0.474$ and $\hat{\beta} = 0.441$ in the two U.S. Census samples. The next columns report variants of the grouping estimator, with group means defined over surnames in columns (2)-(4) or first names in columns (5)-(7). For comparability, the grouping estimators are based on the same sample as the direct estimator.

In columns (2) and (5), we report estimates based on equation (8) and "*inclusive*" means. These grouping estimates are always larger than the corresponding direct estimates ($\hat{\pi} > \hat{\beta}$). The gap is larger in the Finnish compared to the U.S. data, and greater for first than for surnames. The grouping estimator is however not necessarily larger than the direct estimator, contrary to such suggestions in the prior literature.[15] In columns (4) and (7), we report estimates based on equation (9) and "*leave-out*" means with no overlap between the parent and child sample. These estimates are always smaller, and often much smaller, than the corresponding estimates based on the inclusive mean ($\hat{\kappa} < \hat{\pi}$).[16] They

[14] We approximate such settings with the leave-out mean to abstract from changes in sample size. In addition, we verified our results using a split-sample grouping estimator.

[15] See also Olivetti and Paserman (2015), who highlight important sources of downward bias in their grouping estimator, such as measurement error induced by imputed father's occupational status or the intergenerational transmission of unobservable characteristics not captured by first names.

[16] The size of this gap depends also on sample size as we discuss below. We therefore find an even larger gap when using a a split-sample IV estimator that splits our sample into separate first- and second-stage samples with zero overlap – in particular for surname groups, of which many small ones are dropped from the analysis due to unsuccessful matches between the first stage sample and the second stage sample. For example, using the IPUMS linked representative sample the grouping estimates are 0.095 and 0.195 for

Table 5: Direct vs. Grouping Estimators

	Dependent variable: Son's occupational status						
	Direct	Surnames			First names		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Group definition	–	inclusive	partial	leave-out	inclusive	partial	leave-out
Overlap		100%	75%	0%	100%	75%	0%
Panel A: Finnish Longitudinal Veteran Database							
Father's occupational	0.600	0.640	0.560	0.322	0.763	0.739	0.607
status (HISCAM)	(0.014)	(0.017)	(0.022)	(0.028)	(0.029)	(0.035)	(0.032)
AR2	0.246	0.199	0.144	0.033	0.102	0.092	0.057
N	5,986	5,986	3,871	3,852	5,986	4,446	5,821
Panel B: IPUMS Linked Representative Sample 1880-1900							
Father's log	0.474	0.479	0.384	0.179	0.501	0.425	0.224
occupational income	(0.012)	(0.015)	(0.019)	(0.024)	(0.027)	(0.032)	(0.036)
AR2	0.149	0.103	0.067	0.011	0.038	0.026	0.005
N	9,076	9,076	5,666	5,119	9,076	6,530	8,051
Panel C: Linked 1915 Iowa State Census Sample							
Father's log	0.441	0.446	0.425	0.381	0.533	0.460	0.215
occupational income	(0.021)	(0.024)	(0.027)	(0.032)	(0.037)	(0.047)	(0.057)
AR2	0.142	0.112	0.103	0.073	0.053	0.041	0.006
N	3,204	3,204	2,080	2,317	3,204	2,255	2,781

Note: The table reports the coefficients from a regression of son's occupational status (HISCAM) (Panel A) or log occupational income (Panels B and C) on the father's corresponding occupational status (column 1) or the mean of the father's status in the name group, defined by son's surname (columns 2-4) or first name (columns 5-7). Panel A reports estimates from the Finnish Longitudinal Veteran Database (White Guard only). Panel B reports estimates from the IPUMS Linked Representative Sample 1880-1900 of the U.S. Census (Olivetti and Paserman, 2015). Panel C reports estimates from a digitized sample of the 1915 Iowa State Census (Goldin and Katz, 2000) linked to the 1940 US Federal Census (Feigenbaum, 2018). Standard errors in parentheses.

are either greater (Panel A, first names) or smaller than the direct estimates $\hat{\beta}$ (all other cases). The gap between the inclusive and leave-out variants is particularly large in the linked U.S. Census samples (Panel B and Panel C). In the IPUMS Linked Representative Sample 1880-1900, the surname-based grouping estimator is nearly three times larger when constructed from inclusive means ($\hat{\pi} = 0.479$ vs. $\hat{\kappa} = 0.179$).

The inclusive mean $\bar{x}_j$ with full overlap and leave-out mean $\bar{x}_{(i)j}$ with zero overlap represent the two polar cases. In applications, generations are often defined by repeated cross-sections, with partial overlap between the parent and child samples. To illustrate such intermediate cases, columns (3) and (6) report split-sample grouping estimates based on parent and child samples that overlap by 75 percent ($p = 0.75$).[17] As expected, the

surnames and first names, respectively.

[17] Starting from the sample used for the direct estimator of size N (e.g. $N = 9,076$ observations in the IPUMS linked representative sample 1880-1900), we draw without replacement two sub-samples of size

estimates are in between the two polar cases. These findings are robust to specification choice and hold for alternative outcome variables.[18]

5.2 The Grouping Estimator in Overlapping Samples

The interpretation of the grouping estimator – and its comparability to the conventional estimator – depends therefore crucially on sampling properties of the underlying data. We formalize the relation between the direct and grouping estimators to rationalize this observation. Our arguments resemble arguments from the literature on peer effects, in which grouping estimators have often been misinterpreted (Angrist 2014). For simplicity, the exposition is in terms of population moments.

Consider first the "*short*" group-level regression equation (7), based on the "inclusive" mean $\bar{x}_{ij}$ with full overlap ($p = 1$). As we observe both direct family links and names we can also estimate the corresponding "*long*" regression,

$$y_{ij} = \pi_0 x_{ij} + \pi_1 \bar{x}_{ij} + v_{ij}, \tag{10}$$

which includes both direct (individual) and group (name-level) effects. As we discuss in Section 6.4, many different models could rationalize why names have *added informational content* (i.e., $\pi_1 \neq 0$). The outcomes y_{ij} and group means $\bar{x}_{ij}$ in these regressions are sampled from the same families i. Using the omitted variable formula, the relationship between the short and long regression equations can therefore be derived as

$$\pi = \frac{Cov(y_{ij}, \bar{x}_{ij})}{Var(\bar{x}_{ij})} = \frac{Cov(\pi_0 x_{ij} + \pi_1 \bar{x}_{ij}, \bar{x}_{ij})}{Var(\bar{x}_{ij})} = \pi_0 \frac{Cov(x_{ij}, \bar{x}_{ij})}{Var(\bar{x}_{ij})} + \pi_1 = \pi_0 + \pi_1, \tag{11}$$

where the last step follows because the slope coefficient in a regression of an individual variable on its group means equals one by definition. Similarly, the relation between the direct and long regressions is

$$\beta = \frac{Cov(y_{ij}, x_{ij})}{Var(x_{ij})} = \frac{Cov(\pi_0 x_{ij} + \pi_1 \bar{x}_{ij} + v_{ij}, x_{ij})}{Var(x_{ij})} = \pi_0 + \pi_1 \frac{Cov(\bar{x}_{ij}, x_{ij})}{Var(x_{ij})}. \tag{12}$$

The combination of equations (11) and (12) yields

$$\pi = \pi_0 + \pi_1 = \beta + \pi_1 \left(1 - \frac{Cov(\bar{x}_{ij}, x_{ij})}{Var(x_{ij})}\right), \tag{13}$$

$N_s = pN$. For example, for $p = 0.75$ we have $N_s = 6,807$ in the IPUMS linked representative sample. Since some names are not included in both samples, the effective number of observations as reported in Table (5) is below this theoretical upper bound.

[18] Table A2 presents robustness checks in which we replace the log occupational income with log annual earnings or years of education, in regressions that are otherwise analogous to the ones presented in Panel C of Table 5. We again find that the leave-out grouping estimator is smaller than the inclusive variant, and either larger or smaller than the direct estimates.

where the ratio in brackets is smaller than one, because x_{ij} varies within name groups. Accordingly, the *inclusive* grouping estimator will be larger than the direct estimator ($\pi > \beta$) if and only if names have added informational content over and above a parent's observed socioeconomic outcome ($\pi_1 > 0$). It cannot be smaller than the direct estimator as long as π_1 is non-negative, which is plausible for both first and last names. These implications hold regardless of sample size and the extent to which names predict socioeconomic status. Moreover, they follow mechanically, irrespectively of the underlying model of intergenerational transmission.[19]

The results reported in Table 5 are therefore not specific to our samples, but exemplify a general point: the grouping estimator will tend be larger than the direct estimator if child and parent samples contain the same families. For example, this observation rationalizes why the grouping estimator is always larger than the direct estimator in linked U.S. tax data (Chetty et al., 2014).

Surprisingly, names do not need to have *any* informational content for the inclusive grouping estimator to capture intergenerational mobility. To motivate the grouping estimator, previous studies emphasize that names carry systematic information about socioeconomic status. This condition is indeed necessary in some settings but not if the parent and child samples overlap. This result relates to the observations that a TS2SLS estimator applied to fully overlapping samples equates to a conventional 2SLS estimator, and that the 2SLS estimator is biased towards the OLS estimator if the instrument (here: name means) is only a weak predictor of the regressor of interest (see also Section 5.4). Such bias is undesirable in most applications, but not in the context considered here – after all, the TS2SLS is meant to approximate the (infeasible) OLS estimator. Indeed, if names have no systematic informational content, the grouping estimator has the same probability limit as the OLS estimator.

Equation (13) has been similarly derived in Adermon, Lindahl and Palme (2019). It also underlies a critical review of grouping estimators by Güell, Rodríguez Mora and Solon (2018), who argue that π_1 could vary substantially across studies, and that "*[t]his finding sheds light on a puzzle in the existing literature: why do some researchers (such as Clark, 2014) estimate group-level coefficients much larger than the usual individual-level coefficients while others [...] do not?*" However, the equation underlying this argument holds only in one particular setting – if the parent and child samples overlap *completely*. Most studies are instead based on partially overlapping samples, and the grouping estimator behaves very differently in such settings (as illustrated in Table 5 and shown formally in the next section). Accordingly, equation (13) does not always apply, and differences in

[19] To assign a particular interpretation to the observation that a grouping estimator is larger than the direct estimator is therefore conceptually equivalent to assigning a particular interpretation to the observation that names have added informational content. However, this observation may reflect very different theoretical mechanisms (see Section 6.4).

sampling properties might be another important reason why grouping estimates differ so much across studies.

5.3 The Grouping Estimator in Independent Samples

In other applications, the parent and child samples do not overlap, or do not overlap much ($p = 0$). The statistical properties of the grouping estimator turn out to be very different in such settings. To illustrate, consider the "*short*" group-level regression based on the leave-out mean $\bar{x}_{(i)j}$ in equation (9), and the corresponding "*long*" regression

$$y_{ij} = \kappa_0 x_{ij} + \kappa_1 \bar{x}_{(i)j} + u_{ij}. \tag{14}$$

Following the same steps as in the previous section, the relationship between the short and long regression equations is then

$$\kappa = \kappa_0 \frac{Cov(x_{ij}, \bar{x}_{(i)j})}{Var(\bar{x}_{(i)j})} + \kappa_1, \tag{15}$$

and between the direct and the long regression,

$$\beta = \frac{Cov(\kappa_0 x_{ij} + \kappa_1 \bar{x}_{(i)j}, x_{ij})}{Var(x_{ij})} = \kappa_0 + \kappa_1 \frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(x_{ij})}. \tag{16}$$

Finally, combining equations (15) and (16) yields

$$\kappa = \beta + \kappa_1 \left(1 - \frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(x_{ij})}\right) - \kappa_0 \left(1 - \frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}\right) \tag{17}$$

where the ratios in the brackets are again smaller than one.

Equation (17) characterizes the relation between the grouping and the direct estimator when the child and parent samples do not overlap.[20] It suggests that this relation is ambiguous. On the one hand, the added informational content of names κ_1 is likely to be small compared to the informational content in the parent's observed socioeconomic outcomes κ_0. On the other hand, the ratio $\frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(x_{ij})}$ is necessarily smaller than the ratio $\frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}$. As a result, the *leave-out* grouping estimator can either be larger or smaller than the direct estimator (cf. columns (1), (4) and (7) in Table 5) – in contrast to the *inclusive* grouping estimator, which tends to be larger.[21]

[20] Similarly, Olivetti and Paserman (2015) derive the relation between the grouping and the direct estimator under the assumption that the parent and child samples are independent. As such, our equation (17) corresponds therefore closely to equation (2) in their article. Olivetti and Paserman also recognized the distinction between overlapping and independent samples in an earlier draft of their study.

[21] This sharp contrast is somewhat counterintuitive. In sufficiently large samples, the inclusive and leave-out mean should be highly correlated, so why would it matter if one considers one or the other? The two means indeed tend to be highly correlated, even in our modestly sized samples. For example,

If names have no *added* informational content ($\kappa_1 = 0$) the "long" equation (14) collapses into the direct one ($\kappa_0 = \beta$), and equation (17) simplifies to

$$\kappa = \beta \frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}. \tag{18}$$

The leave-out grouping estimator understates the direct estimator in this scenario – again in contrast to the "inclusive" grouping estimator, which collapses into the direct estimator ($\pi = \beta$) if names have no added informational content ($\pi_1 = 0$).

Equation (18) further illustrates that the leave-out variant of the grouping estimator is closely related to the R^2 estimator. If names have low informational content, the leave-out mean $\bar{x}_{(i)j}$ and parental status x_{ij} will not correlate much, and the attenuation factor $\frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}$ and therefore the grouping estimator κ will be close to zero. These implications correspond to the observation that the two-sample grouping estimator applied to non-overlapping samples equates to a split-sample IV estimator, which is biased towards zero when the instruments are weak (Choi, Gu and Shen, 2018).[22] Even if names are very predictive of socioeconomic status, the grouping estimator may still severely understate the direct intergenerational coefficient if the means $\bar{x}_{(i)j}$ are constructed over too few individuals.

5.4 The Grouping Estimator in General Settings

The sampling scheme of many applications falls in between the two polar cases of either complete or zero overlap between the parent and child samples. We therefore consider such cases here, and provide an analytical expression for the grouping estimator in settings with intermediate overlap. The name-based grouping estimator corresponds to a TS2SLS estimator in which the main and auxiliary samples are not independent (contrary to standard assumptions, as e.g. in Inoue and Solon, 2010). As shown by (see Khawand and Lin, 2015), in this more general setting the TS2SLS estimator can be decomposed into a weighted average of the 2SLS and SSIV estimators,

$$\hat{\delta} = \hat{W}\hat{\pi} + \left(1 - \hat{W}\right)\hat{\kappa}, \tag{19}$$

the correlation between the inclusive and leave-out means based on first names is 0.95 in the Finnish and 0.89 in the IPUMS Linked Representative Sample. But while the difference between the two means, $\bar{x}_j - \bar{x}_{(i)j} = \frac{1}{N_j - 1}(x_{ij} - \bar{x}_j)$, becomes small in large name groups, this difference becomes increasingly predictive of child outcomes (because the slope coefficient in the within-name group regression of child outcome y_{ij} on $\bar{x}_j - \bar{x}_{(i)j}$ increases linearly in the name group size N_j). As a result, the properties of the inclusive and leave-out estimator can differ substantially, consistent with the observation that 2SLS and JIVE estimators can be quite different in finite samples (Kolesár et al., 2015).

[22] Alternatively, it can be viewed as an ordinary least square (OLS) estimator with a generated regressor that suffers from classical measurement error (Choi, Gu and Shen, 2018).

where the weight $\hat{W}$ corresponds, approximately,[23] to the share of individuals in the parent sample whose children are contained in the child sample,

$$plim\, \hat{W} = P(i \in \text{parent sample} \,|\, i \in \text{child sample}) = p.$$

The grouping estimator thus behaves as a weighted average of the *inclusive* and *leave-out* versions of the grouping estimator as presented in the previous two sections. In overlapping samples ($p = 1$), the grouping estimator corresponds to a 2SLS estimator and is biased toward the direct (OLS) estimator. In non-overlapping samples ($p = 0$), the grouping estimator corresponds to a SSIV estimator and is instead attenuated towards zero. The overlap p is therefore a key object to interpret grouping estimates. It differs substantially between studies, and should be reported or estimated.

5.5 The Bias-Corrected Grouping Estimator

Moreover, equation (19) points to a simple way to correct grouping estimates for the attenuation bias in partially or non-overlapping samples.[24] Two objects are required for such correction, the overlap p and the attenuation bias from using *other* parents instead of the parents from sampled children to construct the name group means $\bar{x}_j$. The former follows from the way the samples were constructed, while an estimate of the latter is readily available by regressing parent status x_{ij} on its *leave-out* mean $\bar{x}_{(i)j}$.

Specifically, if names have no *added* informational content ($\pi_1 = \kappa_1 = 0$), combining the probability limit of equation (19) and equations (13) and (17) yields

$$\delta = \beta \underbrace{\left(p + (1-p) \frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})} \right)}_{\text{Attenuation Factor}}, \tag{20}$$

where p is the degree of overlap between the parent and child samples, and $\frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}$ is the slope coefficient in a regression of parental outcomes x_{ij} on their leave-out group mean $\bar{x}_{(i)j}$).[25] Dividing a grouping estimator by the term in brackets therefore adjusts for

[23] In finite samples $\hat{W} = \sum_{i \in N_{11}} \bar{x}^2_{j,11} / \left(\sum_{i \in N_1} \bar{x}^2_{j,1} \right)$, where $\bar{x}^2_{j,1}$ are the name group means in the main (e.g., child) sample and $\bar{x}^2_{j,11}$ are the name means among families i that are sampled both in the parent and the child sample. See Khawand and Lin (2015) for a detailed discussion of finite sample properties.

[24] See also Choi, Gu and Shen (2018), who adapt weak-instrument robust inference to the case of two-sample instrumental variable regressions and illustrate that they lead to much larger grouping estimates in intergenerational data. While Choi et al consider the "classic" two-sample setting in which the two samples are assumed to be independent, we allow for partial overlap between the parent and child samples.

[25] For example, if we use a 1% extract from a population-wide Census for both parent and child generation then $p = 0.01$, and we would estimate the attenuation factor $\frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}$ by regressing individual outcomes in the parent sample on their leave-out group mean (using only those names that contribute to the grouping estimator, i.e., dropping those names that feature in the parent but not in the child's

the attenuating effects from limited overlap and sample size. In the next section, we show that this correction procedure works well even in small samples, and when names have added informational content.

5.6 Simulation Evidence

The properties of the grouping estimator depend therefore on many factors, such as the (i) size of the parent and child samples and the overlap between them, (ii) the informational content of names, (iii) the name frequency distribution, and (iv) their *added* informational content (which depends on the underlying data generating process). Moreover, the various factors interact. To illustrate these interdependencies, we provide simulation-based evidence in Figure 2.

We consider two data generating processes. Subfigures (a) and (b) are based on an AR(1) process, i.e., we impose a structural interpretation on the standard parent-child regression in equation (1) and set $\beta = 0.5$. This process is a natural baseline, in that names play no role in the transmission process and therefore have no *added* information content ($\pi_1 = \kappa_1 = 0$). Subfigures (c) and (d), are instead based on a latent factor model, which is underlying the arguments by Clark (2014), and which could rationalize recent evidence on multigenerational correlations across multiple generations (Lindahl et al., 2015; Braun and Stuhler, 2018). The parameters of this model determine the rate of transmission of latent advantages and the signal-to-noise ratio of observed to latent advantages. We choose these parameters such that the implied value for β is similar to the AR(1) model (see figure notes). In a first step, we generate parent and child status for the entire population. We then draw sub-samples of size p separately for the parent and child generations, where p is the population share and therefore the overlap between the parent and child samples.[26] Finally, we estimate the grouping estimator within each sub-sample.

Subfigure (a) is based on a uniform name frequency distribution. We consider three variants of the AR(1) process, with the parental socioeconomic status x_{ij} being randomly distributed such that names have no informational content (*no IC*), name fixed effects explaining ten percent of the variance in parental status (*IC 10%*), or names explaining twenty percent (*IC 20%*). If names have no informational content, the grouping estimator increases linearly in the overlap p between the parent and child samples, fluctuates around zero if there is no overlap ($p = 0$), and has the same probability limit as the direct estimator under full overlap ($p = 1$) – consistent with the analytic expression for the inclusive and leave-out estimators in equations (13) and (18). Intuitively, the grouping estimator always

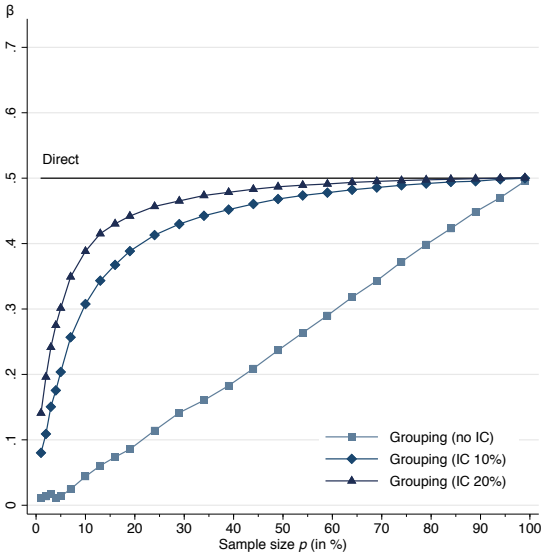
sample).

[26] For simplicity, sample size and overlap are determined by the same parameter. To verify our results, we also considered a variant in which sample size and overlap are set separately.

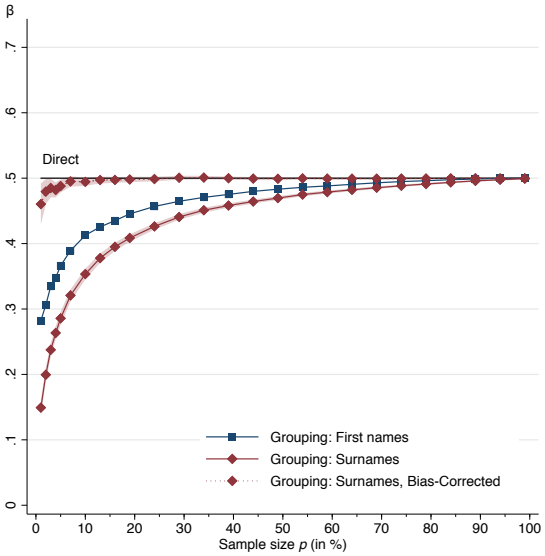
Figure 2: The Grouping Estimator vs. Sampling Probability

Panel A: AR(1) model

(a) Names uniformly distributed

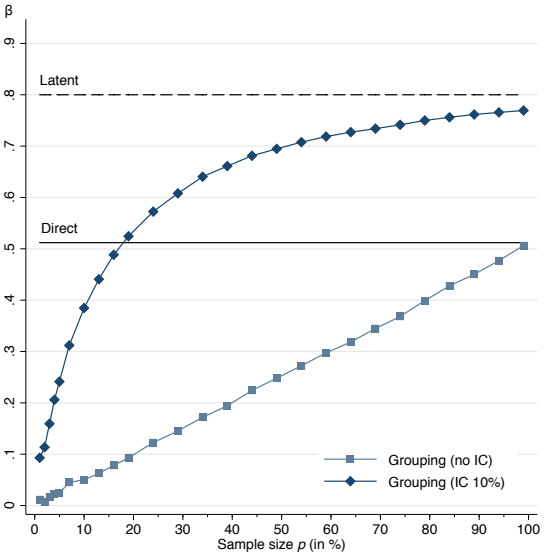


(b) Name distributions from U.S. Census

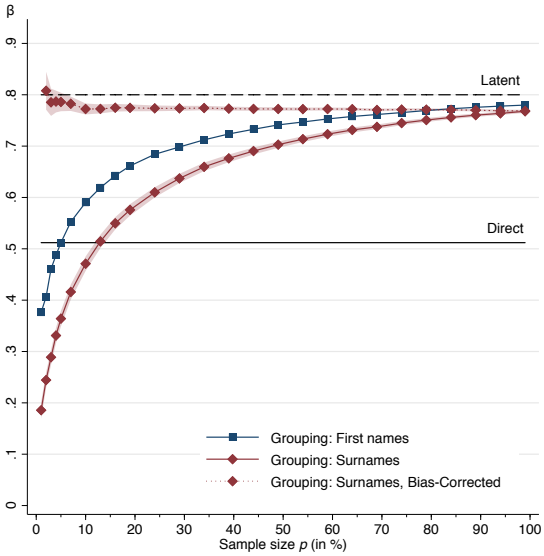


Panel B: Latent factor model

(c) Names uniformly distributed



(d) Name distributions from U.S. Census



Note: Estimates from separate regressions based on differently sized samples (x-axis). Sub-figures (a) and (b) are based on an AR(1) process with slope $\beta = 0.5$, in which parent status is normally distributed with name fixed effects ($\rightarrow$ IC). Sub-figures (c) and (d) are based on a latent factor model given by the equations $y_{it} = \rho e_{it} + u_{it}$ and $e_{it} = \lambda e_{it-1} + v_{it}$, with y_{it} and e_{it} standardized at mean zero and variance one, and $\rho = \lambda = 0.8$ (such that $\beta = 0.8^3 = 0.512$). Sub-figures (a) and (c) are based on a simulated name distribution (30,000 names, uniformly distributed frequency between 1 and 250), while sub-figures (b) and (d) are based on the frequency of female first names and male surnames as observed in the 1920 U.S. Census (see Olivetti and Paserman, 2015). Shaded areas represent 95% confidence intervals.

captures the intergenerational transmission for "complete" parent-child pairs, even if names are not systematically related to socioeconomic status in the cross-section.

If names have informational content (*IC 10%*), the grouping estimator remains positive even when the parent and child have limited overlap. Intuitively, the grouping estimator then also captures part of the intergenerational transmission among "incomplete" pairs, in which either the parent or the child is included in the sample, but not both. That the grouping estimator increases in the informational content of names follows implicitly from equation (18). Figure 2a however illustrates that this increase is highly non-linear in the sample size and overlap p. For example, the group-level estimate is still around 0.4 when sampling 10% of the parent and child generations but drops to 0.25 when taking 5% random samples. The grouping estimator can therefore be rather insensitive to sample size in some settings while being highly sensitive in others, if sample size falls below a certain threshold (the suddenness of this switch depends on how informative names are in the cross-section, cf. *IC 10%* and *20%*). We therefore recommend that researchers test the sensitivity of their estimates to sample size (see Section 6.3).

To study how sensitive the grouping estimator is to the marginal distribution of names, we switch to more realistic name distributions. Specifically, we import the distribution of *surnames* and *female first names* from the 1% sample of the 1920 U.S. Census (as used by Olivetti and Paserman, 2015). We first approximate the name frequency distribution in the complete-count census based on the observed distribution of names in the 1% sample.[27] Figure 2b plots the grouping estimates as based on first names (blue squares) or surnames (red diamonds). The data generating process is again the AR(1) model, with names explaining 10% of the variation in socioeconomic status. The surname-based grouping estimator is more sensitive to sample size than the corresponding estimator based on first names, as the average frequency is much lower for surnames than for first names. The name frequency distribution does not always have a strong impact on the grouping estimator (in line with evidence by Olivetti and Paserman, 2013, Section 7.1). In smaller samples, though, the difference between surname and first name distributions matters, consistent with the observation that the surname-based estimates reported by Olivetti and Paserman (2015) are much smaller than the corresponding estimates based on first names. This leads to the question if researchers should restrict their sample to frequent names. We return to this question in Section 6.1.

In sub-figure (b), we also illustrate the performance of our bias-corrected estimator,

[27] In a first step, we draw from the binomial distribution (with success probability 1%) the simulated frequency of a name in the complete-count Census given the observed frequency of the name in the 1% sample. In a second step, we use the negative binomial distribution to compute the probability that a name with n observations in the complete-count Census is *not* contained in the 1% sample, and create the missing names accordingly. To verify the plausibility of this simulated name distribution we again draw a 1% sample. This simulated 1% sample has a similar name count (12,486 names vs. 12,895 first names) and average frequency per name (11.1 vs. 10.4) as the actual 1% sample.

which adjusts the grouping estimator for the attenuation factor as derived in equation (20). Specifically, we compare the grouping estimates based on surnames as observed in the 1920 U.S. Census with its bias-adjusted version. The standard grouping estimator is heavily attenuated when only sub-samples of the Census are observed. In contrast, the bias-corrected estimator precisely recovers the intergenerational coefficient based on direct parent-child links. If the samples are very small, the bias-corrected estimates have noticeably larger standard errors than the unadjusted estimates. However, the bias-corrected estimates still perform much better in terms of bias and mean-squared error. By reducing the influence of application-specific properties, this correction procedure would also allow researchers to compare their grouping estimates more directly.

Finally, we repeat our analysis using the latent factor model as our data generating process. Figure 2c is again based on a simulated name distribution with uniform name frequency, while Figure 2d is based on the actual name distributions as observed in the U.S. Census. Names have added informational content (AIC) in this model, such that the grouping estimator can now be larger than the direct (conventional) estimator. The intuition is that the grouping "averages away" idiosyncratic variation in status, such that the group means approximate the mean latent status of the respective group (see Clark, 2014, and Section 6.4). Accordingly, in large samples with high overlap, the grouping estimates reflect the persistence in latent advantages (assumed to be $\lambda = 0.8$ in our simulation). However, this result depends again critically on sample size and overlap. In small samples with little overlap, the grouping estimator is heavily attenuated and can be substantially below the direct estimator. These results might help to explain why some authors have found much smaller grouping estimates than others: differences in sampling properties may help to reconcile some of the contrasting results in the literature.

Figure 2 illustrates that the bias-correction procedure also works well when names have added informational content (i.e., when the name instruments do not satisfy the exclusion restriction). The grouping estimator is attenuated because the name mean in the parent sample is an imperfect predictor of the corresponding mean in the child sample. This bias can be reduced by focusing on frequent names, for which the sample means are less noisy. However, frequent names also have less informational content (see Section 6.1). Our bias-correction procedure allows researchers to retain the entire sample, and to correct for the bias that arises from the noisiness of sample means in small name groups. The bias-corrected grouping estimates are only slightly lower than the *latent* rate of persistence, even in small samples. Of course, they are much larger than the *observed* rate of persistence as captured by the direct estimator. Whether the latent or the observed rate of persistence is the primary object of interest varies across studies (and depends on one's assumptions about the transmission process). Our point here is that grouping estimates can be adjusted for the influence of sampling properties, irrespectively

of the underlying transmission model.

5.7 Grouping Estimates in the Prior Literature

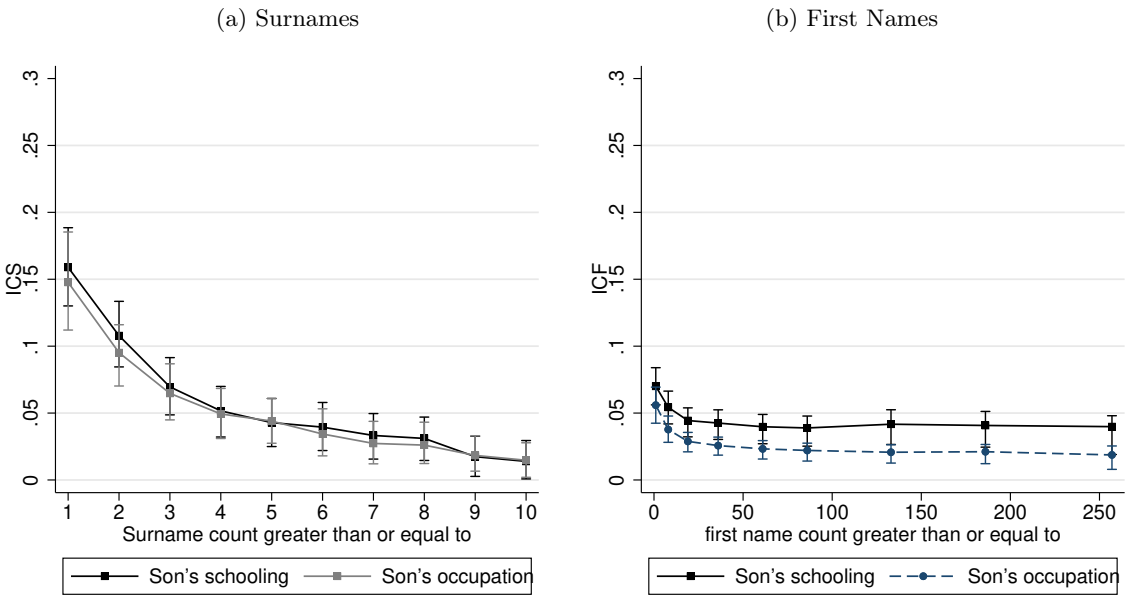
The statistical properties of the grouping estimator depend therefore critically on the probability with which a child's parent (or more generally, ancestor) is included in the sample. We note that this question is related to but distinct from the issue of sample attrition from out-migration, which has received particular attention in the literature. For example, Barone and Mocetti (2020) compare the socioeconomic status of surnames in the city of Florence across six centuries, in a sample that excludes descendants who out-migrated from the city and includes non-descendants who in-migrated from other areas. As the authors note, the exclusion of out-migrants might affect their estimates if the probability to migrate covaries with socioeconomic mobility. However, our findings imply that migration attenuates the grouping estimator even if the decision to migrate is random, because it pushes the estimator from the "*inclusive*" (2SLS) towards its "*leave-out*" (two-sample) version – which mechanically decreases the grouping estimates, even if the intergenerational patterns were similar for migrants and non-migrants.[28]

In comparison, the grouping estimates based on first names in Feigenbaum (2018) are less susceptible to out-migration, thanks to the shorter analysis period and the use of a complete-count Census in the descendant's generation. However, the group means in the parent generation are based on the IPUMS 1-percent excerpt of the 1910 Census, so the overlap p between the parent and child samples is very small. So while the grouping estimator in Barone and Mocetti (2020) corresponds to a mixture of its "*inclusive*" and "*leave-out*" variants, the estimators used by Feigenbaum (2018) and Olivetti and Paserman (2015) correspond more closely to the latter. Neither variant is preferable per se, but the distinction affects how grouping estimates should be interpreted, and may explain why their size varies so much across studies. Moreover, a high degree of overlap between the parent and child samples simplifies the interpretation of the grouping estimator, as it reduces the influence of sample size.[29] We therefore propose that researchers report or estimate the overlap in their application.

[28] This argument could potentially explain why the estimates of Barone and Mocetti (2020) increase when they take measures to reduce sample attrition from migrants.

[29] For example, in Feigenbaum's linked 1915 Iowa State Census Sample the "inclusive" grouping estimate (0.533, see Table 5) is larger than the direct estimate (0.441) and more so than the grouping estimate as reported in Feigenbaum (2018) as based on the group means of fathers' status from Olivetti and Paserman (2015) that have little overlap with the sons in his data (0.353, Panel (b) of Table 7). We actually find the leave-out grouping estimator based on the smaller set of fathers included in his own linked sample for Iowa to be even more attenuated (0.215).

Figure 3: Informational Content vs. Name Frequency



Note: The figures plot estimates of the ICS and ICF and corresponding bootstrap intervals, based on a regression of son's years of schooling (solid line) or son's occupational status (dashed line) on a set of surname dummies (sub-figure a) for name groups with a name count greater than or equal to $n = \{1, ..., 10\}$ or first name dummies (sub-figure b) for name groups with frequencies at or above percentiles $p = \{0, 10, 20, 30, 40, 50, 60, 70, 80\}$. Finnish Longitudinal Veteran Database, White Guards only.

6 Properties and Caveats

We have hitherto reviewed how the R^2 and grouping estimators are constructed, and illustrated their basic functions and properties. In this section, we develop some more detailed arguments on how characteristics of the name and sampling distributions affect each estimator. As will become clear, name-based estimators are subject to quite many influences, some of which can have dramatic effects on their level and interpretation. The discussion also highlights that the R^2 and grouping estimators, while closely linked in some aspects, are susceptible to different conceptual caveats.

6.1 Name Frequency

Name distributions are heavily skewed, with a large share of names being held by few individuals and a small share being held by many individuals. This skewness has a first-order effect on the R^2 estimator, which is primarily identified from rare names, and *decreases* in name frequency. In contrast, the grouping estimates *increase* in name frequency in our samples. This positive correlation is however not universal as it represents the net result of two countervailing effects.

Consider first the R^2 estimator. We found that the informational content is substan-

Table 6: Most and Least Prestigious First Names

Rank	Most prestigious		Least prestigious	
	Red Guard	White Guard	Red Guard	White Guard
1	Maurits	Harald	Joose	Hemmi
2	Rudolf	Jarl	Juha	Aate
3	Klaus	Carl	Manu	Nikodemus
4	Reinhold	Harry	Eemeli	Sulho
5	Konrad	Bror	Jooseppi	Eeli

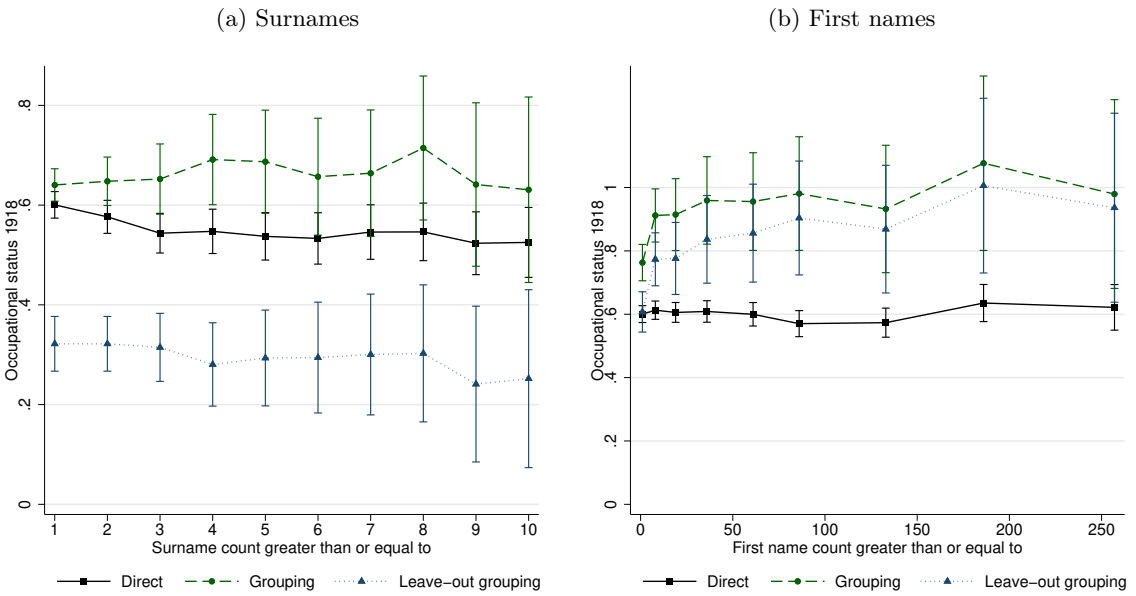
Note: Names ranked by mean father's occupational status. We drop name groups with less than five observations.

tially lower for first names than for surnames (see Table 4), consistent with the much higher mean frequency for first names. Figure 3 shows that the informational content of both surnames and first names decreases with name frequency, but the rate of decay is much faster for the former. For surnames, it is indeed the rare surnames that drive the informational content, and the ICS becomes small or zero in larger surname groups (Figure 3a). Güell, Rodríguez Mora and Telmer (2015) document a similar pattern in data from Catalonia, and explain that it is a natural consequence of the birth-death process for surnames: rare names are indicative of family links, while common surnames are less indicative of such links and therefore less informative. Turning to Figure 3b, we see that rare first names have much lower informational content than rare surnames but frequent first names remain remarkably informative – the ICF is comparatively flat across the frequency distribution. Frequent first names may be informative because of aspirational naming (Olivetti and Paserman, 2015) or because name preferences vary across socioeconomic groups (Lieberson and Bell, 1992).[30] So while the mechanical transmission process underlying surnames washes out in larger name groups, the choice process underlying first names remains relevant. The comparatively flat gradient of the ICF with respect to name frequency suggests that estimators based on first names are based on a more representative part of the population than surname-based estimators.

The low informational content of common surnames also has implications for the grouping estimator. On the one hand, the grouping estimator depends on the extent with which one can predict a person's socioeconomic status based on the status of others sharing the same name (see equation (17)) – which is closely related to the informational content of names. Frequent names have lower informational content, which tends to attenuate the grouping estimator (in particular its *leave-out* variant). On the other hand, the sample

[30] For example, names with a royal or noble connotation may be generally popular though *more* popular among families with high socio-economic status – and as such maintain their informational content. In our sample, name preferences differ between members of the White and Red Guard (see Table 6). Among the White Guard, none of the top-5 most prestigious names (as measured by mean occupational status) are of Finnish origin, and all use the Swedish spelling form (e.g., Eric vs. Erkki).

Figure 4: The Grouping Estimator vs. Name Frequency



Note: The figures plot the estimate and corresponding confidence intervals from a regression of son's years of schooling on father's occupational score (black solid line) or on the imputed occupational score based on surnames (sub-figure a) for name groups with a name count greater than or equal to {1,....,10}, or first names (sub-figure b) for name groups with frequencies at or above percentiles $p = \{0, 10, 20, 30, 40, 50, 60, 70, 80\}$. Finnish Longitudinal Veteran Database, White Guards only.

mean is a more precise estimate of the population mean for the more frequent names, increasing the grouping estimates. It is therefore ambiguous if the grouping estimator decreases or increases when restricting the sample to frequent names.

For illustration, Figure 4 plots the direct, "*inclusive*" grouping and "*leave-out*" grouping estimates from sub-samples with varying minimum name frequency (as indicated on the x-axis). The *leave-out* grouping estimator based on *surnames* is fairly insensitive to name frequency in our sample. As this zero net effect depends on two countervailing effects, it may vary across settings. Indeed, Clark finds large grouping estimates in rare surnames across several sources, while Chetty et al. (2014) show that the grouping estimator increases with the frequency of surnames in U.S. tax data. In contrast, the *leave-out* grouping estimator based on *first names* increases substantially in name frequency, from about 0.25 in the full sample to more than 0.5 when restricting our attention to the most frequent first names. The reason becomes clear when going back to Figure 3: the informational content of first names declines less with name frequency than the informational content of surnames. The net effect of name frequency is therefore positive, and the grouping estimator becomes quite large when considering only the most popular first names. We find a similar pattern using the *IPUMS Linked Representative Sample*.

In sum, all name-based estimators are sensitive to the name frequency distribution.

The R^2 estimator based on surnames is most sensitive, while the grouping estimator is less sensitive, in particular when the parent and child samples overlap. How does this affect their interpretation? The debate in the literature has focused primarily on whether intergenerational transmission varies systematically with name frequency – the name-based estimator's dependency on rare surnames would be a concern if the transmission process is suspected to be different for individuals with rare surnames. But in our data, this appears not to be the case (see Figure 4) and mobility estimates also appear fairly insensitive to the frequency of names in U.S. tax data (Chetty et al., 2014, Online Appendix).

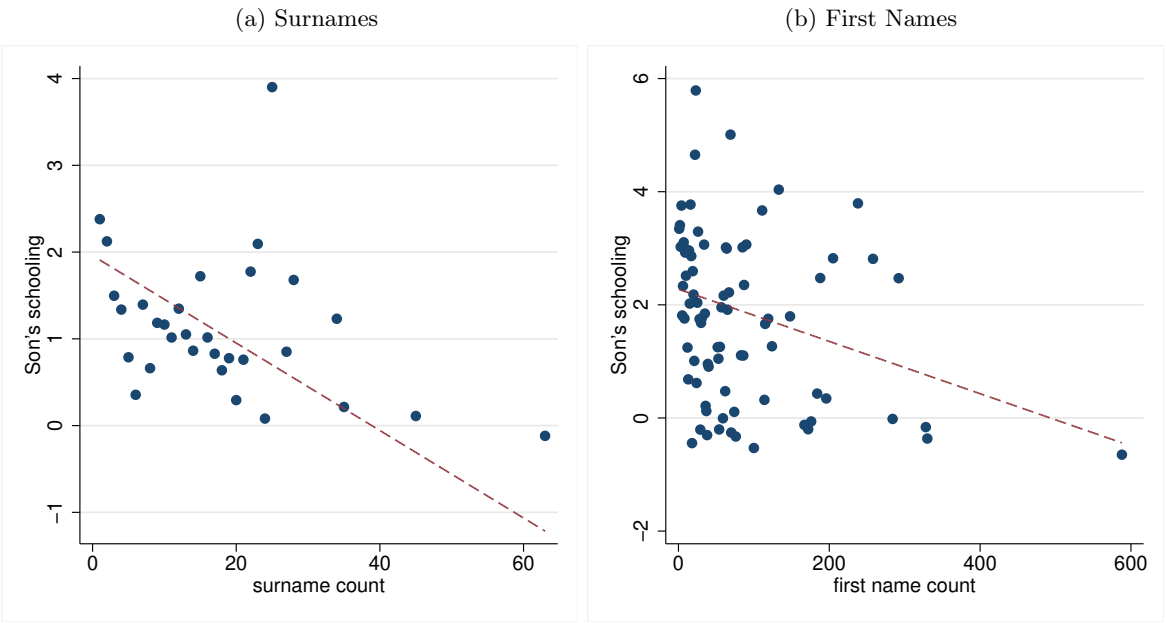
However, because the informational content decreases mechanically in name frequency, the frequency distribution needs to be standardized for the R^2 estimator to be used for comparative purposes. For example, Güell et al. (2018) standardize the name frequency distribution across regions in Italy. The relation between name frequency and the grouping estimator is more complex and has different implications depending on the purpose of the study. For comparative purposes, it may be enough to ensure that the name frequency distributions underlying the group means are comparable between subgroups. If however the *level* of intergenerational mobility is the object of interest, one needs to understand the various mechanisms via which name frequency affects the grouping estimator. Routinely, authors have restricted their attention to rare surnames. But as illustrated by equation (17), the grouping estimator may in fact be attenuated in rare surname groups, in particular when the parent and child samples do not overlap. We therefore propose that researchers report how their grouping estimates vary with name frequency, and correct for the attenuation bias that originates from the use of small samples and infrequent names (see Section 5.5).

6.2 The Socioeconomic Gradient in Name Frequency

A related caveat is that the average socioeconomic status decreases systematically with name frequency. Figure 5 plots the average years of schooling of sons across bins of the name frequency distribution. The socioeconomic gradient is substantial – the most common surnames (first names) have on average 3 (2.5) fewer years of schooling than rare names, compared to a standard deviation of 4.8 years.

The negative relation between status and the frequency of *first* names is easily understood. As shown by Fryer and Levitt (2004), affluent parents tend to choose names for their offspring that are new, and different from the most common ones in their society. This observation also holds in our data. Why status decreases in the frequency of *surnames* is less obvious. Migration is one of the primary ways of introducing new surnames into a population, and migrants often have lower socioeconomic status (Güell, Rodríguez Mora and Telmer, 2015). On the other hand, the deliberate choice of a new surname – i.e., name "*mutations*" – might be more common among individuals with high socioeconomic status.

Figure 5: Socioeconomic Status Decreases in Name Frequency



Note: Binscatter plot of the sons' average years of schooling against the frequency of the surname (sub-figure a) or first name (sub-figure b). Finnish Longitudinal Veteran Database, White Guards only.

Collado, Ortín and Romeu (2008) document a negative relation between socioeconomic status and name frequency in Spain, and note that this gradient may reflect signaling behavior by successful dynasties. We confirm this "selective mutation" hypothesis in Section 6.6.

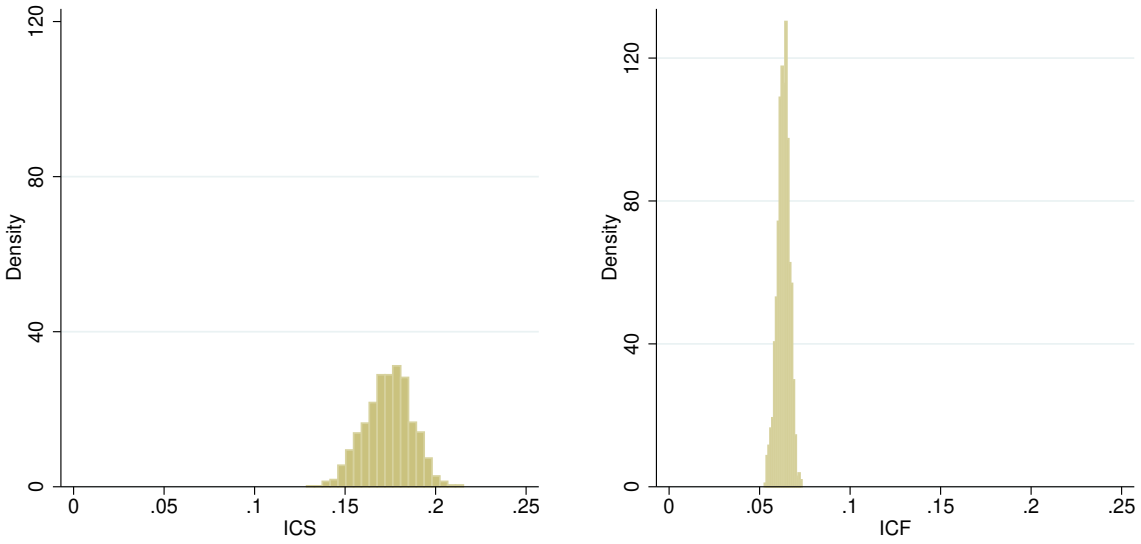
The observation that name-based estimators are identified primarily from rare names (Section 6.1) combined with the observation that people bearing rare names tend to have higher status, suggest that name-based estimators capture mobility within a non-representative subset of the population. This does not necessarily imply that the mobility estimates themselves would be non-representative, and we already noted that the conventional intergenerational coefficient appears fairly similar in rare and more frequent names. However, the socioeconomic gradient in name frequency explains why *singletons* (individuals with a unique name) may contribute to the ICS and ICF as defined in equations (3) and (5). And while it has less obvious implications for the grouping estimator, it may still be good practice to report the socioeconomic gradient in the name frequency distribution in applications.

6.3 Sample Size

How sensitive are name-based estimators to sample size? Since these methods are attractive in settings for which register-based data sources are not available, many applications

Figure 6: The Informational Content and Placebo Distributions

(a) The Informational Content of Surnames (b) The Informational Content of First Names



Note: Histogram of estimated ICS (sub-figure a) and ICF (sub-figure b) in sons' years of schooling across 1,000 placebo distributions.

are confined to samples of limited size. This problem may be accentuated by the nature of the research question. If one focuses on regions, e.g., in order to identify place-based determinants of mobility, the subsamples inevitably become small no matter how large the base population is. We argue that sample size is not a major concern when applying the R^2 estimator, but that uncertainty induced by the reshuffling of names in the placebo distribution needs to be addressed in smaller samples. In contrast, the grouping estimator tends to be very sensitive to sample size. As has been a recurrent theme, this issue is highly intertwined with other factors, such as sampling properties, the name frequency distribution, and the informational content of names.

As Güell, Rodríguez Mora and Telmer (2015) and Güell et al. (2018) develop the R^2 estimator in full-count Census data they do not address sampling uncertainty. We discuss here how their estimation procedures can be adapted to applications in settings with more limited sample size. A first concern is that the R^2 estimator may be upward biased in smaller samples.[31] This issue is typically addressed by reporting the *adjusted* R^2, which rescales the sample R^2 based on sample size and the number of regressors. Different adjustment formulas are in use, with those formulas that are implemented in standard statistical software not necessarily producing the least biased results (Yin and Fan, 2001). Interestingly, Güell, Rodríguez Mora and Telmer (2015) chose an empirical

[31] The OLS estimator is subject to overfitting in finite samples, and the *sample* R^2 is a biased estimate of the coefficient of determination (i.e., the *population* R^2).

instead of an analytical approach, comparing the sample R^2 to the corresponding R^2 from a placebo regression in which the name dummies have been reshuffled across individuals (see equation (3)). While both approaches perform well in larger samples, the empirical approach produces imprecise estimates in small samples. The reshuffling of names across individuals introduces uncertainty, with different draws of the name distribution resulting in quite different estimates of the ICS. As an example, Figure 6 plots the distribution of the R^2 estimator when reshuffling the name distributions 1,000 times. Sub-figure (a) plots the estimated ICS corresponding to the specification reported in column (3) of Table 4, while sub-figure (b) plots the estimated ICF corresponding to column (4). A simple solution to this issue is to report the mean of the R^2 estimator across repetitions, as we do in Table 4.[32]

A related issue is inference. Güell, Rodríguez Mora and Telmer (2015) do not report standard errors given the large sample size underlying their study. But it is important to quantify the precision of estimates in smaller samples. We propose a bootstrap procedure, reporting 95% confidence intervals that are based on 1,000 bootstrap samples. In each round, we draw cluster of observations on the surname level with replacement, assigning different identifiers to surname groups that are drawn multiple times. Within each bootstrap sample, we compute the ICS or ICF as the difference between the actual and a single placebo regression (the repeated reshuffling of the name distributions within each bootstrap sample is computationally intensive and does not affect much the estimated confidence intervals). We then report the 2.5 and 97.5 percentile of the resulting distribution. These confidence intervals account for the uncertainty from reshuffling of names in the placebo regressions, as well as standard sampling uncertainty.

Sample size has more serious implications for the grouping estimator. The *leave-out* variant depends on the attenuation factor $\frac{Cov(\bar{x}_{(i)j},x_{ij})}{Var(\bar{x}_{(i)j})}$, and therefore on the extent with which one can predict a person's socioeconomic status with the status of others sharing the same name (see equation (17)) – which increases in sample size. To reduce the attenuation bias, researchers might be tempted to restrict their sample to name groups that are sufficiently large. However, because the informational content of names is smaller for more frequent names (Güell, Rodríguez Mora and Telmer, 2015), the relation between the grouping estimator and name frequency is ambiguous (see Section 6.1). Instead, we propose that researchers probe the influence of sample size in two ways. First, by providing evidence on how sensitive the grouping estimates are to fluctuations in the size of the sample, e.g., by drawing sub-samples of their main sample. As we illustrate in Section 5.6, if a grouping estimator is heavily attenuated in small samples, it will also be sensitive to changes in sample size. Second, by directly estimating the attenuation factor,

[32] While our approach is a natural extension of the definition in Güell, Rodríguez Mora and Telmer (2015), the question arises as to whether analytical methods to estimate the population R^2 would perform better in smaller samples.

in order to construct a bias-corrected grouping estimator. We provide such correction procedure in Section 5.5.

6.4 The *Added* Informational Content of Names

The motivation for using names is to proxy for socioeconomic status variables that are not contained in the data at hand, but the concern – or attraction, depending on the perspective – is that they might reflect more than just that.[33] Because our data includes direct family linkages, we can address this question explicitly. Table 7 reports the results from a regression of son's years of schooling on (a linear or flexible function of) the occupational score of the father and the mean occupational score in his surname group. If surnames were merely an imprecise proxy for individual occupational status then the coefficient of the group mean should be insignificant.

Instead, names tend to have *added informational content* (AIC). Columns (1)-(3) of Table 7 show that conditional on own father's occupational status, the imputed occupational status of the surname group still has a significant association with son's years of schooling in the Finnish Longitudinal Veteran Database data (Panel A). This pattern is robust to the consideration of other outcome variables or the inclusion of control variables. Columns (4)-(6) of Table 7 provide the corresponding evidence for first names, showing that first names too have AIC. Similar but somewhat weaker evidence of AIC is shown for IPUMS Linked Representative Sample 1880-1900 (Panel B), and for the Linked 1915 Iowa State Census Sample (Panel C). Finally, Appendix Table A1 presents similar evidence based on log earnings or years of education in otherwise similar models as those presented in Panel C of Table 7. For all data sources, the coefficient on the father's occupational status is larger when based on first names instead of surnames, perhaps reflecting the importance of "aspirational" naming (Olivetti and Paserman, 2015) or differences in the name frequency distribution.

The observation that names have added informational content, over and above an individual's observed socioeconomic status, alters the interpretation of name-based estimates. It can be rationalized by very different theoretical mechanisms, such as group-level causal effects (as in Borjas, 1992), observable status being only an imperfect proxy for individual status (Clark, 2014), or, for first names, aspirational naming (Olivetti and Paserman, 2015). Name-based estimators are therefore not directly comparable to the conventional direct estimator, and the question as to why names have additional informational content affects their interpretation.

[33] Most studies use the (feasible) name-based estimators as a second-best alternative to the (infeasible) conventional estimator. In contrast, the argument by Clark (2014) is explicitly based on the assumption that the surname-based grouping estimator captures aspects of the transmission process that are not captured by the conventional estimator.

Table 7: The Added Informational Content of Surnames and First Names

	Dependent variable: Son's occupational status					
	Surnames			First Names		
	(1)	(2)	(3)	(4)	(5)	(6)
Father's occupational status	Linear	Flexible	Flexible	Linear	Flexible	Flexible
Other controls	–	–	Yes	–	–	Yes
Panel A: Finnish Longitudinal Veteran Database						
Father's name mean	0.138	0.162	0.121	0.218	0.199	0.198
(occupational status, HISCAM)	(0.030)	(0.039)	(0.029)	(0.031)	(0.045)	(0.030)
AR2	0.249	0.325	0.386	0.252	0.327	0.389
N	5,986	5,986	5,986	5,986	5,986	5,986
Panel B: IPUMS Linked Representative Sample 1880-1900						
Father's name mean	0.012	0.006	0.009	0.047	0.051	0.033
(log occupational income)	(0.025)	(0.025)	(0.025)	(0.028)	(0.028)	(0.028)
AR2	0.149	0.177	0.194	0.149	0.177	0.194
N	9,076	9,076	9,076	9,076	9,076	9,076
Panel C: Linked 1915 Iowa State Census Sample						
Father's name mean	0.021	-0.027	-0.024	0.124	0.113	0.114
(log occupational income)	(0.050)	(0.050)	(0.051)	(0.043)	(0.043)	(0.043)
AR2	0.142	0.169	0.170	0.143	0.171	0.172
N	3,204	3,204	3,175	3,204	3,204	3,175

Note: The table reports the coefficients from a regression of son's occupational status (HISCAM) (Panel A) or log occupational income (Panels B and C) on the father's corresponding occupational status (column 1) and the mean of the father's status in the name group, defined by son's surname (columns 1-3) or first name (columns 4-6). Panel A reports estimates from the Finnish Longitudinal Veteran Database (White Guard only). Other controls include dummies for ethnicity, year of birth and region of birth (10 synthetic counties). Panel B reports estimates from the IPUMS Linked Representative Sample 1880-1900 of the U.S. Census (Olivetti and Paserman, 2015). Other controls include dummies for foreign born, year of birth and state of residence in 1880. Panel C reports estimates from a digitized sample of the 1915 Iowa State Census (Goldin and Katz, 2000) linked to the 1940 US Federal Census (Feigenbaum, 2018). Other controls include dummies for foreign born, year of birth and state of birth. Standard errors in parentheses.

6.5 Control Variables

A recurring concern regarding name-based estimators, related to the previous caveat, is that they weight group-level and individual-level transmission processes differently than the conventional direct estimator. In particular, an important criticism of surname-based estimates is that they are heavily influenced by the role of ethnic or national origin in the transmission of advantages (Chetty et al. 2014, Torche and Corvalan, 2018). We note that those criticisms do not only apply to the surname-based grouping estimator (as used by Clark, 2014), but to all name-based estimators. A potential strategy to address them is to either exclude names that are very indicative of origins, or to include indicators of those origins as a control variables. Indeed, the inclusion of such controls has been standard in applications based on the R^2 estimator proposed by Güell, Rodríguez Mora and Telmer

Table 8: Stability of Name-based Estimators to the Inclusion of Controls

	Dependent Variable: Son's occupational status				
	(1)	(2)	(3)	(4)	(5)
Direct estimator:	0.600	0.582	0.575	0.527	0.423
	(0.014)	(0.014)	(0.013)	(0.013)	(0.015)
R^2 *estimator:*					
Surnames (ICS)	0.272	0.240	0.229	0.178	0.103
First names (ICF)	0.117	0.092	0.108	0.073	0.031
Grouping estimator:					
Surnames	0.640	0.616	0.605	0.543	0.415
First names	0.763	0.702	0.736	0.617	0.420
Finnish		Yes	Yes	Yes	Yes
Year of birth			Yes	Yes	Yes
Region of birth (county)				Yes	
Region of birth (parish)					Yes
Observations	5,986	5,986	5,986	5,986	5,986

Note: Members of the White Guard only. The direct estimator in column (1) refers to a univariate regression of son's occupational status score (HISCAM) in 1918 on his father's occupational status score. The implied ICS and ICF are the difference between the adjusted R-squared of a model including a complete set of name dummies and an otherwise identical regression in which names are randomly reshuffled. The grouping estimator imputes father's occupational status based on surnames and first names. The control variables added gradually to the models (columns (2)-(5)) include an indicator for ethnicity, year of birth, and region of birth classified based on geographic coordinates into 10 synthetic counties (coarse) or 583 parishes (fine). Standard errors in parentheses. Source: The Finnish Longitudinal Veteran Database.

(2015), and similar strategies could be adopted in all name-based methods.

We therefore explore the stability of the mobility estimates in our main sample to the inclusion of control variables, across all estimators discussed in this study. Column (1) of Table 8 reports the mobility estimates in the benchmark case, in which no controls are included. Columns (2)-(5) explores specifications that gradually include ethnicity, year of birth, and regional fixed effects at a coarse county or finer parish level. We find that all estimators are sensitive to the inclusion of controls. As expected, the conventional estimator using direct family links is the most stable one. The R^2 estimators are most sensitive, in particular if based on first names. The grouping estimator is attenuated by 25-50 percent when controlling for ethnicity and region of birth, depending on the choice of socioeconomic outcome for sons. These results are in line with Feigenbaum (2018), who also finds the direct coefficient to be most stable to the inclusion of control variables. They also support the argument that name-based estimates overweight ethnic and regional factors as compared to the conventional estimates, and sensitivity analyses such as the

one presented here may help to disentangle individual-level transmission processes from such group-level processes.

6.6 Name Mutations

While the transmission of surnames is a fairly deterministic affair, name changes or "mutations" do occur. In the short run, name mutations are a nuisance for researchers using surnames to infer intergenerational mobility, as they sever the link between parents and children. But in the long run, mutations are necessary for surnames to retain their informational content. Güell, Rodríguez Mora and Telmer (2015) conjecture that in the absence of mutations surnames would eventually collapse into one universal surname, and hence no longer contain any socioeconomic information. Instead, a mutation infuses the mutated surname with informational content and secures the functionality of the surname as a proxy for kinship for some generations to come.

The frequency of surname mutations varies substantially over time and particularly active periods have often been spurred by political and nationalistic movements.[34] The frequency also varies in contemporary contexts.[35] The birth cohorts sampled in our data coincide with the aforementioned particularly active period of name changes in Finland. Moreover, we observe both the prior (pre-mutation) and the mutated (post-mutation) surname. Thanks to these two advantages we can explore the birth-death process of names in detail, and we compare our findings to related evidence from Spain provided by Collado, Ortín and Romeu (2008). The mutation rate is roughly 8.7 percent in our data, for both White and Red Guard (see Table A3 in the Appendix). For comparison, the estimated lifetime mutation rate in Güell, Rodríguez Mora and Telmer (2015) is only about 0.25 percent. We observe nearly 600 name mutations among the Red Guard, and more than 800 name mutations among the White Guard.

Table 9 shows that the estimated ICS is higher when using the current (post-mutation) surnames in the estimations. Replacing the mutated surnames in the sample with the prior (pre-mutation) surnames decreases the ICS by about 10 percent (the drop is significant at $p = 0.01$). As illustrated in Figure 7a, post-mutation surnames tend to be infrequent surnames, with the share of individuals who actively chose their surname being five times

[34] Paik (2014) reports that during the Japanese occupation, many Koreans strategically changed their clan lineage. In Finland, name changes were particularly frequent during the romantic nationalist movement for independence from Imperial Russia around the turn of the twentieth century. Most name changers switched from ethnic Swedish to Finnish-sounding names, but switches from one Finnish-sounding name to another were also common. In particular, names that were common among sharecroppers were converted to national romantic names with references to nature. Another example of a surge in name mutations is the aftermath of the emancipation of slaves in the U.S.; many American ex-slaves adopted the surname of the slaveholder for whom they used to work or the names of former presidents (Baiardi, 2016).

[35] In Sweden, surnames with more than 2,000 holders were deregulated in 2017. Anyone can attain such a surname at a cost of 1,800 SEK ($204). While the Swedish Patent and Registration Office received between 5000 - 10,000 applications annually before the reform, the number spiked threefold in 2017.

Table 9: Informational Content of Surnames Pre/Post-Mutation

	Post-mutation		Pre-mutation	
	Occ. status	Schooling	Occ. status	Schooling
	(1)	(2)	(3)	(4)
Surname dummies	Yes	Yes	Yes	Yes
AR2	0.308	0.410	0.293	0.392
Implied ICS	0.147	0.158	0.133	0.141
95% CI	[0.116, 0.181]	[0.131, 0.191]	[0.102, 0.166]	[0.115, 0.170]
N	14,754	12,846	14,754	12,846

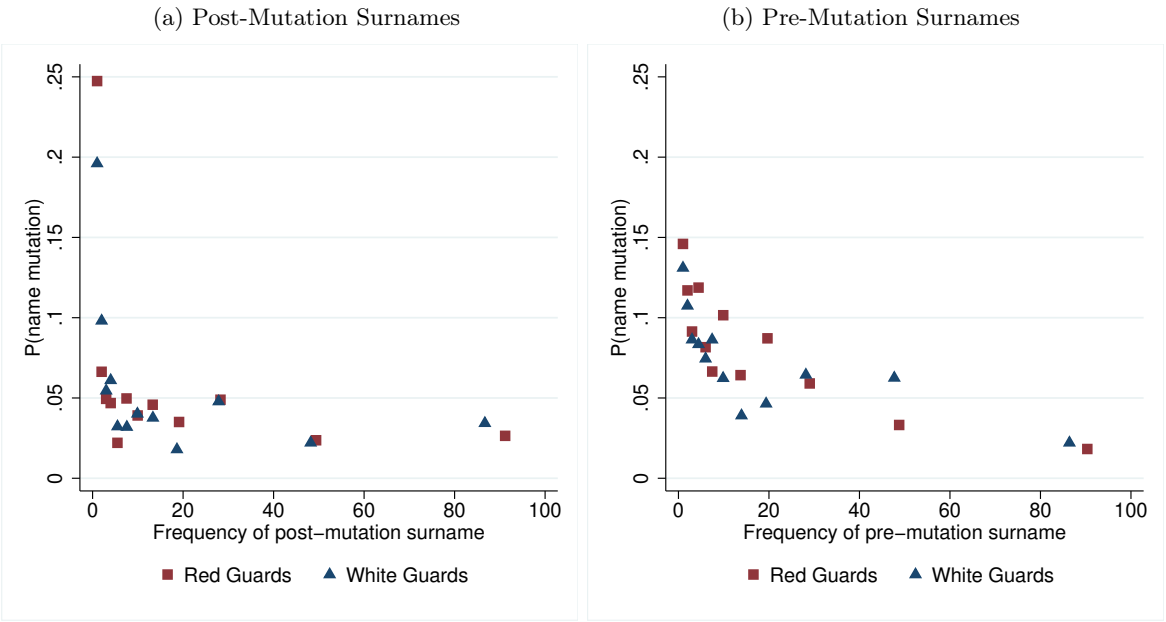
Note: Columns (1) and (2) report estimates of the ICS for occupational status and for years of schooling based on post-mutation surnames. Columns (3) and (4) report estimates of corresponding models that replace post-mutation names with pre-mutation names for all name switchers. All regressions include a dummy for ethnicity (Finnish-sounding name), a dummy for White Guard (the reference category being the Red Guard), year of birth and region of birth (10 synthetic counties). 95% confidence interval across 1,000 bootstrap samples in brackets. Source: The Finnish Longitudinal Veteran Database.

higher among rare than among common surnames. That is useful for mobility research, as it is rare names from which most information on socioeconomic status can the extracted (see Section 6.1).

Still the effect of those name mutations on the ICS appears surprisingly limited, given the frequent name changes in our period of study. The reason for this becomes clear from Figure 7b: name switchers tend to have rare surnames even prior to their name change. For both Red and White Guard, individuals in the lowest quartile of the name frequency distribution are about four times as likely to change their surname as individuals with more common surnames. We hypothesize this observation can be rationalized by the same argument as the observed relation between name changes and post-mutation name frequency. Rare names have a higher informational content, which may either create incentives to pick them (if the signal is intended) or to abandon them (if the signal is unintended). Figures 7b and 7a are then mirror implications of the same basic insight, that rare surnames are more informative. Given this symmetry it is not obvious if episodes in which large shares of the population change their surname will necessarily increase the informational content of surnames (although our empirical result supports the presumption that typically they do).

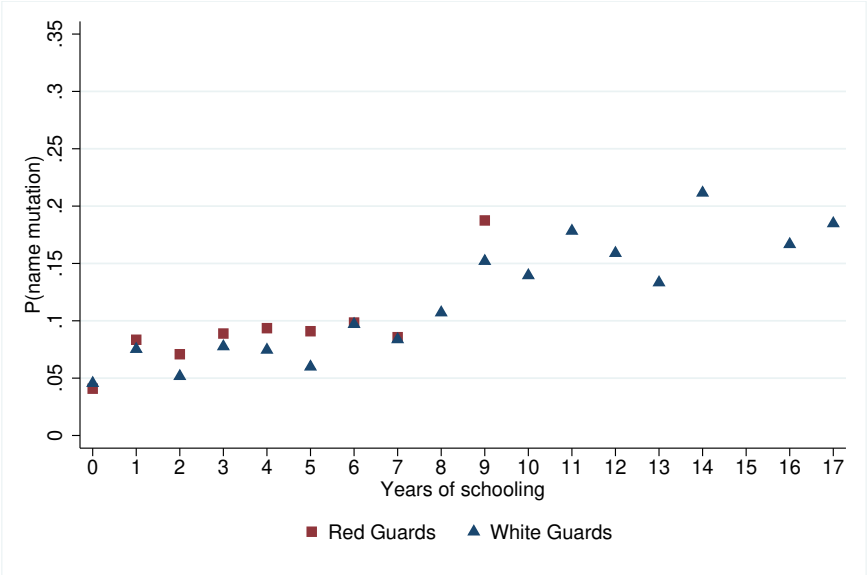
Mutations also update the socioeconomic content of a name, which may explain why the frequency of a surname is inversely correlated with socioeconomic status (Section 6.2). The observation of name changes allows us to directly test whether there is a socioeconomic bias in the probability to change names, as has been hypothesized by Collado, Ortín and Romeu (2008). Figure 8 shows that surname mutations are indeed selective, with

Figure 7: Name Mutations vs. Name Frequency



Note: Binscatter plot of indicator for name mutation against the frequency of the pre-mutation (sub-figure a) or post-mutation (sub-figure b) surname.

Figure 8: Socioeconomic Bias in Name Mutations



Note: Scatter plot of mean indicator for name mutation against sons' years of schooling. Only cells with more than 10 observations plotted.

the probability to change names increasing four-fold over the distribution of educational attainment. Deliberate mutations might in this sense be a means of strengthening the signal of economic status that a surname sends.[36] For example, Collado et al. show that in Spain, many of the rarer surnames in the 20th century did not exist in the 19th century, and note that surnames act as a signaling device for successful dynasties. Figure 8 demonstrate that there exists a socioeconomic gradient in name mutations in Finland as well.

7 Intergenerational Mobility and the Finnish Civil War

We have highlighted the R^2 and grouping estimators' close relationship to the conventional estimator based on direct family links but also illustrated a number of caveats. Since we observe first and surnames, direct family links, and occupation and schooling for two generations, we can evaluate the performance of all aforementioned estimators in our data. We compare intergenerational mobility between distinct groups of the Finnish society during the beginning of 20th century. For Finland, this was an interesting period of transformation from an agrarian towards an industrialized society but also one plagued by political unrest, World War I and – in its aftermath – the Finnish Civil War in 1918. We first shed light on the prewar mobility patterns of the Reds and Whites, the two antagonistic sides of the Civil War. The name-based methods are then put to test by estimating within-group mobility for the two samples and comparing the mobility pattern with the one that the direct estimator produces. These methods are compared in the type of setting for which they are designed, namely one in which linkage based on individual identifiers is uncommon.

7.1 Historical Background

The Finnish Civil War was fought between 27 January and 15 May 1918, with the two antagonistic parties being the troops of the Social Democrats led by the People's Commission of Finland (the *Red Guard*), and the troops led by the conservative government (the *White Guard*). The Red Guard consisted predominantly of industrial and agrarian workers whereas the White Guard were supported by the farmers and middle- and upper class factions of the population. Each side had roughly equally many men in their ranks and received back up from their respective allies (the newly formed Soviet Russia backed

[36] Güell, Rodríguez Mora and Telmer (2015) note that immigrants are more likely to mutate their names, sometimes unintentionally through transliterations or misspellings by the authorities in the host country. Immigration may therefore reduce the correlation between name mutations and status, if immigrants tend to have lower status. See also the related literature on the economic incentives of name changes for immigrants and the positive consequences of cultural assimilation (Algan, Mayer and Thoenig, 2013; Arai and Skogman Thoursie, 2009; Carneiro, Lee and Reis, 2016; Abramitzky, Boustan and Eriksson, 2020).

the Red Guard while the Germany Empire backed the White Guard) but eventually the better equipped and more organized White Guard came out victorious.

The conflict that ultimately led to civil war was partly rooted in the country's economic decline after the outbreak of World War I and disparities in land ownership (Jantti, Saari and Vartiainen, 2006; Arosalo, 1998). In the cities, the export-driven industrial production declined rapidly from 1914 onwards, leading to wage cuts and high unemployment among factory workers. In the rural areas, land ownership was becoming increasingly polarized along with the commercialization of land that followed from the expansion of forestry industry and rationalization of farming. The conditions of tenant farmers and agricultural workers deteriorated further during World War I and many tenant leases were discontinued. As a consequence, regions with particularly uneven land distributions and industrial towns were most affected by the economic downturn between 1914 and 1916. They also saw more violence (Arosalo, 1998).

These findings suggest that in addition to the general unrest in Europe in the aftermath of World War I, and the political turmoil in Russia, the state (or lack) of social mobility might have been one factor contributing to the outbreak of the Finnish Civil war. One particular question is whether intergenerational mobility was different among the individuals who joined the Red Guard as compared to the individuals who joined the White Guard.

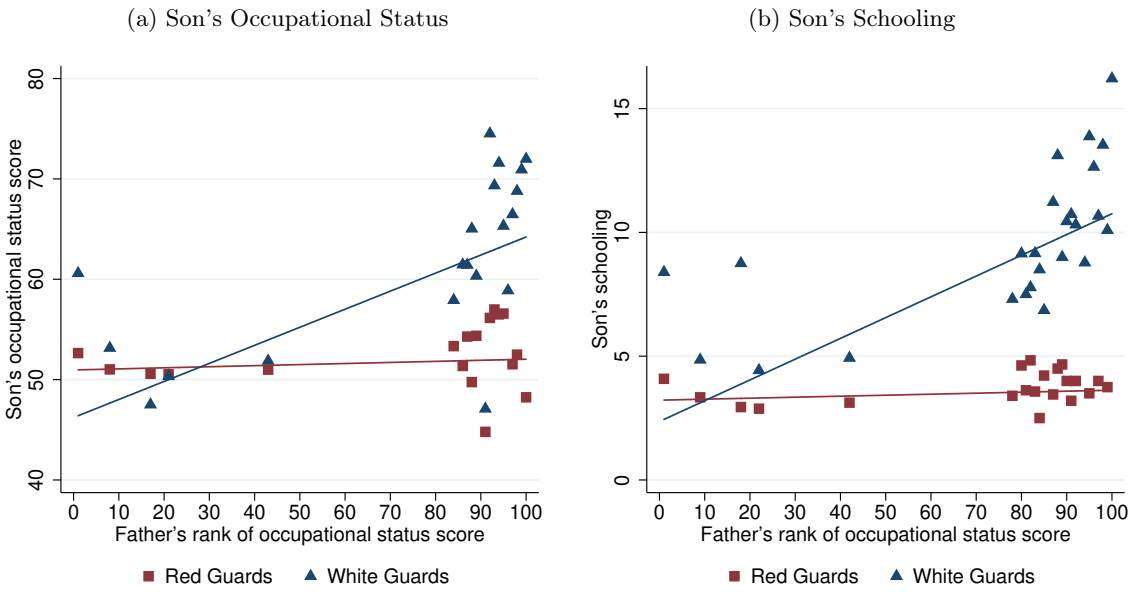
7.2 Descriptive Statistics and Measurement

Members of the White Guard have on average more schooling ($t = 53.61$), higher occupational status ($t = 17.23$) and less common first and surnames than the Red Guard (see Table 3). In the individual-level records of the National Archives, the father's occupational status is only measured for the members of the White Guard (self-reported through home interviews). We therefore complement these data with a matched subset of digitized birth records that contain the father's occupation, which allows us to directly compare the socioeconomic background and mobility of the Red and White Guard. Not surprisingly, those who joined the Red Guard tend to come from families with lower socioeconomic status than those who joined the White Guard. However, the gap is not as large as one might expect. As shown in Table 3, the mean of the occupational status score among fathers of the Red Guard is only half a standard deviation below the corresponding mean among fathers of the White Guard.

7.3 Direct Mobility Estimates

To compare the prewar intergenerational mobility of members of the socialist Red Guard and the conservative White Guard, we first consider standard measures based on direct links between fathers' and sons' outcomes. Figure 9 plots the average of son's schooling

Figure 9: Intergenerational Mobility of Red and White Guard (Direct Estimates)



Notes: Scatter plot of mean of indicated outcome variable against the percentile rank of the father's occupational status score (HISCAM). The lines correspond to predicted value from a regression of outcome on percentile rank on the individual level.

or son's occupational status scores (based on occupation in 1918) against the percentile rank of the father's occupational status score, separately for members of the Red Guard and White Guard. Among the White Guard we observe a standard pattern of fairly strong intergenerational persistence that appears similar to the degree of intergenerational mobilityin other populations in the early 20th century.

In contrast, intergenerational and in particular *downward* intergenerational mobility is remarkably high among members of the Red Guard. The gap to the White Guard is particularly striking among Reds from advantaged backgrounds, who tend to have as low schooling and occupational status as those from low-status fathers. For example, the expected years of schooling for a son born to a father at the 20th percentile of the occupational status distribution is about 3.3 years, compared to 3.5 years for a son born to a father at the 80th percentile. In contrast, the gap is much larger for members of the White Guard, around 4 years at the 20th percentile compared to more than 9 years at the 80th percentile.

To quantify these pattern in more detail, the first panel of Table 10 reports results from direct intergenerational regressions of son's years of schooling and occupational status score (measured in 1918) on father's occupational score, separately for members of the Red Guard and the White Guard. As Figure 9, these regressions are based on the unrestricted sample of all direct links identified from son's birth certificates. The results

confirm that members of the Red Guard have substantially higher intergenerational mobility than members of the White Guard. The estimated slope coefficient among members of the Red Guard is nearly zero, irrespectively of whether son's socioeconomic status is approximated by years of schooling or occupational status score. The contrasting pattern between Red and White Guard, and the high degree of downward mobility among Red Guard, provide an opportunity to test the performance of the different name-based methods. An interesting question is whether the methods will reproduce the same pattern of mobility between and within groups as the direct family links.

7.4 Name-based Mobility Estimates

The last two panels of Table 10 present estimates from the name-based methods, starting with the R^2 estimators based on first names or surnames. Consistent with the conventional (direct) estimator, we find substantially lower ICS estimates for members of the Red Guard as compared to members of the White Guard. This result is consistent with the argument by Güell, Rodríguez Mora and Telmer (2015) that the ICS is monotonically increasing in the intergenerational persistence on the individual, so that its ordering is informative about mobility differences between groups. More surprisingly, the same pattern holds when using the R^2 estimator based on first names: the ICF is systematically lower for members of the Red Guard as compared to the corresponding estimate for members of the White Guard, irrespectively of which socioeconomic measure we consider.

The R^2 estimator has been primarily used for comparative purposes but in this application it interestingly captures the fact that the level of intergenerational mobility is near-perfect among members of the Red Guard. While it may in general be difficult to map the ICS to more standard intergenerational correlations, it provides a good approximation in extreme cases such as the one considered here.

The last panel of Table 10 considers the grouping estimator based on surnames or first names. In order to estimate the group-level regressions, we impute father's occupational status scores for each first name and surname based on the father's schooling or occupation as observed among members of the White Guard.[37] Because 14 percent of surnames are unique to members of the Red Guard, the sample size is slightly lower for surnames than for first names. Our grouping estimates consistently confirm that members of the Red Guard have substantially higher mobility than members of the White Guard.

In sensitivity analyses we impose additional sampling restrictions to make the two groups more comparable. To ensure that differences in the ICS are not due to differences in the surname distributions between the two groups, we harmonize their distributions by

[37] This introduces an asymmetry in the definition of the group mean between the Red and White Guard. However, the results are qualitatively the same when imputing the occupational status distribution from digitized and matched birth records, which are available for both Red and White Guard.

Table 10: Intergenerational Mobility of White and Red Guard

	White Guard		Red Guard	
	(1) Son's schooling	(2) Son's occ. Score	(3) Son's schooling	(4) Son's occ. score
Direct estimator				
Father's occupational status (BR)	0.211	0.469	-0.002	0.055
	(0.016)	(0.061)	(0.008)	(0.043)
	n=905	n=1,010	n=979	n=1,091
R^2 estimator				
Surnames	0.180	0.201	0.071	0.080
	[0.137, 0.230]	[0.157, 0.254]	[0.029, 0.119]	[0.033, 0.132]
First names	0.097	0.072	0.016	0.009
	[0.074, 0.115]	[0.053, 0.092]	[0.0009, 0.032]	[-0.005, 0.023]
	n=7,021	n=8,282	n=5,825	n=6,472
Grouping estimator				
Surnames	0.088	0.235	0.002	0.022
	(0.011)	(0.040)	(0.004)	(0.021)
	n=3,249	n=3,852	n=3,372	n=3,891
First names	0.184	0.472	0.026	0.119
	(0.013)	(0.045)	(0.005)	(0.028)
	n=5,154	n=5,821	n=5,655	n=6,306

Note: The direct estimator refers to a regression of son's years of schooling (columns 1 and 3) or son's occupational status score in 1918 (columns 2 and 4) on father's occupational status score (HISCAM). The direct estimator is based on son's birth records that were obtained by linking of individuals of the Finnish Longitudinal Veteran Database to their birth records (BR) available at www.anscestry.fi. The R2 estimator is the difference between the adjusted R-squared of a model including a complete set of name dummies and an otherwise identical regression in which names are randomly reshuffled. The grouping estimator is based on the mean occupational status score in a name group among members of the White Guard. To enhance comparability, the grouping estimator for the White Guard are based on leave-out means. All regressions include dummies for ethnicity and year and region of birth (10 synthetic counties). Standard errors in parentheses, 95% confidence intervals across 1,000 bootstrap samples in squared brackets. Source: The Finnish Longitudinal Veteran Database.

making the Gini coefficient and the count of individuals per surname as good as identical across the groups (in the spirit of Güell et al., 2018). The results corresponding to the ICS estimates of Table 10 remain qualitatively the same (Table A5 of the Appendix). The original samples were subject to different sampling frames, in that members of the Red Guard were assembled from a registry of pension applications in 1973, whereas members of the White Guard were assembled from a registry of White Guard veterans recorded during mobilization in the mid-1930s.[38] One concern is that the attrition of Red Guard who did

[38]See Section A of the Appendix for a detailed description of the data acquisition.

not survive until 1973 might be systematically related to intergenerational mobility. In order to address this concern, we identified the sampled members of the White Guard in the Population Registers, based on their first names, surname, date of birth and place of birth.[39] Based on the Red Guard data, we know that conditional on being alive in 1973, the match rate at the Population Registers is very high (99.5 percent for the Red Guard pension applications from 1973). Thus, restricting the White Guard sample to only those who were identified by the Population Registers as having survived until 1973 harmonizes the attrition of the two subgroups. Table A7 in the Appendix shows that members of the Red Guard have substantially higher mobility than members of the White Guard also in this trimmed sample.

7.5 Intergenerational Mobility and Political Preferences

Overall, our estimates tell a fairly consistent story. Among members of the Red Guard, intergenerational mobility, in particular intergenerational downward mobility, was remarkably high before the Finnish Civil War 1918, and markedly higher than among the members of the White Guard. This high mobility is a consequence of the Reds attaining little education and placing into low status occupations irrespectively of the status of their fathers. In particular, the Reds were doing worse than their White counterparts conditional on their father having high socioeconomic status.

These findings relate to a theoretical literature on the link between political preferences and intergenerational mobility.[40] Previous work suggests that an individual's *support for redistributive policies* may be a function of expectations about future socioeconomic status. For example, Piketty (1995) argues that individuals's views about social mobility and the relative role of effort in the determination of socioeconomic outcomes depend on their personal experience, generating heterogenous beliefs about intergenerational mobility and the socially optimal redistribution rate in equilibrium. In a two-by-two mobility table representing mobility across adjacent generations, a society may be characterized by a stable majority of left-wing voters in the lower class, a stable minority in the upper class, and intermediate levels of left-wing voters among the upward or downwardly mobile in the off-diagonal cells. In contrast, Bénabou and Ok (2001) show that even for the poor it can be rational to support low levels of redistribution under certain premises,

[39] The Population Registers can only identify individuals who were alive as of 1970 (in fact the year of the first full digitized Finnish Census) or later.

[40] Of course, Civil Wars cannot be simply rationalized by conflicting political ideologies (in the Finnish case, direct violence was denounced by the majority of the Social Democratic Party; Paavolainen, 1966). As mentioned before, important factors leading up to the Finnish Civil War were the general unrest in the aftermath of World War I and the Russian revolution of 1917, polarization of land ownership during the transformation from an agrarian towards an industrialized society, and financial distress among factory workers due to an economic recession. The stark ideological difference between the Red and White Guard, in particular with respect to redistributive policies, were however a particular salient aspect of the conflict.

Table 11: Membership in the Red Guard by Intergenerational Mobility

		Son's occupational status		
		Low	Farmer	High
Father's occupational status	Low	0.71	0.15	0.75
		(n=328)	*(n=47)*	*(n=351)*
	Farmer	0.25	0.05	0.17
		(n=81)	*(n=438)*	*(n=168)*
	High	0.75	0.00	0.34
		(n=52)	*(n=31)*	*(n=195)*

Note: The table reports the share of Red Guards in the pooled sample of Red and White Guards by occupational status (HISCAM) of the son and occupational status of the father. Number of observations in brackets.

i.e., low risk aversion and sufficiently high optimism about prospects of future upward mobility. In their study, the agents are assumed to be informed about the true mobility processes and therefore anyone who is poor can be in favor of low levels of redistributions, conditional on their beliefs about their prospects of upward mobility. In Piketty (1995), agents instead learn about mobility processes through their own experience, rationalizing differential behavior among the immobile and downwardly mobile poor. These studies provide therefore a theoretical reason as to why the rate of intergenerational mobility may differ systematically between political groups. Recent empirical work confirms their key underlying assumption, that *beliefs* about socioeconomic mobility and political preferences are causally related. In particular, Alesina, Stantcheva and Teso (2018) show in experimental and survey data that those who have more pessimistic views about intergenerational mobility also tend to support more aggressive government intervention and more redistribution. They demonstrate that this link is causal, as the exposure to pessimistic information about social mobility tends to increase the support for redistributive policies.

Our results contribute to this literature in two aspects. First, they show that socioeconomic mobility is not only related to political preferences, but also to political *action* and conflict. Second, they confirm a specific prediction from the theoretical literature that the *experienced* rate of intergenerational mobility may differ systematically between political groups. We find that this difference can be substantial, and that *downward* intergenerational mobility may contribute to left-wing political action. To illustrate this further, Table 11 reports the share of Red Guard in a 3×3 mobility table that distinguishes between three classes in both the parent and child generation: farmers (HISCAM score=49.91), low occupational status (HISCAM<50), and high occupational status (HISCAM$\geq$50). Because our sample is not representative for the overall population, only the relative size of the cells can be interpreted. Three observations stand out. First, children from farmers (second row) or those who are farmers themselves (second column) are much more likely

to become members of the White Guard. Second, children from high-status parents who themselves achieve only low occupational status (bottom-left cell) are much more likely to be members of the Red Guard than those who achieve high status themselves (bottom-right cell), i.e. the downwardly mobile are more likely to join the Reds. Third, children from low-status parents who themselves achieve high occupational status (top-left cell) are as likely to be members of the Red Guard as those who remain in the lower class (top-right cell), i.e., the upwardly mobile are *not* less likely to join the Reds. The second and third observation explain why the average occupational score is similar for Red and White Guard from low-status parents but very different for those from high-status parent (as shown in Figure 9).

8 Conclusions

We reviewed name-based estimators of intergenerational mobility, based on newly digitized data from Finland and U.S. Census data. To conclude, we summarize our findings that appear most directly relevant for applications:

First, all name-based estimators are predominantly identified from rare names, which are more informative about socioeconomic status than frequent names. This can be problematic if individuals with rare names are not representative for the wider population. Indeed, the average socioeconomic status tends to vary with name frequency, as may the intergenerational process. Because the informational content varies less with name frequency for first names than for surnames, estimators based on first names will be less sensitive to this issue. Irrespectively of the chosen estimator, we propose that researchers test the robustness of their findings with respect to the exclusion or re-weighting of rare names in applications.

Second, name-based estimators weight transmission mechanisms differently than conventional estimators based on direct family links. The intergenerational persistence of status reflects multiple mechanisms, including some that play out at an aggregate or group level. The concern is that name-based estimators weight the latter more heavily than conventional estimators, and as such provide little insight on mobility processes on the individual level. It can be partly addressed by testing how sensitive estimates are to the inclusion of group-level controls, such as location and ethnicity, or other variables that capture group-level processes and covary with names. While implemented in some studies the underlying issue extends to all name-based estimators. We thus propose that applications report evidence on the stability of the results to the inclusion of group-level controls.

Third, names have *added informational content* beyond proxying for a particular socioeconomic status variable. Different studies have different takes on its interpretation.

For some, the insight that name-based estimators capture more than the conventional parent-child estimates is their principal attraction, as it may suggest that they capture something more fundamental (e.g. Clark, 2014). Others use the name-based estimators simply as a feasible "drop-in" replacement for settings in which the direct estimator is infeasible. and consider the added informational content of names a nuisance. Either way, the question *if* names have added informational content is central for the interpretation of any name-based estimator, and should be discussed accordingly – by providing evidence to the extent possible with the data at hand, and by discussing how added informational content would alter the interpretation of the name-based estimates.

Fourth and related, the direct and grouping estimators capture systematically different objects. It is sometimes assumed that because names have added informational content, the name-based grouping estimator must necessarily be larger than the direct estimator. We show that this is true only under certain assumptions on the sampling properties of the data. The grouping estimator will indeed be larger if the offspring and ancestor sample *overlap*, i.e. if parents and their offspring are both sampled (as in complete-count census data). But the grouping estimator can be smaller than the direct estimator if the offspring and ancestor samples do not overlap fully (as in repeated cross-sectional data with only partial coverage of the population). The grouping estimator is indeed highly sensitive to these sampling properties in our example application.

The grouping estimator identifies therefore different conceptual objects depending on the properties of the underlying data. As a consequence, estimates are not necessarily comparable across studies, even if based on the *same* name-based estimator. This in turn may be one reason why some authors find very high rates of intergenerational persistence, while others do not – we conjecture that the variability of grouping estimates across studies is partly a consequence of their sensitivity to properties of the underlying data. We therefore propose that researchers study and report those sampling properties explicitly, and discuss how these properties affect the interpretation of their estimates. In settings where this is possible, it would be useful to report and compare the size of both the "*inclusive*" and "*leave-out*" variants of the grouping estimator, and researchers should always probe the sensitivity of the grouping estimates to sample size.

Finally, researchers can explicitly estimate the attenuation bias in the "leave-out" version that arises from the imperfect measurement of group means. We presented a bias-corrected version of the grouping estimator that is straightforward to construct, and which is based on two observations. First, the grouping estimator corresponds to a TS2SLS estimator, and the finite sample properties of the TS2SLS estimator depend critically on the overlap between the main (i.e., child) and auxiliary (parent) samples (Khawand and Lin, 2015). Second, in non-overlapping samples, the TS2SLS estimator is biased towards zero when the instruments are weak (Choi, Gu and Shen, 2018), and the extent of this bias can

be approximated by running a regression of individual outcomes on their leave-out means in the name group. These observations together imply a simple bias-correction formula for the grouping estimator that could be applied in future applications.

To assess the performance of name-based estimators in a typical example we compared the intergenerational mobility of the two antagonistic parties of the Finnish Civil War 1918, the *Red Guard* and *White Guard*. We find that *all* name-based estimators – both the R^2 and the grouping estimators, based on either first names or surnames – align with the conventional estimator based on direct links, showing that intergenerational mobility and in particular *downward* intergenerational mobility was much higher among the Reds as compared to the Whites. The application illustrates that name-based estimators can be very useful for comparative purposes, despite being subject to many conceptual and interpretational issues.

References

Abramitzky, R., L. Boustan, and K. Eriksson (2020): "Do Immigrants Assimilate More Slowly today than in the past?," *American Economic Review: Insights*, 2(1), 125–41.

Abramitzky, R., L. P. Boustan, and K. 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, R., R. Mill, and S. Pérez (2018): "Linking Individuals Across Historical Sources: a Fully Automated Approach," Working Paper 24324, National Bureau of Economic Research.

Acciari, P., A. Polo, and G. Violante (2016): "'And Yet, It Moves': Intergenerational Mobility in Italy," Discussion paper, mimeo.

Adermon, A., M. Lindahl, and M. Palme (2019): "Dynastic Human Capital, Inequality and Intergenerational Mobility," Discussion paper, mimeo, Uppsala University.

Alesina, A., S. Stantcheva, and E. Teso (2018): "Intergenerational Mobility and Preferences for Redistribution," *American Economic Review*, 108(2), 521–54.

Algan, Y., T. Mayer, and M. Thoenig (2013): "The Economic Incentives of Cultural Transmission: Spatial Evidence from Naming Patterns across France," CEPR Discussion Papers 9416, C.E.P.R. Discussion Papers.

Angrist, J. D. (2014): "The perils of peer effects," *Labour Economics*, 30(C), 98–108.

Angrist, J. D., and J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 1 edn.

Arai, M., and P. Skogman Thoursie (2009): "Renouncing Personal Names: An Empirical Examination of Surname Change and Earnings," *Journal of Labor Economics*, 27(1), 127–147.

Arosalo, S. (1998): "Social Conditions for Political Violence: Red and White Terror in the Finnish Civil War of 1918," *Journal of Peace Research*, 35(2), 147–166.

Baiardi, A. (2016): "The Persistent Effect of Gender Division of Labour: African American Women After Slavery," Discussion paper, mimeo.

Barone, G., and S. Mocetti (2020): "Intergenerational Mobility in the Very Long Run: Florence 1427-2011," *Review of Economic Studies*.

Bénabou, R., and E. A. Ok (2001): "Social Mobility And The Demand For Redistribution: The Poum Hypothesis," *The Quarterly Journal of Economics*, 116(2), 447–487.

Borjas, G. (1992): "Ethnic capital and Intergenerational Mobility," *The Quarterly Journal of Economics*, pp. 123–150.

Braun, S. T., and J. Stuhler (2018): "The Transmission of Inequality Across Multiple Generations: Testing Recent Theories with Evidence from Germany," *The Economic Journal*, 128(609), 576–611.

Carneiro, P., S. Lee, and H. Reis (2016): "Please Call Me John: Name Choice and the Assimilation of Immigrants in the United States, 1900-1930," CReAM Discussion Paper Series 1608, Centre for Research and Analysis of Migration (CReAM), Department of Economics, University College London.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014): "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States," *Quarterly Journal of Economics*, 129(4), 1553–1623.

Choi, J., J. Gu, and S. Shen (2018): "Weak-instrument Robust Inference for Two-sample Instrumental Variables Regression," *Journal of Applied Econometrics*, 33(1), 109–125.

Clark, G. (2014): *The Son Also Rises: Surnames and the History of Social Mobility*. Princeton University Press.

——— (2018): "Estimating Social Mobility Rates from Surnames: Social Group or Dynastic Transmission versus Family Effects," Discussion paper, mimeo.

Clark, G., and N. Cummins (2012): "What is the True Rate of Social Mobility? Surnames and Social Mobility, England 1800-2012," Unpublished working paper.

——— (2014): "Intergenerational Wealth Mobility in England, 1858–2012: Surnames and Social Mobility," *The Economic Journal*, 125(582), 61–85.

Clark, G., N. Cummins, Y. Hao, and D. D. Vidal (2015): "Surnames: A New Source for the History of Social Mobility," *Explorations in Economic History*, 55, 3–24.

Colagrossi, M., B. d'Hombres, and S. V. Schnepf (2019): "Like (Grand)Parent, Like Child? Multigenerational Mobility Across the EU," IZA Discussion Papers 12302, Institute for the Study of Labor.

Collado, M. D., I. O. Ortín, and A. Romeu (2008): "Surnames and Social Status in Spain," *Investigaciones Economicas*, 32(3), 259–287.

Collado, M. D., I. Ortuño-Ortín, and A. Romeu (2012): "Intergenerational Linkages in Consumption Patterns and the Geographical Distribution of Surnames," *Regional Science and Urban Economics*, 42(1-2), 341–350.

Collado, M. D., I. Ortuño-Ortín, and A. Romeu (2014): "Long-run Intergenerational Social Mobility and the Distribution of Surnames," Discussion Paper 2, UMUFAE Economics Working Papers.

Collado, M. D., I. Ortuño-Ortín, and J. Stuhler (2019): "Estimating Intergenerational and Assortative Processes in Extended Family Data," Discussion paper, mimeo.

Durante, R., G. Labartino, and R. Perotti (2016): "Academic Dynasties: Decentralization and Familism in the Italian Academia," Discussion paper, Manuscript, Revise and Resubmit at American Economic Journal: Economic Policy.

Feigenbaum, J. A. (2016): "A Machine Learning Approach to Census Record Linking," Discussion paper, mimeo.

Feigenbaum, J. J. (2018): "Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940," *The Economic Journal*, 128(612), F446–F481.

Fryer, R., and S. Levitt (2004): "The Causes and Consequences of Distinctively Black Names*," *Quarterly Journal of Economics*, 119(3), 767–805.

Goldin, C., and L. Katz (2010): "The 1915 Iowa State Census Project," *Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]*, pp. 12–14.

Güell, M., M. Pellizzari, G. Pica, and J. V. R. Mora (2018): "Correlating Social Mobility and Economic Outcomes," *The Economic Journal*, 0(0).

Güell, M., J. V. Rodríguez Mora, and G. Solon (2018): "New Directions in Measuring Intergenerational Mobility: Introduction," *The Economic Journal*.

Güell, M., J. V. Rodríguez Mora, and C. I. Telmer (2015): "The Informational Content of Surnames, the Evolution of Intergenerational Mobility and Assortative Mating," *The Review of Economic Studies*, 82(2), 693–735.

Hull, P. (2017): "Examiner Designs and First-Stage F Statistics: A Caution," Discussion paper, mimeo.

Inoue, A., and G. Solon (2010): "Two-Sample Instrumental Variables Estimators," *The Review of Economics and Statistics*, 92(3), 557–561.

Jantti, M., J. Saari, and J. Vartiainen (2006): "Growth and Equity in Finland," Discussion Paper DP2006/06.

Johnson, D. S., C. Massey, and A. O'Hara (2015): "The Opportunities and Challenges of Using Administrative Data Linkages to Evaluate Mobility," *The ANNALS of the American Academy of Political and Social Science*, 657(1), 247–264.

Khawand, C., and W. Lin (2015): "Finite Sample Properties and Empirical Applicability of Two-Sample Two-Stage Least Squares," Discussion paper, mimeo.

Kolesár, M., R. Chetty, J. Friedman, E. Glaeser, and G. W. Imbens (2015): "Identification and Inference With Many Invalid Instruments," *Journal of Business & Economic Statistics*, 33(4), 474–484.

Lambert, P. S., R. L. Zijdeman, M. H. D. V. Leeuwen, I. Maas, and K. Prandy (2013): "The Construction of HISCAM: A Stratification Scale Based on Social Interactions for Historical Comparative Research," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 46(2), 77–89.

Lieberson, S., and E. O. Bell (1992): "Children's First Names: An Empirical Study of Social Taste," *American Journal of Sociology*, 98(3), 511–554.

Lindahl, M., M. Palme, S. Sandgren Massih, and A. Sjögren (2015): "Long-term Intergenerational Persistence of Human Capital: An Empirical Analysis of Four Generations," *Journal of Human Resources*, 50(1), 1–33.

Long, J., and J. Ferrie (2013): "Intergenerational Occupational Mobility in Great Britain and the United States since 1850," *American Economic Review*, 103(4), 1109–37.

Miles, A., M. Leeuwen, and I. Maas (2002): *HISCO. Historical International Standard Classification of Occupations*. Leuven University Press, Belgium.

Modalsli, J. (2017): "Intergenerational Mobility in Norway, 1865–2011," *The Scandinavian Journal of Economics*, 119(1), 34–71.

Neidhöfer, G., and M. Stockhausen (2019): "Dynastic Inequality Compared: Multigenerational Mobility in the United States, the United Kingdom, and Germany," *Review of Income and Wealth*, 65(2), 383–414.

Nye, J., G. Mason, M. Bryukhanov, S. Polyachenko, and V. Rusanov (2016): "Social Mobility in the Russia of Revolutions, 1910-2015: A Surname Study," Discussion paper, mimeo.

Olivetti, C., and D. Paserman (2013): "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1930," Discussion Paper 18822, NBER.

Olivetti, C., and M. D. Paserman (2015): "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940," *American Economic Review*, 105(8), 2695–2724.

Olivetti, C., M. D. Paserman, and L. Salisbury (2018): "Three-generation mobility in the United States, 1850–1940: The role of maternal and paternal grandparents," *Explorations in Economic History*.

Paavolainen, J. (1966): *Poliittiset väkivaltaisuudet Suomessa 1918. I. 'Punainen terrori' [Political Violence in Finland 1918. I: 'Red Terror'*. Tammi, Helsinki.

Paik, C. (2014): "Does lineage matter? A study of ancestral influence on educational attainment in Korea," *European Review of Economic History*.

Piketty, T. (1995): "Social Mobility and Redistributive Politics," *The Quarterly Journal of Economics*, 110(3), 551–584.

Solon, G. (2018): "What Do We Know So Far about Multigenerational Mobility?," *The Economic Journal*, 128(612), F340–F352.

Torche, F., and A. Corvalan (2018): "Estimating Intergenerational Mobility With Grouped data: A Critique of Clark's the Son Also Rises," *Sociological Methods & Research*, 47(4), 787–811.

Vosters, K. (2018): "Is the Simple Law of Mobility Really a Law? Testing Clark's Hypothesis," *The Economic Journal*, 128(612), F404–F421.

Vosters, K., and M. Nybom (2017): "Intergenerational Persistence in Latent Socioeconomic Status: Evidence from Sweden and the United States," *Journal of Labor Economics*, 35(3), 869–901.

Yin, P., and X. Fan (2001): "Estimating R 2 shrinkage in multiple regression: a comparison of different analytical methods," *The Journal of Experimental Education*, 69(2), 203–224.

A Appendix

A.1 Red Guard data set

Our sample of members of the Red Guard was constructed by linking two data sources, namely a registry of compensation claims by former members of the Red Guard combined with an archive of individual-level prosecution acts dating back to 1918 from the State Court of Clemency.

In 1973 the Prisoners of War (POW) of the Red Guard were rehabilitated and granted compensation by the Finnish Government. Everyone who was prosecuted by State Court of Clemency and imprisoned in the aftermath of 1918 was entitled to this compensation. The amount varied from a baseline sum of 1,000 Finnish markka (≈ 1,150 Euros in 2018) to 2,500 Finnish markka (≈ 2,900 Euros) depending on the duration of imprisonment.[41]The base population of the Red Guard data set is a registry stored at the National Archives of Finland containing all filed compensation claims in 1973 that were received by Ministry of Social Affairs. After a screening of the received 12,000 pension applications roughly 11,000 claims were approved. We linked registry of pension claims manually based on first names, second names, birth date and birth place to the registry of State Court of Clemency Acts in which all individual acts of the prosecutions in 1918 and 1919 of Red Guardists are included. In total 7,939 successful linkages were made i.e., an act for the individual dating back to 1918-1919 was found in the Registry of State Court of Clemency. From these acts, all individual-level information available, such as sociodemographic background, occupation, and complete name were acquired. We identified the individuals at the Population Register of Finland (PRF) and were able to link them to their relevant social security number with an identification rate of 99.6 percent. More exactly, of the 6,858 cases in our data, only 22 individuals were unidentified. Further, 350 of the identified individuals turned out to be duplicates (due to existence of multiple acts or multiple pension applications of the same individual), and thus 175 excessive rows were deleted. Hence, in total, 6,661 unique individuals were linked to their social security numbers. Our analytic sample includes these individuals.

A.2 White Guard data set

In 1934 the collecting of a registry of White Guard veterans was commenced on the initiative of the Civil Guard, a hybrid of civil war veteran corps and home guard with the aim at assembling a complete registry of White Guard veterans. By the end of 1938 9,917 home interviews were conducted recording individual-level information on sociode-

[41]Everyone who were imprisoned were entitled to a the base compensation of 1,000 marks and the ones who were still imprisoned by the end of the year 1918 received an additional 500 marks for each additional 6 months of imprisonment until a maximum total amount of 2,500 marks.

mographic background, civil war, current occupation and complete name. This registry is administered by the National Archives of Finland. We acquired all individual-level variables for all individual interviews available in this registry and digitized these records in 2015-2016. These individual interviews were matched at the Population Register to social security numbers (issued in 1970). This enables us to measure sample attrition and make the sample of members of the White Guard more comparable the members of the Red Guard in our sample who all survived until 1973.

A.3 Merging harmonized variables from two sources into one data set

Pooling the the two data collections into a pooled data set comprising veterans of the Finnish Civil War in 1918 of both sides was substantially facilitated by the availability of precisely the same key variables for both groups. First, the same socioeconomic outcomes were available for both groups, i.e., highest completed education and occupational status in 1918. Second, names were recorded in the same way for both groups, i.e., a maximum of three first names, the surname including the former surname in the event of a name mutation. Third, both data sets contained the and sociodemographic characteristics such as place of birth and year of birth were recorded in both data entities. The ethnicity of a name was fairly simple to infer for both data entities as Finnish and Swedish belong to different language families (Swedish being an Indoeuropean language and Finnish an Uralic language).

The distinguishing feature of the White Guard data set is the availability of self-reported father's occupation. In order to balance the two data sets we ascertained information on father's occupation through matching individuals to their birth records (which contain information on father's occupation) by a matching algorithm that used complete names, date of birth and place of birth as matching criteria.

As mentioned, in the individual-level records of the National Archives, the father's occupational status is only measured for the members of the White Guard (self-reported through home interviews). We therefore complement our data with a matched subset of digitized birth records that contain the father's occupation. This matching exercise is done for both members of the Red Guard and White Guard in order to be able to evaluate the accuracy of the birth records. We have two independent measures of father's occupation status score: one based on the self-reported occupation by the son through home interviews (that are our main data source) and the other from the matched sons' birth records. The two measures have similar moments and are highly correlated. Moreover, the probability of matching a birth record is uncorrelated with socioeconomic variables (see Appendix Table A4). The match probability differs across regions, which is a consequence of the regionally uneven state of the digitization of Finnish genealogy records for the relevant

cohorts.[42] Because socioeconomic mobility in the sample matched to birth records is close to the average rate in the full sample, this regional selection is not a concern for our analysis (see Appendix Table A6). Taken together, these results suggest that results based on the father's occupational status score from birth records are reliable and comparable to results based on the self-reported status from our main data.

[42] The universe of birth certificates for the years 1850-1900 are digitized for 41 parishes out of 194 parishes in total. For the cohorts considered in our sample, most parishes with digitized birth records are located in two out of ten regions.

Table A1: The Added Informational Content with Other Socioeconomic Outcomes

	Surnames			First Names		
	(1)	(2)	(3)	(4)	(5)	(6)
Father's status	Linear	Flexible	Flexible	Linear	Flexible	Flexible
Other controls	–	–	Yes	–	–	Yes
	Dependent variable: Son's log earnings					
Father's name mean	0.041	0.050	0.045	0.134	0.152	0.151
(log earnings)	(0.082)	(0.089)	(0.088)	(0.069)	(0.073)	(0.074)
AR2	0.024	0.031	0.047	0.026	0.034	0.050
N	2,041	2,041	1,958	2,041	2,041	1,958
	Dependent variable: Son's education					
Father's name mean	0.118	0.090	0.100	0.172	0.173	0.175
(years of education)	(0.051)	(0.051)	(0.052)	(0.051)	(0.049)	(0.049)
AR2	0.057	0.069	0.080	0.059	0.072	0.082
N	3,378	3,378	3,338	3,378	3,378	3,338

Note: The table reports estimates from a digitized sample of the 1915 Iowa State Census (Goldin and Katz, 2000) linked to the 1940 US Federal Census (Feigenbaum, 2018). The first panel reports the coefficients from a regression of son's annual log earnings in 1940 on the father's log annual earnings in 1915 and the mean of the fathers' log annual earnings in the name group, defined by son's surname (columns 1-3) or first name (columns 2-4). The second panel reports the corresponding coefficients from a regression of son's years of education on father's years of education. Other controls include dummies for foreign born, year of birth and state of birth. Standard errors in parentheses.

Table A2: Direct v. Grouping Estimator with Other Socioeconomic Outcomes

	Direct	Surnames			First names		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Group definition	–	inclusive	partial	leave-out	inclusive	partial	leave-out
Overlap		100%	50%	0%	100%	50%	0%
	Dependent variable: Son's log earnings						
Father's log earnings	0.209	0.219	0.157	0.171	0.307	0.205	0.250
	(0.032)	(0.035)	(0.046)	(0.045)	(0.061)	(0.070)	(0.079)
AR2	0.025	0.02	0.008	0.011	0.015	0.005	0.007
N	2,041	2,041	1,252	1,446	2,041	1,427	1,775
	Dependent variable: Son's education						
Father's education	0.264	0.298	0.270	0.237	0.397	0.291	0.214
	(0.023)	(0.027)	(0.027)	(0.035)	(0.047)	(0.045)	(0.053)
AR2	0.056	0.051	0.043	0.029	0.029	0.017	0.006
N	3,378	3,378	2,183	2,452	3,378	2,381	2,942

Note: The table reports estimates from a digitized sample of the 1915 Iowa State Census (Goldin and Katz, 2000) linked to the 1940 US Federal Census (Feigenbaum, 2018). The first panel reports the coefficients from a regression of son's annual log earnings in 1940 on the father's log annual earnings in 1915 (column 1) or the mean of the fathers' log annual earnings in the name group, defined by son's surname (columns 2-4) or first name (columns 5-7). The second panel reports the corresponding coefficients from a regression of son's years of schooling on father's years of schooling. Standard errors in parentheses.

Table A3: Descriptive Statistics of Name Mutations

	Red Guard	White Guard
Number of mutations	582	838
Mutation rate	8.7	8.7
Mutation of name ethnicity	68.4	74.6
Pre-mutation name:		
Mean frequency	6.5	3
Percent unique	16.8	23.7
Post-mutation name:		
Mean frequency	4.1	2.8
Percent unique	11.3	19.7

Source: The Finnish Longitudinal Veteran Database.

Table A4: Sampling of Birth Records

	Birth record observed yes/no			
	White Guards		Red Guards	
	(1)	(2)	(3)	(4)
Years of schooling	0.000	-0.001	0.003	-0.001
	(0.002)	(0.002)	(0.003)	(0.003)
HISCAM score in 1918	-0.000	-0.000	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Father's HISCAM	0.001	0.000		
	(0.001)	(0.001)		
Surname count	0.001	0.001	0.000	0.000
	(0.001)	(0.001)	(0.000)	(0.000)
Region 2		0.088		0.135
		(0.020)		(0.024)
Region 3		0.466		0.454
		(0.023)		(0.027)
Region 4		0.178		0.180
		(0.028)		(0.016)
Region 5		0.114		0.108
		(0.020)		(0.008)
Region 6		0.096		0.174
		(0.023)		(0.012)
Region 7		0.024		0.023
		(0.015)		(0.018)
Region 8		0.034		0.194
		(0.015)		(0.012)
Region 9		0.105		0.106
		(0.016)		(0.021)
Region 10		0.303		0.329
		(0.029)		(0.053)
AR2	0.005	0.154	0.000	0.052
N	4,366	4,366	5,708	5,708

Note: The dependent variable equals one if a son's digitized birth record successfully links father's occupation to the son at www.genealogy.fi. All regressions include a dummy for ethnicity (Finnish sounding name). Robust standard errors in parentheses. Source: The Finnish Longitudinal Veteran Database.

Table A5: White and Red Guard: R^2 Estimator with Truncated Name Distribution

	White Guards		Red Guards	
	(1) Son's schooling	(2) Son's occupational score	(3) Son's schooling	(4) Son's occupational score
R^2 estimator based on surname groups with ≤30 individuals				
Surnames	0.184	0.201	0.077	0.037
	[0.143, 0.239]	[0.155, 0.252]	[0.022, 0.145]	[-0.005, 0.101]
	N=6,851	N=8,079	N=4,825	N=5,344
R^2 estimator based on all surname groups (benchmark from Table 11)				
Surnames	0.180	0.197	0.070	0.037
	[0.139, 0.231]	[0.150, 0.250]	[0.026, 0.121]	[-0.001, 0.083]
	N=7,032	N=8,294	N=5,821	N=6,469

Note: The R2 estimator is the difference between the adjusted R-squared of a model including a complete set of name dummies and an otherwise identical regression in which names are randomly reshuffled. In the first panel, cross-group comparability is enhanced by dropping the most frequent (i.e., least informative) surnames. Based on the harmonization method proposed by Guell et al. (2018), we drop the right tail of the name distribution with more than 30 individuals per name. 95% confidence intervals across 1,000 bootstrap samples in brackets. Source: The Finnish Longitudinal Veteran Database.

Table A6: Mobility Using Alternative Measures of Father's Occupational status score

	Son's Schooling		Son's occ. status 1918		Son's occ. status 1930s	
	(1)	(2)	(3)	(4)	(5)	(6)
Father's occupational status (S)	0.203		0.505		0.603	
	(0.012)		(0.035)		(0.042)	
Father's occupational status (BR)		0.233		0.546		0.636
		(0.013)		(0.041)		(0.051)
AR2	0.280	0.289	0.236	0.212	0.200	0.165
N	879	879	907	907	964	964

Note: The table reports the slope coefficients from a regression of the respective son's variable (top row) on father's occupational status score (HISCAM) as measured by self-reports (S) or by linking digitized birth records of sons at www.anscestry.fi that include father's occupation (BR) in a restricted sample in which both variables are observed. All regressions control for ethnicity (Finnish sounding name). Robust standard errors in parentheses. Source: The Finnish Longitudinal Veteran Database.

Table A7: Intergenerational Mobility of White and Red Guard: Restricted Sample

	White Guard		Red Guard	
	(1) Son's schooling	(2) Son's occ. score	(3) Son's schooling	(4) Son's occ. score
Direct estimator				
Father's occupational status (BR)	0.215	0.400	-0.002	0.055
	(0.024)	(0.086)	(0.008)	(0.043)
	N=273	N=268	N=979	N=1,091
R^2 estimator				
Surnames	0.216	0.146	0.071	0.080
	[0.101, 0.394]	[-0.024, 0.382]	[0.029, 0.119]	[0.033, 0.132]
First names	0.103	0.140	0.016	0.009
	[0.054, 0.158]	[0.078, 0.207]	[0.0009, 0.032]	[-0.005, 0.023]
	N=1,551	N=1,544	N=5,825	N=6,472
Grouping estimator				
Surnames	0.092	0.081	0.002	0.022
	(0.026)	(0.078)	(0.004)	(0.021)
	N=433	N=436	N=3,372	N=3,891
First names	0.130	0.296	0.026	0.119
	(0.021)	(0.089)	(0.005)	(0.028)
	N=1,048	N=1,103	N=5,655	N=6,306

Note: Replication of Table 10, but restricting the estimation sample for the White Guards to those individuals who survived until 1973 in order to make attrition comparable across groups. Standard errors of the in parentheses or 95% confidence intervals across 1,000 bootstrap samples in brackets. Source: The Finnish Longitudinal Veteran Database.