Name-Based Estimators of Intergenerational Mobility: Evidence from Finnish Veterans

Torsten Santavirta[*] and Jan Stuhler[†‡]

September 12, 2019

Abstract

A fascinating development in intergenerational research is the use of *names* – first and surnames – to overcome data limitations. Name-based estimators are underlying innovative research on mobility across multiple generations, historical periods, or regions. However, it remains unclear how different methods relate to each other, and how reliable they are. This paper reviews name-based methods and validates them empirically, based on newly digitized data from Finland that contain names, name mutations, and direct family links. We show that the different name-based methods are closely related, but that their interpretation depends on sampling properties of the data that differ across studies. To demonstrate the reliability of name-based methods 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 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 seminar participants at SOFI at Stockholm University, Uppsala University, Purdue University, Universidad Autonoma de Barcelona, 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 and fascinating development in intergenerational research is the use of *names* to overcome data limitations. When direct family links are not available, names – first and surnames – can serve as a proxy for those links. Based on this insight, researchers have developed different name-based estimators that enabled path-breaking work on the "long-run" persistence of inequality across multiple generations, on historical trends in intergenerational mobility, and its pattern across regions, among other questions (see review in next section). Table 1 provides a partial list of recent contributions.

All listed studies are motivated by the observation that names contain socioeconomic information, but they exploit that information in different ways and rely on different methods. While researchers have emphasized the innovative aspects of their respective approach, few connections have been established between the various methods. This methodological diversity complicates the interpretation of name-based estimators, and impairs their further development. It may also mask to what degree criticisms leveled against one particular approach extend to the other methods.

It is therefore timely to provide a systematic review of name-based estimators of intergenerational mobility. We discuss their respective properties, strengths and weaknesses, and describe links and differences between the various methods that have not been made explicit before. Our arguments are backed up empirically with evidence from U.S. Census data and newly digitized historical data from Finland that are uniquely suited for this purpose. They include all the required ingredients to compare name-based and conventional estimators, namely (i) socioeconomic outcomes across two generations, (ii) direct parent-child links, (iii) first and surnames for both generations, (iv) as well as explicit information on name *mutations*. We conclude our study by comparing the performance of the name-based estimators in estimating mobility rates of the two antagonistic parties in the Finnish Civil War of 1918, the socialist *Red Guard* and the conservative *White Guard*.

We argue that the vast majority of name-based methods can be classified in a simple two-by-two diagram, as shown in Table 2 – with names (*first names* vs. *surnames*) on the horizontal axis and the type of estimator on the vertical one (R^2 vs *grouping* estimators).[1] Although labeled differently by different authors, we argue that most studies use what is fundamentally the same *surname-based grouping estimator*. The estimates from this estimator, however, are not necessarily comparable across studies, as its properties depend critically on the sampling properties of the underlying data – which differ substantially across applications.

[1] We will not review the innovative ways in which names have been used to impute direct links between parents and their offspring (e.g., machine learning algorithms), as described in Long and Ferrie (2013), Johnson, Massey and O'Hara (2015), Modalsli (2015), Feigenbaum (2016), Ruggles, Fitch and Sobek (2017), or Abramitzky, Mill and Pérez (2018). We touch upon some of them further ahead but otherwise focus on name-based estimators of intergenerational mobility that make direct use of names.

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	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
Barone and Mocetti	2016	Working Paper	Surnames, Two-sample Two-stage IV	Repeated cross-section of surname averages	Multigenerational mobility in Florence, Italy (1427-2011)
Nye, Mason, Bryukhanov, Polyachenko, 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

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), Barone and Moretti (2016), Feigenbaum (2018)

Note: The table classifies name-based intergenerational mobility studies according to their empirical methodology, with the exception of frequency-based methods (see Table 1).

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 completely circumventing the need to link generations. If both surnames and status are transmitted from one generation to the next, then rare surnames 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.

The R^2 estimator proposed by Güell et al. is based on the naming process of *surnames*. We note that it can be also usefully applied to measure the informational content of *first names*. As the naming process for first names is very different than for last names, the conceptual motivation of the R^2 estimator given by Güell et al. does not apply to first names. However, the value of first names for mobility research has been demonstrated before in the context of other approaches (Olivetti and Paserman, 2015), and the R^2 estimator based on first names appears to perform well empirically. Its use could therefore be considered in future research.

A more commonly used estimator is the *grouping estimator*, which can be viewed as a two-step estimator. In the first step, the average socioeconomic status within each name group is computed. In the second step, a variant of the conventional intergenerational regression is estimated in which parent socioeconomic status is imputed by the group-level means. The recent studies by Gregory Clark and co-authors are well-known examples (e.g. Clark, 2014, or Clark et al., 2015). Using historical sources that span across several countries and centuries, they document that socioeconomic status regresses only slowly at the surname level. These findings have potentially far-reaching implications and have triggered a lively debate, which we briefly review in Section 2.

Other recent studies impute socioeconomic status by surname as well. Barone and Mocetti (2016) study mobility in the very long run across *six centuries*, using medieval census data for the Italian city of Florence. Using a two-sample IV method, they report that the earnings elasticity of contemporary descendants with respect to medieval ancestors is around 0.04. While small in absolute value, this estimate is much higher than extrapolations of contemporary estimates of the parent-child elasticity would suggest. Under additional assumptions, the relative frequency of surnames in elite groups can also be informative about intergenerational transmission.[2]

Finally, 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.[3] A key advantage of their approach is that first names do not change upon marriage, and therefore remain informative in parental and maternal lineages, and 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 socioeconomic status by first name and the second stage regresses child socioeconomic outcomes against their parental group mean. Using U.S. Census data they find that the rate of social mobility remained flat during the nineteenth century, but saw a drastic break at about 1900, when mobility began to decrease, until around 1920, after which it increased slightly until 1940. Olivetti, Paserman and Salisbury (2018) extend this approach to track paternal and maternal lineages in a multigenerational context.

We argue that these name-based methods are more closely related than what has previously been noted. 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 (2016) for surnames. These two-stage estimators in turn belong to the same class of *grouping estimators* as earlier studies that directly relate surname averages across generations, such as Clark (2014). The literature can therefore be classified into three distinct but closely related approaches, as shown in Table 2. While these approaches are not directly comparable to the conventional estimates in individual-level data, they can be used to study mobility variation across groups or time, or to inform about aspects of the intergenerational process

[2] For example, Clark et al. (2015) study the relative frequency of names on admissions lists of two elite academic institutions, the Universities of Oxford and Cambridge, in data dating back to 1170. Paik (2014) constructs a measure of average regional prestige based on the share of Korean family clan lineages by use of historical data of civil servant exam passers throughout the Joseon Dynasty (1392-1897). He finds a strong correlation between the average historical clan lineage status and the contemporary average educational level of the region. Comparing this correlation across cohorts over time (1955-2000), Paik (2014) shows a substantial decline. The correlation as such and its decline is attributed to intergenerational transmission within families and the weakening of it.

[3] As Clark et al. (2015) note, first names are based on parents' active choices and may thus carry more information about family background than surnames. A surnames link individuals to a *distant* ancestor, unless a name mutation occurred in the recent past. First names are instead chosen by close ancestors.

that are not captured by the conventional estimates.

The grouping and R^2 estimators are subject to similar conceptual issues. First, they are dominated by the influence from rare names. The surname-studies by Clark and co-authors have been criticized on these grounds, but the observation applies to all name-based estimators, including those based on first names (although, as we show, to a lesser extent). Second, name-based estimators weight the underlying transmission mechanisms differently than the conventional (individual-level) estimator. Intergenerational correlations reflect a number of distinct transmission processes within the family (e.g., parental resources and behavior) and on more aggregate levels (e.g., on the ethnic or regional level). Because names vary systematically with the latter, name-based estimators weight these more heavily than the individual-level estimator. We provide evidence on these concerns and describe various other issues that apply to all named-based estimators.

However, the commonalities that we describe do not imply that all name-based studies estimate the same object. In particular, the grouping estimator will estimate 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 being included in the child sample (i.e., the "overlap" between samples). This probability differs widely across studies, hence the estimates from existing name-based studies are unlikely to be directly comparable – even among studies that use the same type of estimator. We confirm this theoretical result by imitating different sampling probabilities within our own data. In many cases, the grouping estimator doubles in size when switching from non-overlapping to fully overlapping samples. The interpretation of the grouping estimator is therefore highly dependent on the sampling properties of the underlying data. A corollary of this finding is that the two-sample IV perspective (TSIV) on the grouping estimator emphasized in recent studies can be problematic. Not only is its exclusion restriction unlikely to hold (Olivetti and Paserman 2015), but the TSIV perspective also implies a polar assumption on the joint sampling probability of parents and their children that will not be appropriate in some settings.

Curiously, if the parent and child samples overlap (i.e. cover the same families), the grouping and direct estimators are particularly closely linked if names had *no* informational content. This is a surprising result, as previous studies emphasize that the grouping estimator relies on a systematic relationship between names and socioeconomic status (see Section 4). The intuition for why this is not necessarily true is that names can inform about intergenerational transmission via two different channels – either because socioeconomic status varies *systematically* across names, or because names link parents and children in a probabilistic sense. Both channels are related, and both are needed to motivate the R^2 estimator (as formalized in Güell, Rodríguez Mora and Telmer, 2015), but in some scenarios only the latter channel is needed to motivate the grouping estimator. Moreover, the grouping estimator may capture different aspects of intergenerational transmission depending on which of the two channels dominates in any given sample.

The development of name-based estimators has led to important insights, but many questions remain. They have not been unconditionally embraced by other scholars, even in interesting historical and other contexts in which they are the only option to study intergenerational mobility. But while there is a lively debate on the validity and interpretation of specific name-based studies (see Chetty et al., 2014; Torche and Corvalan, 2015; Braun and Stuhler, 2018; Güell, Mora and Solon, 2018; Solon, 2018; Adermon, Lindahl and Palme, 2019; Clark, 2018; Choi, Gu and Shen, 2018), no systematic review of all name-based methods has been provided so far. Feigenbaum (2018) comes closest in spirit to our work. Matching fathers from the Iowa State Census of 1915 to their sons in the 1940 Federal Census, he estimates intergenerational mobility based on earnings, education, occupation, and names. He finds that the grouping estimator based on first names and the direct estimator arrive at qualitatively the same conclusion, i.e, that mobility was high in Iowa during the first decade of the 20th century. However, the comparison of name-based methods is not the focus of Feigenbaum's study, and his data is, in some respects, not as suitable for that purpose as our primary source (as it contains only imputed instead of direct family links, and does not include information on name mutations).

In our empirical analyses we use U.S. Census data from the early 20th century as well as historical data on 16,318 veterans of the Finnish 1918 Civil War. These data contain both first names and surnames, name mutations (i.e. both the pre- and post-mutation names), years of schooling for sons, and occupations for both fathers and sons. In addition, we know on what side the individuals fought in the war. We compare the prewar mobility of the members of the socialist Red Guard veterans and the members of the conservative White Guard. We find that all name-based methods consistently estimate lower prewar mobility estimates for the members of the Red Guard as compared to members of the White Guard – consistent with the direct estimator, which shows that the mobility rates and in particular downward intergenerational mobility are far higher among the Red Guard. All name-based methods therefore "pass the test", including the so far unexploited R^2 estimator based on first names.

The remainder of this paper is organized as follows. Section 2 reviews some of the main insights from recent name-based studies. Section 3 presents the data. Section 4 explores the informational content of first and surnames and reviews the R^2 estimators. Section 5 reviews the grouping estimators. Section 6 provides insight about the different method's stance to mutations and evidence using data on mutations. Section 7 contains 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 (2016) 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.[4] However, recent multigenerational studies that directly track family members across three or more generations 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).

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 (2016) 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),

[4] Chetty et al. (2014) in their Online Appendix D and Güell et al. (2018) in their Online Appendix A estimate mobility using Clark's method, but find similar results only when using very common surnames. Other contributions to this debate include Torche and Corvalan (2015), Braun and Stuhler (2018), Solon (2018), Adermon, Lindahl and Palme (2019), Vosters and Nybom (2017), Vosters (2018), and Clark (2018).

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.

Finally, names can be also used to impute direct family linkages in repeated cross-sections (Long and Ferrie 2013, Johnson, Massey and O'Hara 2015, Modalsli 2015). As novel and automated algorithms are being developed for the linkaging process, the availability of multigenerational data containing socioeconomic status and direct links between multiple generations will likely improve further (Feigenbaum, 2016; Abramitzky, Mill and Pérez, 2018). Nevertheless, the aforementioned name-based estimation methods are not at risk of becoming obsolete by the development of automated linking algorithms, given the enormous data requirements of many intergenerational, let alone multigenerational, studies. For example, a study such as Barone and Mocetti (2016) that compares mobility across six centuries would not be feasible using direct linkages. Another advantage of name-based methods is their potential scale. For instance, the full non-anonymous 1940 Federal Census is the first U.S. Census that contains individual-level information on highest completed education and on income for the whole population. While direct parental-offspring links can be recovered for some subsets of the population for which prior censuses are available (as demonstrated by Feigenbaum, 2018), comparisons based on name-based estimators could be conducted for the entire United States.

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 to 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 for both generations renders a sample size of 9,076 observations. Finally, we use the digitized Iowa State Census 1950 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 (based on 1950 occupational income).

3.1 Finnish Longitudinal Veteran Database

We assemble our main sample by combining individual-level data on veterans of the Finnish Civil War 1918 from 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.[5] 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

[5] 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.

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, 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).[6] 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=14,811), the HISCAM score based on occupation in 1918 has a mean of 53.1 (s.d. 10.3). The HISCAM score as of the 1930s was only recorded for members of the White Guard (n=8,723) and has a mean of 59.7 (s.d. 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.

[6] 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

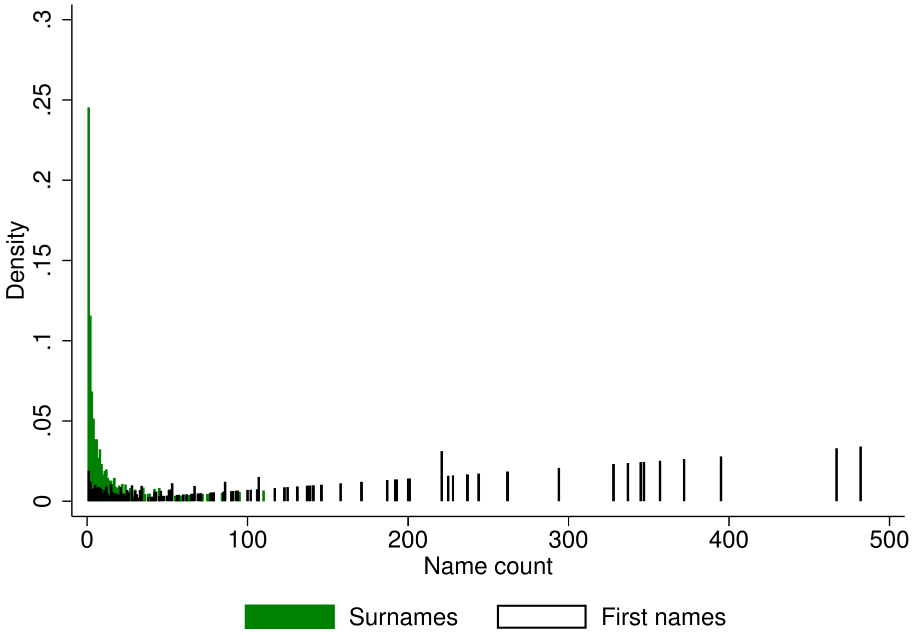
	Red Guards	White Guards
Number of sons	6,666	9,652
Linked fathers (self-reported)		7,051
Linked fathers (birth records)	700	1,164
First names		
Number of distinct names	426	585
Mean frequency per name	15.6	16.5
Sons with singleton first name	2.5%	2.2%
Top-50 names	76.1%	66.5%
Surnames		
Number of distinct 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)	6,507	8,304
mean (std. dev.)	51.47 (7.51)	54.38 (11.92)
Son's occupational status (1930)		8,723
mean (std. dev.)		59.67 (11.85)
Father's occupational status		7,051
mean (std. dev.)		55.17 (12.12)
Father's occupational status (BR)	700	1,164
mean (std. dev.)	48.71 (5.56)	53.22 (9.62)

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

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.

Figure 1: The Sample Frequency of First and Surnames



4 The Informational Content of Names

The starting point for most name-based mobility studies is the observation that names predict socioeconomic status. Both first and surnames are informative about status, but for different reasons. The informational content of surnames stems from what is predominantly 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 – parents *choose* names for their children, and those choices correlate with status. Hence, first names capture parental socioeconomic status, but also differences in name preferences conditional on status. Because these preferences might be intertwined with the mobility process itself, more detailed arguments are necessary to motivate the use of first names in mobility research (Olivetti and Paserman 2015).

Surnames are therefore more straightforward to use, and have been the more popular choice for mobility studies (see Table 2). However, the contrast between first and surnames is not as sharp as it may seem. First, even surnames are eventually subject to individual choice. In fact, Güell, Rodríguez Mora and Telmer (2015) note that name *mutations* – deliberate or accidental name changes – are essential for surnames to retain their informational content. Intuitively, in the absence of such mutations the distribution of surnames would eventually collapse into a small number of frequent and uninformative surnames such as “Smith” and “Jones”. The informational content of surnames depends therefore also on choices, albeit less directly. Second, while individual choice creates conceptual difficulties, it is attractive from a purely predictive perspective. Due to their mechanical transmission, the predictive power of surnames becomes negligible for frequent surnames. In contrast, first names may in principle retain informational content irrespective of their frequency. It is therefore apriori not obvious whether first or surnames are more useful for mobility research.

4.1 The R^2 Estimator

Most name-based studies estimate 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 approach. However, researchers can in principle make inference about the extent of intergenerational mobility without running a single intergenerational regression, by analyzing the informational content of surnames in cross-sectional data (Güell, Rodríguez Mora and Telmer 2015). Intuitively, if socioeconomic status is more strongly transmitted, then surnames should explain a larger share of its variance – the R^2 in a regression of socioeconomic status on name dummies is increasing in the degree of intergenerational status transmission. This *"R^2 estimator"* has been also applied in Barone and Mocetti (2016) and Güell et al. (2018).

We argue that there is a close relation between the R^2 estimator on one hand, and the grouping estimator on the other. First, recent criticisms of the grouping estimator apply similarly to the R^2 estimator, and measures to address those criticisms can be adopted in either approach. For example, Güell et al. note that Clark's method will be prone to bias due to correlates of the grouping principle that themselves may affect economic outcomes, such as ethnicity. However as noted by the same authors, the R^2 estimator reflects those correlates as well.[7] We study this particular issue in Section 5.5. Second, R^2 estimators can be constructed from first instead of surnames, motivated by similar arguments as the two-step estimator proposed by Olivetti and Paserman (2015). We therefore provide evidence on both first name and surname-based R^2 estimators in this section.

4.2 The Informational Content of Surnames

Güell, Rodríguez Mora and Telmer (2015) develop a method that exploits surnames to estimate intergenerational mobility in the absence of direct child-parent links. The method has both intuitive and counterintuitive aspects. The idea that surnames contain socioeconomic information is straightforward: as surnames are passed on from parents to their

[7] An example for the potential influence of ethnicity is the mobility comparison across Italian regions by Güell et al. (2018), in which the region with the highest informational content of surnames is the bilingual region of South Tyrol.

children, they are correlated with other characteristics, such as economic status, that likewise get passed on from one generation to another. The insight that regular disruptions in this process are imperative for surnames to retain their socioeconomic content is perhaps less intuitive.

While surnames can be subject to choice (Collado, Ortín and Romeu, 2008), familial linkages will account for much of the partition into surnames. The rarer a surname, the more informative it is about familial linkages. 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 large share of individuals. This skewness of the surname distribution, or at least its persistence, 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). In the absence of name mutations, a stationary process of intergenerational transmission would over time dilute the informativeness of surnames.

To capture the informational content of surnames, Güell, Rodríguez Mora and Telmer (2015) estimate a linear regression of the economic status y_{is} of individual i with surname s on a set of surname dummies,

$$y_{is} = \beta^{'} Surname_s + \gamma^{'} X_{is} + \varepsilon_{is}. \tag{1}$$

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 adjusted R^2 (AR^2) obtained from this regression is contrasted against a placebo AR^2_P from an otherwise identical regression as (1), in which surnames are reshuffled across individuals (while maintaining their marginal distribution). The *informational content of surnames (ICS)* is defined as the difference between the two measures

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

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

Table 4 reports the ICS estimates in 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'

Table 4: The Informational Content of Surnames

	Son's occupational status				Sons' schooling			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Surname dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic dummies		Yes	Yes	Yes		Yes	Yes	Yes
Region of birth (county)			Yes				Yes	
Region of birth (parish)				Yes				Yes
AR2	0.191	0.253	0.263	0.326	0.256	0.360	0.377	0.453
Implied ICS	0.190	0.151	0.136	0.096	0.256	0.191	0.173	0.129
Bootstrapped 95% CI	[0.145,	[0.117,	[0.103,	[0.084,	[0.201,	[0.151,	[0.133,	[0.111,
	0.240]	0.193]	0.178]	0.149]	0.315]	0.241]	0.217]	0.179]
N	11.505	11.505	11.505	11.505	11.505	11.505	11.505	11.505

Note: The dependent variable is the occupational status score (HISCAM) in 1918 in columns (1)-(4) and years of completed schooling in columns (5)-(8). Demographic dummies include an indicator for ethnicity (Finnish-sounding name), White Guard, and year of birth. Region of birth is classified based on geographic coordinates into 10 synthetic counties (coarse) or 583 parishes (fine). The implied ICS is the difference between the adjusted R-squared reported in the column and the adjusted R-squared of an otherwise identical regression in which the surname dummies are randomly reshuffled. 95% confidence interval across 1,000 bootstrap samples in brackets. Source: The Finnish Longitudinal Veteran Database.

geocoded parishes of birth into 10 synthetic counties by k-medoid clustering.[8] Column (4) replaces these county dummies with indicators at a finer regional aggregation level, distinguishing 582 parishes of birth. Columns (5)-(8) report otherwise identical models with years of schooling as the socioeconomic outcome instead of occupational status.

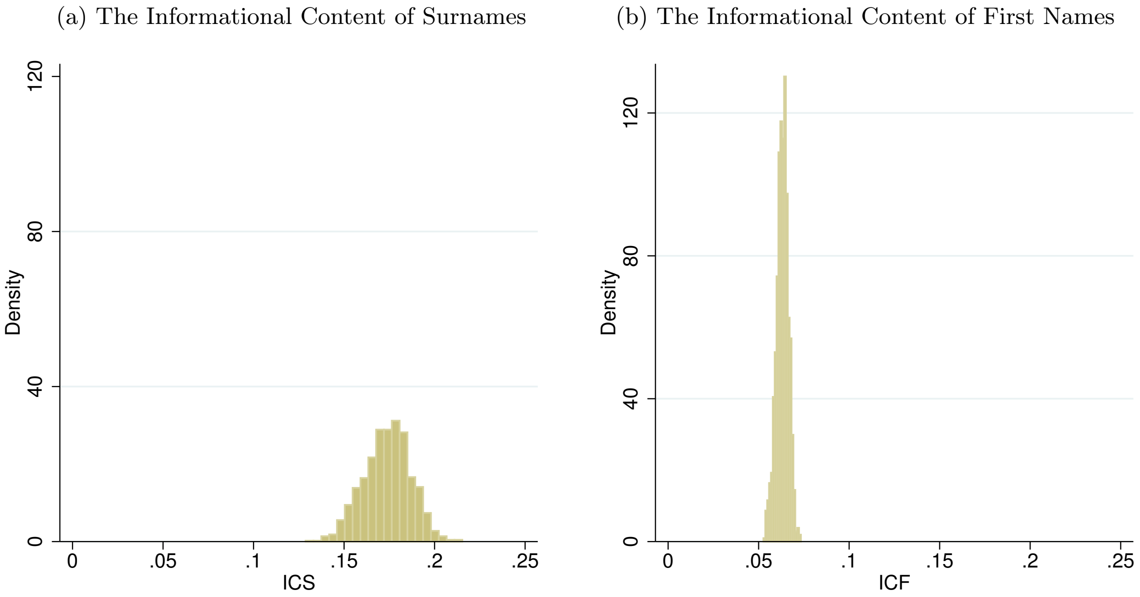
The implied ICS reported in the table is the difference between the adjusted R^2 from these regressions (reported) and the adjusted R^2 from the corresponding placebo regression in which the surname dummies are randomly reshuffled (not reported). To account for the limited size of our sample, we reshuffle the surname dummies 1,000 times and report the mean ICS across these repetitions. As an example, Figure 2a plots the distribution for the estimated ICS from column (7) of Table 4. The estimated ICS is quite stable, indicating that sampling uncertainty is not a major concern. In addition, we report 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 IDs to surname groups that are drawn multiple times). We then report the 2.5 and 97.5 percentile of the resulting distribution.[9]

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

[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, see http://www.sthda.com/english/articles/27-partitioning-clustering-essentials/88-k-medoids-essentials/#pam-concept

[9] These confidence interval account for the uncertainty from the placebo regressions that are specific to the ICS estimator (as illustrated in 2a) as well as standard sampling uncertainty.

Figure 2: The Informational Content and Placebo Distributions


(a) The Informational Content of Surnames

(b) The Informational Content of First Names

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

by a dummy for a Finnish sounding surname, are particularly important. When including dummies for county of birth the ICS is estimated to be 13.6 percent for occupational status and 17.3 percent for years of schooling, but falls to 9.6 and 12.9 percent respectively when controlling for place of birth at the parish level. Ethnicity does not seem to affect the ICS much once region is accounted for in our context, but this is partly because region and ethnicity are overlapping, as the ethnic Swedish minority lived mostly along the coastline. The exclusion of a dummy for Finnish sounding names from X_{is} does not seem to change the ICS markedly when parish fixed effects are included. For occupational status the equivalent to the ICS in column (4) becomes 9.7 and for years of schooling the equivalent to the ICS of column (8) becomes 12.3 percent.[10]

The informational content of surnames is therefore partially due to their covariance with region of birth and ethnicity. As a qualitative result this is not necessarily a concern, as intergenerational persistence on the individual-level likewise reflects variation across regions and ethnic groups. The concern is however that the ICS weights these factors more heavily than the direct estimator (we will return to these considerations in Subsection 5.5). Practitioners should therefore check whether comparative findings are robust to the inclusion of control variables, in particular those capturing ethnicity and regional and

[10] Our ICS estimates for both occupational status and years of schooling are substantially larger than the ones estimated for years of schooling by Güell, Rodríguez Mora and Telmer (2015), which could be either due to differences in the intergenerational process or differences in the distribution of surnames.

cultural differences in naming conventions. Surnames still retain substantial explanatory power when abstracting from such factors.

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

$$\text{ICF} \equiv AR^2 - AR_P^2. \tag{3}$$

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.

Table 5 confirms that a substantial 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 indicator variables for individuals belonging to the White guard (vs. the Red guard), for ethnicity as measured by having a Finnish sounding name, and year of birth to the model. Column (3) further controls for county of birth, while Column (4) instead includes region dummies at the finer parish level. Columns (5)-(8) report otherwise identical models with years of schooling as the socioeconomic outcome. 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 follow an otherwise similar pattern – both are larger for years of schooling than for occupational status, and decrease substantially when place of birth fixed effects are included. When place of birth is defined on a coarser level by aggregating parishes into 10 synthetic counties the ICF is estimated to be 4.8 percent for occupational status and 6.4 percent for years of schooling, but those estimates fall to 2.6 and 3.6

Table 5: Informational Content of First Names

	Son's occupational status				Son's schooling			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
First name dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic dummies		Yes	Yes	Yes		Yes	Yes	Yes
Region of birth (county)			Yes				Yes	
Region of birth (parish)				Yes				Yes
AR2	0.058	0.161	0.174	0.254	0.088	0.244	0.267	0.358
Implied ICF	0.058	0.059	0.048	0.026	0.088	0.080	0.064	0.036
Bootstrapped 95% CI	[0.037,	[0.041,	[0.032,	[0.014,	[0.062,	[0.062,	[0.049,	[0.026,
	0.080]	0.076]	0.062]	0.038]	0.114]	0.098]	0.079]	0.047]
N	11.505	11.505	11.505	11.505	11.505	11.505	11.505	11.505

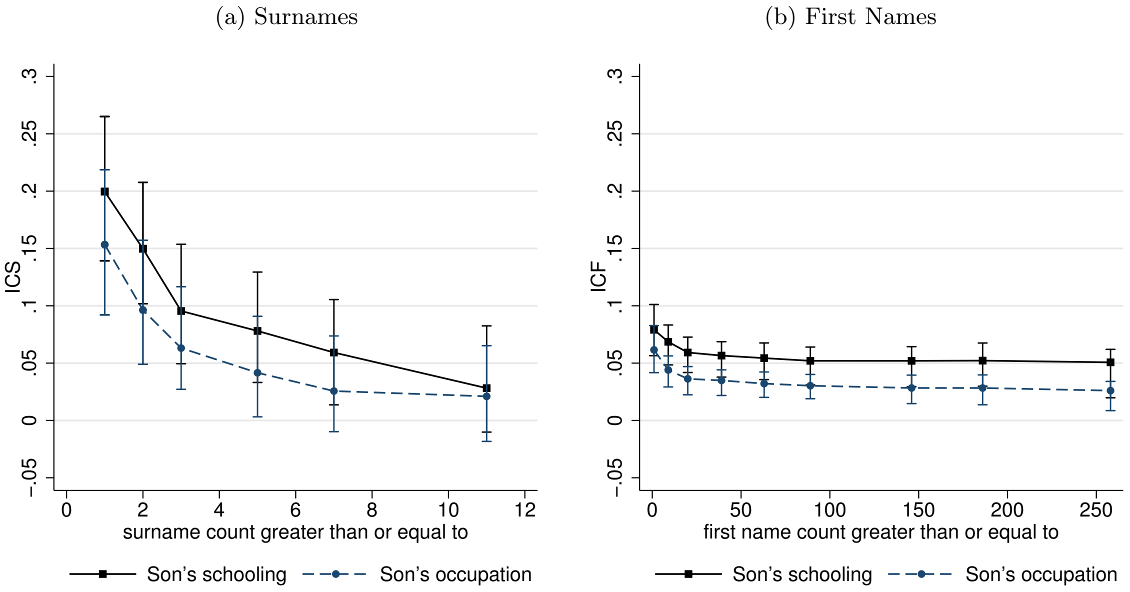
Note: The dependent variable is the occupational status score (HISCAM) in 1918 in columns (1)-(4) and years of completed schooling in columns (5)-(8). Demographic dummies include an indicator for ethnicity (Finnish-sounding name), White Guard, and year of birth. Region of birth is classified based on geographic coordinates into 10 synthetic counties (coarse) or 583 parishes (fine). The implied ICF is the difference between the adjusted R-squared reported in the column and the adjusted R-squared of an otherwise identical regression in which first name dummies are randomly reshuffled. 95% confidence interval across 1,000 bootstrap samples in brackets. Source: The Finnish Longitudinal Veteran Database.

percent, respectively, when considering place of birth at the finer parish level. That the informational content of first and surnames follow similar patterns is remarkable, given that their distributions and transmission processes are so different. As shown in Figure 1, the frequency count of first names is much less right-skewed than the distribution of surnames. Moreover, the naming process of first names follows a polar opposite principle, with "mutations" occurring in every generation. In this sense, the informational content of first names is updated in each generation, and does not depend on family links from the distant past. While the theoretical motivation for the ICS given by Güell et al. does not generalize to the ICF, both measures are informative in practice. We confirm below that in a typical setting, both can capture mobility differences between groups.

4.4 The Properties of Name-based Mobility Measures

Both the ICS and ICF are promising measures for intergenerational mobility research, and we further test their reliability in an empirical application (see Section 7). Even so, name-based estimators such as the ICS and ICF are subject to several conceptual caveats that warrant particular attention, three of which we highlight here. First, they are sensitive to the frequency of names. Second, this name frequency correlates with socioeconomic status itself. And third, names have *added informational content* over and above their role as proxies for the observed socioeconomic status of parents.

Figure 3: Informational Content vs. Name Frequency


(a) Surnames

(b) First Names

Notes: The figures plot the R^2 estimates and corresponding bootstrap intervals from regressions of son's years of schooling (solid line) or son's Hiscam score (dashed line) on a set of surname dummies (sub-figure a) or first name dummies (sub-figure b), separately for name groups with frequencies at or above percentiles $p = \{0, 10, 20, 30, 40, 50, 60, 70, 80\}$. White Guards only.

4.4.1 Name Frequency and Informational Content

A stable finding in our sample is that the informational content is substantially lower for first names than for surnames (see Tables 4 and 5). This result may be a mechanical consequence of the differences in name distributions between first names and surnames, in particular their skewness. Intuitively, infrequent names should be more informative. As to surnames, rare names reflect a higher concentration of kinship linkages. Indeed, Güell, Rodríguez Mora and Telmer (2015) show that rare surnames have substantially higher informational content than more frequent surnames. As to first names, the socioeconomic distribution of fathers is less compressed in larger name groups, as many popular first names tend to be popular among both low and high-status parents.

Figure 3a illustrates how the ICS for son's years of schooling and son's occupational score varies with the frequency of the name groups, Figure 3b shows the equivalent analyses for the ICF. Both measures decrease with the frequency of names, but the rate of decay is much faster for the ICS than the ICF. For surnames, it is indeed the rare surnames, i.e. names that are more directly informative about family links, that drive the informational content, and the ICS becomes small or zero in larger surname groups (as shown in Güell, Rodríguez Mora and Telmer, 2015). This dependency on rare surnames *may* be a concern if the mobility process is suspected to be different for rare than for more frequent surnames.

We show in Section 5 that this appears not to be the case in our data; the intergenerational coefficient from direct parent-child links is remarkably insensitive to the frequency of either first names or surnames. The observation othat common surnames have low informational content has however important implications for the grouping estimator, as we discuss below.

Rare first names have much lower informational content than rare surnames, but the ICF remains remarkably constant for the more frequent first names – the ICF is comparatively flat across the frequency distribution. First names may remain informative in larger name groups because of aspirational naming or differences in name preferences across groups. For example, names that have a royal or noble connotation may be generally popular, but may still be correlated with parental characteristics, and as such maintain their informational content. Lieberson and Bell (1992) and Olivetti and Paserman (2015) document differential naming patterns by parental socioeconomic status. We show in Table 6 that they also differ between members of the White and Red Guard. Strikingly, among the White Guard, none of the top-5 most prestigious names (as measured by mean occupational status) is of Finnish origin, and all use the Swedish spelling form (e.g., Eric vs. Erkki).

The difference in the decay of the informational content with name frequency reflects the different mechanisms via which first names and surnames carry information, with the mechanical transmission process underlying surnames washing out in larger name groups while the choice process underlying first names remains relevant. The flatness of the ICF with respect to name frequency is a potential advantage, and suggests that estimators based on first names are based on a more representative part of the population than surname-based estimators.

4.4.2 Name Frequency and Socioeconomic Status

Another potential caveat of name-based mobility estimators is that socioeconomic status tends to decrease with the frequency of a person's first or surname. Figure 4 plots the average years of schooling of sons across bins of the name frequency distribution (the pattern is similar for son's and fathers' occupational scores). The magnitude of this relationship 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. When grouping the data by name frequency, the most common first and surnames in our sample have indeed the lowest average status.

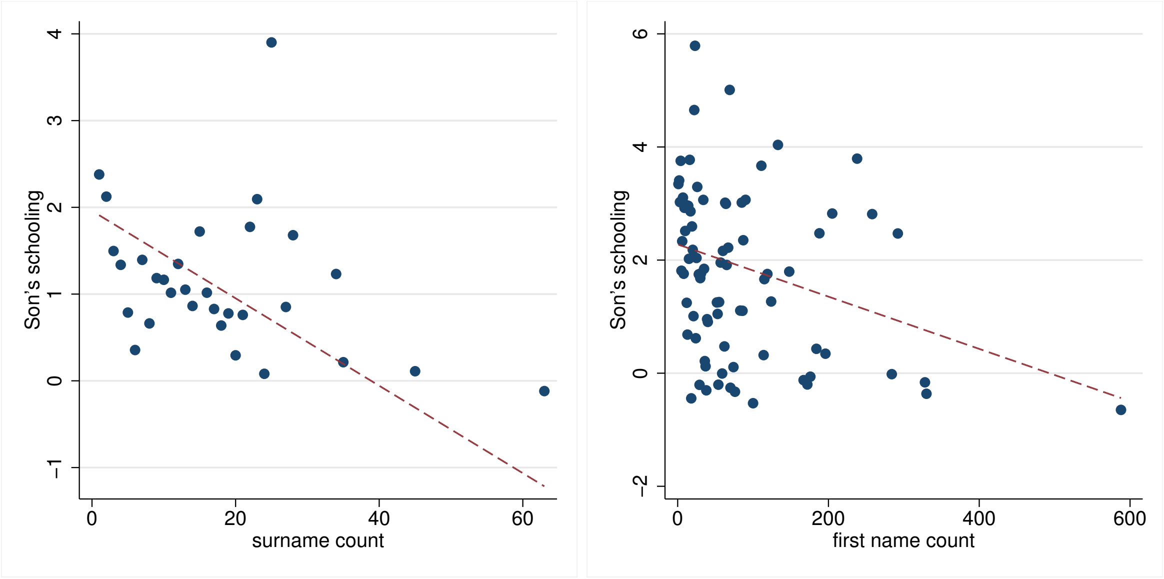
This pattern suggests that new surnames are predominantly created by individuals with high socioeconomic status. The existence of a link between socioeconomic status and the birth process of surnames (e.g. as signaling behavior by successful dynasties) has been hypothesized by Collado, Ortín and Romeu (2008), who find a negative relation

Table 6: Most and Least Prestigious First Names

	Red Guard	White Guard
Rank:	Most prestigious	
1	Adolf	Harald
2	Harald	Eric
3	Elis	Jarl
4	Maurits	Carl
5	Rudolf	Harry
Rank:	Least prestigious	
1	Juhana	Sulho
2	Antton	Emmanuel
3	Jooseppi	Nikodemus
4	Eelis	Aate
5	Manu	Eeli

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

Figure 4: Socioeconomic Status Decreases in Name Frequency


(a) Surnames

(b) First Names

Notes: Binscatter plot of the sons' average years of schooling against the frequency of the surname (sub-figure a) or first name (sub-figure b). White Guards only.

between socioeconomic status and name frequency in Spain. We provide direct evidence on this "selective mutation" hypothesis in Section 6. Similarly, our evidence suggests that families with low socioeconomic status choose more common first names. The observation that people bearing rare names are systematically different can be problematic, given that name-based mobility measures are identified primarily from such infrequent names.

4.4.3 The *Added* Informational Content of Names

The motivation for using names in mobility research 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.[11] Because our data also 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 father's imputed occupational score based on his surname (i.e. the mean occupational status in the surname group). If surnames were merely an imprecise proxy for the occupational status of the father then the coefficient of the imputed occupational score should be insignificant.

Instead, we find that surnames 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 statistically and substantially 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 added informational content over and above the occupational status of the parent. The coefficients on name group means are also positive in the IPUMS Linked Representative Sample 1880-1900 (Panel B), but the relation is weaker and not statistically significant. In the Linked 1915 Iowa State Census Sample, only first names are found to have added informational content (Panel C). Finally, Appendix Table A1 presents results 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 economically and statistically more significant 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.

These findings imply that sons' names are correlated with a number of parental background variables, only one of them being the observed socioeconomic status. It is unclear

[11] 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	0.135	0.140	0.105	0.228	0.193	0.205
(occupational status, HISCAM)	(0.027)	(0.034)	(0.026)	(0.028)	(0.037)	(0.027)
AR2	0.240	0.313	0.373	0.246	0.320	0.377
N	5,996	5,996	5,996	5,996	5,996	5,996
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.

how failure to weight these unobservable correlates affects comparisons, e.g., regional comparisons of the degree of mobility based on the R^2 of names. The observation that names have added informational content, over and above an individual's observed socioeconomic status, alters however the interpretation of any name-based estimator. This observation 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. The question as to why names have additional informative content is key for the interpretation of any name-based estimator, and should be addressed more systematically in applications.

5 The Grouping Estimator

A more commonly used approach imputes the average socioeconomic status of each name, in order to then estimate a variant of the standard intergenerational regression in which name-group averages replace the often unavailable individual socioeconomic outcomes. While different studies present their empirical specifications in different ways, we note that fundamentally they all use the same type of estimator – a Wald or *grouping* estimator, in which groups are defined by first names or surnames. Despite this similarity, different name-based studies have produced very different estimates, an observation that we aim to rationalize here.

We first link the grouping estimator to the *direct* (individual-level) estimator and show that their relative size depends crucially on (i) the degree to which names have *added* informational content (see Section 4.4.3) and (ii) the sampling scheme, especially the conditional probability that a parent is sampled when his or her child is being sampled (i.e., the degree to which the parent and child sample *overlap*). The grouping estimator will generally be larger than the direct estimator if the offspring and ancestor samples overlap, i.e., if an ancestor is sampled whenever his or her offspring is sampled. In contrast, the grouping estimator can be smaller than the direct estimator when the offspring and ancestor samples do not overlap fully, as in repeated cross-sectional data with partial coverage of the population.

The observation that name-based estimates are often larger than individual-level estimates holds therefore not unconditionally, but depends on the way the intergenerational data is sampled. The grouping estimator behaves very differently in data in which the parent and child samples overlap, and in data in which they do not. Accordingly, the sampling scheme could be an important factor in explaining why different name-based studies have found very different patterns. While grouping estimates are much larger than direct estimates in some studies, others find the reverse pattern (Clark 2014; Clark and Cummins 2014; Olivetti and Paserman 2015). A case in point is Feigenbaum (2018), who within the same study finds both upward biased and downward biased grouping estimates as compared to the corresponding direct estimates.[12]

In overlapping samples, the grouping estimator is remarkably insensitive to other properties of the sample, such as sample size, the name frequency distribution, or the informational content of names. If the parent and child samples do not overlap, these properties however gain importance. While ceteris paribus a greater name frequency has a direct positive effect on the grouping estimator (Olivetti and Paserman, 2015), the informational content of names tends to decline with name frequency (Güell, Rodríguez Mora and Telmer, 2015), such that its net effect is ambiguous. Moreover, because the influence

[12] We are comparing the direct estimates of Table 4 and Table 6 in Feigenbaum (2018) with the corresponding grouping estimates in Table 7.

of sample size and name frequency are highly non-linear, the grouping estimator can be stable in some settings while being unstable in others. Fortunately, its stability can be easily tested.

Finally, we link the grouping estimator to the R^2 estimator presented in the previous section. A low informational content of names as captured by the R^2 estimator corresponds to a "weak" first stage in the grouping estimator. Curiously, that is not much of a concern if the parent and child samples overlap. In such settings, the grouping is a standard 2SLS estimator, and is biased towards the OLS estimator if the name instruments are weak – but such bias is desirable if the (feasible) grouping estimator is meant to approximate the (infeasible) OLS estimator. When the parent and child samples do not overlap, the grouping estimator corresponds instead to a two-sample IV estimator and is biased towards *zero* if the instruments are weak (see Choi, Gu and Shen, 2018). While a low informational content of names becomes more consequential in such settings, it has only a small attenuating effect as long as the sample size is sufficiently large.

These arguments suggest that the different estimators proposed in the literature are more closely related than has been noted before. This has a number of interesting implications. For example, while the grouping estimator based on surnames as used by Clark (e.g., Clark 2014; Clark and Cummins 2014) has been the focus of methodological scrutiny, similar criticisms apply in fact to all name-based estimators. This link however also applies to potential refinements of the methodology. In particular, we argue that a control variable strategy as used by Güell, Rodríguez Mora and Telmer (2015) for the R^2 estimator should also be adopted in applications based on the grouping estimator.

5.1 The Grouping Estimator

The evidence in Clark (2014) and related studies (e.g. Clark and Cummins 2012a; Clark 2012; Clark and Cummins 2014) is primarily based on regression to the mean of surname groups. 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 (or earlier) generation. The evidence in Olivetti and Paserman (2015) and Barone and Mocetti (2016) is instead based on a two-sample two-stage least squares (TS2SLS) estimator. However, a two-stage least squares estimator based on dummy variables is analytically tantamount to running a weighted linear regression on a set of group means.[13] For this reason, an instrumental variable estimator based on a set of dummy variables is also called the *grouping estimator*. The approach by Clark and co-authors is therefore equivalent to the IV approach used in more recent studies, as

[13] 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. 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).

long as group means are appropriately weighted. Accordingly we adopt the label *grouping estimator* for either approach.

The two-sample IV perspective is useful in so far as (i) uncertainty from the first stage is taken into account in the estimation of standard errors, and (ii) name groups are automatically weighted by their frequency (if the second stage is estimated on the individual level). These issues would need to be separately addressed in alternative implementations of the grouping estimator. 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. Name-based grouping estimators are constructed as a TS2SLS estimator, but are not used for identifying a particular causal effect. We highlight another reason why the TS2SLS perspective can be problematic – the properties of name-based grouping estimators depend critically on the extent to which the parent and child samples overlap (i.e., are drawn from the *same* families), and the TS2SLS perspective implicitly imposes a polar assumption on this probability.

5.2 Grouping vs. Direct Estimator

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},

$$y_{ij} = \beta x_{ij} + \epsilon_{ij} \tag{4}$$

with the corresponding grouping estimator, in which x_{ij} is replaced by a group mean defined by son's first names (as in Olivetti and Paserman, 2015) or surnames (see references in Table 2).

It will be crucial if the group mean is defined over the parents of the sampled children, or over *other* individuals who merely share the name j. We consider the two polar cases. We consider first the "*short*" group-level regression

$$y_{ij} = \pi \bar{x}_j + v_{ij}, \tag{5}$$

where $\bar{x}_j$ represents the "*inclusive*" mean, which averages over the parents of sampled children, i.e. $\bar{x}_j = \frac{\sum_i x_{ij}}{N_j}$. Equation (5) is the relevant object if families in the parent and child samples *overlap*. Such overlap will occur if the grouping estimator is applied in complete-count census data, or data that track families according to some fixed criteria.

In other settings, the parent and child samples might not overlap, or at least not fully overlap – for a family i in name group j we might observe an ancestor or a descendant, but not both. To illustrate how this affects the grouping estimator, consider the group-level

Table 8: 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%	50%	0%	100%	50%	0%
Panel A: Finnish Longitudinal Veteran Database							
Father's occupational	0.585	0.625	0.575	0.322	0.747	0.660	0.590
status (HISCAM)	(0.014)	(0.017)	(0.022)	(0.027)	(0.029)	(0.033)	(0.032)
AR2	0.238	0.193	0.153	0.033	0.198	0.084	0.054
N	5,996	5,996	3,887	3,864	5,996	4,449	5,831
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	2080	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.

regression

$$y_{ij} = \kappa \bar{x}_{(i)j} + u_{ij} \tag{6}$$

in which $\bar{x}_{(i)j}$ represents the "*leave-out*" mean in which each descendant's own ancestor is excluded, i.e. $\bar{x}_{(i)j} = \frac{N_j \bar{x}_j - x_{ij}}{N_j - 1}$. Equation (6) represents the grouping estimator in settings in which there is zero or only negligible overlap between the parent and child samples, for example because they are small and independent draws from the overall population.[14] Note that $\bar{x}_{(i)j}$ corresponds to the predicted value of x_{ij} underlying the jackknife instrumental variables (JIVE) estimator (Kolesár et al., 2015).

[14] In practice, researchers do not construct leave-out means, but draw y_{ij} and $\bar{x}_j$ from separate samples. The advantage of using the leave-out means to represent such settings is that the sample size remains more comparable to the inclusive grouping estimator. We also tested the performance of split-sample grouping estimators with zero overlap. Because of the greater reduction in sample size, estimates from the split-sample estimator are generally smaller than estimates from the corresponding leave-out estimator.

Empirical Evidence. Table 8 reports estimates from equations (4) to (6), 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 (4), which are $\hat{\beta} = 0.585$ 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 different versions of the grouping estimator, which for comparability are based on the same sample as the direct estimator. The group means are defined over surnames in columns (2)-(4) and over first names in columns (5)-(7). Estimates based on equation (5) and inclusive means (with full overlap between the parent and child sample) are reported in columns (2) and (5), respectively. The estimates based on "inclusive" group means 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 generally larger than the direct estimator, contrary to what one might infer from some prior comparisons (e.g., Güell, Mora and Solon, 2018).[15] In columns (4) and (7), we report estimates based on equation (6) and leave-out means (with zero 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 are either greater (Panel A, first names) or smaller than the direct estimates $\hat{\beta}$ (all other cases). The gaps between the inclusive and leave-out variants are particularly large in the linked U.S. Census samples (Panel B and Panel C, first names). 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. $\hat{\kappa} = 0.179$). In contrast, the direct and surname-based grouping estimates are quite similar in the linked 1915 Iowa State Census Sample, irrespectively of which sampling scheme the mean is based on.

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 that are spaced 20 or 30 years apart, which contain at least some family lineages. A partial overlap between the descendant and ancestor samples yields estimates in between these two extremes. For illustration, columns (3) and (7) report estimates from a grouping estimator based on parent and child samples that overlap by 50 percent.[17] We

[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, 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 1880-1900, the grouping estimator is estimated to be 0.095 and 0.195 for 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

provide more detailed evidence on how the grouping estimator varies with intermediate degrees of overlap below.

These findings are robust to specification choice and also hold in alternative outcome variables. 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 8. We again find that the leave-out grouping estimator is smaller than the inclusive variant, and either larger or smaller than the direct estimates. Estimates based on the grouping estimator are therefore not directly comparable to direct estimates, and their interpretation depends on the sampling properties of the underlying data. We formalize these arguments in the next sections.

The Grouping Estimator with Inclusive Means. By formalizing the relation between the direct and grouping estimators, we illustrate that their relative size depends on multiple factors, including the sampling properties of the underlying data. Our arguments resemble arguments from the literature on peer effects, in which grouping estimators have sometimes been misinterpreted (Angrist 2014). For simplicity, the exposition is in terms of population moments.

Consider first the "*short*" group-level regression based on the inclusive mean $\bar{x}_j$ in equation (5). With both direct family links and names we can also estimate the corresponding "*long*" regression,

$$y_{ij} = \pi_0 x_{ij} + \pi_1 \bar{x}_j + v_{ij}, \tag{7}$$

which includes both direct and group effects.[18] The outcomes y_{ij} and group means $\bar{x}_j$ 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

$$\pi = \frac{Cov(y_{ij}, \bar{x}_j)}{Var(\bar{x}_j)} = \frac{Cov(\pi_0 x_{ij} + \pi_1 \bar{x}_j, \bar{x}_j)}{Var(\bar{x}_j)} = \pi_0 \frac{Cov(x_{ij}, \bar{x}_j)}{Var(\bar{x}_j)} + \pi_1 = \pi_0 + \pi_1, \tag{8}$$

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

IPUMS linked representative sample 1880-1900), we draw without replacement two sub-samples of size xN that have the maximum possible size given intended overlap p. Since $p = max\,(2(x - 0.5), 0)$ we have (for $p > 0$) that $x = (p + 1)/2$. For example, for an intended overlap of 50% ($p = 0.5$) we draw two samples of size $x = 0.75$ (corresponding to about $N = 6,807$ in the IPUMS linked representative sample 1880-1900). Since some names are not included in both samples, the effective number of observations as reported in Table (8) is below this theoretical upper bound.

[18] As noted in Section 4.4.3, many different models could rationalize why names have *added* informational content (i.e., $\pi_1 \neq 0$). We do not need to impose any specific model to derive our main arguments, but provide an example below.

direct and long regressions is

$$\beta = \frac{Cov(y_{ij}, x_{ij})}{Var(x_{ij})} = \frac{Cov(\pi_0 x_{ij} + \pi_1 \bar{x}_j + v_{ij}, x_{ij})}{Var(x_{ij})} = \pi_0 + \pi_1 \frac{Cov(\bar{x}_j, x_{ij})}{Var(x_{ij})}. \tag{9}$$

The combination of equations (8) and (9) then yields

$$\pi = \pi_0 + \pi_1 = \beta + \pi_1 \left(1 - \frac{Cov(\bar{x}_j, x_{ij})}{Var(x_{ij})}\right), \tag{10}$$

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 (see Section 4.4.3). 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] Our empirical results as reported in Table 8 are therefore not specific to our samples, but exemplify a general point – the grouping estimator will tend to be larger than the direct estimator if the child outcomes y_{ij} and parent mean outcomes $\bar{x}_j$ are sampled from the same families, as in complete-count census data, or data that track families according to some fixed criteria. For example, this finding rationalizes why the grouping estimator is always larger than the direct estimator in linked U.S. tax data (see Online Appendix Table V in Chetty et al. 2014).

Equation (10) has been similarly derived in Adermon, Lindahl and Palme (2019). It also underlies a critical review of grouping estimators by Güell, Mora and Solon (2018), who note that π_1 could be substantial in some studies. Güell et al. argue 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, we emphasize that equation (10) holds only in a particular setting – if the parent and child samples overlap *completely*. Most studies are instead based on partially overlapping samples, and as seen in Table 8 (and derived formally in the next section), the grouping estimator behaves very differently in such settings. Accordingly, equation (10) is only of limited use for cross-study comparisons, and differences in sampling properties might be the primary reason why group-level estimates differ so much across studies.

[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, such as group-level causal effects (as in Borjas, 1992), observable status being only an imperfect proxy for latent advantages (as in Clark, 2014), or "aspirational" naming (Olivetti and Paserman, 2015).

The inclusive grouping estimator collapses into the direct estimator ($\pi = \beta$) if names have no added informational content ($\pi_1 = 0$). Interestingly, with full overlap, names do not need to have *any* informational content (as defined in Section 4) for the grouping estimator to work. Several previous studies have emphasized that the grouping estimator relies on the assumption that names carry information about socioeconomic status. This condition is indeed necessary in some settings (see below), but it is not necessary if the parent and child samples overlap. The intuition behind this result is that names approximate direct family links, and therefore capture the *intergenerational* transmission of socioeconomic status – even if they have no systematic relation to the *cross-sectional* variation in socioeconomic status within any given generation (e.g. even if names are randomly distributed). Moreover, in this particular case the grouping estimator is consistent for the direct intergenerational coefficient.[20] Again, this result holds only if the parents of sampled children are included in the estimation sample, i.e., if there are actual parent-child links that names can approximate.[21]

The Grouping Estimator with Leave-Out Means. In many other applications however, the parent and child samples do not overlap, or do not fully overlap. 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 (6), and the corresponding "*long*" regression

$$y_{ij} = \kappa_0 x_{ij} + \kappa_1 \bar{x}_{(i)j} + u_{ij}. \tag{11}$$

The relationship between the short and long regression equations is then

$$\kappa = \frac{Cov(y_{ij}, \bar{x}_{(i)j})}{Var(\bar{x}_{(i)j})} = \frac{Cov(\kappa_0 x_{ij} + \kappa_1 \bar{x}_{(i)j}, \bar{x}_{(i)j})}{Var(\bar{x}_{(i)j})} = \kappa_0 \frac{Cov(x_{ij}, \bar{x}_{(i)j})}{Var(\bar{x}_{(i)j})} + \kappa_1, \tag{12}$$

and between the direct and the long regression,

$$\beta = \frac{Cov(\kappa_0 x_{ij} + \kappa_1 \bar{x}_{(i)j}, x_{ij})}{Var(x_{ij})} = \kappa_0 + \kappa_1 \frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(x_{ij})}. \tag{13}$$

[20] This implication corresponds to the observation that IV estimators are biased towards OLS when the instruments are weak. While that is an undesirable property in most applications, it is desirable in the context here, in which the grouping estimator is used as a replacement for the (infeasible) OLS estimator.

[21] In a two-sample IV setting, the limited informational content of names would equate to a weak first stage regression of parent's socioeconomic status on names, which biases the TSIV estimator towards zero (see Choi, Gu and Shen, 2018). However, this argument does not apply here, because the parent and child samples are not independent draws.

Finally, combining equations (12) and (13) yields

$$\kappa = \beta + \kappa_1 \left(1 - \frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(x_{ij})}\right) - \kappa_0 \left(1 - \frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}\right) \tag{14}$$

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

Equation (14) characterizes the relation between the grouping and the direct estimator when the child and parent samples do not overlap.[22] 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 8). This pattern is in contrast to the estimator based on the "inclusive" mean $\bar{x}_{ij}$, which under plausible assumptions is always larger ($\pi > \beta$). The properties of the grouping estimator depend therefore crucially on the sampling scheme – it will tend to be larger than the direct estimator if parent and child generations are sampled from a similar set of families, but can be smaller than the direct estimator if not.

If names have no *added* informational content ($\kappa_1 = 0$) then the "long" equation (11) collapses into the direct one ($\kappa_0 = \beta$), and equation (14) collapses to

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

The leave-out grouping estimator understates the direct estimator, again in contrast to the "inclusive" grouping estimator, which collapses on the direct estimator ($\pi = \beta$). If names have low predictive value (i.e., low informational content) for own socioeconomic status, the leave-out mean $x_{(i)j}$ will be approximately uncorrelated with x_{ij}, and the grouping estimator will be close to zero. Even if names do have predictive value for socioeconomic status (high informational content), the grouping estimator will severely understate the direct intergenerational coefficient if the means $\bar{x}_{(i)j}$ are constructed over too few individuals (such that the denominator in the ratio in equation (15) is large compared to its numerator).[23] These implications correspond to the observation that

[22] 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. That sampling assumption is comparable to the assumption that a child's own parent is left out in the construction of the group mean, as we assumed here. As such, equation (14) corresponds closely to equation (2) in Olivetti and Paserman (2015). The contrast to equation (10) illustrates that the relation between the direct and grouping estimator depends strongly on sampling properties, an argument that we return to below.

[23] The ratio in equation (15) can be close to one even if names are not very predictive, as long as the group means are constructed within sufficiently large name groups. The grouping estimator can therefore behave quite differently if names have *no* vs. only *minor* informational content.

two-sample IV estimators are biased towards zero when the instruments are weak (Choi, Gu and Shen, 2018).

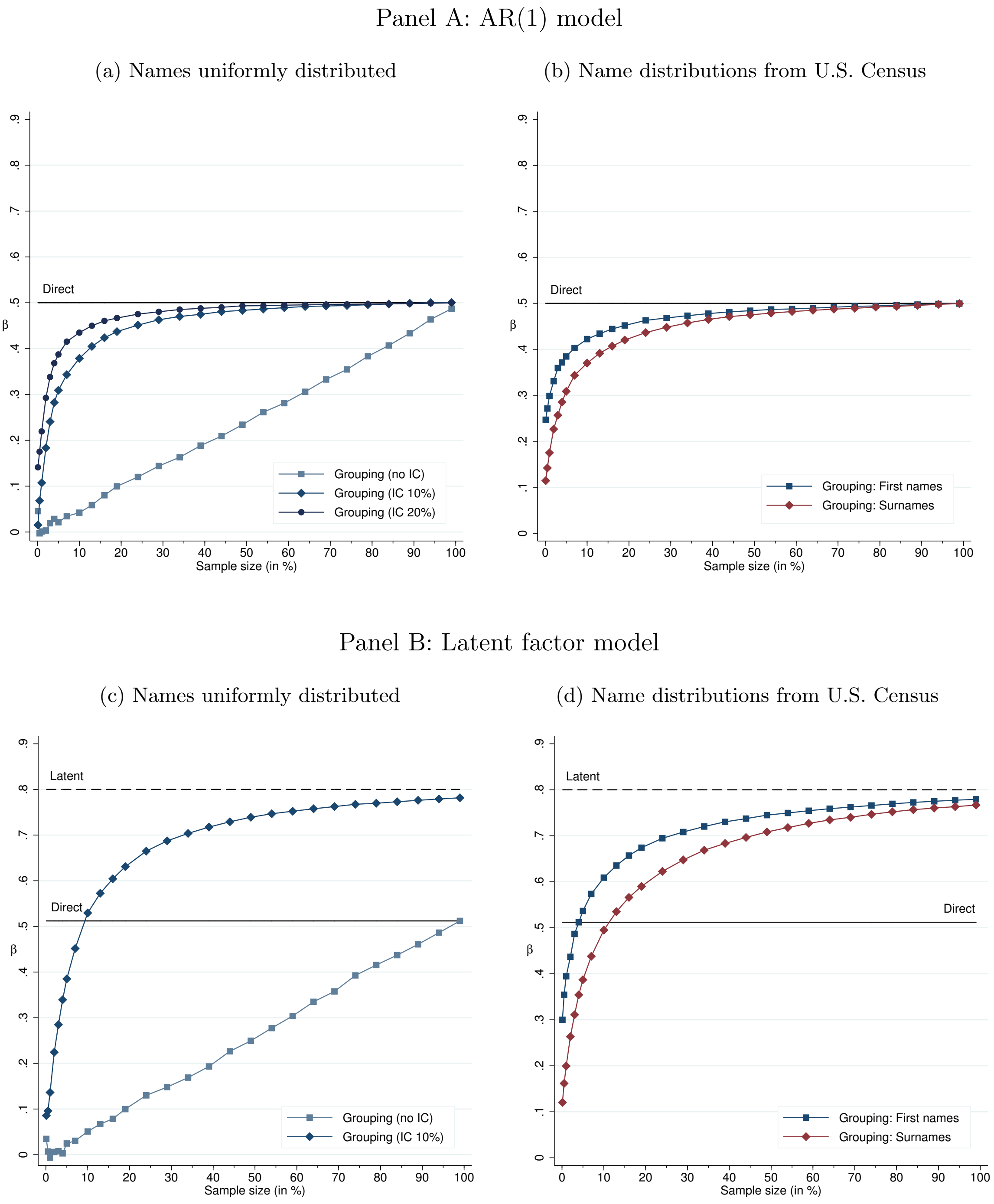
Comparison of equation (15) to equations (10) and (14) however illustrates that these results apply only to some applications of the grouping estimator, as they depend on the assumptions that (i) the parent and child samples do not overlap (i.e. that the two samples in the TSIV setup are independent), and that (ii) names do not have added informational content (i.e., that the instrument exclusion restriction holds). The name frequency distribution, the sample size, and the strength of the name instruments strongly affect the probability limit of the grouping estimator in such settings, but they matter less if the parent and child samples overlap.

The sharp contrast between the "inclusive" and "leave-out" version of the grouping estimator is somewhat counterintuitive. The inclusive and leave-out mean should be highly correlated in large samples, so why would it matter if the grouping estimator is based on one or the other? The two means indeed tend to be highly correlated, even in our modestly sized samples. For example, 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).

To reduce the apparent attenuation bias in equation (15), researchers might aim to maximize the ratio $\frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}$ by restricting the sample to name groups that are sufficiently large. Limiting the sample to more frequent names would indeed decrease $Var(\bar{x}_{(i)j})$. However, the observation that the informational content of names is smaller for more frequent names (Güell, Rodríguez Mora and Telmer, 2015, and Section 4) suggests that $Cov(\bar{x}_{(i)j}, x_{ij})$ will also decrease in name frequency. It is therefore apriori not certain that restricting the sample to more frequent names would reduce the attenuation bias. Instead, the attenuating term $\frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}$ could be directly estimated, even in a two-sample setting. Accordingly, researchers can study which sampling restrictions reduces the attenuation bias most effectively, or correct for that bias. For example, Olivetti and Paserman (2015) estimate the size of the attenuating term in their decomposition of the grouping estimator.

Simulation Evidence. Our comparison between the "inclusive" (full overlap) and "leave-out" (no overlap) variants of the grouping estimator illustrates the polar cases, but the

Figure 5: 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

Notes: The figures plots the slope coefficient of the grouping estimator from separate regressions based on differently sized samples (x-axis). The data generating process underlying sub-figures (a) and (b) is an AR(1) process with $\beta = 0.5$. Generated variables are normally distributed, and may include name fixed effects (→ IC). The data generating process underlying sub-figures (c) and (d) is a latent factor model given by the equations $y_{it} = \rho e_{it} + u_{it}$ and $e_{it} = \lambda e_{it-1} + v_{it}$, standardized such that y_{it} and e_{it} have mean zero and variance one for all t, 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 (20,000 names, uniformly distributed frequency between 1 and 500). Sub-figures (b) and (d) are instead based on the frequency of female first names and male surnames as observed in the 1920 U.S. Census (see Olivetti and Paserman, 2015).

sampling scheme of many applications falls in between these two extremes. For example, if the researcher observes random samples for the parent and the child generation, some children will have their parent sampled while others will not (partial overlap). To understand such intermediate cases, we provide simulation-based evidence on how the grouping estimator varies with (i) sampling probabilities, the (ii) name frequency distribution, the (iii) informative content of names, and (iv) the underlying data generating process.

We consider two different intergenerational processes. In subfigures (a) and (b) of Figure 5, we simulate data according to an AR(1) process with $\beta = 0.5$ (i.e., we use the standard parent-child regression as our data generating process). This model is a natural baseline, in that it abstracts from any independent role of names in the transmission of status (i.e., the AIC is approximately zero, $\pi_1 = \kappa_1 = 0$). In subfigures (c) and (d), we instead consider the latent factor model that is underlying the argument by Clark (2014), and which is a potential candidate to rationalize recent evidence on multigenerational correlations across multiple generations (Braun and Stuhler 2018). The key parameters of this model determine the rate of transmission of latent advantages, and the signal-to-noise ratio of observed status as a measure for latent advantages. We choose these parameters such that implied value for β is similar to the simpler AR(1) model (see table notes).

Subfigure (a) is based on a simulated name distribution with uniformly distributed name frequencies. We consider three variants of the AR(1) process, with socioeconomic status being randomly distributed across parents such that names have no informative content (*no IC*), name fixed effects explaining ten percent of the total variance in status (*IC 10%*), or names explaining twenty percent (*IC 20%*). We generate parent and child status for the entire population, draw sub-samples of different sizes as indicated on the x-axis, and then estimate the grouping estimator within each of these sub-samples.

If names have no informational content, the grouping estimator grows linearly in the sampling probability. Intuitively, the grouping estimator always captures the intergenerational transmission for "complete" parent-child pairs, even if names are not systematically related to socioeconomic status in the cross-section. The grouping estimator therefore increases in the overlap between the parent and child samples, yields an estimate of zero if there is no overlap, and coincides with the direct estimator under full overlap – consistent with the analytic expression for the inclusive and leave-out estimators in equations (10) and (15).

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 informative content of names followed already from equation (15), but Figure 5a illustrates that this increase is highly non-linear in sample size. If names are reasonably informative about socioeconomic status, the grouping estimator is fairly stable with respect to the sampling probability. However, it drops steeply when the sampling probability drops below some critical threshold. For example, in our simulation the group-level estimate is around 0.4 when sampling 10% of the parent and child generations (i.e. 1% overlap), but it drops to 0.25 in 5% random samples (0.25% overlap).

The grouping estimator can therefore be extremely sensitive to sample size in some settings, while being stable in others. The suddenness with which the grouping estimator switches from stable to unstable is striking. The switch is even more sudden when names have higher informational content (*IC 20%*). We therefore recommend that researchers always test the sensitivity of their estimates to sample size, e.g., by applying the grouping estimator to sub-samples of their main sample. If the estimates drop strongly, they are likely in the unstable range, and the grouping estimator will substantially understate persistence. If the estimates remain stable, one is operating in the stable range, and the grouping estimator will be less attenuated.

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 *female first names* and *surnames* from the 1% sample of the 1920 U.S. Census, as also used by Olivetti and Paserman (2015). To approximate the distribution of names in the complete-count Census, we scale up the number of observations per name, but also the number of names itself (as many less frequent names will not have been contained in the 1% sample).[24] Sub-figure (b) plots the grouping estimates as based on first names (blue line) or surnames (red line). As data generating process we again use the AR(1) model, with names explaining 10% of the variation in socioeconomic status. The pattern is qualitatively similar as for uniformly distributed name frequencies. However, the surname-based grouping estimator is more sensitive to sample size than the grouping estimator based on first names. The reason is that the average frequency is much lower for surnames than for first names, such that – for a given informational content – the mean status in the name group is more noisy and less correlated with the actual parent's status. Consistent with evidence in Olivetti and Paserman (2013, Section 7.1), we find that the name frequency distribution has a only a limited impact on the grouping estimator (sub-figure b). In smaller samples, though, the difference between surname and first name

[24] 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 X 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 female first names) and average frequency per name (11.1 vs. 10.4) as the actual 1% sample of the 1920 U.S. Census.

distributions can become sizableconsistent with the observation by Olivetti and Paserman that the surname-based grouping estimator is very small in their samples.

Finally, we repeat this analysis using the latent factor model as our data generating process. Sub-figure (c) is again based on a simulated name distribution with uniform name frequency, while sub-figure (d) is based on the actual name distributions as observed in the U.S. Census. Because names have added informational content (AIC) in the latent factor model, the grouping estimator can now be larger than the direct (conventional) estimator. Intuitively, the grouping "averages away" idiosyncratic variation in status, such that the group means provide a good approximation of the mean latent status of the respective group (see Clark, 2014). Accordingly, in large samples with high overlap, the grouping estimator approximates the persistence in latent advantages (assumed to be 0.8 in our simulation). However, this result depends again critically on sample size. In small samples with little overlap, the grouping estimator will understate latent persistence, and also be substantially below the direct estimator. This observation might help to explain why some authors have found much smaller estimates on the surname level than Clark and co-authors: the joint sampling probability may be the key to reconcile some of the apparent contrasting results in the literature.

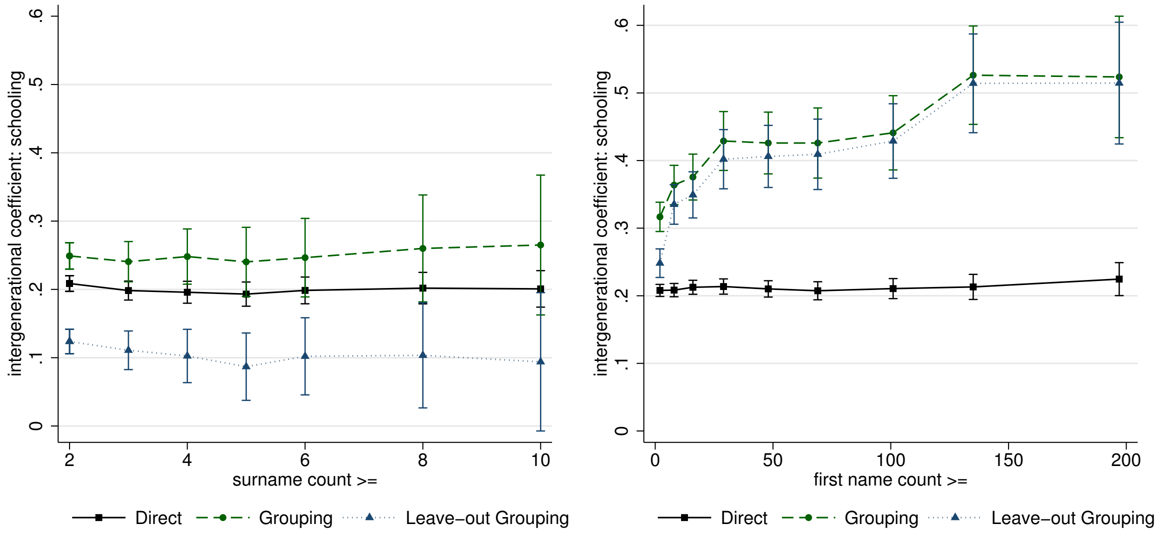
5.3 Grouping vs. R^2 Estimator

How does the grouping estimator relate to the informational content of names, as captured by the R^2 estimator from Section 4? The answer depends again on the extent to which the parent and child samples overlap. In fully overlapping samples, the "inclusive" grouping estimator is weakly greater than the direct estimator regardless of the informational content of names. If names have no *added* informational content ($\pi_1 = 0$), the informational content of names has no influence on the probability limit of the grouping estimator. This is illustrated in sub-figure (a) of Figure 5, in which the group-level estimates are similar in 100% (i.e., fully overlapping) samples regardless of the informational content of names. However, the precision of the grouping estimates does increase in the informational content of names, which amplifies the variance of the group means.

The informational content of names does affect the probability limit of the grouping estimator in overlapping samples if names have *added* informative content ($\pi_1 > 0$), because the ratio $\frac{Cov(\bar{x}_j,x_{ij})}{Var(x_{ij})}$ in equation (10) corresponds to the R^2 in a regression of parent outcomes on name dummies.[25] This is seen in sub-figure (c) of Figure 5: in samples generated by the latent factor model, the grouping estimator is much larger if names have

[25] The R^2 of a bivariate regression is equal to the slope coefficient from the forward regression (e.g., individual outcomes on group means) times the coefficient from the reverse regression (group means on individual outcomes), i.e., $R^2 = \frac{Cov(\bar{x}_j,x_{ij})}{Var(\bar{x}_j,)}\frac{Cov(\bar{x}_j,x_{ij})}{Var(x_{ij})}$. Because the regression of a variable on its group means produces a coefficient of one, we have $R^2 = \frac{Cov(\bar{x}_j,x_{ij})}{Var(x_{ij})}$.

Figure 6: The Grouping Estimator vs. Name Frequency


(a) Surnames

(b) First names

Notes: 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), on the imputed occupational score based on surnames (sub-figure a), or first names (sub-figure b), separately for name groups with frequencies at or above percentiles $p = \{0, 10, 20, 30, 40, 50, 60, 70, 80\}$. White guards only.

informational content, even if parent and child samples overlap completely.

The grouping estimator tends to be much more sensitive to the informational content of names in non-overlapping samples, because the ratio $\frac{Cov(\bar{x}_{(i)j},x_{ij})}{Var(\bar{x}_{(i)j})}$ in equations (14) and (15) increases in the informational content of names. To show this, decompose parental status x_{ij} as the sum of a name fixed effect and an idiosyncratic component, $x_{ij} = \mu_j + \tilde{x}_{ij}$, such that $\frac{Cov(x_{ij},\bar{x}_{(i)j})}{Var(\bar{x}_{(i)j})} = \frac{Var(\mu_j)}{Var(\mu_j)+\frac{1}{N-1}Var(\tilde{x}_{ij})}$. This ratio is therefore close to one if $Var(\mu_j)$ is large compared to $Var(\tilde{x}_{ij})$, *or* if the name group size N is large. Accrodingly, the grouping estimator can be large even if the informational content of names is very low, if the sample size is sufficiently large (see also Olivetti and Paserman, 2015). We face therefore a trade-off between the informational content and sample size, and a lack of informational content becomes more consequential in smaller samples. These implications are again visible in sub-figure (a) of Figure 5. With limited overlap and sample size, the grouping estimator depends critically on the informational content of names. In such settings, the grouping estimator also becomes sensitive to the name frequency distribution, as illustrated in sub-figure (b).

The R^2 and the grouping estimators are therefore linked, and insights from one have implications for the other. Particularly interesting is the observation that the informational content of names drops with name frequency (Güell, Rodríguez Mora and Telmer, 2015, and Figure 3). On the other hand, for a given informational content, the ratio $\frac{Cov(x_{ij},\bar{x}_{(i)j})}{Var(\bar{x}_{(i)j})}$ increases in the frequency with which a name is observed. It is therefore ambiguous if the grouping estimator increases if the sample is limited to more frequent names. To illustrate, Figure 6 plots the direct estimator, the "inclusive" grouping and the "leave-out" estimates in our Finnish sample in different samples with varying minimum name frequency (as indicated on the x-axis). The direct estimates are insensitive to name frequency, and as large in small as in larger name groups. The grouping estimator based on surnames is likewise stable. In contrast, the grouping estimator based on first names increases in name frequency, because the informational content declines less strongly for first names than for surnames (see Figure 3).[26]

5.4 Grouping Estimators in the Prior Literature

As we show above, the statistical properties of the grouping estimator depend on the probability with which a child's parent (or an ancestor's descendant) is included in the sample. This does not imply that the "inclusive" variant is necessarily better than the "leave-out" variant. Rather, because their statistical properties can differ so strongly, it would be helpful if researchers indicated which version is identified in their setting (or which mixture between the two), and how it maps into their structural or descriptive parameter of interest.

As an example, consider the issue of sample attrition from out-migration, which has received some attention in the literature. Barone and Mocetti (2016) 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 over that long period (and includes non-descendants who in-migrated from other areas).[27] As the authors note, the exclusion of migrant descendants might affect their estimates if the decision to migrate covaries with socioeconomic mobility. However, our findings imply that it will affect the grouping estimator even if the decision to migrate is random. Excluding a migrant excludes an ancestor, and therefore pushes the grouping estimator from the "inclusive" towards its "leave-out" version – which mechanically decreases the grouping estimates, even if the socioeconomic mobility of migrants were not different from non-migrants.[28] This issue is specific to the grouping estimator, and does not apply to the individual-level direct estimator (in contrast

[26] Consistent with our findings here, individual-level estimates of income mobility are insensitive to name frequency in U.S. tax data, while the grouping estimator based on surnames increases with sample size (see Chetty et al. 2014, Online Appendix, Table V).

[27] Migration and name mutations are important sources of sample attrition when considering outcomes across multiple generations. For example, Barone and Mocetti 2016 link socioeconomic outcomes across six centuries or about *twenty* generations, demonstrating the potential reach of named-based mobility studies. In this setting, only half of the surnames in the historic (full) census are found also in the modern (full) records, and some of that gap will be due to same attrition from migration.

[28] This argument could potentially explain why the estimates of Barone and Mocetti (2016) increase when they take measures to reduce sample attrition from migrants.

to the non-random selection of migrants, which affects both).

In comparison, the sample underlying the first-name grouping estimator in Feigenbaum (2018) is less susceptible to out-migration, thanks to the shorter analysis period and the use of a nation-wide Census with full-population coverage in the descendant's generation. The parent of a sampled child is nevertheless unlikely to be included in the group means. To impute the average occupation score among fathers, Feigenbaum uses the data provided by Olivetti and Paserman (2015) as based on the IPUMS excerpt of the 1910 Census. Because this excerpt only captures 1 percent of the full Census, the vast majority of children will not have their father included in the estimation sample used to impute father's status.

While the grouping estimator in Barone and Mocetti (2016) corresponds to a mix between the "inclusive" and "leave-out" variants, the grouping estimator used by Feigenbaum (2018) and Olivetti and Paserman (2015) will correspond more closely to the "leave-out" version. To be clear, neither variant is preferable per se, our argument just affects the question as to how existing estimates in the literature should be interpreted and compared. For example, a low degree of overlap may be one reason why the estimates in Olivetti and Paserman are small compared to some of the surname-based estimates from other studies. While it remains an open question as to which variant tracks the direct estimator better, a take-away from our comparison is that one should not deliberately avoid overlap between the parent and child samples, because the grouping estimator is much less sensitive to sample size when based on inclusive means.[29] Such categorizations are not always as straightforward as in the studies discussed here, and we propose that researchers discuss explicitly the degree to which the ancestor and descendants' overlap in their samples.

5.5 Sensitivity to the Inclusion of Control Variables

A recurring concern in the debate around name-based estimators is that they might weight group-level and individual-level transmission processes differently than the conventional individual-level estimator. In particular, an important criticism of surname-based methods is that they may primarily capture the role of ethnic or national origin in the transmission of advantages (Torche 2015, Chetty et al. 2014). We note that those criticisms do not only apply to the the surname-based grouping estimator (as used by Clark, 2014) at which they were explicitly aimed at, but to all name-based estimators. A potential strategy to address such concerns is to either exclude those names that are most indicative of origins, or to include indicators of those origins as a control. Indeed, the inclusion of such controls has been standard in applications based on the R^2 estimator proposed by Güell,

[29] For example, in Feigenbaum's linked 1915 Iowa State Census Sample the inclusive grouping estimator (0.533) is slightly closer to the direct estimator (0.441) than the grouping estimator reported in his study (0.353, Panel (b) of Table 7), which is based on the group means from Olivetti and Paserman (2015) with little overlap. It is much closer than the leave-out grouping estimator based on the smaller set of fathers included in his own linked sample for Iowa (0.215).

Table 9: Stability of Name-based Estimators to the Inclusion of Controls

	Dependent variable: Son's schooling				
	(1)	(2)	(3)	(4)	(5)
Direct estimator					
Father's occupational status	0.211	0.207	0.205	0.189	0.157
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)
R^2 estimator					
Surnames (ICS)	0.256	0.210	0.227	0.171	0.133
	–	–	–	–	–
First names (ICF)	0.162	0.140	0.128	0.093	0.047
	–	–	–	–	–
Grouping estimator					
Surnames	0.224	0.220	0.215	0.195	0.156
	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
First names	0.286	0.274	0.267	0.233	0.168
	(0.009)	(0.009)	(0.010)	(0.009)	(0.010)
Finnish		Yes	Yes	Yes	Yes
Year of birth			Yes	Yes	Yes
Region of birth (county)				Yes	
Region of birth (parish)					Yes
Observations	5,343	5,343	5,343	5,343	5,343

Note: Members of the White Guard only. The direct estimator in column (1) refers to a univariate regression of son's years of schooling on 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 (Finnish-sounding name), 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.

Rodríguez Mora and Telmer (2015). 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 9 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. The R^2 estimators are most sensitive, in particular if based on first names, whereas the conventional intergenerational regression using direct links is the most stable one. 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.

6 Name Mutations

While the transmission of male surnames is often 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, name 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.

While surname mutations tend to occur in most Western societies, their frequency vary much over time. Paik (2014) reports that during the Japanese occupation, many Koreans strategically changed their clan lineage originating in the imperial period to rarer surnames. 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 also to ethnic Finnish names. 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 while others took names of former presidents Baiardi (2016). Episodes such as the aforementioned are extreme events, but it is not unusual in contemporary contexts to observe particularly active periods of name changes.[30]

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. Because that process might be context-specific, 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,

[30] 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).

Table 10: Informational Content of Surnames Pre/Post-Mutation

	Post-mutation	Pre-mutation
Dependent variable: Son's occupational status		
Surname dummies	Yes	Yes
AR2	0.263	0.249
Implied ICS	0.136	0.123
95% CI	[0.103, 0.178]	[0.088, 0.159]
N	11,505	11,505
Dependent variable: Son's schooling		
Surname dummies	Yes	Yes
AR2	0.377	0.364
Implied ICS	0.173	0.161
95% CI	[0.133, 0.217]	[0.124, 0.200]
N	11,505	11,505

Note: Replication of columns (3) and (7) in Table 4 using the post-mutation names (column 1) and corresponding estimates using pre-mutation names (column 2). All regressions include dummies for year and region of birth (10 synthetic counties). 95% confidence interval across 1,000 bootstrap samples in brackets. Source: The Finnish Longitudinal Veteran Database.

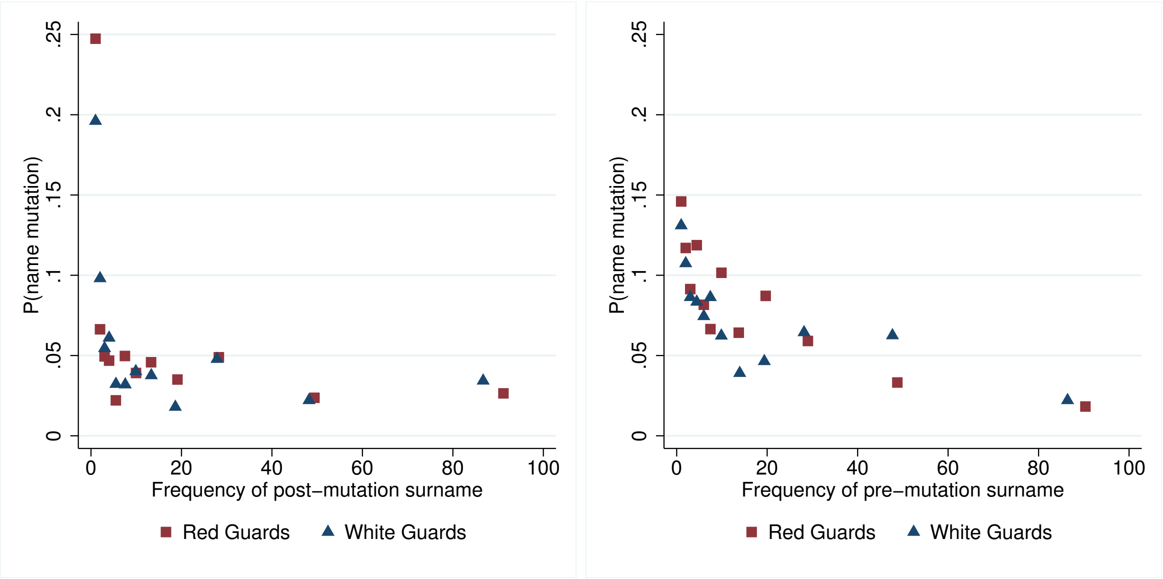
for both White and Red Guard (see Table A3 in the Appendix). For comparison, the estimated lifetime mutation rate in Güell, Rodríguez Mora and Telmer (2015) is only about 0.25 percent. We observe nearly 600 name mutations among the Red Guard, and more than 800 name mutations among the White Guard.

Table 10 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, with the drop being statistically significant at the 1-percent level. As illustrated in Figure 7a, post-mutation surnames are likely to be infrequent surnames, with the share of individuals who actively chose their surname being five times 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 4).

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 more likely to change their surname than 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

Figure 7: Name Mutations vs. Name Frequency

(a) Post-Mutation Surnames (b) Pre-Mutation Surnames



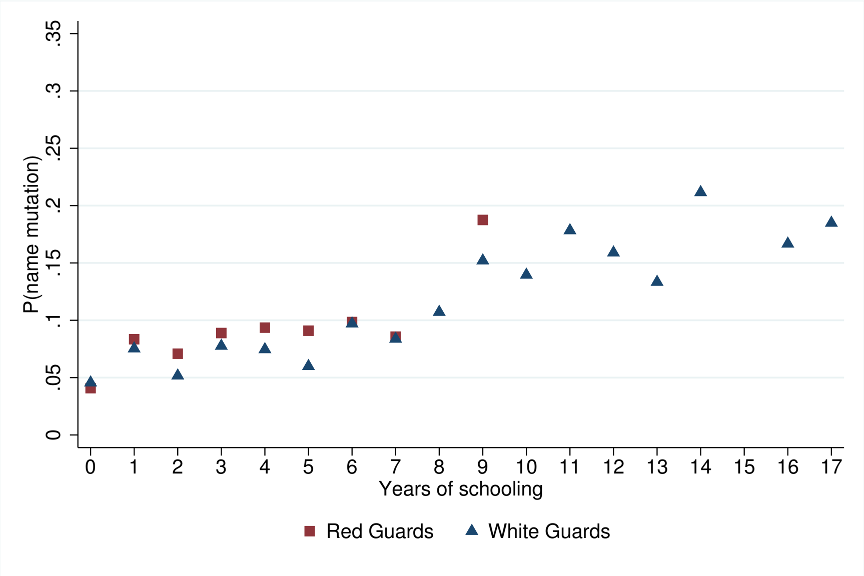
Notes: Binscatter plot of indicator for name mutation against the frequency of the pre-mutation (sub-figure a) or post-mutation (sub-figure b) surname.

frequency. Rare names have a higher informational content, which may either create incentives to pick such name (if the signal is intended) or to abandon it (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).

In addition to a shift in the surname distribution, mutations also update the socioeconomic content of a name. Important for name-based mobility studies is the observation that the frequency of a surname is inversely correlated with socioeconomic status (Section 4). Because we observe name changes, we can directly test if 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 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.[31] Collado, Ortín and Romeu (2008) show that

[31] 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 occupational status, as immigrants tend to have lower occupational status. See also the related literature on the economic incentives of name changes for immigrants and the positive consequences of cultural assimilation

Figure 8: Socioeconomic Bias in Name Mutations



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

in Spain, people bearing uncommon surnames tend to enjoy a higher socioeconomic status than people bearing more common surnames, and note that many of the rarer surnames in the 20th century did not exist in the 19th century. They argue that surnames act as a signaling device for successful dynasties, in particular by combination of the two previous surnames into a new surname. Such combinations are specific to Spanish naming convention, but Figure 8 suggests that name mutations are socioeconomically biased in Finland as well.

7 Applications

We have highlighted the R^2 estimator's and grouping estimator's close relationship to the conventional estimator based on direct links between generations, 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.

7.1 Intergenerational Mobility and the Finnish Civil War 1918

Our data allow us to compare mobility between distinct groups of the Finnish society during the beginning of 20th century. For Finland, this was an interesting period of

(Abramitzky, Boustan and Eriksson, 2016; Algan, Mayer and Thoenig, 2013; Arai and Skogman Thoursie, 2009; Carneiro, Lee and Reis, 2016).

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. We then put the name-based methods to test. We therefore compare these methods in the type of setting for which they were designed for, in which linkages based on individual identifiers are uncommon.

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. The strength of the Red Guard was roughly 80,000 men (including enforcements from the newly formed Soviet Union) while the White army consisted of 70,000 men. Owing to its better organization, trained officers (some of whom had received their training in Germany during World War I) and soldiers, the White Guard were militarily superior and eventually managed to defeat the Red Guard.

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 prewar social mobility in Finland might have been an important factor contributing to the outbreak of the Finnish Civil war. A related and interesting 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.

Descriptive statistics and measurement

Members of the White Guard have on average more schooling ($t = 53.61$), higher occupational status ($t = 17.23$) and more compressed first and surname distributions 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.

In order to evaluate the accuracy of this source, we also match birth records to a subset of the members of the White Guard. We therefore 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 state of the digitization of Finnish genealogy records for the relevant cohorts.[32] 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.

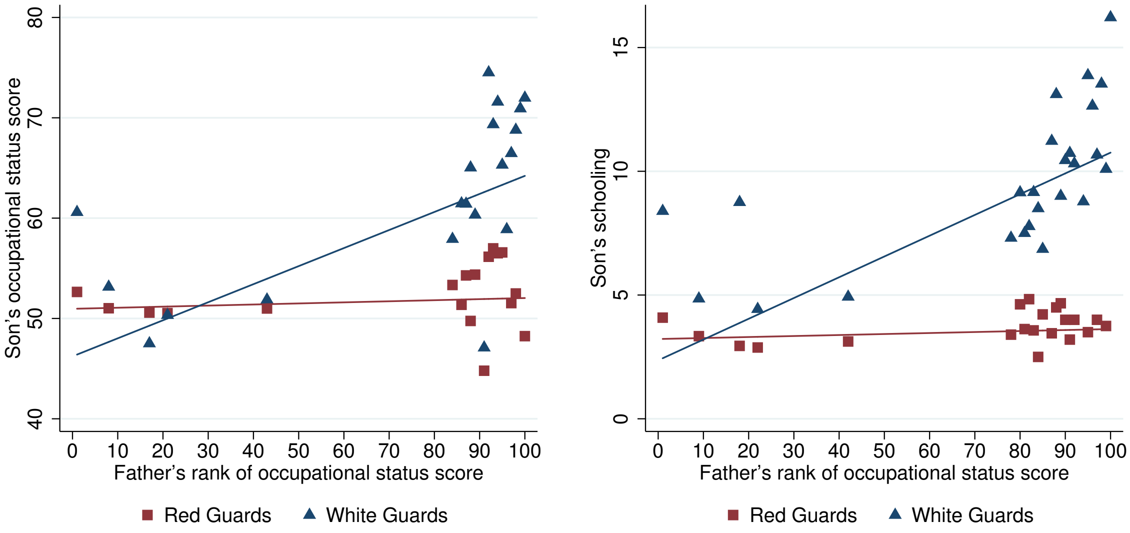
We can therefore directly compare the socioeconomic background and intergenerational mobility of Red Guard and White Guard. Not surprisingly, those who joined the Red Guard tend to come from families with lower socioeconomic status than those who joined the White Guard. However, the gap is not as large 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.

Red vs. White Guard: 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 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. Members of the White Guard follow a standard pattern of modest

[32] 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.

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



(a) Son's Occupational Status

(b) Son's Schooling

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.

intergenerational persistence, with sons of academics and professionals becoming on average academics and professionals and sons of skilled workers becoming on average skilled workers. As we show in more details below, the degree of intergenerational mobility in occupational status appears similar as for other populations in the early 20th century.

Intergenerational and in particular *downward* intergenerational mobility, however, is remarkably high for members of the Red Guard. The gap to the White Guard is particularly striking among Reds from high socioeconomic 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 of schooling at the 20th percentile compared to more than 9 years at the 80th percentile. Intergenerational mobility appears therefore as good as perfect among the members of the Red Guard, with only a negligible association between the mean status of sons and their fathers. As we discuss below, this pattern may suggest that selection into the Red and White Guard may depend on intergenerational mobility itself.

To quantify these pattern in more detail, the first panel of Table 11 reports results from direct intergenerational regressions of son's years of schooling and occupational status score (measured in 1918) on father's occupational status 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 if 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.

Red vs. White Guard: Name-based mobility estimates

The last two panels of Table 11 present estimates from the name-based methods, starting with the R^2 estimators based on first names or surnames. The pattern by and large confirms our results from the conventional (direct) intergenerational regressions. 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. As the ICS, 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.

Cross-sectional estimators such as the ICS have been primarily used for comparative purposes, but it is interesting to note that in this application they capture 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 coefficients, this finding suggests that it may provide a good approximation in extreme cases such as the one considered here.

The last panel of Table 11 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.[33] Because 14 percent of surnames are unique to members of the Red Guard, the sample size is slightly lower for surnames than

[33] 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 11: Intergenerational Mobility of White and Red Guard

	White Guards		Red Guards	
	(1)	(2)	(3)	(4)
	Son's schooling	Son's occupational score	Son's schooling	Son's occupational score
<u>Direct estimator</u>				
Father's occupational status (BR)	0.212	0.453	0.006	0.003
	(0.016)	(0.060)	(0.008)	(0.056)
	N=906	N=1,010	N=616	N=681
<u>R^2 estimator</u>				
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
First names	0.097	0.070	0.016	-0.001
	[0.075, 0.117]	[0.051, 0.089]	[0.000, 0.031]	[-0.016, 0.006]
	N=7,032	N=8,294	N=5,821	N=6,469
<u>Grouping estimator</u>				
Surnames	0.088	0.241	0.026	-0.003
	(0.011)	(0.040)	(0.005)	(0.020)
	N=3,263	N=3,864	N=3,370	N=3,890
First names	0.185	0.460	0.026	0.093
	(0.014)	(0.045)	(0.005)	(0.023)
	N=5,165	N=5,831	N=5,652	N=6,303

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 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 of the in parentheses or 95% confidence intervals across 1,000 bootstrap samples in brackets. Source: The Finnish Longitudinal Veteran Database.

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 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 11 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.[34] One concern is that the attrition of Red Guard who did 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.[35] 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.

Interpretation

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 occupational scores 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 previous work on the relation between political orientations and social mobility.

Although 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), ideology typically plays an important role. It is outside

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

[35] 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.

the scope of this study to provide an explanation for the factors that triggered the Civil war, or which side an individual chose to join. As mentioned before, according to previous work the main causes were the general unrest in the aftermath of World War I and the Russian revolution of 1917, polarization of land ownership as a side-effect of the transformation from an agrarian towards an industrialized society and financial distress of factory workers due to the economic recession. However, previous work suggests that preferences for redistribution also depend on knowledge about the socioeconomic status of previous generations, the current socioeconomic status, and expectations about future socioeconomic status.

For example, Piketty (1995) argues that individuals may infer the relative role of effort in the determination of socioeconomic outcomes by looking at familial socioeconomic trajectories, generating between-family variation in beliefs about intergenerational mobility and consequentially in beliefs about the socially optimal redistribution rate. In steady state, a society may be characterized by a stable majority of left-wing voters in the lower class (as compared to a stable minority in the upper class), and intermediate levels of left-wing voters in off-diagonal cells in the two-by-two mobility table representing mobility across adjacent generations (i.e., upward and downward mobility). 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, 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 own experiences, rationalizing differential behavior among the immobile and downwardly mobile poor.[36]

These studies provide therefore a theoretical reason as to why the rate of intergenerational mobility may differ systematically between political groups. Our evidence suggests that these differences can indeed be substantial, and that *downward* intergenerational mobility may contribute to left-wing political action. To illustrate this further, Table (12) 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≥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

[36] This work also relates to work on the relation between social mobility and political behavior in sociology and political sciences (e.g. Weakliem, 1992). An important hypothesis from that literature is that the upwardly mobile adopt the political orientation of their destination comparatively quickly, while the downwardly mobile tend to retain the behavior typical of their origin.

Table 12: 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.

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).

7.2 Regional Variation in Mobility

A higher bar for the name-based estimators is to document regional variation in intergenerational mobility. The literature on regional differences in mobility levels and trends has burgeoned rapidly in recent years (e.g. Chetty et al., 2014, Güell et al., 2018, Mazumder and Davis, 2018). However, sample size becomes a bottleneck when splitting data into regions, and researchers rarely have access to rich panel data with direct family links in order to estimate direct intergenerational models across regions. Name-based methods are attractive in this context, as they only require on cross-sectional data – and large-scale cross-sectional sources are available for most countries (such as population or electoral censuses). But while attractive from a data perspective, it is not obvious if name-based methods will work in this context. Both observed and unobserved sociodemographic characteristics vary across regions, and as we have shown (Table 4) the ICS is susceptible to omitted variable bias in the name regressions. Further, Solon (2018) points out the pitfalls of excluding potential group-level confounders from grouping regressions.

Table 13: Correlation of Name Shares and Socioeconomic Status Across Regions

	Surnames		First Names	
	Actual	Placebo	Actual	Placebo
Name Shares	0.37	0.57	0.83	0.96
Father's Hiscam	0.03	0.25	0.45	0.51
Son's schooling	0.13	0.19	0.45	0.50

Note: The table reports the pairwise correlations in the indicated variables across three regions. Pairwise correlations weighted by name size and averaged across all region combinations.

We therefore test if name-based methods can capture the regional pattern in intergenerational correlations. We restrict this analysis to the White Guard subsample due to the availability of self-reported father's occupation for this subgroup. We consider a flexible definition of regions and cluster individuals into different regional aggregation levels based on the geocoded place of birth, varying from splitting the country into a minimum of two regions to a maximum of 15 regions. We first show evidence based on three regions defined by the observed level of parent-child mobility.

We begin by exploring the distribution of names and their informational content across regions. The first row of Table 13 shows regional variation in the share of first and surnames in the early 1900s. To distinguish regional differences from sampling variation we also report the corresponding placebo correlation, in which the regional dummies are allocated randomly across individuals (while keeping their marginal distribution constant). The frequency of surnames in our sample differs quite systematically across regions, consistent with the idea that surnames can be specific to a particular city or area (Barone and Mocetti, 2016; Güell et al., 2018). In contrast, first names are distributed much more uniformly, with the actual share of first names being nearly as strongly correlated across regions as across the placebo definition of regions. The second and third row of Table 13 show a similar contrast in the socioeconomic content of names. While the region x surname average of father's occupational status score or son's schooling is only weakly correlated across regions, the region x first name averages are strongly correlated. In sum, both the frequency and socioeconomic prestige of surnames varies regionally, while the frequency and prestige of first names hardly varies across regions.

[to be extended]

8 Conclusions [work in progress]

We reviewed name-based methods from a conceptual and empirical perspective, based on newly digitized data from Finland that records names and name changes ("name mutations"), but also contains direct family links. These data, combined with linked U.S. Census data used in prior studies, allowed us to provide direct evidence on the performance of different name-based estimators, and to provide a more comprehensive account of their properties than what has been available so far. 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. Average socioeconomic status may vary with name frequency, as may the intergenerational process. Because the informational content of first names varies less with name frequency than the informational content of surnames, estimators based on first names will be less sensitive to this issue than those based on last names. Irrespective of the preferred 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. Intergenerational persistence reflects multiple transmission mechanisms, including some that play out at an aggregate level, such as those related to regions or ethnicity. 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. However, this concern can be partly addressed by testing how sensitive estimates are to the inclusion of controls such as location and ethnicity, or other variables that (i) relate to group-level processes and that (ii) correlate with names. Some name-based studies have implemented such tests, but the underlying issue extends to all name-based estimators. We therefore propose that applications based on name-based estimators should always contain evidence on the stability of the results to the inclusion of those group-level controls that vary particularly strongly across names groups.

Third, names have *added informative content* beyond serving as a proxy for a particular socioeconomic status variable for parents. Name-based studies are often not explicit on this fact, or have different takes on its interpretation. For some authors, the idea that name-based estimators capture more than the conventional parent-child estimates is their principal attraction, as it may suggest that the former capture something more fundamental than the latter (e.g. Clark, 2014). That is, some authors assign a very particular interpretation to the fact that names have added informative content. Others use the name-based estimators simply as a feasible "drop-in" replacement for settings in which the direct estimator is infeasible. From this perspective, the observation that names have added informative content is a nuisance, and is often left unexplored. While its interpretation is up for debate, the question *if* names have added informative 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 informative content would alter the interpretation of the name-based estimates.

Fourth, the direct and grouping estimators capture systematically different objects. It is sometimes assumed that because names have added informative content, the estimates from the grouping estimator are necessarily larger than those from 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 than the direct estimator if names have added informative content and 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 sampling properties of the 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 in name-based estimates , while others do not. We therefore propose that researchers study and report the sampling properties of their data explicitly, and discuss how these properties affect the interpretation of their estimates. In settings where this is possible, it would also be useful to report and compare the size of both the "*inclusive*" and "*leave-out*" variants of the grouping estimator, and to explicitly estimate the attenuation bias in the "leave-out" version that arises from the imperfect measurement of group means.

Finally, a number of assumptions could be made more explicit. For example, name-based estimators are typically motivated by the argument that names have informational content. But that is not always necessary – while the R^2 estimator is directly based on this argument, the grouping estimator does not require that names have informative content if the parent and child samples overlap. In fact, we note that in such settings, its interpretation may greatly simplify if names do not have informational content. We propose that researchers estimate the informational content of names in their application (e.g. by reporting the R2 estimator), even when their analysis is otherwise based on the grouping estimator, and explain to what extent the former affects the latter in their setting (which depends on sampling properties, see above).

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. An interesting question is why mobility differed so markedly between the two groups. Our findings relate to previous studies that document breaks in mobility patterns in populations around the turn of the twentieth century that were exposed to similar shocks, and to work on the interaction between intergenerational processes and political preferences more generally. In particular, our findings suggest that the experience of downward intergenerational mobility may be one factor distinguishing right and left-wing political groups.

The application illustrates that name-based estimators can be very useful in practice, despite the many conceptual and interpretational issues that we highlight in this study. We however propose that the the different types of estimators should be described and implemented more consistently, their underlying assumptions clarified, and their robustness to sampling variations and the inclusion of control variables more systematically tested.

References

ABRAMITZKY, R., L. P. BOUSTAN, AND K. ERIKSSON (2016): "Cultural Assimilation during the Age of Mass Migration," NBER Working Papers 22381, National Bureau of Economic Research, Inc.

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.

ADERMON, A., M. LINDAHL, AND M. PALME (2019): "Dynastic Human Capital, Inequality and Intergenerational Mobility," Discussion paper, mimeo, Uppsala University.

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 (2016): "Intergenerational Mobility in the Very Long Run: Florence 1427-2011," Temi di discussione (Economic working papers) 1060, Bank of Italy, Economic Research and International Relations Area.

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. (2012): "What is the True Rate of Social Mobility in Sweden? A Surname Analysis, 1700-2012," Discussion paper, mimeo.

——— (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 (2012a): "Are there Ruling Classes? Surnames and Social Mobility in England, 1800-2011," Discussion paper.

——— (2012b): "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-Ortin, and A. Romeu (2014): "Long-run Intergenerational Social Mobility and the Distribution of Surnames," Discussion Paper 2, UMUFAE Economics Working Papers.

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.

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

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 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.

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.

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.

Mazumder, B., and J. Davis (2018): "Racial and Ethnic Differences in the Geography of Intergenerational Mobility," Discussion paper, mimeo.

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

Modalsli, J. (2015): "Intergenerational Mobility in Norway, 1865-2011," Discussion papers, Statistics Norway, Research Department.

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.

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.

Ruggles, S., C. Fitch, and M. Sobek (2017): "Building a National Longitudinal Research Infrastructure," Discussion paper, MPC Working Paper No. 2017-2.

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

Torche, F. (2015): "Analyses of Intergenerational Mobility: An Interdisciplinary Review," *The Annals of the American Academy.*

Torche, F., and A. Corvalan (2015): "Estimating Intergenerational Mobility with Grouped Data: A Critique of Clark's The Son Also Rises," Discussion Paper Working Paper No. 2015-22, NYU Population Center Working Paper Series.

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.

Weakliem, D. L. (1992): "Does social mobility affect political behaviour?," *European Sociological Review*, 8(2), 153–165.

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.[37]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-

[37] 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 of the two data entities 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.

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. 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	0.001	0.001	0.004	0.001
	(0.002)	(0.002)	(0.003)	(0.003)
HISCAM score in 1918	-0.001	-0.000	-0.001	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)
Father's HISCAM	0.000	0.000		
	(0.001)	(0.001)		
Surname count	0.001	0.002**	-0.001***	-0.000
	(0.001)	(0.001)	(0.000)	(0.000)
Region 1		0.000		0.000
		(.)		(.)
Region 2		0.029		0.065***
		(0.015)		(0.018)
Region 3		0.471***		0.452***
		(0.023)		(0.027)
Region 4		0.088***		0.111***
		(0.022)		(0.013)
Region 5		0.02		0.051***
		(0.014)		(0.006)
Region 6		0.022		0.052***
		(0.016)		(0.008)
Region 7		0.019		-0.000
		(0.014)		(0.004)
Region 8		0.015		0.126***
		(0.013)		(0.010)
Region 9		0.077***		0.066***
		(0.014)		(0.017)
Region 10		0.302***		0.322***
		(0.029)		(0.052)
ar2	0.001	0.215	0.003	0.100
N	4,380	4,380	5,735	5,735

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 whether the surname was Finnish. 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)	(2)	(3)	(4)
	Son's schooling	Son's occupational score	Son's schooling	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 occupational status			
	Son's Schooling		in 1918		in 1930s	
	(1)	(2)	(3)	(4)	(5)	(6)
Father's occupational status (S)	0.221		0.529		0.679	
	(0.013)		(0.038)		(0.045)	
Father's occupational status (BR)		0.244		0.534		0.667
		(0.015)		(0.046)		(0.057)
ar2	0.293	0.281	0.199	0.160	0.216	0.145
N	735	735	780	780	817	817

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 Guards		Red Guards	
	(1)	(2)	(3)	(4)
	Son's schooling	Son's occupational score	Son's schooling	Son's occupational score
Direct estimator				
Father's occupational status (BR)	0.203	0.397	0.006	0.003
	(0.029)	(0.092)	(0.008)	(0.056)
	N=218	N=225	N=616	N=681
R^2 estimator				
Surnames	0.249	0.182	0.070	0.037
	[0.103, 0.489]	[-0.024, 0.392]	[0.026, 0.121]	[-0.001, 0.083]
	N=1,552	N=1,545	N=5,821	N=6,469
First names	0.102	0.145	0.016	-0.001
	[0.050, 0.147]	[0.073, 0.210]	[0.000, 0.031]	[-0.016, 0.006]
	N=1,552	N=1,545	N=5,821	N=6,469
Grouping estimator				
Surnames	0.092	0.074	0.002	-0.003
	(0.026)	(0.080)	(0.004)	(0.020)
	N=434	N=437	N=3,402	N=3,915
First names	0.130	0.296	0.026	0.087
	(0.021)	(0.090)	(0.005)	(0.024)
	N=1,049	N=1,104	N=5,662	N=6,310

Note: Replication of Table 11, 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.