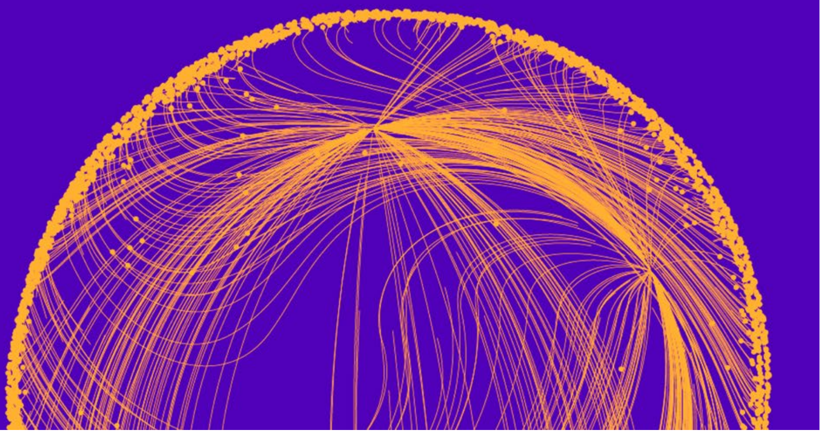
Helsinki GSE

HELSINKI GSE DISCUSSION PAPERS 24 · 2024

Name-Based Estimators of Intergenerational Mobility

Torsten Santavirta
Jan Stuhler



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

HANKEN

Aalto University

Helsinki GSE Discussion Papers

Helsinki GSE Discussion Papers 24 · 2024

Torsten Santavirta and Jan Stuhler:
Name-Based Estimators of Intergenerational Mobility

ISBN 978-952-7543-23-8 (PDF)
ISSN 2954-1492

Helsinki GSE Discussion Papers:
https://www.helsinkigse.fi/discussion-papers

Helsinki Graduate School of Economics
PO BOX 21210
FI-00076 AALTO
FINLAND

Helsinki, April 2024

Name-Based Estimators of Intergenerational Mobility

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

18th April 2024

Abstract

Recent studies use names—first and surnames—to estimate intergenerational mobility in sources that lack direct family links. While generating novel evidence on intergenerational transmission processes, it remains unclear how different estimators compare and how reliable they are. This paper evaluates the most popular name-based methods, using newly digitised records from Finland and U.S. Census data. We illustrate that their interpretation depends on sampling properties of the data, such as the overlap between the parent and child samples, which differ widely across studies. We correct for the attenuation bias from limited overlap and address other common problems encountered in applications.

JEL classification: J62.

[*] University of Helsinki and Institute for Housing and Urban Research (IBF), Uppsala University (email: torsten.santavirta@helsinki.fi)

[†] Universidad Carlos III de Madrid (email: jan.stuhler@uc3m.es)

[‡] We thank Ilkka Jokipii and Virva Liski for excellent data collection and data preparation for this project. We are very grateful to James Feigenbaum, Claudia Olivetti and Daniele Paserman for the provision of samples and replication files to link records in the U.S. Census. We also thank Giorgia Estefani, Maia Güell, Sevi Rodríguez Mora and seminar participants at SOFI at Stockholm University, Uppsala University, Purdue University, Universidad Autonoma de Barcelona, Statistics Norway, UC Davis, the Nuffield Historical Social Mobility seminar, the 2019 SOLE conference, and the SSHA 2022 conference for comments. Support from KONE Foundation and the Ministerio de Ciencia e Innovación (Spain, RYC2019-027614-I, ECO2017-87908-R and IJCI-2016-30011) and Comunidad de Madrid (AM-EPUC3M11 and H2019/HUM-5891) is gratefully acknowledged.

1 Introduction

A recent development in research on intergenerational mobility is the use of *names* to overcome data limitations. Conventional measures, such as the slope coefficient in a regression of the child's socioeconomic status y_i in family i on the parent's status x_i,

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

require linked data across two generations. In the absence of family links, this "*direct*" regression is infeasible, but names—first names and surnames—can serve as a proxy for these links. Based on this insight, researchers have developed different *name-based* estimators of intergenerational mobility, which have become instrumental in several strands of the literature. Examples include recent work on the long-run persistence of inequality across multiple generations (Clark and Cummins 2014, Barone and Mocetti 2020), on trends in intergenerational mobility (Clark 2014, Olivetti and Paserman 2015a, Güell *et al.* 2015) or its pattern across regions (Güell *et al.* 2018a).

While these studies are all motivated by the observation that names contain socioeconomic information, they exploit that information in different ways (Table A.1 provides a non-exhaustive list of recent contributions). Most authors focus on the innovative features of their respective application, while the conceptual similarities that link different methods have received less attention. This diversity complicates the interpretation of name-based estimators and their further development, and masks the degree to which insights and criticisms regarding one method extend to another—or the general approach as such.

We therefore provide a systematic review of name-based estimators of intergenerational mobility in this article.[1] Specifically, we (i) provide an overview of the proposed methods, (ii) evaluate their properties, strengths, and weaknesses, and (iii) describe how the various methods are linked. Our arguments are supported by evidence from U.S. Census data and newly digitised historical data from Finland that contain all the required elements (including direct family links) necessary for comparing name-based and conventional estimators.

We first provide an overview of the various methods that have, to date, been developed. While different authors use different labels, most studies can be categorised within the simple two-by-two diagram shown in Table 1, with name type (*first names* vs. *surnames*) and type of estimator (R^2 vs. *grouping*) on the horizontal and vertical axes.[2]

[1] A lively debate has ensued on the validity and interpretation of specific name-based studies. Recent contributions include Chetty *et al.* (2014), Vosters and Nybom (2017), Torche and Corvalan (2018), Braun and Stuhler (2018), Güell *et al.* (2018a), Solon (2018), Adermon *et al.* (2021), Vosters (2018), Clark (2018), and Choi *et al.* (2018). However, no systematic review has thus far been conducted. Feigenbaum (2018) comes closest in spirit, showing that in a U.S. sample the grouping and direct estimators yield the same qualitative conclusion.

[2] Names have been used in other creative ways, which we do not review here. Some studies com-

Table 1: A Classification of Name-Based Methods

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

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

Starting from the top-right cell, the R^2 *estimator* developed by Güell *et al.* (2007, 2015) considers the joint distribution of names and socioeconomic status in a given generation, thereby circumventing the need to link generations. If both surnames and status are transmitted from one generation to the next, then surnames should explain status in the cross-section. The R^2 of a regression of individual-level outcomes on a set of surname dummies summarises this *informational content of surnames.* Because a high R^2 implies strong status inheritance, the estimator can be used to rank groups or regions by their level of intergenerational mobility.

Most studies however use what is fundamentally the same type of *grouping estimator* (bottom row), which can be implemented as a two-step process. First, the average socioeconomic status within each name group and generation is computed. We then estimate regression (1), replacing the parent's socioeconomic status x_i by the group-level means for their generation. Prominent examples include Clark (2014) and related studies (such as Clark *et al.*, 2015), who find that status regresses only slowly at the surname level, or Barone and Mocetti (2020), who show that intergenerational correlations can persist across *six centuries* (i.e., over the very long run).

Olivetti and Paserman (2015a) develop a grouping estimator based on *first names*, in which the individual's given name serves as a proxy for family background. One major advantage is that first names do not change upon marriage, and therefore remain informative for daughters and maternal lineages. Olivetti and Paserman describe their empirical

pare the representation of rare surnames in elite institutions and the general population. For example, Clark *et al.* (2015) study their relative frequency in historical admissions lists of Oxford and Cambridge Universities. Paik (2014) shows that in Korea, the historical representation of clan lineages in civil service exams is predictive of contemporary educational levels. Other recent applications using a relative-representation estimator include Clark *et al.* (2020), Halder (2020), Jaramillo-Echeverri *et al.* (2021), Álvarez and Jaramillo-Echeverri (2022), Bukowski *et al.* (forthcoming) and Häner and Schaltegger (2022). Collado *et al.* (2012) circumvent a lack of intergenerational data by comparing the spatial distributions of consumption behaviour and surnames. Names can also be used to impute direct links between parents and their offspring, see for example Ferrie (1996) and Abramitzky *et al.* (2021a).

strategy as a two-sample two-stage least squares (TS2SLS) estimator, in which the first stage groups parental status by first name and the second stage regresses child socioeconomic outcomes against their parental group mean. Their study provides evidence on U.S. mobility trends in a previously unexplored period at the turn of the 20th century, while Olivetti *et al.* (2018) extend their approach to track paternal and maternal lineages in a multigenerational context.

Finally, the R^2 estimator proposed by Güell *et al.* (2015) can be adapted to measure the informational content of *first names*. The conceptual motivation given by Güell and co-authors relates to the naming process for surnames (which are *inherited*), not first names (which are *chosen*). Nevertheless, the potential value of first names for mobility research has been demonstrated by Olivetti and Paserman (2015a), and we study the R^2 estimator based on first names in our data. Finally, we show that the various estimators are closely related. The R^2 estimator proposed by Güell *et al.* (2015) is approximately the (adjusted) R^2 from the first stage of the two-stage estimators proposed by Olivetti and Paserman (2015a) for first names or Barone and Mocetti (2020) for surnames—which in turn belong to the same class of *grouping estimators* that directly relate surname averages across generations, such as Clark (2014).

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

This dependency on properties of the underlying data also links the grouping to the R^2 estimator. Specifically, a low informational content of names corresponds to a "weak" first stage in the grouping estimator. Interestingly, this does not pose much of an issue if the parent and child samples overlap fully. In such settings, the grouping corresponds to a standard 2SLS estimator and is biased towards OLS; yet such bias is desirable if the (feasible) grouping estimator is used as a substitute for the (infeasible) OLS estimator. However, if the parent and child samples do not overlap, the grouping estimator instead corresponds to a split-sample IV estimator, and is biased towards zero (Choi *et al.* 2018, Khawand and Lin 2015).

We show that the resulting bias can be large in typical applications, and propose a simple bias correction procedure that accounts for (i) the degree of overlap between the parent and child samples, and (ii) the extent to which the name group means in the parent sample predict the corresponding means in the child sample. This correction procedure could improve the comparability of estimates in the intergenerational literature. Moreover, the degree of overlap between main and auxiliary samples varies also in other contexts, such that the problems that we describe in the intergenerational context extend to other applications using the TS2SLS estimator.

The grouping and R^2 estimators are otherwise subject to similar conceptual issues. First, both are identified primarily from rare names and might therefore not be representative for the population as a whole. The grouping estimator proposed by Clark (2014) has been criticised on these grounds, although the concern is universal to all name-based estimators, including those based on first names (although, to a lesser extent). Second, socioeconomic status tends to decrease with the frequency of names, amplifying such concerns further. Third, the grouping estimator tends to increase in name frequency and sample size, while the R^2 estimator does not. Fourth, name-based estimators weight the underlying transmission mechanisms differently than conventional estimators.

The remainder of this paper is organised as follows. Section 2 reviews the main insights from recent name-based studies. Section 3 introduces the data. Section 4 explores the informational content of first names and surnames and reviews the R^2 estimators. Section 5 assesses the grouping estimators, while Section 6 examines the properties and various conceptual caveats affecting both estimators. Section 7 concludes.

2 Recent Applications

Name-based estimators have opened up promising new research areas by enabling the exploitation of historical and cross-sectional sources that lack direct family links. For illustration, we describe their role in three active strands of the literature, which are changing our understanding of intergenerational processes in a number of key aspects.

First, they are informative about intergenerational mobility in the very long run. Studies such as Clark (2014), Clark and Cummins (2014) or Barone and Mocetti (2020) show that the average socioeconomic status of surnames can be highly persistent across generations, much more so than the status of individual families as captured by conventional estimators based on direct family links. Clark (2014) notes that this observation is consistent with the idea that conventional measures understate intergenerational persistence because they do not capture the transmission of latent characteristics that affect future generations. This interpretation has triggered a lively debate, and some scholars remain decidedly critical towards the grouping estimator itself (see footnote 1). However, recent studies that directly link distant relatives largely confirm that conventional parent-child measures understate the transmission of economic advantages.[3]

[3] See for example Lindahl *et al.* (2015), Braun and Stuhler (2018), Neidhöfer and Stockhausen (2019), Adermon *et al.* (2021) or Collado *et al.* (2023).

Second, name-based studies shed light on the extent of mobility for countries and periods for which intergenerational panels with direct family links are not available. For example, using Census data, Long and Ferrie (2013) and Olivetti and Paserman (2015a) find that while the U.S. may have been characterised by high intergenerational mobility in the 19th century, mobility was lower in the early 20th century. Clark (2014) and others estimate mobility rates for a number of countries and time periods for which few if any other estimates are available. Barone and Mocetti (2020) show that intergenerational mobility in the Italian city of Florence may have been much lower during the 15th century than in modern times, and that socioeconomic differences persist over nearly six centuries. Name-based studies therefore expand our knowledge about how intergenerational processes vary across time and countries.

Third, name-based methods can help characterising the geography of intergenerational processes in greater detail. Following Chetty *et al.* (2014), a number of studies compare mobility rates across regions within countries, based on large-scale administrative data. This regional evidence is interesting from a descriptive perspective, but also opens the door for causal research designs exploiting regional differences. Unfortunately, the type and quality of administrative data used in these studies is not available for most countries. Name-based methods allow researchers to use more standard data sources, such as Census data. For example, Güell *et al.* (2018a) employ the R^2 estimator in cross-sectional data to study how mobility rates differ between Italian provinces.

3 Data

We use historical data sources from Finland and the United States to compare name-based estimators in the type of setting for which they were designed. Our main source are longitudinal records from the turn of the 19th century and the 20th century in Finland, which are well suited for our purposes. First, they include socioeconomic outcomes for individuals of two 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 name-based methods.

Finnish Longitudinal Veteran Database. We assembled our main sample by digitising and linking various individual-level data sources from the National Archives of Finland on veterans of the Finnish Civil War, which was fought in 1918 between the socialist Red Guards and the conservative White Army (also commonly known as the *White Guards*; see Upton, 1980). Our dataset, named the *Finnish Longitudinal Veteran Database* (Santavirta and Stuhler, 2024), contains 16,212 individuals born between 1865 and 1904 who survived the Civil War. It includes information on first names, surnames, schooling, occupation, father's occupation, demographic characteristics, and the side on which the individual fought in the war. We focus on the first of up to three given names, henceforth, first name.[4] Surnames are cleaned from obvious spelling mistakes. After dropping all females and males with missing occupation our analytic sample contains 14,734 individuals, of which 6,452 fought in the Red Guard and 8,282 fought in the White Guard. We observe father's occupation for 7,012 father-son pairs through the son's self-report of his father's occupation. We probe the quality of the self-reported father's occupation by matching occupational information from digitised genealogy records.[5] Section B in the Online Appendix describes this matching process and the individual registries from which the variables were acquired.

Table 2 reports summary statistics, separately for veterans of the White Guard and Red Guard. We use two quantitative measures of socioeconomic status: occupational status and years of schooling. We observe occupation as of 1918, at the time of enrolment 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 developed by Miles *et al.* (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 standardised to have a mean of 50 and a standard deviation (std. dev.) of 15 in a nationally representative sample of individuals. In our full sample (n=14,734), the HISCAM score based on occupation in 1918 has a mean of 51.3 (std. dev. 10.8). The HISCAM score as of the 1930s was only recorded for members of the White Guard (n=8,680) and has a mean of 59.6 (std. dev. 15.8). Schooling is coded as number of completed years of schooling based on the highest completed level of education.[7] The first name distribution is more compressed among the Red Guard, with 76.5% of the individuals having a first name that ranks within the 50

[4] We further harmonised 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.

[5] We matched our sample to digitised birth certificates from www.ancestry.com that are maintained by the Genealogical Society of Finland (http://hiski.genealogia.fi/hiski/93id4x?en) using a matching algorithm developed specifically for this purpose by Eric Malmi. We can impute father's occupation for a total of 2,506 veterans from these sources.

[6] Individuals who are socially close to one another are more likely to interact and form marriages than individuals who are socially far apart (http://www.camsis.stir.ac.uk/index.html).

[7] Each educational category, e.g. compulsory schooling, was coded according to 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, so incomplete schooling careers are observed.

Table 2: Summary Statistics

	Red Guards	White Guards
Number of sons (N)	6,610	9,602
Linked fathers (self-reported)	–	7,012
Linked fathers (birth records)	1,096	1,389
First names		
Number of distinct names	400	583
Mean frequency per name	16.5	16.5
Sons with singleton first name	2.2%	2.2%
Top-50 names	76.5%	67.0%
Surnames		
Number of distinct names	2,852	4,394
Mean frequency per name	2.3	2.2
Sons with singleton surname	30.5%	29.7%
Top-50 surnames	26.6%	12.3%
Socioeconomic Outcomes		
Son's years of schooling (N)	5,803	7,021
mean (std. dev.)	3.24 (1.56)	6.77 (4.85)
Son's occupational status 1918 (N)	6,452	8,282
mean (std. dev.)	47.76 (7.58)	54.01 (12.07)
Son's occupational status 1930s (N)		8,680
mean (std. dev.)		59.64 (15.83)
Father's occupational status (N)		7,012
mean (std. dev.)		55.01 (12.14)
Father's occupational status BR (N)	1,096	1,389
mean (std. dev.)	47.81 (5.30)	52.99 (10.18)

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

most popular names compared to 67% among the White Guard. Rare surnames are more common than rare first names, and roughly 30% of individuals have a unique surname. As for first names, the surname distribution is more compressed among the Red Guard veterans, with 26.6% of all individuals having a top-50 ranked surname as compared to 12.3% among the White Guard. The difference in the first name and surname distributions is illustrated further in Figure E.1 in the Online Appendix, which shows that surnames have a right-skewed distribution while first names do not.

Other samples. We replicate our key results in linked records from the U.S. Census as used in other recent studies. 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 and the occupational mean income of fathers and their sons. Olivetti and Paserman (2015a) provide replication files to reconstruct their samples (Olivetti and Paserman, 2015b), 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 occupational income renders a sample size of 9,076 observations. Finally, we use the digitised Iowa State Census 1915 Sample (Goldin and Katz, 2000) linked to the 1940 U.S. Federal Census by Feigenbaum (2018) and restricted to father-son pairs in which the son was aged 3-17 in 1915, resulting in 3,204 father-son pairs with non-missing occupational income.[8] For comparability we use the same 1950-based occupational income score (calculated by IPUMS from a 1956 Census Report) as used by Olivetti and Paserman (2015a) and Feigenbaum (2018). To test the robustness of our results to different income definitions, we follow Collins and Wanamaker (2022a) and also consider occupational income scores based on the complete count of the 1940 Census.[9]

4 The Informational Content of Names

Most name-based mobility studies start from the observation that names predict status. Both first names and surnames are informative, though for different reasons. The informational content of surnames stems from a mechanical process—children *inherit* their surname, along with other factors that influence their socioeconomic prospects. In contrast, the informational content of first names results from parental *choices*, which correlate with status. However, as those choices may be intertwined with the mobility process itself, more detailed arguments are necessary to justify the use of first names in mobility research (Olivetti and Paserman, 2015a). Surnames have therefore been the more popular option in economics (see Table 1) and other fields (Collado *et al.*, 2008). It is however not obvious whether first or surnames are more useful. First, the informational content of surnames also depends on choice, albeit less directly. Güell *et al.* (2015) note that name *mutations*—be they intentional or accidental—are essential for surnames to retain their informational content, which would otherwise collapse into a small number of frequent and uninformative names such as “Smith” and “Jones”. Second, while deliberate name choice may be intertwined with the mobility process itself, it is attractive from a predictive perspective: while the predictive power of common surnames tends to be negligible, first names can remain informative irrespective of their frequency.

[8] We thank James Feigenbaum for sharing his code and replication files.

[9] The 1940 income variable excludes income from self-employment, which leads to measurement concerns for farmers. We compute farmers’ income using a method developed by Collins and Wanamaker (2022a). We thank the authors for sharing replication files (Collins and Wanamaker, 2022b), using data from IPUMS (Ruggles *et al.*, 2015).

4.1 The R^2 Estimator

Most studies estimate intergenerational regressions on the name-group level, in which the informational content of names plays only an implicit role (see Section 5). However, Güell *et al.* (2007, 2015) show that researchers can make inference about mobility without running a single intergenerational regression, by quantifying the informational content of surnames in cross-sectional data. Intuitively, if socioeconomic status is strongly transmitted, then surnames should explain a large share of its variance—the R^2 in a regression of status on name dummies is increasing in the degree of intergenerational transmission. Specifically, Güell et al. regress outcome y_{ij} of individual i with (sur)name j on a set of name dummies (our notation here may refer to surnames or first names),

$$y_{ij} = \beta' name_j + \gamma' X_{ij} + \varepsilon_{ij}. \tag{2}$$

The vector X_{ij} may include demographic characteristics such as region of birth, year of birth or ethnicity. To estimate the true *incremental* information that surnames carry, the R^2 from this regression is contrasted against a placebo R^2_P from an otherwise identical regression, in which surnames are reshuffled across individuals (while maintaining their marginal distribution). Güell *et al.* (2015) define the *informational content of surnames (ICS)* as the difference between the true and the placebo R^2,

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

While not directly comparable with conventional measures, Güell et al. argue that the ICS is monotonically increasing in the degree of status persistence on the individual level, and can therefore be used to compare intergenerational mobility across time, regions or groups. Indeed, they show that the ICS evolves similarly over time as a more conventional sibling correlation in their data.[10]

The rationale for why surnames contain socioeconomic information is straightforward: surnames are passed on from parents to the next generation, along with other characteristics that affect economic status. The insight that disruptions in this process help surnames to retain their socioeconomic content is perhaps less intuitive. If surnames were transmitted perfectly they would tend to loose their informational content over time (unless status was also transmitted perfectly). The surname distribution is however almost universally skewed, with a large share of surnames being held by few individuals and conversely, a small share of surnames held by a large share of individuals. This skewness is generated by a birth-death process through which some surnames become extinct (e.g., because families fail to reproduce on the male lineage) and new ones are being created by migration or

[10] See also Acciari *et al.* (2022), who show that direct estimates of intergenerational mobility from Italian tax data correlate with name-based estimates presented in Güell *et al.* (2018a) across regions.

name mutations (see Section 6.7).

Table 3: The Informational Content of Surnames and First Names

	Dependent variable: Son's occupational status						
	ICS			ICF			IC
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Name dummies	Surname	Surname	Surname	First	First	First	First+Surn
Demographic controls		Yes	Yes		Yes	Yes	Yes
Region of birth			Yes			Yes	Yes
Panel A: Finnish Longitudinal Veteran Database (N=14,754)							
Adjusted R-squared	0.205	0.292	0.307	0.065	0.197	0.215	0.330
Implied informational content	0.205	0.163	0.147	0.065	0.069	0.056	0.169
Bootstrapped 95% CI	[0.162, 0.248]	[0.128, 0.199]	[0.113, 0.181]	[0.046, 0.091]	[0.053, 0.086]	[0.041, 0.071]	[0.139, 0.213]
Panel B: IPUMS Linked Representative Sample 1880-1900 (N=9,076)							
Adjusted R-squared	0.126	0.175	0.223	0.046	0.092	0.146	0.275
Implied informational content	0.083	0.079	0.069	0.037	0.028	0.022	0.108
Bootstrapped 95% CI	[0.038, 0.131]	[0.039, 0.129]	[0.039, 0.125]	[0.017, 0.057]	[0.009, 0.046]	[0.005, 0.043]	[0.066, 0.182]
Panel C: Linked 1915 Iowa State Census Sample (N=3,841)							
Adjusted R-squared	0.171	0.170	0.192	0.013	0.021	0.094	0.193
Implied informational content	0.171	0.160	0.100	0.013	0.013	0.003	0.101
Bootstrapped 95% CI	[0.102, 0.246]	[0.094, 0.238]	[0.037, 0.175]	[-0.025, 0.049]	[-0.024, 0.050]	[-0.030, 0.040]	[0.011, 0.205]

Note: Regression of son's occupational HISCAM score (Panel A) or log occupational income (Panels B and C) on a set of name dummies and control variables. The implied informational content for first names (columns 1-3), surnames (columns 4-6) or first and surnames combined (column 7) is the difference between the adjusted R-squared from this and an otherwise identical regression in which the name dummies are randomly reshuffled. Panel A reports estimates from the Finnish Longitudinal Veteran Database. Demographic controls include dummies for ethnicity (Finnish-sounding name), White Guard (reference category: Red Guard), and region and year of birth. Panel B reports estimates from the IPUMS Linked Representative Sample 1880-1900 of the U.S. Census (Olivetti and Paserman, 2015). Controls include dummies for immigrant status, county (or state) of residence in 1880 and year of birth. Panel C reports estimates from the Linked 1915 Iowa State Census Sample (Goldin and Katz, 2000) linked to the 1940 US Federal Census (Feigenbaum, 2018). Controls include dummies for immigrant status, county of residence in 1915 and year of birth. 95% confidence interval across 1,000 bootstrap samples in brackets.

The key advantages of the R^2 estimator are its modest data requirements and its broad applicability. First, the estimator requires only cross-sectional data rather than the type of linked intergenerational panels that conventional estimators require. Even a single cross-section may be sufficient for the estimation of mobility trends. For example, Güell *et al.* (2015) study the evolution of intergenerational mobility and assortative mating in Catalonia using a single Census. Second, the R^2 estimator is less sensitive to sampling properties than its main alternative, the grouping estimator. In particular, it is much less sensitive to sample size (see Section 6.4). Güell *et al.* (2015) do not address sampling uncertainty, as their analysis is based on complete Census counts. To adapt the estimator to settings in which only part of the population is observed, we make two adjustments. First, we reshuffle the name dummies repeatedly and report the *mean* ICS across many placebo draws. Second, we implement a bootstrap procedure that draws clusters of observations on the name level to estimate confidence intervals (see Section 6.4 and Online Appendix E.2).

4.2 The Informational Content of Surnames

We report the R^2 estimates of the informational content of surnames in explaining occupational status in the first three columns in Table 3. In the Finnish Longitudinal Veteran Database (Panel A), the estimated ICS is 20.6% in a regression without controls, but falls to 14.7% when controlling for county of birth and ethnicity, as proxied by a dummy for a Finnish-sounding surname.[11] The informational content of surnames is therefore partially due to the fact that names differ systematically across regions and ethnicities. As a qualitative result this is not a concern, as intergenerational persistence on the individual level likewise reflect regional and ethnic factors. The concern is however that name-based estimators weight these factors more heavily than conventional estimators. Practitioners should therefore follow Güell *et al.* (2015) and study whether findings are robust to the inclusion of group-level control variables that may vary systematically across names (we return to these considerations in Subsection 6.6). Surnames still retain substantial explanatory power when abstracting from such factors. Panels B and C in Table 3 report the corresponding estimates from the IPUMS Linked Representative Sample 1880-1900 of the U.S. Census (Olivetti and Paserman, 2015a) and the Linked 1915 Iowa State Census Sample (Feigenbaum, 2018). The ICS in these other samples is smaller and less sensitive to demographic and regional controls.[12] As in the Finnish sample, surnames explain a substantial share of the variation in occupational status.

4.3 The Informational Content of First Names

The R^2 estimator proposed by Güell *et al.* (2015) captures the informational content of surnames. However, first names also carry informational content, as parent's name choice correlates with their socioeconomic characteristics. It is a priori not clear whether first names or surnames are more informative. On the one hand, first names are more selective. As noted by Clark *et al.* (2015), first names carry more information "*[...] because the surname links someone to the status of some distant ancestor, while the first name gives*

[11] We proxy ethnicity by a dummy for a Finnish-sounding surname. 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. We aggregate the individuals' geocoded parishes of birth into 10 synthetic counties by k-medoid clustering.

[12] We replace county with state whenever there is less than ten observations per county, a rule that renders 96 counties/states.

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 Finnish sample.

Following the same empirical strategy as outlined in equations (2) and (3), we report estimates of the *informational content of first names (ICF)* in columns (4) to (6) of Table 3. A non-negligible share of the variation in socioeconomic status across individuals can be explained by their first names. While first names are less informative than surnames in our sample, the two estimators otherwise follow a similar pattern—both decrease substantially when place-of-birth fixed effects are included. When including the full set of controls the ICF in our Finnish sample is estimated to be 5.6%. The ICF is smaller in the two U.S. samples, but remains positive. Finally, in column (7) of Table 3 we report the informative content of first names *and* surnames (IC). In all three samples, first names and surnames together contain more informative content than either separately.

While the mechanisms underlying the ICS and the ICF differ, both are empirically informative. Both variants of the R^2 estimator are therefore promising measures for intergenerational research. However, while the inheritance of surnames follows specific rules that are fairly stable over time or space, the process by which parents choose the first name of their child may fluctuate more strongly. Moreover, all name-based estimators are subject to certain conceptual caveats that warrant attention. We review the R^2 estimator's sensitivity to the frequency of names and several other caveats in Section 6.

5 The Grouping Estimator

A more common approach uses name-group averages to impute the unavailable parental outcomes. The informational content of names motivates the first step of this estimator, which corresponds to the name dummy regression (2) underlying the R^2 method. While they may be differently framed in the literature, we note that all studies use the same type of estimator—a Wald or *grouping* estimator, which groups by either first names or surnames. However, despite using the same estimator, prior works have reported very different estimates, a fact that we aim to rationalise here.

We first link the grouping estimator to the conventional *direct* estimator in equation (1), and show that their relation crucially depends on the sampling properties of the underlying data. In particular, the grouping estimator tends to be larger than the direct estimator if the parent and child samples *overlap*, such that an offspring is sampled whenever his or her parents are sampled as well. In contrast, the grouping estimator can be much smaller than the direct estimator when the two samples do not overlap fully, as in repeated cross-sectional data with partial coverage of the population. Moreover, in overlapping samples, the grouping estimator is not very sensitive to other properties of the data, such as sample size, name frequencies, or the informational content of names. Meanwhile, these properties matter greatly in non-overlapping samples. The fact that the estimator behaves very differently in different settings complicates comparisons of grouping estimates and helps to explain why they are larger than direct estimates in some studies while others find the reverse pattern.

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

5.1 The Grouping Estimator

The grouping estimator appears in various forms in the literature. Clark (2014) and related studies (such as Clark and Cummins, 2014) consider regression to the mean on the *surname* level. In a first step, the average socioeconomic status across individuals within each name and generation is computed. In a second step, the mean status in one generation is regressed on the mean in the previous generation. Others consider rather a two-sample two-stage least squares (TS2SLS) estimator, instrumenting parent's status in equation (1) with a set of first name (Olivetti and Paserman, 2015a) or surname dummies (Barone and Mocetti, 2020). However, a two-stage least squares estimator based on a set of dummy variables is tantamount to running a weighted linear regression on a set of group means, and is therefore also called the "grouping" estimator.[13] The approach by Clark and co-authors is therefore equivalent to the instrumental variable approach used in more recent studies, as long as group means are appropriately weighted.[14] Accordingly, we

[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. Indeed, a weighted regression on group means can be understood as a linear combination of all Wald estimators that can be constructed from pairs of means (Angrist and Pischke 2008).

[14] Since the two approaches produce numerically identical coefficients, it might seem computationally simpler to directly construct the group means for each name rather than running a 2SLS regression with a large set of name instruments. Considering this question in the context of quasi-experimental "examiner" designs, Hull (2017) however notes that the "manual" approach leads to inflated first-stage F-statistics, as it understates the true dimensionality of the underlying instruments—a severe bias from weak instruments may therefore go undetected. This argument is particularly relevant in the intergenerational context in which a large set of name instruments is only weakly predictive of socioeconomic status.

adopt the label *grouping estimator* for either approach. The TS2SLS perspective remains useful, and we return to it below.[15]

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},[16]

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

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

$$y_{ij} = \delta \bar{x}_j + u_{ij}. \tag{5}$$

We argue that the properties of the group coefficient δ depend crucially on if the group mean is defined over the parents of the sampled children, or over *other* individuals who merely share the same name.[17] Specifically, its level and interpretation depends on the *overlap* between the parent and child samples, defined as the probability p that a parent is sampled if his or her child is contained in the child sample, i.e.

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

We begin by considering the two polar opposites: the case of complete overlap and the case of no overlap. Consider first the "*short*" group-level regression

$$y_{ij} = \pi \bar{x}_{ij} + v_{ij}. \tag{6}$$

where $\bar{x}_{ij}$ with subscript i represents the "*inclusive*" mean that averages over the parents of sampled children (including i). Equation (6) is the relevant object if families in the parent and child samples *overlap* completely ($p = 1$), as for example in Chetty *et al.* (2014). The overlap might also be near-perfect if the grouping estimator is applied in complete-count Census data, or data that track families according to some fixed criteria.[18]

This scenario of complete overlap has become more relevant along with the increased availability of historical Census data. For example, the U.S. decennial censuses are made

[15] The TS2SLS estimator (i) takes uncertainty from the first stage into account and (ii) automatically weights name groups by their frequency. However, the TS2SLS perspective also has its pitfalls. As Olivetti *et al.* (2018) note in response to a critique by Choi *et al.* (2018), names are unlikely to be a valid instrument in the sense of satisfying the exclusion restriction. Moreover, we show that the properties of the grouping estimator depend critically on the extent to which the parent and child samples overlap, on which the TS2SLS perspective imposes a polar assumption.

[16] We abstract from the intercept by expressing all variables as deviations from their mean.

[17] Note that the grouping estimates are insensitive to whether we use the individual outcome y_{ij} or the group mean $\bar{y}_j$ as the dependent variable.

[18] Even in complete-count data, the overlap might not be perfect because of deaths, international migration, variation in fertility across families, and so on.

available to the public after 72 years, and complete-count microdata for the 1860-1940 Censuses are provided by IPUMS.[19] This has in turn spurred the development of algorithms that link individuals or families across Censuses (e.g., Ferrie 1996; Abramitzky *et al.* 2021a). However, typically only a fraction of individuals can be linked to their parents, and the match rates are lower for women and ethnic minorities (Bailey *et al.*, 2020; Collins and Wanamaker, 2022a; Helgertz *et al.*, 2022). The grouping estimator, therefore tends to have an advantage in terms of sample size and representativeness and continues to be actively used (see Table A.1). Of course, this supposed advantage depends on what data are used to construct the group means; in some cases, the data are based on a selected sample of the population or a subset of specific names.

In other settings, the parent and child samples might not overlap fully—for a family i in name group j we might observe an ancestor or a descendant but not both. We can approximate such settings by

$$y_{ij} = \kappa \bar{x}_{(i)j} + w_{ij} \tag{7}$$

in which $\bar{x}_{(i)j} = \frac{N_j \bar{x}_{ij} - x_{ij}}{N_j - 1}$ represents the "*leave-out*" mean in a name group of size N_j, in which each descendant's own ancestor is excluded. It corresponds to the predicted value of x_{ij} underlying the jackknife instrumental variables (JIVE) estimator (Kolesár *et al.*, 2015). Equation (7) represents the grouping estimator in settings in which there is zero or only negligible overlap between the parent and child samples ($p = 0$), for example because each sample is a small and independent draw from the population.[20]

Table 4 compares direct and grouping estimates with varying degrees of overlap for our three samples. Column (1) reports the direct estimates based on equation (4), which are $\hat{\beta} = 0.600$ in the Finnish sample, and $\hat{\beta} = 0.474$ and $\hat{\beta} = 0.441$ in the two U.S. samples. The next columns report variants of the grouping estimator, with group means defined over surnames in columns (2)-(4) or first names in columns (5)-(7). For comparability, the grouping estimators are based on the same sample as the direct estimator.

Columns (2) and (5) report grouping estimates based on equation (6) and "*inclusive*" means. They are always larger than the corresponding direct estimates ($\hat{\pi} > \hat{\beta}$). The gap is greater in the Finnish compared to the U.S. data, and greater for first than for surnames. The grouping estimator is however not necessarily larger than the direct estimator, contrary to such suggestions in the prior literature.[21] In columns (4) and (7), we report

[19] Advances in optical scanning techniques and large scale digitising efforts have also made other sources of population records available, e.g., marriage certificates. For example, Craig *et al.* (2021) use both automated linking and name-based methods to link couples identified in the complete count of marriage certificates of Massachusetts in 1850-1910 to childhood and adult Census records of complete count Censuses in the late 1800s.

[20] We approximate such settings with the leave-out mean to reduce changes in sample size, but also verified our results using a split-sample grouping estimator.

[21] See also Olivetti and Paserman (2015a), 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.

Table 4: 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%	66%	0%	100%	66%	0%
Panel A: Finnish Longitudinal Veteran Database							
Father's occupational	0.600	0.640	0.602	0.322	0.763	0.726	0.607
status (HISCAM)	(0.014)	(0.017)	(0.020)	(0.028)	(0.029)	(0.033)	(0.032)
Adjusted R-squared	0.246	0.199	0.181	0.033	0.102	0.096	0.057
N	5,986	5,986	4,043	3,852	5,986	4,456	5,821
Panel B: IPUMS Linked Representative Sample 1880-1900							
Father's log	0.474	0.479	0.453	0.179	0.501	0.432	0.207
occupational income	(0.014)	(0.017)	(0.023)	(0.026)	(0.030)	(0.037)	(0.041)
Adjusted R-squared	0.149	0.103	0.092	0.011	0.038	0.029	0.004
N	9,076	9,076	6,010	5,119	9,076	6,642	8,029
Panel C: Linked 1915 Iowa State Census Sample							
Father's log	0.441	0.446	0.428	0.381	0.533	0.431	0.215
occupational income	(0.021)	(0.024)	(0.029)	(0.032)	(0.037)	(0.045)	(0.057)
Adjusted R-squared	0.141	0.112	0.102	0.073	0.053	0.034	0.006
N	3,204	3,204	2,194	2,317	3,204	2,323	2,781

Note: The table reports the coefficients from a regression of son's occupational HISCAM score (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 clustered at the household level in parentheses.

estimates based on equation (7) and "*leave-out*" means with no overlap between the parent and child sample. These estimates are always smaller, and often much smaller, than the corresponding estimates based on the inclusive mean ($\hat{\kappa} < \hat{\pi}$).[22] They are either greater (Panel A, first names) or smaller than the direct estimates $\hat{\beta}$ (all other cases). The gap between the inclusive and leave-out variants is particularly large in the linked U.S. Census samples (Panel B and Panel C). In the IPUMS Linked Representative Sample 1880-1900, the surname-based grouping estimator is nearly three times larger when constructed from inclusive means ($\hat{\pi} = 0.479$ vs. $\hat{\kappa} = 0.179$).

The inclusive mean $\bar{x}_{ij}$ with full overlap and the leave-out mean $\bar{x}_{(i)j}$ with zero overlap represent the two polar cases. But many applications are based on repeated cross-sections with partial overlap between the parent and child samples. To illustrate such intermediate

[22] The size of this gap depends also on sample size, as we discuss below, and is therefore even larger when using a split-sample IV estimator that splits our sample into separate first- and second-stage samples.

cases, columns (3) and (6) report split-sample grouping estimates based on parent and child samples that overlap by 66%.[23] As expected, the estimates are in between the two polar cases. As shown in Appendix C, these findings hold for alternative outcome variables (earnings or education, Appendix Table C.1) and are robust to specification choice. In Appendix Table C.2 we present grouping estimates using first and surnames jointly, which are slightly lower than using surnames only but follow otherwise the same pattern. Finally, in Appendix Table C.3 we replicate our baseline table using the 1940 rather than the 1950 Census as the basis for constructing occupational scores, implementing also the correction method by Collins and Wanamaker (2022a) for farmers' income. While the level of the estimates changes slightly, the pattern across columns is again similar.

5.2 The Grouping Estimator in Overlapping Samples

To rationalise why the grouping estimator is so sensitive to sampling properties we formalise its relation to the direct estimator. Our arguments resemble arguments from the literature on peer effects, in which grouping estimators have often been misinterpreted (Angrist 2014). For simplicity, the exposition is in terms of population moments.

Consider first the "*short*" group-level regression equation (6), based on the "inclusive" mean $\bar{x}_{ij}$ with full overlap ($p = 1$). As we observe both direct family links and names we can also estimate the corresponding "*long*" regression,

$$y_{ij} = \pi_0 x_{ij} + \pi_1 \bar{x}_{ij} + v_{ij}, \tag{8}$$

which includes both direct (individual) and group (name-level) effects. As we discuss in Section 6.1, many different models could rationalise why names have *added informational content* (i.e., $\pi_1 \neq 0$). Using the omitted variable formula, the relationship between the short and long regression equations can be derived as

$$\pi = \frac{Cov(y_{ij}, \bar{x}_{ij})}{Var(\bar{x}_{ij})} = \frac{Cov(\pi_0 x_{ij} + \pi_1 \bar{x}_{ij}, \bar{x}_{ij})}{Var(\bar{x}_{ij})} = \pi_0 \frac{Cov(x_{ij}, \bar{x}_{ij})}{Var(\bar{x}_{ij})} + \pi_1 = \pi_0 + \pi_1, \tag{9}$$

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

$$\beta = \frac{Cov(y_{ij}, x_{ij})}{Var(x_{ij})} = \frac{Cov(\pi_0 x_{ij} + \pi_1 \bar{x}_{ij} + v_{ij}, x_{ij})}{Var(x_{ij})} = \pi_0 + \pi_1 \frac{Cov(\bar{x}_{ij}, x_{ij})}{Var(x_{ij})}. \tag{10}$$

[23] Beginning with the sample used for the direct estimator (e.g., $N = 9,076$ observations in the IPUMS linked representative sample 1880-1900), we sort the data randomly and draw two partially overlapping sub-samples of size x in such way as to maximise the overlap between them. In this exercise we choose $x = 0.75$ ($N_{sub} = 6,807$ for each generation) such that half of the original sample is overlapping (implying a 66% overlap of the two sub-samples). Since some names are not included in both sub-samples, the effective number of observations as reported in Table (4) is below this theoretical upper bound.

The combination of equations (9) and (10) yields

$$\pi = \pi_0 + \pi_1 = \beta + \pi_1 \left(1 - \frac{Cov(\bar{x}_{ij}, x_{ij})}{Var(x_{ij})}\right), \tag{11}$$

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 own socioeconomic outcome ($\pi_1 > 0$). It cannot be smaller than the direct estimator as long as π_1 is non-negative, a plausible scenario for both first and last names (see Section 6.1). 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.[24] The results reported in Table 4 are therefore not specific to our samples, rather they exemplify a general point: the grouping estimator will tend to be larger than the direct estimator if child and parent samples contain the same families, as in linked U.S. tax (Chetty *et al.*, 2014) or Census data (Ward, 2023) with nearly full overlap between fathers and sons.

The grouping estimator is also linked to the R^2 estimator, as the ratio $\frac{Cov(\bar{x}_{ij},x_{ij})}{Var(x_{ij})}$ in equation (11) is the (population) R^2 in a regression of x_{ij} on a full set of name dummies. Surprisingly, names do not need to have *any* informational content for the inclusive grouping estimator to capture intergenerational mobility: while most studies motivate the grouping estimator with the observation that names do carry systematic information, this condition is in fact not necessary if the parent and child samples overlap. This result relates to the observations that a TS2SLS estimator applied to fully overlapping samples equates to a conventional 2SLS estimator, and that the 2SLS estimator is biased towards the OLS estimator if the instruments are only weak predictors of the regressor of interest (see also Section 5.4). Such bias towards OLS is usually undesirable, but not in the specific context considered here—after all, the TS2SLS is meant to approximate the (infeasible) OLS estimator. Indeed, whenever names have *no* informational content, such that $\pi_1 = 0$, the grouping estimator has the *same* probability limit as the OLS estimator.

According to equation (11), the direct and grouping estimators would also be similar at the other extreme, when names are *very* informative about individual status such that the (population) R^2 defined as $\frac{Cov(\bar{x}_{ij},x_{ij})}{Var(x_{ij})}$ is close to one. However, it is difficult to characterise the relation between the grouping and the R^2 estimators more generally, as this relation depends on the specific transmission model for status and names. For example, Güell *et al.* (2015) show that the R^2 estimator is monotonically increasing in the inheritance

[24] To assign a particular interpretation to the observation that a grouping estimator is larger than the direct estimator is therefore conceptually equivalent to assigning a particular interpretation to the observation that names have added informational content. However, this observation may reflect very different theoretical mechanisms (see Section 6.1).

parameter of a simple autoregressive process; and in their setting the R^2 estimator would also be monotonically increasing in the grouping estimator π (as names have no added informational content if status follows an autoregressive process, so $\pi_1 = 0$ and $\pi = \beta$; see also Section 6.1).

Equation (11) has been similarly derived in Adermon *et al.* (2021). It also underlies a critical review of grouping estimators by Güell *et al.* (2018b), who note that π_1 could vary substantially across studies, and 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, the equation underlying this argument holds only if the parent and child samples overlap completely. Most studies are instead based on partially overlapping samples, and the grouping estimator behaves very differently in such settings (as illustrated in Table 4 and shown formally in the next section). Differences in sampling properties might therefore be another reason why grouping estimates differ so much across studies.

5.3 The Grouping Estimator in Independent Samples

In other applications, the parent and child samples do not overlap ($p = 0$). The statistical properties of the grouping estimator turn out to be very different in such settings. To illustrate, consider the "*short*" group-level regression based on the leave-out mean $\bar{x}_{(i)j}$ in equation (7), and the corresponding "*long*" regression

$$y_{ij} = \kappa_0 x_{ij} + \kappa_1 \bar{x}_{(i)j} + u_{ij}. \tag{12}$$

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

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

and between the direct and the long regression,

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

Finally, combining equations (13) and (14) yields

$$\kappa = \beta + \kappa_1 \left(1 - \frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(x_{ij})}\right) - \kappa_0 \left(1 - \frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}\right) \tag{15}$$

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

Equation (15) characterises the relation between the grouping and the direct estimator when the child and parent samples do not overlap.[25] 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 coefficient on the parent's 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 4)—in contrast to the *inclusive* grouping estimator, which tends to be larger.[26]

Whenever names have no *added* informational content ($\kappa_1 = 0$) the "long" equation (12) collapses into the direct one ($\kappa_0 = \beta$), and equation (15) simplifies to

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

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

Equation (16) further illustrates that the leave-out grouping estimator is closely related to the R^2 estimator: if names have low informational content, the leave-out mean $\bar{x}_{(i)j}$ and parental status x_{ij} will not correlate much, and the attenuation factor $\frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}$ and therefore the grouping estimator κ will be near zero. These implications correspond to the observation that the two-sample grouping estimator applied to non-overlapping samples equates to a split-sample IV estimator, which is biased towards zero when the instruments are weak (Choi *et al.*, 2018).[27] Even if names are very predictive of socioeconomic status, the grouping estimator may still severely understate the direct intergenerational coefficient if the means $\bar{x}_{(i)j}$ are constructed over only few individuals.

[25] Similarly, Olivetti and Paserman (2015a) derive the relation between the grouping and the direct estimator under the assumption that the parent and child samples are independent. As such, our equation (15) corresponds closely to equation (2) in their article. In an earlier draft of their study, Olivetti and Paserman also discussed the distinction between overlapping and independent samples.

[26] In large samples, the inclusive and leave-out mean should be highly correlated, so why would the distinction matter? The two means indeed tend to be highly correlated, even in our modestly sized samples. For example, for first names the correlation is 0.95 in the Finnish and 0.89 in the IPUMS Linked Representative Sample. But while the difference $\bar{x}_{ij} - \bar{x}_{(i)j} = \frac{1}{N_j - 1}(x_{ij} - \bar{x}_{ij})$ becomes small in large name groups, it also becomes increasingly predictive of child outcomes (because the coefficient in the within-name group regression of child outcome y_{ij} on $\bar{x}_{ij} - \bar{x}_{(i)j}$ increases 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 estimates can be quite different in finite samples (Kolesár *et al.*, 2015).

[27] Alternatively, it can be viewed as an ordinary least square (OLS) estimator with a generated regressor that suffers from classical measurement error (Choi *et al.*, 2018).

5.4 The Grouping Estimator in General Settings

The sampling scheme of many applications falls in between the two polar cases of either complete or zero overlap between the parent and child samples. More generally, the name-based grouping estimator corresponds to a TS2SLS estimator in which the main and auxiliary samples are not independent (contrary to standard assumptions, as e.g. in Inoue and Solon, 2010). As shown by Khawand and Lin (2015), in such settings the TS2SLS estimator can be decomposed into a weighted average of the 2SLS and SSIV estimators,

$$\hat{\delta} = \hat{W}\hat{\pi} + \left(1 - \hat{W}\right)\hat{\kappa}, \tag{17}$$

where the weight $\hat{W}$ corresponds, approximately,[28] to the share of individuals in the parent sample whose children are contained in the child sample,

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

In partially overlapping samples, the grouping estimator thus behaves as a weighted average of its *inclusive* and *leave-out* variants as presented earlier. In fully overlapping samples ($p = 1$), the grouping estimator corresponds to a 2SLS estimator and is biased toward the direct (OLS) estimator. In non-overlapping samples ($p = 0$), the grouping estimator corresponds to a SSIV estimator and is instead attenuated towards zero.

The interpretation of grouping estimates depends therefore critically on the overlap p between the parent and child samples. This observation is related to but distinct from the issue of sample attrition from out-migration. Various authors express concern that the exclusion of migrants from their sample may introduce bias, as the socioeconomic mobility of migrants may differ from non-migrants. And indeed, it is easy to show that migration and socioeconomic mobility processes are intertwined.[29] Our argument however implies that migration attenuates the grouping estimator irrespectively of whether migrants are selected, pushing it from the "*inclusive*" (2SLS) towards its "*leave-out*" variant.[30]

Moreover, the overlap p differs substantially between studies. For example, the main estimates in Olivetti and Paserman (2015a) are based on 1-percent excerpts of the Census, so the overlap between the parent and child samples is very small. In contrast, the estim-

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

[29] In supplemental analysis, we restricted the Olivetti and Paserman (2015a) linked representative sample to the 4,803 immobile individuals who did not change county of residence between 1880 and 1900. Both the direct and the inclusive grouping estimates increase by roughly 0.1 (i.e., 20%).

[30] For example, the sample spanning six centuries considered by Barone and Mocetti (2020) excludes descendants who out-migrated from Florence and includes non-descendants who in-migrated from other areas. Our argument here may rationalise why their estimates increase when taking measures to reduce sample attrition from migrants.

ates in Barone and Mocetti (2020) likely correspond to an intermediate level of overlap. Feigenbaum (2018) reports different estimates with varying degrees of overlap, and the name-based estimates in Chetty *et al.* (2014) correspond to fully overlapping samples. Moreover, the argument extends to other settings than the intergenerational context that we focus on here. To illustrate this point, Table C.4 in the Online Appendix lists some recent TS2SLS studies, showing that the overlap tends to vary widely.

5.5 Simulation Evidence

The properties of the grouping estimator depend on many factors, including the (i) size of and overlap between the parent and child samples, (ii) the informational content of names, (iii) the name frequency distribution, and (iv) their *added* informational content. To illustrate their interdependencies, we provide simulation-based evidence in Figure 1.

We consider two data generating processes. Subfigures (a) and (b) are based on an AR(1) process,

$$y_{ij} = bx_{ij} + u_{ij}, \tag{18}$$

which corresponds to a structural interpretation of the standard parent-child regression in equation (1). This process is a natural baseline, in which names play no role in the transmission process and have no *added* informational content ($\pi_1 = \kappa_1 = 0$). Subfigures (c) and (d), are instead based on a latent factor model as considered in Clark (2014) or recent multigenerational studies (Braun and Stuhler, 2018), given by

$$x_{ij} = \rho e^x_{ij} + u^x_{ij} \tag{19}$$

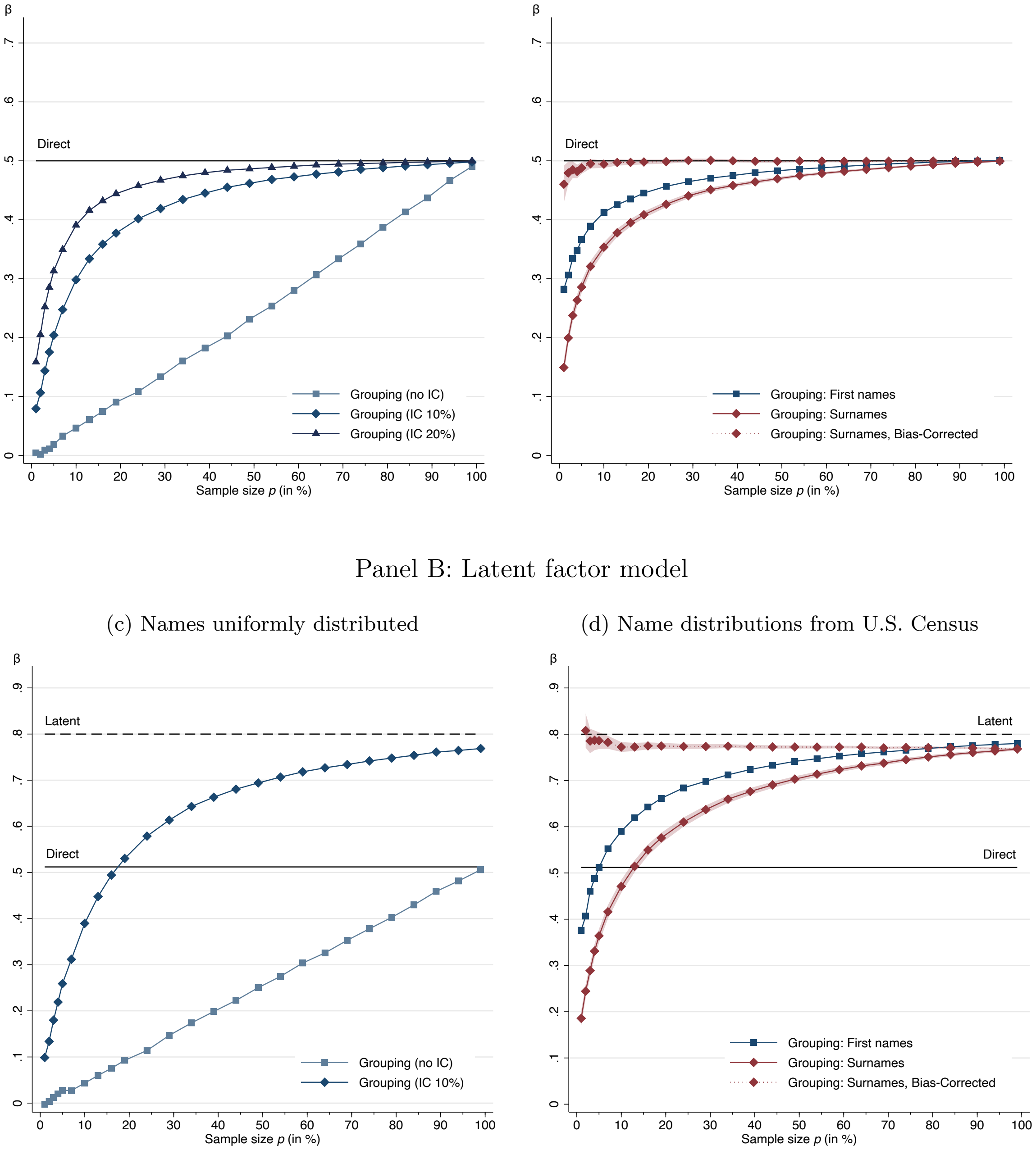
$$y_{ij} = \rho e^y_{ij} + u^y_{ij} \tag{20}$$

$$e^y_{ij} = \lambda e^x_{ij} + v_{ij}, \tag{21}$$

where u and v are white-noise error terms, and e^y and e^x are the latent endowments of child and parent, respectively. The parameters λ and ρ of this model determine the rate of transmission of latent advantages and the signal-to-noise ratio of observed to latent advantages. We set these parameters such that the implied parent-child correlation in y is approximately $\beta \approx 0.5$ (see figure notes). In a first step, we generate parent and child status for the entire population. We then draw sub-samples of size p separately for the parent and child generations, where p also determines the overlap between the parent and child samples. Finally, we estimate the grouping estimator within each sub-sample.

Figure 1a is based on a uniform name frequency distribution. We consider three variants of the AR(1) process in which the parental socioeconomic status x_{ij} is randomly distributed such that names have no informational content (*no IC*), in which name fixed effects explain 10% of the variance in parental status (*IC 10%*), or explain20%

Figure 1: The Grouping Estimator vs. Sampling Probability (Simulation)



Panel A: AR(1) model

(a) Names uniformly distributed

(b) Name distributions from U.S. Census

Panel B: Latent factor model

(c) Names uniformly distributed

(d) Name distributions from U.S. Census

Note: Estimates from differently sized samples (x-axis). Sub-figures (a) and (b) are based on the AR(1) process (18) with slope $b = 0.5$. Parent status x_{ij} contains name fixed effects (→ IC). Sub-figures (c) and (d) are based on a latent factor model given by equations (19)-(21), with x_{ij} and e_{ij} standardised at mean zero and variance one, and $\rho = \lambda = 0.8$ (such that $\beta = \lambda\rho^2 = 0.8^3$). Sub-figures (a) and (c) are based on a simulated name distribution (30,000 names, uniformly distributed frequency between 1 and 250), while sub-figures (b) and (d) are based on the frequency of female first names and male surnames as observed in the 1920 U.S. Census (see Olivetti and Paserman, 2015). Shaded areas represent 95% confidence intervals.

(*IC 20%*). When names have no informational content, the grouping estimator increases linearly in the overlap p between the parent and child samples, fluctuates around zero if there is no overlap ($p = 0$), and converges to the direct estimator under full overlap ($p = 1$)—consistent with the analytic expression for the inclusive and leave-out estimators in equations (11) and (16). Intuitively, the grouping estimator always captures the intergenerational transmission for "complete" parent-child pairs in the sample, even if names are not systematically related to socioeconomic status in the cross-section.

When names have informational content (*IC 10%*), the grouping estimator remains positive even when the parent and child samples have limited overlap; consistent with equation (16), it then also captures transmission among "incomplete" pairs, in which either the parent or the child is sampled, but not both. However, Figure 1a illustrates that the relation between the sampling rate p, the informational content (cf. *IC 10%* and *20%*) and the grouping estimator is highly non-linear. For example, the group-level estimate is still around 0.4 when sampling 10% of each generation but drops to 0.25 when only 5% random samples are taken. The grouping estimator can therefore be (locally) insensitive to sample size in some settings but highly sensitive in others, and we suggest that researchers study the sensitivity of their estimates to sample size (see Section 6.4). Population size matters for similar reasons: in Appendix D.1 we show that the attenuation bias is amplified in smaller populations with fewer individuals per name.

To study how sensitive the grouping estimator is to the marginal distribution of names, we move on to more realistic distributions. Specifically, we consider the distributions of *surnames* and *female first names* in the 1920 U.S. Census (as studied in Olivetti and Paserman, 2015a).[31] Figure 1b plots the grouping estimates as based on first names (blue squares) or surnames (red diamonds) in simulated data from the AR(1) model (with IC 10%). The surname-based grouping estimator is more sensitive to sample size than the one based on first names, as the average frequency is much lower for surnames.[32]

Finally, we switch to the latent factor model as our data generating process, considering again uniformly distributed names (Figure 1c) or the actual name distributions as observed in the U.S. Census (Figure 1d). Names have *added informational content* (AIC) in this model, so the grouping estimator can be larger than the direct (conventional) estimator. The intuition is that the grouping "averages away" idiosyncratic variation in status,

[31] We approximate their distribution in the complete-count census based on the 1% sample. We first draw from a binomial distribution the simulated frequency of a name in the complete-count Census given its observed frequency in the 1% sample. We then use a negative binomial distribution to compute the probability that a name with n observations in the complete-count Census is *not* contained in the 1% sample, and create the missing names accordingly. To verify the plausibility of our simulated distribution we again draw a 1% sample. This simulated 1% sample has a similar name count (12,486 vs. 12,895 first names) and average frequency per name (11.1 vs. 10.4) as the actual 1% sample.

[32] The name frequency distribution does not always have a strong impact on the grouping estimator (see also Olivetti and Paserman, 2013, Section 7.1), but the difference between surname and first name distributions matters in smaller samples.

such that the group means approximate the mean latent status of the respective group (see Clark, 2014, and Section 6.1). Accordingly, in large samples with high overlap, the grouping estimates reflect the persistence in latent advantages ($\lambda = 0.8$ in our simulation). However, in smaller samples with little overlap, the grouping estimator is heavily attenuated and can be substantially below the direct estimator. This sensitivity to sampling properties might explain why some authors have found much smaller grouping estimates than others, and help to reconcile some of the contrasting results in the literature.

5.6 The Bias-Corrected Grouping Estimator

Equation (17) points to a simple way to correct grouping estimates for the attenuation bias in partially or non-overlapping samples.[33] Two objects are required for such correction, the overlap p and the attenuation bias from using *other* parents instead of the parents from sampled children to construct the name group means $\bar{x}_j$. The former follows from the way the samples were constructed, while an estimate of the latter is readily available by regressing parent status x_{ij} on its leave-out mean $\bar{x}_{(i)j}$.

In particular, if names have no *added* informational content ($\pi_1 = \kappa_1 = 0$), combining the probability limit of equation (17) and equations (11) and (15) yields

$$\delta = \beta \underbrace{\left(p + (1-p)\,\frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})} \right)}_{\text{Attenuation Factor}}, \tag{22}$$

where p is the degree of overlap between the parent and child samples, and $\frac{Cov(\bar{x}_{(i)j}, x_{ij})}{Var(\bar{x}_{(i)j})}$ is the slope coefficient in a regression of parental outcomes x_{ij} on their leave-out mean $\bar{x}_{(i)j}$.[34] Dividing a grouping estimator by the term in brackets therefore adjusts for the attenuating effects from limited overlap and finite sample size.

Figure 1b illustrates the performance of this bias-corrected estimator in simulated data based on the 1920 U.S. Census (see previous section). The standard grouping estimator is heavily attenuated when only sub-samples of the population are observed. In contrast, the bias-corrected estimator recovers the intergenerational coefficient based on direct parent-child links. While the confidence intervals are larger for the corrected than for the raw grouping estimates, the mean-squared error is greatly reduced even in small samples. However, the interpretation of the grouping estimator, as well as the proposed bias correction, depend on the underlying data generating process (see Section 6.1). Fig-

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

[34] Considering only observations that contribute to the grouping estimator, i.e., dropping names that feature in the parent but not in the child sample.

ure 1d is based on a latent transmission process, in which names have *added informational content*. In this setting, the correction based on equation (22) on partially overlapping samples yields estimates that are even closer to the true latent rate persistence than the inclusive grouping estimator applied to the full population.

These examples illustrate that irrespectively of which assumptions a researcher invokes on the underlying transmission model, grouping estimates can be corrected for the influence of sampling error and imperfect overlap—which would also improve the comparability of grouping estimates across applications.

6 Properties and Caveats

We have discussed the basic functions and properties of the R^2 and grouping estimators. In this section, we compare how characteristics of the name and sampling distributions affect each estimator (relegating supporting evidence to Online Appendix E). The R^2 and grouping estimators share many conceptual caveats, but differ sharply in some aspects.

6.1 The *Added* Informational Content of Names

The rationale for using names is to proxy for socioeconomic 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.[35] Because our data includes direct family linkages, we can address this question explicitly. Table 5 reports the results from a regression of son's occupational score on the father's own score and the mean score in his surname group (all scores are standardised). If surnames were merely an imprecise proxy for individual status then the coefficient of the group mean should be insignificant.

Instead, names tend to have *added informational content (AIC)*; conditional on a father's own score, the mean score of his name group predicts his son's score. The pattern is most pronounced in the Finnish data (Panel A), where the coefficient on the standardised mean score of father's name group is up to one fifth of the size of his own direct signal. The pattern is also visible in some of the U.S. regressions, regardless of whether the dependent variable is occupational status (Panels B and C in Table 5), earnings, or education (Table E.2 in Online Appendix).

However, this added informational content of names could be generated by different causal processes, leading to different potential interpretations. As it is difficult to distinguish between those alternative processes, we have thus far focused on properties of the grouping estimator that matter irrespectively of the underlying structural model—such as

[35] Most studies use the (feasible) name-based estimators as a second-best alternative to the (infeasible) conventional estimator. In contrast, Clark (2014) argues that the surname-based grouping estimator may capture important aspects of the transmission process that are not captured by the conventional estimator.

Table 5: The Added Informational Content of Surnames and First Names

	Dependent variable: Son's standardized occupational status					
	Surnames			First Names		
	(1)	(2)	(3)	(4)	(5)	(6)
Father's occupational status	Linear	FEs	FEs	Linear	FEs	FEs
Other controls	–	–	Yes	–	–	Yes
Panel A: Finnish Longitudinal Veteran Database						
Father's standardized	0.564			0.612		
occupational status (HISCAM)	(0.039)			(0.026)		
Father's standardized	0.109	0.128	0.086	0.103	0.094	0.085
name mean	(0.032)	(0.031)	(0.023)	(0.023)	(0.021)	(0.014)
Adjusted R-squared	0.249	0.325	0.388	0.252	0.327	0.390
N	5,986	5,986	5,986	5,986	5,986	5,986
Panel B: IPUMS Linked Representative Sample 1880-1900						
Father's standardized occupational	0.375			0.375		
status (log occupational income)	(0.018)			(0.012)		
Father's standardized	0.010	0.006	0.008	0.018	0.019	0.013
name mean	(0.018)	(0.017)	(0.016)	(0.012)	(0.012)	(0.011)
Adjusted R-squared	0.149	0.177	0.248	0.150	0.177	0.248
N	9,076	9,076	9,076	9,076	9,076	9,076
Panel C: Linked 1915 Iowa State Census Sample						
Father's standardized occupational	0.364			0.351		
status (log occupational income)	(0.037)			(0.021)		
Father's standardized	0.016	-0.020	-0.030	0.054	0.049	0.052
name mean	(0.038)	(0.038)	(0.037)	(0.019)	(0.019)	(0.018)
Adjusted R-squared	0.141	0.169	0.180	0.143	0.17	0.181
N	3,204	3,204	3,204	3,204	3,204	3,204

Note: The table reports the coefficients from a regression of son's standardised occupational HISCAM score (Panel A) or log occupational income (Panels B and C) on the standardised father's occupational status and the standardised mean status in his name group. Father's status is controlled for linearly in columns 1 and 4 and flexibly (occupational fixed effects) in all other columns. 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 county (or state) of residence in 1880. Panel C reports estimates from a digitised 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 county of residence in 1915. Standard errors clustered at the household level in parentheses.

the overlap between parent and child sample. Here, to be more explicit, we compare three distinct models: (i) a simple autoregressive process, (ii) a latent factor model, and (iii) an autoregressive process with regional persistence. The first two processes were already introduced in Section 5.5. Given the AR(1) process in equation (18), the direct estimator would identify $\beta = \frac{Cov(y_{ij}, x_{ij})}{Var(x_{ij})} = b$ while the grouping estimator would capture

$$\delta = \frac{Cov(y_{ij}, \bar{x}_j)}{Var(\bar{x}_j)} = b\frac{Cov(x_{ij}, \bar{x}_j)}{Var(\bar{x}_j)}, \tag{23}$$

where the covariance-variance ratio is one in overlapping samples (as also follows from equation (11)), or smaller than one if the parent and child samples do not overlap (as in equation (16)). In this model names have no AIC, so the grouping estimator serves merely as a—potentially attenuated—proxy of the direct estimator, as is also illustrated in Figures 1a and 1b in our simulation.

Alternatively, consider the latent factor model in equations (19)-(21) as also considered by Clark (2014), where outcomes y_{ij} and x_{ij} depend on latent endowments e^y_{ij} and e^x_{ij}, transmitted from parents to children at rate λ. For simplicity we assume that these variables are standardised with mean zero and variance one. In this model, the direct estimator identifies

$$\beta = \frac{Cov(\rho e^y_{ij} + u^y_{ij}, \rho e^x_{ij} + u^x_{ij})}{Var(x_{ij})} = \lambda\rho^2, \tag{24}$$

while the grouping estimator with full overlap instead identifies

$$\delta = \frac{Cov(\rho e^y_{ij} + u^y_{ij}, \rho \bar{e}^x_j + \bar{u}^x_j)}{Var(\bar{x}_j)} = \frac{Cov(\rho\lambda e^x_{ij}, \rho \bar{e}^x_j)}{Var(\rho\bar{e}^x_j) + Var(\bar{u}^x_j)} = \lambda\frac{\rho^2 Var(\bar{e}^x_j)}{\rho^2 Var(\bar{e}^x_j) + Var(\bar{u}^x_j)}, \tag{25}$$

where $\bar{e}^x_j$ and $\bar{u}^x_j$ are the name group means of e^x_{ij} and u^x_{ij}, respectively, and where the last step follows because a regression of a variable on its group mean equals one. The key insight is that the ratio $\frac{\rho^2 Var(\bar{e}^x_j)}{\rho^2 Var(\bar{e}^x_j)+Var(\bar{u}^x_j)}$ approaches one for larger name groups, as the (independent) noise u^x_{ij} tends to average out more quickly across individuals within a surname than the latent endowment e^x_{ij} (which is not independent, as it depends on common ancestors). Online Appendix D.2 provides an illustration. The direct and grouping estimator therefore identify different parameters; while the former depends on both ρ and λ, the latter can isolate λ, the model's primary determinant of long-run persistence across many generations. The grouping estimator would therefore be larger than the direct estimator in overlapping samples, as is illustrated in Figures 1c and 1d in our simulation.[36]

Finally, consider a third model, in which we allow for systematic differences in the outcome across regions. As noted by Güell *et al.* (2018a, Appendix A2), if a surname is frequent in some region but rare in others, its average outcome will tend to be similar to the average outcome of that region. To illustrate this idea, we augment the autoregressive process (18) by assuming that the outcome y_{ij} contains a region-specific effect r^y_j that

[36] However, the grouping estimator may suffer from weak-instrument bias and therefore understate persistence in non-overlapping samples (see Section 5.6 and Choi *et al.*, 2018).

varies systematically across name groups,

$$y_{ij} = bx_{ij} + r_j^y + \tilde{u}_{ij}, \tag{26}$$

as does the outcome x_{ij} for parents, $x_{ij} = r_j^x + \tilde{x}_{ij}$, where $\tilde{x}_{ij}$ is assumed to be uncorrelated to r_j^y. Assume region-specific effects are linked across generations according to $E[r_j^y|r_j^x] = \mu r_j^x$, where μ might be close to one if children tend to stay in the same region as their parents and socioeconomic differences between regions are stable. In this model, the direct estimator identifies

$$\beta = \frac{Cov(bx_{ij} + r_j^y + \tilde{u}_{ij}, x_{ij})}{Var(x_{ij})} = b + \mu\frac{Var(r_j^x)}{Var(x_{ij})}, \tag{27}$$

and would therefore not be very sensitive to regional inequalities if they explain only a small share of the individual variation in status (if $Var(r_j^x) \ll Var(x_{ij})$). However, the grouping estimator would weight those regional inequalities more heavily, identifying

$$\delta = \frac{Cov(bx_{ij} + r_j^y + \tilde{u}_{ij}, \bar{x}_j)}{Var(\bar{x}_j)} = b + \mu\frac{Var(r_j^x)}{Var(\bar{x}_j)}, \tag{28}$$

where the weight on the regional persistence μ is now much greater (as $Var(\bar{x}_j) < Var(x_{ij})$). The intuition here is that if family members locate in the same region, then slow regression to the mean of surname averages might simply reflect a high persistence of regional differences. This applies in particular to frequent surnames, whose group means are less dispersed (such that $Var(\bar{x}_j) \ll Var(x_{ij})$); see Güell *et al.* (2018a) for a detailed discussion of this argument and its implications.[37]

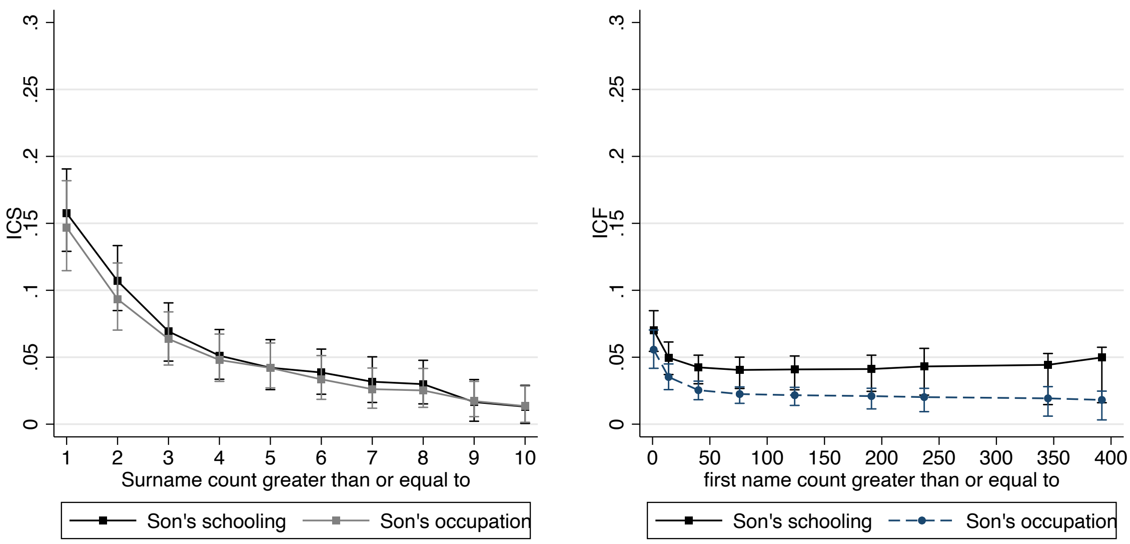
These examples illustrate that the grouping estimator weights the underlying transmission processes differently than the direct estimator. Moreover, different explanations as to why names have added informational content imply different interpretations of the grouping estimator, but many studies focus on only one possible interpretation. Regardless of which interpretation one favours, the bias originating from sampling properties could be more systematically addressed (see Section 5.6). Moreover, certain group-level mechanisms could be systematically controlled for (see Section 6.6).

6.2 Name Frequency

Name distributions are heavily skewed, with a large share of names being held by few individuals and vice versa. This skewness has a first-order effect on the R^2 estimator, which

[37] More generally, names might proxy for geographic, ethnic or other factors that vary systematically across surnames (Chetty *et al.* 2014, Torche and Corvalan 2018, Solon 2018), and while conventional parent-child correlations also reflect such group-level mechanisms (Borjas, 1992), name-based estimators weight them more heavily.

Figure 2: Informational Content vs. Name Frequency



(a) Surnames

(b) First Names

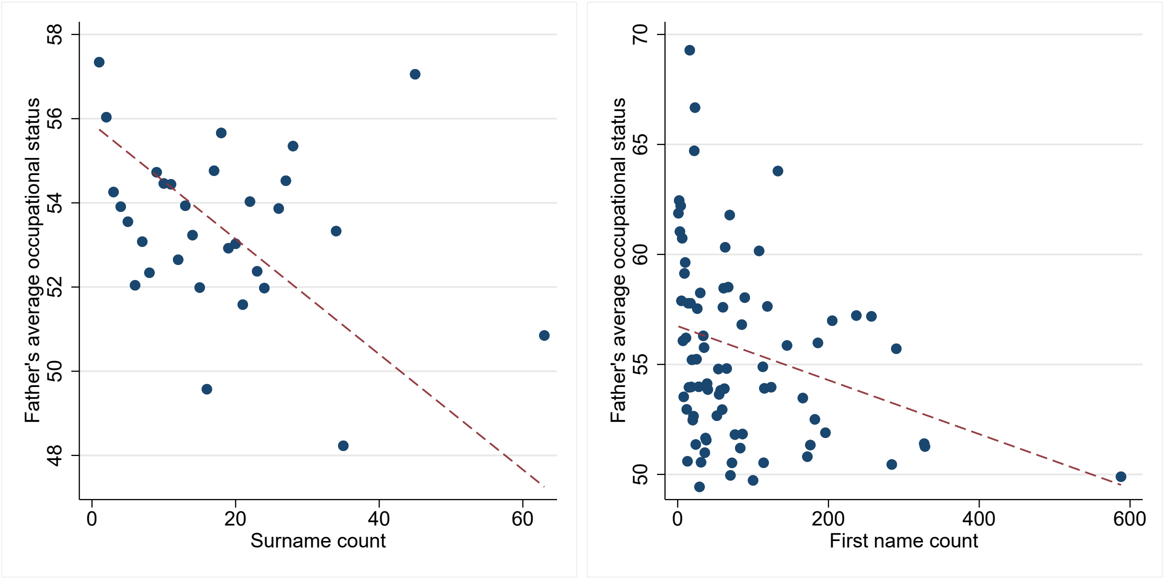
Note: The figures plot estimates of the ICS and ICF and corresponding bootstrap intervals, based on a regression of son's years of schooling (solid line) or son's occupational status (dashed line) on a set of surname dummies (sub-figure a) for name groups with a name count greater than or equal to $n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

is primarily identified from rare names, and *decreases* in name frequency. In contrast, the grouping estimates tend to *increase* in name frequency in our samples. This latter result is however not universal as it represents the net result of two countervailing effects.

Figure 2 shows that the informational content of both surnames and first names decreases with name frequency, but the decay is much faster for the former; the ICS becomes small or zero in larger surname groups. Güell *et al.* (2015) document a similar pattern in data from Catalonia, and attribute it to the natural birth-death process for surnames: rare names are indicative of family links while common surnames are less so and therefore less informative. First names have lower informational content than surnames overall, but frequent first names remain informative—the ICF is comparatively flat across the frequency distribution (see Online Appendix E.1 for further discussion).

These patterns have also implications for the grouping estimator, which depends on the extent to which one can predict a person's socioeconomic status based on the status of others sharing the same name (see equation (15)). Frequent names have lower informational content, which attenuates the grouping estimator (unless the parent and child samples overlap fully). However, the sample mean is a more precise estimate of the population mean for more frequent names. The net effect of those two countervailing forces is therefore ambiguous and may vary across settings. In our Finnish sample, the grouping

Figure 3: Socioeconomic Status Decreases in Name Frequency



(a) Surnames

(b) First Names

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

estimator does not vary much with the frequency of surnames, but increases in the frequency of first names (see Figure E.2, Online Appendix)—as their informational content remains more stable.

How do these observations affect the interpretation of name-based estimators? The debate in the literature has focused on whether the intergenerational transmission process varies systematically with name frequency, but direct mobility estimates actually appear insensitive to name frequency in our data (see Figure E.2) as well as in U.S. tax data (Chetty *et al.*, 2014, Online Appendix). However, name-based estimators tend to be sensitive to name frequency even if the transmission process within families is not. One concern is the relation between name frequency and the informational content of names. For comparative studies based on the R^2 estimator, it is useful to standardise the name frequency distributions (as in Güell *et al.*, 2018a). For the grouping estimator, researchers should report how their estimates vary with name frequency and sample size, and correct for the attenuation bias from "weak" informational content (see Section 5.6).

6.3 The Socioeconomic Gradient in Name Frequency

A related caveat is that socioeconomic status decreases systematically with name frequency. Figure 3 plots the average occupational status in our Finnish sample across bins of the name frequency distribution. The gradient is substantial—the most common surnames (first names) have on average 6.5 (12.0) lower occupational score than rare names, compared to a standard deviation of 12.1.[38] The negative relation between status and the frequency of *first* names is easily understood. As shown by Fryer and Levitt (2004), affluent parents are more likely to choose names for their offspring that are new or different from the most common ones in their society. This observation also holds in our data. Why status decreases in the frequency of *surnames* is less obvious. One hypothesis is that the deliberate choice of a new surname—name "*mutations*"—is more common among high-status individuals, as a form of signalling behaviour by successful dynasties (Collado *et al.*, 2008). We examine this "selective mutation" hypothesis in Section 6.7.

The observation that name-based estimators are identified primarily from rare names (Section 6.2), combined with the observation that people bearing rare names tend to have higher status, suggest that those estimators capture mobility within a non-representative subset of the population. However, they might still approximate the population-average mobility rate if parent-child mobility is similar in rare and more frequent names—which appears to be the case in our Finnish data (as seen in Figure E.2 of Online Appendix E) as well as U.S. data (see Chetty *et al.*, 2014, Online Appendix Table V). While this observation is reassuring, it may still be useful to report the socioeconomic gradient in the name frequency distribution in applications.

6.4 Finite-Sample Properties and Sample Size

How sensitive are name-based estimators to sample size? As they are attractive in settings in which register-based sources are not available, many applications are based on small samples. This problem may be accentuated by the nature of the research question. For example, in mobility comparisons across regions, subsamples inevitably become small no matter how large the base sample is.

Sample size has serious implications for the grouping estimator. Its *leave-out* variant depends on the attenuation factor $\frac{Cov(\bar{x}_{(i)j},x_{ij})}{Var(\bar{x}_{(i)j})}$, and therefore on the extent with which one can predict a person's socioeconomic status with the status of others sharing the same name (see equation (15))—which increases in sample size. To reduce this bias, researchers might be tempted to restrict their sample to name groups that are sufficiently large. However, as frequent names have less informational content (see Section 6.2), the relation between the grouping estimator and name frequency is ambiguous. Instead, we propose that researchers implement two tests. First, to report how sensitive the grouping estimates are to fluctuations in sample size, motivated by the observation that a heavily-attenuated grouping estimator also tends to be sensitive to further reductions in sample size (see Section 5.5). Second, by estimating the attenuation factor directly in order to

[38] Jaramillo-Echeverri *et al.* (2021) find a negative correlation between surname frequency and socioeconomic status in Chile, but not in Colombia.

construct bias-corrected estimates (see Section 5.6).

Sample size is less of a concern for the R^2 estimator. Its definition in equation (3) as the difference between the true R^2 and a placebo R_P^2 already accounts for overfitting in finite samples (see Online Appendix E.2). However, this version of the estimator produces noisy estimates in smaller samples, as different placebo draws may result in very different estimates of R_P^2. A simple solution is to report the mean of the R^2 estimator across multiple placebo draws, as we do in Table 3. Moreover, researchers may wish to quantify the precision of their estimates, and we propose a simple bootstrap procedure that accounts for the uncertainty from the reshuffling of placebo names as well as standard sampling uncertainty (see Online Appendix E.2 for details).

6.5 Selective Samples

While bias from non-random sample selection is a general problem, there are reasons to suspect that it is a particular issue in name-based studies. Many studies are based on historical data or registries, which can be selective; for example, a wealth register might more likely list those who have high wealth. Moreover, many data sources exclude migrants, and their intergenerational mobility tends to deviate from non-migrants (see footnote 29). In addition, name-based estimators tend to respond differently to certain sampling choices than the "direct" estimator. They are particularly sensitive to the frequency of names (see Section 6.2), and some studies focus on a subset of rare surnames (e.g., Clark, 2014) while others consider both frequent and infrequent names (e.g., Benveniste, 2023). On the one hand, a restriction to rare names is intuitive; individuals sharing a rare surname are likely related, while the relation of those sharing a common surname is less certain. On the other hand, the attenuation bias from limited overlap between parent and child samples (see Section 5) tends to be worse in rare name groups. Sampling choices may matter even in fully overlapping samples: for example, if the true data generating process is the latent factor model presented in Section 5.5, a grouping estimator implemented on frequent names would identify its latent persistence while the same estimator implemented on infrequent surnames might not (see Online Appendix D.2 for an illustration). Conversely, if surnames are clustered in certain regions and regional differences are pronounced, the grouping estimator applied to frequent surnames might capture regional persistence (see Section 6.1 or Güell *et al.* 2018a, Appendix A2). There are therefore conflicting reasons why one might want to focus on frequent or infrequent names, and we abstain from taking a position on this question here. Instead, we suggest that researchers show how their estimates vary with the frequency of names and try to rationalise the observed pattern.

Table 6: Stability of Name-based Estimators to the Inclusion of Controls

	Dependent Variable: Son's occupational status				
	(1)	(2)	(3)	(4)	(5)
Panel A: Finnish Longitudinal Veteran Database (N=5,986, White Guards only)					
Direct estimator:	0.600	0.582	0.575	0.527	0.423
R^2 estimator:					
Surnames (ICS)	0.283	0.284	0.237	0.185	0.113
First names (ICF)	0.128	0.165	0.118	0.082	0.044
Grouping estimator:					
Surnames	0.640	0.616	0.605	0.543	0.415
First names	0.763	0.702	0.736	0.617	0.420
Panel B: IPUMS Linked Representative Sample 1880-1900 (N=9,076)					
Direct estimator:	0.474	0.460	0.458	0.414	0.411
R^2 estimator:					
Surnames (ICS)	0.081	0.077	0.076	0.073	0.067
First names (ICF)	0.035	0.031	0.027	0.019	0.021
Grouping estimator:					
Surnames	0.479	0.459	0.460	0.407	0.401
First names	0.501	0.461	0.456	0.386	0.385
Panel C: Linked 1915 Iowa State Census Sample (N=3,204)					
Direct estimator:	0.441	0.437	0.439	0.365	0.362
R^2 estimator:					
Surnames (ICS)	0.167	0.161	0.164	0.101	0.102
First names (ICF)	0.015	0.016	0.015	0.001	0.003
Grouping estimator:					
Surnames	0.446	0.442	0.444	0.339	0.332
First names	0.533	0.533	0.536	0.380	0.376
Immigrant/ethnic origin		Yes	Yes	Yes	Yes
Year of birth			Yes	Yes	Yes
Region (coarse)				Yes	
Region (fine)					Yes

Note: The direct estimator in column (1) refers to a regression of son's occupational score on his father's occupational score (see Table 4 for details). 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 following control variables were added gradually to the models (columns 2-5): Column (2) adds an indicator for immigrant/ethnic group measured as a dummy for native Finnish speaking in Panel A and dummies for share of grandparents immigrant status in Panels B and C; Column (3) adds year of birth; Column (4) adds a coarse definition of region of birth (or residence during childhood) measured as 10 synthetic counties (classified based on geographic coordinates) in Panel A, 48 states (of residence in 1880) in Panel B or a dummy for whether household's region (of residence in 1915) was urban in Panel C; Column (5) adds a fine definition of region of birth based on parish in Panel A and county in Panels B and C.

6.6 Control Variables

A recurring critique is that name-based estimators weight group-level transmission processes differently than conventional estimators. In particular, Chetty *et al.* (2014) and Torche and Corvalan (2018) argue that the estimates of Clark (2014) might reflect the influence of ethnic or national origin in the transmission of advantages, as surnames vary systematically along those lines. We note that this criticism does not only apply to the surname-based grouping estimator as used by Clark, but to all name-based estimators. One strategy to address such criticisms is to include indicators of ethnic or national origins as a control variables. Indeed, the inclusion of such controls has been standard in applications of the R^2 estimator (Güell *et al.*, 2015), and could be adopted for all name-based methods.

In Table 6, we thus explore the stability of mobility estimates in all three samples to three demographic control variables: ethnic origin, age and region of birth (see table notes for variable definitions in each data set). All estimators are sensitive to the inclusion of demographic controls. As expected, the conventional estimator using direct family links is less sensitive than the name-based estimators (Feigenbaum, 2018, obtains a similar result in the Linked 1915 Iowa State Census). The R^2 estimators are most sensitive, in particular if based on first names. Depending on which socioeconomic outcome, the grouping estimates decline by 15-45% when controlling for ethnicity and region of birth. These results confirm that name-based estimators overweight ethnic and regional factors as compared to conventional methods, but also suggest that those factors can partially be controlled for.

6.7 Name Mutations

While the transmission of surnames is a fairly deterministic affair, name changes or "*mutations*" do occur. In the short run, they are a nuisance for researchers using surnames to infer intergenerational mobility, as they sever the link between parents and children. In the long run however, name mutations are necessary for surnames to retain their informational content. Güell *et al.* (2015) conjecture that a mutation infuses the mutated surname with informational content, securing its functionality as a proxy for kinship for some generations to come. The birth cohorts sampled in our data coincide with a particularly active period of name changes in Finland. Moreover, we observe both the prior (pre-mutation) and the mutated (post-mutation) surname, allowing us to study the birth-death process of names and its impact on name-based estimators quite directly.

We find that mutated surnames tend to be infrequent surnames, and that mutations are highly selective, with the mutation rate increasing four-fold over the distribution of educational attainment (see Online Appendix E.4). Name mutations should therefore increase the informational content of names even in the short run. Indeed, the estimated ICS is about 10% higher when based on post-mutation rather than pre-mutation surnames (Appendix Table E.4). However, name switchers tend to have rare names even prior to their name change, with individuals in the lowest quartile of the name frequency distribution being four times more likely to change their surname than individuals with more common surnames (Figure E.4). We argue that this observation can be rationalised along the same lines as the observation that post-mutation names tend to be infrequent names: rare names have a higher informational content, which may create incentives to pick them (if the signal is intended, see Collado *et al.*, 2008) but also reason to abandon such names (if the signal is unintended).

7 Conclusions

To conclude, we summarise some of the key issues affecting the interpretation of name-based estimators. First, they are predominantly identified from rare names, which are more informative about socioeconomic status than frequent names. Second, name-based estimators weight intergenerational transmission mechanisms differently than conventional estimators based on direct family links. Third, names have *added* informational content beyond proxying for a particular socioeconomic characteristic. Different studies propose different interpretations: For some, the idea that name-based estimators capture more than the conventional parent-child estimates is their principal attraction. Others use them as a feasible “drop-in” replacement for settings in which the direct estimator is infeasible, and consider the added informational content a nuisance.

The interpretation of name-based estimators depends not only on the data generating process, but also properties of the underlying sample. A key property is the “overlap” between the parent and child samples, i.e. the conditional probability that a parent is sampled when his or her child is included in the child sample—which differs widely across applications. As a consequence, estimates are often not comparable across studies, even if based on the same estimator. This in turn may be one reason why some authors find very high intergenerational persistence on the surname level, while others do not— the variability of estimates across studies also reflects their sensitivity to sampling properties. We discussed how to address this issue and to correct for the attenuating effects from limited overlap and sample size. These findings from the intergenerational literature also represent a cautionary tale for two-sample instrumental variable applications in other contexts: while the two samples will be independent when drawn from survey data, the estimator can behave very differently in the type of large-scale census or administrative data that are becoming increasingly common in recent work.

While name-based estimators are therefore subject to many conceptual issues, most can be addressed. To guide practitioners, Table A.2 summarises our main suggestions for the grouping estimator. The list is meant as a starting point, and the precise interpretation of name-based estimators remains open to debate, even after accounting for the statistical issues that we raised here. Still, the innovative use of names has already generated many new insights, and we hope our suggestions will prove to be useful for future applications.

References

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J. and Pérez, S. (2021a). 'Automated linking of historical data', *Journal of Economic Literature*, vol. 59(3), pp. 865–918, doi:10.1257/jel.20201599.

Abramitzky, R., Boustan, L., Jacome, E. and Perez, S. (2021b). 'Intergenerational mobility of immigrants in the United States over two centuries', *American Economic Review*, vol. 111(2), pp. 580–608, doi:10.1257/aer.20191586.

Acciari, P., Polo, A. and Violante, G.L. (2022). 'And yet it moves: Intergenerational mobility in Italy', *American Economic Journal: Applied Economics*, vol. 14(3), pp. 118–63.

Adermon, A., Lindahl, M. and Palme, M. (2021). 'Dynastic human capital, inequality, and intergenerational mobility', *American Economic Review*, vol. 111(5), pp. 1523–48.

Álvarez, A. and Jaramillo-Echeverri, J. (2022). 'Segregation in education: The role of historical groups and the marriage market in Colombia', mimeo.

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

Angrist, J.D. and Pischke, J.S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, ISBN 0691120358.

Bailey, M.J., Cole, C., Henderson, M. and Massey, C. (2020). 'How well do automated linking methods perform? lessons from US historical data', *Journal of Economic Literature*, vol. 58(4), pp. 997–1044, doi:10.1257/jel.20191526.

Barone, G. and Mocetti, S. (2020). 'Intergenerational mobility in the very long run: Florence 1427-2011', *Review of Economic Studies.*

Benveniste, S. (2023). 'Like father, like child: Intergenerational mobility in the French Grandes Écoles throughout the 20th century', AMSE Working Papers 2318, Aix-Marseille School of Economics.

Borjas, G. (1992). 'Ethnic capital and intergenerational mobility', *The Quarterly Journal of Economics*, vol. 107(1), pp. 123–150.

Braun, S.T. and Stuhler, J. (2018). 'The transmission of inequality across multiple generations: Testing recent theories with evidence from Germany', *The Economic Journal*, vol. 128(609), pp. 576–611, ISSN 1468-0297, doi:10.1111/ecoj.12453.

Bukowski, P., Clark, G., Gáspár, A. and Peto, R. (forthcoming). 'Social mobility and political regimes: Intergenerational mobility in Hungary', *J Population Economics*, doi: 10.3368/jhr.0621-11749R2.

Carneiro, P., Lee, S. and Reis, H. (2020). 'Please call me John: Name choice and the assimilation of immigrants in the United States, 1900–1930', *Labour Economics*, vol. 62, p. 101778.

Chetty, R., Hendren, N., Kline, P. and Saez, E. (2014). 'Where is the land of opportunity? The geography of intergenerational mobility in the United States', *Quarterly Journal of*

Economics, vol. 129(4), pp. 1553–1623.

Choi, J., Gu, J. and Shen, S. (2018). ‘Weak-instrument robust inference for two-sample instrumental variables regression’, *Journal of Applied Econometrics*, vol. 33(1), pp. 109–125, doi:10.1002/jae.2580.

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

Clark, G. (2018). ‘Estimating social mobility rates from surnames: Social group or dynastic transmission versus family effects’, mimeo.

Clark, G. and Cummins, N. (2012). ‘What is the true rate of social mobility? Surnames and social mobility, England 1800-2012’, .

Clark, G. and Cummins, N. (2014). ‘Intergenerational wealth mobility in England, 1858–2012: Surnames and social mobility’, *The Economic Journal*, vol. 125(582), pp. 61–85.

Clark, G., Cummins, N., Hao, Y. and Vidal, D.D. (2015). ‘Surnames: A new source for the history of social mobility’, *Explorations in Economic History*, vol. 55, pp. 3–24, ISSN 0014-4983, doi:http://dx.doi.org/10.1016/j.eeh.2014.12.002.

Clark, G. and Diaz-Vidal, D. (2015). ‘How strong is assortative mating? A surname analysis’, Working Paper.

Clark, G., Leigh, A. and Pottenger, M. (2020). ‘Frontiers of mobility: Was Australia 1870-2017 a more socially mobile society than England?’, *Explorations in Economic History*, vol. 76, p. 101327, ISSN 0014-4983, doi:https://doi.org/10.1016/j.eeh.2020.101327.

Collado, M.D., Ortín, I.O. and Romeu, A. (2008). ‘Surnames and social status in Spain’, *Investigaciones Economicas*, vol. 32(3), pp. 259–287.

Collado, M.D., Ortuño-Ortín, I. and Romeu, A. (2012). ‘Intergenerational linkages in consumption patterns and the geographical distribution of surnames’, *Regional Science and Urban Economics*, vol. 42(1-2), pp. 341–350.

Collado, M.D., Ortuño-Ortín, I. and Romeu, A. (2013). ‘Long-run intergenerational social mobility and the distribution of surnames’, UMUFAE Economics Working Papers.

Collado, M.D., Ortuño-Ortín, I. and Stuhler, J. (2023). ‘Estimating intergenerational and assortative processes in extended family data’, *Review of Economic Studies*, vol. 90(3), pp. 1195–1227.

Collins, W.J. and Wanamaker, M.H. (2022a). ‘African american intergenerational economic mobility since 1880’, *American Economic Journal: Applied Economics*, vol. 14(3), pp. 84–117, doi:10.1257/app.20170656.

Collins, W.J. and Wanamaker, M.H. (2022b). ‘Data and code for: African american intergenerational economic mobility since 1880’, Https://doi.org/10.3886/E128442V1.

Connor, D.S. and Storper, M. (2020). ‘The changing geography of social mobility in the United States’, *Proceedings of the National Academy of Sciences*, vol. 117(48), pp. 30309–30317, doi:10.1073/pnas.2010222117.

Craig, J., Eriksson, K. and Niemesh, G.T. (2021). 'Marriage and the intergenerational mobility of women: Evidence from marriage certificates 1850-1910', mimeo.

Durante, R., Labartino, G. and Perotti, R. (2016). 'Academic dynasties: Decentralization and familism in the Italian academia', Working paper.

Feigenbaum, J.J. (2018). 'Multiple measures of historical intergenerational mobility: Iowa 1915 to 1940', *The Economic Journal*, vol. 128(612), pp. F446–F481, doi:10.1111/ecoj.12525.

Ferrie, J.P. (1996). 'A new sample of males linked from the public use microdata sample of the 1850 U.S. federal census of population to the 1860 U.S. federal census manuscript schedules', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 29(4), pp. 141–156, doi:10.1080/01615440.1996.10112735.

Fryer, R. and Levitt, S. (2004). 'The causes and consequences of distinctively black names', *Quarterly Journal of Economics*, vol. 119(3), pp. 767–805.

Goldin, C. and Katz, L.F. (2000). 'Education and income in the early twentieth century: Evidence from the prairies', *The Journal of Economic History*, vol. 60(3), pp. 782–818, doi:10.1017/S0022050700025766.

Güell, M., Pellizzari, M., Pica, G. and Mora, J.V.R. (2018a). 'Correlating social mobility and economic outcomes', *The Economic Journal*, vol. 128(612), pp. F353–F403, doi: 10.1111/ecoj.12599.

Güell, M., Rodríguez Mora, J.V. and Solon, G. (2018b). 'New directions in measuring intergenerational mobility: Introduction', *The Economic Journal*, doi:10.1111/ecoj.12607.

Güell, M., Rodríguez Mora, J.V. and Telmer, C.I. (2007). 'Intergenerational mobility and the informative content of surnames', C.E.P.R. Discussion Papers.

Güell, M., Rodríguez Mora, J.V. and Telmer, C.I. (2015). 'The informational content of surnames, the evolution of intergenerational mobility and assortative mating', *The Review of Economic Studies*, vol. 82(2), pp. 693–735.

Halder, T. (2020). 'Caste, reservation and social mobility in India: 1856 - 2017', mimeo.

Häner, M. and Schaltegger, C.A. (2022). 'The name says it all. Multigenerational social mobility in Switzerland, 1550-2019', *Journal of Human Resources*, pp. 0621–11749R2.

Helgertz, J., Price, J., Wellington, J., Thompson, K.J., Ruggles, S. and Fitch, C.A. (2022). 'A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel', *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 55(1), pp. 12–29, doi:10.1080/01615440.2021.1985027.

Hull, P. (2017). 'Examiner designs and first-stage f statistics: A caution', mimeo.

Inoue, A. and Solon, G. (2010). 'Two-sample instrumental variables estimators', *The Review of Economics and Statistics*, vol. 92(3), pp. 557–561.

Jácome, E., Kuziemko, I. and Naidu, S. (2021). 'Mobility for all: Representative intergenerational mobility estimates over the 20th century', National Bureau of Economic

Research, Working Paper No. 29289.

Jaramillo-Echeverri, J., Álvarez, A. and Bro, N. (2021). 'Surnames and social rank: Long-term traits of social mobility in Colombia and Chile', Banco de Desarrollo de America Latina (CAF), CAF Working Paper 2021/17.

Khawand, C. and Lin, W. (2015). 'Finite sample properties and empirical applicability of two-sample two-stage least squares', mimeo.

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E. and Imbens, G.W. (2015). 'Identification and inference with many invalid instruments', *Journal of Business & Economic Statistics*, vol. 33(4), pp. 474–484, doi:10.1080/07350015.2014.978175.

Lambert, P.S., Zijdeman, R.L., Leeuwen, M.H.D.V., Maas, I. and Prandy, K. (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*, vol. 46(2), pp. 77–89, doi:10.1080/01615440.2012.715569.

Lieberson, S. and Bell, E.O. (1992). 'Children's first names: An empirical study of social taste', *American Journal of Sociology*, vol. 98(3), pp. 511–554, ISSN 00029602, 15375390.

Lindahl, M., Palme, M., Sandgren Massih, S. and Sjögren, A. (2015). 'Long-term intergenerational persistence of human capital: An empirical analysis of four generations', *Journal of Human Resources*, vol. 50(1), pp. 1–33.

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

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

Neidhöfer, G. and Stockhausen, M. (2019). 'Dynastic inequality compared: Multigenerational mobility in the United States, the United Kingdom, and Germany', *Review of Income and Wealth*, vol. 65(2), pp. 383–414, doi:10.1111/roiw.12364.

Nye, J., Mason, G., Bryukhanov, M., Polyachenko, S. and Rusanov, V. (2016). 'Social mobility in the Russia of revolutions, 1910-2015: A surname study', mimeo.

Olivetti, C. and Paserman, D. (2013). 'In the name of the son (and the daughter): Intergenerational mobility in the United States, 1850-1930', NBER.

Olivetti, C. and Paserman, M.D. (2015a). 'In the name of the son (and the daughter): Intergenerational mobility in the United States, 1850-1940', *American Economic Review*, vol. 105(8), pp. 2695–2724.

Olivetti, C. and Paserman, M.D. (2015b). 'Replication data for: In the name of the son (and the daughter): Intergenerational mobility in the United States, 1850-1940', Https://doi.org/10.3886/E112929V1-144541.

Olivetti, C., Paserman, M.D. and Salisbury, L. (2018). 'Three-generation mobility in the United States, 1850–1940: The role of maternal and paternal grandparents', *Explorations in Economic History*, ISSN 0014-4983, doi:https://doi.org/10.1016/j.eeh.2018.07.

001.

Olivetti, C., Paserman, M.D., Salisbury, L. and Weber, E.A. (2020). 'Who married, (to) whom, and where? Trends in marriage in the United States, 1850-1940', National Bureau of Economic Research, doi:10.3386/w28033.

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

Ruggles, S., Genadek, K., Goeken, R., Grover, J. and Sobek, M. (2015). 'Integrated public use microdata series: Version 6.0 [database]', Http://doi.org/10.18128/D010.V6.0.

Santavirta, T. and Stuhler, J. (2024). 'Replication data for: Name-based estimators of intergenerational mobility', .

Solon, G. (2018). 'What do we know so far about multigenerational mobility?', *The Economic Journal*, vol. 128(612), pp. F340–F352, doi:10.1111/ecoj.12495.

Torche, F. and Corvalan, A. (2018). 'Estimating intergenerational mobility with grouped data: A critique of Clark's the son also rises', *Sociological Methods & Research*, vol. 47(4), pp. 787–811.

Upton, A.F. (1980). *The Finnish Revolution 1917-1918*, Minneapolis: University of Minnesota Press.

Vosters, K. (2018). 'Is the simple law of mobility really a law? Testing clark's hypothesis', *The Economic Journal*, vol. 128(612), pp. F404–F421, ISSN 0013-0133, doi:10.1111/ecoj.12516.

Vosters, K. and Nybom, M. (2017). 'Intergenerational persistence in latent socioeconomic status: Evidence from Sweden and the United states', *Journal of Labor Economics*, vol. 35(3), pp. 869–901.

Ward, Z. (2023). 'Intergenerational mobility in American history: Accounting for race and measurement error', *American Economic Review*, vol. 113(12), pp. 3213–48.

Yin, P. and Fan, X. (2001). 'Estimating r2 shrinkage in multiple regression: A comparison of different analytical methods', *The Journal of Experimental Education*, vol. 69(2), pp. 203–224.

A Appendix

Table A.1: Selected Intergenerational Studies based on Names

Authors	*Year*	*Publication*	*Method*	*Data*	*Main Application*
Clark	2014	Princeton University Press	Surnames, Grouping	Repeated cross-section of rare surnames	Inter- and multi-generational mobility in various countries
Olivetti and Paserman	2015	American Economic Review	First names, TS2SLS	Repeated cross-section	Historical mobility trends in the United States
Güell, Rodríguez and Telmer	2015	Review of Economic Studies	Surnames, R2	Single cross-section	Intergenerational mobility level and trends in Catalonia
Clark	2012	Working Paper	Surnames, Name frequencies	Repeated cross-section of surname frequencies	Multigenerational mobility in Sweden
Clark and Cummins	2012	Working Paper	Surnames, Grouping (region)	Repeated cross-section of rare surnames	Multigenerational mobility in England
Collado, Ortuño and Romeu	2012	Reg. Science and Urban Econ.	Surnames, Grouping	Single cross-section across areas	Intergenerational consumption mobility in
Collado, Ortuño and Romeu	2013	Working Paper	Surnames, Grouping	Repeated cross-section of surname averages	Multigenerational mobility in Spanish provinces
Clark and Cummins	2014	Economic Journal	Surnames, Grouping	Repeated cross-section of rare surnames	Multigenerational wealth mobility in England
Clark and Diaz-Vidal	2015	Working Paper	Surnames, Grouping	Repeated cross-section of surname averages	Multigenerational and assortative mobility in Chile
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	Surnames, Name frequencies	Single cross-section of surname frequencies	Family connections at Italian universities
Feigenbaum	2018	Economic Journal	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, TS2SLS	Repeated cross-section	Multigenerational mobility in the United States
Barone and Mocetti	2020	Review of Economic Studies	Surnames, TS2SLS	Repeated cross-section of surname averages	Multigenerational mobility in Florence, Italy (1427-2011)
Clark, Leigh and Pottenger	2020	Explorations in Economic History	Surnames, Grouping	Repeated cross-section of rare surnames	Multigenerational mobility in Australia and England
Olivetti, Paserman, Salisbury and Weber	2020	Working Paper	First names, TS2SLS	Repeated cross-section	Mobility trends of women and marital sorting in the US
Eriksson, Craig and Niemesh	2020	Working Paper	First names, TS2SLS	Repeated cross-section of marriage certificates	Intergenerational mobility of women in Massachusetts
Ward	2022	Working Paper	First and Surnames, TS2SLS	Repeated cross-section	Historical mobility trends in the United States
Eriksson, Lake and Niemesh	2022	AEA Papers and Proceedings	First names, TS2SLS	Repeated cross-section	Immigration and marital sorting in the United States

Note: The table lists selected intergenerational mobility research that uses first names or surnames to overcome the lack of direct parent-child links. The top three entries are the key references for the methods discussed in this paper. The table is ordered by the year of publication, not by the timing of contributions. TS2SLS=Two-sample two-stage least squares.

Table A.2: A Practitioner's Guide to the Grouping Estimator

Input:	Repeated cross-sections with individual-level information on socioeconomic outcomes and first or surnames.
Steps:	
Step 1	Estimate the informational content of names, and how it varies with name frequency (Section 4).
Step 2	Report the overlap between the parent and child samples (Sections 5.1-5.3).
Step 3	Implement the grouping estimator (Section 5) in its "manual" or TS2SLS variant.
Step 4	Correct for attenuation bias in partially or non-overlapping samples (Section 5.6).
Step 5	Document how sensitive the grouping estimator is to name frequency (Section 6.2), sample size (Section 6.4) or the inclusion of group-level controls (Section 6.6).

Name-Based Estimators of Intergenerational Mobility

Online Appendix

B Finnish Longitudinal Veteran Database

B.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 political crime courts.

In 1973 the Prisoners of War (POW) of the Red Guard were rehabilitated and granted compensation by the Finnish Government. Everyone who was prosecuted for instigating rebellion by a political crime court and imprisoned in the aftermath of 1918 was entitled to this compensation. The amount varied from a baseline sum of 1,000 Finnish markka ($\approx$ 1,150 Euros in 2018) to 2,500 Finnish markka ($\approx$ 2,900 Euros) depending on the duration of imprisonment.[39] 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 the political crime court in which all individual acts of the prosecutions in 1918 and 1919 of Red Guards 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 the political crime court. From these acts, all individual-level information available, such as sociodemographic background, occupation, and complete name were acquired. In order to remove potential duplicates, we identified 7,907 individuals at the Population Register of Finland (PRF) and were able to link them to their relevant social security number for an identification rate of 99.6%. Excluding either record of the verified 209 duplicates renders a total of 7,698 unique individuals in our data. We restrict the sample to males and remove all females (n=1,076) from our data, rendering a total of 6,632 unique males who fought with the Red Guard in 1918. Our analytic sample includes these individuals.

B.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 civil guard with the aim at assembling a complete registry of White Guard veterans. By the end of 1938, 9,602 home interviews were conducted recording individual-level information on sociodemographic background, civil war, current occupation and complete name. The interviews

[39] 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.

Table B.1: Sampling of Birth Records

	Birth record observed yes/no			
	White Guards		Red Guards	
	(1)	(2)	(3)	(4)
Years of schooling	0.000	-0.001	0.003	-0.001
	(0.002)	(0.002)	(0.003)	(0.003)
HISCAM score in 1918	-0.000	-0.000	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Father's HISCAM	0.001	0.000		
	(0.001)	(0.001)		
Surname count	0.001	0.001	0.000	0.000
	(0.001)	(0.001)	(0.000)	(0.000)
Adjusted R-squared	0.005	0.154	0.000	0.051
N	4,366	4,366	5,688	5,688

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

also contained retrospective questions referring to 1918, for instance, the interviewees were asked to recall their occupation as of 1918. This registry of unique individual interviews is administered by the National Archives of Finland. We acquired all individual-level variables for all individual interviews available in this registry and digitised these records in 2015-2016. Here, since all veterans were interviewed individually, duplicates are not a concern.

B.3 Merging Harmonised Variables From Two Sources

Pooling 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 (the 1934-1938 interviews inquired about current occupation but also occupation as of 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 sociodemographic characteristics such as place of birth and year of birth. Inferring the ethnicity of a name is a fairly simple affair as Finnish and Swedish belong to different language groups (Swedish being an Indoeuropean language and Finnish an Uralic language).

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 inter-

Table B.2: Mobility Using Alternative Measures of Father's Occupational status score

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

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

views). We therefore complement our data by imputing father's occupation from digitised birth records for a subset of the 1918 veterans. This imputation exercise was done for both members of the Red Guard and White Guard in order to probe the accuracy of the birth records. We have two independent measures of father's occupation status score: one based on the self-reported occupation by the son through home interviews (that are our main data source) and the other from the imputed sons' birth records. The two measures have similar moments and are highly correlated. Moreover, the probability of linking a birth record to an individual included in our main data set is uncorrelated with socioeconomic variables (see Appendix Table B.1). The link probability differs across regions, which is a natural consequence of the regionally unbalanced state of the digitisation of Finnish genealogy records for the relevant cohorts.[40] 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 B.2). Reassuringly, this evidence suggests that the self-reported father's occupation of our main data is reliable and accurately reported.

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

C Grouping Estimator: Additional Evidence

C.1 The Grouping Estimator with other Outcomes

Table C.1: 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%	66%	0%	100%	66%	0%
	Dependent variable: Son's log earnings						
Father's log earnings	0.209	0.219	0.181	0.172	0.307	0.190	0.250
	(0.032)	(0.035)	(0.040)	(0.045)	(0.061)	(0.070)	(0.079)
Adjusted R-squared	0.025	0.020	0.015	0.010	0.014	0.005	0.006
N	2,041	2,041	1,377	1,446	2,041	1,491	1,775
	Dependent variable: Son's education						
Father's education	0.264	0.298	0.290	0.237	0.397	0.339	0.214
	(0.023)	(0.027)	(0.031)	(0.035)	(0.047)	(0.048)	(0.053)
Adjusted R-squared	0.055	0.050	0.049	0.028	0.028	0.022	0.005
N	3,378	3,378	2,330	2,452	3,378	2,468	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 clustered at the household level in parentheses.

Table C.1 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 4. We again find that the leave-out grouping estimator is smaller than the inclusive variant, and either larger or smaller than the direct estimates.

C.2 The Grouping Estimator using First *and* Surnames

Table C.2 presents grouping estimates using first and surnames jointly (columns 2-4). For comparison, the table also reports grouping estimates based on surnames only (columns 5-7). The estimates in columns (6) and (7) differ slightly from those reported in Table 4 due to different sample size; all regressions are estimated using the TS2SLS estimator as the leave-out mean is not defined in the case of more than one grouping principle. The estimates are similar when the parent and child sample are fully overlapping (columns 2 and 5), but using first and surnames leads to lower grouping estimates in partial samples.

Table C.2: Grouping Estimator Based on First Name and Surname

	Dependent variable: Son's occupational status						
	Direct	First names and surnames			Surnames alone		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Group definition	–	inclusive	partial		inclusive	partial	
Overlap		100%	66%	0%	100%	66%	0%
Panel A: Finnish Longitudinal Veteran Database							
Father's occupational	0.600	0.637	0.585	0.152	0.640	0.607	0.163
status (HISCAM)	(0.014)	(0.016)	(0.019)	(0.036)	(0.017)	(0.020)	(0.045)
Adjusted R-squared	0.246	0.214	0.190	0.011	0.199	0.184	0.008
N	5,986	5,986	4,026	1,481	5,986	4,026	1,481
Panel B: IPUMS Linked Representative Sample 1880-1900							
Father's log	0.521	0.484	0.427	0.151	0.479	0.452	0.171
occupational income	(0.013)	(0.016)	(0.020)	(0.028)	(0.017)	(0.023)	(0.037)
Adjusted R-squared	0.171	0.125	0.105	0.018	0.103	0.093	0.013
N	9,076	9,076	5,899	1,704	9,076	5,899	1,704
Panel C: Linked 1915 Iowa State Census Sample							
Father's log	0.441	0.449	0.417	0.286	0.446	0.430	0.414
occupational income	(0.021)	(0.022)	(0.027)	(0.046)	(0.024)	(0.029)	(0.046)
Adjusted R-squared	0.141	0.127	0.111	0.064	0.112	0.102	0.097
N	3,204	3,204	2,130	720	3,204	2,130	720

Note: The table reports the coefficients from a regression of son's occupational HISCAM score (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 first name and surname (columns 2-4) or surname alone (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 clustered at the household level in parentheses.

C.3 The Grouping Estimator using 1940 Occupation Scores

The most recent historical American social mobility literature favours the complete count of the 1940 Census as the basis for occupational scores instead of the 3.3% sample of the 1950 Census for at least two reasons (Connor and Storper, 2020; Jácome *et al.*, 2021; Abramitzky *et al.*, 2021b; Collins and Wanamaker, 2022a). First, the complete count of the 1940 Census has been made available to researchers since 2013, enabling the estimation of precise occupational scores even in specific region, race and gender cells. Second, 1940 is closer in timing to the relevant time period that much of the literature that makes use of historical Censuses (1860-1940) concern. We present robustness checks in Table C.3 in which Panels B and C of Table 4 are replicated with occupation scores proposed by Collins and Wanamaker (2022a) calculated as region-specific occupational average earnings by

Table C.3: Grouping Estimator, Occupational Income Based on 1940 Occupation Scores

	Dependent variable: Son's occupational status, 1940 basis						
	Direct	Surnames			First names		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Group definition	–	inclusive	partial	leave-out	inclusive	partial	leave-out
Overlap		100%	66%	0%	100%	66%	0%
<u>IPUMS Linked Representative Sample 1880-1900</u>							
Father's log occupational	0.392	0.389	0.378	0.158	0.443	0.391	0.245
income, 1940 basis	(0.010)	(0.012)	(0.014)	(0.020)	(0.021)	(0.027)	(0.031)
Adjusted R-squared	0.185	0.125	0.121	0.016	0.053	0.039	0.010
N	9,076	9,076	5,991	5,119	9,076	6,586	8,029
<u>Linked 1915 Iowa State Census Sample</u>							
Father's log occupational	0.522	0.538	0.530	0.450	0.642	0.573	0.267
income, 1940 basis	(0.022)	(0.026)	(0.031)	(0.036)	(0.041)	(0.048)	(0.058)
Adjusted R-squared	0.186	0.152	0.150	0.096	0.074	0.065	0.009
N	2,774	2,774	1,887	1,924	2,774	2,024	2,388

Note: The table reports the coefficients from a regression of son's log occupational income 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). The occupational income sore is calculated based on average occupational earnings in the complete count of the 1940 US Census. See Section 5.1 and the Online Appendix of Collins and Wanamaker (2022) for a detailed description of the construction of the occupational income scores. Standard errors clustered at the household level in parentheses.

Census division based on the complete count of the 1940 US Census. As to occupational income of the sons (as measured in the 1900 Census) in the IPUMS Linked Representative Sample 1880-1900 (top panel), we follow the correction method by Collins and Wanamaker (2022a) for farmer's income. Specifically, we distinguish land-owning farmers from tenant farmers and scale up farmers', farm managers' and farm labourers' incomes to account for in-kind transfers (see the Online Appendix of Collins and Wanamaker (2022a) for a detailed description of the correction method). Since farm ownership status is neither recorded in the 1880 Census (when fathers' outcome was measured in the top panel) nor in the entire Linked 1915 Iowa State Census Sample (fathers' and sons' outcomes of the bottom panel) we cannot apply the farmer income correction for other outcomes than the mentioned 1900 occupational income (sons' outcome in Panel B). The in-kind transfer scaling is however done for agricultural occupations in all relevant US Census cross sections. All results are qualitatively similar to the results in Table 4; the leave-out grouping estimator is consistently smaller than the inclusive estimator. As to the bottom panel, the comparison is somewhat muddled by the difference in sample size between Tables C.3 and 4 due to missing "occ1950" codes in the Feigenbaum (2018) data.

C.4 Overlap in other Applications

Table C.4: Overlap Between Main and Auxiliary Samples in TS2SLS Applications

Authors	*Year*	*Publication*	*Method*	*Overlap*	*Application*
Miguel	2007	Review of Economic Studies	TS2SLS	9% (a panel of 11 years and 67 villages, first stage for only one year, n=67)	Impact of income shocks on witch murders in rural Tanzania
Feldman	2010	Review of Economics and Statistics	TS2SLS	0% (data from 1991 in first stage, data from 1992 in second stage)	Effect of income tax shifting on mental accounting
Brunner, Cho and Reback	2012	Journal of Public Economics	TS2SLS	65% (sub-sample of 1,699 observations in first stage, 2,617 observations in second sample)	Effect of school choice programs on mobility and housing markets
Siminski	2013	Review of Economics and Statistics	TS2SLS	78% (first stage includes full population 868,605, second stage includes 22% fewer, n=675,832)	Effect of vietnam-era service and veteran status on labor market outcomes
Currie and Yelowitz	2000	Journal of Public Economics	TS2SLS	<1% (CPS March 1990-95 in first stage, IPUMS 1% and 5% samples of 1990 Census in second stage)	The effect of linving in housing projects on educational attainment

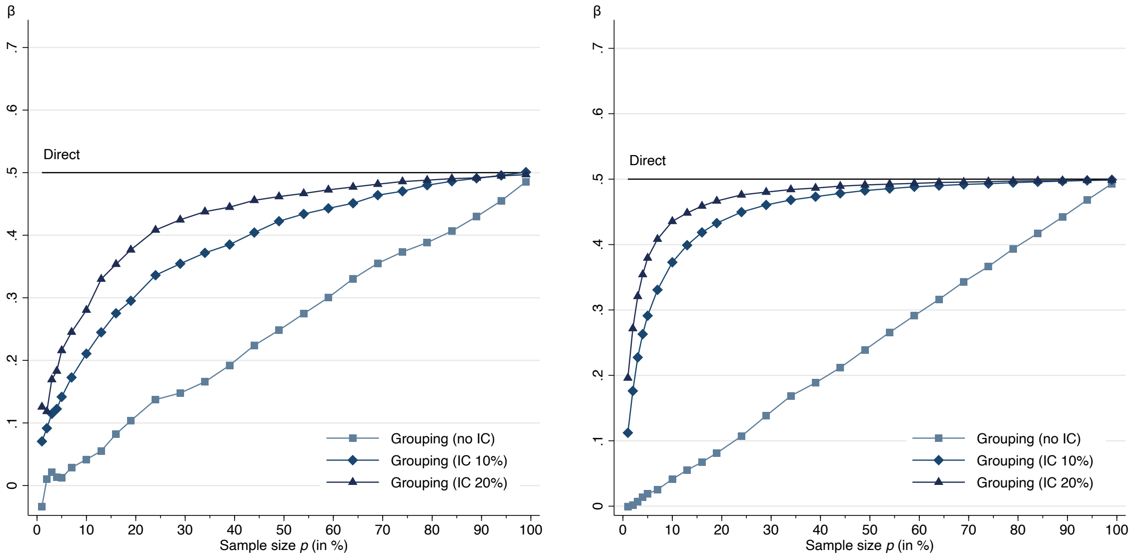
Note: The table lists selected research in Economics outside the context of intergenerational mobility that applies the TS2SLS estimator. Source: Choi et al. (2016).

One key result in Section 5 is that the *overlap* between the parent and child samples has important implications for the properties of the grouping estimator. Because this overlap differs across applications, grouping estimates might not be directly comparable, even if the underlying methodology is the same. Such variation in overlap may also occur in other settings than the intergenerational context that we are focused on here. To illustrate this point, Table C.4 lists several recent studies that use the TS2SLS estimator (a subset of studies considered by Choi *et al.* 2018). Based on the information of the sample sizes of the auxiliary and primary data reported in these papers, we also report the implied degree of overlap. We see that it varies widely, with near-complete overlap between the main and auxiliary samples in some studies and essentially no overlap in others.

D Additional Simulation Evidence

D.1 The Grouping Estimator and Population Size

Figure D.1: The Grouping Estimator vs. Population Size (Simulation)



(a) AR(1), small population (5,000 names)

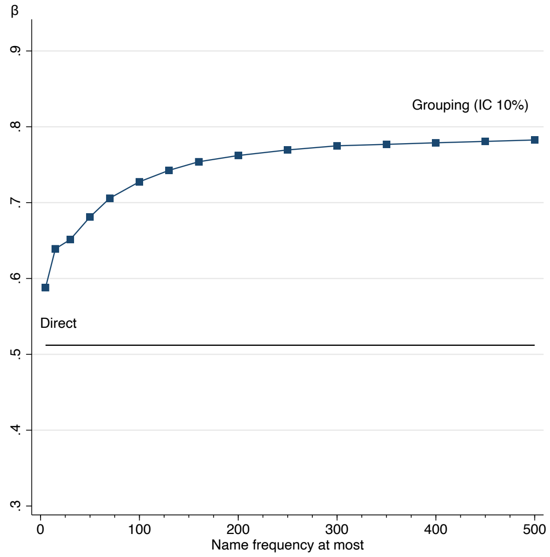
(b) AR(1), large population (50,000 names)

Note: Grouping estimates from differently sized samples (x-axis). The simulated data is based on an AR(1) process with slope $\beta = 0.5$, in which parent status is normally distributed with name fixed effects explaining some of the variation (→ IC). Sub-figure (a) is based on a simulated name distribution with 5,000 names, uniformly distributed frequency between 1 and 100, sub-figure (b) is based on 50,000 names, uniformly distributed frequency between 1 and 500.

To probe the role of population size, Figure D.1 replicates Panel (a) of Figure 1 for two alternative population sizes: a "small" population (5,000 names, uniformly distributed frequency between 1 and 100) and a "large" population (50,000 names, uniformly distributed frequency between 1 and 500). The data-generating intergenerational process is an AR(1) model with $\beta = 0.5$, in which names have no added informational content. In a first step, we generate parent and child status for the entire population. We then draw sub-samples of size p separately for the parent and child generations, where p also determines the overlap between the parent and child samples. Finally, we estimate the grouping estimator within each sub-sample. We find a similar qualitative pattern in both cases, but the attenuation in the grouping estimates is more severe in the small population, in particular for intermediate values of the overlap. The intuition is that in smaller populations, the leave-out mean in the name group is a worse predictor for own status. Formally, the attenuation term in equation (16) tends to be smaller in small populations, or when there are fewer individuals per name group.

D.2 The Grouping Estimator and Name Frequency

Figure D.2: The Grouping Estimator vs. Name Frequency (Simulation)



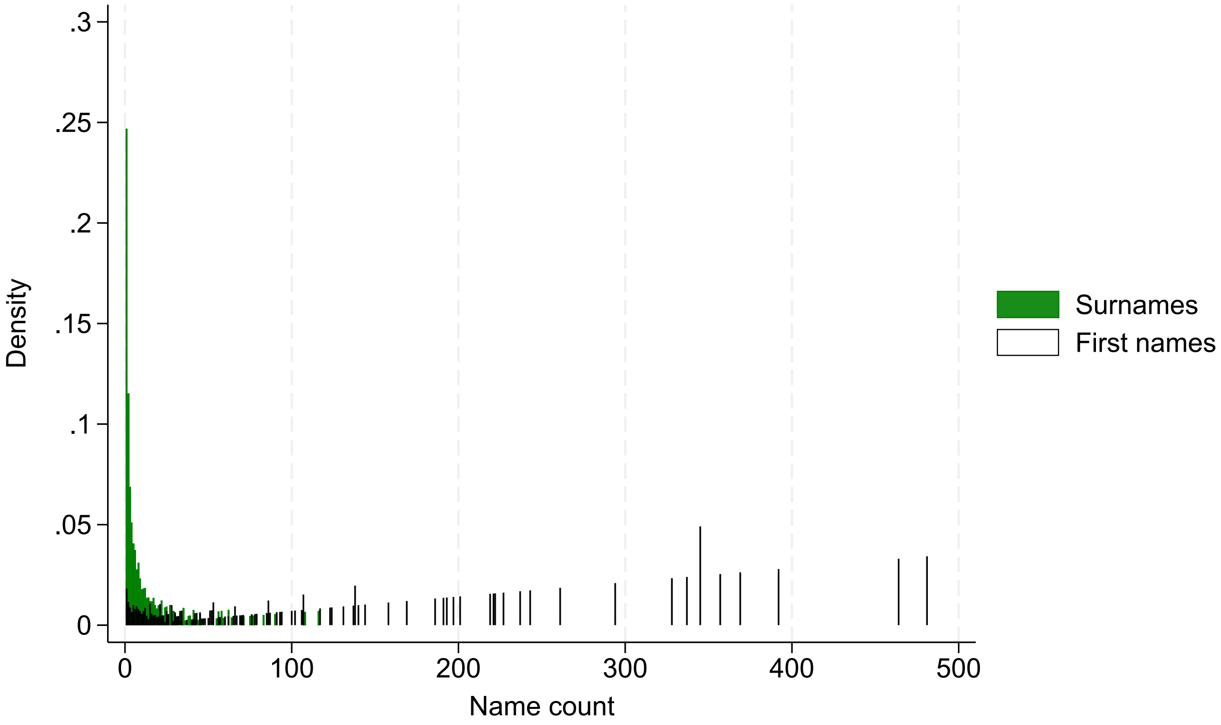
Note: Grouping estimates using more or less frequent names (x-axis). Simulated data based on a latent factor model given by equations (19)-(21) and $\rho = \lambda = 0.8$ (such that $\beta = \lambda\rho^2 = 0.8^3$). Based on 50,000 names, uniformly distributed frequency between 1 and 500.

To probe the role of name frequency, Figure D.1 implements the grouping estimator in fully overlapping samples using name groups of different sizes. The simulated data is based on the same latent factor model with $\rho = \lambda = 0.8$ as Panel (a) of Figure 1, for 50,000 names, uniformly distributed frequency between 1 and 500. Including frequent names, the grouping estimate is just below 0.8 (our chosen value of λ). However, if we restrict the sample to smaller names, the grouping estimates is below 0.6, and therefore much closer to the direct estimate. In particular, the grouping estimator collapses on the direct estimator when the frequency of each name converges to one. The example illustrates that choosing only frequent or only rare names may yield systematically different estimates, depending on the underlying structural model.

E Properties and Caveats: Additional Evidence

E.1 Name Frequency: Additional Evidence

Figure E.1: The Sample Frequency of First and Surnames in Finnish Sample



As shown in Figure E.1, the frequency of names in our Finnish sample differs widely, in particular for surnames. While 76% of the individuals in our sample have a first name that ranks within the 50 most popular names, the corresponding share for surnames is below 20%. The informational content of both surnames and first names decreases with name frequency (see Figure 2), but the rate of decay is much faster for the former. While the informativeness of surnames washes out in larger name groups, the choice process underlying first names remains relevant. Frequent first names may be informative because of aspirational naming (Olivetti and Paserman, 2015a) or because name preferences vary across socioeconomic groups (Lieberson and Bell, 1992). For example, names with a royal or noble connotation may be generally popular though *more* popular among families with high socio-economic status—and as such maintain their informational content. In our sample, name preferences differ between members of the White and Red Guard (see Table E.1).[41]

[41] Among the White Guard, none of the top-5 most prestigious names (as measured by mean occupational

Table E.1: Most and Least Prestigious First Names

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

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

These observations have also important implications for the grouping estimator. As shown in equation (15), the estimator depends on the extent to which one can predict a person's socioeconomic status based on the status of others sharing the same name —which is closely related to the informational content of names. Frequent names have lower informational content, which tends to attenuate the grouping estimator, in particular when the parent and child samples do not overlap and contain individuals from different families (what we call the *leave-out* variant of the grouping estimator). However, the sample mean is a more precise estimate of the population mean in larger name groups. It is therefore ambiguous if the grouping estimator decreases or increases in name frequency.[42] Figure E.2 illustrates that in our main sample, the grouping estimator does not vary much with the frequency of surnames, but does increase in the frequency of first names—because the informational content declines less with name frequency for the latter. We find a similar pattern using the *IPUMS Linked Representative Sample.*

E.2 Finite Sample Properties of the R^2 Estimator

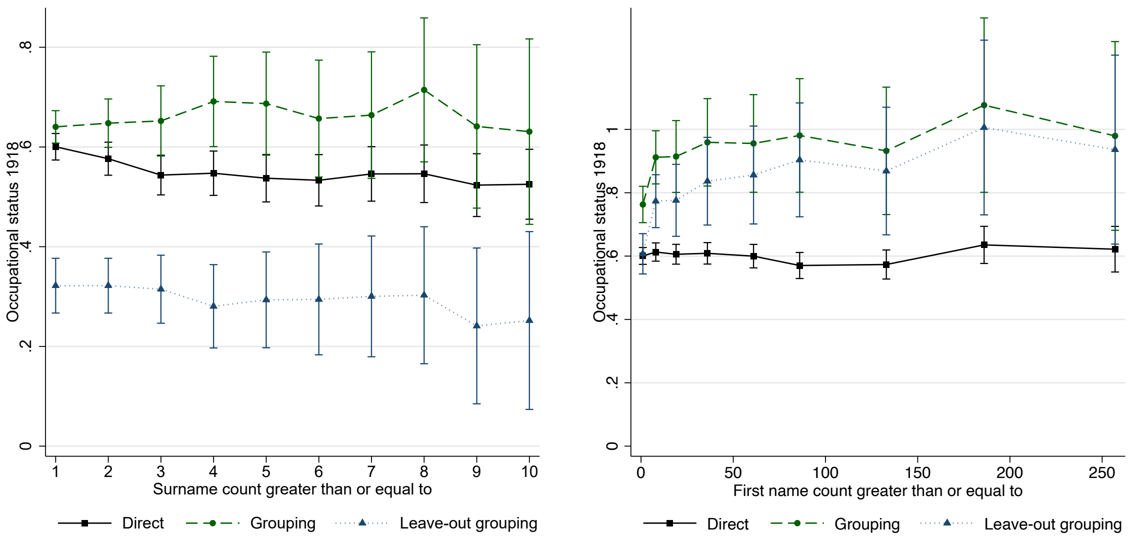
As Güell *et al.* (2015) and Güell *et al.* (2018a) develop the R^2 estimator in complete-count Census data they do not address sampling uncertainty. This appendix discusses how their estimation procedures can be adapted to applications in settings with more limited sample size.

A first concern is that the R^2 estimator may be upward biased in smaller samples. OLS regressions are subject to overfitting in finite samples, particularly when including a large set of (name) dummy variables. Accordingly, the *sample* R^2 is a biased estimate of the coefficient of determination (i.e., the *population* R^2). This issue is typically addressed

status) are of Finnish origin, and all use the Swedish spelling form (e.g., Eric vs. Erkki).

[42] As this zero net effect depends on two countervailing effects, it may vary across settings. Indeed, Clark finds large grouping estimates in rare surnames in several sources, while Chetty *et al.* (2014) find that in U.S. tax data, the grouping estimator increases with the frequency of surnames.

Figure E.2: The Grouping Estimator vs. Name Frequency



(a) Surnames

(b) First names

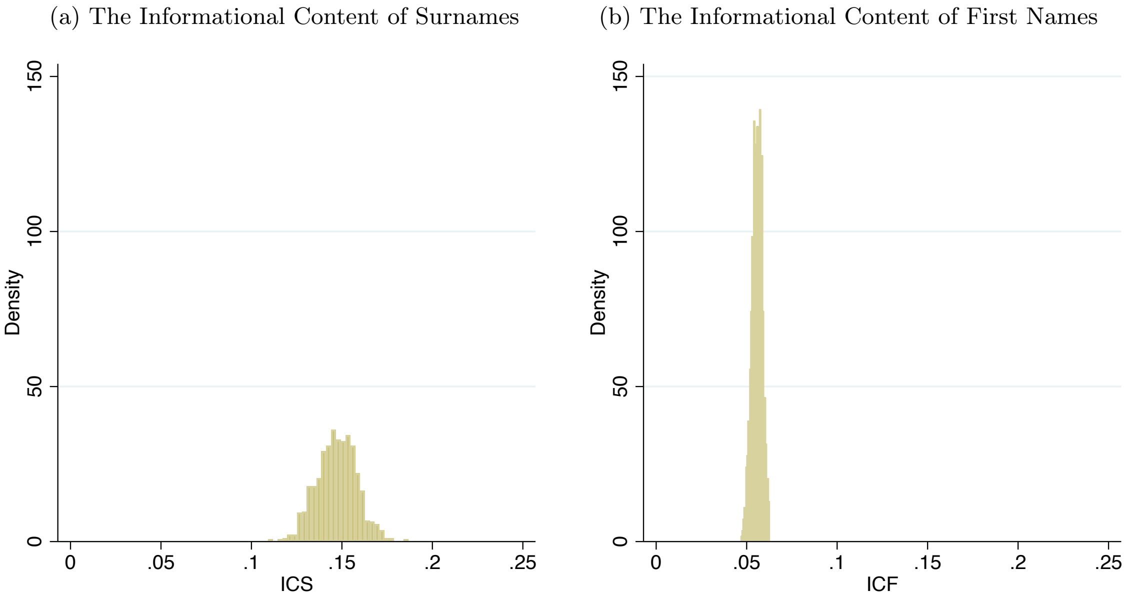
Note: The figures plot the estimate and corresponding confidence intervals from a regression of son's occupational status on father's occupational score (black solid line) or on the imputed occupational score based on either surnames (sub-figure a) for name groups with a surname count greater than or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

by reporting the *adjusted* R^2, which rescales the sample R^2 based on sample size and the number of regressors. Different adjustment formulas are in use, and the formulas implemented in standard statistical software may not produce the least biased results (Yin and Fan, 2001). Interestingly, Güell *et al.* (2015) chose an empirical instead of an analytical bias adjustment, comparing the sample R^2 to the corresponding R^2_P from a placebo regression in which the name dummies have been reshuffled across individuals (see equation (3)). The difference between R^2 and $R^2 - R^2_P$ can be interpreted as an estimate for the extent of overfitting in finite samples.

However, this empirical approach produces imprecise estimates in small samples. The reshuffling of names across individuals introduces uncertainty, with different draws of the name distribution resulting in quite different estimates. For illustration, Figure E.3 plots the distribution of the R^2 estimator when reshuffling the name distributions 1,000 times. Sub-figure (a) plots the estimated ICS corresponding to the specification reported in column (3) of Table 3, while sub-figure (b) plots the estimated ICF corresponding to column (6). A simple solution to this issue is to report the mean of the R^2 estimator across repetitions, as we do in Table 3.[43]

[43] While this is a natural extension of the definition in Güell *et al.* (2015), the question arises as to whether analytical methods to estimate the population R^2 would perform better in smaller samples.

Figure E.3: The Informational Content and Placebo Distributions


(a) The Informational Content of Surnames

(b) The Informational Content of First Names

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

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

E.3 The *Added* Informational Content of Names: Additional Evidence

In Section 6.1, we show that names have *added informational content (AIC)*. In Table E.2 we provide additional evidence from alternative outcome variables in the 1915 Iowa State Census (linked to the 1940 U.S. Federal Census by Feigenbaum, 2018). Panel A reports results from a regression of son's log earnings on a linear or flexible function of

Table E.2: The Added Informational Content with Other Socioeconomic Outcomes

	Dependent variable: Son's standardized occupational status					
	Surnames			First Names		
	(1)	(2)	(3)	(4)	(5)	(6)
Father's status	Linear	FEs	FEs	Linear	FEs	FEs
Other controls	–	–	Yes	–	–	Yes
	Dependent variable: Son's earnings (standardised)					
Father's earnings	0.163			0.158		
(standardised)	(0.067)			(0.033)		
Father's name mean	0.031	0.037	0.039	0.060	0.068	0.059
(standardised)	(0.061)	(0.066)	(0.062)	(0.031)	(0.033)	(0.031)
Adjusted R-squared	0.024	0.031	0.095	0.026	0.034	0.097
N	2,041	2,041	2,041	2,041	2,041	2,041
	Dependent variable: Son's years of schooling (standardised)					
Father's years of schooling	0.160			0.200		
(standardised)	(0.038)			(0.023)		
Father's name mean	0.089	0.068	0.046	0.073	0.073	0.063
(standardised)	(0.038)	(0.038)	(0.037)	(0.022)	(0.020)	(0.020)
Adjusted R-squared	0.058	0.069	0.111	0.060	0.072	0.113
N	3,378	3,378	3,378	3,378	3,378	3,378

Note: The table reports the coefficients from a regression of son's standardised earnings in the top panel (years of education in the bottom panel) on the standardised father's earnings (years of education) and the standardised mean of the fathers' earnings (years of education) in the name group. Father's status is controlled for linearly in columns 1 and 4 and flexibly by including earnings (years of education) fixed effects in all other columns. All estimates are based on 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 birthplace. Standard errors clustered at the household level in parentheses.

his father's earnings and the mean score in his surname group. Panel B reports the corresponding evidence for educational outcomes. If surnames were merely an imprecise proxy for individual status then the coefficient of the group mean should be insignificant. Instead, the imputed mean status of a father's name group tends to have a significant association with the sons's status, even conditional on the father's own status.

E.4 Name Mutations: Additional Evidence

Güell *et al.* (2015) conjecture that without name *mutations*, surnames would eventually collapse into one universal surname, and hence no longer contain socioeconomic information. Instead, name mutations infuse surnames with informational content and secure their functionality as a proxy for kinship for some generations to come. The frequency of surname mutations varies substantially over time and particularly active periods have

Table E.3: Descriptive Statistics of Name Mutations

	Red Guards	White Guards
Number of mutations	556	831
Mutation rate (in %)	8.4	8.7
Mutation of name ethnicity	69.8	75.0
Pre-mutation name:		
Mean frequency	4.4	3
Percent unique	11.3	17.3
Post-mutation name:		
Mean frequency	3.8	2.7
Percent unique	18.1	24.7

Source: The Finnish Longitudinal Veteran Database.

often been spurred by political and nationalistic movements.[44] The frequency also varies in contemporary contexts.[45] The birth cohorts sampled in our data coincide with the aforementioned particularly active period of name changes in Finland. Moreover, we observe both the prior (pre-mutation) and the mutated (post-mutation) surname. Thanks to these two advantages we can explore the birth-death process of names in detail, and we compare our findings to related evidence from Spain provided by Collado *et al.* (2008). The mutation rate is more than 8% in our data on Finnish Civil War veterans, for both White and Red Guard (see Table E.3). For comparison, the estimated lifetime mutation rate in Güell *et al.* (2015) is only about 0.25%. We observe nearly 600 name mutations among the Red Guard, and more than 800 name mutations among the White Guard.

Table E.4 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% (the drop is significant at $p = 0.01$). As illustrated in Figure E.4a, post-mutation surnames tend 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

[44] Paik (2014) reports that during the Japanese occupation, many Koreans strategically changed their clan lineage. In Finland, name changes were particularly frequent during the romantic nationalist movement for independence from Imperial Russia around the turn of the twentieth century. Many name fennicised Swedish or Russian-sounding names to ethnic Finnish-sounding names (e.g., the typical Swedish surname Gustafsson was changed to Lainio after a well kown fell in the Finnish part of Lappland and Russian-sounding Bordakoff became Nuotio, which is the Finnish word for bonfire), but switches from one Finnish-sounding name to another were also common (e.g., a typical peasant name, Peltonen was changed to Linnankoski). In particular, names that were common among sharecroppers were converted to national romantic names with references to nature.

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

research, as it is rare names from which most information on socioeconomic status can the extracted (see Section 6.2). Still the effect on the ICS appears surprisingly limited, given the frequent name changes in our period of study. The reason for this becomes clear from Figure E.4b: name switchers tend to have rare surnames even prior to their name change. For both Red and White Guard, individuals in the lowest quartile of the name frequency distribution are about four times as likely to change their surname as individuals with more common surnames. We hypothesise this observation can be rationalised by the same argument as the observed relation between name changes and post-mutation name frequency. Rare names have a higher informational content, which may either create incentives to pick them (if the signal is intended) or to abandon them (if the signal is unintended). Figures E.4b and E.4a 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).

Table E.4: Informational Content of Surnames Pre/Post-Mutation

	Post-mutation		Pre-mutation	
	Occ. status	Schooling	Occ. status	Schooling
	(1)	(2)	(3)	(4)
Surname dummies	Yes	Yes	Yes	Yes
Adjusted R-squared	0.307	0.409	0.292	0.392
Implied ICS	0.147	0.159	0.132	0.141
95% CI	[0.113, 0.181]	[0.113, 0.181]	[0.103, 0.165]	[0.113, 0.169]
N	14,734	12,824	14,734	12,824

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

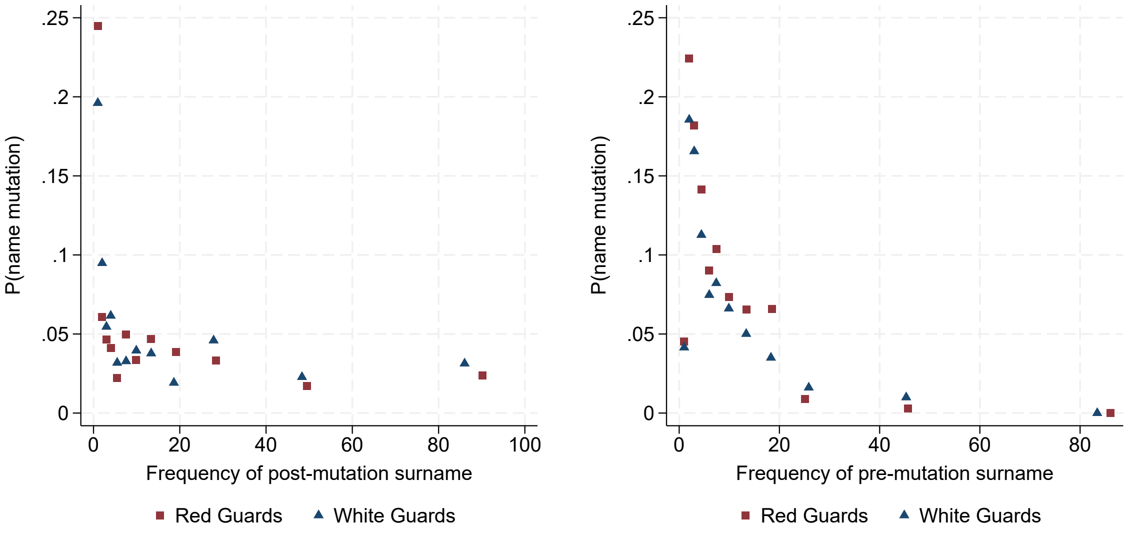
The observation of name changes allows us to directly test whether there is a socioeconomic bias in the probability to change names, as has been hypothesised by Collado *et al.* (2008). Figure E.5 shows that surname mutations are indeed selective, with the probability to change names increasing four-fold 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.[46] For example, Collado et al. show that in Spain,

[46] Güell *et al.* (2015) note that immigrants are more likely to mutate their names, sometimes uninten-

Figure E.4: Name Mutations vs. Name Frequency

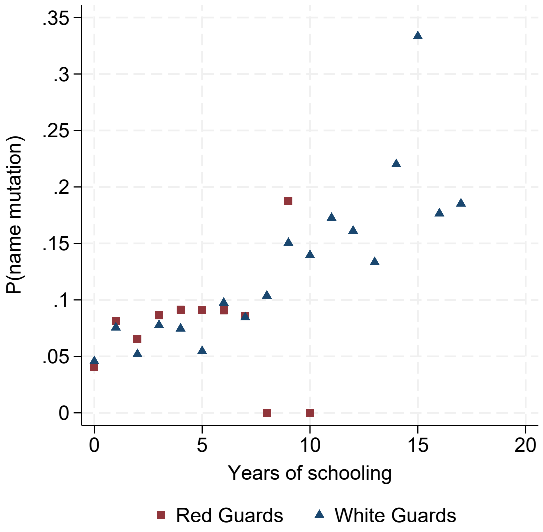
(a) Post-Mutation Surnames

(b) Pre-Mutation Surnames



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

Figure E.5: Socioeconomic Bias in Name Mutations



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

many of the rarer surnames in the 20th century did not exist in the 19th century, and note that surnames act as a signalling device for successful dynasties. Figure E.5 demonstrate that there exists a socioeconomic gradient in name mutations in Finland as well

tionally through transliterations or misspellings by the authorities in the host country. Immigration may therefore reduce the correlation between name mutations and status, if immigrants tend to have lower status. See also the related literature on the economic incentives of name changes for immigrants and the positive consequences of cultural assimilation (Carneiro *et al.*, 2020).