THREE ESSAYS IN LABOR ECONOMICS

By

Kelly Noud Vosters

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics - Doctor of Philosophy

2016

ABSTRACT

THREE ESSAYS IN LABOR ECONOMICS

By

Kelly Noud Vosters

The first chapter tests a recently proposed hypothesis regarding rates of social mobility. Recent work by Gregory Clark and coauthors uses a new surnames approach to examine intergenerational mobility, finding much higher persistence rates than traditionally estimated. Clark proposes a model of social mobility to explain the diverging estimates, including the crucial but untested assumption that traditional estimates of intergenerational persistence are biased downward because they use only one measure (e.g., earnings) of underlying status. I test for evidence of this using an approach from Lubotsky and Wittenberg (2006), incorporating information from multiple measures into an estimate of intergenerational persistence with the least attenuation bias. Contrary to Clark's prediction, I do not find evidence of substantial bias in prior estimates.

The second chapter, coauthored with Martin Nybom, further examines this hypothesis using rich administrative data for Sweden. We exploit detailed proxy measures to test the proposition regarding attenuation bias in prior estimates for Sweden, and also conduct a Sweden-U.S. comparison. We find no evidence of substantial bias in prior estimates, or that the Sweden-U.S. difference in persistence is smaller than found in previous research. We further explore the concept of family status by incorporating mothers, thereby also contributing to the literature on intergenerational transmission for women. We find that while mothers' income is a poor proxy for status, incorporating information on mothers' occupation improves the ability to capture transmission from mothers to both sons and daughters.

The third chapter, coauthored with Cassandra Guarino and Jeffrey Wooldridge, examines the SAS® EVAAS® models for estimating teacher effectiveness, which are used by several states

and districts in teacher evaluation programs despite little attention in the evaluation literature. The EVAAS approach involves using one of two distinct models, the Multivariate Response Model (MRM) or the Univariate Response Model (URM). The MRM jointly models scores from multiple subjects, grades, and cohorts in a 5-year period; it is generally limited to within-district purposes due to the large computational burden and is sometimes not feasible if data requirements cannot be met. Hence, the URM was developed for these situations. The URM models a single subject, and thus is less intensive computationally and more flexible with respect to data requirements. The method involves the computation of a composite score on several lagged scores in multiple subjects, and then using this composite score as the only regressor in empirical Bayes' estimation of the teacher effects. In this paper, we discuss and illustrate advantages and disadvantages of the EVAAS approach relative to the other widely used and studied value-added methods. We perform simulations to evaluate their ability to uncover true teacher effects under various teacher assignment scenarios. We also use administrative data to illustrate the extent of agreement between the URM and other common value-added approaches. Although the differences are small in our administrative data, we show with theory and simulations that standard linear regression using OLS performs at least as well as—and sometimes better than—the more complicated EVAAS URM.

ACKNOWLEDGEMENTS

Many thanks to Gary Solon for his extensive guidance and support. I am also grateful to Jeff Wooldridge and Leslie Papke for providing helpful advice and encouragement. I have been very fortunate to have such a wonderful committee and have learned a tremendous amount from each of them. I would like to acknowledge Martin Nybom, with whom the second chapter is coauthored. The third chapter is coauthored with Cassie Guarino and Jeff Wooldridge, to whom I am grateful for helpful guidance and collaboration. I am also thankful for financial support from the Institute of Education Sciences Grants R305B090011 and R305D10002 to Michigan State University.

Thanks to my classmate Margaret Brehm, whose daily conversations have helped improve my research and also made my time at Michigan State far more enjoyable. I am extremely grateful to my family for their endless love and support throughout my graduate studies as well as my endeavors that led me here. I am also grateful to my husband, Brian, whose unwavering love, patience, support, and sense of humor have given me strength and motivation through the ups and downs of graduate school.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Is the Simple Law of Mobility Really a Law? Testing Clark's Hypothesis

1.1 Introduction

There has been long-standing interest in the persistence of outcomes across generations, from earlier theoretical work by Becker and Tomes (1976, 1979), to the development of intergenerational datasets enabling expansions of empirical work. These studies aim to describe, for instance, the extent to which inequalities are passed on from one generation to the next, or the extent to which opportunities or outcomes have been equalized for children from various family backgrounds. The typical approach to studying intergenerational mobility begins with a basic model relating children's outcomes to parents' outcomes:

$$y_{it+1} = \beta y_{it} + \epsilon_i \tag{1}$$

where i indexes family, t indicates parent's generation and $t+1$ indicates the child's generation.[1] Generally, y_{it+1} and y_{it} represent a measure such as income, wealth, or education. The regression coefficient, β, then provides a measure of persistence, or *immobility*, in the outcome from the parent's generation to the child's generation. Hence, the quantity 1-β can be interpreted as a measure of mobility. For the U.S., the persistence parameter relating a child's log income to parent's log income (hence, an income elasticity) is estimated to be about 0.4 to 0.6 (Solon, 1999; Mazumder, 2005; Lee & Solon, 2009; Black & Devereux, 2011), while for Nordic countries the estimate is lower at 0.1 to 0.3 (Black & Devereux, 2011).[2] These estimates are taken to be summary statistics, describing the extent to which income differences persist from one generation to the next in a country or society. Among the explanations for the lower persistence observed in Nordic countries

[1] In equation (1), along with the remaining equations in the paper, the intercept is suppressed by considering the variables in deviation-from-mean form.

[2] This paper uses intergenerational income regressions as a point of departure, thus extending the income mobility literature, but there is a broader literature that looks at other outcomes. For example, Hertz et al. (2007) is an oft cited recent example providing intergenerational correlation and regression coefficients in educational attainment for 42 countries; Björklund and Salvanes (2011) also provide a succinct review of related literature. Additionally, another subset of the literature is concerned with intergenerational persistence in occupation or occupational prestige. Hodge (1966) is an early example studying intergenerational occupational mobility in the U.S., while Long & Ferrie (2007, 2013) are more recent examples; see also Black & Devereux (2011) for a brief discussion of related studies.

relative to the U.S. is one that highlights why so much attention is given to such differences in mobility: higher mobility may reflect policy differences, such as more redistributive tax structures and generous social welfare programs.

In a recently published book, though, Gregory Clark makes the provocative claim that these estimates are substantially biased downward, and that "true" persistence in social status is much higher—approximately 0.75—and is uniform across all countries and over time (Clark, 2014). The latter part of Clark's assertion, regarding lower mobility, draws on a body of work by Clark and his coauthors, including an article in this journal, that uses innovative methods and a variety of creative names data sources covering many societies over several centuries.[3] The methods exploit the information content of *rare* surnames in these societies to explore social mobility, *without* having actual intergenerational family links.[4] The basic idea is that if inheritance matters, then rare surnames contain information on economic status, and they also indicate some family lineage given naming conventions and the inheritance of paternal surnames (or in some countries both maternal and paternal surnames).[5]

The first part of Clark's controversial claim—regarding bias in prior estimates—is based on a model proposed to explain the discrepancies between mobility estimates. Clark (2014) postulates that the higher persistence rate (0.75) governs a law of social mobility, and summarizes the general intuition underlying the hypothesized downward bias in traditional estimates as:

> "Families turn out to have a general social competence or ability that underlies partial measures of status such as income, education, and occupation. These partial measures are linked to this underlying, not directly observed, social competence only with substantial random components. The randomness with which underlying status produces particular observed aspects of status creates the illusion of rapid social mobility using conventional measures." (Clark, 2014, p.8)

[3] See Clark (2014) for a comprehensive list of these studies, as well as the more recent papers Clark & Cummins (2015) and Clark et al. (2015).

[4] For the data sources containing explicit socioeconomic measures, such as probated wealth at death, equation (1) is estimated using the group averages of wealth for rare surnames. For data without such measures, the approach instead looks at persistence in the representation of the rare surname in an "elite" group relative to representation in the population as a whole.

[5] Güell et al. (2014) show that rare surnames do contain such information, and propose a method using the joint distribution of surnames and economic status to explore intergenerational transmission of status in Spain.

More formally, Clark & Cummins (2015) and Clark (2014) present a simple model for mobility:

$$x^*_{it+1} = bx^*_{it} + e_{it} \tag{2}$$

where x^* represents underlying social status, and b the "true" persistence rate. The hypothesized attenuation bias in prior estimates is thought to arise from the focus on a single "noisy" measure, y_{it} (e.g., income, wealth, or education), of the underlying social status, x^*_{it}, where this relationship is assumed to be of the form:

$$y_{it} = x^*_{it} + u_{it} \tag{3}$$

where u_{it} is idiosyncratic error.[6] Additionally, Clark claims to be able to measure the "true" persistence rate by using surname group averages in equation (1), or $\bar{y}_{zt+1} = b\bar{y}_{zt} + \bar{u}_{zt}$, where z indexes surname (instead of i indexing family). The argument relies on classical measurement error assumptions so that $\bar{y}_{zt} \simeq \bar{x}^*_{zt}$ because $\bar{u}_{zt} \simeq 0$ when the surname samples are sufficiently large.[7]

In a recent article in this journal, Clark & Cummins (2015) present both traditional and surname estimates of social mobility in England using wealth measures to illustrate the discrepancies in mobility estimates, and also test implications of one dimension of the proposed model—the AR(1) form of the law of motion for social mobility in equation (2). However, they do not test the proposed explanation for the discrepancies:

> "... if we were to measure the social status of families as an aggregate of earnings, wealth, education, occupation, and health, then observed social mobility even in parent child studies would decline. For such an aggregation would reduce the variance of the error component in measured status. Thus the measured rate of persistence, even in one generation, will be much closer to that of the underlying latent variable."
>
> (Clark & Cummins, 2015)

[6] Specifically, the assumption is that traditional estimates are biased downward by the usual classical measurement error attenuation factor $\frac{\sigma^2_x}{\sigma^2_x+\sigma^2_u}$, where σ^2_x is var(x^*) and σ^2_u is var(u).

[7] Clark (2014) notes that any group averaging over individuals would similarly reduce measurement error and reveal true status, thus resulting in much higher estimates of persistence (Clark, 2014, p.110). As noted by Solon (2015) however, many of the intergenerational mobility studies that use group averages do not actually find such results. For example, Chetty et al. (2014) show in Appendix D that using surname group averages from administrative U.S. income tax data results in estimates similar to the individual-level regressions.

I fill this gap by exploring this hypothesis that when information from multiple measures is aggregated and then used to obtain traditional estimates, the lower mobility rates will be revealed.[8]

This paper empirically tests the proposed existence, magnitude, and nature of a downward bias in traditional estimates. Conveniently, the theoretical setup for the law of social mobility laid out in equations (2) and (3) translates nicely into a latent variables framework, and the attenuation bias portion of this intriguing theory can be easily tested using publicly available data. Considering x^* the latent status, equation (2) can be interpreted as the structural equation. For each of the particular measures mentioned in the first quotation above (i.e., income, education, and occupation) we can write a separate measurement equation of the form presented in equation (3). Under the strong classical measurement error assumptions maintained in Clark's theory, instrumental variables (IV) using one noisy measure to instrument for another noisy measure produces a consistent estimate of the intergenerational coefficient (IGC), b. If the classical assumptions are relaxed to allow for slope coefficients in the measurement equations as well as unrestricted correlations among the measurement errors, IV estimation is inconsistent. The magnitude and direction of the inconsistency is potentially unknown, depending on the assumptions and measures used. However, an approach proposed by Lubotsky & Wittenberg (2006) is particularly well suited for addressing the case of multiple noisy measures, and under less stringent assumptions. While not identifying b, the method allows one to obtain an estimate with the least attenuation bias—so in this case a greatest lower bound on b—by incorporating information from all of the suggested measures (i.e., income, education, occupation) into a single estimate of b.

In this paper, I employ these approaches using a sample of fathers and sons from the Panel Study of Income Dynamics to test the attenuation bias assumption underlying the law of social mobility. I find little evidence supporting the hypothesized downward bias in prior estimates, and show that incorporating additional measures such as education and occupation has no meaningful impact on the estimated persistence rates obtained from traditional models focused on single measures.

[8] Other recent papers have been testing other hypotheses put forth in Clark's work. For example, in footnote 7, I mentioned the estimates from Chetty et al. (2014) that do not support Clark's assertion that mobility estimates based on any group averages over individuals will result in higher estimates. Clark (2014) also advocates that his results explain why findings from multigenerational regressions indicate a positive grandparent coefficient. In fact, as Solon (2015) points out, the papers do not all find positive coefficients. For instance, Lucas and Kerr (2013) find little evidence of non-zero grandparent coefficients in multigenerational regressions using administrative income data for Finland. Similarly, Braun & Stuhler (2015) use survey data on education and occupation in Germany and find that after controlling for parents' outcomes, they cannot reject a zero coefficient for the grandparents' outcome.

Considering intergenerational persistence in this more comprehensive sense does not reveal higher persistence estimates, but rather confirms the picture of mobility obtained from prior studies that focused on a single measure of socioeconomic status. The paper is organized as follows. In the next section I describe the data and sample. Then I outline the empirical approach, and next present the results. In the last section I summarize the results and conclude.

1.2 Data

I use data from the Panel Study of Income Dynamics (PSID), as this data is ideally suited for my study. The data contains the requisite intergenerational links and also includes information on *multiple* measures of socioeconomic status, which is crucial for testing the attenuation bias claim.[9] Further, I am able to select a sample of individuals very similar to prior PSID studies about which the attenuation bias claims are made, thereby facilitating an appropriate comparison.[10]

The PSID is a longitudinal study that began in 1968 with a sample of approximately 5,000 families in the U.S., with interviews conducted annually through 1997, and biennially since then. Children from these original families are followed when they start their own households, and one can observe family links and follow multiple generations, which is key for traditional intergenerational mobility studies. This paper focuses on the Survey Research Center (SRC) part of the sample[11], in particular during the 1968-1972 surveys for fathers and 1992 survey for sons.[12] While more recent years are available, this time period allows for more direct comparability to prior estimates targeted by the proposed bias, lessens concerns about deterioration of data quality in later years, and still allows sons' ages to be appropriate for measuring earnings outcomes.

[9] Although administrative datasets such as the income tax records used by Chetty et al. (2014) have much larger samples, the data would not suffice for the tests conducted in this paper because information on other status measures such as educational attainment or occupation is not available.

[10] For example, Solon (1992) and Chadwick & Solon (2002) use similar father-son samples. Their sample selections differ in that son's earnings is observed starting at age 25. I restrict my sample to sons for whom I observe earnings starting at age 30 (up to age 40), to minimize life-cycle bias, as discussed below.

[11] The SRC sample was designed to be nationally representative in 1968, while the other component—the Survey of Economic Opportunity (SEO) sample—oversampled low income households.

[12] Focusing on father-son persistence in status rather than parent-child (or mother-daughter, etc.) is more straightforward given female labour force participation patterns, and the resulting issues with defining and measuring earnings and occupation outcomes. The surnames work also focused primarily on patrilineal lines of inheritance, given naming conventions (Clark, 2014, p.15), but still posited this general *law*. Hence, the proposed *law* of mobility should be just as evident using only fathers and sons as would be the case if mothers or daughters were included.

My analysis sample is comprised of sons who were members of the original 1968 sample and are male heads of their household in the 1992 survey, restricted to those who were born in 1951-1961. The lower bound on birth year ensures that the sons were 17 years of age or younger in 1968, avoiding selecting older children still living at home. Further, the sons' birth year restrictions minimize life-cycle bias in annual earnings by ensuring that sons are 30 to 40 years old for the 1991 earnings measure (reported during the 1992 survey).[13] Fathers are identified as the male heads of the household in which the son lived in 1968. The earnings outcome for both fathers and sons is measured as log annual earnings, so the sample excludes any observations with non-positive earnings or earnings which were imputed by major assignment (for sons, this refers to earnings in 1991, and for fathers, earnings in each of the years 1967–71). Fathers missing data on educational attainment are also excluded. The earnings exclusions apply to 24 sons and 28 fathers, with 11 additional fathers excluded due to missing education, amounting to excluding a total of 46 father-son pairs, and leaving a final sample of 415 sons from 293 fathers.[14]

Table A1 provides summary statistics describing this sample. The sample is predominately white, with only five percent black. Given the age exclusions for sons (and lack thereof for fathers), the fathers are observed, on average, at an older age than sons, with fathers' average age just over 40 in 1967 and sons' average age approximately 35 in 1991. Average annual earnings are slightly lower for sons than fathers, and are also more variable for sons, consistent with the well-documented life-cycle profile in earnings.[15] Approximately 25 percent of the fathers have at least a four-year college degree.

For the empirical analysis, I define the education measure of father's latent status as father's educational attainment as of the 1968 survey, coded as 1-16 for years of schooling up to a 4 year

[13] Haider & Solon (2006) show that the measurement error in men's current earnings as an indicator of lifetime earnings is non-classical at younger and older ages, causing intergenerational persistence estimates to be biased downward (as also illustrated in Lee and Solon (2009)). They find that observing men's earnings from the early thirties through the early forties best avoids this life-cycle bias, as this is when the measurement error is approximately classical. Findings presented by Nybom & Stuhler (*forthcoming*) show similar results using Swedish earnings data.

[14] It is possible to construct larger PSID samples, but I choose a sample similar to those in prior intergenerational studies since these were used to produce the U.S. estimates which Clark purports are biased downward, and are thus germane to the explorations in this paper. Further, Nybom and Vosters (2015) use Swedish administrative data to conduct similar tests as well as supplementary analyses examining the robustness of the results in this paper, showing that the results are not unique to this sample or the measures used.

[15] All earnings variables are expressed in 1991 dollars (adjusted for inflation using the CPI-U) for illustration purposes, but this transformation does not affect IGC estimates since the log of earnings is being used.

college degree, with a value of 18 indicating any graduate school completed. The occupation measure refers to the main job discussed in the 1969 survey, and is incorporated in the form of occupational category indicators. As listed in Table A1, there are seven categories: 1) professional, technical; 2) managers, businessmen, self-employed; 3) clerical, sales; 4) craftsman, foreman; 5) operatives; 6) labourers, service workers, farmers, and farm managers; 7) miscellaneous (includes armed services members, protective services workers, those not currently employed, and those missing an occupation category). To further illustrate the composition of the occupation categories for fathers, Table A2 provides average education and earnings by category. Average earnings and education are generally monotonically decreasing from occupation categories 1 to 6. The final category, 7–miscellaneous, is similar to categories 3 and 4, though with few observations and substantial variability in earnings. Hence, I take a flexible approach in the analysis, incorporating the occupation measure as a vector of indicators for each of the first six occupation categories (with category 7 the omitted reference group), taking no stance on the relative social status of the categories, but rather assuming that each contains some information on the underlying latent status.

1.3 Empirical Approach

To test the hypothesis that traditional estimates of intergenerational persistence suffer from attenuation bias, I begin by providing a baseline traditional estimate from this PSID sample. I use the five-year average of log earnings from 1967-71 as the measure of father's status in equation (1), similar to previous studies (e.g., Solon, 1992; Zimmerman, 1992; Chetty et al., 2014).[16] Given that the proposed attenuation bias is thought to come from the focus on a single noisy measure of an underlying latent social status, and that incorporating additional measures such as education and occupation should reveal greater persistence in status, I extend the model by adding these other measures of father's status. I then estimate these intergenerational regressions using the typical ordinary least squares (OLS) approach, an instrumental variables (IV) approach, and the approach

[16] With classical noise in annual earnings measures, estimating equation (1) using OLS results in an IGC estimate that is biased downward by the well-known attenuation factor of $\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$, where σ_x^2 is var(x^*) and σ_u^2 is var(u). Taking the five-year average of earnings mitigates the attenuation bias, reducing the attenuation factor to $\frac{\sigma_x^2}{\sigma_x^2+(\sigma_u^2/5)}$. The attenuation factor becomes more complicated when one incorporates serial correlation in earnings from one year to the next.

proposed by Lubotsky & Wittenberg (2006), to look for evidence of attenuation bias. In all estimations, I control for a quadratic in father's age and a quadratic in son's age to account for the life-cycle profile in earnings.[17]

To more clearly illustrate Clark's theory discussed above in the context of my empirical approach, I present a more formal latent variables framework, with the intergenerational equation (1) now represented by the so-called structural equation:

$$y_{it+1} = \beta x^*_{it} + \epsilon_i \tag{4}$$

where y_{it+1} is son's log earnings, and x^*_{it} is father's underlying social status. Then we can consider equation (3) expanded to comprise the system of measurement equations:

$$y_{1it} = \rho_1 x^*_{it} + u_{1it} \tag{5}$$

$$y_{2it} = \rho_2 x^*_{it} + u_{2it} \tag{6}$$

$$\vdots$$

$$y_{jit} = \rho_j x^*_{it} + u_{jit} \tag{7}$$

In these measurement equations, y_{1it} represents the average of father's log annual earnings in 1967-71, y_{2it} is father's education, and y_{3it} is father's occupation (specifically, a vector of occupation category indicators).[18] Further, this framework allows for slope coefficients in the measurement equations, relaxing the theory presented earlier, which took these ρ_j to be equal to 1.[19]

This notation reflects the fact that I do not directly address the latent status for sons. If we were to take literally the simple law's assumption of classical measurement error on the left-hand side, there would be no concern of this limitation inducing bias. More generally, with any status measure on the left-hand side, we should still see growth in the intergenerational coefficient

[17] Including quadratics in both father's and son's age as controls arises from taking models of current earnings of the form $y_{it} = y_i + a_{i0} + a_{i1}Age_{it} + a_{i2}Age^2_{it} + v_{it}$, for i = father or son and t = time (e.g., year), then solving for the long run component of earnings y_i, and substituting each into equation (1). Taking the five-year average of log earnings implies using the five-year average of age. See Solon (1992) for explicit derivations.

[18] The occupation indicators are generally referred to as one measure—occupation—even though occupation is flexibly accounted for by including an indicator for each occupation category. This implementation is similar to the drinking water proxy for wealth used in one of the examples presented in Lubotsky and Wittenberg (2006).

[19] Intercepts are omitted because the outcome, measures, and latent variable should all be considered to be demeaned, which is consistent with the implementation of the Lubotsky and Wittenberg (2006) approach discussed below.

towards 0.75 as we add measures for fathers on the right-hand side if the proposed attenuation bias argument holds. Further, addressing status for sons is tricky, as there is no basis for obtaining optimal weights (discussed below) for son's measures on the left-hand side. Even so, I perform a robustness check by applying the weights determined for fathers' measures to those for sons to obtain a more comprehensive status measure for sons, and get very similar results.[20]

Also under the perhaps unwise assumptions of classical measurement error, one method for consistently estimating β is instrumental variables (IV). One can use any y_j to instrument for another measure y_k and consistently estimate β, provided that $\sigma_{jk} \equiv cov(u_j, u_k) = 0$ and $\rho_k = 1$ (otherwise, the estimate converges to β/ρ_k). Hence, this IV approach is slightly robust to failure of classical assumptions, allowing some $\rho_j \neq 1$. In the case where $\sigma_{jk} = 0$ fails, the IV estimator is no longer consistent for β, but the direction of bias may be intuitively inferred based on belief about the sign of $cov(u_j, u_k)$.[21] Although Clark's simple law assumes the measurement errors are uncorrelated (i.e., $cov(u_j, u_k) = 0$) there are obvious reasons to believe this assumption is violated in the setting considered here.[22] Thus, I next turn to my preferred approach which allows for this correlation.

The approach proposed by Lubotsky & Wittenberg (2006) (henceforth LW), not only produces a single estimate of β while incorporating multiple measures, but does so in an optimal way such that the estimate asymptotically provides the greatest lower bound on β. The approach results in the least attenuation bias by extracting the strongest combined signal out of all of the measures.[23] Hence, I can directly test the attenuation bias argument by observing whether estimates are converging to the hypothesized persistence rate of 0.75 as I incorporate additional noisy measures of father's status. Not only does the method allow for incorporating all available measures, it also relaxes the strong assumptions that $cov(u_j, u_k) = 0$ for all $j \neq k$ and $\rho_j = 1$ for all j, allowing these

[20] As discussed below with the results on robustness checks, the intergenerational coefficient obtained from this regression based on using income, education, and occupation for sons and fathers is 0.433, which is not significantly different from the estimate of 0.473 based on only income for sons.

[21] When $\rho = 1$, β_{IV} converges to $\beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{jk}}$, where σ_x^2 is var(x^*), implying upward bias if $\sigma_{jk} < 0$ or downward bias if $\sigma_{jk} > 0$. When $\rho \neq 1$, β_{IV} converges to $\beta \frac{\rho_j \sigma_x^2}{\rho_k \rho_j \sigma_x^2 + \sigma_{jk}}$, with a more complicated inconsistency factor.

[22] For example, it is plausible that an idiosyncratic shock may affect father's income and occupation, inducing correlation among the measurement errors. Allowing for unrestricted correlation among the measurement errors thus permits error structures that contain a common factor, so that shocks may affect all observable measures for an individual.

[23] Specifically, the LW estimate achieves a greatest lower bound among a class of estimators, but other estimates can simply be mapped into this class for comparing magnitudes.

to be mostly unrestricted (subject to a normalization on ρ).[24]

The normalization on ρ is needed to identify this vector of slope coefficients in the measurement equations. I normalize ρ_1 to equal 1, which simply sets the scale of the latent x^* to that of y_1 (earnings). Clearly latent status has no scale, but given that I am positioning this paper using intergenerational income regressions as the point of departure, the natural normalization to adopt is to the scale of father's income. With this normalization, the equation for the remaining ρ_j can be shown to be:

$$\rho_j = \frac{cov(y_{it+1}, y_{jit})}{cov(y_{it+1}, y_{1it})} \tag{8}$$

This ratio can be estimated directly using IV estimation, instrumenting for y_{1it} (father's income) using y_{it+1} (son's income), with y_{jit} (the measure we are estimating ρ_j for) as the dependent variable. LW show that an auxiliary ordinary least squares regression of y_{it+1} on the measures y_{1it}, y_{2it}, ..., y_{jit}, produces the vector of coefficient estimates, $\hat{\phi}$, which provides information on the noisiness of the measures and on the conditional covariance of each measure with y_{it+1} (conditional on the other measures). Then, these coefficient estimates, $\hat{\phi}$, combined with the estimates, $\hat{\rho}$, form an optimal linear combination of the information from the j measures.[25] This optimal linear combination provides a greatest lower bound on β. Explicitly, the LW estimator is[26]:

$$\beta_{LW} = \hat{\rho}_1\hat{\phi}_1 + \hat{\rho}_2\hat{\phi}_2 + \cdots + \hat{\rho}_j\hat{\phi}_j \tag{9}$$

To control for other covariates (namely the quadratics in father's and son's age), these covariates are included in the auxiliary regression of son's earnings, y_{it+1}, on father's status measures, y_{1it}, y_{2it}, ..., y_{jit}, (to obtain ϕ) as well as in the IV estimations of the ρ_j.[27] Standard errors for the β_{LW} estimates are bootstrapped with 1,000 repetitions, using a block/panel bootstrap to account

[24] The approach assumes that $cov(u_j, \epsilon) = 0$, although very small deviations from this will not substantially alter the results.

[25] Linearity is adopted throughout this discussion and is relied upon for the LW approach, but this is a reasonable approximation for the measures considered here and the hypothesis being examined.

[26] Note that each element of β_{LW} (i.e., $\rho_j\phi_j$) can be considered as a product of ratios $\frac{cov(y_{it+1},y_{jit})}{cov(y_{it+1},y_{1it})}\frac{cov(\widetilde{y}_{it+1},\widetilde{y}_{jit})}{var(\widetilde{y}_{jit})}$, where $\frac{cov(\widetilde{y}_{it+1},\widetilde{y}_{jit})}{var(\widetilde{y}_{jit})}$ is conditional on the other measures in y_t, so the estimated $\hat{\beta}_{LW}$ will be monotonically increasing in magnitude as measures are added only in the case where the conditional covariance has the same sign as the unconditional covariance.

[27] This implementation strategy is theoretically (and numerically) equivalent to that suggested in Lubotsky & Wittenberg (2006)—to first regress each measure and the dependent variable on the other covariates and use these residualized variables for estimation of ρ and ϕ.

for clustering within family.

1.4 Results

1.4.1 Main Results

First, I establish a baseline estimate using the traditional approach with this PSID sample of 415 father-son pairs. Next, I explore the sensitivity of this estimate to including other measures of status in the regression. Panel A of Table A3 provides the results from this series of ordinary least squares (OLS) regressions. The baseline estimate of the intergenerational coefficient (IGC) is 0.439, using the traditional approach of regressing son's log earnings on the five-year average of father's log earnings (hence this can also be interpreted as an income elasticity). As expected, this is in the range (0.4–0.6) of traditional IGC estimates for the U.S. (Solon, 1999; Black & Devereux, 2011). Moving along columns 2-4 of Table A3, I present the OLS results from the augmented models. Adding education to the model, the coefficient on father's earnings falls slightly to 0.398, but the coefficient on education is essentially zero. Similarly, when I add occupation categories instead of education, the coefficients on these occupation category indicators are not jointly significant (F=0.54, p-value=0.779); in this case, however, the coefficient on earnings rises slightly to 0.480. When education and occupation are both incorporated, the coefficient on earnings is similar to the baseline estimate. Again, neither the coefficient on education nor the coefficients on the occupational category indicators (F=0.66, p-value=0.682) are significant.[28]

The next panel in Table A3 shows the results from an IV approach, which is commonly used to

[28] Similar to my results, other studies also find that when the variable used for the parent is the same as that used for the offspring in the dependent variable, then additional variables for the parents do not have practically or statistically significant coefficients. Sewell & Hauser (1975, p.86) find this result in analysis based on the Wisconsin Longitudinal Study. With son's earnings as the dependent variable, they note that the coefficients on father's education and occupation are not statistically significant after conditioning on father's income. Using the PSID, Corcoran et al. (1992) also use son's earnings as the dependent variable and similarly find that after accounting for parental income, the coefficients for several other family or community background characteristics are not practically or statistically significant. Duncan et al. (2005) find similar results for intergenerational associations for 17 outcome measures (traits and behaviors) in the National Longitudinal Survey of Youth (NLSY). After accounting for the same measure for parents, the coefficients on the other trait or behavioral measures are not statistically significant in 84 percent of the 272 cases. Further, two very recent studies find this result using large administrative datasets: Boserup, Kopczuk, & Kreiner (2014) estimate the wealth elasticity in Denmark, and upon adding parental and child income find that these coefficients are not practically significant; Nybom & Vosters (2015) perform analyses analogous to those in this paper using Swedish administrative data and show that, with son's income as the dependent variable, after conditioning on father's income the coefficient on father's education is not practically or statistically significant.

address classical measurement error. With two "noisy" measures of status (earnings and education) I use education to instrument for earnings.[29] The estimated IGC is 0.497, which still falls in the range of traditional estimates for the U.S. and does not indicate substantial attenuation bias in the baseline estimate.

Finally, in Panel C, I present the estimates of the intergenerational persistence coefficient obtained using the LW approach to minimize the attenuation bias from using multiple noisy measures of status. All of the IGC estimates themselves are statistically significant, so I focus the discussion on changes in the estimates across specifications. The first estimate is simply the OLS estimate (0.439), as this is a special case of the LW approach when one uses a single measure. Adding father's education as an additional measure of status produces only a slight increase in the estimated IGC to 0.445. When occupation information is added instead of education, the IGC estimate is larger, at 0.465. And, when both education and occupation measures are simultaneously included, the IGC estimate increases slightly to 0.473, but again there is not a substantial increase in the estimated persistence.[30] Note that the OLS coefficient estimates presented in Panel A are identical to the auxiliary coefficient estimates, $\hat{\phi}_j$, used in the LW approach. Given the lack of practical or statistical significance of these estimates discussed above, it is unsurprising that we do not see large changes in the LW estimates of the intergenerational correlation. Attempting to incorporate additional information on social status causes the IGC to fluctuate some, but all estimates remain in the range of prior estimates for the U.S. Figure A1 shows that even when considering the precision of the estimates and looking at the 95 percent confidence intervals (the bars) around the estimates (the dots), neither indicate IGC estimates increasing to the hypothesized underlying persistence rate of 0.75. The plots show the estimates and confidence intervals for each specification listed in Table A3, beginning with the baseline estimate, then adding education, occupation, and both. The upper bounds on the confidence intervals are, respectively, 0.585, 0.585 0.622, and 0.629, still falling short of the hypothesized persistence rate. The precision of these estimates is hampered by the

[29] As noted above, instrumenting in this fashion produces an IV estimate that converges to β/ρ_1 where ρ_1 is the coefficient in the earnings measurement equation (and assuming the measurement errors are uncorrelated), thus enabling the comparability to our LW estimate based on latent status set to the scale of father's income.

[30] When a more flexible approach is taken using the five annual earnings years as separate variables as well as separate education category variables (high school graduate, some college, four-year degree, at least some graduate school), the LW estimate of intergenerational persistence is still quite similar at 0.485 but less precise with a standard error of 0.095.

PSID sample size, but Nybom and Vosters (2015) find strikingly similar—and more precise—results for Sweden, with similarly small increases in persistence estimates, even after using more detailed measures and incorporating analogous measures for mothers.

1.4.2 Robustness Checks

In the main analysis, I focus on adding measures of status for fathers, but do not directly address son's latent status. As discussed in the empirical approach section, this should not substantially alter the results. However, I still perform a sensitivity check in which son's latent status is explicitly addressed. I apply the weights determined by the LW approach for father's latent status to the measures for both fathers and sons, creating index measures of status for each generation. Then I regress the composite measure for sons on the composite measure for fathers. This results in an IGC estimate of 0.433 with a (bootstrapped) standard error of 0.071, which is similar to, albeit slightly smaller than, the main LW estimates reported in Table A3.

My LW results are also robust to adjusting several of the sample restrictions, as shown in Table A4. The first row of results, with the estimates in bold and standard errors in italics underneath, simply provides the main results from Panel C of Table A3 for comparison. Allowing sons who are 25-29 years old at their 1991 earnings measure to also be included in the sample, the sample grows to 582 father-son pairs (with sons aged 25-40 years old). The IGC estimates of 0.402–0.464 are slightly smaller relative to the main results, consistent with the life-cycle effects literature (Haider & Solon, 2006; Nybom & Stuhler, *forthcoming*), but the pattern of minimal increases remains unchanged as additional measures of status are included. The same pattern is revealed when instead of adjusting the restrictions on son's age, I do so for father's age, limiting the fathers to those aged 30-50 in 1968 and obtaining IGC estimates ranging 0.457–0.494. Incorporating both of these sample adjustments at the same time also produces the same pattern, as expected, which is shown in the next row of results with estimates ranging 0.420–0.452. Returning to the original sample restrictions, except now including mother-son pairs from female-headed 1968 households (so single mothers) in the sample, the IGC estimates are smaller in magnitude (0.360–0.410) but still follow the same pattern as more measures of status are added. Finally, the last row of estimates in Table A4 presents results from changing the functional form of the father's earnings measure from the *average of log earnings*

for 1967–71 to the *log of average earnings*. These results yet again exhibit the same pattern as the main results, with IGC estimates ranging 0.463–0.495.

1.5 Conclusions

Several recent studies by Gregory Clark and coauthors have examined intergenerational mobility using a new method based on surnames and newly developed datasets, finding higher persistence rates (i.e., lower mobility) than previously estimated (e.g., Clark, 2014; Clark & Cummins, 2015). In these studies, the hypotheses presented to explain the discrepancy use a simple measurement error argument that is consistent with the proclaimed higher persistence rate of approximately 0.75 from surname methods and the smaller estimates from traditional studies. I am the first to empirically test the proposition that prior estimates are attenuated from focusing on a single measure such as income and should rise when additional information is incorporated.

I use Lubotsky & Wittenberg's (2006) approach designed for scenarios such as this, where multiple measures of a latent variable (i.e., status) are available, but the measurement errors are likely correlated. The method combines the information from available measures of the latent variable in a way that produces a single persistence estimate with the least attenuation bias. I aggregate information from income, education, and occupation—three recommended measures of father's social status—using the LW method, yet I see no indication of the persistence rates approaching 0.75 as the additional measures are added. There are small increases in the persistence estimates as additional measures are incorporated, but these changes are not meaningful in a statistically significant or practical sense. In fact, all of the estimates presented in the main results, as well as in robustness checks, range from 0.360 to 0.491, quite similar to the prior estimates for the U.S. The pattern of small increases with additional measures is robust to adjusting sample restrictions as well as measure definitions. And, although my sample size is not conducive to assessing the statistical significance of these small changes in the point estimates, the sample I use facilitates relevant comparisons to prior literature. I am able to obtain a baseline estimate analogous to the prior studies about which the attenuation bias claims are made, which is an appropriate starting point for then incorporating information from other measures. I find no evidence that adding information from other status measures produces estimates that are converging

to a substantially greater level of intergenerational persistence.

My findings reject Clark's measurement error interpretation of his results relative to those from prior literature, but they do not shed light on why his estimates based on surnames are higher than traditional estimates. Averaging over surnames does not always produce higher persistence estimates, as shown with U.S. income tax data in Appendix D of Chetty et al. (2014). Further work is needed to gain a more nuanced understanding of discrepancies between Clark's estimates using the surname-average method and traditional methods, and what each method might be identifying. As noted by Solon (2015) and Chetty et al. (2014), the traditional approach may be correctly identifying individual-level mobility, while the surnames method may be identifying group-level mobility for these particular groups of surnames. This is further developed in a recent exposition by Torche & Corvalan (2015), which shows that estimating surname-level regressions captures between-group persistence in average outcomes for the particularly "elite" or "underclass" surnames chosen, rather than Clark's interpretation of using group averages to eradicate measurement error and reveal individual-level mobility.

APPENDIX

Table A1: Summary Statistics for Analysis Sample

	Mean	*Std. Dev.*	*Min*	*Max*
Race - black	0.05	0.22	0	1
Sons' age in 1991	34.92	3.14	30	40
Sons' 1991 individual earnings	35,695	26,251	300	335,000
Fathers' age in 1967	40.47	6.81	27	67
Fathers' Individual earnings				
Annual earnings 1967	39,684	24,409	1,101	244,671
Log annual earnings 1967	10.43	0.60	7.00	12.41
5-year-avg of log earnings, 1967-71	10.46	0.59	7.79	12.65
Fathers' Educational attainment				
Less than HS graduate	0.33	0.47	0	1
High school graduate	0.32	0.47	0	1
Some college	0.11	0.31	0	1
Bachelor's degree	0.14	0.34	0	1
At least some graduate school	0.11	0.31	0	1
Fathers' 1969 Occupation categories				
1 - Professional, technical	0.23	0.42	0	1
2 - Manager/businessmen	0.14	0.35	0	1
3 - Clerical, sales	0.09	0.29	0	1
4 - Craftsman, foreman	0.23	0.42	0	1
5 - Operatives	0.17	0.38	0	1
6 - Laborers, service, farmers	0.12	0.32	0	1
7 - Not currently employed/missing	0.02	0.15	0	1

Notes. The sample includes 415 sons and 293 fathers. All earnings are expressed in 1991 dollars.

Table A2: Fathers' Average Earnings and Education by Occupation Category

	Earnings in 1969		Educational attainment		
Occupation Category	*Mean*	*Std. Dev.*	*Mean*	*Std. Dev.*	*N*
1 - Professional, technical	61,382	40,129	15.67	1.76	66
2 - Manager/businessmen	54,983	41,871	12.83	2.73	42
3 - Clerical, sales	38,379	10,920	12.96	1.89	27
4 - Craftsman, foreman	38,212	15,203	10.79	2.62	67
5 - Operatives	30,044	11,108	9.76	2.53	50
6 - Laborers, service, farmers	20,614	10,468	9.97	2.55	34
7 - Not employed or missing	38,379	31,544	10.00	3.61	7
Overall	42,419	30,209	12.09	3.27	293

Notes. The sample includes 293 fathers. All earnings are expressed in 1991 dollars.

Table A3: OLS, IV, And LW Results

Fathers' noisy measures of status	[1] Earnings	[2] Earnings, education	[3] Earnings, occupation	[4] Earnings, education, occupation
Panel A: OLS results				
Five-year average of log earnings: 1967-71	0.439	0.398	0.480	0.438
	0.075	*0.098*	*0.100*	*0.120*
Educational attainment		0.010		0.016
		0.013		*0.016*
Occupation categories				
1 - Professional, technical			0.002	-0.077
			0.228	*0.236*
2 - Manager/businessmen			-0.029	-0.064
			0.233	*0.222*
3 - Clerical, sales			0.001	-0.051
			0.256	*0.258*
4 - Craftsman, foreman			0.066	0.052
			0.233	*0.218*
5 - Operatives			-0.027	-0.032
			0.229	*0.211*
6 - Laborers, service, farmers			0.181	0.152
			0.254	*0.244*
Panel B: IV results (education to IV for 5-yr-avg earn)				
First stage	0.105			
	0.006			
Second stage	0.497			
	0.090			
Panel C: LW estimates of IGC				
	0.439	0.445	0.465	0.473
	0.075	*0.072*	*0.080*	*0.080*
N	415	415	415	415

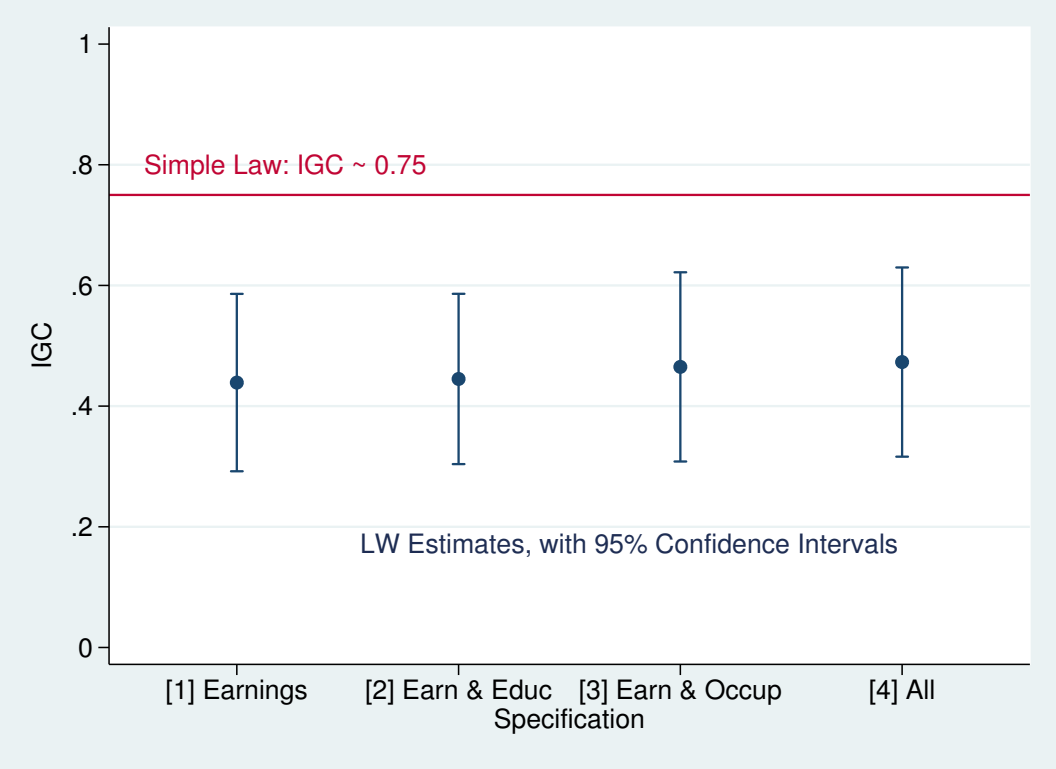
Notes. All specifications use log of son's 1991 earnings as the dependent variable and include as controls a quadratic in son's earnings and a quadratic in father's age (the average age during the five years of earnings observations). The omitted occupation category is "7 - Not employed or missing". The sample size for all estimations is 415 father-son pairs, from 293 families. OLS and IV standard errors are clustered by family. LW standard errors are computed using a block bootstrap to account for within family correlation (1,000 repetitions).

Table A4: Robustness of LW Results

	N	[1] Earnings	[2] Earnings, education	[3] Earnings, occupation	[4] Earnings, education, occupation
Main results	415	0.439	0.445	0.465	0.473
		0.075	*0.072*	*0.080*	*0.080*
Adjusting sample exclusions:					
Son's age 25-40	582	0.402	0.422	0.446	0.464
		0.065	*0.061*	*0.075*	*0.077*
Father's age 30-50	380	0.457	0.463	0.484	0.494
		0.083	*0.083*	*0.090*	*0.089*
Son's age 25-40 and Father's age 30-50	483	0.420	0.426	0.444	0.452
		0.076	*0.073*	*0.083*	*0.082*
Include 1968 female-headed households	444	0.360	0.375	0.392	0.410
		0.072	*0.064*	*0.072*	*0.069*
Adjusting earnings measure:					
Log of father's 5-yr avg. of annual earnings 1967-71	415	0.463	0.466	0.490	0.495
		0.075	*0.073*	*0.081*	*0.081*
Father's status measures					
Earnings (5-yr-avg)		x	x	x	x
Educational attainment			x		x
Occupational categories				x	x

Notes. The dependent variable is log of son's 1991 earnings, and the measure of father's earnings is the 5-year average of log earnings from 1967-71. All specifications include as controls a quadratic in son's earnings and a quadratic in father's age (the average age during the five years of earnings observations). The omitted occupation category is "7 - Not employed or missing". Standard errors are computed using a block bootstrap to account for within family correlation (1,000 repetitions).

Figure A1: LW Results



Notes. The sample includes 415 fathers and 293 fathers. Standard errors are computed using a block bootstrap to account for within family correlation (1,000 repetitions).

REFERENCES

REFERENCES

Becker, G. & Tomes, N. (1976). Child endowments, and the quantity and quality of children. *Journal of Political Economy*, 84(4)2, S143-S162.

Becker, G. & Tomes, N. (1979). An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy*, 87, 1153-189.

Björklund, A., & Salvanes, K. G. (2011). Education and family background: Mechanisms and policies, in E. Hanushek, S. Machin, and L. Woessmann (eds.), *Handbook of the Economics of Education*, 3(3), 201-247.

Black, S. E. & Devereux, P. J. (2011). Recent developments in intergenerational mobility, in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, 4, 1487-1541, Amsterdam: Elsevier.

Braun, S., & Stuhler, J. (2015). The transmission of inequality across multiple generations: Testing recent theories with evidence from Germany. Mimeo.

Boserup, S.H., Kopczuk, W., & Kreiner, C.T. (2014). Intergenerational wealth mobility: Evidence from Danish wealth records of three generations. Working paper, October 2014.

Chetty, R., Hendren, N., Kline, P. & Saez, E. (2014). Where is the Land of Opportunity? The geography of intergenerational mobility in the United States, *Quarterly Journal of Economics*, 129(4), 1553-1623.

Clark, G. (2014). *The son also rises: surnames and the history of social mobility.* Princeton University Press.

Clark, G. and Cummins, N. (2015). Intergenerational wealth mobility in England, 1858-2012. Surnames and social mobility. *Economic Journal*, 125, 61-85.

Clark, G., Cummins, N., Hao, Y. & Vidal, D.D. (2015). Surnames: a new source for the history of social mobility. *Explorations in Economic History*, 55, 3-24.

Corcoran, M., Gordon, R., Laren, D., & Solon, G. (1992). The association between men's economic status and their family and community origins. *Journal of Human Resources*, 27(4), 575-601.

Duncan, G., Kalil, A., Mayer, S. E., Tepper, R., and Payne, M. R. (2005). The apple does not fall far from the tree, in Samuel Bowles, Herbert Gintis and Melissa Osborne Groves (eds.), *Unequal Chances: Family Background and Economic Success*, pp.23-79, Princeton University Press.

Güell, M., Rodríguez Mora, J. V. & Telmer, C. (2014). The informational content of surnames, the evolution of intergenerational mobility and assortative mating. *Review of Economic Studies*, Advance Access published online December 10, 2014, doi:10.1093/restud/rdu041.

Haider, S. J. & Solon, G. (2006). Life-cycle variation in the association between current and lifetime earnings. *American Economic Review*, 96(4), 1308-1320.

Hertz, T., Jayasundera, T., Piraino, P., Selcuk, S., Smith, N., & Verashchagina, A. (2007). The inheritance of educational inequality: International comparisons and fifty-year trends. *The BE Journal of Economic Analysis and Policy*, 7(2).

Lee, C. I. & Solon, G. (2009). Trends in intergenerational income mobility. *The Review of Economics and Statistics*, 91(4), 766-772.

Long, J., & Ferrie, J. (2007). The path to convergence: Intergenerational occupational mobility in Britain and the U.S. in three eras. *The Economic Journal*, 117(519), C61-C71.

Long, J., & Ferrie, J. (2013). Intergenerational occupational mobility in Great Britain and the United States since 1850. *The American Economic Review*, 103(4), 1109-1137.

Lubotsky, D. & Wittenberg, M. (2006). Interpretation of regressions with multiple proxies. *The Review of Economics and Statistics*, 88(3), 549-562.

Lucas, R. E., & Kerr, S. P. (2013). Intergenerational income immobility in Finland: Contrasting roles for parental earnings and family income. *Journal of Population Economics*, 26(3), 1057-1094.

Mazumder, B. (2005). Fortunate sons: New estimates of intergenerational mobility in the United States using social security earnings data. *The Review of Economics and Statistics*, 87(2), 235-255.

Nybom, M. & Stuhler, J. (*forthcoming*).Heterogeneous income profiles and life-cycle bias in intergenerational mobility estimation. *Journal of Human Resources*.

Nybom, M. & Vosters, K. (2015). Intergenerational persistence in latent socioeconomic status: Evidence from Sweden. SOFI Working Paper 3/2015.

Panel Study of Income Dynamics, public use dataset. Produced and distributed by the Institute for Social Research, University of Michigan, Ann Arbor, MI (accessed Dec 2013).

Sewell, W. H., & Hauser, R. M. (1975). *Education, occupation, and earnings. Achievement in the early career.* New York: Academic Press.

Solon, G. (1992). Intergenerational income mobility in the United States. *The American Economic Review*, 82(3), 393-408.

Solon, G. (1999). Intergenerational mobility in the labor market', in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, 3A, pp. 1761-1800, Amsterdam: North-Holland.

Solon, G. (2015). What do we know so far about multigenerational mobility? NBER Working Paper No. w21053.

Torche, F. & Corvalan, A. (2015). Estimating intergenerational mobility with grouped data: a critique of Clark's *The Son Also Rises*. Working Paper 15-22, NYU Population Center.

Zimmerman, D. J. (1992). Regression toward mediocrity in economic stature. *The American Economic Review*, 82(3), 409-429.

Chapter 2

Intergenerational Persistence in Latent Socioeconomic Status: Evidence from Sweden[1]

2.1 Introduction

Researchers and policymakers have long shown a great deal of interest in understanding the degree of socioeconomic mobility within and across societies, resulting in a large body of economic research examining the extent to which income differences are passed on from parents to their children. One of this literature's most notable results is that intergenerational mobility in the Nordic countries is substantially higher than in countries such as the United States. However, recent work by Gregory Clark and coauthors has led to surprisingly contrary conclusions, suggesting that the "true" rate of mobility is generally very low and also steady across time and countries with vastly different social and economic contexts, including Sweden and the United States (Clark 2014, p.107).

The descriptive literature on intergenerational income mobility generally estimates an equation resembling a basic AR(1) process:

$$y_{it+1} = \beta y_{it} + \varepsilon_i, \qquad (1)$$

where y_{it+1} is offspring log income of family i, y_{it} is parental (typically fathers') log income, and ε_i an idiosyncratic error; β is then interpreted as the intergenerational elasticity.[2] This process is not necessarily taken literally, nor is the estimate believed to be causal, but instead the goal is to obtain a

[1] This chapter is coauthored with Martin Nybom from the Institute for Labor Market and Education Policy Evaluation (IFAU) and the Swedish Institute for Social Research (SOFI), Stockholm University.

[2] This parameter thus measures persistence, whereas $1-\beta$ is a measure of mobility. For this equation and all those that follow, variables are considered in deviation-from-mean form, allowing intercepts to be suppressed.

summary statistic describing how differences in economic status persist from one generation to the next. For Sweden, the estimated persistence in income is around 0.2-0.3, compared to 0.4-0.6 in the U.S. (see Solon, 1999, Björklund & Jäntti, 2009, Black & Devereux, 2011). The greater mobility in the Nordic countries is often attributed to policy differences, such as more redistributive tax structures, which facilitate public human-capital investments in terms of subsidized pre-school and college education.[3] Others point out that characteristics of the labor market also matter, such as differences in the returns to skills and the intergenerational transmission of employers (Björklund et al., 2012; Corak & Piraino, 2011). However, Clark (2014, p.5) follows the former argument, boldly interpreting the low and constant rates of mobility as evidence of large policy failure.

The creative methods underlying this recent work exploit the information content of *uncommon* surnames in lieu of actual intergenerational family links, and the results paint an extraordinarily different picture of mobility for Sweden as well as for other countries.[4] The persistence rate for underlying status is estimated to be as high as 0.7-0.8 across a wide range of societies and time periods, leading to the conclusion that for Sweden, "The implied social mobility rates are as low as those of modern England or the United States" (Clark 2014, p. 41).

These claims are quite controversial, with important implications for the interplay between policy and socioeconomic mobility. They also clearly contradict conclusions from prior intergenerational studies. Acknowledging this incongruity, Clark and coauthors suggest that conventional methods have been limited in their measures of socioeconomic status. The main argument is that families have a *general social status* that underlies imperfect status measures such as income, education, or occupation, and these measures are linked to this underlying and unobserved

[3] Public investment in children's human capital is put forth as one of the key determinants of the size of the reduced-from intergenerational income relationship in Solon's (2004) log-linear version of Becker and Tomes' (1979) model of parental investments.

[4] Güell et al. (2015) show that rare surnames do contain such information, developing a different method using the joint distribution of surnames and economic status to examine intergenerational transmission of status in Spain.

latent factor with substantial random components. Formalized into a simple model, social mobility is reduced to a universal *law of mobility*, $x_{it+1} = bx_{it} + e_{it}$, where x_{it} is the underlying status of family *i* in generation *t* (Clark, 2014). A single measure such as income is then assumed to be related to status with some additive random noise, $y_{it} = x_{it} + u_{it}$, whereby substituting this into the conventionally estimated equation (1) leads to the classical errors-in-variables attenuation bias. Averaging within a surname, *z*, then reveals true status, $\bar{y}_{zt} = \bar{x}_{zt}$, as $\bar{u}_{zt}$ is approximately zero for large enough surname groups. For data without surnames, Clark & Cummins (2015) propose that if the information from multiple measures—for example, income, education, and occupation—were combined for an individual, then conventionally estimated persistence would rise.

Applying an approach proposed by Lubotsky & Wittenberg (2006) to optimally aggregate information from multiple measures, Vosters (2015) tests this proposition using data from the Panel Study of Income Dynamics (PSID). The estimated persistence rates remain just under 0.5 and are insignificantly different from conventional estimates, even after accounting for multiple partial measures of underlying status. While this study shows that this approach does not substantially raise estimated persistence for the U.S., the question remains as to what information could be extracted from multiple status measures in a country with a more redistributive welfare state, such as Sweden. In fact, according to Clark's hypothesis, this approach should have a *greater* impact on estimated persistence in settings where persistence is conventionally estimated to be quite low.

Therefore, we follow the above approach, performing similar tests to look for any evidence of this asserted attenuation bias in conventional estimates for Sweden. We first provide estimates using measures constructed to take full advantage of the rich Swedish data. We construct nearly career-long income measures, which mitigates biases stemming from transitory fluctuations (Mazumder, 2005) and life-cycle effects (Haider & Solon, 2006; Nybom & Stuhler, *forthcoming*) in short-run incomes. Our data also have more detailed occupation categories available, allowing us to better

examine the degree to which information on status can be extracted from an individual's occupation. Moreover, the small sample size in Vosters (2015) yields low statistical precision and only very large attenuation biases can be formally rejected. In contrast, our sample consists of more than 167,000 parent-child pairs, which provides much greater statistical power. We also examine the claim that persistence is uniform across countries. For this, we provide estimates using variables constructed similarly to those based on the PSID, facilitating a test of whether persistence in underlying status is indeed of equal magnitude in Sweden and the U.S. In doing so, we also indirectly address implications of various data limitations with the U.S. data, such as inaccurate measurement of long-run income and occupations. As such, not only do we obtain results comparable to those for the U.S. to evaluate the applicability of the *simple law of mobility* across countries, we also obtain more robust results on the magnitude of the hypothesized attenuation bias in the Swedish estimates.

We find no evidence to support the *simple law of mobility*, as persistence estimates remain around 0.25-0.30 even after multiple measures are combined. Further, our comparison with the U.S. confirms the prior perception that mobility is indeed substantially higher in Sweden. These results are robust across a variety of specifications and methods for constructing the measures, and the country difference in persistence appears even greater when using measures constructed to mimic those used for the comparable U.S. study. Our findings thus support those in Vosters (2015), as well as the discussion in Chetty et al. (2014), suggesting that the very low mobility rates provided by the surname approach strongly underestimate the degree of mobility in the population as a whole.[5]

Although much of our evidence reaffirms results from existing literature rather than lending support to Clark's conclusions, we do find that the latent status framework of the *simple law* can be empirically relevant for certain groups (e.g., mothers). Motivated primarily by the concept of *family*

[5] See also Braun & Stuhler (2015), who discuss the surname approach in the context of mobility across multiple generations.

status, we add the analogous status measures for mothers as we incorporate those for fathers. We find mothers' occupation to be the most important addition, though producing only a nominal rise in persistence. Further exploring this result, we examine persistence with respect to mothers' status alone, finding that both the estimates of mother-son and mother-daughter relationships increase substantially when multiple measures are used, though rising from low levels compared to those typically found for father-child relationships. Still, these results highlight the unintentional implications of this framework for measurement issues specific to women, showing that combining multiple proxy measures can provide more informative estimates in cases where appropriate income data is not available.

Our paper thus extends the literature on the measurement of intergenerational mobility. To date, research has mostly focused on the measurement of specific status indicators, with the approximation of lifetime (or permanent) income being the prime example. Inspired by the work by Gregory Clark and others, we complement this research by providing new evidence on whether such status indicators themselves, even when accurately measured, suffice to capture a broader concept of socioeconomic status. Our findings imply that for men detailed measures of long-run income are indeed good proxies for latent status. In contrast, for women combining individual income information with occupation improves the measurement of status substantially. We also add to the large literature on cross-national mobility differences (e.g., Solon, 2002). The finding based on surnames data, that social mobility is constant across countries, is put into question by our results; these show a Sweden-U.S. country differential that is in line with previous income-based evidence.

The rest of the paper unfolds as follows. Section 2 describes the data, before Section 3 discusses our empirical approach. Section 4 presents our main findings. Section 5 presents our extension to both parents, and the intergenerational associations related to mothers and daughters. The final section offers some concluding remarks.

2.2 Data

2.2.1 Sources and Sample Selection

We use administrative data from various sources, which have been merged by Statistics Sweden using unique personal identifiers. A multigenerational register links children to their biological parents; censuses provide data on parents' occupation and education; income tax declarations for both parents and offspring provide data on total *individual* income. Our main sample is based on a random draw of 35 percent of all children born in Sweden 1951-1961 and their biological parents.[6] We restrict our analysis to these cohorts for a couple of reasons. Given the available income data, we can observe long-run prime-age incomes of both these offspring cohorts and their parents. Moreover, these are the cohorts used in Vosters (2015), so this selection further facilitates the comparison between our estimates for Sweden with those for the U.S.

2.2.2 Construction of Status Measures

For annual income, we use administrative data covering the years 1968-2007. The data are based on individual income-tax declarations and we define the income measure separately for fathers and mothers. Our measure includes income before taxes from all sources except means-tested benefits and universal child benefits. These data come with a number of advantages: they are almost entirely free from attrition and reporting error, pertain to all jobs, and are not censored.[7] For parents, we approximate log lifetime income by the average of log annual income over ages 30-60. For offspring,

[6] We exclude those with parents who were more than 40 years old at the birth of the child.

[7] In contrast to many other administrative data sources, our data are not censored (nor truncated) in the top of the income distribution. Further, the Swedish system provides strong incentives to declare some taxable income since doing so is a requirement for eligibility to most social insurance programs. Hence, we expect very little missing data in the low end of the distribution.

we construct measures of log lifetime income as the average of log income over ages 27-46. We require parents and offspring to have at least five non-missing annual income measures.[8] Throughout all analyses we control flexibly for parental and offspring birth year using cohort dummies.

While the tests we conduct focus on the measurement of "status", our income measure is particularly important because it also minimizes the potential for two well-known sources of bias in estimated intergenerational associations; bias arising from transitory shocks to income and from life-cycle effects. By using a long-run average of annual income observations, potential attenuation biases from transitory fluctuations are greatly reduced (Mazumder, 2005). In our sample, 88 percent of sons have 20 non-missing log income observations and 88 percent of fathers have at least 10 non-missing log income observations. Further, we measure income as long-run averages during mid-life in order to minimize so-called life-cycle bias (Nybom & Stuhler, *forthcoming*). While there are slightly fewer mid-life income observations for fathers, 91 percent of them have at least one annual income observation from before age 50.

We use occupation data from national censuses conducted every five years between 1960 and 1990. The occupational classification employed in the censuses builds on the Nordic Occupational Classification (NYK), which is based on the International Standard Classification of Occupations (ISCO). The NYK categorizes occupations according to the end result of the tasks and duties undertaken in the job. Hence, level of education and professional status are typically not considered in the categorization (Statistics Sweden, 2004). The classification has a hierarchical structure, allowing for analyses at different aggregation levels. Three-digit codes denote unique occupations, two-digit codes denote minor occupation groups and one-digit codes denote major occupation

[8] Missing income is rare in our sample, and such occurrences could be due to quite different reasons; individuals could be living abroad, they could fail to file their tax declaration, or it might arise due to coding errors.

groups. To fully exploit the available information, we use the unique occupation indicators in our main analyses, but also test the sensitivity of our results to using the broader classification levels.

We define a parent's occupation as the occupation he or she had in the 1970 census. Fathers in our sample are, on average, about 44 years old in 1970, so this census provides a good prime-age occupation measure. If occupation is missing in this census, however, we use the corresponding data from the 1975 or the 1980 censuses.[9] For those with occupation still not coded, we include indicators for missing and undefined in our main specifications to flexibly account for these special cases. Including missing and undefined as separate categories, the resulting sample holds 270 unique occupations classified into 61 minor occupation groups, or 12 major occupation groups. To demonstrate the nature of the classification, the major occupation groups are: 1) Professional work (arts and sciences); 2) Managerial work; 3) Clerical work; 4) Wholesale, retail, and commerce; 5) Agriculture, forestry, hunting, and fishing; 6) Mining and quarrying; 7) Transportation and communication; 8) Manufacturing; 9) Services; 10) Military/Armed forces; 11) Undefined; 12) Missing.

For parental education, we use data on final education in 1970 according to the data from Statistics Sweden's education register, which is based on a standard conversion translating each level into years of education. The measures of parental education reflect their highest educational attainment, with the levels including: less than nine years of primary school, nine years of primary school, two-year secondary school, three-year secondary school, less than three years of post-secondary school, three years or more of post-secondary school, and graduate school. We also perform a set of robustness tests in which we control for education more flexibly. First, we again use the above measure but now by including a dummy variable for each of the different levels. Second,

[9] Incorporating the later censuses is primarily beneficial in obtaining more accurate information on occupation for mothers, who are more likely to have missing data in 1970. Very few fathers have missing occupation in 1970.

we also exploit more detailed information on educational attainment from the same data source. In doing so, we include a large set of dummies reflecting length and *type* of education, distinguishing between various tracks within high school as well as a large number of different academic and vocational post-secondary educational categories.

Because we are exploring implications of aggregating the information on parental income, education, and occupation, we only include parent-child pairs for which the parents have non-missing information on all of these measures and the child has the requisite non-missing income measures. Table B1 provides descriptives for our resulting main sample of 167,552 sons matched to 153,920 fathers.

2.2.3 Alternative Measures for U.S. Comparison

We also construct alternative measures to facilitate a Sweden-U.S. comparison based on comparable findings in Vosters (2015). The analysis by Vosters is based on data from the nationally representative part of the Panel Study of Income Dynamics (PSID), which began with a sample of about 3,000 families in 1968. Importantly, the PSID includes family links and follows original sample members and their children over time. Fathers are identified as the male head of the household in which the child resided at the time of the initial survey, which does not necessarily represent a biological link. Thus, our Swedish sample differs slightly in that we use biological rather than cohabitating fathers.[10]

To enable a credible cross-country comparison, we construct alternative measures for Sweden that are analogous to those from the PSID. For offspring income, we use the log of annual income

[10] For approximately 95 percent of the sons in the Vosters (2015) PSID sample, the identified cohabitating father is in fact the biological father, so this difference is minor.

in 1991. Fathers' income is defined as the average of log income in 1968-72.[11] Our education measure is very similar, reflecting the highest level of attainment. For occupation, we use the major groups described above, which differ slightly but not much from the seven groups used in the PSID (see Vosters, 2015). To better match the last "residual" category in the PSID, we add missing and undefined occupations to our military/armed services category, resulting in 10 major categories for the Swedish sample. In the U.S. data, education and occupation are from the 1968-1969 surveys, while our corresponding data are from 1970. Although there are minor differences in some variable definitions across the two countries, they are marginal at most and should have very little effect on our results. The sample with non-missing data on these measures includes 146,783 sons matched to 135,020 fathers.

We provide descriptive statistics for both the full sample and this restricted U.S. comparison sample in Table B1. The samples are very similar across all observables. Sons' average income is slightly higher than that for fathers; in logarithmic form, these averages are 12.22-12.29.[12] Fathers' average education of just over 9 years, as well as the distribution among the various levels of attainment, is nearly identical across samples, as are the proportions in each occupation category. Professional work and manufacturing comprise much of the sample of fathers, with 19 and 38 percent in the respective categories.

2.3 Empirical Approach

Our empirical approach is designed to test the hypothesis that estimates of intergenerational persistence in socioeconomic status approach 0.7-0.8 as we add the proposed partial measures. We

[11] Since our income data start in 1968, this measure is marginally different from the U.S. data that are based on earnings in 1967-71.

[12] Income is provided in 2005 Swedish kronor (SEK).

then proceed to contrast our results with comparable estimates for the U.S. to also test the claim that persistence in latent status is the same across countries. We begin by obtaining a baseline estimate of persistence by estimating the usual intergenerational income equation above in (1). To gauge the degree of attenuation bias in this estimate, we then add the additional measures of parental status to this equation. Although this provides insights into the sensitivity of conventional estimates to accounting for other status measures, it does not provide a *single* estimate of persistence in underlying status that *combines* information from all measures.

Our preferred method, proposed by Lubotsky & Wittenberg (2006), estimates the persistence in *latent status*, aggregating the information in the included proxy measures. To better illustrate our methodological approach, we first present the hypothesis in a simple latent variables framework, writing measurement equations for each of the partial measures, y_{jit}, of the form:

$$y_{jit} = \rho_j x^*_{it} + u_{jit}, \qquad (2)$$

where *j* indexes the measure, *i* indexes family, and *t* generation. We generalize the measurement equations from the *simple law* to allow for slope coefficients, ρ_j. Our main empirical specifications include equations for y_{1it} for parental (e.g., fathers') income, y_{2it} for parental education, and y_{3it}-y_{kit} for the *k-2* parental occupation indicators. x^*_{it} is the unobserved latent status and the u_{jit} are the measurement errors. The so-called structural equation can then be written:

$$y_{it+1} = \beta x^*_{it} + \varepsilon_{it}, \qquad (3)$$

where β is the measure of intergenerational persistence in underlying latent status. This notation shows that we do not explicitly address offsprings' latent status with multiple partial measures.[13] However, the outcome variable we do use—a twenty-year average of annual incomes during mid-life—is likely one of the best *single* measures of socioeconomic status available. Further, the simple conditions underlying the *simple law* of mobility rely on the assumptions of a classical errors-in-variables model, under which measurement error on the left-hand side is innocuous.

Under the classical assumptions, the measurement errors (u_{jit}) are all uncorrelated, and the coefficients ρ_j are equal to one. In this simple case, there are several econometric methods available. For example, instrumental variables (IV) estimation using one measure to instrument for another is common solution. We provide one such estimate, using father's education to instrument for father's income, which under the proposed *law* should estimate persistence levels in the 0.7-0.8 range.[14] Other possible approaches include the MIMIC (multiple indicators, multiple causes) or LISREL frameworks (see, e.g., Jöreskog & Goldberger, 1975, and Bollen, 1989). More recently, Black & Smith (2006) propose a GMM estimator with potential efficiency gains. However, each of these approaches relies critically on the assumption of uncorrelated measurement errors, and we find this restriction to be particularly concerning in the setting considered here.[15] First, the nature of the suggested measures (income, education, and occupation) makes the case of zero correlation among measurement errors unlikely. Second, the anecdotal examples used to motivate the concept of

[13] To assess sensitivity to this choice, we performed two different tests. First, we created omnibus measures of status for fathers *and* sons (applying fathers' weights to sons' measures) and obtained an estimate of 0.237, which is nearly identical to the comparable estimate (0.238) using only fathers' measures. Second, we used average log incomes across same-sex siblings as measure of offspring status (excluding those without same-sex siblings from the sample). While baseline estimates and thus the scaling differ in the latter case, the estimated decrease in attenuation bias is very similar.

[14] Note that in this particular IV setup, consistency requires only the coefficient in the income measurement equation to equal one (and the measurement errors still being uncorrelated), which is not problematic as this is the normalization we adopt for our preferred approach described below. This normalization simply sets the scale of latent status to be on that of fathers' income.

[15] If the measurement errors were positively correlated, Black & Smith (2006) point out that the IV estimate from using one measure to instrument for the other provides a benchmark for a lower bound. In our case though, the measurement errors may be negatively correlated, which would leave the IV estimate biased upward.

underlying latent status directly imply correlation among the measurement errors.[16] Further, our main purpose is not to point identify β, rather we seek to test whether attenuation bias *decreases* as multiple proxies for latent status are taken into account. The LW approach is in this respect superior, allowing us to compare different lower bounds without making restrictive assumptions on cross-correlations of the measurement errors.

In addition to relaxing the assumption of zero correlation among the measurement errors (u_{jit}), we also allow the coefficients, ρ_j, in the measurement equations to be mostly unrestricted (subject to a normalization discussed below). The approach from Lubotsky & Wittenberg (2006; henceforth, LW) is ideally suited for this scenario, as it actually exploits the correlation in the measurement errors and estimates the coefficients in the measurement equations. In fact, the LW approach incorporates the information from all included measures of status in an optimal fashion, producing the estimate of persistence with the least attenuation bias. The LW estimator can be written as:

$$\beta_{LW} = \hat{\rho}_1\hat{\phi}_1 + \hat{\rho}_2\hat{\phi}_2 + \cdots + \hat{\rho}_J\hat{\phi}_J, \qquad (4)$$

where the $\hat{\rho}_j$'s are estimates of the slope coefficients in the measurement equations, and the $\hat{\phi}_j$'s are obtained from an auxiliary OLS regression described below. Hence, actual computation entails a multistep process.

The first step of the LW approach is to obtain the auxiliary OLS coefficient estimates of ϕ_j from regressing the dependent variable on all measures of status:

[16] Clark (2014, p.11) refers to education being a poor measure of status for Bill Gates (who presumably has high status), as he is a college dropout but has incredibly high income. Conversely, the other example posits that income would be a poor measure of status for a philosophy professor, whose education would be a more appropriate measure. These scenarios imply a negative correlation among the measurement errors for income and education.

$$y_{it+1} = \phi_1 y_{1it} + \phi_2 y_{2it} + \cdots + \phi_J y_{Jit} + \vartheta_i. \quad (5)$$

To identify the coefficients in the measurement equations, we need a normalization assumption on one of the ρ_j's. We normalize $\rho_1 = 1$, which simply sets the scale of the latent status to be on that of fathers' log income.[17] This implies the following formula to obtain estimates of the ρ_j's:

$$\rho_j = \frac{cov(y_{it+1}, y_{jit})}{cov(y_{it+1}, y_{1it})} \quad (6)$$

Estimating this ratio can be done in a single step via IV estimation, with y_{jit} as the outcome variable and using y_{it+1} to instrument for y_{1it}. We obtain standard errors for the β_{LW} estimate using a block bootstrap (100 replications) to account for within-family correlation. While not identifying β itself, this estimator provides an estimate of β with the least attenuation bias based on the joint information in the proxy measures of status. If the *simple law of mobility* does hold, we should see estimated persistence levels rising as we add these measures of status.

In addition to the proclaimed elevated persistence (i.e., lower mobility), the other controversial aspect of the *simple law* is the assertion that rates of mobility are constant across countries. To facilitate a cross-country comparison between Sweden and the U.S., we estimate analogous specifications using a Swedish sample with the measures constructed similarly to those used for the U.S. by Vosters (2015). From this we can also examine the consequences of various data limitations within the Swedish setting, thus providing indirect evidence on whether the U.S. estimates would change if based on richer data. We also conduct various robustness checks with other constructs of the income, education, and occupation measures.

[17] This normalization hence allows the LW estimate to be directly comparable to the conventionally estimated intergenerational income elasticity. In fact, in the case where income is the only status measure used, it is easily seen from equations (4) and (5) that the LW estimate is identical to this conventional estimate.

Further, because the hypothesized simple law relies on the social status of *families*, we extend our analysis to other family members, by adding the analogous measures for mothers, as well as estimating specifications with only mothers' measures. This exercise provides some suggestive evidence not only on the role of mothers but also on new methods for measuring mothers' status. In addition, given the paucity of evidence on intergenerational persistence for daughters, we also extend our analysis to daughters.

2.4 Empirical Results

2.4.1 Main Results

We first examine the conventionally estimated intergenerational persistence of income in Sweden, and whether adding additional partial measures affects the estimated coefficient on log income. In these and all other estimations, we control flexibly for cohorts of each generation using birth-year dummies. For the results presented in Table B2, we use the long-run average of sons' log income as the dependent variable, and fathers' measures of status include the long-run average of fathers' log income, educational attainment, and unique occupation indicators. The first set of results in Table B2 provides OLS estimates (omitting those for the 269 occupation indicators for brevity), with columns [1]-[4] progressively adding measures of fathers' status. Note that these estimates also correspond to the OLS components (ϕ_j) of the LW estimate obtained in the auxiliary regression.

The baseline OLS estimate of equation (1) is 0.23. This estimate of the intergenerational income elasticity is of similar magnitude to previous estimates for Sweden. Moving to column [2], fathers' educational attainment is added to the regression, but the coefficient estimate on fathers' income remains nearly identical. When instead fathers' occupation indicators are added to the regression in column [3], the coefficient on income does fluctuate some, falling to 0.21. This estimate is hardly

affected by the inclusion of education, as shown in column [4], indicating that while there is some sensitivity to the addition of the occupation measure, we see very little sensitivity to the addition of educational attainment.

With two noisy measures of status, and assumptions of classical measurement error, IV estimation provides a consistent estimate of intergenerational persistence in underlying status. Considering this scenario, the next rows of Table B2 include first and second stage results when instrumenting for fathers' income using fathers' education. This estimate of persistence is 0.24, similar to conventionally estimated persistence for Sweden. However, it is important to recognize the possibility that the assumptions for consistency may be violated. In particular, the measurement error in income as a measure of social status may be correlated with the measurement error in educational attainment, leaving the direction of bias unknown without further information on the nature of the correlation.

The final estimation approach, proposed by Lubotsky & Wittenberg (2006), exploits such violations by using the information on the relationships among the measurement errors and providing the greatest lower bound on persistence in underlying status. The LW estimate in column [1] is identical to the OLS estimate (by construction). However, as we incorporate more measures of status, this approach provides a single estimate from an optimal aggregation of the information from all measures. Given that the OLS estimates shown in the top of Table B2 are underlying components of the LW coefficient estimate, it is unsurprising that adding education does not change the LW estimate, as shown in column [2]. Similarly, given the sensitivity of the OLS coefficients to adding the occupation indicators, the increased persistence with the inclusion of occupation in column [3] is somewhat expected. However, the nominal rise from the conventional estimate of 0.23 to 0.26 when all suggested partial measures are included (column [4]) does not support the hypothesis of substantial attenuation bias in prior estimates. This pattern of results is similar to that

found for the U.S. (Vosters, 2015), exhibiting minimal increases in persistence when more partial measures of status are included, despite claims of elevated persistence in all countries. However, an important difference here is the statistical certainty. Due to our much smaller standard errors, we can reject even moderate drops in attenuation bias (for this specific model).

2.4.2 A Comparison of Sweden and the United States

Next we turn to directly address the hypothesis that persistence in status is in fact constant across countries. Our main results (provided again in Table B3) show that the persistence estimates for Sweden remain in the previously cited range of 0.20-0.30. Further, these estimates are substantially lower than the U.S. estimates of 0.44-0.47 found by Vosters (2015), illustrating a meaningful distinction in persistence between the two countries. However, the Swedish measures are constructed differently (e.g., the long-run income measure and the unique occupation indicators). While the measurement differences would likely bias the U.S. estimates towards the Swedish estimates, we carefully construct our measures to mimic those used by Vosters to allow for a more sound comparison. Using the Swedish data, we also indirectly gauge what the estimated persistence in status might look like in the U.S. if based on richer data.

With the five-year average of log income and broad occupation categories constructed to match those for the U.S., we find that estimated persistence in Sweden is lower at 0.19-0.22. To check whether this might be due to sample composition differences between our main sample and this smaller sample, we also analyze the same sample using our original measure constructs, and find estimates (0.23-0.26) nearly identical to our main results. Sample composition does not appear to be driving the differences. While our results do show that the U.S. estimates may be somewhat attenuated, possibly by some 10-20 percent, we can also see that the increase in estimates as additional measures are added does not change regardless of how measures are constructed; in no

cases do the estimates rise substantially when additional measures of status are included. Further, the estimates for Sweden remain in the approximate range previously asserted in the literature, albeit at the low end around 0.20, and clearly differ from the estimates for the U.S.[18] Thus, our results fail to support either aspect of the *simple law of mobility*.

2.4.3 Robustness of Main Results

Next we examine the sensitivity of our main results to various modifications to the measures of status. For our measure of income, we did see some sensitivity to the number of yearly income observations included in the average, as the five-year average used for the Sweden-U.S. comparison produced lower estimates than our longer-term measure. Another more arbitrary aspect of our measure construction is the choice to use the average of the annual log earnings rather than the log of average annual earnings. We provide estimates based on this alternative income measure construction in Table B4. While these estimates are slightly higher than our main results, they still remain in the typical range of estimates for Sweden. Moreover, the general pattern of the estimates as additional measures of status are added remains unchanged.

The other adjustments to the income measure, as well as the education and occupation measures, are motivated in part by our chosen empirical approach. For example, our long-term income measure gives equal weight to each annual measure from age 30 to 60, while each annual measure entering separately would allow the LW method to optimally choose these weights, which may vary over the life cycle. However, since the LW method also excludes any observations with a missing covariate and several of the fathers in our sample have incomplete income histories, we

[18] That the estimate for this specification is slightly lower than previous ones in the literature is not unexpected. While previous estimates have been based on long-run income measures and an optimal use of existing data, our goal here is to use data constructs comparable to the U.S.

estimate specifications that include annual log income from age 40 to 50, to reduce the data requirements while still focusing on income measures during mid-life. The persistence estimates are higher, ranging between 0.25-0.30, but are based on a much smaller and presumably more homogeneous sample of fathers that have log income observed in each of these eleven consecutive years. For comparison, we also estimate persistence based on the average of these annual log incomes, finding persistence estimates to be slightly lower (0.24-0.30), suggesting only trivial gains from allowing the LW method to determine the weights on the separate annual income measures. For another point of reference, the corresponding estimates using our baseline income measure for this sample are 0.28-0.33, which are even higher. So it appears that this sample exhibits more persistence than the full sample, but we also see that our longer-term average is serving as a better proxy for status than using the more flexible annual income measures when limited to fewer years.

We also adjust the education and occupation measures. For our main specifications, educational attainment enters under the assumption of a linear relationship in years of schooling. We relax this by using indicators for each level of highest attainment. Even with this flexible approach, education does not appear to provide substantial information on status (conditional on income), with estimates increasing by less than 0.01. We also estimate specifications indicating the *type* of education along with each level, again with increases of less than 0.01 in the estimate. Our main specification used the most flexible measure available for occupation. However, these detailed occupation indicators can be grouped into minor or major occupation groups (similar to those used for our U.S. comparison), resulting in estimates of 0.25 and 0.24, respectively. We thus see some numerical sensitivity of the estimates in this regard, though not to an extent that would affect the conclusions reached with our main analysis. In our main analysis, we included observations with occupation missing or undefined, accounted for using separate category indicators. When excluding these two groups, the baseline estimates increase by around 0.05. However, this modest numerical change has

little effect on our main conclusions regarding the level of persistence in Sweden (or the comparison to the U.S.).

In general, our robustness checks in Table B4 show that while there is some sensitivity of the estimates to how the partial measures are constructed, none of the changes are meaningful qualitatively. In particular, they do not change our conclusion that estimated persistence is not converging to 0.7-0.8 as additional measures are included, nor the conclusion of higher mobility in Sweden relative to the U.S.

2.4.4 Extension to Mothers and Daughters

Our results thus far have focused on male lineages, as is common in the intergenerational literature (including Vosters' and Clark's work). However, the *simple law* is described to pertain to underlying latent *family* status. To more appropriately address the concept of *family* status, we perform tests analogous to those above but including mothers' income, education, and occupation in addition to the same measures for fathers. This extension is warranted both by the specific hypothesis we are testing, but also by the dearth of evidence pertaining to mothers. To supplement the limited evidence in the literature, we also estimate persistence based on only mothers' status, and then attempt to disentangle contributions of status measures separately for mothers and fathers, in determining their child's later socioeconomic status. Since intergenerational associations for daughters are also much less common in the literature—especially mother-daughter associations in individual income—we conduct all of these tests for daughters as well.[19]

[19] Chadwick & Solon (2002) for the U.S. along with Rauum et al. (2007) for several different countries look at intergenerational income associations for daughters, circumventing the labor force participation issues by using a family income measure. Altonji & Dunn (1991) comprehensively looks at associations in family income and individual income, for all parent-child pairs, using U.S. data.

Similar to our main analysis, we begin by obtaining a baseline estimate via OLS and further augment the regression with additional measures. The first set of results in Table B5 replicates the main analysis for fathers and sons, only now restricting the sample to sons matched to both a father and mother, to facilitate comparisons with the different parent-offspring samples considered here.[20] For sons, the coefficient on fathers' income is not affected by the addition of education, while for daughters, adding fathers' education does seem to matter. The estimates for both daughters and sons are somewhat sensitive to accounting for fathers' occupation. When we add the corresponding measures for mothers to each of these specifications, the changes in the coefficient estimates are negligible for both sons and daughters (comparing the first panel to the second). The last set of results is for specifications using only mothers' measures. As the coefficient on mothers' income is very low, these estimates illustrate why mothers are generally not considered in studies of intergenerational *income* persistence. While today Sweden indeed has a high rate of female labor force participation, it was much lower for the cohorts of mothers in our sample (born before 1940), and thus individual income is a very noisy indicator of socioeconomic status.

In Table B6 we present the LW results, which aggregate information from additional status measures for each of the different parent-child samples. For fathers and sons, the results are nearly identical to the main results from the full sample, with persistence estimates ranging 0.23-0.26. For daughters, the intergenerational persistence in status with regard to their fathers is slightly lower, with estimates ranging 0.15-0.19. An important difference is that fathers' education does matter for the association in status with daughters, while it did not for sons. Fathers' occupation is similarly important for persistence in status with daughters and sons. The results for mothers are more striking, showing that mothers' occupation is crucial for measuring mothers' status. This holds

[20] Descriptives for these samples can be found in Table B7. OLS and LW results for the full mother-offspring and father-offspring samples can be found in Tables B8 and B9, respectively.

especially when considering intergenerational associations with sons, as the estimated persistence rises from 0.03 to 0.24. For daughters, the corresponding increase is from 0.06 to 0.13. These estimates are similar to the mother-son association found for the U.S. by Altonji & Dunn (1991), though they obtained a larger mother-daughter estimate. Clearly income is a very poor measure of status for mothers, and this is further confirmed by the results in Table B6; what was not obvious in the OLS results in Table B5 is the substantial impact of accounting for mothers' occupation, which is made apparent by the LW method's aggregation of all information contained in mothers' income, education, and occupation. Education is also salient to mothers' status, as shown in columns [6] and [8], though less so than occupation.

Next we include mothers' and fathers' measures jointly, to consider how persistence might change if we take more literally the concept of *family* status. When we compare these estimates to those accounting for only fathers' status (i.e., estimates reflecting the same information as most of the literature), we see that mothers' occupation does contain additional information on family status with respect to intergenerational transmission to sons, and even more so for daughters. Further, mothers' income seems salient to transmission of family status for daughters, a result consistent with Altonji & Dunn's (2000) finding that factors underlying earnings had stronger intergenerational associations along gender lines.

To further assess the relative importance of mothers' and fathers' status measures, we also attempt to separate the relative contributions of each parent to the intergenerational persistence estimate. Decomposing the estimate into portions due to mothers' and fathers' status, we see in the bottom portion of Table B6 that the vast majority of the persistence for sons is coming from fathers' measures, with only 4-5 percent from mothers in the income and income/education

specifications.[21] Mothers' occupation appears more important though, shifting more weight to mothers so they account for 15-16 percent. For daughters, the role of mothers' status is more substantial, accounting for 32-43 percent of estimated persistence in underlying status.

Whether mothers should contribute (conditional on fathers) to intergenerational persistence in family status is an empirical question. Theoretically, one could posit several stories. For example, if we believe there to be substantial positive assortative matching on latent status in the marriage market, then we might expect mothers' or fathers' status measures to serve as equally suitable measures of family status. Indeed, for the sample of sons, LW estimates from specifications including all measures for fathers are very similar to those including all measures for mothers (0.26 and 0.25, respectively). However, this is not the case when occupation indicators are omitted; nor does it hold as convincingly for the sample of daughters (with estimates of 0.19 and 0.14). While previous studies have found evidence of positive assortative matching in both Sweden (Hirvonen, 2008; Nakosteen et al., 2004) and the U.S. (Chadwick & Solon, 2002), this does not seem to explain our results here. In auxiliary correlational analyses, we find the mother-father correlation in educational attainment to be the highest at 0.55, but the correlation in long-run income is low (0.06). The correlations between mothers' and fathers' estimated latent status is also low (0.08), which is not surprising given that income both weights heavily into the status measures and exhibits low parental matching.

More likely, our results are explained by the well-known issues with using mothers' income, some of which we mentioned above. For education, it is less clear what the explanation is; educational attainment is both believed to suffer less from measurement problems and exhibit smaller male-female differences than what is the case for income. However, we do see that

[21] The decomposition is done by separating the sum $\beta_{LW} = \hat{\rho}_1\hat{\phi}_1 + \hat{\rho}_2\hat{\phi}_2 + \cdots + \hat{\rho}_J\hat{\phi}_J$ into the sum of elements from fathers' measures and the sum of elements from mothers' measures.

combining information from income *and* education can mitigate these measurement issues, and adding occupation is especially helpful. So Clark & Cummins' (2015) proposition that persistence estimates will rise when combining information from multiple measures seems to have some merit for capturing intergenerational associations with mothers. Each of our noisy measures contributes to measuring mothers' status, however not to the extent of raising persistence estimates to the levels proposed in the *simple law*.

2.5 Conclusions

Clark's work shifts the focus to be on *underlying socioeconomic status,* which is described to be a slightly different—presumably more general—concept relative to the purely economic ones economists have thus far considered. While it is not entirely clear to what extent these concepts should differ, Clark's work is painting an entirely different landscape for intergenerational persistence, provoking a new set of studies (such as this one) testing the surname results and associated hypotheses. Very few of these papers are confirming the results found with the surnames approach or the proposed reasoning for the contradictory results, as in the present paper (e.g., Chetty et al., 2014; Braun & Stuhler, 2015; Vosters, 2015).

We tested two facets of the hypothesized simple law of mobility, failing to find evidence to support either claim. We first looked for evidence of substantially increased intergenerational persistence in underlying social status in Sweden when information from several partial measures of parental status was combined. Incorporating information on educational attainment has almost no effect on the conventionally estimated persistence rate of 0.23. When occupation is included, the estimate increases slightly to 0.26, but does not come close to the hypothesized "true" persistence rate of about 0.7-0.8. We then investigated the claimed uniform persistence across countries, by comparing our Swedish estimates with those for the U.S. (presented in Vosters, 2015). Even after

harmonizing our sample and variables as to mimic those used in the U.S. study, our estimates still differ substantially, with the U.S. estimates of persistence being more than twice as large as the Swedish. Our analysis thus confirms the previously established higher levels of intergenerational mobility for Sweden relative to the US.

Prior studies, such as Goldberger (1989), also recognized that non-income measures may be important in measuring persistence in socioeconomic status.[22] However, Clark's theory formalizes this notion with a very simple measurement error framework and proposes an easily testable hypothesis. So while Clark is not the first to emphasize the importance of non-income measures, the exercise of considering a more general latent status has also prompted various extensions to the literature. For example, ours along with Vosters (2015) is one of the first studies to aggregate information from different dimensions of status into a single measure of persistence. While sociologists and economists have included, say, income and education in the same regression, these have been attempts to identify mechanisms, or simply reactions to data limitations, rather than for the purposes of obtaining one aggregate persistence estimate.

Coupled with our method for obtaining an aggregate estimate, Clark's theory regarding latent *status* unintentionally inspires another important contribution to both the measurement and intergenerational literatures, enabling further examination of intergenerational associations related to mothers. Studies rarely consider status transmission from mothers to children, or even fathers to daughters, due to data limitations stemming from lower labor force participation rates for females. In the context of Clark's work, despite the underlying theory being presented in the realm of male lineages, the latent variable approach might be more relevant for females. Hence, we first extended our analysis to more carefully consider the concept of *family* status by accounting for mothers' in

[22] Sociologists also consider non-income measures, instead often focusing on social "class" and various measures of occupational prestige.

addition to fathers' status measures, which had very little impact on estimated persistence, especially for sons, with persistence rising to only 0.28. Although beyond the scope of the *simple law*, we do find the framework to be more relevant to females, especially mothers. We show that a modified version of the measurement error framework presented by Clark proves useful in estimating intergenerational associations between mothers and their offspring. In contexts where income is a very noisy measure of socioeconomic status, as is often the case for mothers, supplementing this information with additional noisy measures can make an important difference, as shown by our analyses incorporating mothers' education and occupation.[23] In fact, for daughters the intergenerational persistence estimates accounting for all measures are only slightly lower for mothers relative to fathers. While these results warrant future research for a better understanding of these transmission channels, our results here illustrate what information might be gained by considering other estimation approaches, and other measures of status.

[23] If the available income data is of low quality (or observed only as short snapshots), the same approach could also be potentially useful when studying men.

APPENDIX

Table B1: Summary Statistics for Full Sample and U.S. Comparison Sample

	Full sample		U.S. comparison sample	
Variable	**mean**	*std dev*	**mean**	*std dev*
Offspring				
Year of birth	1956	*3.14*	1956	*3.12*
Average income, age 27-46	250,584	*198,680*	253,318	*201,623*
Average log income, age 27-46	12.22	*0.53*	12.25	*0.50*
Non-missing incomes, age 27-46	19.52	*1.81*	19.66	*1.41*
Log income in 1991			12.23	*0.70*
Years of education	11.79	*2.40*	11.83	*2.41*
Number of offspring (N)	167,552		146,783	
Fathers				
Age when offspring born	30.05	*5.13*	30.22	*5.06*
Year of birth	1926	*6.53*	1925	*6.29*
Average income, age 30-60	245,013	*166,478*	249,144	*164,416*
Average log income, age 30-60	12.26	*0.48*	12.29	*0.45*
Non-missing incomes, age 30-60	17.78	*6.56*	17.71	*6.39*
Average log income 1968-72			12.28	*0.48*
Years of education	9.14	*2.88*	9.15	*2.91*
Educational attainment (years)				
< 9 years of primary school	0.58	*0.49*	0.58	*0.49*
9 years of primary school	0.04	*0.21*	0.04	*0.20*
2-year secondary school	0.18	*0.38*	0.17	*0.38*
3-year secondary school	0.11	*0.31*	0.11	*0.31*
< 3 years of post-secondary school	0.03	*0.17*	0.03	*0.17*
3+ years of post-secondary school	0.06	*0.23*	0.06	*0.24*
Graduate school	0.01	*0.08*	0.01	*0.08*
Occupation category				
1. Professional work (arts & sciences)	0.19	*0.40*	0.20	*0.40*
2. Managerial work	0.04	*0.21*	0.05	*0.21*
3. Clerical work	0.04	*0.19*	0.04	*0.19*
4. Wholesale, retail, & commerce	0.08	*0.28*	0.08	*0.27*
5. Agriculture, forestry, hunting, & fishing	0.10	*0.30*	0.10	*0.30*
6. Mining & quarrying	0.01	*0.08*	0.01	*0.08*
7. Transportation & communication	0.09	*0.29*	0.09	*0.29*
8. Manufacturing	0.38	*0.48*	0.38	*0.49*
9. Services	0.04	*0.20*	0.04	*0.19*
10. Military / armed forces	0.01	*0.10*	0.01	*0.10*
Undefined	<0.00	*0.01*	0.00	*0.01*
Missing	0.02	*0.13*	0.01	*0.10*
Number of fathers (N)	153,920		135,020	

Notes. The main sample is the full sample used for our main analysis as well as robustness checks. The U.S. comparison sample is the subset that has the income measures needed to compute the PSID comparable income measures (average of log income in years 1968-72 for fathers and, for sons, log income in 1991).

Table B2: OLS, IV, and LW Estimates for Full Sample (Fathers and Sons)

	[1]	[2]	[3]	[4]
OLS estimates				
Fathers' log average income	0.231	0.230	0.208	0.207
	0.003	*0.004*	*0.004*	*0.004*
Fathers' years of education		0.000		0.000
		0.001		*0.001*
Fathers' unique occupation (indicators)			x	x
IV estimates				
First stage (educ. IV for log income)	0.083			
	0.000			
Second stage	0.235			
	0.006			
LW estimates of the IGE				
B	0.231	0.231	0.260	0.260
	0.004	*0.004*	*0.004*	*0.004*
Observations (N)	167,552	167,552	167,552	167,552

Notes. All specifications use the average of sons' log income as the dependent variable and include birth-year dummies of fathers and sons as controls. The noisy measures of status for fathers included in each model are: [1] income; [2] income and education; [3] income and occupation; [4] income, education and occupation. Because the occupation measure is 270 unique occupation categories, the OLS coefficients and standard errors for occupations are omitted from the table. OLS and IV standard errors are clustered by family and LW standard errors are computed using a block bootstrap to account for within-family correlation (100 repetitions).

Table B3: Comparison of LW Estimates - Sweden and the U.S.

	N	[1]	[2]	[3]	[4]
Sweden estimates					
Main results (full sample)	167,552	0.231	0.231	0.260	0.260
		0.004	*0.004*	*0.004*	*0.004*
Main specifications for restricted sample used in U.S. comparable specification	146,783	0.231	0.231	0.262	0.262
		0.003	*0.003*	*0.004*	*0.004*
Sweden estimates using U.S. (PSID) comparable specification	146,783	0.194	0.194	0.215	0.215
		0.004	*0.004*	*0.005*	*0.005*
U.S. estimates (from Vosters, 2014)	415	0.439	0.445	0.465	0.473
		0.075	*0.072*	*0.080*	*0.080*

Notes. The noisy measures of status for fathers included in each model are: [1] income; [2] income and education; [3] income and occupation; [4] income, education and occupation. The main specifications use the average of sons' log income (age 27-46) as the dependent variable, average of log income (age 30-60) for father's income, unique occupation indicators for fathers' occupation, and include birth-year dummies of fathers and sons as controls. The PSID-comparable measures are: sons' log income in 1991; father's average log income 1968-1972; indicators for fathers' major occupation category. All specifications use years of education as the measure of educational attainment. LW standard errors are computed using a block bootstrap to account for within-family correlation (100 repetitions).

Table B4: Robustness of LW Estimates to Construction of Status Measures

	N	[1]	[2]	[3]	[4]
Main results	167,552	0.231	0.231	0.260	0.260
		0.004	*0.004*	*0.004*	*0.004*
Adjusting the occupation measure					
Indicators for minor occupation (2-digit)	167,552	0.231	0.231	0.247	0.247
		0.004	*0.004*	*0.004*	*0.004*
Indicators for major occupation	167,552	0.231	0.231	0.238	0.238
		0.004	*0.004*	*0.004*	*0.004*
Excluding "undefined" and missing	164,678	0.235	0.235	0.265	0.265
		0.003	*0.003*	*0.004*	*0.004*
Adjusting the education measure					
Indicators for each education level	167,552	0.231	0.233	0.260	0.261
		0.004	*0.004*	*0.004*	*0.004*
Indicators for level/type of attainment	167,552	0.231	0.241	0.260	0.265
		0.004	*0.004*	*0.004*	*0.004*
Adjusting the income measure					
Log (average annual income)	167,550	0.270	0.274	0.296	0.297
		0.003	*0.003*	*0.003*	*0.003*
Separate log annual income measures, age 40-50	57,728	0.247	0.247	0.304	0.304
		0.008	*0.008*	*0.009*	*0.010*
Average of log annual income, age 40-50	57,728	0.241	0.241	0.298	0.298
		0.008	*0.008*	*0.009*	*0.009*
Main specification using this restricted sample	57,728	0.279	0.279	0.333	0.333
		0.007	*0.007*	*0.009*	*0.009*

Notes. All specifications use the average of sons' log income as the dependent variable and include birth-year dummies of fathers and sons as controls. The noisy measures of status for fathers included in each model are: [1] income; [2] income and education; [3] income and occupation; [4] income, education and occupation. LW standard errors are computed using a block bootstrap to account for within-family correlation (100 repetitions).

Table B5: OLS Estimates from Extensions with Mothers' Measures of Status

	Sons				Daughters			
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Fathers' measures								
Log average income	0.231	0.229	0.208	0.207	0.152	0.128	0.130	0.123
	0.003	*0.004*	*0.004*	*0.005*	*0.003*	*0.004*	*0.004*	*0.004*
Education		0.001		0.001		0.008		0.006
		0.001		*0.001*		*0.001*		*0.001*
Fathers' & Mothers' measures								
Fathers' log avg. income	0.225	0.227	0.203	0.203	0.142	0.125	0.125	0.120
	0.003	*0.004*	*0.005*	*0.005*	*0.003*	*0.004*	*0.004*	*0.004*
Mothers' log avg. income	0.024	0.023	0.010	0.009	0.059	0.053	0.048	0.046
	0.002	*0.002*	*0.002*	*0.002*	*0.002*	*0.002*	*0.003*	*0.003*
Fathers' education		-0.001		-0.001		0.003		0.002
		0.001		*0.001*		*0.001*		*0.001*
Mothers' education		0.002		0.003		0.005		0.005
		0.001		*0.001*		*0.001*		*0.001*
Mothers' measures								
Log average income	0.034	0.021	0.002	-0.001	0.064	0.052	0.043	0.040
	0.002	*0.002*	*0.002*	*0.002*	*0.002*	*0.002*	*0.003*	*0.003*
Education		0.016		0.011		0.015		0.012
		0.001		*0.001*		*0.001*		*0.001*
Observations (N)	152,486	152,486	152,486	152,486	145,256	145,256	145,256	145,256

Notes. All specifications use the average of sons' or daughters' log income as the dependent variable and include birth-year dummies of included parents and offspring as controls. The noisy measures of status for parents included in each model are: [1], [5] income; [2], [6] income and education; [3], [7] income and occupation; [4], [8] income, education and occupation. Because the occupation measure is 270 unique occupation categories, the OLS coefficients and standard errors for occupations are omitted from the table. Standard errors are clustered by family to account for within-family correlation.

Table B6: LW Estimates from Extensions with Mothers' Measures of Status

	Sons				Daughters			
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Fathers	0.231	0.231	0.262	0.262	0.152	0.163	0.188	0.193
	0.003	*0.003*	*0.004*	*0.004*	*0.003*	*0.003*	*0.004*	*0.004*
Mothers	0.034	0.096	0.235	0.252	0.064	0.096	0.132	0.142
	0.002	*0.006*	*0.014*	*0.015*	*0.002*	*0.003*	*0.004*	*0.004*
Fathers & Mothers	0.234	0.235	0.283	0.283	0.209	0.218	0.273	0.276
	0.003	*0.003*	*0.004*	*0.004*	*0.004*	*0.005*	*0.006*	*0.006*
Fathers' portion	0.225	0.222	0.242	0.239	0.142	0.140	0.160	0.157
	96%	95%	85%	84%	68%	64%	59%	57%
Mothers' portion	0.009	0.012	0.041	0.044	0.066	0.078	0.113	0.118
	4%	5%	15%	16%	32%	36%	41%	43%
Observations (N)	152,486	152,486	152,486	152,486	145,256	145,256	145,256	145,256

Notes. These estimation samples have non-missing data on all measures for mothers and fathers. All specifications use the average of sons' or daughters' log income as the dependent variable and include birth-year dummies of included parents and offspring as controls. The noisy measures of status for parents included in each model are: [1], [5] income; [2], [6] income and education; [3], [7] income and occupation; [4], [8] income, education and occupation. Because the occupation measure is 270 unique occupation categories, the OLS coefficients and standard errors for occupations are omitted from the table. Standard errors are computed using a block bootstrap to account for within-family correlation (100 repetitions).

Table B7: Summary Statistics for Mothers & Fathers (Balanced Samples)

Variable	mean	*std dev*	mean	*std dev*	mean	*std dev*	mean	*std dev*
Offspring	*Sons*				*Daughters*			
Year of birth	1956	*3.13*			1956	*3.12*		
Average income, age 27-46	252,496	*203,293*			169,955	*78,675*		
Average log income, age 27-46	12.23	*0.53*			11.84	*0.54*		
Non-missing incomes, age 27-46	19.52	*1.79*			19.57	*1.66*		
Years of education	11.81	*2.39*			12.21	*2.30*		
Number of offspring (N)	152,486				142,020			
Parents	*Fathers*		*Mothers*		*Fathers*		*Mothers*	
Age when offspring born	29.93	*5.12*			29.92	*5.13*		
Year of birth	1926	*6.38*	1929	*6.16*	1926	*6.38*	1929	*6.16*
Average income, age 30-60	245,518	*161,788*	121,359	*65,144*	245,259	*182,143*	121,419	*64,393*
Average log income, age 30-60	12.26	*0.48*	11.38	*0.78*	12.26	*0.48*	11.39	*0.78*
Non-missing incomes, age 30-60	18.22	*6.45*	19.49	*6.60*	18.22	*6.45*	19.53	*6.61*
Years of education	9.18	*2.90*	8.53	*2.36*	9.17	*2.89*	8.53	*2.37*
Educational attainment (years)								
< 9 years of primary school	0.57	*0.49*	0.63	*0.48*	0.57	*0.49*	0.63	*0.48*
9 years of primary school	0.04	*0.21*	0.11	*0.31*	0.04	*0.20*	0.10	*0.31*
2-year secondary school	0.18	*0.38*	0.18	*0.38*	0.18	*0.39*	0.18	*0.38*
3-year secondary school	0.11	*0.31*	0.02	*0.14*	0.11	*0.31*	0.02	*0.14*
< 3 years of post-secondary school	0.03	*0.17*	0.03	*0.17*	0.03	*0.17*	0.03	*0.17*
3+ years of post-secondary school	0.06	*0.24*	0.03	*0.18*	0.06	*0.24*	0.03	*0.18*
Graduate school	0.01	*0.08*	0.00	*0.02*	0.01	*0.08*	0.00	*0.02*
Occupation category								
1. Professional work	0.20	*0.40*	0.18	*0.38*	0.20	*0.40*	0.18	*0.38*
2. Managerial work	0.04	*0.21*	0.01	*0.08*	0.04	*0.20*	0.01	*0.08*
3. Clerical work	0.04	*0.19*	0.16	*0.36*	0.04	*0.19*	0.16	*0.36*
4. Wholesale, retail, & commerce	0.09	*0.28*	0.11	*0.31*	0.09	*0.28*	0.11	*0.32*
5. Agriculture, forestry, hunting, fishing	0.09	*0.29*	0.06	*0.23*	0.09	*0.29*	0.06	*0.23*
6. Mining & quarrying	0.01	*0.08*	0.00	*0.02*	0.01	*0.08*	0.00	*0.02*
7. Transportation & communication	0.09	*0.29*	0.04	*0.19*	0.09	*0.29*	0.04	*0.19*
8. Manufacturing	0.38	*0.48*	0.09	*0.29*	0.38	*0.49*	0.10	*0.29*
9. Services	0.04	*0.20*	0.26	*0.44*	0.04	*0.20*	0.26	*0.44*
10. Military / armed forces	0.01	*0.10*	0.00	*0.00*	0.01	*0.10*	0.00	*0.00*
Undefined	0.00	*0.01*	0.00	*0.04*	0.00	*0.02*	0.00	*0.04*
Missing	0.02	*0.13*	0.10	*0.30*	0.02	*0.13*	0.10	*0.29*
Number of parents (N)	140,052		140,234		133,884		134,108	

Table B8: OLS Estimates from Extensions with Mothers' Measures of Status, for All Parent-Child Samples

	Sons				Daughters			
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Fathers' measures								
Log average income	0.231	0.230	0.208	0.207	0.153	0.129	0.130	0.122
	0.003	*0.004*	*0.004*	*0.004*	*0.003*	*0.004*	*0.004*	*0.004*
Education		0.000		0.000		0.008		0.006
		0.001		*0.001*		*0.001*		*0.001*
Observations (N)	167,552	167,552	167,552	167,552	159,172	159,172	159,172	159,172
Fathers' & Mothers' measures								
Fathers' log avg. income	0.225	0.227	0.203	0.203	0.142	0.125	0.125	0.120
	0.003	*0.004*	*0.005*	*0.005*	*0.003*	*0.004*	*0.004*	*0.004*
Mothers' log avg. income	0.024	0.023	0.010	0.009	0.059	0.053	0.048	0.046
	0.002	*0.002*	*0.002*	*0.002*	*0.002*	*0.002*	*0.003*	*0.003*
Fathers' education		-0.001		-0.001		0.003		0.002
		0.001		*0.001*		*0.001*		*0.001*
Mothers' education		0.002		0.003		0.005		0.005
		0.001		*0.001*		*0.001*		*0.001*
Observations (N)	152,486	152,486	152,486	152,486	145,256	145,256	145,256	145,256
Mothers' measures								
Log average income	0.032	0.019	-0.002	-0.005	0.062	0.049	0.039	0.036
	0.002	*0.002*	*0.002*	*0.002*	*0.002*	*0.002*	*0.002*	*0.002*
Education		0.016		0.010		0.015		0.012
		0.001		*0.001*		*0.001*		*0.001*
Observations (N)	173,608	173,608	173,608	173,608	165,161	165,161	165,161	165,161

Notes. All specifications use the average of sons' or daughters' log income as the dependent variable and include birth-year dummies of included parents and offspring as controls. The noisy measures of status for parents included in each model are: [1], [5] income; [2], [6] income and education; [3], [7] income and occupation; [4], [8] income, education and occupation. Because the occupation measure is 270 unique occupation categories, the OLS coefficients and standard errors for occupations are omitted from the table. Standard errors are clustered by family to account for within-family correlation.

Table B9: LW Estimates from Extensions with Mothers' Measures of Status, for All Parent-Child Samples

	Sons				Daughters			
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Fathers	0.231	0.231	0.260	0.260	0.153	0.164	0.190	0.194
	0.004	*0.004*	*0.004*	*0.004*	*0.004*	*0.004*	*0.005*	*0.005*
Observations (N)	167,552	167,552	167,552	167,552	159,172	159,172	159,172	159,172
Mothers	0.032	0.098	0.246	0.263	0.062	0.049	0.039	0.036
	0.003	*0.007*	*0.012*	*0.012*	*0.003*	*0.004*	*0.006*	*0.007*
Observations (N)	173,608	173,608	173,608	173,608	165,161	165,161	165,161	165,161

Notes. All specifications use the average of sons' or daughters' log income as the dependent variable and include birth-year dummies of included parents and offspring as controls. The noisy measures of status for parents included in each model are: [1], [5] income; [2], [6] income and education; [3], [7] income and occupation; [4], [8] income, education and occupation. Because the occupation measure is 270 unique occupation categories, the OLS coefficients and standard errors for occupations are omitted from the table. Standard errors are computed using a block bootstrap to account for within-family correlation (100 repetitions).

REFERENCES

REFERENCES

Altonji, J. G. & Dunn, T. A. (1991). Relationships among the family incomes and labor market outcomes of relatives (No. 3724). National Bureau of Economic Research.

Altonji, J. G. & Dunn, T. A. (2000). An intergenerational model of wages, hours, and earnings. *Journal of Human Resources*, 35(2), 221-258.

Becker, G. & Tomes, N. (1979) An equilibrium theory of the distribution of income and intergenerational Mobility. *Journal of Political Economy,* (87), 1153-1189.

Björklund, A. & Jäntti, M. (2009). Intergenerational income mobility and the role of family background, in (W. Salverda, B. Nolan and T. Smeeding eds.) Oxford Handbook of Economic Inequality, Oxford University Press.

Björklund, A. Jäntti, M., & Nybom, M. (2015). The contribution of early-life vs. labour-market factors to intergenerational income persistence: a comparison of the UK and Sweden. Mimeo, Stockholm University.

Black, D. A. & Smith, J. A. (2006). Estimating the returns to college quality with multiple proxies for quality. *Journal of Labor Economics*, 24(3), 701-728.

Black, S. E. & Devereux, P. J. (2011). Recent developments in intergenerational mobility, in (O. Ashenfelter and D. Card, eds.), *Handbook of Labor Economics*, (4), 1487-1541, Amsterdam: Elsevier.

Bollen, K. A. (1989). Structural equations with latent variables. Wiley, New York, NY.

Braun, S. & Stuhler, J. (2015). The transmission of inequality across multiple generations: Testing recent theories with evidence from Germany. Mimeo.

Chadwick, L. & Solon, G. (2002). Intergenerational income mobility among daughters. *American Economic Review*, 92(1), 335-344.

Chetty, R. Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *Quarterly Journal of Economics*, *129*(4), 1553-1623.

Clark, G. (2014). The son also rises: surnames and the history of social mobility. Princeton University Press.

Clark, G. & Cummins, N. (2015). Intergenerational wealth mobility in England, 1858–2012: Surnames and social mobility. *The Economic Journal*, 125(582), 61-85.

Corak, M. & Piraino, P. (2011). The intergenerational transmission of employers. *Journal of Labor Economics*, 29(1), 37-68.

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

Goldberger, A.S. (1989). Economic and mechanical models of intergenerational transmission. *American Economic Review,* 79(3), 504-513.

Haider, S. J. & Solon, G. (2006). Life-cycle variation in the association between current and lifetime earnings. *American Economic Review*, 96(4), 1308-1320.

Hirvonen, L. H. (2008). Intergenerational earnings mobility among daughters and sons: Evidence from Sweden and a comparison with the United States. *American Journal of Economics and Sociology*, *67*(5), 777-826.

Jöreskog, K. G. & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351a), 631-639.

Lubotsky, D. & Wittenberg, M. (2006). Interpretation of regressions with multiple proxies. *The Review of Economics and Statistics*, 88(3), 549-562.

Mazumder, B. (2005). Fortunate sons: New estimates of intergenerational mobility in the United States using social security earnings data. *The Review of Economics and Statistics*, 87(2), 235-255.

Nakosteen, R. A., Westerlund, O., & Zimmer, M. A. (2004). Marital Matching and Earnings: Evidence from the Unmarried Population in Sweden. *Journal of Human Resources*, *39*(4), 1033-1044.

Nybom, M. & Stuhler, J. (*forthcoming*). Heterogeneous Income Profiles and Life-Cycle Bias in Intergenerational Mobility Estimation. *Journal of Human Resources*.

Raaum, O., Bratsberg, B., Røed, K., Österbacka, E., Eriksson, T., Jäntti, M., & Naylor, R.A. (2007). Marital Sorting, Household Labor Supply, and Intergenerational Earnings Mobility Across Countries. *The BE Journal of Economic Analysis & Policy*, *7*(2).

Solon, G. (1999). "Intergenerational mobility in the labor market," in (O. Ashenfelter and D. Card, eds.), *Handbook of Labor Economics*, (3A), 1761-1800, Amsterdam: North-Holland.

Solon, G. (2002). Cross-country differences in intergenerational earnings mobility. *The Journal of Economic Perspectives*, 16(3), 59-66.

Solon, G. (2004). A model of intergenerational mobility variation over time and place. In M. Corak, *Generational Income Mobility in North America and Europe* (pp. 38-47). Cambridge University Press.

Statistics Sweden (2004). Jämförelse mellan yrkesuppgifter i FoB och yrkesregistret. URL: http://www.scb.se/sv_/Hitta-statistik/Statistik-efter-amne/Arbetsmarknad/Sysselsattning-forvarvsarbete-och-arbetstider/Yrkesregistret-med-yrkesstatistik/59068/Jamforelser-mot-aldre-yrkesstatistik/ (Accessed: 2015-05-28)

Vosters, K. (2015). Is the simple law of mobility really a law? Testing Clark's hypothesis. Unpublished manuscript.

Chapter 3

Understanding and Evaluating SAS® EVAAS® Models for Measuring Teacher Effectiveness[1]

3.1 Introduction

A large literature examines many of the statistical methods that states or districts are using to estimate teacher effectiveness based on their students' test scores. However, one of the methodological approaches that has been adopted by several states and districts—the SAS® EVAAS® model—has experienced relatively limited exposure in these studies, in large part due to the proprietary nature of the analysis. Still, the EVAAS estimates have been incorporated into formal teacher evaluation programs used for accountability, including high stakes policies such as tenure, dismissal, or incentive pay. With high stakes programs such as these relying on estimated effectiveness, it is important to understand the strengths and limitations of the underlying methods.

The prevalence of such policies has grown in recent years, but the SAS EVAAS approach itself has a much longer history. The current name, EVAAS, stands for Education Value-Added Assessment System, which is a variant on the earlier and perhaps more familiar name Tennessee Value-Added Assessment System (TVAAS), as Tennessee was where it was developed and used since the early 1990's.[2] In addition to the name change, documentation of the EVAAS methods has evolved over the years but the details that allow researchers to easily replicate the approach remain somewhat elusive. The nature of the documentation combined with proprietary programs and data likely impede the implementation of EVAAS in many evaluation studies (Kupermintz, 2003; Amrein-Beardsley, 2008).

The EVAAS methods include two options for estimating teacher effectiveness; the multivariate response model (MRM) and the univariate response model (URM). The MRM, also referred to as the "layered" teacher model, involves joint modeling of scores from multiple tested subjects for multiple grades and cohorts in a 5-year period. Jointly modeling the test scores aims to improve

[1] This chapter is coathored with Cassandra Guarino and Jeffrey Wooldridge.

[2] The name is often modified in a similar fashion in states which adopt the EVAAS methods, such as "PVAAS" for Pennsylvania (e.g., www.portal.state.pa.us, accessed 1/12/2015).

efficiency, and using the complete set of scores available for a student attempts to account for any other student characteristics that might affect achievement. This model is generally limited to within-district purposes due to the large computational burden, and is sometimes not feasible if data requirements cannot be met. Hence, the URM was developed for these situations. The URM focuses on a single subject, and thus is less intensive computationally and more flexible with respect to data requirements. The method involves the computation of a single composite score for each student based on their lagged scores in the same subject as well as others, and then using this composite score as the only regressor in empirical Bayes' estimation of the teacher effects.

A number of studies have addressed signature features of the MRM, such as the omission of student covariates or joint modeling of subjects, typically focusing on a comparison to a generalized or modified version of the model (e.g., Ballou, Sanders, & Wright, 2004; McCaffrey et al., 2004; Lockwood et al., 2007). The URM has received less attention, with the exception of a recent report by Rose, Henry, & Lauen (2012) that studies the performance of nine estimators, one of which is the URM, in simulations and with administrative data. They find that under random assignment of students to teachers, a three-level hierarchical linear model (HLM) and the URM outperform several other popular estimation approaches and that under certain nonrandom assignment scenarios, the HLM approach outperforms the URM by a fair margin.

Our paper also focuses primarily on the URM and we include both simulations and the analysis of actual data. We build on the work of Rose et al. (2012), although our simulations are designed somewhat differently, and our results diverge from theirs. While we confirm that random effects approaches such as HLM are best under random assignment (a result we have found in prior work—see Guarino, Reckase, and Wooldridge 2015), we find that under the type of nonrandom assignment that we simulate, approaches that assume fixed rather than random teacher effects are better suited to capturing true teacher effects than the URM. The URM assumes random teacher effects and is thus inconsistent if teacher assignment is related to students' prior test scores. In contrast, OLS estimation of the regression of student achievement on teacher fixed effects and control variables including lagged student achievement scores is consistent even when nonrandom assignment based on lagged achievement generates correlation between the teacher dummy variables and the control variables.

OLS, however, assumes fixed teacher effects and is still consistent under this type of teacher assignment. In other nonrandom assignment scenarios in which both the URM and OLS are inconsistent, OLS performs at least as well as the URM.

Our paper further contributes to the literature by drawing key theoretical connections between the URM and other types of estimation approaches. In particular, we show areas of overlap between the URM and OLS or empirical Bayes' estimation of typical value-added models, and we also show how and where the various estimation approaches differ. Through the theoretical discussion, simulations, and empirical work, we show that standard linear regression techniques perform very similarly—and in certain cases better—under plausible data scenarios. In addition, our detailed descriptions of the URM help make it more readily available for other researchers to implement and include in future evaluation studies.

We begin by describing common value-added model (VAM) approaches as well as the EVAAS approaches in Section 2, providing details on both the MRM and URM, and then we review relevant literature in Section 3. In Section 4, we discuss our simulation design and present results from the simulation. Section 5 describes our empirical analysis and results using administrative data. We summarize and conclude in Section 6.

3.2 Value-Added Models

Teacher value-added models (VAMs) are generally derived from or motivated by a so-called "education production function" (Hanushek, 1979; Todd & Wolpin, 2003; Guarino, Reckase & Wooldridge, 2015). In its most general formulation, academic achievement at any point in time is written as a function of all current and past child, family, and school inputs:

$$A_{it} = f(E_{it}, ..., E_{i0}, X_{it}, ..., X_{i0}, c_i, u_{it}) \tag{1}$$

where A_{it} is current achievement at time t for student i, E_{it},...,E_{i0} represent current and past education (school) inputs, X_{it},...,X_{i0} represent current and past student or parent inputs, c_i is unobserved student heterogeneity (e.g., motivation or some form of time-invariant innate ability), and u_{it} is an idiosyncratic error term. Given that we cannot measure each of these elements during each

time period—at least not in available data—researchers typically adopt a more parsimonious model with a simple (estimable) functional form. For example, with a set of simplifying assumptions, the education production function is reduced to an estimating equation such as:

$$A_{it} = \tau_t + \lambda A_{it-1} + X_{it}\beta + E_{it}\gamma + c_i + e_{it} \tag{2}$$

where τ_t allows for a different intercept in each time period to capture time (e.g., year) effects, A_{it} is the current test score at time t, A_{it-1} is the lagged test score from the previous year, E_{it} is a vector of *observed* education inputs at time t (e.g., teacher assignment indicators), and X_{it} is a vector of observed individual student characteristics.[3]

The simplifying assumptions that facilitate the transition from equation (1) to equation (2) include linearity and geometric decay in the parameters; see Guarino, Reckase, & Wooldridge (2015) for a detailed discussion and derivations. We cannot measure the individual student heterogeneity, c_i, so this is generally left in the error term in commonly used approaches. While there are methods to eliminate this term in panel data settings (e.g., adding student indicators, or fixed effects estimation), we seldom compute teacher value-added measures with multiple years of data on the same students, which would be required to identify these individual student effects.[4] Rather, teacher effects are typically obtained using up to a few years of data on *teachers* (so multiple cohorts of different students).

Even with this relatively parsimonious model, administrative data may be missing test scores or characteristics for some students, or some students may not be linked to teachers. In traditional regression analysis such as OLS estimation, student observations missing these data are omitted from the estimation sample, but consistent estimates can still be obtained. For consistency, whether data (on the outcome or the regressors) are observed or missing for a student can be related to the observed covariates that we control for (e.g., the lagged score, A_{it-1}, or student characteristics, X_{it}) but not unobserved elements of the error term (see Wooldridge, 2010, Ch 19). This is similar to the "missing at random" (MAR) assumption EVAAS methods are said to rely on (Wright et

[3] In the empirical work presented later, the set of student characteristics includes race/ethnicity, gender, free- and reduced-price lunch eligibility, limited English proficiency, disability, and days absent.

[4] Such approaches actually performed quite poorly in the simulations conducted in Guarino, Reckase, & Wooldridge (2015). See the paper for details on the reasons for this for each grouping/assignment scenario.

al., 2010), with the distinction that MAR generally assumes that the covariates related to whether data are missing are always observed themselves (Wooldridge, 2010, Ch. 19).

3.2.1 Common Methods for Estimating Teacher Effects

Given that the student heterogeneity term in equation (2) is generally ignored when estimating value-added models, the estimating equation for a given subject s can be written as:

$$A_{ist} = \tau_t + \lambda A_{ist-1} + X_{it}\beta + E_{ist}\gamma + v_{ist} \tag{3}$$

where $v_{ist} = c_i + e_{ist}$ is the composite error term. OLS on this equation will estimate teacher effects, $\hat{\gamma}$. We call this estimator *DOLS*, to reflect the OLS estimation of the teacher effects and acknowledge the dynamic (D) specification containing the lag score on the right-hand side. This can easily be extended to incorporate multiple lagged scores in multiple subjects. With this approach, to consistently estimate the vector γ, we need teacher assignment (E_{ist}) to be uncorrelated with the student heterogeneity term, c_i. This means, for example, that principals cannot assign students with higher (or lower) unobserved ability to more effective teachers.

The next two methods omit the teacher assignment dummies (E_{ist}); we then obtain estimates of teacher effectiveness from the student-level residuals. One approach is to estimate the abbreviated version (omitting E_{ist}) of equation (3) via OLS, and then calculate the teacher effects as the within-teacher averages of the student-level OLS residuals. We refer to this as the average residual (AR) method. Again, consistency requires that teacher assignment is not be based on the student heterogeneity. However, also note that any correlation between the lagged test score A_{ist-1} and the teacher assignment is not being partialled out of the teacher effects, so assignment based on prior scores also becomes problematic.

The last approach, which we will abbreviate to *EB*, involves empirical Bayes' estimation of this more parsimonious equation, obtaining the teacher effects from the shrunken residuals. The empirical Bayes' method is essentially a GLS or random effects approach, where the teacher effect estimates are effectively "shrunken" towards the mean teacher effect (Guarino et al., 2015).[5] The

[5] As described in Guarino et al. (2015), this method involves two stages, but is easily implemented in Stata with

so-called shrinkage takes teachers' class sizes into account, and thus aims to reduce the noisiness of the estimates from a small number of observations contributing to the estimation of the teacher effects. Like the AR method, consistent estimation relies on teacher assignment being uncorrelated with student heterogeneity and student-level covariates contained in the model (including prior achievement). The latter is also relevant to the EVAAS URM approach we focus on in this paper.

3.2.2 EVAAS Methods

3.2.2.1 EVAAS Univariate Response Model (URM)

Similar to the OLS and EB approaches discussed above, the URM estimates teacher effectiveness for a single grade and subject (e.g., 5th grade math). There are two key differences between the common approaches just described and the URM. First, the URM uses prior test scores from multiple years and subjects in lieu of student characteristics to account for past student achievement or other student characteristics affecting current achievement. Second, the URM allows for students to be missing some of these prior test scores. The URM's strategy for allowing incomplete test score data generates the complex nature of the approach, but the complicated steps do not necessarily develop a more robust estimator.

For instance, the consistency of the URM estimates relies on very similar assumptions regarding the nature of these missing data to the assumptions needed for OLS estimates to be consistent. In fact, when there are no missing data, there is a direct relationship between the URM and simpler standard linear regression techniques. Consider the simplest case where students have no missing test score data, students are randomly assigned to teachers, teachers have identical class sizes, and estimation is based on one cohort of students for teachers. Then the teacher effect estimates from the URM are identical (up to a constant) to OLS estimates. When students are nonrandomly assigned to teachers based on the included prior test scores, the estimates diverge. OLS partials out this assignment mechanism and consistently estimates the teacher effects while the URM does

the "xtmixed" command specifying a random component at the teacher level, and then post-estimation using the "predict , reffects" command to get the teacher random effects. The first stage estimates the normal maximum likelihood (with the random teacher effects in the error term) and the second stage applies the shrinkage factor to these teacher effects.

not partial out assignment and consequently produces biased estimates of the teacher effects. One goal of this paper is to derive and demonstrate these relationships.

In the discussion that follows, we first provide a detailed explanation of the URM approach, expanding on the description in Wright et al. (2010), and then illustrate how the URM compares with standard linear regression methods.

The URM estimating equation for subject s is:

$$A_{ist} = \tau + \kappa \hat{A}_{ist} + \gamma + \zeta_{ist} \tag{4}$$

where, compared to equation (3), the intercept τ does not have a time subscript, the lag score and student covariates have been replaced by a "composite score" $\hat{A}_{ist}$, and now the error term ζ_{ist} includes estimation error from using estimated components in $\hat{A}_{ist}$. This equation is estimated using empirical Bayes' to obtain the teacher effects γ. The γ contains the random effect for the student's teacher. Although this appears relatively simple, the composite score $\hat{A}_{ist}$ is the result of a multi-step process using all available lagged test scores (Wright et al., 2010), so the model is not as parsimonious as it appears. The composite score is essentially a different approach to a control, using multiple lagged test scores to predict a student's current score, and this prediction serves as a sort of sufficient statistic for the student's past inputs. The idea is explained by Sanders et al. (2009), "by including all of a student's testing history, each student serves as his or her own control."

The URM involves multiple steps to compute the composite score, with each step performed separately for every year of data (i.e., student cohort) that contributes to the estimated teacher effects. Thus, to estimate teacher effectiveness during a three-year period (i.e., based on three cohorts of students), each of the initial steps—up to and including computing the composite score—is done separately for the first, second, and third years of data. Then the final step—empirical Bayes' estimation of the teacher effects—is performed pooling the three years of data.

In computing the composite scores, the URM allows for many prior test scores across different subjects and years. For clarity, we focus our discussion on an example where we are using 1-year and 2-year lagged test scores for both reading (r) and math (m). The URM computes a composite

score in a specific subject (math shown in the equation below) as a linear combination of demeaned versions of the lagged test scores:

$$\hat{A}_{imt} = \hat{\mu}_{mt} \; + \; \hat{\beta}_{mt-1}\ddot{A}_{imt-1} \; + \; \hat{\beta}_{mt-2}\ddot{A}_{imt-2} \; + \; \hat{\beta}_{rt-1}\ddot{A}_{irt-1} \; + \; \hat{\beta}_{rt-2}\ddot{A}_{irt-2}. \tag{5}$$

In this equation, $\ddot{A}_{ist-y}$ (for the 1-year and 2-year lagged scores in subject s) denotes a "demeaned" y-year lagged test score in subject s for student i,

$$\ddot{A}_{ist-y} = A_{ist-y} - \hat{\mu}_{st-y} \tag{6}$$

In equations (5) and (6), the estimated means $\hat{\mu}_{st-y}$ are not the overall means of the test scores. Rather, each $\hat{\mu}_{st-y}$ (including y=0 for the current score) is the sum of two components: an average across teachers of the teacher-level mean score and an adjustment to account for students with missing test score data. We discuss each of these components in further detail below.

The weights in the composite score equation, $\hat{\beta}_{st-y}$, are coefficient estimates that maximize the correlation between the lagged scores and current score. With no missing data, $\hat{\beta}$ is essentially a vector of OLS coefficient estimates from the regression of A_{imt} on an intercept, A_{imt-1}, A_{imt-2}, A_{irt-1}, A_{irt-2}, and teacher assignment indicators. So, this particular step would produce coefficients on the lags from a DOLS-type equation that includes lagged test scores in multiple subjects, where teacher assignment is partialled out of the coefficient estimates.

Rather than use regression, however, the URM takes a different approach to estimation to allow for certain patterns of missing data. In general, the URM requires a minimum of three lagged scores and one of these must be the most recent lag in the same subject as the dependent variable. In our example, this means students must have records for A_{imt-1} and at least two scores out of the set of $\{A_{imt-2}, A_{irt-1}, A_{irt-2}\}$. The URM uses the EM Algorithm to estimate a variance-covariance matrix, $\mathbf{C}$, for calculating the coefficients $\hat{\beta}$ (rather than estimating these directly with a regression, which would omit observations with missing data).[6]

[6] The EM Algorithm is an optimization algorithm that iterates between the E step (expectation) and the M step (maximization) until the values of all parameters sufficiently converge. The Stata code for estimation as described here is: *mi impute mvn* $\ddot{a}_{m0}$ $\ddot{a}_{m1}$ $\ddot{a}_{m2}$ $\ddot{a}_{r1}$ $\ddot{a}_{r2}$, *emonly*.

The EM Algorithm estimation step of the URM is done separately for each year of data. It uses a transformation of the current and lagged test scores where the *teacher*-level means are subtracted from each score so that $\mathbf{C}$ is a "within-teacher" variance-covariance matrix. We denote these transformed scores used for the EM Algorithm estimation as:

$$\ddot{a}_{isy} = A_{ist-y} - \hat{\mu}_{jst-y} \tag{7}$$

where $\hat{\mu}_{jst-y}$ is the average of A_{ist-y} across the students i assigned to teacher j.

Then the within-teacher variance-covariance matrix obtained via the EM Algorithm, for each year, is:

$$\mathbf{C} = \left[\begin{array}{c|c} c_{\ddot{a}_{m0}\ddot{a}_{m0}} & \mathbf{c}_{\ddot{a}_{sy}\ddot{a}_{m0}} \\ \hline \mathbf{c}_{\ddot{a}_{m0}\ddot{a}_{sy}} & \mathbf{C}_{\ddot{a}_{sy}\ddot{a}_{sy}} \end{array}\right] = \left[\begin{array}{c|cccc} c_{\ddot{a}_{m0}\ddot{a}_{m0}} & c_{\ddot{a}_{m1}\ddot{a}_{m0}} & c_{\ddot{a}_{m2}\ddot{a}_{m0}} & c_{\ddot{a}_{r1}\ddot{a}_{m0}} & c_{\ddot{a}_{r2}\ddot{x}_{m0}} \\ \hline c_{\ddot{a}_{m0}\ddot{a}_{m1}} & c_{\ddot{a}_{m1}\ddot{a}_{m1}} & c_{\ddot{a}_{m2}\ddot{a}_{m1}} & c_{\ddot{a}_{r1}\ddot{a}_{m1}} & c_{\ddot{a}_{r2}\ddot{x}_{m1}} \\ c_{\ddot{a}_{m0}\ddot{a}_{m2}} & c_{\ddot{a}_{m1}\ddot{a}_{m2}} & c_{\ddot{a}_{m2}\ddot{a}_{m2}} & c_{\ddot{a}_{r1}\ddot{a}_{m2}} & c_{\ddot{a}_{r2}\ddot{x}_{m2}} \\ c_{\ddot{a}_{m0}\ddot{a}_{r1}} & c_{\ddot{a}_{m1}\ddot{a}_{r1}} & c_{\ddot{a}_{m2}\ddot{a}_{r1}} & c_{\ddot{a}_{r1}\ddot{a}_{r1}} & c_{\ddot{a}_{r2}\ddot{x}_{r1}} \\ c_{\ddot{a}_{m0}\ddot{a}_{r2}} & c_{\ddot{a}_{m1}\ddot{a}_{r2}} & c_{\ddot{a}_{m2}\ddot{a}_{r2}} & c_{\ddot{a}_{r1}\ddot{a}_{r2}} & c_{\ddot{a}_{r2}\ddot{x}_{r2}} \end{array}\right] \tag{8}$$

where the first matrix shows subdivided "blocks" of the matrix (to be referenced below), with $\ddot{a}_{sy}$ referencing the vector of lagged test scores in both subjects. The second matrix, with the lines for the subdivided blocks, is fully expanded to show each element of $\mathbf{C}$; the diagonal elements are the variance terms and the off-diagonal (symmetric) elements are the covariance terms.

The URM uses the elements of $\mathbf{C}$ to compute the set of within-teacher coefficient estimates, $\hat{\beta}_{st-y}$, by plugging into the familiar formula:

$$\beta_p = \mathbf{C}^{-1}_{\ddot{a}_{sy}\ddot{a}_{sy},p}\mathbf{c}_{\ddot{a}_{sy}\ddot{a}_{m0},p} \tag{9}$$

where p has been added to index each pattern of observed scores. With complete data for all students, the p index is not needed, and this equation would be equivalent to the OLS estimator from the regression of $\ddot{a}_{m0}$ on $\ddot{a}_{m1}$, $\ddot{a}_{m2}$, $\ddot{a}_{r1}$, $\ddot{a}_{r2}$ (or, equivalently, with the original scores, from the regression of A_{mt} on A_{mt-1}, A_{mt-2}, A_{rt-1}, A_{rt-2}, and teacher assignment indicators).[7] When

[7] Another equivalent representation for the case of full data, in the matrix notation often used for the OLS estimator,

students have incomplete records though, the formula in (9) allows us to separately estimate a unique vector of coefficients, $\hat{\beta}_p$, for each pattern of observed scores, using the subset of matrix **C** corresponding to the relevant observed scores. So, in our example, given that the first lag of the math score must be present, we would compute up to four vectors $\hat{\beta}_p$ to account for different missing scores. We could consider $p = 0$ for complete records, $p = 1$ for records missing A_{mt-2}, $p = 2$ for records missing A_{rt-1} and $p = 3$ for records missing A_{rt-2}. For students with $p = 0$ the full matrix is used, while for students with $p = 1$ (missing A_{mt-2}) the 3rd row and 3rd column are dropped.

The EM Algorithm estimation also produces means that contribute to the $\hat{\mu}_{st}$ in the composite score equation and the $\hat{\mu}_{st-y}$ underlying the transformed scores ($\ddot{A}_{ist-y}$) in (6). To be clear, in equations (5) and (6), the estimated mean is $\hat{\mu}_{st-y} = \hat{\mu}^{mtm}_{st-y} + \hat{\mu}^{EMm}_{st-y}$, which is not the overall mean of the lagged test score. The first term on the right-hand-side is the mean-of-teacher-means $\hat{\mu}^{mtm}_{st-y}$ for each y-year lagged score in subject s. In other words, the mean lagged test score is computed for each teacher and then the average over all teachers is taken.[8]

The second term on the right-hand-side is produced by the EM Algorithm.[9] It is an adjustment to the mean of teacher means to account for missing data—i.e., students with incomplete records. Since the EM Algorithm estimation step uses demeaned test scores (specifically, the teacher-demeaned scores $\ddot{a}_{st-y}$), this term is zero when there is complete data for all students. But when some students are missing test scores (and thus not contributing to the mean-of-teacher-means for the missing score), the estimated $\hat{\mu}^{mtm}_{st-y}$ may be biased and the URM includes the mean provided in the EM Algorithm output, $\hat{\mu}^{EMm}_{st-y}$, to reduce potential bias from missing lagged scores.

The transformation in (6) that subtracts these two mean components is similar to removing year effects, which would be done by instead subtracting the overall mean (or by including year dummies in a regression). Subtracting the mean-of-teacher-means ($\hat{\mu}^{mtm}_{st-y}$) instead ensures that the "average" teacher has a teacher effect of zero and the EM Algorithm component ($\hat{\mu}^{EMm}_{st-y}$) corrects

is $\beta = (\ddot{a}'_{sy}\ddot{a}_{sy})^{-1}\ddot{a}'_{sy}\ddot{a}_{m0}$, where $\ddot{a}_{sy}$ contains $\ddot{a}_{m1}$, $\ddot{a}_{m2}$, $\ddot{a}_{r1}$, $\ddot{a}_{r2}$, or $\beta = (X'X)^{-1}X'A_{mt}$ where X includes A_{mt-1}, A_{mt-2}, A_{rt-1}, A_{rt-2}, an intercept and teacher assignment indicators.

[8] To the best of our knowledge–based on the description in Wright et al. (2010)—this average per teacher is across all of the teacher's students, even if the teacher teaches multiple classes. Regardless, this distinction is not important for our theoretical or empirical results and conclusions.

[9] The EM Algorithm estimates both the variance-covariance matrix discussed earlier as well as the means, $\hat{\mu}^{EMm}_{st-y}$, used here.

for potential bias in the mean-of-teacher-means from students missing test scores (Wright et al., 2010).

Finally, we compute the so-called *composite score*, $\hat{A}_{imt}$, according to equation (5). The composite score is the sum of the "adjusted mean" of the current math score ($\hat{\mu}_{st} = \hat{\mu}_{st}^{mtm} + \hat{\mu}_{st}^{EMm}$) plus a weighted average of transformed lagged scores $\ddot{A}_{st-y}$, with the weights being the coefficient estimates, $\hat{\beta}_p$. The composite score is a prediction of the current score (A_{imt}) based on the student's past test scores and assuming the student has the "average" teacher in the current year (Wright et al., 2010).

After the composite scores are obtained, the final step in computing the teacher effects is the empirical Bayes' estimation of equation (4)—as mentioned above.

Note that this discussion has focused on estimating teacher effects for math teachers. If one wished to estimate teacher effectiveness in, say, reading, then the outcome variable would be the current reading score, and the composite score would constitute a predicted reading score. While the same lagged scores could be used to obtain the composite score, the estimated elements (i.e., the $\hat{\mu}_{st}^{mtm}$, $\hat{\mu}_{st}^{EMm}$, and $\hat{\beta}_{st}$) would be different because they would be based on predicting the current *reading* score, using the sample of students satisfying the corresponding data requirements. So, in this respect, the URM is similar to the common VAM approaches that estimate teacher effects separately by subject (and grade).

3.2.2.1.1 Relating the EVAAS URM to Other Approaches

Unlike traditional regression-based VAM methods, the EVAAS approach handles at least some missing data patterns. It also uses empirical Bayes' shrinkage in the final step in order to account for teachers having different numbers of students. But is EVAAS very different from the standard regression estimators? In practice, differences in the estimated teacher VAMs may be minor. In fact, in the simplest scenario the two approaches yield numerically identical teacher effect estimates.

In the simplest setting, there are no missing data and only one year of data is used. Either shrinkage is not used or the number of students per teacher is identical, in which case shrinkage

simply multiplies all of the teacher VAMs by the same constant. With a single year of data, a simple extension of DOLS to allow other lagged test scores comes from OLS estimation of the equation

$$A_i = X_i\beta + E_i\gamma + v_i, \tag{10}$$

where X_i includes all lagged test scores in various subjects and E_i is the vector of teacher assignment dummies. For simplicity, we drop the time subscripts indicating subject and year. Technically, the OLS estimates from (10) are not the DOLS estimates described earlier because (10) includes other lagged test scores. But adding additional lags of the same and other subject test scores is a small modification, and produces no extra conceptual or computational difficulties. We could legitimately refer to the OLS estimates from (10), as the motivation is the same: control for factors that predict current test scores and may be correlated with teacher assignment.

From the Frisch-Waugh partialling-out theorem, the OLS coefficients on the lagged test scores, $\hat{\beta}$, can be obtained in three steps:

(i) Regress A_i on E_i and obtain the residuals, $\ddot{A}_i$. Now, $\ddot{A}_i = A_i - E_i\hat{\eta}$ where, because the E_i are teacher assignment dummies, $\hat{\eta}_j$ is the average of the A_i (current test score) for teacher j. Therefore, $\ddot{A}_i$ is student i's test score deviated from the average test score for the student's teacher.

(ii) Regress each lagged test score in X_i on E_i and collect the vectors of residuals, $\ddot{X}_i$. Just as with $\ddot{A}_i$, each element of $\ddot{X}_i$ is one of student i's lagged test scores deviated from the mean for student i's teacher.

(iii) Run the regression

$$\ddot{A}_i \text{ on } \ddot{X}_i$$

and obtain $\hat{\beta}$.

In other words, when the regression is restricted to a single year, and there are no missing data,

the OLS and URM estimates of β are identical; the URM simply performs the partialling out of teacher assignment in a separate step, rather than using the full regression in (10).

As described earlier, the next step in the URM is to construct the composite score in equation (5). But the composite score $\hat{A}_i$ can be written as

$$\hat{A}_i = X_i\hat{\beta} + \hat{\psi}, \tag{11}$$

where $\hat{\psi}$ depends on $\hat{\beta}$ and the overall means of the test scores. Now, the equation used to obtain the teacher effects is

$$A_i = \kappa\hat{A}_i + E_i\gamma + error_i, \tag{12}$$

where $error_i$ includes estimation error because $\hat{A}_i$ depends on $\hat{\beta}$. The URM approach applies empirical Bayes' to (12), but that is simply to shrink the estimates of γ towards the average teacher effect. Without shrinkage, or with the same number of students per teacher, we just apply OLS to (12). Again, without missing data, we know the result by the algebra of OLS: $\hat{\kappa} = 1$ and $\hat{\gamma}$ will be identical to what is obtained from (10). The argument is simple. We know the DOLS estimates minimize the sum of squared residuals, and yet we know the $\hat{\beta}$ obtained from the URM is identical to the $\hat{\beta}$ from DOLS. So one cannot do any better by choosing $\hat{\kappa}$ different from unity and $\hat{\gamma}$ as the DOLS coefficients. The additive constant in (11) changes nothing because the DOLS regression, with a full set of teacher dummies, effectively estimates an intercept. However, when the coefficient on the composite score is estimated by EB, the coefficient is not unity, which breaks the equivalence. In fact, this seems to cause bias. So if (12) were estimated by OLS then the URM and OLS estimates would be the same.

So how does the EVAAS URM generally differ from OLS? Even if we assume no missing data and ignore shrinkage in the final step of EVAAS, there is a difference with more than one time period. With OLS estimation, typically one would augment (10) by adding year dummy variables, and then the partialling out in steps (i) and (ii) are also done via pooled regression on the year dummies and teacher effects. By contrast, EVAAS does a teacher-year demeaning, which is the

same as allowing a full set of interactive effects between the year dummies and teacher dummies. However, after obtaining the composite test scores, EVAAS then pools the data (say, over three years) to obtain a single teacher effect for each teacher. The resulting URM estimates cannot be characterized as coming from an OLS regression, but the difference may not be great. If one thought that teacher effects vary over time, then one might estimate an equation by OLS separately for every year. This would precisely achieve the partialling out used by the URM. Then, given the $\hat{\gamma}_t$, one must decide how to combine these into single teacher effect estimates. The URM has one way to do that, but there are others, such as using a weighted average with weights chosen to reflect the relative precision of the $\hat{\gamma}_t$ across different years.

Whether allowing for teacher-year specific effects is important is mainly an empirical issue, but it would not be surprising to find that adding year dummies to the OLS regression, and imposing constant teacher effects across time, generally produces similar results. Often OLS will provide good estimates of average partial effects when interaction terms are present but omitted from regression analyses. See, for example, Wooldridge (2010, Chapter 6).

3.2.2.2 EVAAS Multivariate Response Model (MRM)

The MRM is a multivariate, longitudinal, linear mixed model where the full set of observed scores—meaning all subjects and all years—is fitted simultaneously. Hence, this model simultaneously estimates teacher effects for these separate subject/grade/years, whereas the URM and other VAMs discussed earlier estimate teacher effects for a single subject/grade (possibly pooling over multiple years). With the joint modeling of scores across various grades and years, the MRM requires vertically scaled tests, or conversion of scale scores to NCEs (Normal curve equivalents) (Wright et al., 2010). To show this, we begin with a set of equations that illustrate the need for the appropriately scaled test scores as well as the description "layered teacher model".

As portrayed in Ballou, Sanders, & Wright (2004), a student's set of, say, math scores, must

satisfy the following equations:

$$\begin{aligned} y_t^3 &= b_t^3 + u_t^3 + e_t^3 \\ y_{t+1}^4 &= b_{t+1}^4 + u_t^3 + u_{t+1}^4 + e_{t+1}^4 \\ y_{t+2}^5 &= b_{t+2}^5 + u_t^3 + u_{t+1}^4 + u_{t+2}^5 + e_{t+2}^5 \end{aligned} \tag{13}$$

where y_t^g is the test score (gain) for grade g in year t and b_t^g is the district-level average test score for grade g in year t. u_t^g is the grade g teacher's input to the student's test score in year t and e_t^g is a student-level idiosyncratic error term for the grade g score in year t.

The year subscript on the teacher effects show that teacher effects vary over the years, so this approach is estimating a the effect of each teacher in each year (i.e., teacher/year effects). (More precisely, when we consider the full model with multiple subjects, the approach actually estimates teacher/year/subject effects). However, for a given student, the effects of past teachers do not change over time—a student's 3rd grade teacher's contribution to their 4th grade score is the same as that same teacher's contribution was to the student's 3rd grade score. In other words, a teacher's effect on a student's achievement does not diminish as the student progresses through grades. This highlights the importance of using vertically scaled test scores. The meaning of the teacher effect (resulting in test score gain/loss) must be the same in any grade as well as throughout the test score distribution. So moving down the set of equations in (10) from the first line to the second, an additional "layer" (teacher effect from the next teacher and next idiosyncratic shock) is added in each year, motivating the nickname "layered teacher model" commonly used to describe the MRM.[10]

The more technical representation of the MRM begins with presenting the linear mixed model

[10] Another representation of the MRM, as given in Wright et al. (2010), is the algebraic equation, $y_{ijkl} = \mu_{jkl} + (\sum_{k^* \leq k} \sum_{t=1}^{T_{ijk^*l^*}} \omega_{ijk^*l^*t} \times \tau_{ijk^*l^*t}) + \epsilon_{ijkl}$, where the the inner summation adds across all teachers the student has in a given subject/grade/year with the $\omega_{ijk^*l^*t}$ term capturing the fraction of time spent with a particular teacher, and the outer summation is where the "layered" aspect comes in, adding the cumulative teacher effects over previous grades and years in the same subject. Note that this representation highlights the ability to accommodate team-teaching or students switching teachers during the year; this is possible with other approaches as well, but is rarely done in practice.

(with notation similar to earlier sections of our paper):

$$\mathbf{A} = \mathbf{X}\beta + \mathbf{E}\gamma + \epsilon. \tag{14}$$

Now $\mathbf{A}$ contains the set of *all* test scores (gain scores), meaning all subjects tested over all grades and years for all students during the period being studied (up to 5 years). The matrix $\mathbf{X}$ is comprised of subject/grade/year indicators, and β is the vector of coefficients that are treated as fixed. The $\mathbf{E}$ matrix contains teacher/grade/subject/year assignment indicators, and the teacher random effects are contained in γ. The joint distribution of γ and ϵ is such that $E(\gamma) = E(\epsilon) = \mathbf{0}$ and the variance-covariance matrix is block diagonal with $Var(\gamma) = \mathbf{G}$ and $Var(\epsilon) = \mathbf{R}$ and $Cov(\gamma, \epsilon) = \mathbf{0}$. Estimates of β and γ are obtained as solutions to Henderson's mixed model equations (see Wright et al. (2010) for explicit equations) so that the resulting estimator for the teacher effects is:

$$\gamma^* = \mathbf{GE'}(\mathbf{EGE'} + \mathbf{R})^{-\mathbf{1}}(\mathbf{A} - \mathbf{X}\beta^*) \tag{15}$$

where β^* is the GLS estimator for $\mathbf{A}$ on $\mathbf{X}$ (so $\mathbf{A} - \mathbf{X}\beta^*$ is the vector of GLS residuals) and $\mathbf{GE'}(\mathbf{EGE'} + \mathbf{R})^{-\mathbf{1}}$ is the shrinkage factor.

Although the shrinkage factor may look complicated in matrix form, the idea is the same as that for the shrinkage used in the EB and URM approaches.[11] The greater the student-level noise (i.e., the larger the variances along the diagonal elements of $var(\epsilon) = \mathbf{R}$), the more the estimated residuals ($\mathbf{A} - \mathbf{X}\beta^*$) are shrunk towards the mean (zero).[12] [13] Since the district mean (gain) scores are estimated in β, the teacher effects are deviations from the district mean, and gains attributed

[11] For basic intuition, consider Ballou, Sanders, & Wright's (2004) example with the simple case where γ contains one teacher effect so the shrinkage factor reduces to the reliability ratio, $\frac{var(\gamma)}{var(\gamma)+[var(\epsilon)/N]}$.

[12] $\mathbf{R}$ captures the within-student covariances in student test score residuals, ϵ. Sorting the scores, $\mathbf{A}$, by student, $\mathbf{R}$ is block diagonal with a block $\mathbf{R_i}$ for each student, and all other elements zero reflecting the imposed zero correlation *between* students. To form this matrix, consider an overall covariance matrix, $\mathbf{R_0}$, that contains a row and column for each subject/grade, so covariances among subjects and grades are assumed to be the same for all years (cohorts), but is otherwise unrestricted. Similar to the URM's accommodation of incomplete records, in the MRM each student has a block $\mathbf{R_i}$ composed of the subset of elements in the overall covariance matrix, $\mathbf{R_0}$, that correspond to the subject/grades for which the student has test scores, regardless of whether the student is linked to a teacher for these scores. Hence, the $\mathbf{R}$ matrix allows the MRM to incorporate information from all available scores from each student.

[13] $\mathbf{G}$ captures the variance of teacher effects, and is block diagonal with a block for each subject/grade/year. The (block) diagonal form reflects the assumption that teacher effects are not correlated across subjects or years, allowing teachers' effectiveness to vary from year to year and subject to subject. Each block has the form $\sigma^{\mathbf{2}}_{\mathbf{jkl}}\mathbf{I}$ where $\sigma^{\mathbf{2}}_{\mathbf{jkl}}$ is the teacher variance for the j^{th} subject in the k^{th} grade in the l^{th} year, allowing the variance of teacher effects to vary across subjects, grades, and years.

to teachers are estimated by adding the teacher effect to the district mean (gain).

As noted in Wright et al. (2010), when $\mathbf{G}$ and $\mathbf{R}$ are known, γ^* is the best linear unbiased predictor (BLUP) of γ, β^* is the best linear unbiased estimator (BLUE) of β, and the solution is equivalent to GLS. If γ and ϵ are Normal, then the solution is MLE. Generally $\mathbf{G}$ and $\mathbf{R}$ are not known so estimates are used instead; the solution approaches MLE as the estimated $\mathbf{G}$ and $\mathbf{R}$ approach their true population values.

This approach is computationally burdensome—and hence generally limited to district-level analysis rather than state-level—so we do not estimate the MRM in this paper. Further, many characteristics of the approach, such as the joint modeling of scores from different subjects, or the accommodation of missing data, have been evaluated in other studies (e.g., Lockwood et al., 2007; McCaffrey et al., 2011).

3.3 Prior Literature Evaluating EVAAS Methods

While the research literature on estimating teacher effectiveness has been growing rapidly in recent years, only a handful of these studies have ever implemented either of the EVAAS teacher models in simulations or using administrative data (Lockwood et al., 2003; Ballou, Sanders, & Wright, 2004; McCaffrey et al., 2004; Lockwood et al., 2007; McCaffrey et al., 2008; Rose, Henry, & Lauen, 2012). The majority of these studies focus on a specific assumption or characteristic of the EVAAS MRM, such as the complete persistence of teacher effects, and only mention in passing that this is part of the EVAAS method. To our knowledge, Rose, Henry, & Lauen (2012) is the only other study to evaluate the URM, among several other estimators they consider. We focus specifically on the URM, providing a more detailed discussion of the method and also how the method relates to other (simpler) approaches. In particular, we show that standard linear regression using OLS is a simpler—and in some cases more robust—alternative to this EVAAS method.

There are several papers made available by the SAS Institute (SAS White Papers) that discuss the theoretical advantages of the EVAAS methods, and some also evaluate the performance of the EVAAS methods (e.g., Sanders, 2006; Wright, Sanders, & Rivers, 2006; Wright, 2010). These papers tend to focus on the scaling of the test scores, measurement error in the test scores, and

missing data.

The test-score scaling issue stems from the fact that some approaches—such as the MRM and other gain-score VAMs—require test scores to be vertically scaled, not just from grade-to-grade, but in a way that leaves the meaning of a 1-unit change in score the same at any point in the distribution. The URM and other lag-score VAMs, however, do not require such scaling, and, for the approaches that do, the entire issue can be circumvented by converting the scores to normal curve equivalents (NCEs) (Wright et al., 2010).

The second concern is that the measurement error in the lagged test scores will cause bias in the estimates of teacher effects, lending also to instability in the estimates. In a paper aimed at evaluating a standardized gain model and student growth percentile model, the URM and MRM are also estimated for comparison, and their estimated teacher effects are shown to have smaller correlations with the percent of students in a teacher's class who are eligible for free- and reduced-price lunches (Wright, 2010). Hence, the proposed solution for measurement error bias is to include multiple lag scores (at least three) to mitigate the attenuation bias, as the measurement error tends to average out (e.g., Wright, 2010). However, this can worsen missing data issues in some approaches, lending advantage to the EVAAS MRM, as it uses all possible test scores (Wright, 2010).

This leads to the last of the major concerns—students with incomplete test score records. Both the MRM and URM incorporate ways to mitigate missing-data issues, with the MRM including students with any observed test scores, while the URM requires at least three prior scores (Wright et al., 2010).[14] Other noted features include the use of shrinkage estimation (empirical Bayes') (Sanders, 2006) and the MRMs layering of all past, present and "future" test scores (Wright et al., 2010), both of which are thought to improve stability of teacher effect estimates.

In one of the early papers to implement EVAAS, McCaffrey et al. (2004) propose a "general model" which encompasses several other VAMs as a special case. Their approach for the general model is similar to the EVAAS MRM, differing in that it allows for the inclusion of student covariates

[14] The distinction is also made that the EVAAS methods require data to be "missing at random" (MAR) as opposed to other methods, which require the data to be "missing completely at random" (Wright, 2010; Wright et al., 2010). However, as noted above, the consistency of an OLS approach such as DOLS only relies something similar to MAR.

and does not impose complete persistence of teacher effects. In a theoretical discussion, which is supported by their simulation and empirical evidence, they find that omitting student covariates results in biased teacher effect estimates when the distribution of covariates differ by school (and school effects are omitted), but that in other cases—when the distribution of covariates differs, say, by classroom—the use of within-student correlation mitigates this bias. Another feature of the EVAAS MRM is the assumed complete persistence of teacher effects (e.g., the contribution of the 3rd grade teacher persists undiminished for the scores in all subsequent grades). Given that this assumption is not theoretically or empirically justified, it is perhaps unsurprising that McCaffrey et al. (2004) find no evidence to support this, estimating the persistence parameters to be 0.1–0.3 (none of which are significantly different from zero). However, with the small simulation and limited administrative data (678 students from 5 elementary schools in a single suburban district, with free- and reduced-price lunch eligibility as the only covariate), even the authors admit the evidence on both the omission of student covariates as well as persistence is insufficient and warrants future research.

More recently, Lockwood et al., (2007) develop a Bayesian framework which is better suited to scale to large datasets than the maximum likelihood methods used by McCaffrey et al. (2004). Further, they expand the analysis by now jointly modeling reading and math scores, and explore the implications of using different approaches for addressing missing data. They use five years of data on one cohort of students from a large urban district, as well as simulations, and, again, find persistence estimates are substantially less than 1. They conclude that joint versus marginal modeling does not affect teacher effects significantly (rank correlations between teacher effects from joint and marginal models are greater than 0.99 for the variable persistence model and greater than 0.97 for the complete persistence model). They also note that their results are robust to which method is chosen to handle missing data. To further examine the implications of missing data, McCaffrey & Lockwood (2011) extend the approach to explicitly allow for data to be missing not at random, but find little impact on the estimated teacher effects, suggesting that violations of the missing at random assumption (MAR) may not be problematic.

Also using their proposed model (a generalization of the MRM), Lockwood & McCaffrey (2007) use simulations to explore how the potential bias from omitting student covariates changes, depend-

ing on the assumptions one is willing to make about the way in which the student heterogeneity relates to the measures, and whether a random effects or fixed effects approach is taken. They argue that even when omitted student heterogeneity is related to other variables, the GLS estimator arising from the mixed model approach (similar to the MRM) which jointly models test scores from different subjects has additional information available on the heterogeneity, and this increases efficiency and also reduces the bias (relative to modeling a single subject).[15]

Ballou, Sanders, & Wright (2004) also focus on the issue of omitting student covariates, but do so specifically with the EVAAS MRM.[16] The authors obtain the usual EVAAS estimates of the teacher effects as well as estimates from a modified EVAAS approach that controls for student's FRL eligibility, non-white race, gender, and interactions between these covariates. This modification is implemented in a first stage to obtain quasi-residuals from estimation using the gain score as the dependent variable and student characteristics and teacher-by-year indicators as covariates. Then they use these quasi-residuals in the usual EVAAS estimation. They find that the estimated teacher effects do not differ substantially, with high rank correlations between estimates and also similar numbers of teacher classified as "excellent". To explore whether this result is due to the history of prior scores accounting for student covariates, they also compare the $\mathbf{R}$ matrix for the original and modified EVAAS approaches, finding the elements to be approximately 18% smaller in the latter. They conclude that including prior test scores does control for "much" of the information contained in student-level covariates. While this is suggestive evidence in support of omitting student covariates, this and earlier evidence is limited to the MRM, as none of the studies mentioned here have included the URM.

A recent paper by Rose, Henry, & Lauen (2012), however, not only provides comparisons using the EVAAS URM, but actually provides these comparisons with a broader set of (nine) value-added approaches. The paper discusses assumptions and implications of violations, and also provides simulation and statewide empirical evidence using three years of administrative data from

[15] For this paper, they use maximum likelihood methods for the mixed models, but also note that separate simulations using a Bayesian approach does NOT??? change the results substantively.

[16] Interestingly, the authors discuss the omission of student covariates as a virtue of the EVAAS MRM, in that this reduces data requirements. Indeed these data (such as FRL eligibility, gender, race, absences, etc.) can be missing for some students, but these data are generally available, and a similar argument could be made for the many VAMs that utilize only one or two prior test scores, as using a complete history of test scores can be quite onerous when using large administrative datasets. For example, Lockwood et al. (2007) note that only 20% of their sample of students had a complete set of reading and math scores over the 5 years (grades) used.

North Carolina. The estimation approaches include the URM, three HLM approaches which also treat the teacher effects as random, two which take the within-teacher average of the residuals, and three which treat the teacher effects as fixed. The specifications vary with respect to the number of (and subject of) lagged scores, school effects, and time-constant student covariates. They find high agreement among most of the approaches, but overall recommend the URM, the two approaches that treat the teacher effects as random and account for a school random effect, as well as a student fixed effects approach.[17]

3.4 Simulation

3.4.1 Simulation Design

We conduct simulations to assess the performance of the DOLS, EB, AR, and URM estimators under various student grouping and assignment scenarios. This allows us to know the "true" teacher effect (which we generate), and then evaluate the ability of each of the estimators to capture this effect—something not possible with administrative data. We generate data for 3 cohorts of 800 students each, creating a current score and two lagged scores for each student. For our analysis, we focus on a single grade, so using one observation per student, but 3 cohorts of students per teacher. The simulations are designed with elementary grades in mind, so we can think of this setting as looking at 5th grade students and teachers. Class size is set to 20, for a total of 40 teachers.

To generate the test scores, we first obtain a baseline score (i.e., the first grade tested) drawn from a standard normal distribution. Each of the subsequent test scores, A_{it}, is then generated according to the equation below:

$$A_{it} = \lambda A_{i,t-1} + \gamma_{it} + c_i + u_{it} \tag{16}$$

where $A_{i,t-1}$ is lagged achievement, γ_{it} is the teacher contribution to the current score (the true teacher effect), c_i is the time-constant unobserved student effect, and u_{it} the idiosyncratic error. The decay parameter, λ, is set to either 0.5 (substantial decay) or 1 (no decay). The correlation

[17] Rose, Henry, & Lauen (2012) also note that their results (and hence conclusions) for the DOLS estimator may differ from those in Guarino, Reckase & Wooldridge (forthcoming) due to simulation design.

between lagged achievement and the student fixed effect is 0.5. The three random parameters are drawn from normal distributions: student fixed effect $c_i \sim N(0, .5^2)$, teacher effect $\gamma \sim N(0, .25^2)$, and the idiosyncratic error $u_{it} \sim N(0, 1)$ (so their respective proportions of the total variance in test scores are 19%, 5%, and 76%).

To look at nonrandom sorting of students, we make the distinction between *grouping* (how students are grouped into classrooms) and *assignment* (how students are assigned to teachers), allowing for students to be, say, grouped based on prior achievement levels, but then randomly assigned to teachers. We look at grouping based on the lagged score (referred to as dynamic grouping), based on the original baseline score (a form of "static" grouping referred to as baseline grouping), and based on the student individual heterogeneity (another form of static grouping, referred to as heterogeneity grouping). We look at three different assignment mechanisms for each of these grouping scenarios: random assignment, positive assignment (e.g., better students to better teachers), and negative assignment (e.g., struggling students to better teachers). In the cases of nonrandom assignment, the assignment is not perfectly separating students in rank order of, say, lagged achievement, rather assignment is noisy with the noise being drawn from a standard normal distribution.

In addition to varying the grouping and assignment mechanisms and the decay parameter (λ), we also conduct simulations using larger teacher effects, with $\gamma \sim N(0, .6^2)$ (and $c_i \sim N(0, .5^2)$, so their respective proportions of the total variance in test scores are 21% each). We conduct 100 Monte Carlo repetitions for each grouping-assignment-parameter scenario.

We examine the performance of four of the estimators discussed above (DOLS, AR, EB, EVAAS URM). For the first three estimators we consider a "common" specification, similar to equation (3), where the covariates include a lagged test score, and in the case of DOLS, teacher assignment indicators. (We do not incorporate effects for student characteristics into the simulation.) For the URM, we use base the composite score on this same lagged test score as well as a two-year lagged test score.[18] As discussed above, we also estimate specifications that "mimic" the EVAAS URM

[18] Incorporating further lagged scores or lagged scores in other subjects would not contribute substantively to our evaluation of the theoretical implications of sorting or assignment for the URM estimator, as these would constitute the same issues as having one vs. two lags. We prefer to present the simple case of two lags to facilitate transparency in our simulation design and results.

approach, using DOLS, AR, and EB, to illustrate where divergences in the performance of the estimators is coming from. Hence, for the simulations, this means including both the one-year and two-year lagged test scores in the estimating equation. For all estimators and specifications, we estimate the teacher effects first using one year (cohort) of data, and then estimate them pooling over three cohorts (years) as well.

Our first metric for evaluating the performance of these estimators is the Spearman rank correlations between the estimates and the true teacher effects, to examine their ability to uncover the true effect. We also look at the correlations between the estimates obtained via the URM and those from the other estimators, to look at how similarly they rank teachers.

3.4.2 Simulation Results

We first assess the ability of each of the estimators to uncover the true teacher effect, looking at the correlations between the estimated and true effects. For our main results, we focus on the "small" teacher effects, which account for 5 percent of the variation in test scores. In practice, it would be convenient to use one year of data (i.e., one cohort of students) to evaluate teacher effectiveness, so Table C1 provides the rank correlations between the true teacher effects and the estimated teacher effects in this setting. Panel A shows the case of substantial decay ($\lambda = .5$) and Panel B the case of complete persistence ($\lambda = 1$). Within each panel, 10 grouping-assignment scenarios are explored. The estimators considered first are DOLS, AR (average residual), and EB on the "common" specification which controls only for one lag score (in addition to the teacher effects in the case of DOLS). The next set of columns are based on approaches which also include a two-year lagged score, to mimic the information in the composite score of the URM.

Table C1 shows that under random grouping and random assignment, the rank correlations are 0.69 for all estimators, and nonrandom grouping does not cause large departures from this, as long as assignment to teachers is random. The estimators actually perform best in the positive assignment cases, in particular when grouping is based on the student heterogeneity, with rank correlations ranging .78–.80, a result arising from bias that expands the distribution of estimated teacher effects, making it easier to distinguish between teachers (see Guarino, Reckase, & Wooldridge (2015) for a

more detailed discussion of this result). Conversely, the estimators perform the worst when students are grouped on heterogeneity and then negatively assigned to teachers, with rank correlations ranging .41–.43 when $\lambda = .5$ and .45–.46 when $\lambda = 1$.

Also evident in Table C1 is the close relationship between the EVAAS URM and using EB to estimate a specification with the same lagged test scores, as the correlations in the URM and EB-mimic columns are nearly identical. Further, we see that under dynamic grouping with positive or negative assignment, although all estimators perform worse relative to random assignment, DOLS performs substantially better than AR, EB, or URM, regardless of the value of λ. This arises from the fact that these approaches are not correctly partialling out the assignment mechanism from the teacher effects. EB and the URM both are closer to DOLS than AR, though, because as the number of students per teacher gets larger, the Empirical Bayes' estimates of the teacher effects (underlying EB and URM) will get closer to DOLS (see Guarino et al. (2015) for a more detailed discussion of this result that the random effects (RE) estimates will converge to the fixed effects estimates as the sample size increases).

Although using one cohort of students is convenient, in practice multiple cohorts are often used, so we also present results from using three cohorts of students (i.e., three years of data on teachers) to estimate teacher effectiveness. Given that this is increasing the amount of information on teachers (and teacher effects do not vary by year in our simulation), we expect the performance of all of the estimators to improve. The rank correlations in Table C2 show this improved performance, but the results also follow the same relative performance patterns across scenarios and estimators. The correlations under random grouping and random assignment are now larger at .84. In the case of grouping based on student heterogeneity with positive assignment to teachers, the correlations are now .89 for all estimators. When students are instead negatively assigned to teachers (based on heterogeneity), the correlations are .52–.56 when $\lambda = .5$ and .57–.58 when $\lambda = 1$. Under this scenario, the correlations for the "mimic" specification estimators are slightly larger than those from the "common" specifications when $\lambda = .5$, but this result comes from the amount of decay; when $\lambda = 1$ there is no motivation for controlling for a second lagged score. Again we see the nearly identical performance of the URM and EB-mimic. The issue of poor performance of AR, EB, and URM under nonrandom assignment based on the lagged score remains. Again, the URM and EB

estimators perform more similarly to DOLS than AR exhibiting the convergence of the random effects approach (EB, URM) to the fixed effects approach (DOLS). AR performs the worst because the assignment mechanism is not partialled out at all.

Table C3 shows the correlations for the DOLS, AR, and EB estimates compared with the EVAAS URM estimates under the various grouping and assignment scenarios, first when estimation uses one cohort of students and then when estimation is based on three cohorts of students. Agreement with the URM is high (.99–1.00) for all estimators under most scenarios, with the smallest correlations being for the cases of nonrandom assignment based on the lagged test score. In these cases, the correlations between DOLS and the URM are still above .90 (ranging .92–.97), reflecting the difference in accounting for the assignment mechanism discussed above.

3.4.3 Sensitivity of Simulation Results

While some sensitivity analyses were presented as part of the main results (e.g., using one versus three cohorts of students, or choosing $\lambda = .5$ versus $\lambda = 1$), we also conducted simulations with larger teacher effects. In this case, the teacher effect and the student heterogeneity each account for about 21% of the total variation in test scores. With the teacher effects accounting for more of the variation in the test scores, we naturally expect the performance of the estimators to be improve. As shown in Table C4, this is certainly the case. Still, the results follow the same general patterns discussed for the small teacher effects case. Similarly, Table C5 shows the similar agreement between DOLS, AR, EB and the URM, with the correlations being of similar magnitude except for the dynamic grouping and nonrandom assignment cases. The agreement in these cases is higher, but still illustrates the divergence in estimates that arises from how the approaches account for the assignment mechanism.

3.5 Empirical Analysis

3.5.1 Administrative Data

We use administrative data on students in grades 5 and 6 during years 2002–2007 in a large urban anonymous district.[19] Similar to our example used in the EVAAS URM discussion, we focus on math scores as the outcome and use one-year and two-year lagged math and reading scores as covariates in some specifications. The data contain information on student race/ethnicity, days absent, gender, disability, limited English proficiency (LEP), and free- or reduced-price lunch eligibility (FRL). We exclude students who are not linked to mathematics teachers, students who are assigned to classes (i.e., teacher/year groups) with fewer than 10 students, and students who were retained or have duplicate grade-year observations. All estimations also require that students have, at a minimum, a current math score and a one-year lagged math score.

Sample characteristics and average scores for the fifth and sixth grade samples are provided in Table C6 for the students with data satisfying the minimum sample inclusion requirements just described; these estimation samples cover years 2002–2007. The first set of descriptives in Panel A are for the sample of fifth grade students, while Panel B contains the descriptives for the sixth grade sample. Across grade, the average student characteristics are very similar, with about 61% of the students being Hispanic, 28% Black, and 50% are female. Approximately 52% are classified as limited English proficient (LEP) and about 70% are FRL-eligible. The sample sizes are also provided for each variable in the table, to illustrate how the samples could change depending on which lagged test scores are included. For example, adding a two-year lagged math score in a regression would mean 3,433 (3.1%) fifth grade students are omitted. For sixth grade, the sample falls by 3,412 (3.4%) with the addition of two-year lagged math. This indicates that including a longer history of scores does impose data restrictions, though as discussed above, the URM is able to relax these restrictions somewhat.

We estimate teacher effects separately for 5th and 6th grade, focusing on math teachers only (so we use math scores as our outcome variable). We compute estimates using one or two years of data for teachers (i.e., three cohorts of students). Similar to the approach for the simulations,

[19] Our data sharing agreement does not allow us to name the district or state.

we estimate several specifications using AR, DOLS, and EB. One set of specifications includes student characteristics and either the 1-year or both the 1-year and 2-year lagged math scores. The other specifications are designed to be more similar to the URM (and omit student characteristics); one specification includes the 1-year and 2-year lagged scores in math and reading, and the other specification uses the composite scores as the only regressor.

3.5.2 Empirical Results

With the administrative data, we estimate a few specifications with each of the three "common" estimation methods considered in this paper (DOLS, AR, and EB). The first specification is the lag-score specification shown in equation (3), which controls for the 1-year lagged math score, other student-level covariates, and year effects.[20] The second specification is augmented with a 2-year lagged math score also. The third specification omits student covariates but includes the same lagged scores as the composite score computed for the URM, hence attempting to "mimic" the information used in the URM estimation. The last specification uses the composite score itself as the only covariate (so when using EB estimation, this is identical to the URM). We consider teacher effects computed based on one year of data or pooled over two years of data, covering the years 2002-2007. We then examine agreement among the estimators in each year and present results on average agreement during this time period.

In Table C7, we provide average Spearman correlations between the EVAAS URM estimates and those from each of the other estimator/specification combinations. Within each specification, the rank correlations do not change significantly when pooling over an additional year of data for estimation and also do not differ substantially between estimators. In column [1], the correlations show that agreement with the URM is slightly better in the 6th grade analysis for all estimators, and there we also see that agreement is highest for EB, slightly lower for DOLS, and lowest for AR.

When we add a 2-year lagged math score to the specification (column [2]), the rank correlations all improve substantially, around .97 for 5th grade and slightly higher around .98 for 6th grade

[20] The student-level covariates include controls for days absent, race/ethnicity, disability, LEP, FRL-eligibility, and female.

(with the exception of AR, which is lower at .96 for 6th grade). In column [3] we omit student characteristics and use 2 lag scores each in reading and math, and now find even greater agreement with the URM estimates with rank correlations above .99. Finally, in column [4], we use the composite score as the only regressor, and now the rank correlations are even higher. (The rank correlations for EB are exactly 1 because this is the URM approach itself.)

Within each grade/specification combination, the EB rank correlations are at least as large as those for DOLS or AR, which indicates that the estimation approach matters somewhat. However, the specification seems to be more important in our data. Agreement with the URM increases for all estimators as we get closer to using the same specification as the URM (moving left to right from columns [1]-[4]); when we use the composite score as the only regressor, all of the rank correlations are very close to 1.

The results in column [3] also show that the differences between the URM and the regression based approaches using the same lag scores are not large. The complicated nature of the URM stems from taking extra steps to include students with certain patterns of partially missing test score records, since regression-based methods omit these students from estimation. Given that consistent estimation for DOLS and the URM requires very similar (if not identical) assumptions regarding the way in which data are missing, it is not surprising that the two approaches reach similar results. The estimates from simple DOLS estimation of a similar specification with teacher indicators correlates very highly (.99) with the complicated multi-step EVAAS URM estimation.

The high agreement between DOLS and the URM also suggests that there is not substantial nonrandom assignment based on prior achievement in our data. Our simulation results showed that DOLS is robust to this type of nonrandom assignment while the URM (along with EB and AR) is not.

For another illustration related to a policy context, Table C8 shows the average percent (and number) of teachers for which each of the other estimators would disagree with the URM on their classification of teachers in the top decile in the distribution of estimated teacher effects. So this could represent a scenario where the top 10 percent of teachers received a pay increase or bonus. The disagreement rates range from 0.3%–2.6%, with the smallest for EB estimation of the specification

that "mimics" the URM, which is expected. In this case, during the 2002-2007 period only 4 or 5 sixth grade (9 or 10 for fifth grade) teacher effects were classified in the top 10 percent with the URM estimates, but classified as below the 90th percentile with the EB-mimic estimates. The analogous results in column [3] for DOLS show disagreement rates on the top decile are .7% (30 or 34 teacher effects) for fifth grade and 1.2%–1.6% (22 or 29 teacher effects) for sixth grade.

3.6 Summary and Conclusions

We have shown how, in a simplified setting, the multi-step EVAAS URM estimation approach relates very closely to simple OLS estimation using the same lagged test scores. While this exact relationship is more difficult to see when we extend to settings with missing data or multiple years, we show how similar the estimates are, and under what conditions they are expected to diverge, using both simulations and administrative data.

Our simulation evidence shows that the URM exhibits similar performance patters to those seen with empirical Bayes' estimation in Guarino et al. (2015). While the URM and EB perform similarly to DOLS under the ideal conditions of random assignment and random grouping, DOLS is most robust to nonrandom assignment, especially assignment based on the lagged score, which is certainly a plausible assignment mechanism. Our results based on administrative data suggest that there may not be substantial sorting in this district, given the similarity between the URM/EB and DOLS, regardless of specification.

Although our simulations showed that OLS generally does as well—or better—than the more complicated EVAAS URM in recovering true teacher effects, our analysis of administrative data suggests the extent of the differences may not be extremely problematic in practice. This is perhaps reassuring given that the EVAAS methods are already used in several states and districts for teacher evaluation purposes, in some cases for high-stakes decision making.

APPENDIX

Table C1: Correlations Between Estimated and True Teacher Effects (1 Cohort of Students)

1 cohort of students		"common"			"mimic"			
Small teacher effects		1-yr lag score			1-yr and 2-yr lag scores			
Grouping	Assignment	DOLS	AR	EB	URM	DOLS	AR	EB
PANEL A - Substantial decay *(lambda = 0.5)*								
Random	Random	0.69	0.69	0.69	0.69	0.69	0.69	0.69
Dynamic	Random	0.70	0.70	0.70	0.70	0.70	0.70	0.70
	Positive	0.67	0.49	0.53	0.53	0.68	0.50	0.53
	Negative	0.70	0.53	0.58	0.58	0.70	0.53	0.57
Baseline	Random	0.67	0.67	0.67	0.68	0.68	0.68	0.68
	Positive	0.75	0.72	0.73	0.71	0.73	0.69	0.71
	Negative	0.50	0.49	0.50	0.55	0.55	0.53	0.55
Heterogeneity	Random	0.64	0.64	0.64	0.65	0.64	0.65	0.65
	Positive	0.80	0.79	0.79	0.79	0.79	0.78	0.79
	Negative	0.41	0.41	0.41	0.43	0.43	0.43	0.43
PANEL B - Complete persistence *(lambda = 1)*								
Random	Random	0.69	0.69	0.69	0.69	0.69	0.69	0.69
Dynamic	Random	0.68	0.68	0.68	0.68	0.68	0.68	0.68
	Positive	0.65	0.43	0.46	0.46	0.65	0.43	0.46
	Negative	0.70	0.49	0.53	0.52	0.70	0.49	0.53
Baseline	Random	0.69	0.69	0.69	0.69	0.69	0.69	0.68
	Positive	0.69	0.64	0.66	0.66	0.69	0.63	0.65
	Negative	0.63	0.59	0.61	0.62	0.64	0.59	0.61
Heterogeneity	Random	0.65	0.65	0.65	0.65	0.65	0.65	0.65
	Positive	0.79	0.78	0.79	0.79	0.79	0.78	0.79
	Negative	0.46	0.45	0.45	0.45	0.45	0.45	0.45

Notes. This table provides the Spearman rank correlations with the true teacher effects. The URM and the "mimic" DOLS, AR, and EB are based on specifications with a 1-year and 2-year lagged score, while the "common" DOLS, AR, and EB estimates are based on the specification with just the 1-year lagged score. These results are based on simulations with small teacher effects, and 1 cohort of students.

Table C2: Correlations Between Estimated and True Teacher Effects (3 Cohorts of Students)

3 cohorts of students		"common"			"mimic"			
Small teacher effects		1-yr lag score			1-yr and 2-yr lag scores			
Grouping	Assignment	DOLS	AR	EB	URM	DOLS	AR	EB
PANEL A - Substantial decay *(lambda = 0.5)*								
Random	Random	0.84	0.84	0.84	0.84	0.84	0.84	0.84
Dynamic	Random	0.84	0.84	0.84	0.84	0.84	0.84	0.84
	Positive	0.84	0.66	0.76	0.76	0.84	0.66	0.76
	Negative	0.83	0.68	0.77	0.77	0.83	0.68	0.77
Baseline	Random	0.82	0.82	0.82	0.83	0.83	0.83	0.83
	Positive	0.88	0.87	0.88	0.87	0.87	0.85	0.87
	Negative	0.65	0.65	0.65	0.71	0.72	0.70	0.71
Heterogeneity	Random	0.81	0.81	0.81	0.82	0.82	0.82	0.82
	Positive	0.89	0.89	0.89	0.89	0.89	0.89	0.89
	Negative	0.52	0.52	0.52	0.55	0.56	0.55	0.55
PANEL B - Complete persistence *(lambda = 1)*								
Random	Random	0.84	0.84	0.84	0.84	0.84	0.84	0.84
Dynamic	Random	0.84	0.84	0.84	0.83	0.84	0.84	0.84
	Positive	0.84	0.59	0.70	0.70	0.84	0.59	0.70
	Negative	0.84	0.62	0.72	0.72	0.84	0.62	0.72
Baseline	Random	0.84	0.84	0.84	0.84	0.84	0.84	0.84
	Positive	0.85	0.81	0.84	0.84	0.85	0.80	0.84
	Negative	0.80	0.76	0.79	0.79	0.81	0.76	0.79
Heterogeneity	Random	0.82	0.82	0.82	0.82	0.82	0.82	0.82
	Positive	0.89	0.89	0.89	0.89	0.89	0.89	0.89
	Negative	0.58	0.57	0.58	0.58	0.58	0.57	0.58

Notes. This table provides the Spearman rank correlations with the true teacher effects. The URM and the "mimic" DOLS, AR, and EB are based on specifications with a 1-year and 2-year lagged score, while the "common" DOLS, AR, and EB estimates are based on the specification with just the 1-year lagged score. These results are based on simulations with small teacher effects, and 3 cohorts of students.

Table C3: Correlations - URM vs. Other Estimators (Small Teacher Effects)

"Common" estimating equation Small teacher effects		1 cohort of students			3 cohorts of students		
Grouping	Assignment	DOLS	AR	EB	DOLS	AR	EB
PANEL A - Substantial decay *(lambda = 0.5)*							
Random	Random	0.99	0.99	0.99	0.99	0.99	0.99
Dynamic	Random	0.99	0.99	0.99	0.99	0.99	0.99
	Positive	0.94	0.98	0.99	0.97	0.96	0.99
	Negative	0.96	0.99	0.99	0.97	0.97	0.99
Baseline	Random	0.99	0.99	0.99	0.99	0.99	0.99
	Positive	0.99	0.99	0.99	0.99	1.00	0.99
	Negative	0.99	0.99	0.99	0.98	0.98	0.98
Heterogeneity	Random	0.99	0.99	0.99	0.99	0.99	0.99
	Positive	0.99	0.99	1.00	1.00	1.00	1.00
	Negative	0.99	0.99	0.99	0.99	0.99	0.99
PANEL B - Complete persistence *(lambda = 1)*							
Random	Random	1.00	1.00	1.00	1.00	1.00	1.00
Dynamic	Random	0.98	1.00	1.00	0.99	0.99	0.99
	Positive	0.92	0.99	1.00	0.93	0.97	1.00
	Negative	0.93	0.99	1.00	0.94	0.97	1.00
Baseline	Random	1.00	1.00	1.00	1.00	1.00	1.00
	Positive	0.99	1.00	1.00	1.00	0.99	1.00
	Negative	1.00	1.00	1.00	1.00	0.99	1.00
Heterogeneity	Random	1.00	1.00	1.00	1.00	1.00	1.00
	Positive	1.00	1.00	1.00	1.00	1.00	1.00
	Negative	1.00	1.00	1.00	1.00	1.00	1.00

Notes. This table provides the Spearman rank correlations with the URM estimates, for the DOLS, AR, and EB estimates. The URM composite score uses a 1-year and 2-year lagged score, while the DOLS, AR, and EB estimates are based on the "common" specification with just the 1-year lagged score. These results are based on simulations with small teacher effects.

Table C4: Correlations - Estimated vs. True Teacher Effects (Large Teacher Effects)

OLS, AR, EB on "common" specification
Large teacher effects

		1 cohort of students				3 cohorts of students			
Grouping	Assignment	DOLS	AR	EB	URM	DOLS	AR	EB	URM
PANEL A - Substantial decay *(lambda = 0.5)*									
Random	Random	0.90	0.90	0.90	0.90	0.95	0.95	0.95	0.95
Dynamic	Random	0.90	0.89	0.90	0.90	0.95	0.95	0.95	0.95
	Positive	0.89	0.68	0.83	0.83	0.95	0.75	0.93	0.94
	Negative	0.90	0.76	0.87	0.87	0.95	0.83	0.94	0.94
Baseline	Random	0.89	0.89	0.89	0.90	0.95	0.95	0.95	0.95
	Positive	0.91	0.88	0.90	0.90	0.96	0.94	0.96	0.95
	Negative	0.82	0.82	0.82	0.84	0.89	0.89	0.89	0.91
Heterogeneity	Random	0.86	0.87	0.86	0.87	0.94	0.94	0.94	0.94
	Positive	0.93	0.92	0.93	0.93	0.96	0.95	0.96	0.96
	Negative	0.77	0.76	0.77	0.78	0.84	0.83	0.84	0.85
PANEL B - Complete persistence *(lambda = 1)*									
Random	Random	0.90	0.90	0.90	0.90	0.95	0.95	0.95	0.95
Dynamic	Random	0.89	0.89	0.90	0.90	0.95	0.95	0.95	0.95
	Positive	0.89	0.63	0.77	0.77	0.95	0.69	0.92	0.92
	Negative	0.90	0.70	0.84	0.84	0.95	0.77	0.94	0.94
Baseline	Random	0.90	0.90	0.90	0.90	0.95	0.95	0.95	0.95
	Positive	0.90	0.85	0.89	0.89	0.95	0.91	0.95	0.95
	Negative	0.87	0.85	0.87	0.87	0.93	0.92	0.93	0.94
Heterogeneity	Random	0.87	0.88	0.87	0.87	0.94	0.94	0.94	0.94
	Positive	0.93	0.92	0.93	0.93	0.96	0.95	0.96	0.96
	Negative	0.79	0.78	0.79	0.79	0.86	0.85	0.86	0.86

Notes. This table provides the Spearman rank correlations with the true teacher effects. The URM composite score uses a 1-year and 2-year lagged score, while the DOLS, AR, and EB estimates are based on the "common" specification with just the 1-year lagged score. These results are based on simulations with large teacher effects.

Table C5: Correlations - URM vs. Other Estimators (Large Teacher Effects)

"Common" estimating equation Large teacher effects		1 cohort of students			3 cohorts of students		
Grouping	Assignment	DOLS	AR	EB	DOLS	AR	EB
PANEL A - Substantial decay *(lambda = 0.5)*							
Random	Random	1.00	0.99	1.00	1.00	1.00	1.00
Dynamic	Random	0.99	0.99	0.99	0.99	0.99	0.99
	Positive	0.97	0.94	1.00	0.99	0.88	1.00
	Negative	0.98	0.94	0.99	0.99	0.91	0.99
Baseline	Random	0.99	0.99	0.99	1.00	0.99	1.00
	Positive	0.99	0.99	1.00	1.00	0.99	1.00
	Negative	0.99	0.99	0.99	0.99	0.99	0.99
Heterogeneity	Random	1.00	0.99	1.00	1.00	1.00	1.00
	Positive	1.00	0.99	1.00	1.00	0.99	1.00
	Negative	0.99	0.99	0.99	0.99	0.99	0.99
PANEL B - Complete persistence *(lambda = 1)*							
Random	Random	1.00	1.00	1.00	1.00	1.00	1.00
Dynamic	Random	0.99	0.99	1.00	1.00	0.99	1.00
	Positive	0.94	0.95	1.00	0.99	0.86	1.00
	Negative	0.97	0.94	1.00	0.99	0.88	1.00
Baseline	Random	1.00	1.00	1.00	1.00	1.00	1.00
	Positive	1.00	0.98	1.00	1.00	0.98	1.00
	Negative	1.00	0.99	1.00	1.00	0.99	1.00
Heterogeneity	Random	1.00	1.00	1.00	1.00	1.00	1.00
	Positive	1.00	0.99	1.00	1.00	0.99	1.00
	Negative	1.00	1.00	1.00	1.00	1.00	1.00

Notes. This table provides the Spearman rank correlations with the URM estimates, for the DOLS, AR, and EB estimates. The URM composite score uses a 1-year and 2-year lagged score, while the DOLS, AR, and EB estimates are based on the "common" specification with just the 1-year lagged score. These results are based on simulations with large teacher effects.

Table C6: Descriptive Statistics for Students in Sample, by Grade

	Obs	Mean	Std. Dev	Min	Max
Panel A: Grade 5					
Math score	110,147	1638.62	232.26	569	2456
Reading score	109,878	1572.35	314.13	474	2713
1-yr lag Math	110,147	1485.78	254.84	569	2330
2-yr lag Math	106,714	1344.95	287.95	375	2225
1-yr lag Reading	109,879	1523.12	317.68	295	2638
2-yr lag Reading	106,510	1297.28	350.45	86	2514
Disability	110,147	0.11	0.32	0	1
LEP	110,147	0.51	0.50	0	1
Female	110,147	0.50	0.50	0	1
FRL	110,147	0.70	0.46	0	1
Black	110,147	0.28	0.45	0	1
Hispanic	110,147	0.60	0.49	0	1
Panel B: Grade 6					
Math score	101,307	1652.63	242.93	770	2492
Reading score	101,122	1635.41	302.95	539	2758
1-yr lag Math	101,307	1634.20	220.28	569	2456
2-yr lag Math	97,895	1460.65	247.73	569	2330
1-yr lag Reading	101,008	1550.82	306.45	474	2713
2-yr lag Reading	97,572	1504.67	313.31	86	2638
Disability	101,307	0.06	0.24	0	1
LEP	101,307	0.52	0.50	0	1
Female	101,307	0.51	0.50	0	1
FRL	101,307	0.71	0.46	0	1
Black	101,307	0.28	0.45	0	1
Hispanic	101,307	0.61	0.49	0	1

Table C7: Spearman Rank Correlations, Comparing EVAAS URM to Other Estimators

	1-yr lag Math, Student Char. [1]	1-yr & 2-yr lag Math, Student Char. [2]	1-yr & 2-yr lags in Math & Reading [3]	Composite score [4]
Panel A: 5th grade				
1-year estimates				
DOLS	0.918	0.971	0.997	0.999
AR	0.922	0.972	0.995	0.998
EB	0.920	0.972	0.998	1.000
N	5016	5016	5016	5016
2-year estimates				
DOLS	0.918	0.971	0.997	0.999
AR	0.921	0.971	0.994	0.997
EB	0.920	0.972	0.998	1.000
N	4203	4203	4203	4203
Panel B: 6th grade				
1-year estimates				
DOLS	0.941	0.982	0.995	0.997
AR	0.931	0.964	0.987	0.990
EB	0.944	0.984	0.998	1.000
N	1814	1814	1814	1814
2-year estimates				
DOLS	0.945	0.982	0.995	0.997
AR	0.935	0.963	0.986	0.990
EB	0.948	0.984	0.998	1.000
N	1536	1536	1536	1536

Notes. This table provides the average Spearman rank correlation between the EVAAS URM estimate and the other estimator/specifications. N = the number of teacher effect observations underlying the average. For the 1-year estimates this corresponds to the number of Teacher-Year observations from 2002-2007. For the 2-year estimates this corresponds to the number of teacher effects computed (for each estimator/specification) during 2003-2007.

Table C8: Disagreement with the URM in Classification of Teachers Above the 10th Percentile

	1-yr lag Math, Student Char. [1]	1-yr & 2-yr lag Math, Student Char [2]	1-yr & 2-yr lags in Math & Reading [3]	Composite score [4]
Panel A: Grade 5				
1-year estimates (N=5,016 teacher effect estimates)				
DOLS	2.6%	1.5%	0.7%	0.7%
n	129	73	34	35
AR	2.6%	1.4%	0.9%	0.8%
n	130	71	47	41
EB	2.6%	1.4%	0.2%	0.0%
n	128	72	10	0
2-year estimates (N=4,203 teacher effect estimates)				
DOLS	2.5%	1.4%	0.7%	0.7%
n	106	58	30	31
AR	2.5%	1.4%	1.0%	0.8%
n	108	58	42	35
EB	2.5%	1.4%	0.2%	0.0%
n	106	57	9	0
Panel B: Grade 6				
1-year estimates (N=1,814 teacher effect estimates)				
DOLS	2.1%	1.1%	1.2%	1.1%
n	38	21	22	20
AR	2.0%	2.0%	1.6%	1.5%
n	36	36	29	27
EB	2.0%	0.8%	0.3%	0.0%
n	36	15	5	0
2-year estimates (N=1,536 teacher effect estimates)				
DOLS	2.0%	1.2%	1.2%	1.0%
n	31	19	18	16
AR	1.9%	2.0%	1.6%	1.5%
n	29	31	25	23
EB	1.8%	0.8%	0.3%	0.0%
n	28	12	4	0

Notes. This table provides the average percent of teachers whose classification changes from the top 10 percent in the distribution of EVAAS URM estimated teacher effects to below the top 10 percent in the distribution of teacher effects based on the other estimator/specification combinations. The average is taken as the simple average of the percent misclassified in each year 2002-2007 for the 1-year estimates and 2003-2007 for the 2-year estimates. n = the number of teachers for whom classification changes in this way.

REFERENCES

REFERENCES

Boyd, D., Lankford, H., Loeb, S., & James, W. (2012). Measuring test measurement error: a general approach, in Working paper prepared for the National Conference on Value-Added Modeling (University of Wisconsin-Madison).

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System (EVAAS). *Educational Researcher*, 37(2), 65-75.

Dieterle, S., Guarino, C., Reckase, M. & Wooldridge, J. (2015). How do principals assign students to teachers? Finding evidence in administrative data and the implications for value-added. *Journal of Policy Analysis and Management*, 34(1), 32-58.

Guarino, C., Maxfield, M., Reckase, M., Thompson, P., & Wooldridge, J. (2015). An evaluation of empirical Bayes' estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics*, (40), 190-222.

Guarino, C., Reckase, M. & Wooldridge, J. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, 10(1), 117-156.

Hanushek, E. (1979). Conceptual and empirical issues in the estimation of educational production functions. *The Journal of Human Resources*, 14(3), 351-388.

Lockwood, J. R. & McCaffrey, D. F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics*, (1), 223-252.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, 32(2), 125-150.

McCaffrey, D. F., & Lockwood, J. R. (2011). Missing data in value-added modeling of teacher effects. *The Annals of Applied Statistics*, 5(2A), 773-797.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics*, 29(1), 67-101.

Rose, R., Henry, G., & Lauen, D. (2012). Comparing value-added models for estimating individual teacher effects on a statewide basis: Simulations and empirical analyses. Consortium for Educational Research and EvaluationNorth Carolina. http://cerenc.org

Sanders, W. (2006). Comparisons among various educational assessment value-added models. Presented at The Power of two: National Conference on Value-Added, Columbus, OH, October 16. SAS® White Paper. Cary, NC: SAS Institute.

Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009). A response to criticisms of SAS® EVAAS®. SAS® White Paper. Cary, NC: SAS Institute.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485), F3-F33.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data, 2nd ed. Cambridge, MA: MIT Press.

Wright, S. P., Sanders, W. L., & Rivers, J. C. (2006). Measurement of academic growth of individual students toward variable and meaningful academic standards. SAS® White Paper. Cary, NC: SAS Institute.

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). SAS® EVAAS® statistical models. SAS® EVAAS® Technical Report. Cary, NC: SAS Institute.

Wright, S. P. (2010). An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education. SAS® EVAAS® Technical Report. Cary, NC: SAS Institute.