University of California
Santa Barbara

Essays in Applied Econometrics

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Economics

by

Antoine Deeb

Committee in charge:

Professor Peter Kuhn, Co-Chair
Professor Douglas Steigerwald, Co-Chair
Professor Clément de Chaisemartin
Professor Dick Startz
Professor Gonzalo Vazquez-Bare

June 2022

The Dissertation of Antoine Deeb is approved.

Professor Clément de Chaisemartin

Professor Dick Startz

Professor Gonzalo Vazquez-Bare

Professor Peter Kuhn, Committee Chair

Professor Douglas Steigerwald, Committee Chair

April 2022

Essays in Applied Econometrics

by

Antoine Deeb

Acknowledgements

This dissertation would not have been possible without the incredible support and encouragement from a large number of people. I would first like to thank my dissertation committee, Peter Kuhn, Douglas Steigerwald, Clément de Chaisemartin, Dick Startz, and Gonzalo Vazquez-Bare. I would not be the researcher I am today without the countless hours of advice, support, and feedback they have all provided throughout my time at UCSB. I am sincerely grateful to have had the opportunity to learn from you all.

Second, I would like to thank Serena Canaan and Pierre Mouganie. Working with you has taught me so much, I am grateful to have you as friends and coauthors.

Third, I would also like to thank all the faculty members and staff at UCSB, especially Mark Patterson for patiently helping me navigate many administrative hurdles over the past few years.

Fourth, I would like to thank my friends in Santa Barbara including Richard Uhrig, Ryan Sherrard, Molly Schwarz, Danae Hernandez Cortes, Matthew Fitzgerald, Hazem Alshaikhmubarak, Ryan Anderson, Ryan Ley, Nick Combs, Jordan Smith, and Emma Kurtenbach along with many others. Thank you for making these past few years so fun and memorable.

Fifth, I would like to thank my friends and family back home for all their love and support. I am eternally grateful for my parents Liliane and Nabil Deeb, and my brother Elie Deeb, whose unconditional love and support made me the person I am today and allowed me to pursue this endeavor and see it through.

Finally, I would like to thank my amazing partner Allegra Cockburn for her unconditional love and support these past few years. Thank you for endlessly going over my work with me, for standing by me and helping me through the difficult moments, and for celebrating with me during the happy ones. I could not have done this without you.

Curriculum Vitæ
Antoine Deeb

Education

2022	Ph.D. in Economics (Expected), University of California, Santa Barbara
2017	M.A. in Economics, University of California, Santa Barbara
2016	M.A. in Economics, American University of Beirut
2014	B.A. in Economics, American University of Beirut

Publications

Advisor Value-Added and Student Outcomes: Evidence from Randomly Assigned College Advisors, with Serena Canaan and Pierre Mouganie. (*Forthcoming,* **American Economic Journal: Economic Policy**)

Working Papers

A Framework for Using Value-Added in Regressions

Clustering and External Validity in Randomized Controlled Trials, with Clément de Chaisemartin

The Impact of Religious Diversity on Students' Academic and Behavioral Outcomes, with Serena Canaan and Pierre Mouganie

Computer Programs

`twowayfeweights`, with Clément de Chaisemartin and Xavier D'Haultfoeuille. Stata command computing some of the estimators proposed in de Chaisemartin, C. and D'Haultfoeuille, X. (2020) "Two-way fixed effects estimators with heterogeneous treatment effects". 14,581 downloads from the SSC repository as of September 2021

Honors and Awards

2021	University of California, Berkeley , David P. Gardner Fellow
2021	University of California, Santa Barbara, Graduate Division Dissertation Fellowship
2020	University of California, Santa Barbara, Dept. Research Quarter Fellowship Spring

2020	University of California, Santa Barbara, Outstanding Undergraduate Course TA
2019	University of California, Santa Barbara, Best Graduate Course TA
2019	University of California, Santa Barbara, Gretler Fellowship
2018	University of California, Santa Barbara, Deacon Fellowship
2017	University of California, Santa Barbara, Microeconomics Prelim Distinction
2017	University of California, Santa Barbara, Williamson Fellowship
2016	American University of Beirut, President Elias Hraoui Endowed Student Award in Economics for a Graduate Student
2014	American University of Beirut, Nadim Khalaf Memorial Award for highest GPA in Economics

UCSB Teaching Experience

Instructor, Stata Programming Class, Summer 2020

Head TA, Economics 10A: Intermediate Microeconomic Theory, Fall 2019, Winter 2019, Spring 2022

TA, Economics 241A: Econometrics I (Ph.D Level), Fall 2018

TA, Economics 241B: Econometrics II (Ph.D Level), Winter 2019, Winter 2021

TA, Economics 241C: Econometrics III (Ph.D Level), Spring 2019, Spring 2021

TA, Economics 10A: Intermediate Microeconomic Theory, Fall 2017, Winter 2018, Spring 2018

Conferences and Seminars

2021	Applied Seminar, UCSB and Applied Micro Lunch, UCSB
2020	5th IZA Workshop: The Economics of Education and The David P. Gardner Seminar

Programming Skills

Stata, Python, R, LaTeX, Matlab, EViews

Languages

English (Fluent), French (Native), Arabic (Native)

Permissions and Attributions

1. The content of Chapter 2 and Appendix B is the result of a collaboration with Serena Canaan and Pierre Mouganie, and is forthcoming in the American Economic Journal: Economic Policy. The journal grants authors the right to republish, post on servers, redistribute to lists and use any component of their work in other works. A pre-print version is available at `https://www.aeaweb.org/articles?id=10.1257/pol.20200778`

2. The content of Chapter 3 and Appendix C is the result of a collaboration with Clément de Chaisemartin.

Abstract

Essays in Applied Econometrics

by

Antoine Deeb

This dissertation consists of three essays that use and develop econometric methods to causally investigate topics in education and development economics. In the first chapter, I develop an econometric framework to correctly and efficiently draw inference in models where estimated value-added (VA) is an explanatory variable (and models where it appears as the dependent variable). Estimated VA measures have become increasingly popular metrics of worker and institutional quality, and they are now widely used in regressions by researchers seeking to establish links between worker quality and a broad range of outcomes. Although consistent standard errors are crucial to obtain correct confidence intervals and assess the validity of conclusions drawn by studies using VA measures in regressions, the literature has not yet tackled this issue. I contribute to this literature by setting up an econometric framework that allows me to show why naïve standard error estimators are inconsistent in such models, derive consistent standard error estimators, propose a more efficient estimator for models using VA measures as explanatory variables, and propose a testable condition under which naïve standard errors are consistent for models that use VA measures as dependent variables. Then, in an application using data from North Carolina public schools, I find that the increase in standard errors resulting from the required correction that I propose is larger than the impact of clustering standard errors. In the second chapter, based on joint work with Serena Canaan and Pierre Mouganie, we use VA measures to provide the first causal evidence on the impact of college advisor quality on student outcomes. To do so, we exploit a unique setting where students are randomly assigned to faculty advisors during their first year of college. We find that having a higher grade VA advisor reduces time to complete freshman year and increases four-year graduation rates by 2.5 percentage points. It also raises high-ability students' likelihood of enrolling and graduating with a STEM degree by 4 percentage points. The magnitudes of our estimated effects are comparable to those from successful financial aid programs and proactive coaching interventions. We also show that non-grade measures of advisor VA predict student success. In particular, advisors who are effective at improving students' persistence and major choice also boost other college outcomes. Our results indicate that allocating resources towards improving the quality of academic advising may play a key role in promoting college success. In the third chapter, based on joint work with Clément de Chaisemartin, we consider the case of a randomized controlled trial with individual-level treatment assignment, and we allow for individual-level and cluster-level (e.g. village-level) shocks to affect the units' potential outcomes. We show that one can draw inference on two estimands: the ATE conditional on the realizations of the cluster-level shocks, using heteroskedasticity-robust standard errors; the ATE netted out of those shocks, using cluster-robust standard errors. We show that by clustering, researchers can test if the treatment would still have had an effect, had the stochastic shocks that occurred during the experiment been different. Then, the decision to cluster or not depends on the level of external validity one would like to achieve.

Contents

Chapter 1

A Framework for Using Value-Added in Regressions

1.1 Introduction

Assessing the quality of workers and institutions and the impact of this quality on various outcomes is a common topic in economics. For instance, it is common to assess the impact of teachers' value added (VA) on long-run student outcomes, often by regressing those outcomes on the estimated VA measures (Chetty et al., 2014b; Jackson, 2018; Canaan et al., 2021).Furthermore, some studies ask whether observable characteristics of teachers can predict VA and regress VA measures on those characteristics. In this paper, I start by showing that when VA is the outcome or the explanatory variable in a regression, the regression's robust standard errors researchers routinely use to draw inference are invalid. I then construct consistent standard errors for the estimators used in such studies. Finally, for models using VA as an explanatory variable I propose a more efficient estimator, construct optimal instruments under certain assumptions, and discuss a specification test.

For ease of exposition, I present my results in the context of teacher test-score value added, estimated using leave-year out measures. However, the results can be extended to different VA measures, and to different methods of estimating VA. While I consider the case of linear regressions, the framework discussed in this paper can also be extended to non-linear models.

The main insight underlying the results in this paper is that the assumptions underpinning VA models naturally lead to a generalized method of moments (GMM) framework. After demonstrating this, I show how one can use that framework for estimation and inference in models that employ VA measures in regressions. Specifically, in models using VA measures as explanatory variables, I show that the estimation of the VA measures, and correlations between the observable characteristics used to construct them and true teacher quality, lead to incorrect inference due to inconsistent standard error estimators. I then propose corrected standard errors from a GMM framework, and discuss other possible solutions under stronger assumptions.

Next I show that models using VA measures as explanatory variables are often overidentified systems of moment conditions resembling instrumental variables systems, and use that fact to propose a more efficient estimator of the impact of having a high test-score VA teacher on long-run outcomes. Finally, I also provide corrected standard errors for models using VA measures as a dependent variable.

My main theoretical findings are as follows. First, I describe the current practices used to estimate the impact of teacher test-score VA on student earnings. This is typically done using a multi-step procedure where the effect of observable characteristics is removed from test-scores and earnings, the best linear predictor of current year VA is constructed using the measures from other years, and residualized earnings are regressed on the best linear predictor. I set up the treatment effect model underlying those practices, and show how the assumptions for the residuals of that model imply the higher level assumptions typically made in the literature, and lead to identification through a system of moment equations. Specifically, I show that these models rest on three important assumptions for the residuals of the treatment effect model. Teachers' VA has to be mean independent the unobserved determinants of students' test scores and earnings, and the average unobserved determinants of test-scores and earnings of students matched to a given teacher have to be uncorrelated across years. These assumptions imply the forecast unbiasedness assumption used to justify the use of VA measures as explanatory variables.

Second, I show that the aforementioned assumptions lead to the identification of the parameters of interest using a system that contains four sets of moment equations that map to the steps researchers routinely use to estimate VA's effect on long-run outcomes: a first set of moments to remove the effect of covariates from the test-scores, a second set of moments used to construct the best linear predictor of current year VA using the measures from other years, an optional third set of moments to remove the effect of covariates from the outcome, and a fourth set to get the impact of VA on the residualized outcome.[1]

Third, I use the fact that the asymptotic results for GMM estimators in this context naturally capture the aforementioned multi-step procedure to derive the asymptotic distribution of the estimator of the impact of true teacher VA on earnings used in the literature. I show that while an OLS regression will yield a consistent estimate of the coefficient, the associated standard error estimators will be inconsistent. Indeed, these standard error estimators take into account neither the correlations between the observable characteristics of students and true teacher VA, nor the construction of the best

[1] Whether the outcome needs to be residualized or not depends on the context of each specific application. In some VA studies, the outcome is residualized, but in other studies it is not.

linear predictor. Given that most VA papers rely on a base assumption of selection on observables, the aforementioned correlations are likely to be strong, and thus accounting for them and the estimation of the best linear predictor is important.

Fourth, I show that if a researcher is interested in constructing confidence intervals using the OLS coefficients and doesn't wish to estimate a system by GMM, then corrected standard errors obtained from the GMM formula need to be calculated using the estimated parameters.[2] I also show that in settings with random assignment of students to teachers, correct standard errors can be obtained by a 2SLS regression of earnings on current year VA while instrumenting the current year's VA by other year's VAs.

Fifth, I demonstrate that simply regressing the outcome on the best linear predictor of VA fails to take advantage of overidentification. I show that the previous system of moments is nested in a more general and overidentified system. Indeed, under the same assumptions required to use the best linear predictor of current year VA as an explanatory variable, we have that the VA measures in years $s \neq t$ are valid instruments for year t's VA. Then when we have more than two years of data, we have more than one valid instrument for the endogenous year t measure. I use this to show that one can then obtain a more precise estimator of the effect of VA on earnings using an optimal weighting matrix. Furthermore, in a constant effect framework, one can test the validity of the model using an overidentification test.

Sixth, I consider the construction of optimal instruments under the stronger assumption of random assignment of students to teachers, when researchers also choose to include covariates in their analysis. I find that if one is willing to assume that the conditional mean of the current year's VA given other years' VA is linear and errors are homoskedastic, a 3SLS estimator using the best linear predictor can be optimal.

[2] Another possibility is bootstrapping the estimation of the system. In practice this can be done without using GMM by bootstrapping the entire analysis starting with the estimation of value added.

Finally, I focus on cases where VA is used as a dependent variable. I derive corrected standard errors for coefficients from such regressions using a GMM framework and provide a testable condition under which using OLS with unadjusted standard errors can also lead to valid inference.

The preceding theoretical results are confirmed by simulations and an application. The simulations focus on the importance of adjusting standard errors when estimating an OLS regression using VA measures as explanatory variables. Indeed, even in a simple model with constant true value added over time, 95% confidence intervals constructed using OLS coefficients and standard errors only adjusted for clustering have a coverage rate of 72.4%. On the other hand, using the multi-step OLS estimator with the proposed corrected standard errors from GMM yields correct inference with a coverage rate of 94.2%. Furthermore, the proposed optimal GMM estimator has a variance that is 1.3% lower than the multi-step OLS estimator.

My application uses data on third graders in North Carolina public schools. I document the presence of correlations between the variables used to predict VA and true teacher VA. I find an effect of sorting on lagged test scores at the student level that is similar in magnitude to Chetty et al. (2014a), but I also find that these correlations are much stronger for the classroom and school-year level lagged test scores used as controls to estimate VA. Indeed, sorting on lagged test scores at the classroom and school-year level is respectively five and three times larger than at the individual level.[3] I then illustrate my theoretical findings by showing the impact that this sorting has on the standard errors of coefficients in regressions using value added. The GMM standard errors I propose are between 37 and 70% larger than the standard errors currently used in the literature. Notably, in this application, the increase in standard errors resulting from my adjustment

[3] Rothstein (2017) also finds significant sorting at the school level.

is larger on average than the impact of clustering standard errors.

This paper contributes to various literatures. First, for applied researchers, this paper provides a simple-to-implement framework to correctly and efficiently draw inference when using VA measures in regressions. Papers using value added as an explanatory variable are quite common in the economics of education literature, with notable examples being Chetty et al. (2014b) and Jackson (2018) which use VA to highlight the importance of school teachers in improving students' adult outcomes. Rose et al. (2019) use VA to study the link between teacher quality and future student criminal behavior. Canaan et al. (2021) use VA to show that advisors who raise short-run student achievement, such as GPA, improve subsequent long-run outcomes such as graduation. Mulhern (2019) uses VA to show that good high-school counselors tend to improve all measures of educational attainment for students. Liu and Loeb (2021) use VA to explore the link between a teacher's impact on student's attendance and their long-run outcomes. Opper (2019) uses VA as an explanatory variable when testing for spillover effects to determine whether the impact of teachers extends beyond the students in their classrooms.

This paper contributes to the applied literature by providing a theoretical framework that underpins the methods used in the literature while providing guidelines for the need to correct standard errors when using VA in regressions. Indeed, this paper identifies the correlations between the student characteristics and true teacher quality as one of the main causes of incorrect inference, and the construction of the best linear predictor as another. While most applied researchers are aware of the importance of properly accounting for these factors in order to obtain unbiased estimates of teacher VA, this paper documents the presence of these correlations in practice and stresses the importance of accounting for them when conducting inference as well.

Second, this paper is also related to the methodological literature on VA. This literature has mostly focused on different ways to estimate and shrink VA measures. For example Angrist et al. (2017) show how researchers can leverage lotteries to create better VA estimates by combining non-experimental and quasi-experimental methods to obtain VA estimates with lower MSE, while Gilraine et al. (2020) propose a non-parametric method to shrink VA estimates. Finally, Opper (2019) develops a method of moments estimator to construct VA measures that account for spillovers. This paper fills a different gap in the literature by providing a flexible framework for estimation and correct inference when using VA measures in regressions, as well as providing a guide for efficient estimation when using value-added measures as explanatory variables.

Third, the study of the properties of estimators in this paper is also related to the work of Pagan (1986) who provides a unified framework for the properties of two-step estimators with a focus on the possibility of invalid inference, and the work of Newey (1984) and Newey and McFadden (1994) which shows that the properties of multi-step estimators can be obtained by framing the estimators as method of moment estimators. This paper uses the results of the latter to derive its findings.

Finally, the conceptual link between VA models and GMM outlined in this paper, and the associated causes of incorrect inference and their solutions, can be extended to different settings using estimated measures of quality in regressions. Indeed, models resembling the VA framework described in this paper are commonly used in economics. Examples include studies of the effectiveness of healthcare providers (Currie and Zhang, 2021), the effectiveness of bosses (Lazear et al., 2015; Bertrand and Schoar, 2003), and the impact of individuals working as part of a team (Arcidiacono et al., 2017). While the standard error estimators derived in this paper might not directly apply in those settings because the procedures used to estimate VA measures can sometimes depart from the ones described here, researchers could draw correct inference in such cases by following the general approach outlined in this paper and jointly modeling the estimation of the VA measures and subsequent regression analysis together in a GMM framework.

The rest of this chapter is organized as follows. Section 2 presents the results for using VA as an explanatory variable. Section 3 presents the results for using VA as a dependent variable. Section 4 presents the simulation results. Section 5 goes over an empirical example and section 6 concludes.

1.2 Using Value Added as an Explanatory Variable

Researchers are often interested in the effect of teachers on the long-run outcomes of students. For example, how does the quality of a teacher affect the adult earnings of their students? To estimate this effect, one constructs a value-added measure of teacher quality and collects the outcomes of students that were matched with each teacher.

1.2.1 Setup

To begin, we must define the value added of a teacher. The common definition is the improvement in a student's test score attributable to the teacher. Let R_{it} be the test score for student i in year t. Then the potential test score, if the student were matched to a teacher who has value added μ is

$$R_{it}^{pot}(\mu) = \mu + R_{it}^{pot}(0)$$

where the test score is normalized to have a mean of 0 and a variance of 1. Because the test score is normalized, the value-added measure is also normalized to have a mean of

0. The part of the test score that is not attributable to the teacher's value added is

$$R_{it}^{pot}(0) = X_{it}' \beta_0 + \epsilon_{it},$$

where X_{it} captures the observed characteristics of the student and ϵ_{it} captures the unobserved characteristics of the student that determine test scores.

The value added of a teacher, while defined in terms of a short-run outcome, may also affect long-run outcomes, such as adult earnings. Let Y_i be the adult earnings for student i. Consider a model in which the value added of a teacher has a linear effect on earnings that is constant for all students (I relax the constant effect assumption in Appendix A.3). In this setting, the potential outcome function for the adult earnings of student i is given by

$$Y_i^{pot}(\mu) = \kappa_0 \mu + Y_i^{pot}(0)$$

and the potential outcome for an average quality teacher is

$$Y_i^{pot}(0) = X_{it}' \beta_0^Y + \eta_{it}$$

The student-level characteristics that determine this relation are measured at the time of exposure to the teacher: X_{it} captures the observed characteristics of the student and η_{it} captures the unobserved characteristics of the student and teacher.

To implement this framework, I allow for the possibility that the value added of a teacher varies over time, which for teacher j is denoted μ_{jt}. It follows from the potential outcome framework that the observed earnings for student i who was matched to teacher j in year t are equal to:

$$Y_i^{obs} = X_{it}'\beta_0^Y + \kappa_0\mu_{jt} + \eta_{it}. \tag{1.1}$$

The coefficient κ_0 measures the effect a teacher has on future earnings that arises from the teacher's contribution to the student's test scores. However, it is important to note that the effect captured in κ_0 might not be due solely to the direct effect on student test scores. There may be an indirect effect that can arise, for example, when a high value-added teacher in year t leads parents to select better teachers in years after t, which also positively affects earnings.[4]

I now describe a four step procedure in terms of population quantities used to identify κ_0 which I formalize in a system of moment conditions in the next section. Each step will be followed by a brief description of how it is implemented in practice. Before doing so, I introduce some helpful notation. There are J teachers and each teacher is observed for T years. A balanced panel across teachers and years simplifies the presentation, but is not required for the results. Each teacher has class size n_j, which is assumed to be constant over time.

In step one, we remove the effect of covariates from both test-scores and earnings. To do so, we begin with observed student test scores

$$R_{it}^{obs} = X_{it}'\beta_0 + \mu_{jt} + \epsilon_{it} \tag{1.2}$$

and remove the effect of student characteristics yielding

$$R_{it} = \mu_{jt} + \epsilon_{it} \tag{1.3}$$

As we will see later, (1.3) leads to a set of K moment conditions, where K is the dimension

[4] For further discussion about interpreting κ_0 when modeling earnings in a linear setting see Chetty et al. (2014b)

of β_0. In practice $\widehat{\beta}_0$ is estimated using within teacher variation by running a regression of observed test-scores R_{it}^{obs} on the covariates X_{it} and teacher fixed effects, one then obtains an estimate $\widehat{R}_{it} = R_{it}^{obs} - X_{it}'\widehat{\beta}_0$. The teacher fixed effect is included to correct for possible sorting on observable characteristics. If students with highly educated parents select the most gifted teacher, then estimates of the student characteristics would be biased by the omitted teacher effect. To remove this bias, we include teacher fixed-effects. Importantly, the teacher fixed effect is only used to estimate β_0, it is not used to construct $\widehat{R}_{it}$. Note that if we removed the teacher fixed effect from $\widehat{R}_{it}$, we would effectively remove the teacher's value added.

Next we remove the effect of the same set of covariates from earnings. We have:

$$Y_{it} = Y_i^{obs} - X_{it}'\beta_0^Y \tag{1.4}$$

such that Y_{it} is the earnings residual of student i, matched to teacher j in year t. This step will lead to another additional K moment conditions, where K is the dimension of β_0^Y. In practice $\widehat{\beta}_0^Y$ is estimated using within teacher variation by running a regression of observed earnings Y_i^{obs} on the covariates X_{it} and teacher fixed effects, one then obtains an estimate $\widehat{Y}_{it} = Y_i^{obs} - X_{it}'\widehat{\beta}_0^Y$. As before the teacher fixed effect is included to correct for possible sorting on observable characteristics when estimating β_0^Y, it is not used to construct $\widehat{Y}_{it}$.

In step two, we construct a preliminary measure of value added. Averaging R_{it} over all students in a class would yield the preliminary measure of value added

$$\overline{R}_{jt} = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{it} = \mu_{jt} + \overline{\epsilon}_{jt}, \tag{1.5}$$

with the idea that, on average, the $\overline{\epsilon}_{jt}$ are close to 0. This measure we construct contains both the value-added measure, μ_{jt}, and the effect of unobserved student characteristics,

ϵ_{it}.

Given that to estimate the effect of teacher value added on long-run outcomes, we want to use value added as an explanatory variable, there are two issues with using this initial constructed measure $\overline{R}_{jt}$. The first is that it is a noisy measure of value added, and the second is that students with positive unobserved determinants of test scores are likely to have positive unobserved determinants of earnings. In other words if we want to use $\overline{R}_{jt}$ as an explanatory variable then we have mechanical endogeneity from using the same students to form both the value-added measures and the outcome (Jacob et al., 2010). To address these concerns there are two possible solutions. The first, which is currently used in applied research and will be discussed in step three, involves constructing for each teacher the best linear predictor of preliminary value added for the current year $\overline{R}_{jt}$, from the preliminary value-added measures for all years $s \neq t$. The second solution which I discuss in section 1.2.3 will involve using the preliminary value-added measures for all years $s \neq t$ as instruments in an overidentified system of moment equations. I will show that there are efficiency gains from using the latter solution.

As previously mentioned, step three improves upon the preliminary measures by constructing the best linear predictor of value added for the current year t, from the value-added measures for all other years s. The best linear predictor removes the endogeneity because we assume that the unobserved factors that influence tests scores are uncorrelated over time as the classes have no students in common and all sorting of students to teacher is captured by the observable characteristics. In the context of our model it is saying that $\overline{\epsilon}_{jt}$ and $\overline{\eta}_{jt}$ are correlated but $\overline{\epsilon}_{js}$ is uncorrelated with $\overline{\epsilon}_{jt}$ and $\overline{\eta}_{jt}$ for $s \neq t$. Under the assumptions of our model the best linear predictor also shrinks the value-added measure towards a mean of zero and reduces the noise, Appendix A.4 illustrates this with a simple example.

It is reasonable to assume that value added is stationary. This then reduces the number of parameters to be estimated. This requires that mean teacher value added does not vary across calendar years and the correlation of value added across any pair of years depends only on the amount of time which elapses between those years. We can then define this improved measure as:

$$\mu_{jt}^* = \sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js}. \tag{1.6}$$

where

$$\boldsymbol{\phi_0} = \underset{\phi_{|s-t|}}{argmin}\mathbb{E}\left(\left(\overline{R}_{jt} - \sum_{|s-t|\neq 0} \phi_{|s-t|}\overline{R}_{js}\right)^2\right). \tag{1.7}$$

Under the assumption of stationarity, this step will lead to an additional $T-1$ moment conditions, where $T-1$ is the dimension of $\boldsymbol{\phi_0}$. In practice $\boldsymbol{\phi_0}$ can be estimated by regressing $\widehat{R}_{jt} = \frac{1}{n_j}\sum_{i=1}^{n_j}\widehat{R}_{it}$ on $\widehat{R}_{js}$, and the estimate of μ_{jt}^* is then a fitted value from an OLS regression of $\widehat{R}_{jt}$ on the residuals from all other years (Chetty et al., 2014a).

In step four we consider residualized earnings. They contain both the effect of teacher value added on student earnings, $\kappa_0\mu_{jt}$, and the effect of other unobserved characteristics of the student and teacher, η_{it}:

$$Y_{it} = \kappa_0\mu_{jt} + \eta_{it}. \tag{1.8}$$

Averaging over all students in a class yields:

$$\overline{Y}_{jt} = \kappa_0 \mu_{jt} + \overline{\eta}_{jt}. \tag{1.9}$$

which leads to one additional moment condition. In practice, the unobserved value μ_{jt} must be replaced with an estimate.

Specifically the common practice estimator of κ_0 is obtained by estimating the following sample counterpart of (1.9) using OLS:

$$\widehat{Y}_{jt} = \kappa \widehat{\mu}_{jt} + \zeta_{jt}, \tag{1.10}$$

where $\widehat{\mu}_{jt} = \sum_{|s-t|\neq 0} \widehat{\phi}_{|s-t|} \widehat{R}_{js}$ is an estimate of μ^*_{jt} and $\widehat{Y}_{jt} = \frac{1}{n_j} \sum_i Y_i^{obs} - X'_{it} \widehat{\beta}_0^Y$. I will refer to this estimator of κ_0 as the multi-step OLS estimator.

The next section formalizes the identification of κ_0 using a system of moment equations that follows the steps outlined above.

1.2.2 Identification

I now discuss Assumptions 1 and 2 which give interpretable primitive conditions on the residuals of the treatment effect model under which κ_0 is identified.

Assumption 1

1. *For all i, j, t:* $\mathbb{E}[\epsilon_{it} | \boldsymbol{\mu}_j] = \mathbb{E}[\epsilon_{it}] = 0$ *where* $\boldsymbol{\mu}_j = \begin{pmatrix} \mu_{j1} \\ \vdots \\ \mu_{jT} \end{pmatrix}$.

2. *For all* $s \neq t$: $\mathbb{E}(\eta_{it} | \mu_{js}) = 0$.

3. *For all* $s \neq t$, $\left(\bar{\epsilon}_{jt}, \overline{\eta}_{jt}\right) \perp\!\!\!\perp \left(\bar{\epsilon}_{js}, \overline{\eta}_{js}\right)$

Point 1 of Assumption 1 requires that students be sorted to teachers based only on observable characteristics so that teacher quality is unrelated to unobservable determinants of student short-run outcomes. Points 2 and 3 underpin the leave-one-out procedure described in the previous section. Specifically, Point 2 requires that the unobserved determinants of earnings in year t be mean independent of the true test-score value added of teacher j in years $s \neq t$. The mean independence assumption in point 2 can be weakened to be μ_{js} and $\overline{\eta}_{jt}$ being uncorrelated, but if teacher value-added is constant then this assumption must hold for $s = t$. Note that Point 2 does not rule out all sorting on long-run outcomes, it only requires that any sorting of students to teacher based on long-run outcomes be independent of a teacher's test-score value added. Finally Point 3 requires that the observable characteristics used to residualize short-run outcomes be sufficiently rich such that the average unobserved determinants of short-run and long-run achievement within teacher, be independent over time.

I impose the following additional assumption before establishing my identification result. To state the assumption compactly, for any variable H, let $\overline{H}_{jt} = \frac{1}{n_j}\sum_i H_{it}$ and $\ddot{H}_{jt} = \overline{H}_{jt} - \frac{1}{T}\sum_t \overline{H}_{jt}$. For each j: $\boldsymbol{H}_j$ is a matrix stacking $\overline{H}_{jt}$ over t and $\ddot{\boldsymbol{H}}_j$ is a matrix stacking $\ddot{H}_{jt}$ over t.

Assumption 2

1. $\mathbb{E}(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{X}}_j)$ *is finite and invertible.*

2. $0 < Var(\mu^*_{jt}) < \infty$.

3. $\mathbb{E}\left(\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\ddot{\boldsymbol{\mu}}_{\boldsymbol{j}}\right) = 0$, $\mathbb{E}(\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\ddot{\boldsymbol{\epsilon}}_{\boldsymbol{j}}) = 0$ *and* $\mathbb{E}(\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\ddot{\boldsymbol{\eta}}_j) = 0$.

Assumption 2 contains the assumptions required for identification. Point 1 requires no perfect multicolinearity in the covariates. Point 2 requires that the linear projection of residual test scores in the current year on other years have non-zero and finite variance. Point 3 requires that fluctuations in covariates be uncorrelated with fluctuations in unobserved shocks over time as well as fluctuations in value added over time. This third point is required to identify $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_0^Y}$ in a regression with teacher fixed effects.

I will now show κ_0 is identified by a set of four moment conditions.

Result 1 *If Assumptions 1 and 2 hold, then* $(\boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \boldsymbol{\phi_0})$ *and* κ_0 *are identified by the following system of moments:*

$$\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right) = 0 \tag{1.11}$$

$$\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0^Y}\right)\right) = 0 \tag{1.12}$$

$$\mathbb{E}\left(\boldsymbol{R_j^{(-t)\prime}}\left(\boldsymbol{R}_j - \boldsymbol{R_j^{(-t)}}\boldsymbol{\phi_0}\right)\right) = 0 \tag{1.13}$$

$$\mathbb{E}\left(\boldsymbol{\phi_0'}\boldsymbol{R_j^{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R_j^{(-t)}}\boldsymbol{\phi_0}\right)\right) = 0 \tag{1.14}$$

where $\boldsymbol{Y}_j$ *is a vector stacking* $\overline{Y}_{jt}$ *for teacher* j, $\overline{R}_j^{(-t)}$ *is a* $T-1$ *row vector stacking the* $\overline{R}_{js}$ *excluding* $\overline{R}_{jt}$, *and* $\boldsymbol{R_j^{(-t)}}$ *is a* $T \times (T-1)$ *matrix stacking the* $\overline{R}_j^{(-t)}$.

To map these moments into the four step procedure from the previous section, note that the first set of moment conditions will be used to estimate $\boldsymbol{\beta_0}$ and maps to the first step of the procedure. The second set of moment conditions will be used to estimate $\boldsymbol{\beta_0^Y}$, this also maps to the first step of the procedure. The third set of moment conditions will be used to obtain the best linear prediction of $\overline{R}_{jt}$ as a function of other years. This maps to the second and third step of the procedure with $\boldsymbol{R_j^{(-t)}}\boldsymbol{\phi_0} = \boldsymbol{\mu_j^*}$. The fourth set of moment conditions will be used to estimate κ_0 and maps to the last step.

For further intuition on this result, Appendix A.5 also presents an exposition of this identification result using variances and covariances.

Finally note that Assumptions 1 and 2 imply the assumptions commonly made when using value added as a regressor. To see that, consider $\mu^*_{jt} = \sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js}$ from (1.6). Note that μ^*_{jt} is the best linear predictor of $\overline{R}_{jt} = \mu_{jt} + \overline{\epsilon}_{jt}$ as a function of other years. We can write:

$$\overline{R}_{jt} = \sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js} + \theta_{jt}, \tag{1.15}$$

where θ_{jt} is the error from the best linear prediction. Then under Assumptions 1 and 2 we have the following result:

Result 2 *If Assumptions 1 and 2 hold, then:*

$$\begin{aligned} Cov\left(\overline{\eta}_{jt}, \mu^*_{jt}\right) &= 0 \\ \frac{Cov\left(\mu_{jt}, \mu^*_{jt}\right)}{Var\left(\mu^*_{jt}\right)} &= 1. \end{aligned}$$

Now notice that it follows from (1.9) that κ_0 is identified under Result 2:

$$\frac{Cov(\overline{Y}_{jt}, \mu^*_{jt})}{Var(\mu^*_{jt})} = \kappa_0 \frac{Cov(\mu_{jt}, \mu^*_{jt})}{Var(\mu^*_{jt})} + \frac{Cov(\overline{\eta}_{jt}, \mu^*_{jt})}{Var(\mu^*_{jt})} = \kappa_0,$$

since the conditions in Result 2 are the population equivalent of the assumptions usually stated in the current literature. For example, Chetty et al. (2014b) suggest that the reduced form coefficient from an OLS regression of earnings on estimated value added would identify κ_0, if an implicit and infeasible regression of true value added on estimated value added yields a coefficient of one. They show that one can recover the parameter of interest under the following conditions, which they call forecast unbiasedness for an estimate $\widehat{\mu}_{jt}$ of μ_{jt} and selection on observables:

$$\frac{Cov\left(\mu_{jt}, \widehat{\mu}_{jt}\right)}{Var\left(\widehat{\mu}_{jt}\right)} = 1 \text{ and } Cov\left(\eta_{ijt}, \widehat{\mu}_{jt}\right) = 0.$$

Therefore Assumptions 1 and 2 are primitive assumptions that ensure that the higher level assumptions currently made in the literature hold.

1.2.3 Estimation and Inference

This section will focus on drawing correct inference on κ_0. In many applications, neither n_j (class size) nor T (the number of years over which a teacher is observed) are very large, therefore I will consider asymptotics where J (the number of teachers) goes to infinity and n_j and T are fixed.

In practice, researchers use the multi-step OLS estimator $\widehat{\kappa}$ along with $\widehat{s}$, a consistent estimator of $\widetilde{\sigma}^2 = (G_{\kappa}^{-1})^2\mathbb{E}\left(g(\boldsymbol{Z})^2\right)$, in order to draw inference on κ_0.[5] In this section, I will show that $\widehat{s}$ will not be a consistent estimator of the true variance of $\widehat{\kappa}$, and propose an alternative and consistent estimator of the variance of $\widehat{\kappa}$. Then, I will show that one can construct an optimal GMM estimator that is more efficient than $\widehat{\kappa}$. Next, I will show that if students are randomly assigned to teachers, one can draw valid inference on κ_0 using a 2SLS procedure. Finally under the assumption of random assignment, I will show that given some distributional assumptions a 3SLS estimator of κ_0 is optimal if the researcher chooses to include covariates in their analysis.

[5] The actual form of $\widehat{s}$ used varies depending on whether one uses heteroskedasticity-robust variance estimators or cluster-robust variance estimators.

Asymptotic Distribution of the Multi-Step OLS Estimator

The multi-step OLS estimator of κ_0, described using the four steps in section 2.1, is an exactly identified method-of-moments estimator where the moments correspond to (1.11), (1.12), (1.13), and (1.14). Thus, asymptotic results for GMM estimators in this context naturally capture the multi-step OLS estimator. Therefore, one can treat $\widehat{\kappa}$ as part of a joint GMM estimator of the system of moments in Result 1 in order to derive its asymptotic distribution.

To establish consistency and asymptotic normality of the GMM estimators, regularity conditions are required - they are listed and discussed in Appendix A.6 as Assumptions A.6.1 and A.6.2.

The object of interest here is the asymptotic distribution of $\widehat{\kappa}$, the multi-step OLS estimator of the impact of value added on earnings. Theorem 1 below establishes the asymptotic distribution of this estimator. One should note that Theorem 1 is a subset of the more general Theorem A.7.1 which establishes the asymptotic normality and consistency of the joint estimators $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{Y}}, \widehat{\kappa})$. Given that $(\boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y})$ are often regarded as nuisance parameters required to estimate value-added measures and lack any meaningful causal interpretation, my discussion will focus only on the asymptotic distribution of $\widehat{\kappa}$.

Theorem 1 *If Assumptions 1, 2, A.6.1, and A.6.2 hold, then:*

1.

$$\sqrt{J}(\widehat{\kappa} - \kappa_0) \rightsquigarrow \mathcal{N}\left(0, \sigma^2\right) \tag{1.16}$$

where

$$\sigma^2 = (G_\kappa^{-1})^2 \mathbb{E}\left(\left(g(\boldsymbol{Z}) + G_{\beta^Y}\psi_3(\boldsymbol{Z}) + G_\phi\psi_2(\boldsymbol{Z}) + G_\beta\psi_1(\boldsymbol{z}) - G_\phi M_{2\phi}^{-1} M_{2\beta}\psi_1(\boldsymbol{Z})\right)^2\right)$$

and $\boldsymbol{Z_j} = \left(\boldsymbol{X_j}, \boldsymbol{R_j^{obs}}, \boldsymbol{Y_j^{obs}}\right)$*,* $\boldsymbol{Z}$ *stacks the* $\boldsymbol{Z_j}$*,* $g(\boldsymbol{Z}) = \boldsymbol{\phi_0'}\boldsymbol{R_j^{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R_j^{(-t)}}\boldsymbol{\phi_0}\right)$ *is the moment function used to estimate* κ_0*,* $G_\kappa = \mathbb{E}[\nabla_\kappa g(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)]$ *and* G_β*,* G_ϕ*,* G_{β^Y} *are defined analogously. The remaining terms are defined in the proof.*

2. *Let* $\widehat{\sigma}^2$ *correspond to an estimator of* σ^2*, constructed by replacing the population moments in* σ^2 *by averages and the parameters by the GMM estimators. Then:*

$$\widehat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2.$$

Theorem 1 provides the asymptotic distribution of $\widehat{\kappa}$, and notably its variance σ^2. Crucially this variance depends on G_β, G_ϕ, G_{β^Y}, the expected values of the gradient of the moment conditions used to identify κ_0 with respect to $\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\beta^Y}$ evaluated at $(\boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)$. Intuitively, we need to consistently estimate $(\boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y})$ in order to consistently estimate κ_0, therefore the variance of κ_0 must reflect the fact that the parameters used to construct the dependent variable and the value-added measure are estimated (Newey and McFadden, 1994).

The first implication of Theorem 1 is that GMM estimation of the system of moments from Result 1 will yield a consistent estimator of κ_0 and correct standard errors for that estimator. These standard errors can be used for hypothesis testing and confidence intervals. This is especially convenient as GMM estimation and the associated variance estimator are standard routines in most statistical software.

Another implication of Theorem 1 is that $\widehat{s}$, the variance estimator routinely used in practice, will be an inconsistent estimator of σ^2. Indeed, $\widehat{s}$ is a consistent estimator of $\widetilde{\sigma}^2$, and it follows from point one of the theorem that $\widetilde{\sigma}^2 \neq \sigma^2$. It is then worth examining what terms cause the difference between the two variances.

The first term I consider arises from the need to consistently estimate μ_{jt}^* and will depend on G_ϕ. As discussed in section 2.1, the construction of μ^*_{jt} when using the multi-step OLS estimator plays a crucial role in circumventing endogeneity issues arising from correlations between $\bar{\epsilon}_{jt}$ and $\overline{\eta}_{jt}$. Mathematically, we can show that G_ϕ is only zero in trivial cases, so $\widehat{s}$ can only be consistent in trivial cases. Indeed:

$$G_\phi = -\kappa_0 \boldsymbol{\phi_0'} \mathbb{E}\left(\boldsymbol{R_j^{(-t)\prime} R_j^{(-t)}}\right)$$

and under point 3 of Assumption A.6.2 we have $\mathbb{E}\left(\boldsymbol{R_j^{(-t)\prime} R_j^{(-t)}}\right) \neq 0$ and $\boldsymbol{\phi_0} \neq 0$. Therefore G_ϕ is only equal to zero in the trivial case where $\kappa_0 = 0$. Then we have the following corollary:

Corollary 1 *If Assumptions 1, 2, A.6.1, and A.6.2 hold, then $\widetilde{\sigma}^2$ cannot be equal to σ^2 unless $\kappa_0 = 0$.*

A direct implication of Corollary 1 is that under our assumptions, $\widehat{s}$ will be an inconsistent estimator of σ^2 in all non-trivial cases. Therefore, to correctly draw inference on κ_0 when estimating it using the multi-step OLS estimator, one needs to manually construct the variance estimator following the formula for σ^2 in Theorem 1. This can be done by replacing population moments with sample averages and parameters with their estimator. For example G_ϕ can be replaced with $\widehat{G}_\phi = -\widehat{\kappa}\boldsymbol{\widehat{\phi}'}\left(\frac{1}{J}\sum_{j=1}^{J} \boldsymbol{\widehat{R}_j^{(-t)\prime} \widehat{R}_j^{(-t)}}\right)$.

To further examine the difference, I consider the remaining two terms. The second term arises from the need to consistently estimate the relationship between covariates and test scores. It will depend on:

$$G_\beta = \mathbb{E}\left[-\left(\boldsymbol{Y_j'} - \boldsymbol{2\kappa_0 \phi_0' R_j^{(-t)\prime}}\right) A\right]$$

where A is a function of the covariates in years $s \neq t$.[6] Crucially, G_β depends on the covariance between the unobserved determinants of earnings $\overline{\eta}_{jt}$ and the covariates $\overline{X}_{js}$ for $s \neq t$, the covariance between covariates $\overline{X}_{jt}$ and true value added μ_{jt} in all years including t, and the covariance between $\overline{X}_{jt}$ and the unobserved determinants of test scores $\bar{\epsilon}_{jt}$ in all years including t. Given that most value-added models rest on assumptions of selection on observables, it is reasonable to expect that the covariance between covariates and value added is non-zero and thus $G_\beta \neq 0$.[7] The application to North Carolina data in section 1.5.2 and previous studies (Rothstein, 2017) show that these correlations are large and significant.

The final term that causes the difference between the two variances arises from the need to consistently estimate the relationship between earnings and covariates, and will depend on:

$$G_{\beta^Y} = \mathbb{E}\left[-\left(\boldsymbol{\phi_0'} \boldsymbol{R_j^{(-t)'}} \boldsymbol{X_j}\right)\right].$$

Similarly to G_β, G_{β^Y} will depend on the covariance between the covariates $\overline{X}_{jt}$ and test-score value added μ_{js} for $s \neq t$, and the covariance between $\overline{X}_{jt}$ and the unobserved determinants of test scores $\bar{\epsilon}_{js}$ for $s \neq t$. Again since value-added models rest on assumptions of selection on observables, one can expect that $G_{\beta^Y} \neq 0$.

Therefore, the multi-step nature of the procedure used to estimate κ_0, and the correlations between true value added and the observables used to estimate it, make the standard errors that researchers routinely use inconsistent estimators of σ^2. Appendix A.8 presents

[6] A is a $T \times K$ matrix such that each row consists of $(\phi_1 \overline{X}_{jt-1} + \phi_2 \overline{X}_{jt-2} + ...)$.

[7] Note that point 3 of Assumption 2 only rules out correlations between $\boldsymbol{\ddot{X}_j}$ and $\boldsymbol{\ddot{\eta}_j}$, and correlations between $\boldsymbol{\ddot{X}_j}$ and $\boldsymbol{\ddot{\epsilon}_j}$. So under the assumptions of the model it need not be that the remaining terms in G_β are zero either.

a comparison between σ^2 and $\widetilde{\sigma}^2$, which suggests that one could expect $\widetilde{\sigma}^2$ to be smaller. The next section will focus on the most efficient way to estimate κ_0.

Overidentification and Efficiency

If one is interested in estimating κ_0, then the multi-step OLS estimator will not be the estimator with the lowest variance. Indeed, under Assumptions 1 and 2, the exactly identified system consisting of (1.11), (1.12), (1.13), and (1.14) masks a more general overidentified system when $T > 2$. To see that note that (1.14) only requires that a linear combination of $\mathbb{E}\left(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R_j}\right)\right)$ be equal to zero.[8] However, under Assumptions 1 and 2 we have the stronger conditions:

$$\mathbb{E}\left(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R}_j\right)\right) = 0.$$

Given that $\mathbb{E}\left(\boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R_j}\right)\right) = 0$, any linear combination of these moments will also be zero which makes estimating $\boldsymbol{\phi_0}$ and thus (1.13) redundant. Then, by replacing (1.13) and (1.14) with $\mathbb{E}\left(\boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R_j}\right)\right) = 0$ we instead have the following system of moment conditions:

$$\begin{aligned}
&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right) = 0\\
&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0^Y}\right)\right) = 0\\
&\mathbb{E}\left(\boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R}_j\right)\right) = 0
\end{aligned} \tag{1.17}$$

with $2K + 1$ parameters to estimate and $2K + (T - 1)$ moments to estimate them with.

[8] See section A.9.14 for a proof.

Therefore when $T > 2$, the system is overidentified. Intuitively, the system consisting of (1.11), (1.12), and (1.17) is akin to an instrumental variables system where κ_0 is a scalar parameter that can be identified from $T - 1$ moments. In essence, under the same set of assumptions required to use the multi-step OLS estimator of κ_0, we have that the preliminary measures of value added in years $s \neq t$ are valid instruments for the preliminary measure of value added in year t from (1.3). As such when we have more than two years of data, we have more than one valid instrument for the endogenous current year measure. As a result, we can rewrite the system as an overidentified systems of moment conditions resembling an instrumental variables framework. The first two set of moments are taken from the previous setup and are used to create the dependent variable (residualized earnings), the instruments (preliminary measures of value added in years $s \neq t$), and the endogenous variable (preliminary measure of value added in year t). The third set of moments provide the overidentifying information.

Given that we have shown that the system of moments in Result 1 is nested in the more general overidentified system consisting of (1.11), (1.12), and (1.17), we can now establish that the GMM estimator with the optimal weighting matrix will be a more efficient estimator of κ_0. The optimal GMM estimator will then have a variance that is no greater than the multi-step OLS estimator.[9]

Let $\widehat{W}^*$ be an estimate of the optimal weighting matrix $W^* = \mathbb{E}[\widetilde{g}_1(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \kappa_0)\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \kappa_0)']^{-1}$ that replaces population moments with sample averages and parameters with their estimator. Then we have that the estimators

[9] The GMM system consisting of (1.11), (1.12), and (1.17) weighed using

$$W_1 = \begin{pmatrix} I_{K\times K} & 0 & 0 \\ 0 & I_{K\times K} & 0 \\ 0 & 0 & \boldsymbol{\phi_0}\boldsymbol{\phi_0'} \end{pmatrix}.$$

where $\boldsymbol{\phi_0} = \mathbb{E}\left(\boldsymbol{R_j^{(-t)\prime}}\boldsymbol{R_j^{(-t)}}\right)^{-1}\mathbb{E}\left(\overline{\boldsymbol{R}}\boldsymbol{{}_j^{(-t)\prime}}\boldsymbol{R_j}\right)$ would have the same objective function as the multi-step OLS.

resulting from GMM minimization with an optimal weighting matrix are consistent and asymptotically normal. Theorem 2 formalizes the result for the optimal estimator $\widehat{\kappa}^*$. It is a subset of the more general Theorem A.7.2 which formalizes the result for the joint estimators $(\widehat{\boldsymbol{\beta}}^{\boldsymbol{*}}\boldsymbol{,}\, \widehat{\boldsymbol{\beta}}^{\boldsymbol{Y*}}, \widehat{\kappa}^*)$.

Theorem 2 *If Assumptions 1, 2, and A.6.3 hold, then*

$$\sqrt{J}\left[\widehat{\kappa}^* - \kappa_0\right] \rightsquigarrow \mathcal{N}\left(0, \sigma_*^2\right).$$

where $\widehat{\kappa}^*$ is the estimate resulting from a GMM minimization using $\widehat{W}^*$ as a weighting matrix. One has that $\sigma_*^2 \leq \sigma^2$.

Theorem 2 shows that researchers can use an optimal GMM procedure to obtain more precise estimates of κ_0, such that $\sigma_*^2 \leq \sigma^2$ without imposing any additional assumptions. Furthermore, the validity of Assumptions 1, 2, and A.6.3, and the specifications of (1.2), (1.4), and (1.9) is falsifiable using Hansen's overidentification test under the assumption of constant treatment effects. This is formalized in Result A.7.1.

Estimation and Inference in the Presence of Random Assignment

I now turn to the case where students are randomly assigned to teachers. In such settings, one would not need to adjust for covariates in the analysis and drawing inference on κ_0 is a simpler problem. Indeed, while $\widehat{s}$ is still an inconsistent estimator of σ^2, one needs to account only for the estimation of the linear projection μ_{jt}^*. Then one can use a 2SLS regression of the outcome on the preliminary value-added measure in year t, $\overline{R}_{jt}$, while instrumenting it by the preliminary measures in years $s \neq t$.

To formalize this result and ease exposition, consider the following assumption similar to Canaan et al. (2021):

Assumption 3 *In every year students are randomly assigned to teachers such that one can model observed scores and earnings as:*

$$R_{it}^{obs} = \alpha_t + \mu_{jt} + \epsilon_{it}$$
$$Y_i^{obs} = \alpha_t^Y + \kappa_0 \mu_{jt} + \eta_{it}.$$

Under Assumption 3, given that students are randomly assigned to teachers, there is no need to control for covariates aside from year fixed effects. Indeed, (1.11) and (1.12) are now redundant since the time effects α_t and α_t^Y can be removed by subtracting the averages of R_{it}^{obs} and Y_i^{obs} for year t. Then we are left with (1.17) and the system of moments becomes:

$$\mathbb{E}\left(\boldsymbol{R_j^{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0 \boldsymbol{R_j}\right)\right) = 0,$$

which now aligns with a traditional instrumental variables problem. We can then estimate and draw inference on κ_0 using the following 2SLS regression:

$$\widehat{R}_{jt} = \sum_{|s-t|\neq 0} \phi_{|s-t|} \widehat{R}_{js} + \iota_{jt} \tag{1.18}$$
$$\widehat{Y}_{jt} = \kappa \sum_{|s-t|\neq 0} \widehat{\phi}_{|s-t|} \widehat{R}_{js} + \zeta_{jt} \tag{1.19}$$

where $\widehat{R}_{jt} = R_{jt}^{obs} - \frac{1}{J}\sum_{j=1}^{J} R_{jt}^{obs}$ where $R_{jt}^{obs} = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{it}^{obs}$, and $\widehat{Y}_{jt}$ is defined analogously. This corresponds to a regression of the outcome on current year value added while instrumenting year t value added by value added in years $s \neq t$. Furthermore, let $\widehat{s}_{2SLS}$ be the variance estimator that is computed by statistical softwares when using 2SLS.

This leads to the following result:

Result 3 *If Assumptions 1, 2, 3, A.6.1, and A.6.2 hold: Valid inference can be drawn on κ_0 by estimating the 2SLS regression defined by* (1.18) *and* (1.19) *while using $\widehat{s}_{2SLS}$ as a variance estimator.*

The intuition behind Result 3 is simple, if the covariates are uninformative of teacher value added or the unobserved determinants, then they are not required for a consistent estimation of the residualized outcome or the value-added measures. In that case, standard errors need only account for the fact that the current year preliminary value added measure was instrumented by the preliminary value added measures in other years. However, given that Assumption 3 is unlikely to hold in most applied settings, constructing a consistent estimator of σ^2 or GMM estimation of the system in Result 1 is required to construct confidence intervals using $\widehat{\kappa}$.

Finally, it follows from standard results that this 2SLS estimator will be efficient under homoskedasticity. The next section will show that if a researcher wants to include covariates in their analysis, then this 2SLS estimator is no longer efficient under homoskedasticity.

Optimal Instruments in the Presence of Random Assignment

Even in the presence of random assignment, researchers sometimes include covariates in the analysis, often to improve statistical precision. For such cases, I will now construct the optimal instruments for the system defined by (1.11), (1.12), and (1.17), and show that the traditional 3SLS estimator is optimal under a homoscedasticity assumption and if $\mathbb{E}[\boldsymbol{R}_j|\boldsymbol{R_j^{(-t)}}]$ is linear.[10]

For the remainder of this section I assume that students are randomly sorted to teachers and the following assumption holds:

Assumption 4

1. $\boldsymbol{X}_j \perp\!\!\!\perp (\boldsymbol{\mu_j}, \boldsymbol{\epsilon_j^{(-t)}}, \boldsymbol{\eta_j^{(-t)}})$.

2. $\boldsymbol{\mu_j} \perp\!\!\!\perp (\boldsymbol{\epsilon_j}, \boldsymbol{\eta_j})$.

3. $\mathbb{E}(\boldsymbol{\epsilon_j}|\boldsymbol{\ddot{X}_j})$=0.

Point 1 of Assumption 4 requires that the observable characteristics of students matched to teacher j in year t be independent of teacher j's value-added in year t and the unobservable characteristics of students assigned to teacher j in years $s \neq t$. Point 2 requires that the value-added of teacher j be independent of all student unobservables. These conditions should hold by design if students are randomly assigned to teachers. The third point slightly strengthens point 3 of Assumption 2 to require that the unobservable determinants of test-scores be mean independent of within teacher fluctuations in covariates.We can now construct the optimal instruments for the system:

$$\begin{aligned}
\mathbb{E}\left(\boldsymbol{\ddot{X}}_j'\left(\boldsymbol{\ddot{R}}_j^{obs} - \boldsymbol{\ddot{X}_j\beta_0}\right)\right) &= 0\\
\mathbb{E}\left(\boldsymbol{\ddot{X}}_j'\left(\boldsymbol{\ddot{Y}}_j^{obs} - \boldsymbol{\ddot{X}_j\beta_0^Y}\right)\right) &= 0\\
\mathbb{E}\left(\boldsymbol{R_j^{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R_j}\right)\right) &= 0.
\end{aligned}$$

[10] I use the term traditional 3SLS to refer to the 3SLS estimator that uses the linear projections of the endogenous variables as instruments. This is in contrast with the GMM 3SLS estimator which is optimal under homoskedasticity without further assumptions. The traditional 3SLS estimator is equivalent to GMM 3SLS when all equations use the same instruments (Wooldridge, 2010).

Let $\boldsymbol{u} = \begin{pmatrix} \ddot{\boldsymbol{\epsilon}}_{\boldsymbol{j}} + \ddot{\boldsymbol{\mu}}_{\boldsymbol{j}} \\ \ddot{\boldsymbol{\eta}}_{\boldsymbol{j}} + \kappa_0 \ddot{\boldsymbol{\mu}}_{\boldsymbol{j}} \\ \boldsymbol{\eta}_{\boldsymbol{j}} - \kappa_0 \boldsymbol{\epsilon}_{\boldsymbol{j}} \end{pmatrix}$, then we have the following result:

Result 4 *If Assumptions 1, 2, and 4 hold, then:*
If $\mathbb{E}[\boldsymbol{R}_j | \boldsymbol{R}_{\boldsymbol{j}}^{(-t)}]$ *is linear such that* $\mathbb{E}[\boldsymbol{R}_j | \boldsymbol{R}_{\boldsymbol{j}}^{(-t)}] = \boldsymbol{R}_{\boldsymbol{j}}^{(-t)} \boldsymbol{\phi}_{\boldsymbol{0}}$ *and* $\mathbb{E}[\boldsymbol{u}\boldsymbol{u'} | \ddot{\boldsymbol{X}}_{\boldsymbol{j}}, \boldsymbol{R}_{\boldsymbol{j}}^{(-t)}] = \mathbb{E}[\boldsymbol{u}\boldsymbol{u'}]$ *then the optimal moment conditions are:*

$$\mathbb{E}\left[\left(\mathbb{E}[\boldsymbol{u}\boldsymbol{u'}]^{-1} \begin{pmatrix} \ddot{\boldsymbol{X}}_{\boldsymbol{j}} & 0 & 0 \\ 0 & \ddot{\boldsymbol{X}}_{\boldsymbol{j}} & 0 \\ 0 & 0 & \boldsymbol{R}_{\boldsymbol{j}}^{(-t)} \boldsymbol{\phi}_{\boldsymbol{0}} \end{pmatrix}\right)' \boldsymbol{u}\right] = 0.$$

Those are the moment conditions satisfied by the 3SLS estimator. Then in this case the optimal estimator is the 3SLS estimator which first estimates $(\boldsymbol{\beta}_{\boldsymbol{0}}, \boldsymbol{\beta}_{\boldsymbol{0}}^{\boldsymbol{Y}})$ *using OLS and constructs* $\boldsymbol{R}_j$ *and* $\boldsymbol{Y}_j$*, then estimates* κ_0 *by a 2SLS regression of* $\boldsymbol{Y}_j$ *on* $\boldsymbol{R}_j$ *while instrumenting* $\boldsymbol{R}_j$ *with* $\boldsymbol{R}_{\boldsymbol{j}}^{(-t)}$*, uses those estimates to construct an estimate of* $\mathbb{E}[\boldsymbol{u}\boldsymbol{u'}]$*, and finally estimates the entire system again using GLS.*

Result 4 states that under random assignment, if the errors are homoskedastic and the conditional expectation of the preliminary value-added measures $\boldsymbol{R}_j$ in year t given preliminary value-added measures $\boldsymbol{R}_{\boldsymbol{j}}^{(-t)}$ in years $s \neq t$ is actually linear, then estimating the system composed of (1.11), (1.12), and (1.17) by 3SLS is efficient. Result 4 also rules out the system underlying the multi-step OLS estimator being optimal, as that would require the three components in $\boldsymbol{u}$ to be uncorrelated.

It is then useful to consider under what distributional assumptions the conditions from Result 4 hold. The first condition to consider is homoskedasticity, namely the possibility that $\mathbb{E}[\boldsymbol{u}\boldsymbol{u'} | \ddot{\boldsymbol{X}}_{\boldsymbol{j}}, \boldsymbol{R}_{\boldsymbol{j}}^{(-t)}] = \mathbb{E}[\boldsymbol{u}\boldsymbol{u'}]$. Homoskedasticity requires that the variances and covariance of the unobservable determinants and value added $\boldsymbol{\epsilon_j}, \boldsymbol{\eta_j}, \boldsymbol{\mu_j}$ do not vary across teachers with different levels of the preliminary measures of value added in years $s \neq t$, $\boldsymbol{R_j^{(-t)}} = \boldsymbol{\mu_j^{(-t)}} + \boldsymbol{\epsilon_j^{(-t)}}$, and with different levels of covariate fluctuations $\boldsymbol{\ddot{X}_j}$.

The second condition is $\mathbb{E}[\boldsymbol{R}_j|\boldsymbol{R_j^{(-t)}}]$ being linear. Suppose we have:

$$\begin{bmatrix} \boldsymbol{\mu_j} \\ \boldsymbol{\epsilon_j} \end{bmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} \boldsymbol{\Sigma_\mu} \otimes I_J & 0 \\ 0 & \sigma_\epsilon^2 I_{JT} \end{pmatrix}\right) \tag{1.20}$$

such that teacher value added and the average unobservable determinants of student test scores are joint normally distributed. Under (1.20) teacher value added is independent across teachers but correlated within teacher, and it is independent of the average unobservable determinants of student test scores which are i.i.d across teacher-years (this second requirement is likely to hold under random assignment).

Then we have:

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{R}_j|\boldsymbol{R_j^{(-t)}}] \\
=&\mathbb{E}[\boldsymbol{\mu_j} + \boldsymbol{\epsilon_j}|\boldsymbol{\mu_j^{(-t)}} + \boldsymbol{\epsilon_j^{(-t)}}] \\
=&\mathbb{E}[\boldsymbol{\mu_j}|\boldsymbol{\mu_j^{(-t)}} + \boldsymbol{\epsilon_j^{(-t)}}] + \mathbb{E}[\boldsymbol{\epsilon_j}|\boldsymbol{\mu_j^{(-t)}} + \boldsymbol{\epsilon_j^{(-t)}}] \\
=&\mathbb{E}[\boldsymbol{\mu_j}|\boldsymbol{\mu_j^{(-t)}} + \boldsymbol{\epsilon_j^{(-t)}}] \\
=&(\boldsymbol{\mu_j^{(-t)}} + \boldsymbol{\epsilon_j^{(-t)}})\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\mu_j^{(-t)}}+\boldsymbol{\epsilon_j^{(-t)}}}\boldsymbol{\Sigma_{\mu\mu^{(-t)}}} \\
=&\boldsymbol{R_j^{(-t)}}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{R_j^{(-t)}}}\boldsymbol{\Sigma}_{\boldsymbol{R}_j\boldsymbol{R_j^{(-t)}}} && (1.21) \\
=&\boldsymbol{R_j^{(-t)}}\boldsymbol{\phi_0} && (1.22)
\end{aligned}$$

where $\boldsymbol{\Sigma_{\mu\mu^{(-t)}}}$ is the covariance matrix of $\boldsymbol{\mu}$ and $\boldsymbol{\mu^{(-t)}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{R_j^{(-t)}}}$ is the covariance matrix of $\boldsymbol{R_j^{(-t)}}$. The third equality follows from the fact that $\boldsymbol{\epsilon_j}$ is independent of

$\boldsymbol{\mu_j^{(-t)}} + \boldsymbol{\epsilon_j^{(-t)}}$ by (1.20). The fourth equality follows from the conditional expectation formula for multivariate normal distributions, and the fifth equality follows from the fact that the covariance matrix of $\boldsymbol{R_j}$ and $\boldsymbol{R_j^{(-t)}}$ is the same as $\boldsymbol{\Sigma_{\mu\mu^{(-t)}}}$ by (1.20).

Then if teacher value added and the average unobservable determinants of student test scores are joint normally distributed following (1.20), then $\mathbb{E}[\boldsymbol{R}_j|\boldsymbol{R_j^{(-t)}}]$ is linear. To gain some intuition for (1.22), consider the case with $T = 2$ where

$$\mathbb{E}[\boldsymbol{R}_j|\boldsymbol{R_j^{(-t)}}] = \frac{Cov(\mu_{jt}, \mu_{j(t-1)})}{Var(\mu) + Var(\epsilon)}\overline{R}_{j(t-1)}.$$

The distributional assumption in (1.20) could be plausible in a setting with random assignment of students to teachers, a large number of teachers J, a large number of students per teachers n_j, and where value added is either constant over time for every teacher or $\mu_{jt} = \mu_j + \omega_{jt}$ where μ_j and $\boldsymbol{\omega_j}$ are joint normal and independent. Overall, if one is willing to assume that students are randomly assigned to teachers then the 3SLS estimator is efficient if one assumes homoskedasticity and joint normality of teacher value added and unobserved determinants.

1.3 Using Value Added as a Dependent Variable

Researchers are also often interested in examining whether the observable characteristics of teachers predict their value added. For example, do teachers with National Board Certification or more experience have higher value added? One can also look into whether the implementation of a given policy is linked to an increase in teacher value added. For instance, does a training program for teachers raise their value added? To answer such questions, researchers regress their estimated value-added measures on a set of explanatory variables.

In our setting, suppose that one is interested in the relationship between some teacher-year level variables D_{jt} and teacher value-added μ_{jt}. The parameters of interest in this section are the coefficients from the linear projection of μ_{jt} on D_{jt}:

$$\mu_{jt} = D_{jt}'\boldsymbol{\alpha_0} + \zeta_{jt}. \tag{1.23}$$

Given that μ_{jt} is unobserved, one cannot estimate (1.23) as it is. In practice, researchers often replace μ_{jt} by the estimated measure $\widehat{\mu}_{jt} = \sum_{|s-t|\neq 0} \widehat{\phi}_{|s-t|} \widehat{R}_{js}$ and proceed by estimating (1.23) using OLS.

This section will go over how to identify, consistently estimate, and draw inference on $\boldsymbol{\alpha_0}$ using an exactly identified GMM procedure. The asymptotic result for this GMM estimator will again naturally capture the cases in which the steps are separately estimated using OLS.

I impose the following assumption with additional regularity conditions in Assumption A.6.4 in Appendix A.6:

Assumption 5

1. $\mathbb{E}(\boldsymbol{D}_j'\boldsymbol{D}_j)$ *is finite and invertible.*

2. $\mathbb{E}(\boldsymbol{D}_j'\boldsymbol{\epsilon}_j) = 0.$

Point 1 requires no perfect multicolinearity in the $\boldsymbol{D}_j$. Point 2 requires that the average shocks be uncorrelated with the variables D_{jt}. Similarly to Assumption 1, this assumption requires that the observable characteristics used to residualize short-run outcomes be sufficiently rich such that the remaining unobservables, excluding value added, be uncorrelated with our variables of interest. Note that point 2 can be weakened to require that the average unobserved determinants of test scores in years $s \neq t$ be uncorrelated with D_{jt}, such that $\mathbb{E}(\boldsymbol{D}_j'\boldsymbol{\epsilon}_j^{\boldsymbol{(-t)}}) = 0$.

We can then show that $\boldsymbol{\alpha_0}$ is identified by the following system of moment conditions:

Result 5 *If Assumptions 1, 2, and 5 hold, then* $(\boldsymbol{\beta_0}, \boldsymbol{\alpha_0})$ *are uniquely identified by the following system of moments:*

$$\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right) = 0 \tag{1.24}$$

$$\mathbb{E}\left(\boldsymbol{D_j'}\left(\boldsymbol{R}_j - \boldsymbol{D_j}\boldsymbol{\alpha_0}\right)\right) = 0 \tag{1.25}$$

Result 5 shows that the coefficient from a linear projection of a set of variables D_{jt} on teacher value-added μ_{jt} is identified by two sets of moment conditions. The first set of moments is used to identify $\boldsymbol{\beta_0}$ and construct preliminary measures of value added. They are discussed in sections 2.1 and 2.2. The second set of moments is used to identify $\boldsymbol{\alpha_0}$. Note that unlike the previous sections, under point 2 of Assumption 5, we do not need to create a leave-year out measure of value added to recover the relationship between μ_{jt} and $\boldsymbol{D}_{jt}$. Instead, we can use the preliminary measures of value added $\overline{R}_{jt} = \mu_{jt} + \overline{\epsilon}_{jt}$ as the outcome since the unobserved determinants of student test scores are assumed to be uncorrelated with the variables of interest D_{jt}. If one suspects that point 2 of Assumption 5 is unlikely to hold, and is instead willing to assume that $\mathbb{E}(\boldsymbol{D}_j'\boldsymbol{\epsilon}_j^{\boldsymbol{(-t)}}) = 0$, then $\overline{R}_{jt}$ can be replaced by a simple leave-year out average $\widetilde{\mu}_{jt} = \frac{1}{T-1}\sum_{s\neq t}\overline{R}_{js}$ with the idea that $\boldsymbol{D}_{jt}$ is unlikely to be correlated with the unobservable determinants of test scores of students in different years.

To estimate $(\boldsymbol{\beta_0}, \boldsymbol{\alpha_0})$, let the GMM weighting matrix be the identity matrix $W = I$. Then Theorem A.7.3 shows that we can consistently estimate and draw inference on $\boldsymbol{\alpha_0}$ using

GMM. Using partitioned inversion and Theorem A.7.3, we can obtain the asymptotic variance of $\widehat{\boldsymbol{\alpha}}$.

Theorem 3 *If Assumptions 1, 2, and 5 hold, then*

$$\sqrt{J}(\widehat{\alpha}-\alpha_0) \rightsquigarrow \mathcal{N}\left(0, V_1\right) \tag{1.26}$$

where $V_1 = \mathbb{E}(\boldsymbol{D}_j'\boldsymbol{D}_j)^{-1}\mathbb{E}\left(\Gamma\Gamma'\right)\mathbb{E}(\boldsymbol{D}_j'\boldsymbol{D}_j)^{-1'}$, *and*
$\Gamma = \left(\boldsymbol{D_j'}\left(\boldsymbol{R}_j - \boldsymbol{D_j}\boldsymbol{\alpha_0}\right) - \mathbb{E}\left(\boldsymbol{D_j'}\boldsymbol{X_j}\right)\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{X}}_j\right)^{-1}\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right)$

Theorem 3 shows that a GMM estimation of the system in Result 5 will yield a consistent and asymptotically normal estimator of $\boldsymbol{\alpha_0}$, and the standard errors from that procedure can be used to correctly draw inference. Furthermore, because the GMM estimator captures the estimator of $\boldsymbol{\alpha_0}$ obtained from an OLS regression of an estimate of $\boldsymbol{R}_j$ on D_{jt}, one has to construct a consistent estimator of V_1 in order to construct confidence interval using the OLS estimator of $\boldsymbol{\alpha_0}$. Indeed since the uncorrected OLS variance estimator is a consistent estimator of

$$\mathbb{E}(\boldsymbol{D}_j'\boldsymbol{D}_j)^{-1}\mathbb{E}\left(\boldsymbol{D_j'}\left(\boldsymbol{R}_j - \boldsymbol{D_j}\boldsymbol{\alpha_0}\right)\left(\boldsymbol{R}_j - \boldsymbol{D_j}\boldsymbol{\alpha_0}\right)'\boldsymbol{D_j}\right)\mathbb{E}(\boldsymbol{D}_j'\boldsymbol{D}_j)^{-1'},$$

it will only consistently estimate V_1 if the covariates used to create the value-added measures and the characteristics of interest are uncorrelated such that $\mathbb{E}\left(\boldsymbol{D_j'}\boldsymbol{X_j}\right) = 0$. Given that both $\boldsymbol{D_j}$ and $\boldsymbol{X_j}$ are observable, this condition can be tested using the sample equivalent of $\mathbb{E}\left(\boldsymbol{D_j'}\boldsymbol{X_j}\right)$. A key point here is that $\boldsymbol{R}_j$ is a noisy measure of value added containing true teacher value added $\boldsymbol{\mu_j}$ that is correlated with $\boldsymbol{X_j}$ and the noise term $\boldsymbol{\epsilon_j}$. If $\boldsymbol{D_j}$ and $\boldsymbol{X_j}$ are correlated, consistent estimation of the first step which removes the effect of $\boldsymbol{X_j}$ is required to obtain a consistent estimator of $\boldsymbol{\alpha_0}$. Since ignoring the first steps in calculating standard errors is valid only if inconsistency in the first steps doesn't lead to inconsistency in later steps (Newey and McFadden, 1994), one needs to account for the estimation of current year value added if $\boldsymbol{D_j}$ and $\boldsymbol{X_j}$ are correlated.

1.4 Simulations

I now illustrate my findings using a simulation study. This simulation will show that using unadjusted standard errors with the multi-step OLS estimator will lead to coverage rates that are too low for 95% confidence intervals. It will also show that coverage deteriorates further as correlations between the covariates and true value added increase. On the other hand, the 95% confidence intervals from using the multi-step OLS estimator with corrected standard errors from GMM perform well in all cases. To demonstrate, I draw 1000 replication samples with a total sample size of $n = 900,000$ observations holding the number of students per class n_j and classes per teacher T constant at 30 and 10 respectively, so that the sample contains $J = 3,000$ distinct teachers who each teach 1 class a year for 10 years. The parameters of the simulations are as follows to allow the covariate to be correlated with value added:

1. $\mu_{jt} \sim N(0, 0.01)$.
2. $\rho = 0.5$.
3. $X = \frac{\rho\mu_{jt}+(1-\rho)\mathcal{N}(0,0.01)}{\sqrt{\rho^2+(1-\rho)^2}}$.
4. $R_{it}^{obs} = X_{it} + \mu_{jt} + \epsilon_{it} * U[0,2]$, where $\epsilon \sim \mathcal{N}(0, 0.81)$.
5. $Y_i^{obs} = 5 + 10X_{it} + \kappa_0\mu_{jt} + \eta_{it} * U[0,2]$, where $\eta \sim \mathcal{N}(0, 100)$.
6. $\kappa_0 = 100$.

μ_{jt} has a standard deviation of 0.1, which is in line with estimates in Chetty et al. (2014a) and the simulation in Chetty et al. (2017). Test scores are constructed to be mean zero and standard deviation one, with $U[0, 2]$ allowing for heteroskedasticity. Earnings and κ_0 are chosen to resemble the simulation in Chetty et al. (2017) while allowing for heteroskedasticity. Finally, given that the simulation only uses one covariate, the correlation of this covariate with VA is set to be high at 0.5 to obtain results that are in line with this paper's empirical application.

Each replication first estimates value added as described above and then estimates (1.10), clustering standard errors at the teacher level.

Column 1 of Table A.1 presents the results when the true effect of test score value added on earnings is set to be 100, and κ_0 is estimated using the multi-step OLS estimator and the standard errors are not corrected. Results show that the standard errors obtained from simply clustering and making no other adjustments when using OLS are incorrect. The estimated standard errors were far too small - on average, the standard error estimates were less than two thirds of the correct value. The inconsistent standard errors lead to incorrect coverage rates for 95% confidence intervals, with these intervals containing the true value κ_0 (100) only 72.4% of the time.

Column 2 of Table A.1 presents the results when the true effect of test score value added on earnings is set to be 100, and κ_0 is estimated using the multi-step OLS estimator with the corrected standard errors obtained from the GMM formula. Confidence intervals constructed using this estimator provide correct coverage. The corrected GMM standard errors account for the added variability resulting from the correlations between covariates and true value added, as such the estimated variance is close to the true variance of the estimator.

To better illustrate the role of the correlations between the observable characteristics of students used to estimate value added and true teacher quality, I let ρ be equal 0 , 0.25, 0.5, and 0.75 and draw a 1000 replications for each value. For each set of replications, I repeat the previous exercise and obtain the coverage rate of the estimated confidence intervals from using the multi-step OLS estimator with the incorrect and corrected standard errors. Figure A.1 presents the results of this exercise, while the actual coverage rate of confidence intervals obtained from OLS deteriorates as the correlation between observable characteristics and true quality increases, the confidence intervals constructed using the from the GMM standard errors perform well even when the correlation is set to be unrealistically high with $\rho = 0.75$. To understand why the confidence intervals from OLS provide coverage well below their nominal rates as the correlation increases, I plot for each set of replications the actual standard deviation of $\widehat{\kappa}_{OLS}$ from the Monte-Carlo and the average standard errors obtained from OLS. Figure A.2 presents the results. While the increasing correlation leads the actual standard deviation of $\widehat{\kappa}_{OLS}$ to increase, the average standard errors obtained from OLS do not change since they do not take into account the correlation. This results in a larger drop of coverage for confidence intervals as the correlation increases.

Finally, Table A.2 presents the monte-carlo variances of the multi-step OLS estimator of κ_0 and the optimal GMM estimator of κ_0 from section 1.2.3. The optimal GMM has a slightly a lower variance than the multi-step OLS estimator, specifically it is 1.3% lower.[11]

[11] The small magnitude of the difference could be due to the fact that the simulations have only a moderate amount of heteroskedasticity, no correlations between ϵ and η, and constant VA over time which implies $\mathbb{E}[\boldsymbol{R}_j|\boldsymbol{R}_{\boldsymbol{j}}^{(-\boldsymbol{t})}]$ is linear.

1.5 Application

To further illustrate my results, I draw on administrative data for students in the North Carolina public schools, in the years 2000-2005. Specifically, I focus on third grade students in those years.[12] Students in grade three in North Carolina take end-of-grade tests in math as well as pre-tests in the Fall which will be used as lagged test scores. I standardize all scores within year-grade cell.

I start with 611,870 distinct students, in 1,313 schools. After excluding students with missing test scores, special education classes, classes with fewer than 10 students, and students that are not matched to their teachers, I am left with 444,262 distinct students matched to 8,210 teachers in 22,295 distinct classrooms. Given that the procedure described in section 1.2.1 is a leave-year out procedure, it excludes all teachers who only teach for a year. As such my final sample consists of 388,191 students matched to 5,266 teachers in 19,351 classrooms. I draw long-run outcomes for these students from high-school transcripts (graduation, GPA, class rank), end-of-course algebra scores in high-school, and exit surveys (college plans).

The summary statistics for the sample are presented in Table A.3. Half of the students in my sample are female, around 34% of students are Black or Hispanic, and around 61% of students are white. Only 3% of students are English language learners, whereas 10% are special education students. Of students matched to their long-run outcomes, about 91% graduate from high school, 78% plan on attending college, and 41% plan on attending a 4-year college.

[12] I limit my focus to third grade students from 2000 to 2005 in order to remain as close as possible to the theoretical setting of this paper. I am able to match over 70% of those students to their teacher. Of those matched, none are missing covariates used for VA estimation, and I can observe long-run outcomes for over half of them. Furthermore, the results found in this paper are comparable to those reported in Rothstein (2017) who considers students in grades 3 through 5 for the years 1997-2011.

The remainder of this section is organized as follows. First, I will illustrate the results of section 1.2.3 by showing that that the standard errors routinely used in the literature are incorrect. Next, to understand why the unadjusted standard errors are incorrect, I will show that there is evidence of unconditional sorting i.e strong correlations between the variables used to predict VA and teacher VA. In doing so I will also illustrate the results of section 1.3.

1.5.1 Correcting Inference

To begin, I estimate value-added measures following the procedure laid out in section 1.2.1, notably estimating β_0 using only within teacher variation. I use a rich vector of student controls that is similar to Rothstein (2017) and contains: cubic polynomials in prior scores, gender, age, indicators for special education, limited English, year, lunch eligibility, ethnicity, as well as class- and school-year means of those variables. Table A.4 describes the generated measures, there are 19,351 distinct measures corresponding to the 19,351 classrooms. The measures are mean zero and have a standard deviation of 0.177 such that a one standard deviation increase in estimated VA corresponds to a 17.7% increase in test scores.[13]

It then follows from Theorem 1 that when using the estimated VA measures in a regression on long-run outcomes, we must adjust standard errors to obtain the true variance of κ_0. I focus on estimating the impact of teacher VA on a set of long-run outcomes for the third grade students, namely: high-school algebra scores, high-school graduation, plan to attend college, plan to attend a 4 year college, high-school GPA, and high-school class rank.

[13] Assuming that the VA measures are forecast unbiased.

I estimate the effect of teacher VA on those outcomes following the methodology in section 1.2.1. The results are found in Table A.5. I find that a one standard deviation increase in teacher VA in third grade leads to a 3.84 percent of a standard deviation increase in high-school algebra scores, a 0.5 percentage point increase in high-school graduation, a 0.88 percentage point increase in college enrollment plans, a 1.76 percentage point increase in 4-year college enrollment plans, a 0.036 point increase in weighted high-school GPA, and a 0.93 percentage point increase in high-school class rank. The magnitude of these impacts are similar to the ones obtained by Rothstein (2017) using a larger sample of students from North Carolina, and the ones of Chetty et al. (2014b) in New York.

Importantly, the unadjusted standard errors obtained from simply running an OLS regression and clustering standard errors at the teacher level are incorrect. Indeed, comparing the unadjusted standard errors to those obtained from a GMM estimation of the system shows that the unadjusted standard errors are too small. The clustered standard errors from GMM are 37% to 70% larger than their unadjusted clustered OLS counterparts. Given that the magnitude of the effects are relatively large, the coefficients remain statistically significant even though their t-values decrease substantially. For example, the t-statistic for graduation drops from 7.14 to 4.16, and the t-statistic for planning to attend college drops from 8.18 to 5. In applications for which estimated effects are not so large, this drop could mean the difference between statistically significant and insignificant results. To put this change into perspective, Table A.5 also gives the heteroskedasticity robust standard errors obtained from running an OLS regression. For all but one outcome, the difference between clustered standard errors obtained by running GMM and clustered standard errors obtained by running OLS is at least as large as the difference between clustered and heteroskedasticity robust standard errors obtained by running OLS.

To summarize, adjusting standard errors to account for correlations between teacher VA and the controls, and for the estimation of VA is likely to be important in practice. In this application on data from North Carolina, the increase in standard errors resulting from the adjustment is on average larger than the impact of clustering standard errors.[14]

1.5.2 Presence and Effects of Unconditional Sorting

The previous section has shown that the unadjusted standard errors obtained from simply running OLS are too small. The theoretical results imply that this is likely due to significant correlations between teacher VA and the vector of covariates used to estimate the measures. To examine that, I empirically test for the presence of these correlations.

Similar to Chetty et al. (2014a), given that my VA measures are estimated using within teacher variation in the controls, I can use these measures to estimate the correlations between true VA and the controls. Chetty et al. (2014a) do so by running a univariate regression of the VA measures on lagged test-scores and other covariates.[15] They find positive and statistically significant but small evidence of unconditional sorting on lagged test-scores, better students are assigned slightly better teachers. Given that their point estimate for unconditional sorting is small, and that they obtain VA measures estimated using within and between teacher variation that are highly correlated (0.979) with their original measures estimated using only within teacher variation, they conclude that unconditional sorting is relatively minimal in practice.

[14] In this application, I cluster standard errors at the teacher level. If one were to cluster standard errors at a higher level, say school or school-year, the results should be similar. Indeed, Rothstein (2017) finds that correlations between VA and controls is stronger between schools than within schools, stating that schools with higher VA teachers have much higher prior year test scores and better socioeconomic conditions.

[15] To account for the attenuation resulting from the fact that the VA measures are shrunk, they multiply their coefficients by an estimate $\frac{SD(\mu_{jt})}{SD(\widehat{\mu_{jt}})} = 1.56$ in their data. This ratio is 1.17 in my data and I adjust my estimates accordingly when running the analysis.

I conduct a similar analysis. I find that the degree of unconditional sorting is not minimal but occurs at the group rather than individual student level. As such the class and school-year level means included in the estimation of VA are strong predictors of teacher VA, these findings are consistent with Rothstein (2017). Furthermore, I show that although VA measures estimated using both within and between teachers variation can be highly correlated with the baseline measures using only within teacher variation, they significantly understate the effect of certain teachers and are poor predictors of long-run outcomes. Thus I find that using within teacher variation as proposed by Chetty et al. (2014a) is important in practice.

I estimate the degree of unconditional sorting in two ways. First, I follow Chetty et al. (2014a) by regressing the VA measures on different controls then adjusting the coefficient to account for shrinkage, but correct inference by obtaining the standard errors by bootstrapping the GMM system.[16] Second, I make use of Theorem 3 and run a regression of the preliminary VA measures on controls by GMM estimation of the system in Result 5. The results of this are presented in Table A.6.

Panel A presents the results from the regressions of the VA measures on three different controls: student level lagged test scores, classroom mean lagged test score, and school year mean lagged test score. Surprisingly, the highly significant estimate for unconditional sorting on student level lagged test scores of 0.010 (Column (1), Table A.6 Panel A) is similar to the estimate found by Chetty et al. (2014a) of 0.012, even though the two analyses use different data sets. This suggests a small degree of unconditional sorting on student level past test scores; better students are matched with higher VA teachers. I then further examine this by also regressing the VA measures on the class and school-year average of lagged test scores. I find that sorting on test scores at the class and school-year

[16] I use the bootstrap on the GMM system since the standard errors in Theorem 3 are for a system that regresses preliminary VA measures, not the shrunk measures,x on covariates.

level is significantly stronger than at the individual level. Notably the point estimate of 0.056 for classroom level lagged test scores is highly significant and five times larger than the one at the individual level. Finally, one can see that the unadjusted standard errors from OLS are incorrect, and that the standard errors obtained from bootstrapping the GMM system are about 50% larger.

I confirm these findings with a direct application of Theorem 3. I first estimate a regression of preliminary VA, $\widehat{R}_{jt}$, on classroom mean lagged test score, then re-estimate the parameter using the GMM system as described in Theorem 3. The results correspond to Panel B of Table A.6. The point estimate of 0.073 is larger than the one in Panel A, but they are statistically indistinguishable. Again, the unadjusted standard errors are smaller than the ones obtained by GMM, with the GMM standard errors being approximately 80% larger. The difference between the standard errors is directly explained by Theorem 3. Here the vector of variables $\boldsymbol{D}_j$ is a subset of the covariates used to construct VA, $\boldsymbol{X}_j$, as such it must be that $\mathbb{E}\left(\boldsymbol{D_j'X_j}\right) \neq 0$.

In summation, the results in Table A.6 point to a strong correlation between the covariates used to estimate VA and true teacher VA. Consequently, this means that estimating β_0 without teacher fixed effects, using both between and within teacher variation, will lead to biased estimates and therefore incorrect VA measures and point estimates for long run impacts. To show this, I estimate another set of VA measures without including teacher fixed effects. At first glance, it seems that these measures are very similar to the ones obtained using within teacher variation only. Indeed, Table A.7 shows that the measures have a very high correlation of 0.965.

However, further examination shows that the measures estimated without teacher fixed effects will dramatically understate the effectiveness of certain teachers. As shown in the upper left quadrant of Figure A.3, these measures are sometimes negative for teachers that have positive VA when estimated using within teacher variation. This in turn will lead to biased estimates of κ_0, Figure A.4 illustrates this point. It presents the results of regressions of a variety of long-run outcomes on teacher VA, following the procedure laid out in section 1.2.1. One set of results is obtained by estimating β_0 and β_0^Y using teacher fixed effects, and the other set is obtained by estimating β_0 and β_0^Y without fixed effects. The estimates in blue are the estimates of κ_0 for different outcomes using the exact methodology in section 1.2.1, while the estimates in red are the estimates of κ_0 without including teacher fixed effects to estimate β_0 and β_0^Y. It is clear that although the measures obtained without fixed effects are highly correlated with the baseline measures, they systematically underestimate κ_0.

1.6 Conclusion

In this paper, I consider how to correctly and efficiently draw inference in models using value-added measures in regressions. Starting with models using value-added measures as an explanatory variable, I show that they can be reframed as GMM systems, and use that to construct corrected standard errors for regressions with value-added on the right hand side. I then show that these models can also be written as systems resembling instrumental variables where the preliminary value-added measures for years $s \neq t$ serve as instruments for the preliminary measure in year t. Then when one has more than two years of data, there are multiple instruments available for the measure in year t. I use this overidentifying information to propose a more efficient estimator for the impact of value-added on long-run outcomes using optimal GMM, and to propose a specification test for these models. For regressions using value-added measures as an outcome, I derive corrected standard errors from GMM and provide a testable condition under which unadjusted standard errors can also lead to valid inference. The theoretical results of this paper are checked using a simulation study. Finally, I demonstrate the practical implications of my results in an application on data from North Carolina public schools. I first document the presence of correlations between teacher test-score value added and student observables, and then show that adjusting standard errors to account for those correlations and the estimation of VA measures is relevant in practice.

Chapter 2

Advisor Value-Added and Student Outcomes: Evidence from Randomly Assigned College Advisors

College graduates earn significantly more than those with a high school diploma, and this gap has been widening over time (Oreopoulos and Petronijevic, 2013). The type of postsecondary degrees that students pursue is also a strong determinant of their future earnings. For example, earnings of graduates from the fields of science, technology, engineering and math (STEM) largely exceed those with degrees in non-STEM fields (Hastings et al., 2013; Kirkeboen et al., 2016; Canaan and Mouganie, 2018). Despite these substantial labor market returns, college graduation and STEM enrollment rates remain relatively low. In the United States, only 41.6 and 60.4% of students at four-year colleges respectively graduate within 4 and 6 years of initial enrollment (NCES, 2018). Additionally, only half of freshman college students who initially express interest in pursuing a STEM major eventually obtain a STEM bachelor's degree (Malcom and Feder, 2016). These issues have put the question of how to improve college students' outcomes at the center of ongoing policy debates in the U.S, but not many clear solutions have been put forth.

In an effort to understand how to boost postsecondary outcomes, we focus on an overlooked input in the education production function: the quality of academic advising. While academic advising is offered by most U.S. postsecondary institutions to help students navigate the complexities of college, little is known about whether quality of advising matters for students' academic trajectories. In general, the role of an academic advisor at four-year colleges is to provide students with high touch and personalized support throughout the academic year. Specifically, an advisor's duties are to monitor students' academic progress, provide personalized assistance with selecting courses and developing a plan of study, give information on academic programs and majors, and offer academic and career mentoring. Additionally, freshman or pre-major advisors help students select an appropriate field of study. Advising during the freshman year is particularly important since it is a critical period for both the recruitment of STEM majors (President's Council of Advisors on Science and Technology (2012)) and student retention.[1]

This paper provides the first causal evidence on the effects of college advisor quality on student outcomes. To do so, we first estimate freshman advisor value-added (VA), based on students' course grades, using rich administrative data linking students to faculty advisors at the American University of Beirut, a private 4-year university located in Lebanon. An important feature of the freshman advising system at AUB is that students are randomly assigned to academic advisors. This enables us to compute VA estimates that are free from bias inherent to non-random settings (Rothstein, 2009, 2010), where the student-advisor match is most likely correlated with unobservable factors. We then look

[1] The first-year retention rate is 73.9% among U.S. full-time students who entered college in the fall of 2016 (Clearinghouse, 2018).

at the impact of advisor VA on students' academic performance, retention, graduation and major choice. While the random assignment of students to advisors is unique to our setting, AUB is in many ways comparable to a private four-year university in the United States as we detail in section 2.

Our results indicate that being matched to a one standard deviation higher VA advisor increases freshman year GPA by 5.7 percent of a standard deviation. We further find that advisor grade VA has no significant impact on the likelihood that students persist after freshman year, but it does reduce time to complete the freshman year by 3.1%. Importantly, the benefits of having an effective freshman advisor do not fade out, as we document a 2.5 percentage point (or 5.5%) increase in 4-year graduation rates due to a one standard deviation higher freshman advisor VA. The magnitude of this effect is comparable to estimates from recent evaluations of merit aid programs, as well as interventions that offer students proactive coaching. For example, Bettinger et al. (2016) report that eligibility for California's Cal Grant—which offers 4 years of tuition assistance—raises bachelor's degree completion by 2 to 5 percentage points. Bettinger and Baker (2014) further find that a one-year proactive student coaching program raises degree completion by 4 percentage points.

Effective freshman advisors also influence students' major choices. A one standard deviation higher advisor VA raises high-ability students' likelihood of enrolling and graduating with a STEM degree by around 4 percentage points. These effects are driven by high-ability male and female students who respectively experience a 3.2 and 4.9 percentage point (or 7.8 and 16.3%) increase in the likelihood of enrolling in a STEM major, and comparable improvements in STEM graduation rates. These estimates are close in magnitude to the impact of financial incentives on major choice. Indeed, Denning and Turley (2017) show that eligibility for SMART Grants—which provide low-income students with up to $8,000 to major in technical fields—increases enrollment in STEM majors by 3.2

percentage points.

We further show that advisor characteristics, such as gender or faculty rank, do not seem to predict advisor value-added. However, we do find evidence that gender match between advisors and advisees has positive effects on students' outcomes, especially for women. Using detailed course-level data, we rule out that higher VA advisors push students to take "easier" courses, thereby inflating their freshman GPA and changing their subsequent outcomes. Instead, effective advisors seem to act as coaches or mentors, directly influencing students' grades without altering their course composition. We further construct alternative measures of advisor value-added based on non-grade outcomes and show that these measures of advisor quality also predict significant positive impacts on students' college outcomes. Results from this additional analysis suggests that our findings on the longer term impacts of advisors are, for the most part, driven by grade improvements in freshman year.

Finally, we conduct similar grade and non-grade VA analyses for a sample of students who first enroll at AUB as sophomores with a declared major and who are randomly assigned to faculty advisors within their chosen majors. As we detail in section 2.4.6, the way advising is conducted for these students is close to freshman advising. Notably, we show that our main freshmen results replicate using this new sample. Indeed, we find that sophomores experience a 3.7 percent of a standard deviation increase in their first-year GPA from having a one standard deviation higher grade VA advisor. Sophomore students are also 4.3% more likely to graduate on-time due to a one standard deviation higher advisor grade VA. These results highlight the importance of academic advisors for students at different stages of their postsecondary studies.

This paper is the first to document that effective college advisors largely improve students' academic outcomes. Our findings thus relate to a broad literature focused on how to address low college completion rates and increasing time to graduation in the U.S.

A long body of work examines the role of financial aid in raising degree completion. While some need-based programs are promising (Dynarski, 2003; Bettinger et al., 2016; Castleman and Long, 2016; Barr, 2016; Angrist et al., 2020), much of the research on financial aid has found limited impact on degree attainment (Deming, 2017). Another avenue for improving postsecondary outcomes is to increase per student spending and resources (Bound et al., 2010; Deming and Walters, 2017). Deming and Walters (2017) suggest that increased spending is effective because it can be directed towards academic support services such as advising. Our results corroborate this idea and provide a clear policy recommendation on how postsecondary institutions can promote student success. Specifically, our findings indicate that allocating resources towards improving the quality of academic advising may be an effective way to boost student outcomes.

Another related literature examines whether a variety of interventions can be used to address educational barriers. Programs which offer in-person, individualized and proactive college coaching or advising have shown to substantially increase academic performance (Kot, 2014; Oreopoulos and Petronijevic, 2019) and persistence (Bettinger and Baker, 2014; Carrell and Sacerdote, 2017; Barr and Castleman, 2018; Weiss et al., 2019).[2] On the other hand, light-touch interventions that omit the "personal" element have limited impact on student success. These include nudges, email or text message reminders (Dobronyi et al., 2019; Bird et al., 2021), virtual advising (Oreopoulos and Petronijevic, 2019; Phillips and Sarah, 2019; Sullivan et al., 2019; Gurantz et al., 2020) and in-person but non proactive advising (Angrist et al., 2009; Scrivener and Weiss, 2009; Angrist et al., 2014).

To the best of our knowledge, no prior work has examined whether college advising

[2] Prior work also evaluates counseling programs aimed at increasing *high school* students' access to college or financial aid. These studies show that providing students with one-on-one counseling or assistance significantly increases college enrollment, persistence, and financial aid receipt (Bettinger et al., 2012; Avery et al., 2014; Castleman et al., 2014; Castleman and Goodman, 2018; Mulhern, 2019).

quality matters for students' academic trajectories. Previous studies focus on access to advising (i.e., the extensive margin) and not on the quality of advising. Our finding that quality of advising matters for students' success may thus explain why some of the previously studied advising and coaching programs succeeded and others did not. More broadly, our results emphasize that a reason why some interventions have been successful at boosting college outcomes is because they give students access to high-quality advising. Indeed, programs that have shown the most promise at increasing college completion such as the Accelerated Study in Associates Program—which offers comprehensive student support—have repeated interactive advising as a key component (Weiss et al., 2019).

Our findings also relate to the extensive body of research on the education production function, and the role of school resources and teachers in determining student achievement. Recent studies highlight the importance of teacher value-added in predicting students' outcomes (Staiger and Rockoff, 2010; Jackson et al., 2014; Koedel et al., 2015; Chetty et al., 2014a,b; Jackson, 2018).[3] In line with the evidence on teacher VA, we find that advisors who raise contemporaneous student achievement improve subsequent longer-term outcomes such as graduation. Importantly, we add to this literature by offering a first look into the benefits of academic advising, which is an integral part of most U.S. colleges.[4] In particular, our paper is the first to show that college advisors are an important input in the education production function, and may be just as valuable as

[3] An exception is Carrell and West (2010) who show that U.S. Air Force Academy professors who are effective at increasing contemporaneous student achievement, harm subsequent academic performance. This is because teachers inflate their course grades—by for example, "teaching to the test"—in order to maximize student evaluations.

[4] Indeed, little is known about the role of academic advising in students' college trajectories. Previous studies have examined advising or coaching programs that are operated in partnership with universities but not by colleges themselves. A multitude of papers in the education literature have documented positive correlations between academic advising and students' college outcomes (see Tinto (2010) for a review of the literature). However, these studies do not address the issue of selection bias and hence, cannot cleanly identify causal effects.

teachers in predicting students' success.

Our results further add to an emerging literature that evaluates whether a variety of policies influence students' major choice. Prior work has focused on the role of financial incentives (Sjoquist and Winters, 2015; Denning and Turley, 2017; Evans, 2017), differential pricing of academic programs (Stange, 2015) and timing of course-taking (Patterson et al., 2022) in college major decisions. Our paper complements these studies by showing that advising quality largely influences students' major choice.

By showing that effective advisors increase female STEM degree attainment, we also join a growing literature aimed at identifying strategies to address women's persistent underrepresentation in the sciences. Previous work has highlighted that women are more likely to choose STEM majors and persist in STEM careers when they are exposed to female instructors, role models or advisors in the sciences (Blau et al., 2010; Carrell et al., 2010; Canaan and Mouganie, 2022; Porter and Serra, 2020). However, having a sufficient number of women take on the role of mentors might be difficult given the shortage of females in these fields and since on average, women in academia already allocate more time for service than men (Guarino and Borden, 2017; Buckles, 2019). Our findings suggest that investing in quality of academic advising can promote female STEM degree attainment, without requiring women to take on a disproportionate amount of service work compared to men.

The rest of this chapter is organized as follows. Section 2 provides a detailed description of our institutional setting. Sections 3 and 4 outline our data and methodology, respectively. Section 5 presents our randomization tests and main results. We discuss our findings in section 6 and conclude in section 7.

2.1 Institutional Background

2.1.1 The University

To estimate the impacts of academic advisors' value-added, we exploit a unique feature of the advising system at the American University of Beirut (AUB), which randomly assigns students to faculty advisors. AUB is a small nonprofit private university located in the country of Lebanon. It provides a liberal arts education with an emphasis on undergraduate studies, although it does also offer numerous postgraduate degrees. In total, the university has approximately 50 degrees across a variety of disciplines such as humanities, social sciences, sciences, engineering and medicine. AUB is one of the oldest universities in the region and was established by American protestant missionaries in the year 1866. The sole language of instruction at AUB is English and degrees awarded by the university are officially registered with the New York Board of Regents. It is considered a selective university and has a total enrollment of around 7,000 students. Admission into the freshman year is based on a composite score that is a weighted average of SAT1 scores (50%) and high school GPA in grades 10 and 11 (50%). It is also relatively expensive with an average tuition of approximately $14,000, which is large given the country's average yearly income of $14,846.

Along many dimensions, AUB is comparable to an average private nonprofit 4-year college in the United States. The student to faculty ratio is 11 to 1 and the average class size is less than 25 students. Further, approximately 83% of full-time faculty have doctoral degrees and 50% of students and around 40% of full-time faculty are female. These statistics are similar to the average student to faculty ratio of 10 to 1 at private nonprofit 4-year colleges in the United States. Further, females account for around 55% of all undergraduate students and 44% of all full-time faculty at U.S. post-secondary institutions

(NCES, 2018). Additionally, AUB uses a credit hours system in line with the U.S. model of higher education whereby most courses are worth 3 credit hours and students take an average of 15 credits (5 courses) per semester. Starting with the freshman year, most bachelor's degrees require 120 credit hours or four years to completion.[5]

Our focus in this paper is on students who are initially enrolled at AUB as freshmen. Most students in Lebanon have to pass a national exam at the end of high school, upon which they are awarded a baccalaureate degree (or *Baccalauréat*). Those who pursue the baccalaureate track in high school are ineligible to enroll in university as freshmen, rather they enter as sophomore students with a declared major. Freshman students are those who attended Lebanese or foreign schools that follow the U.S. high school education system and curriculum. Students who initially enter AUB as freshmen are sometimes younger than those eligible to directly enroll as sophomores. Indeed freshman students are, on average, 17.8 years old when they first enroll compared to 18 years of age for sophomore students.[6] Students in our sample are thus academically and culturally more comparable to U.S. rather than Lebanese first-year college students. We should also note that although many freshman students in our sample come from foreign high schools, the role of academic advisors in our setting is *not* to facilitate students' transition into a new country (i.e., making them feel less stressed about moving to Lebanon, etc.). Instead, foreign students are assigned to another mentor whose main job is to help them transition into their new life in Lebanon. Furthermore, foreign school students at AUB are mostly Lebanese expatriates so they are likely already familiar with every day life in Lebanon and may not require much assistance with settling in the country.

[5] The only exceptions are engineering and architecture which require five and six years to completion, respectively.

[6] We calculate students' ages using data on year of birth for both samples. We do not have data on month of birth, so these calculations are a rough approximation of their age.

2.1.2 Academic Advising

At the beginning of their freshman year, students are randomly assigned to academic advisors (or pre-major advisors). Advisors are full-time faculty of professorial rank (Assistant, Associate and Full Professors) chosen from various departments within the Faculty of Arts and Sciences. Preference is given to faculty who are not up for tenure the following year and who are not overloaded with service requirements. Academic advising is counted towards faculty members' service, but additional incentives are in place to encourage volunteering, such as extra research funds or a course release. Faculty commit to advising for the full academic year, and most advise for multiple years.

After deciding on the final pool of advisors, university administrators working within the Faculty of Arts and Sciences randomly assign freshman students to their respective advisors. This is done using a simple two step process. First, students are sorted by either their ID numbers or last names and placed on a list. The method of sorting varies by year, i.e. all freshman students are sorted either by name or by ID within the same academic year. Advisors are then randomly ordered and placed on a separate list. Administrators then pick the first name from the student list and match it to the first name on the advisor list. The second student is then matched to the second advisor and so on. This process continues until all students are matched to an advisor. Importantly, no characteristics of either the advisor or student—such as gender, prior academic performance, or even intended major, etc.—are taken into consideration throughout this process. In section 2.4.1, we confirm that this matching procedure is consistent with what we would expect from the random assignment of students to advisors. This unique institutional feature enables us to identify the causal effect of an academic advisor's VA on students' performance, major choice and graduation outcomes.

Students at AUB typically declare a major at the end of their freshman year, after having completed the requirements for admission into their intended majors. Academic advisors' main tasks are to monitor students' academic progress during the freshman year, help them choose a major and courses, as well as develop a plan of study that will allow them to meet the requirements for entry into their intended majors. Students are advised by the same advisor throughout the freshman year. They are required to meet with their advisors one-on-one at least once per semester and prior to course registration.[7] Advisors further have to hold weekly office hours throughout the semester, and students have the option of contacting them to set up additional out of office hours meetings. They are given access to students' full academic records, including their past high school grades and SAT scores, which allows them to tailor their advice to students' interests and abilities. Advisors are notified of any irregularity or change of status of their respective students—such as whenever students are placed on probation. Additionally, students are not allowed to withdraw from any course without first getting advisor approval.

A key part of an advisor's job is to help students decide on a major and importantly meet the requirements for entry into their intended major. Freshman students apply for a major at the end of their first year of college giving them plenty of time to interact with their advisors before selecting a field of study. Admissions into different majors are granted based upon the fulfillment of credit and course requirements set by departments. Appendix Table A1 highlights an example of the requirements for four different majors—engineering, chemistry, business and history. Regardless of their intended majors, all students have to complete a total of 10 courses in a variety of disciplines (sciences, social sciences, humanities) in order to be eligible complete their freshman year and become sophomores. However, the emphasis on courses taken varies across intended majors. For

[7] Students need a PIN code for course registration that can only be provided by their advisors during those one-on-one meetings, ensuring that they actually meet with their advisors. Furthermore, freshman advisors conduct a group advising session prior to the beginning of the academic year where they introduce students to university resources, the code of conduct and the general requirements for completing their first year and declaring a major.

example, students wishing to pursue science majors such as engineering and chemistry are required to take 2 math and 3 science courses during their freshman year. On the other hand, students who intend on enrolling in other majors such as business and history have to complete only one math and 2 science courses—but have to take more humanities and electives than science majors.

Further, some departments require students to take specific courses. In general, science majors—i.e., engineering, computer science, mathematics, physics, chemistry and biology—are the most restrictive as they require that students take a number of difficult science and math courses. For example, students wishing to pursue engineering have to take Calculus I and II, General Chemistry, and Introductory Physics. In contrast, those who plan on pursuing non-science majors have the option of enrolling in easier math and science courses.[8] Finally, some majors impose admission grade requirements. The most selective majors are engineering which require a minimum cumulative freshman-year GPA of 80 for admission.[9] In our analysis, students' final GPA in each freshman course is the main measure used to construct advisor value-added.[10]

2.1.3 Comparison to Academic Advising at Other Universities

In this section, we discuss how the different features of AUB's academic advising system compare to pre-major advising at 4-year colleges in the United States. First, advising at AUB is carried out by full-time faculty, and around 31 students are assigned to each advisor. A survey conducted by the College Board (2011) among U.S. 4-year colleges found

[8] For example, many of them take "Mathematics for Social Sciences" instead of Calculus.

[9] Freshman students' applications are pooled with those entering directly to the sophomore year, and the admission rate for engineering averages around 17%.

[10] Importantly, courses are not graded on a curve at AUB and, unlike teachers, advisors cannot inflate or manipulate students' grades directly. Further, we standardize all course grades at the class-year level to account for differences in course grading across courses and years.

that full-time faculty advise more than three-fourths of first-year students at 52.4% of responding institutions. This number however varies by type of institution. While 84.1% of surveyed baccalaureate-granting institutions reported that three-fourth of students are advised by full-time faculty, this number is 50% at master's-granting institutions and 22.5% at research universities which mostly rely on professional advisors. Additionally, the National Academic Advising Association reports that in U.S. postsecondary institutions where faculty advise students, the median caseload for a faculty advisor is 25 for small institutions and 45 for medium-sized institutions.[11]

Second, the main goals of advising at AUB are to help students choose a major and courses, develop a plan of study and keep track of their academic progress during the freshman year. These tasks are in line with those emphasized in the U.S. 4-year college advising system. Indeed, according to a survey conducted by the National Academic Advising Association, over 91% of 4-year public and private U.S. colleges stated that they have academic advisors whose responsibilities include helping students develop a plan of study, schedule and register in courses, and select a major (Huber and Miller, 2011). Third, AUB advisors are required to meet one-on-one with students at the beginning of each semester and prior to course registration. The College Board survey (2011) indicates that among U.S. 4-year colleges, 69% of responding institutions also required students to meet with their first-year advisor at least once per term.

To give a clearer idea about how AUB's advising system compares to other settings, we collected information on how pre-major advising is conducted at various selective private 4-year colleges in the United States. We chose 5 liberal arts colleges—Amherst College, Middlebury College, Swarthmore College, Wesleyan University and Williams College—and 5 research universities—Duke University, Harvard College, Princeton Uni-

[11] Small institutions are defined as having an undergraduate enrollment head count of less than 5000 students, while medium-sized institutions have between 6000 to 23,999 students.

versity, Vanderbilt University and Yale University. This information is summarized in Online Appendix Tables A2 and A3. Similar to AUB, advising at most of the liberal arts colleges is conducted exclusively by faculty (last column of Table A2). Research-intensive universities have faculty advisors but advising is also conducted by staff members or administrators.

Interestingly, the tasks of pre-major advisors at both liberal arts colleges and research intensive universities (second column of Table A2) are similar to those of AUB advisors. Specifically, advisors are responsible for helping students set academic and career goals, select courses and choose a program of study. The liberal arts colleges further emphasize that advisors should keep track of students' academic progress and problems. Furthermore, all colleges in Table A2 specify that advisors should meet one-on-one with students several times during the academic year and at least once before course registration—as one of the main goals of advising is to help students select courses.

Finally, AUB advisors have access to students' academic records, are notified when their advisees are placed on academic probation and have to approve course withdrawals. We were unable to find aggregate statistics regarding whether U.S. advisors perform these tasks. However, we were able to find this information for some of the colleges that we collected data on. For example, Online Appendix Table A3 shows that advisors at Middlebury, Wesleyan and Swarthmore have access to their students' academic records. The latter two colleges as well as Williams College also notify advisors when students' academic standing is unsatisfactory. On the other hand, amongst colleges shown in Table A3, only Vanderbilt and Amherst require that advisors approve course withdrawals. In sum, evidence from this section indicates that AUB's academic advising system is comparable in many ways to advising at private 4-year colleges in the United States.

2.2 Data

2.2.1 Data Description

This paper uses student level administrative data acquired directly from the Registrar's office at the American University of Beirut (AUB). These data contain detailed student-level longitudinal information on course grades, credits accumulated, sex, semester GPA, class-year (Freshman, Sophomore, etc...) as well as major during every semester enrolled at university. Importantly, these data also contain information on each student's academic advisor including gender, faculty rank and department. These anonymized data were then matched, through an agreement between the registrar's office and the admissions office, to student baseline information. This enables us to also observe students' Verbal and Math SAT scores, year of birth, high school location as well as legacy status. Our data initially included 4,353 incoming freshmen students matched to 46 faculty advisors at AUB for the academic years 2003-2004 to 2015-2016.[12] We exclude all students who have missing baseline information and all advisors who advised for only one academic year.[13] This leaves us with a final sample of 3,857 freshman students matched to 38 academic advisors.

[12] Freshman students entering university before 2003-2004 had a different advising system in place. For results involving graduation outcomes, we also limit our sample to students entering AUB on or before 2012-2013 in order to observe graduation status for all students.

[13] As we discuss in detail in Section 4, our estimate of value-added (VA) for each advisor-year is computed using a leave one-year-out estimation strategy. Thus, we are unable to compute any VA estimate for advisors who served for one year.

2.2.2 Summary Statistics

Our main analysis involves 3,857 freshman students enrolled in 41,121 courses matched to 38 faculty advisors. Summary statistics for all students and advisors used in our analysis are shown in Table B.2. In columns (1) and (2), we present means and standard deviations for key variables with the number of observations reported in column (3) throughout. We begin by summarizing student baseline characteristics in Panel A of Table B.2. Female students constitute around 48% of individuals in our main sample, compared to 52% male. The average Mathematics and English SAT test scores for freshman students are 573 and 494 points respectively. Approximately 20% of all freshman students are legacy admits, defined as those with a close relative who attended AUB.

Next, we present summary statistics for our main student level outcomes in Panel B of Table B.2. The average freshman GPA is 76.5 out of a possible 100 points with a standard deviation of 9. Relative to all students initially enrolled as freshmen, 79.4% complete the requirements of the freshman year and become sophomores. For students who enter sophomore year, the average time to do so is around 2.5 semesters. Approximately 46% of students initially enrolled as freshmen are able to graduate on-time, i.e., within 4 years of initial enrollment at AUB.[14] Further, around 57.5% of freshmen graduate within 6 years of enrollment.

In our analysis, we focus on the likelihood that students pursue science and business majors (henceforth, selective majors) for several reasons.[15] First, these majors impose more course and grade requirements than other fields and hence prospective students may require a great deal of guidance from their advisors in order to meet the admission

[14] For most majors, on-time graduation is defined as graduating within 4 years. The only exceptions are engineering and architecture which require 5 and 6 years to complete on-time.

[15] This includes all fields of engineering, architecture, Biology, Chemistry, Computer Science, Mathematics, Physics, Statistics and Business majors.

requirements.[16] Second, from a policy perspective, these majors have been shown to have the highest labor market returns (Hastings et al., 2013; Kirkeboen et al., 2016), and governments have been increasingly investing in promoting STEM education. 43% of students in our sample enroll in a selective major and 35.5% of all students eventually graduate from a selective major.

In Panel C of Table B.2, we report statistics for advisor level variables matched to our sample of students. In total, 38 unique faculty members served as freshman advisors for the academic years 2003-2004 to 2015-2016. On average, each advisor spends around 3.5 years advising resulting in 131 advisor-year observations. Around 39% of freshman advisors are female faculty members and 61% are male. This is in line with the overall proportion of female faculty at AUB which stands at approximately 40%. Further, 56.5% of advisors are in a science department and 43.5% are in a social sciences or humanities field within the faculty of arts and sciences. The majority of advisors are at the rank of assistant professor. Indeed, 28% are full professors, 22 are associate and 50% are lecturers or assistant professors. On average, each academic advisor has 31 students per academic year.

2.3 Identification Strategy

2.3.1 Methodology—Computing Value-Added Estimates

We construct advisor value-added (VA) following the methodology presented in Chetty, Friedman, and Rockoff (2014a) with slight modifications to fit our framework. During a

[16] While the business school does not require students to take specific courses, its does have a minimum admission freshman-GPA of 77—which is higher than most other majors.

given year, a typical student is enrolled in around 10 classes (5 during the fall semester, 5 during the spring semester). We predict value-added based on freshman course grades since one of the main roles of an advisor is to track and help improve students' performance during the freshman year. Given that advisors are randomly assigned to students each year, for the purpose of creating VA estimates, an advisor can be thought of as an instructor for multiple different classes in a given year. Accordingly, we define a classroom in this setting as an advisor-year-class cell.[17]

Let students be indexed by i, years by t, classes by c, and advisors by j. Then let student i's final freshman course grade, S_{itc}, in year t and class c be equal to:

$$S_{itc} = \boldsymbol{\beta X_{it}} + \eta_{itc}, \tag{2.1}$$

where:

$$\eta_{itc} = \mu_{jt} + \theta_{ict}, \tag{2.2}$$

and $\boldsymbol{X_{it}}$ is a set of student level covariates that includes math and verbal SAT scores, student gender, and whether the student was a legacy admit. The error term η_{itc} is decomposed into two parts, advisor VA: μ_{jt} (scaled such that the average advisor has a VA of zero and a one-unit increase in VA leads to a one-unit increase in course grades) and a student-class idiosyncratic shock θ_{ict} that is unrelated to advisor quality. As we detail in section 5.1, our data are consistent with what we would expect from the random matching of students to advisors. Importantly, under random assignment, $\boldsymbol{X_{it}}$ and θ_{ict} are balanced across advisors with different levels of VA and are thus uncorrelated with μ_{jt}.[18] Thus, one advantage of our setting is that the average course grades of an

[17] A class refers to all sections of a given course; for example, all students taking Calculus I. Additionally, our results are robust to running the analysis using advisor-year cells.

[18] We also assume that μ_{jt} and θ_{ict} are covariance stationary. This requires that mean advisor quality is constant over time and that the correlation between advisor quality and any shocks across years only depends on the amount of time elapsed between the years. We impose this assumption to be able to

advisor's students can be directly used to construct an unbiased estimate of advisor value added—without the need to impose any additional assumptions.[19]

Due to the random nature of advisor assignment, we do not directly estimate equation (1), rather we start by standardizing student course grades at the class-year level and running a regression of this standardized variable on year fixed effects:

$$S_{itc} = \alpha_t + \nu_{itc}. \tag{2.3}$$

We then create the residuals S^*_{itcj} from Equation (2.3) and collapse them to the advisor-year level $\bar{S}^*_{jt}$ using Chetty, Friedman, and Rockoff (2014a) precision weights which give more weight to classrooms with a lower variance of residual course grades.

The value-added $\hat{\mu}_{jt}$ of advisor j in year t is then constructed by predicting the average $\bar{S}^*_{jt}$ using $\bar{S}^*_{js}$ for all $s \neq 0$ where s is the separation between the years in which the classes were taught. Excluding the year $s = 0$ removes the endogeneity associated with using the same students to form both the treatment and the outcome. *This is equivalent to a leave one-year-out (jackknife) estimate*, where the data from different years are weighted using the method presented in Chetty, Friedman, and Rockoff (2014a) with weights only depending on the lag s:[20]

$$\hat{\mu}_{jt} = \sum_{s \neq 0} \hat{\phi}_s \bar{S}^*_{js}, \tag{2.4}$$

where $\hat{\phi}_s$ are obtained from OLS regressions of $\bar{S}^*_{jt}$ on $\bar{S}^*_{js}$ for each lag s.

Finally, our data include students who took more than one year to complete their freshman year. To account for concerns of mechanical correlations that might arise from these

adjust our VA estimates for drift in advisor quality over time (Chetty, Friedman, and Rockoff, 2014a).

[19] Creating VA following the exact methodology of Chetty, Friedman, and Rockoff (2014a) where grades are first residualized using student covariates yields quantitatively similar estimates of VA. It does however lead to a small loss in precision of VA estimates due to a lower number of observations because of missing covariates for certain observations.

[20] We restrict the covariances for lags greater than 3 years to be equal to the covariance for a lag of 3.

students being matched with the same advisor two years in a row, we compute the VA of advisors based only on the grades of freshman students in their first year of university schooling.

2.3.2 Forecast Unbiasedness of VA estimates

Under the random assignment of students to advisors in a given year t, the average effect on final course GPA of a change in our estimated measure of VA is similar to the average effect of a change in actual VA. To see that, note that given random assignment we have that:

$$Cov(S^*_{itcj}, \hat{\mu}_{jt}) \equiv Cov(\mu_{jt}, \hat{\mu}_{jt}), \tag{2.5}$$

the covariance between residual course grade and estimated VA is equal to the covariance between true VA and estimated VA. This relationship holds because random assignment ensures that all observable and unobservable predictors of course performance are balanced across advisors. Following Chetty, Friedman, and Rockoff (2014a), we consider the following regression of residual course grades on estimated VA:

$$S^*_{itcj} = \alpha_t + \lambda\hat{\mu}_{jt} + \zeta_{itc} \tag{2.6}$$

In our setting we then have:

$$\lambda = \frac{Cov(S^*_{itcj}, \hat{\mu}_{jt})}{Var(\hat{\mu}_{jt})} = \frac{Cov(\mu_{jt}, \hat{\mu}_{jt})}{Var(\hat{\mu}_{jt})}, \tag{2.7}$$

and since $\hat{\mu}_{jt}$ is constructed to be the best linear predictor of S^*_{itcj} we have that $\lambda = 1$ and is the causal impact of being assigned an advisor with a one unit higher VA. We check that this holds in our setting by estimating the regression in Equation (2.6) and testing the hypothesis that $\lambda = 1$. The results presented in Table B.2 show that a one unit increase in estimated freshman advisor course grade VA leads to a statistically significant 0.971 unit increase in freshman course grade. Importantly, we are unable to reject the null hypothesis of $\lambda = 1$. This indicates that a one unit change in our *out-of-sample* estimated VA has the same causal effect on course grades as a one unit change in true VA. This ensures that our estimated VA measure captures the true impact of advisor value-added on longer run outcomes. We also show that our measure of freshman advisor VA is forecast unbiased under different sample splits. Namely, we estimate VA in three other ways: 1) Leaving the current and two previous years out, 2) Leaving the current and two future years out, and 3) Randomly splitting the sample in half, estimating leave-year out VA in one half, and then checking for forecast unbiasedness in the second half of the sample. Online Appendix Table A4 presents results from this exercise. We are unable to reject the null hypothesis of $\lambda = 1$ in any specification, despite sample size reductions from this analysis.

2.3.3 Identifying Equation

Our empirical strategy exploits the random assignment of freshman students to academic advisors at the American University of Beirut. Our main focus involves estimating the causal impact of freshman advisor quality on students' academic outcomes. To capture these effects, we regress student outcomes on estimated advisor course grade VA ($\hat{\mu}_{jt}$) from Equation (4). Specifically, we standardize advisor VA by year ($\hat{m}_{jt}$), and run the following linear regression model for all freshman students matched to an academic advisor:[21]

$$Y_{ijt} = \alpha + \gamma \hat{m}_{jt} + \theta X'_{it} + \lambda_t + \epsilon_{ijt} \tag{2.8}$$

where Y_{ijt} refers to our outcomes of interest for student i matched to advisor j in academic year t. γ is our treatment parameter which captures the average impact of advisor value-added on student outcomes. Our simplest specification includes only these variables and λ_t an academic-year fixed effect that controls for unobserved changes across different years. Intuitively, with the inclusion of year fixed effects, we are comparing students during the same year that are matched with advisors having different VA measures. In alternate specifications, and to alleviate concerns over selection, we further add a set of student controls X'_{it} that should improve precision by reducing residual variation in the outcome variable, but should not significantly alter our VA effects. These controls include students' math and verbal SAT scores, gender and legacy admission status. Finally, ϵ_{ijt} represents our error term. Standard errors are clustered at the advisor-year (treatment) level throughout to account for correlations among students exposed to the same advisor in the same year.

2.4 Results

2.4.1 Tests of the Identifying Assumption

To identify the causal effect of an advisor, it is important that freshman students' characteristics are uncorrelated with their advisor's value-added. The ideal experiment to identify such effects free of bias would be to randomly assign advisors to students. While

[21] The full distribution of the standardized advisor course grade VA measure $\hat{m}_{jt}$ is reported in Online Appendix Figure A1.

our institutional setting provides for random assignment of students to advisors, we perform a series of tests to confirm that our data are consistent with such a process. First, we show that students' predetermined baseline characteristics are uncorrelated with their advisor's VA estimate. To do so, we regress advisor grade VA on a host of student controls including Verbal and Math SAT scores, student gender and legacy status. We include year fixed effects in our regressions to account for any common shocks that vary by cohort. The results of this test are summarized in Table B.3. We find no significant relationship between advisor VA and student ability, student gender or legacy status. Indeed, all coefficients on our student controls are statistically insignificant and reasonably precise. For example, we find that scoring 10 points higher on the math SAT test would lead to at most having an advisor with a 0.99 percent of a standard deviation (0.0099) higher VA. We also find that student characteristics are jointly insignificant, as indicated by a p-value of 0.25 from a test of joint significance. These results are in line with our institutional setting and indicate that students who are assigned to a lower or higher value-added advisor are similar in terms of observable characteristics, consistent with random student-advisor matching.

Second, we complement the above results with additional tests of randomization. Specifically, we use resampling techniques, analogous to those conducted in Carrell and West (2010), to empirically test if our data are consistent with what would be observed from a random process. To do so, we randomly draw 10,000 student samples of equal size for each advisor-year combination without replacement. For each randomly sampled advisor-year combination, we calculate the sums of both the verbal and math scores for all students in that sample. We then compute empirical p-values for each advisor-year based on the proportion of simulations with values less than that of the actual advisor-year sum. Under the random assignment of students to advisors, we would expect that any unique p-value is equally likely to be observed—i.e., that the distribution of empirical p-values should be uniform.

Accordingly, we test for the uniformity of this distribution using both a Kolomogrov-Smirnov one-sample of equality of distribution test and a χ^2 goodness of fit test. These results are summarized in Panel A of Table B.4 and indicate that for all 13 years of our data, we fail to reject the null hypothesis of random assignment for all years based on either test of uniformity. These results hold regardless of whether we use the mathematics or verbal SAT test scores as a proxy for academic ability. In summary, we find no evidence of nonrandom assignment of students to advisors based on academic ability. As an additional test, we also regress these empirical p-values on advisor characteristics, such as value-added and academic rank. These results are reported in Panel B of Table B.4 where we find no statistically significant relationship between our computed p-values and advisor characteristics. We must note however that estimates from Panel B are imprecise mostly because they involve regressions from 131 observations corresponding to the 131 advisor-year combinations in our data.

2.4.2 Freshman Year Academic Performance and Retention

As previously discussed, some of the main tasks of an advisor are to monitor students' academic progress and help them stay on track, with the ultimate goal of preparing students to enroll in a major by the end of their freshman year. Accordingly, we start by examining whether advising quality influences students' freshman year GPA. The corresponding regression estimates are reported in column (1) of Table B.5, with and without the addition of student controls.[22] Throughout our analysis, both freshman GPA and advisor grade value-added (VA) are standardized, and all regressions involve the addition of academic year fixed effects. Results presented in Panel A indicate that

[22] These controls include student gender, Math and Verbal SAT scores as well as legacy status.

a one standard deviation increase in advisor VA raises students' freshman-year GPA by 5.7 percent of a standard deviation. Consistent with the random assignment of students to advisors, the addition of student controls in Panel B does not alter this estimate in a meaningful way. Our estimates on GPA are comparable to professor VA estimates found in other university settings. Indeed, Carrell and West (2010) show that a one standard deviation change in professor quality leads to a 5 percent of a standard deviation increase in course grades at the U.S. Air Force Academy. Further, our estimate on academic performance is only slightly smaller than those found in teacher VA studies in school settings (For examples, see, Kane et al. (2008); Chetty, Friedman, and Rockoff (2014a)). The fact that our estimates are comparable to those from teacher VA studies highlights that students can benefit from different types of interactions with educators. Specifically, while teachers may have more repeated interactions with students, advisors have the advantage of meeting with students one-on-one and providing them with high-touch personalized support.

Next, we examine whether advisors impact students in ways that extend beyond grade improvements. In column (2) of Table B.5, we look at the effect of advisor grade VA on the likelihood that students become sophomores. Since students typically become sophomores after completing all course and credit requirements for the freshman year, this outcome captures first-year retention—i.e., the likelihood that students remain at the university after their freshman year. We find that higher advisor VA has no significant impact on the likelihood that students persist until the sophomore year. On the other hand, column (3) reveals that effective advisors reduce the number of semesters that students take to complete the requirements of the freshman year and become sophomores. A one standard deviation improvement in advisor VA decreases the time to become sophomore by 0.078 semesters. This corresponds to an approximate 3.1% reduction from the baseline mean of 2.48 semesters. This finding is robust to the inclusion of student controls, as indicated by the statistically significant -0.072 estimate reported in Panel B. Panels (a) and (b) of Figure 1 summarize the full distributional advisor grade VA effects for our two significant outcomes, where we display average effect sizes on the y-axis and standardized freshman advisor VA on the x-axis. Notably, advisor quality effect sizes (on student GPA) increase in a roughly linear manner across the whole VA distribution, whereas they decrease linearly with respect to students' time to reach sophomore status.

In Online Appendix Table A5, we conduct heterogeneity analysis for freshman GPA and retention. Overall estimates are restated in column (1) and heterogeneous effects by student ability and gender are reported in columns (2) through (5). We use mathematics SAT test scores as a measure of student ability. Specifically, low-ability students are those scoring below the median math SAT score of their cohort, while higher-ability students are those who score above the median of their cohort. Results presented in columns (2) and (3) of Panel A indicate that the effect of advisor VA on freshman GPA increases with student ability. Having a one standard deviation higher grade VA advisor increases low-ability students' GPA by 4.2 percent of a standard deviation, and by 7.2 percent of a standard deviation for higher-ability students. These estimates are robust to the inclusion of students controls. Results reported in columns (4) and (5) indicate that GPA effects do not differ by gender. Male and female students both experience a 5.4 and 5.8 percent of a standard deviation increase in GPA when exposed to a one standard deviation higher VA advisor, respectively.

In Panel B of Table A5, we examine heterogeneous effects for the likelihood that students declare sophomore status. Consistent with our result for the overall sample, we find that advisor grade VA has no significant impact on the probability that students of different abilities or genders complete the freshman year and become sophomores. On the other hand, Panel C reveals that the overall reduction in freshman year completion time is mostly driven by lower-ability students. Specifically, lower-ability students take 0.107

fewer semesters to become sophomores due to a one standard deviation higher advisor VA—i.e., a 4.1% decrease in time to enroll in the sophomore year. Furthermore, we find that exposure to a one standard deviation higher advisor VA reduces freshman completion time for both male and female students by 0.062 and 0.089 semesters (or 2.45 and 3.66%), respectively. Taken together, our findings indicate that advising quality is critical not only for students' academic performance, but also for improving time to complete the freshman year particularly among low-ability students.

2.4.3 College Completion

Findings from the previous section indicate that high quality advisors substantially improve students' academic performance and time to complete the freshman year. We next examine whether these documented gains persist in the long run and focus on whether freshman advisor grade VA influences college completion.[23] We first look at the likelihood of on-time or 4-year graduation in column (4) of Table B.5. We find that a one standard deviation increase in advisor VA raises the probability of on-time graduation by 2.5 percentage points or 5.5%. The addition of student controls, as shown in Panel B, does not alter results in a meaningful way, as the estimate is slightly reduced to 2.2 percentage points and remains significant at the 1% level. Estimates from column (5) show that advisor VA has no statistically significant impact on 6-year graduation rates, albeit we cannot rule out large effects. These findings indicate that while higher quality advisors do not necessarily influence overall graduation rates, they do however have a large impact on the likelihood that students graduate from university on time. This is consistent with

[23] We note that estimates from this section are based on a reduced sample size of freshman students initially enrolled at AUB from the 2003-2004 to 2012-2013 academic year since we cannot observe graduation for more recent cohorts. In Table A6 of the Appendix, we also report estimates of advisor VA on short run outcomes using the sample of freshman students entering AUB for the years 2003-2004 to 2012-2013. Our documented short run effects remain qualitatively similar using this reduced sample.

our finding that higher advisor VA does not affect the likelihood that students declare sophomore status, but significantly reduces time to complete the freshman year. Panels (c) and (d) of Figure 1 summarize these effects across the full VA distribution. Advisor VA effect sizes on four-year graduation rates exhibit an approximately linear increase with significantly negative effects for students matched to freshman advisors at the lower end of the VA distribution. On the other hand, we find no significant pattern of advisor VA effects on six-year graduation rates.

Heterogeneous effects for graduation outcomes are presented in Appendix Table A7. In columns (2) and (3), we report estimates for students with different levels of ability. For on-time graduation (Panel A), both low and higher-ability students are 2.4 and 2.3 percentage points (or 5.7 and 4.6%) more likely to graduate within 4 years when matched with a one standard deviation higher VA advisor. On the other hand, consistent with the effect for the overall sample, we detect no significant impacts on 6-year graduation rates for both low and higher-ability students (Panel B). In columns (4) and (5), we report heterogeneous effects by gender. We find that a one standard deviation improvement in freshman advisor VA increases men's likelihood of graduating on-time by 3 percentage points (or 6.2%) and no significant impact on 6-year graduation. We do not detect any statistically significant effects on female students' 4 and 6-year graduation rates, but reduced precision prevents us from drawing definitive conclusions regarding their graduation outcomes.

2.4.4 Major Choice

One of the main tasks of an academic advisor is to help students select a major and guide them on how to meet the requirements for admission into their preferred field of study. We therefore examine whether advising quality influences the likelihood that students enroll and eventually graduate from selective majors.[24] As discussed in section 2.1.2, selective majors have more stringent entry requirements compared to other fields of study. As a result, students wishing to enroll in these majors may require a lot of guidance from their freshman-year academic advisor. The different columns in Table B.6 report estimates for the impact of advisor grade VA on students' major choice.[25] For our overall sample, results in Panel A and column (1) indicate that a one standard deviation increase in advisor VA raises the probability that students enroll in selective majors by 2.4 percentage points or 5.6%. The estimate for graduating from a selective major is on the order of 1.5 percentage points (or 4.2%) and is only statistically significant at the 10% level.

These overall effects may mask contextual heterogeneities, as selective majors are potentially more accessible to the highest-ability students. We therefore examine heterogeneous effects by student ability in columns (2) and (3). We define top students as those scoring in the top 75th percentile of the math SAT distribution (i.e., above 600), and non-top students as those with a score below 600. Estimates reported in columns (2) and (3) of Panel A confirm that the highest-ability students are indeed driving the overall effects on selective major enrollment. We find that a one standard deviation increase in freshman advisor VA raises top students' likelihood of enrolling in a selective major by a large and statistically significant 4.9 percentage points (or 8.6%). This is coupled with a similar and significant 3.9 percentage points (or 8.4%) increase in top students' probability of graduating from these majors, indicating that the initial enrollment effects persist in the long run and that virtually all students who are shifted into these majors end up

[24] Recall, we define selective majors as those in the sciences and engineering as well as business degrees. These degrees also happen to correspond to those with the highest earnings potential.

[25] All regressions in Table B.6 include student controls and year fixed effects.

graduating.[26]

Heterogeneous effects by gender, presented in columns (4) and (5) of Panel A, reveal that both top female and male students benefit from being matched to an effective advisor. Specifically, top male and female students with a one standard deviation higher advisor VA are 5.1 and 4.4 percentage points more likely to enroll in a selective major, respectively. Men are also 4.8 percentage points more likely to graduate from these majors. We do not detect significant graduation effects for women, albeit estimates are fairly imprecise.

In Panels B and C of Table B.6, we estimate effects separately for STEM and Business majors. For STEM majors, results are consistent with those for selective majors. Estimates in columns (2) through (5) of Panel B indicate that non-top students' STEM outcomes are not positively affected by a higher VA advisor. However, both top female and male students experience significant increases in the likelihood of enrolling and graduating from STEM fields. Indeed, a one standard deviation higher VA advisor increases top students' likelihood of enrolling and graduating from a STEM major by 3.8 and 4.2 percentage points, respectively. For top male students, this corresponds to a 3.8 percentage point (or 11.6%) increase in graduation with a STEM degree. For top female students, both STEM enrollment and graduation are statistically significant and on the order of 4.9 and 4.6 percentage points (or 16.3 and 19.8%), respectively.

Finally, estimates presented in Panel C of Table B.6 show a 1.3 percentage point increase in the likelihood of majoring in Business for the overall sample, and that this effect is concentrated among non-top students and top male students. Put together, our findings indicate that effective advisors shift students toward selective majors, and that these

[26] All panels in Figure 2 summarize distributional VA effects on selective major enrollment and graduation for the overall and the top student. All figures indicate a roughly linear increase in the effects of advisor quality on students' enrollment and graduation rates across the whole VA distribution. Notably, the documented average positive effects on selective major enrollment for top students seem to be driven by the best set of advisors.

effects are driven by an increase in STEM enrollment and graduation for top students and smaller increases in Business enrollment for non-top students.

2.4.5 Non-Grade Measures of Freshman Advisor Value-Added

We document that advisor value-added has a significant impact on students' college outcomes. Our constructed measure of VA is based on freshman course grades as tracking and helping improve student performance is one of the most important roles of an advisor. However, a good freshman advisor may also directly influence students' major choice and help them persist at university. Accordingly, we check whether non-grade measures of student outcomes are also good predictors of advisor quality and to what extent such VA measures correlate with our current measure of advisor VA. To do so, we introduce two new measures of advisor value-added based on student persistence and major choice indices. Specifically, using a principal component analysis (PCA) decomposition, we first create two student-level indices: a *Persistence Index* and a *Selective Index.* The *Persistence Index* is composed of five key university persistence measures: 4-year graduation, 6-year graduation, freshman drop-out, proportion of courses withdrawn and proportion of courses failed during freshman year. The *Selective Index*, which measures major selectivity, is computed using three key major choice variables: selective major enrollment, selective major graduation and proportion of key science and math courses taken during freshman year.[27] Following the Chetty, Friedman, and Rockoff (2014a) method introduced in Section 4.1, we then compute two separate leave one-year-out measures of advisor VA (i.e., a Persistence and a Selective VA) based on the *Persistence Index* and *Selective Index*, respectively.

We start by looking at how these two non-grade VA measures affect student outcomes.

[27] We standardize these indices to have a mean of zero and standard error of one.

Online Appendix Table A8 summarizes results from this exercise. Estimates from column (1) indicate that a one standard deviation increase in persistence and selective advisor VA leads to a significant 4 and 5 percent of a standard deviation improvement in freshman year GPA respectively. Additionally, columns (2) and (3) show that both advisor VA measures predict positive impacts on students' 4 and 6-year graduation rates, although not all estimates are statistically significant at conventional levels. We also find that having an advisor with a higher persistence or selective value-added lowers the likelihood of course withdrawal and failure for students. Finally, we show that being matched to an advisor with a higher persistence or selective VA increases students' chances of enrollment and graduation from selective majors (STEM + Business) as well as the likelihood of taking key science and mathematics courses during freshman year. Put together, these findings indicate that non-grade measures of advisor quality also predict significant positive impacts on students' college outcomes.

To better understand whether advisors who are skilled at raising student grades are also effective in promoting persistence and influencing major choice, we next report correlations between our previously constructed advisor VA measure (Grade VA) and our two new measures (Persistence and Selective VA). Reporting raw correlations between the different VA measures will understate the true correlations since advisor effects may be estimated with error (Beuermann et al., 2020). To correct for this attenuation, we follow Abdulkadiroglu et al. (2020) and Beuermann et al. (2020) and obtain Maximum Likelihood estimates of the true correlations between VA dimensions. Specifically, let μ_{1j} be the persistent VA of advisor j along skill dimension 1, and μ_{2j} be the persistent VA of advisor j along skill dimension 2. Additionally, let ζ_{1jt} be a transitory effect of advisor j along dimension 1, and ζ_{2jt} be a transitory effect of advisor j along dimension 2. Our raw estimated measures of VA along a single dimension in a given year contain both persistent VA and the transitory effect. Under the assumption of joint normality, as in equation (2.9), we can recover the true correlation between two different VA measures (ρ_{12} in equation (2.9)) net of estimation error using Maximum Likelihood.

$$\begin{bmatrix} \mu_{1j} \\ \mu_{2j} \\ \zeta_{1jt} \\ \zeta_{2jt} \end{bmatrix} \sim N \left(0, \begin{pmatrix} \sigma^2_{\mu_{1j}} I_J & \rho_{12} I_J & 0 & 0 \\ \rho_{12} I_J & \sigma^2_{\mu_{2j}} I_J & 0 & 0 \\ 0 & 0 & \sigma^2_{\zeta_{1jt}} I_{JT} & 0 \\ 0 & 0 & 0 & \sigma^2_{\zeta_{2jt}} I_{JT} \end{pmatrix} \right) \tag{2.9}$$

Results reported in Appendix Table A9 indicate positive correlations between all our different measures of advisor VA. Specifically, Grade VA and Persistence VA exhibit a correlation of $\rho = 0.59$, while Grade VA and Selective VA exhibit a correlation of $\rho = 0.60$. Strikingly, the correlation between Persistence VA and Selective VA is quite high (ρ=0.78) suggesting overlapping advisor skills in affecting these outcomes.

To further understand these patterns, we conduct additional explanatory analysis in the spirit of Jackson (2018). Specifically, we separately regress students' Freshman GPA, *Persistence Index* and *Selective Index* (we also refer to these 3 outcomes as skill measures) on their respective advisor VA measures. We present coefficients from these regressions in Table B.7 where both treatment (VA measure) and outcome are standardized. Estimates from columns (1) and (5) indicate that advisors who raise a given skill measure out of sample have large and statistically significant effects on those same student skills. For instance, the 0.04 estimate from Panel A, column (5) indicates that being matched to an advisor with a one standard deviation higher Persistence VA increases students' persistence index by 4 percent of a standard deviation. Next, we separately regress each of our skill measures on advisor VA estimates computed using *other* skill measures. The results from this analysis are in line with the documented positive raw correlations across advisor VA measures. Indeed, estimates from both Panels of column (2) indicate that advisors who are skilled at increasing persistence and access to selective major positively and significantly affect students' GPA. Furthermore, coefficients in column (4), Panels A and B, reveal that advisors who are effective at improving course grades also positively affect persistence and major choice.

Finally, we present estimates from regressions that simultaneously include grade and non-grade VA measures in Table B.7. Estimates from columns (3) and (6) of Panel A, indicate that, conditional on advisors' grade value-added, persistence VA is no longer statistically related to students' GPA (0.023) or *Persistence Index* (0.023). However, grade VA is still predictive of student GPA (0.055) and *Persistence Index* (0.052) even when controlling for advisors' persistence value-added. This suggests that there is overlap between these two advisor skills. On the other hand, regressions that simultaneously control for grade and selective VA result in statistically significant estimates for both VA measures as shown in Panel B, columns (3) and (6).This suggests that advisor skills needed to improve grades and influence major choice, though correlated, seem to be somewhat complementary.

2.4.6 Additional Results Based on Major Advising

In this paper, we focus on pre-major advising conducted during the freshman year. However, an advantage of our data is that we are also able to look at the impacts of *major advisors*—i.e., advisors who mentor students after they declare a major—for a different sample of students. As discussed in section 2.1.1, most Lebanese students pursue a baccalaureate track in high school rendering them ineligible to enroll in college as freshmen. For these students, their final year in high school is considered equivalent to the U.S. college freshman year. Therefore, they enroll in college as sophomore students with a declared major immediately after finishing high school. We refer to students from this sample as the “sophomore sample”. New enrolling sophomore students at AUB are randomly assigned to faculty advisors within their chosen major using a process similar to that used for freshmen. Specifically, all sophomore students are sorted by either their ID or last name within their respective departments. The chosen method of sorting is the same across all departments in the same year. Advisors within each department are then sorted randomly on a separate list. Students are then matched to advisors within their respective departments using the same matching process as the one used for freshman students. Each student has the same advisor for the entire academic year. Sophomore advisors’ (henceforth major advisors) main tasks are to help students select courses and develop a plan of study that allows them to meet the requirements for graduating from their majors. They also monitor students’ academic progress, have access to their academic records, are notified of their students’ change of status, and are required to meet with students one-on-one at least once at the beginning of each semester.

In this section, we examine how major advisor VA impacts sophomore students’ college outcomes. Extending our main analysis to the sophomore sample has several advantages. First, our main freshman sample includes 38 unique advisors. This limited number of advisors may render our results less generalizable to other settings thus potentially limiting their implications for policy discussions. In contrast, our sophomore sample allows us to observe a significantly larger number of unique advisors (194 unique advisors). Importantly, the sophomore advising system at AUB shares many similarities with freshman advising. Sophomores in our setting are comparable to freshmen in that they are both first-year college students and hence, we are essentially capturing the impacts of first-year college advising for both samples. Additionally, the tasks of a sophomore major advisor are similar to those of a pre-major advisor. The only difference between these two types of advising is their end-goal: while freshman advising is intended to keep students on track to declare a major, sophomore advisors help students stay on track to graduate from their chosen major. Second, our sophomore sample analysis provides new insights into the role of *major advising* which is important in and of itself, as it is offered by most U.S. private 4-year colleges. Indeed, all colleges shown in Table A2 offer one-on-one major advising that is comparable in its goals and the way it is conducted to the one in our setting.[28]

Sophomore Sample Summary Statistics

Before presenting our additional analysis, we describe the sophomore sample. Online Appendix Table A10 summarizes key statistics for the sample of 14,055 first-time enrolling sophomore students at AUB for the academic years 2003-2004 to 2015-2016.[29] Importantly, these students are matched to 194 distinct advisors during this time period, thus providing for a much larger number of student-advisor interactions than the freshman sample.[30] Students from this sample are approximately 48 percent female, similar to the freshman sample. Additionally, 25 percent of admissions are legacy students. Students score an average of 644 and 530 points on the Math and Verbal SAT exams, respectively. Notably, these scores—particularly for the Math SAT—are significantly higher than those from the freshman sample. This is not surprising as students enrolling at AUB as sophomores spend an extra year in high school that is considered equivalent to the freshman university year.

Panel B of Appendix Table A10 shows means for the sophomore sample's main outcomes. Only 8.8 percent of students drop out after sophomore year. To examine the impacts of major advisor VA on longer-term outcomes, we focus on 4-year and 6-year graduation rates as in the freshman sample. However, since sophomore students enroll in college

[28] Of course, this is with the exception that sophomores at AUB are first-year college students, while U.S. sophomores are second-year college students.

[29] Freshman students who eventually become sophomores at AUB are dropped from this analysis, since our focus here is on first-time advising effects.

[30] For the graduation sample, this number shrinks to 152 unique advisors.

one year after freshmen, we slightly modify our definition of these variables. Specifically, for the sophomore sample, 4-year graduation—which is our measure of on-time degree completion—is defined as graduating within 3 years from initial enrollment at AUB. Similarly, 6-year graduation, our measure of overall degree completion, is defined in this case as graduating within 5 years from initial enrollment at AUB.[31] Appendix Table A10 reveals that 79.6 percent of students end up graduating overall, while only 52.9 percent complete their degree on time. We also look at whether major advisor VA influences the probability that students graduate from their initial major—i.e., their declared major in the first semester of their sophomore year. 40.5 percent of sophomore students graduate on-time and 55.4 percent graduate overall from their initial majors.

Finally, Panel C summarizes sophomore advisor level characteristics. 31 percent of major advisors are female and around half are in science departments. Additionally, advisors are well represented across all faculty ranks and each major advisor has an average of 19.1 sophomore students to advise per year.

Sophomore Advisor Course Grade VA

We examine whether a higher value-added major advisor predicts improved outcomes for students entering AUB as sophomores. To do so, we first construct course grade value-added measures for major advisors following the Chetty, Friedman, and Rockoff (2014a) method introduced in section 2.3.1. One notable difference between this and our previous analysis involves the need to include department fixed effects to compute unbiased measures of VA. This is because sophomore students are randomly assigned to faculty advisors in the department corresponding to their declared major. Results presented in Online Appendix Table A11 confirm that this procedure results in a forecast

[31] Compared to other majors, engineering and architecture require 1 and 2 more years to complete on-time, respectively. We accommodate for this in our definitions of 4 and 6-year graduation rates.

unbiased measure of advisor grade value-added by showing that a one unit increase in estimated advisor course grade VA leads to a statistically significant 0.991 unit increase in sophomore course grade.[32] To analyze major advisor quality effects, we regress sophomore student outcomes on our constructed and standardized advisor course grade VA estimate ($\hat{\nu}_{jdt}$) using the following linear regression model:

$$Y_{ijdt} = \alpha + \beta\hat{\nu}_{jdt} + \theta X'_{it} + \zeta_d + \lambda_t + \epsilon_{ijdt} \tag{2.10}$$

Here, β captures the effect of major advisor VA on sophomore student i matched to advisor j in department d and cohort t. Our identifying equation now includes department fixed effects ζ_d since randomization of students to advisors occurs within departments.

Before presenting the main findings from this exercise, we show that our data are consistent with what we would expect from the random matching of sophomore students to advisors within departments. In Online Appendix Table A13, we show that advisor grade VA is unrelated to students' baseline characteristics. Results indicate that conditional on department fixed effects, students' Math and Verbal SAT scores, gender and legacy status are not statistically related to their advisors' value-added. Additionally, these variables are jointly unrelated to advisor grade VA, as the p-value from the test of joint significance is equal to 0.462.

Table B.8 presents the main findings from our course grade VA analysis. Results in Panel A indicate that being matched to a major advisor with a one standard deviation higher grade value-added increases students' sophomore-year GPA by 3.7 percent of a standard

[32] We further check that our measure of sophomore advisor grade VA is forecast unbiased under different sample splits. Namely we estimate sophomore advisor VA in three other ways: 1) Leaving the current and two previous years out, 2) Leaving the current and two future years out, and 3) Randomly splitting the sample in half, estimating leave-year out VA in one half, and then checking for forecast unbiasedness in the second half of the sample. Appendix Table A12 presents results from this exercise. We are unable to reject the null hypothesis of $\lambda = 1$ in any specification, despite the reduced sample sizes from this analysis.

deviation, but does not significantly affect their dropout rate one year after the start of their sophomore year. Column (3) further reveals that a one standard deviation higher quality major advisor increases the likelihood of on-time graduation by 2.3 percentage points, i.e. 4.3 percent. It also increases the overall graduation rate by 1.6 percentage points (column (4)), though this estimate is only significant at the 10 percent level. Estimates in columns (5) and (6) suggest that the documented effect for on-time degree completion is mostly due to students being more likely to graduate on time from their initial declared major. Specifically, we find that a one standard deviation higher advisor VA increases the probability that students graduate on time from their initial major by 2 percentage points, but we detect no statistically significant effects on ever graduating from initial major.

In our freshman sample analysis, we showed that having a higher VA advisor increases the probability that students enroll in STEM majors. Since sophomore students enroll at AUB with a declared major, we cannot look at whether advisor VA impacts their STEM enrollment. An analogous analysis in this case is to examine whether effective advisors are most beneficial for sophomore students whose initial declared major is STEM—i.e., those who declared a STEM major in the first semester of their sophomore year. Indeed, students in these fields may require a great deal of assistance from their advisors since they are the most difficult and competitive majors at AUB. Panels B and C of Table B.8 report estimates of the impact of advisor grade VA on students who initially enrolled in STEM and non-STEM majors. We find that being matched to a major advisor who has a one standard deviation higher grade VA is associated with a 4.1 and 2.4 percent of a standard deviation increase in sophomore-year GPA for STEM and non-STEM students respectively. Nonetheless, the documented overall effects for all other outcomes are concentrated among STEM students. Columns (3) to (5) respectively show that a one standard deviation increase in advisor grade VA significantly raises STEM students' on-time degree completion by 3.4 percentage points, overall graduation by 2.6 percentage points and on-time graduation from their initial major by 3 percentage points. On the other hand, no statistically significant effects are detected for any of the non-STEM students' outcomes.[33]

Taken together, these results indicate that major advisor grade VA has significant impacts on sophomore students' outcomes. Importantly, while there are slight differences in magnitudes, findings from this exercise are consistent with those from the freshman sample. This further solidifies the importance of college advisors in the education production function.

Sophomore Advisor Persistence VA

The role of an effective major advisor is not restricted to improving students' academic performance. Effective advisors can also directly impact students' persistence in the major. As a result, we check whether non-grade measures of student outcomes are also good predictors of major advisor quality. To do so, we first construct a measure of advisor value-added based on a sophomore student persistence index. Specifically, using a principal component analysis (PCA) decomposition, we create a sophomore *Persistence Index* composed of five key university persistence measures related to sophomore students' success: 4-year graduation from initial major, 6-year graduation from initial major, dropout after sophomore year, proportion of courses withdrawn and proportion of courses failed during sophomore year. We then construct a non-grade measure of advisor value-added based on this index. Finally, we regress all our outcomes of interest on this constructed sophomore advisor Persistence VA measure.

[33] The full distributional effects of sophomore advisor grade VA are summarized in Figure 3. Notably, the documented increase in average sophomore GPA seems to be driven by a decrease in the GPA of students matched to the worst set of sophomore advisors (Panel (a) of Figure 3). Additionally, the best set of sophomore advisors seem to significantly improve four and six-year graduation rates.

Findings from this exercise are summarized in Online Appendix Table A14. Strikingly, results presented in column (1) indicate that having an advisor with a one standard deviation higher persistence VA has no overall significant effect on students' sophomore GPA. Rather, higher persistence VA advisors only positively impact the GPA of students in STEM majors (0.025), while having no significant effect on those in non-STEM fields (0.003). Estimates from columns (2) through (6) of Table Table A14 indicate that persistence VA measures of advisor quality predict large and robust impacts on all student measures of persistence and graduation. Notably, these effects are significant and comparable for both STEM and non-STEM majors.

Our findings suggest that advising skills that improve students' persistence in their chosen majors are somewhat distinct from those that improve grade performance. We investigate this further by running additional regressions using our two measures of sophomore advisor value-added together. Specifically, we regress students' sophomore GPA and Persistence Index on their respective advisor VA measures separately and jointly. We present estimates from these regressions in Online Appendix Table A15. Notably, results presented in columns (2) and (3) indicate that advisors who are good at increasing students' persistence do not seem to increase their sophomore GPA. Additionally, estimates presented in columns (5) and (6) show that sophomore advisor persistence VA predicts higher student persistence even after controlling for advisors' grade VA. Taken together, results from this analysis indicate that sophomore advisors who increase students' grades have different skills than those who improve persistence. This provides further evidence that post-major advisor skills are distinct for these two measures of value-added. Overall, results presented in this section using the sophomore sample indicate that our documented findings on freshman advisor VA extend, and replicate, to a larger set of students and advisors.

2.5 Discussion

2.5.1 Discrete Treatment—High and Low Grade VA advisors

We have shown that higher grade VA academic advisors improve students' college outcomes. These positive effects could be masking some interesting treatment heterogeneity relevant for policy analysis. For example, how would students be affected if they were matched to a high-performing advisor rather than a low-performing one? Accordingly, we next estimate the impact of being matched to advisors in different quartiles of the grade VA distribution. These estimates are presented graphically in Online Appendix Figures A2 and A3 for the freshman sample and in Figure A4 for the sophomore sample. Specifically, the different panels plot point estimates and 95% confidence intervals representing the effects of being matched to advisors in the bottom and top two quartiles of the VA distribution—with the second quartile as our excluded baseline category.

Estimates presented in Figure A2a indicate that top quartile freshman advisors substantially improve students' first year GPA by approximately 10 percent of a standard deviation relative to advisors in the second quartile (omitted category). However, we must note that these estimates are relatively noisy which precludes us from making any definitive conclusions related to differences between top versus bottom advisors. Indeed, the top 95% confidence interval for bottom advisors overlaps with the bottom 95% confidence interval for top advisors. Estimates for time to declaring sophomore status mirror those for GPA, as shown in Figure A2b. Specifically, students matched to the lowest grade VA freshman advisors take approximately 0.12 more semesters to complete the freshman year compared to those in the second lowest quartile. However, these effects are relatively imprecise as we cannot rule out that effects are similar across the various VA quartiles. We next examine whether the impacts of top and low-performing advisors persist in the long run by focusing on graduation outcomes. One caveat to keep in mind when interpreting graduation effects is that they are based on a reduced sample of students, since we cannot observe graduation outcomes for more recent cohorts, resulting in a loss of precision. Estimates in Figure A2c and A2d suggest that being matched to a top rather than second quartile advisor results in an increase in on-time graduation and six-year graduation respectively, significant at the 10% level only. However, we are unable to determine that graduation estimates are statistically different across quartiles. Panels (a) through (d) of Appendix Figure A3 show how freshman advisors in different quartiles of the VA distribution impact students' enrollment and graduation from selective majors. For both the overall sample (Figures A3a and A3b) and top students (Figures A3c and A3d), going from a second quartile to top advisor increases the likelihood of enrollment and graduation from selective majors, though these effects are not statistically significant for graduation. Taken together, results from the freshman sample suggest that students benefit the most from being matched to advisors in the top quartile of the VA distribution. However, we are unable to draw any strong conclusions from this exercise, particularly as it relates to differences between bottom and top advisors.

Finally, we run a similar analysis on the population of students entering AUB as sophomores, as they are matched to a larger number of advisors, which could help improve precision. Panels (a) through (d) of Appendix Figure A4 summarize findings from this analysis. Notably, estimates for GPA are less noisy as we find that top versus bottom sophomore advisors have significantly different effects on students' first year GPA. These effects seem to be driven by bottom quartile advisors who reduce students' GPA by approximately 7.5 percent of a standard deviation relative to second quartile advisors. We also uncover evidence suggesting that bottom quartile advisors worsen students' four and six-year graduation rates, though effects similar to advisors in the top two VA quartiles cannot be ruled out.

2.5.2 Potential Mechanisms

In this paper, we find that academic advising quality substantially impacts students' college outcomes. In this section, we discuss the mechanisms that could explain the documented effects. Our first set of results show that effective advisors largely improve students' course performance. There are several potential explanations for this finding. First, it is possible that advisors directly improve students' academic performance by providing them with mentoring, coaching and affirmation effects—especially since they have the opportunity to continuously and repeatedly interact with students during their first year. Another possible explanation is that high quality advisors encourage students to enroll in a specific set of courses that maximize first-year grades (or "easy" courses). To understand which of these two explanations is more likely, we make full use of our data and look at the effects of advisor grade VA on students' course-level outcomes.

We start by looking at the impact of advisor VA on the likelihood that students take challenging courses during their first year. We focus on our main analysis sample which includes freshman students matched to pre-major advisors. We do not conduct this analysis for the sample of sophomores matched to post-major advisors, as sophomore students typically have to take courses that are required by their major during their first year and hence do not have much flexibility in terms of first-year course choice. The results from our freshman sample analysis are reported in Table B.9 separately for the first (Panel A) and second (Panel B) semesters of the freshman year. The most challenging courses during the freshman year are math and science courses that are required for entry into selective majors.[34] Strikingly, estimates from column (1) of Table B.9 reveal that advisors do not push students towards or away from core science and math courses.

[34] These include Calculus I and II as well as Physics, Chemistry, Biology and Computer Science courses targeted for students intending to major in these fields.

Importantly, estimates are small in magnitude and reasonably precise. This result is at odds with our second interpretation in which advisors may influence students' grades by changing their course composition.

While freshman advisors do not influence course choice, estimates presented in column (2) indicate that students are 0.9 percentage points less likely to fail courses due to a one standard deviation higher advisor VA. This corresponds to a 13.4 and 12.5 percent reduction in the likelihood of failing a course during the first and second semesters, respectively. A more telling result is that a one standard deviation improvement in advisor VA decreases the likelihood that students withdraw from a course by 0.5 percentage points or 9.4 percent during the first semester of the freshman year (column (3) and Panel A). Students can only withdraw from courses after meeting one-on-one with their pre-major advisors, and advisors have to approve course withdrawals. This suggests that effective advisors encourage students to persist in their courses, and provide positive affirmation and coaching directly influencing students' grades.[35] Interestingly, the estimate in Panel B reveals that advisor grade VA has no significant impact on course withdrawal during the second semester of the freshman year. This potentially indicates that with time, advisors (or students) acquire more information about their students' (own) abilities, pushing students in the second semester to take courses that match their interests and thereby reduce the chances of withdrawing from courses. Taken together, findings from columns (1) through (3) of Table B.9 indicate that the documented improvement in overall Freshman GPA is most likely due to direct coaching and mentoring provided by advisors and not due to behavioral changes in course selection.

Our findings on the importance of academic advisors are not limited to grade improvements, rather they also extend to other college outcomes such as persistence and major

[35] Sophomore students are not required to meet with their advisors or get their approval to withdraw from courses. As such, examining the relationship between post-major advisor VA and sophomore students' course withdrawal does not allow us to understand whether affirmation effects are at play.

choice. We cannot conclusively speak to the exact mechanism behind these longer run effects, but some of our previous analyses can help shed light on what is driving these effects. For freshman students, the effects we document on student persistence measures, such as time to complete the freshman year and 4-year graduation, are most likely explained by the documented improvement in academic performance during the freshman year. Indeed, higher grades and the lower likelihood of failing and withdrawing from courses increase the odds of successfully completing freshman year. This in turn can lead to a positive feedback loop where the documented increase in performance during freshman year enhances students' confidence and learning thus further bettering future academic outcomes such as on-time graduation. This interpretation is in line with estimates reported in column (6) of Panel A, Table B.7 which indicate that, conditional on being matched to an advisor who is good at improving students' grades, being matched to an advisor who is skilled at helping students persist no longer seems to meaningfully impact persistence at university.

On the other hand, our analysis using non-grade measures of VA for the sophomore sample yields different insights. Indeed, results presented in Appendix Table A15 reveal that even after conditioning on having an advisor who is effective at raising students' grades, being assigned to an advisor who is good at improving persistence still significantly increases sophomore students' persistence in the major. Taken together, these results suggest that the main barrier for freshmen, i.e. pre-major, students' persistence is their first-year academic performance. However, students who already declared a major (i.e., sophomore students) may face other barriers to degree completion and require help from advisors who are skilled at improving both performance and persistence.

Finally, regarding the documented increase in STEM and business major enrollment for freshmen, findings from Table B.9 suggest that it is not due to behavioral changes in terms of shifting away or towards certain classes to fulfill course requirements for these majors. Additionally, analysis reported in column (6) (Panel B) of Table B.7 further suggests that this cannot be fully explained by grade improvements either. Rather, the most likely explanation for the documented increase in selective major enrollment is that it is driven by increased grade performance in addition to positive affirmation effects provided by freshman advisors.

2.5.3 Advisor Characteristics and Match Effects

We next examine whether advisors' observable characteristics predict their value-added. To do so, we regress all our constructed measures of advisor VA on advisor gender, rank and type of department. Results in Online Appendix Table A16 reveal no significant relationship between freshman advisors' faculty rank and their predicted grade or non-grade VA score. Specifically, being an associate or full professor as opposed to an assistant professor or lecturer does not predict a significantly higher or lower VA score, suggesting that faculty experience does not play a key role in predicting advisor quality. Additionally, we find that freshman advisor gender and department (i.e., whether the advisor is in a science versus non-science department) are also statistically unrelated to advisor VA. However, one caveat with these results is that they are based on regressions with a low number of observations—corresponding to the number of advisor-years in our freshman sample. For example, regressions involving the use of Grade VA are based on 131 advisor-year observations. Hence, results from Table A16 only provide suggestive evidence that advisors' observable characteristics are not related to VA. To strengthen conclusions from this analysis, we run the same regressions on a larger set of advisors, i.e. academic advisors matched to students from the sophomore sample. Importantly, results presented in Appendix Table A17 are all insignificant and in line with findings found in the freshman advisor sample, but more precise as they are based on 736 advisor-year observations. Overall, findings from both analyses are consistent with those from Barr and Castleman (2019) who show that counselor characteristics are not significantly related to student outcomes. This suggests that it is most likely unobservable characteristics, such as tone of voice, that may predict a large portion of what constitutes an effective advisor.

Results from the previous exercise indicate that advisors' observable characteristics, such as gender, do not predict advisor quality. Another interesting question is whether the match between advisor and student characteristics matters. Accordingly, we next check whether advisor-student gender match affects students' outcomes. To do so, we run the following reduced form regression:

$$Y_{iat} = \beta_0 + \beta_1 Femad_a + \beta_2 Femst_i + \beta_3 Femst_i * Femad_a + X_i'\gamma + \sigma_t + \epsilon_{iat} \tag{2.11}$$

where Y_{iat} is the outcome of interest for student i matched to advisor a in academic year t. $Femad_a$ is a dummy variable that is equal to 1 if advisor a is female and 0 otherwise. $Femst_i$ is another indicator variable for whether student i is female. We further interact both of these indicators. We also include student controls X_i' and year fixed effects σ_t throughout.[36] Our main coefficients of interest which we report in all our tables are β_1 (the effect of having a female versus male advisor for male students) and β_1+ β_3 (the effect of having a female versus male advisor for female students). Finally, standard errors are clustered at the advisor-year level throughout.

Online Appendix Table A18 summarizes the effects of student-advisor gender match for

[36] For regressions involving the sophomore sample, we also include department fixed effects since randomization occurs within department in that context.

male and female students from our main freshman sample. Specifically, estimates from row 1 show impacts on male students who are matched with a female as opposed to male freshman advisor and estimates from row 2 present coefficients from own gender match for female students (being matched with a female rather than male advisor). Results presented in column (1) of Table A18 indicate that own gender match matters for female students in terms of academic performance but not for male students. Indeed, female students' freshman GPA increases by 7.7 percent of a standard deviation when they are matched with a female versus male advisor. Conversely, the gender of a freshman advisor does not have a significant impact on male students' grades. We also find that advisor gender does not affect the likelihood that men or women drop out after freshman year but gender match does increase female students' 4-year graduation rates by 6.4 percentage points. However, we uncover no significant gender-match effects on 6-year graduation rates. In terms of major choice, we find that gender congruence has no significant impact on the likelihood that men or women enroll or graduate from a selective major. We also show that these effects are statistically insignificant for top-performing male and female students in columns (7) and (8).

Online Appendix Table A19 summarizes these same gender match effects for our post-major advising sample (i.e. the sample of students directly enrolled as sophomores). We find that student-advisor gender match matters for both sexes in terms of first-year academic performance. Indeed, results presented in column (1) of Table A19 indicate that male students' first year GPA increases by 4.9 percent of a standard deviation when matched with a male as opposed to female advisor. Additionally, female students' GPA increases by around 5.4 percent of a standard deviation when matched with a same gender advisor. However, we find no statistically significant effect of own-gender match on any of our student persistence outcomes as shown in columns (2) through (6). Overall, results from this section indicate that even though observable advisor characteristics do not predict advisor VA, student-advisor gender match does seem to matter for some student outcomes; mainly students' GPA.

2.6 Conclusion

In this paper, we study the impact of academic advisor VA on student outcomes. To identify causal effects, we exploit a unique setting where college students are randomly assigned to faculty advisors at the beginning of their freshman year. Students interact with their advisors for the full academic year. Advisors assist students with academic planning, monitor their academic progress, and help them decide on a major. We predict advisor value-added based on students' first-year course performance and show that advisors who raise students' grades also reduce freshman year completion time. These effects are long-lasting, as we show that a one standard deviation increase in freshman advisor grade VA raises 4-year graduation rates by 5.5%. Finally, we find that effective advisors have a strong impact on students' major choice. We document that exposure to higher-VA advisors largely increases high-performing students' chances of enrolling and graduating with a STEM degree.

Our finding that college students substantially benefit from high-quality personalized and continuous support has important implications for current debates on how to increase the rates of college completion and STEM degree attainment. In particular, our results indicate that allocating resources towards improving the quality of academic advising may substantially improve such outcomes. This in line with a recent study by Deming and Walters (2017) who find that higher U.S. state funding for public post-secondary institutions raises degree completion, through increased spending on academic support services such as advising.

Our paper presents new evidence showing that advisor quality is an important determinant of students' college success. However, what exactly constitutes a good advisor remains an open question. Our results indicate that observable characteristics—such as advisor rank, gender or department—do not correlate with advisor VA. Further research is needed to identify which advisor attributes increase their VA. Doing so would allow colleges to improve the quality of academic advising through screening for or training faculty to become effective advisors. Importantly, since most colleges already offer some form of academic advising, policies geared towards improving advisor quality may be a scalable way to promote student success.

Chapter 3

Clustering and External Validity in Randomized Controlled Trials

3.1 Introduction

In a randomized controlled trial (RCT), it is well known that one can estimate and draw inference on the average treatment effect, if the potential outcomes of units participating in the experiment are non-stochastic, a commonly-made assumption in the randomization inference literature (see, e.g., Neyman, 1923; Li and Ding, 2017; Abadie et al., 2020). In practice, it is often implausible that units' potential outcomes are fixed. For instance, an agricultural household's investment decisions may be affected by the weather conditions in its village during planting season, or by other stochastic shocks. In a model where units' potential outcomes are not fixed but depend on stochastic shocks, the results in the randomization inference literature still hold, conditional on the realizations of the shocks affecting units' potential outcomes. One can estimate and draw inference on the average treatment effect (ATE) *conditional on the shocks that occurred during the experiment.* This may not be a parameter of interest, as it lacks in external validity. For instance, when evaluating the effect of a cash grant on farmers' investment decisions, one may want to know the grant's effect independent of the specific shocks that arose during the experiment, rather than the grant's effect given those specific shocks.

To fix ideas, we describe our paper in the context of the farmers' cash grant RCT example, but our results apply to all experiments where shocks arising at a more aggregated level than the randomization unit can affect the outcome. We assume that the cash grant is randomly assigned to some households within each village. We relax the assumption of deterministic outcomes and allow household-level as well as village-level shocks to affect farmers' potential investment decisions without and with the grant. Finally, we define two estimands of interest: the ATE conditional on the village-level shocks and the ATE netted out of those shocks.

We start by showing that researchers can draw inference on the conditional ATE, by regressing farmers' investment on whether they received the cash grant, using the heteroskedasticity-robust variance estimator. This variance estimator is conservative for the variance of the ATE estimator conditional on the village-level shocks. On the other hand, to draw inference on the unconditional ATE, researchers need to cluster their standard errors at the village level. Indeed, we show that the village-clustered variance estimator is conservative for the unconditional variance of the ATE estimator. We also show that owing to the conservative nature of both variance estimators, the expectation of the heteroskedasticity-robust estimator may be larger than the expectation of the clustered one, when the treatment effect is more heterogeneous across farmers than across villages. In such cases, clustering may actually increase the ATE's t-statistic.

In a survey of The American Economic Journal: Applied Economics from 2014 to 2016, we found that only 1 out of the 26 published RCTs clustered their standard errors at a level higher than the unit-of-randomization. Therefore, our results provide an easy to implement and often overlooked solution for researchers to assess the external validity of their findings. By external validity, we mean whether results can be extrapolated beyond the specific circumstances that occurred during the experiment. Whether results can be extrapolated to a different population than the one that participated in the experiment is a different question.

To choose at which level to cluster, one first needs to think of which shocks are likely to arise during the experiment. Shocks are post-randomization events that affect the outcome. For example, in the context of the cash-grant RCT, weather events arising after the randomization are shocks. On the other hand, villages' demographic characteristics may affect the outcome, but they are pre-determined, so they are not shocks. In the context of a nationwide job-placement experiment, a post-randomization event affecting the labor market is a shock. Second, one needs to think of the level at which shocks operate. In the cash-grant RCT example, some weather shocks may arise at the village level and may be independent across villages, while other weather shocks may arise at a more aggregated level. In the job-placement experiment example, some labor market shocks arise at the local level (e.g.: a plant closure), while others arise at the national level (e.g.: a change in the Central Bank's interest rate). There could also be some industry-specific shocks, and other shocks affecting all industries, so shocks need not operate at a geographic level. Finally, one needs to cluster at a level where many shocks are likely to operate, while still having sufficiently many clusters to draw valid inference. In the job placement experiment, clustering at a local (e.g. city or regional) level will account for all the shocks taking place at that level, but it will not account for macro-level shocks.[1] Clustering can only account for shocks arising at a more disaggregated level than the

[1] In that example, one may want to account both for local- and industry-level shocks. We conjecture that doing so may be feasible using a multi-way clustering method (see Cameron et al., 2012; Menzel, 2018; Davezies et al., 2019), but showing it goes beyond the scope of this paper.

level at which the experiment took place.[2] The above exercise is not a mechanical one: it leads to concrete, context-specific, recommendations on the level at which one should cluster. Importantly, to avoid specification searching, the level of clustering should be pre-specified.

Our paper shows that the model-based (Cameron and Miller, 2015; Wooldridge, 2003) and design-based (Murray et al., 1998; Donner and Klar, 2000; Abadie et al., 2017) approaches to clustering are not incompatible, and may be fruitfully combined. In RCTs without clustering in the treatment assignment and where the experimental units are not sampled from a larger population, Abadie et al. (2017) have argued that clustering standard errors is not needed. Our results lead to a different recommendation: we consider the very same RCTs (with individual-level treatment assignment, and in a finite population of units that are not sampled from a larger population), and argue that if there are cluster-level shocks affecting the potential outcomes, one may want to cluster if one wants to draw inference on the average treatment effect netted out of the shocks. This difference arises because in Abadie et al. (2017), assignment to treatment is the only source of randomness when experimental units are not drawn from a larger population. In contrast, our setup allows for another source of randomness, the cluster-level shocks, that are not under the investigator's control, and that may alter the outcome. There are a number of contexts where such shocks are likely to arise, and we now review two other recent papers that have documented their existence and proposed methods to take them into account.

Rosenzweig and Udry (2019) have also shown that with aggregate shocks, heteroskedasticity-robust variance estimators may understate the true variance of the ATE estimator in an RCT. As one of their applications, they use an RCT in Ghana conducted from 2009 to

[2] Hahn et al. (2020) propose a framework for dealing with macro shocks in the context of structural MLE models, a setting that differs from ours.

2011 where farmers were given a rainfall insurance and a cash grant. They use those treatments as an instrument for agricultural investment, and they show that returns to investment vary with rainfalls. Using their estimated coefficient for the interaction of investment and rainfalls, they compute the distribution of returns to investment under the rainfall distribution observed over the last 65 years in Ghana. They find that the resulting distribution has a much larger variance than the sampling variance from the experiment would suggest. The solution they propose to account for aggregate shocks differs from ours. First, it requires using additional data (e.g. the distribution of rainfalls in the Ghana example) while ours does not. Second, it is designed to account for specific observable shocks (e.g. rainfall shocks in the Ghana example) while ours can account for any type of cluster-level shock, including unobserved ones. Finally, their method can be used to extrapolate the distribution of the treatment effect under a different distribution of shocks than that observed during the experiment. This extrapolation can be made under the assumption that the shocks interact multiplicatively with the treatment effect. The clustering method we propose does not rely on this assumption; accordingly, it can tell us if there is evidence that the cluster-level shocks that arose during the experiment affected the impact of the intervention, but it cannot tell us anything about the intervention's impact under different shocks. Let us illustrate this important difference through an example. Assume an agricultural experiment took place in a rainy year, with some variation in rainfall across regions, but no drought in any region. Clustering at the regional level, the researcher can test if it is still possible to reject the null of no effect, accounting for the variability in the treatment effect induced by rainfall variations from moderate to high. But clustering cannot tell us whether the treatment would have had an effect during a drought year. The method proposed by Rosenzweig and Udry (2019) can achieve that, under some assumptions.

Riddell and Riddell (2020) have also highlighted an issue similar to that we discuss here. By revisiting the results of the Self-Sufficiency Project, they find that post-randomization events can threaten the validity of experimental designs. They give the following example. In a randomized trial of a chemotherapy treatment conducted at one site only, if an outbreak of C-difficile occurs during treatment, treatment group members will be more likely to die from the outbreak than the control group members due to a weakened immune system. Then without more sites, we can only draw inference on the treatment effect conditional on the occurrence of a C-difficile outbreak, which is not necessarily the parameter of interest. Riddell and Riddell (2020) mention that multiple sites may help researchers to interpret experimental evidence because different post-randomization events may occur in different sites. Our results support that statement, and show that by clustering at the level at which these post-randomization events take place, one can draw inference on the ATE net of these events.

Finally, many other papers have departed from the randomization inference literature, and have allowed potential outcomes to be stochastic in RCTs (see, e.g. Bugni et al., 2018, 2019). However, those papers usually assume that potential outcomes are i.i.d. Instead, we consider the case where units' potential outcomes are correlated due to cluster-level shocks.

We use our results to revisit Karlan et al. (2014), who study the effects of a rainfall insurance and of a cash grant treatment on farmers' investment decisions. That paper was also revisited by Rosenzweig and Udry (2019), who argue that in this context, regional-level weather shocks need to be accounted for. To do so, we cluster standard errors at the regional level at which Rosenzweig and Udry (2019) argue that weather shocks occur. Doing so, we do not find very different results from those Karlan et al. (2014) had obtained using heteroskedasticity-robust standard errors, thus showing that their results are robust to accounting for the aggregate shocks that arose during the experiment.

We also revisit Cole et al. (2013), who study the effects of various treatments on farmers' adoption of a rainfall insurance. Using heteroskedasticity-robust standard errors, the authors found that two of their treatments significantly increased adoption. This experiment took place in 37 villages of two districts of the state of Andra Pradesh in India, so the most aggregated level we can cluster at is the village one. Even clustering at this fairly disaggregated level, we find that only one of the two treatments still has a significant effect on adoption. The effect of the second treatment may have been due to the specific village-level shocks that arose during the experiment, and may not replicate under different circumstances.

The take-aways of our paper for applied researchers are as follows. When one rejects the null of no effect without clustering but not with clustering, one can assert that the treatment had an effect, given the specific shocks that arose during the experiment, though this conclusion may not generalize under different shocks. On the other hand, when one rejects the null of no effect with clustering, one can assert that the treatment had an effect, independent of the specific shocks that arose during the experiment. Then, the decision to cluster or not depends on the level of external validity one would like to achieve.

3.2 Setup and finite-sample results

3.2.1 Setup and Notation

We consider an RCT taking place in a finite population of K villages. Village k has n_k households, and randomization is stratified at the village level. The experiment wants to look at the effect of cash grants on farming households' investment in agriculture. The outcomes of interest are households' investments in agriculture such as land preparation costs, value of chemicals used, and acres cultivated. It is arguably implausible to assume that households' potential outcomes are fixed, they may be affected by a wealth of stochastic events that could take place until the time they make their investment decisions. These shocks could be specific to the households (such as the breadwinner being laid off or injured), or they could be common to all households within a village (such as extreme weather events, economic hardships in the village etc.). We therefore assume that for all $(i,k) \in \{1,...,n_k\} \times \{1,...,K\}$, the potential outcomes of household i in village k without and with the treatment, $Y_{ik}(0)$ and $Y_{ik}(1)$ satisfy the following equations:

Assumption 6 *Stochastic Potential Outcomes*

$$\begin{aligned} Y_{ik}(0) &= y_{ik}(0) + \eta_k(0) + \epsilon_{ik}(0) \\ Y_{ik}(1) &= y_{ik}(1) + \eta_k(1) + \epsilon_{ik}(1). \end{aligned} \tag{3.1}$$

$\epsilon_{ik}(0)$ (resp. $\epsilon_{ik}(1)$) represents a shock affecting the potential outcomes of household i in village k if she is untreated (resp. treated). $\eta_k(0)$ (resp. $\eta_k(1)$) represents a shock affecting all the untreated (resp. treated) households in village k. We assume that $\mathbb{E}(\epsilon_{ik}(0)) = \mathbb{E}(\epsilon_{ik}(1)) = \mathbb{E}(\eta_k(0)) = \mathbb{E}(\eta_k(1)) = 0$, so $y_{ik}(0)$ (resp. $y_{ik}(0)$) represent the expectation of $Y_{ik}(0)$ (resp. $Y_{ik}(1)$), the outcome without (resp. with) treatment that household i in village k will obtain under "average" household- and village-level shocks. In our cash-grant example, $y_{ik}(d)$ is a household's investment under average shocks and treatment d. $\epsilon_{ik}(d)$ represents the effect of household-level shocks, such the breadwinner being laid off, on the household's investment under treatment d. $\eta_k(d)$ represents the effect of village level shocks, such as an extreme weather event, on the household's investment under treatment d. Let $(\boldsymbol{\eta(0)}, \boldsymbol{\eta(1)}) = (\eta_k(0), \eta_k(1))_{1\leq k\leq K}$ be a vector stacking all the village-level shocks, and let $(\boldsymbol{\epsilon(0)}, \boldsymbol{\epsilon(1)}) = (\epsilon_{ik}(0), \epsilon_{ik}(1))_{1\leq i\leq n_k, 1\leq k\leq K}$ be a vector stacking all the household-level shocks.

Assumption 6 requires that the shocks be additively separable, and take place at the level of the experimental strata. These two conditions are not of essence for our results to hold. Our main results still hold if shocks take place at a more aggregated level than the experimental strata. Our main results also still hold if the shocks do not affect the potential outcomes in an additively separable manner, i.e. if $Y_{ik}(d) = f_{ikd}\left(\epsilon_{ik}(d), \eta_k(d)\right)$ for some functions $f_{ikd}(.)$. In that case, one just needs to redefine $ATE(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))$ below as $\frac{1}{n}\sum_{i,k} E(Y_{ik}(1) - Y_{ik}(0)|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)))$, and ATE as $\frac{1}{n}\sum_{i,k} E(Y_{ik}(1) - Y_{ik}(0))$, see Section C.1.7 of the Appendix for more details. We expect most readers to be familiar with the additively separable model, so we stick to it in the paper to facilitate reading.

Let $n = \sum_{k=1}^{K} n_k$ denote the total number of households in the K villages. We may be more interested in learning

$$ATE = \frac{1}{n}\sum_{i,k}[y_{ik}(1) - y_{ik}(0)], \tag{3.2}$$

rather than

$$ATE(\boldsymbol{\epsilon}(0), \boldsymbol{\epsilon}(1), \boldsymbol{\eta(0)}, \boldsymbol{\eta(1)}) = \frac{1}{n}\sum_{i,k}[Y_{ik}(1) - Y_{ik}(0)], \tag{3.3}$$

or

$$ATE(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)) = \frac{1}{n} \sum_{i,k} \left[(y_{ik}(1) + \eta_k(1)) - (y_{ik}(0) + \eta_k(0)) \right]. \tag{3.4}$$

The second parameter is the average effect of the treatment on households' investments, conditional on the specific village and household shocks that arose during the experiment. The third parameter is the average effect of the treatment on households' investments, conditional only on the specific village shocks. The first parameter is the average effect of the treatment, net of those specific shocks. This parameter is more externally valid than the other two, as it applies beyond the specific circumstances that occurred during the experiment.[3]

Let D_{ik} be an indicator for whether household i in village k is treated, let $\mathbf{D}_k$ be a vector stacking the treatment indicators of all households in village k, and let $\mathbf{D}$ be a matrix stacking these vectors. We consider the following assumption:

Assumption 7 *For all i, k,*

1. *$V(\epsilon_{ik}(d)) = \sigma^2_{dik} < +\infty$ for $d = 0, 1$.*

2. *For all $j \neq i$, $(\epsilon_{ik}(0), \epsilon_{ik}(1)) \perp\!\!\!\perp (\epsilon_{jk}(0), \epsilon_{jk}(1))$.*

3. *$\boldsymbol{D} \perp\!\!\!\perp \left((\epsilon_{ik}(0), \epsilon_{ik}(1))_{1 \leq i \leq n_k}, \eta_k(0), \eta_k(1) \right)_{1 \leq k \leq K}$.*

4. *$(\epsilon_{ik}(0), \epsilon_{ik}(1))_{1 \leq i \leq n_k} \perp\!\!\!\perp (\eta_k(0), \eta_k(1))_{1 \leq k \leq K}$.*

5. *$V(\eta_k(1) - \eta_k(0)) < +\infty$.*

[3] On the other hand, ATE still only applies to the villages participating in the experiment. Recent articles that consider treatment effects extrapolation outside of the estimation sample include, e.g., Dehejia et al. (2019) or Bo and Galiani (2019).

Point 1 requires that $(\epsilon_{ik}(0), \epsilon_{ik}(1))$ have a second moment. Point 2 requires that in each village, the household level shocks be independent. Point 3 requires that the household- and village-level shocks be independent of the treatments, which usually holds by design in a RCT. Point 4 requires that the household- and village-level shocks be independent. Finally, Point 5 requires that the variance of $\eta_k(1) - \eta_k(0)$ exist. Assumption 7 does not require that the shocks $(\epsilon_{ik}(0), \epsilon_{ik}(1))$ and $(\eta_k(0), \eta_k(1))$ be identically distributed: the variance of the shocks may for instance vary across households or villages. Assumption 7 also does not require that $\epsilon_{ik}(0)$ and $\epsilon_{ik}(1)$ be independent, or that $\eta_k(0)$ and $\eta_k(1)$ be independent: one may for instance have $\epsilon_{ik}(0) = \epsilon_{ik}(1)$ and $\eta_k(0) = \eta_k(1)$, if the household- and village-level shocks are the same when treated and untreated.

Let n_{1k} and n_{0k} respectively denote the number of households in the treatment and control groups in village k. Let $Y_{ik} = D_{ik}Y_{ik}(1) + (1 - D_{ik})Y_{ik}(0)$ denote the observed outcome of household i. For any variable x_{ik} defined for every $i \in \{1, ..., n_k\}$ and $k \in \{1, ..., K\}$, let $\overline{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik}$ denote the average value of x_{ik} in village k, let $\overline{x}_{1k} = \frac{1}{n_{1k}}\sum_{i=1}^{n_{1k}} D_{ik}x_{ik}$ and $\overline{x}_{0k} = \frac{1}{n_{0k}}\sum_{i=1}^{n_{0k}}(1 - D_{ik})x_{ik}$ respectively denote the average value of x_{ik} among the treated and untreated households in village k, and let $\overline{x} = \frac{1}{n}\sum_{i,k} x_{ik}$ denote the average value of x_{ik} across all households.

Then let $\overline{n} = \frac{n}{K}$, and let

$$\begin{aligned}\widehat{ATE}_k &= \overline{Y}_{1k} - \overline{Y}_{0k} \\ \widehat{ATE} &= \frac{1}{K}\sum_{k=1}^{K}\frac{n_k}{\overline{n}}\widehat{ATE}_k,\end{aligned}$$

respectively denote the standard difference in means estimator of the average treatment effect in village k, and the estimated average treatment effect in the K villages.

For any variable x_{ik} defined for every $i \in \{1, ..., n_k\}$ and $k \in \{1, ..., K\}$, let $S^2_{x,k} =$

$\frac{1}{n_k-1}\sum_{i=1}^{n_k}(x_{ik}-\overline{x}_k)^2$ denote the variance of x_{ik} in village k, and let $S^2_{x,1,k}=\frac{1}{n_{1k}-1}\sum_{i=1}^{n_{1k}}D_{ik}(x_{ik}-\overline{x}_{1k})^2$ and $S^2_{x,0,k}=\frac{1}{n_{0k}-1}\sum_{i=1}^{n_{0k}}(1-D_{ik})(x_{ik}-\overline{x}_{0k})^2$ respectively denote the variance of x_{ik} among the treated and untreated households in village k. Then let,

$$\widehat{V}_{rob}\left(\widehat{ATE}_k\right) = \frac{1}{n_{1k}}S^2_{Y,1,k}+\frac{1}{n_{0k}}S^2_{Y,0,k}$$

denote the robust estimator of the variance of $\widehat{ATE}_k$ (Eicker et al., 1963; Huber et al., 1967; White et al., 1980), and let

$$\widehat{V}_{rob}\left(\widehat{ATE}\right) = \frac{1}{K^2}\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2\widehat{V}_{rob}\left(\widehat{ATE}_k\right),$$

denote the estimator of the variance of $\widehat{ATE}$ one can form using those estimators and assuming the $\widehat{ATE}_k$s are independent.

We assume that the treatment is randomly assigned at the household level in each village:

Assumption 8 *Stratified completely randomized experiment*
For all k, $\sum_{i=1}^{n_k}D_{ik}=n_{1k}$, and for every $(d_1,...,d_{n_k})$ such that $d_1+...+d_{n_k}=n_{1k}$, $P(\boldsymbol{D}_k=(d_1,...,d_{n_k}))=\frac{1}{\binom{n_k}{n_{1k}}}$.

Finally, we make the following assumption:

Assumption 9 *The vectors $(\boldsymbol{D}_k,\eta_k(1),\eta_k(0),(\epsilon_{ik}(0),\epsilon_{ik}(1))_{1\leq i\leq n_k})$ are mutually independent.*

Assumption 9 requires that the variables attached to different villages be mutually independent.

3.2.2 Finite-sample results

We can now state our first result.

Theorem 4 *If Assumptions 6, 7, 8, and 9 hold,*

1. $\mathbb{E}\left(\widehat{ATE}\middle|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right) = ATE(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))$

2. $V\left(\widehat{ATE}\middle|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right) = \dfrac{1}{K^2}\sum_{k=1}^{K}\left(\dfrac{n_k}{\overline{n}}\right)^2 V\left(\widehat{ATE}_k\middle|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right)$, *where*

$$V\left(\widehat{ATE}_k\middle|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right) = \tfrac{1}{n_{0k}}S^2_{y(0),k} + \tfrac{1}{n_{1k}}S^2_{y(1),k} - \tfrac{1}{n_k}S^2_{y(1)-y(0),k} + \tfrac{1}{n_{1k}}\overline{\sigma^2_{1k}} + \tfrac{1}{n_{0k}}\overline{\sigma^2_{0k}}$$

3. $V\left(\widehat{ATE}\middle|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right) \leq \mathbb{E}\left(\widehat{V}_{rob}(\widehat{ATE})\middle|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right)$, *with equality if there is no treatment effect heterogeneity within village:* $S^2_{y(1)-y(0),k} = 0$ *for all* k.

Point 1 of Theorem 4 shows that $\widehat{ATE}$ is an unbiased estimator of $ATE(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))$, conditional on the village-level shocks. Point 2 gives a formula for the variance of $\widehat{ATE}$ conditional on the village-level shocks. It is similar to the variance of $\widehat{ATE}$ in Neyman (1923), derived assuming fixed potential outcomes. However, it contains one more term, $\frac{1}{K^2}\sum_{k=1}^{K}\frac{1}{n_{1k}}\overline{\sigma^2_{1k}} + \frac{1}{n_{0k}}\overline{\sigma^2_{0k}}$, which comes from the added variation created by the individual-level shocks. Point 3 shows that the robust variance estimator is a conservative estimator of that conditional variance.

In our set-up, the result in Neyman (1923) implies that conditional on the household- and village-level shocks, $\widehat{ATE}$ is an unbiased estimator, and the robust variance estimator is a conservative estimator of the variance of $\widehat{ATE}$. Theorem 4 extends this result, by showing that it still holds when one only conditions on the village-level shocks.[4]

[4] It has also been shown (see Imbens and Rubin, 2015) that when the potential outcomes are i.i.d., the

In our cash-grant example, Theorem 4 implies that if the researcher uses robust standard errors and finds a statistically significant effect, she can conclude that $ATE(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)) \neq 0$: the treatment had an effect, given the specific village-level shocks that arose during the experiment. $ATE(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))$ does not depend on the household-level shocks that arose during the experiment, but it does depend on the village-level shocks. Therefore, the researcher cannot say whether the treatment would still have had an effect if different village-level shocks had occurred. To answer that question, one needs to draw inference on ATE. We now show that this can be achieved, by clustering standard errors at the village level. Let

$$\widehat{V}_{clu}(\widehat{ATE}) = \frac{1}{K\,(K-1)} \sum_{k=1}^{K} \left(\frac{n_k}{\overline{n}} \widehat{ATE}_k - \widehat{ATE} \right)^2$$

be the cluster-robust estimator of the variance of $\widehat{ATE}$ (Liang and Zeger, 1986).[5]

Theorem 5 *If Assumptions 6, 7, 8, and 9 hold,*

1. $\mathbb{E}\left(\widehat{ATE}\right) = ATE$
2. $V\left(\widehat{ATE}\right) = \frac{1}{K^2} \sum_{k=1}^{K} \left(\frac{n_k}{\overline{n}}\right)^2 \left[\frac{1}{n_{0k}} S^2_{y(0),k} + \frac{1}{n_{1k}} S^2_{y(1),k} - \frac{1}{n_k} S^2_{y(1)-y(0),k} + \frac{1}{n_{1k}} \overline{\sigma^2_{1k}} + \frac{1}{n_{0k}} \overline{\sigma^2_{0k}} + V\left(\eta_k(1) - \eta_k(0)\right) \right].$
3. $\mathbb{E}\left(V\left(\widehat{ATE}\middle|\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)\right)\right) \leq V\left(\widehat{ATE}\right) \leq \mathbb{E}\left[\widehat{V}_{clu}(\widehat{ATE})\right].$

The second inequality is an equality if $n_k = \overline{n}$ and $ATE_k = ATE$,

robust variance estimator is an unbiased estimator of $V(\widehat{ATE})$. This result can also be obtained from Theorem 4. Assume that $\eta_k(0) = \eta_k(1) = 0$, thus ensuring that the potential outcomes are independent, and that $y_{ik}(d) = y(d)$ and $\sigma^2_{dik} = \sigma^2_d$, thus ensuring that they are identically distributed. Then, Point 3 implies that $V\left(\widehat{ATE}\right) = \mathbb{E}\left(\widehat{V}_{rob}\left(\widehat{ATE}\right)\right)$.

[5] When all strata have the same number of units, $\widehat{V}_{clu}(\widehat{ATE})$ is equal to the cluster-robust estimator of the variance of the treatment coefficient in a regression of the outcome on a constant and the treatment clustered at the strata level, up to a degrees of freedom adjustment.

with $ATE_k = \frac{1}{n_k}\sum_{i=1}^{n_k}[y_{ik}(1) - y_{ik}(0)]$ *for all* $k = 1, ..., K$.

Point 1 of Theorem 5 shows that $\widehat{ATE}$ is an unbiased estimator of ATE. Point 2 gives a formula for the unconditional variance of $\widehat{ATE}$. Point 3 shows that the cluster-robust variance estimator is a conservative estimator of the unconditional and conditional variances of $\widehat{ATE}$.

In our cash-grant example, Theorem 5 states that if the researcher uses the cluster-robust standard errors and finds a statistically significant effect, she can conclude that $ATE \neq 0$. ATE does not depend on the household- and village-level shocks that arose during the experiment. Therefore, this conclusion is not dependent on the specific shocks that arose during the experiment, but holds when the shocks are averaged out.

Our approach comes with a risk. By defining several potential estimands of interest, it may lead researchers to test several null hypothesis, with or without clustering, or clustering at various different levels. This would distort inference. To avoid that risk, researchers should pre-commit to an analysis plan that specifies if and at what level they intend to cluster standard errors.

Corollary 2 *If Assumptions 6, 7, 8, and 9 hold and* $n_k = \overline{n}$ *for all* k,

$$\begin{aligned}
&\mathbb{E}\left[K\widehat{V}_{clu}(\widehat{ATE})\right] - \mathbb{E}\left[K\widehat{V}_{rob}(\widehat{ATE})\right] \\
=&\frac{1}{K}\sum_{k=1}^{K}V\left(\eta_k(1) - \eta_k(0)\right) + \frac{1}{K-1}\sum_{k=1}^{K}\left(\mathbb{E}\left(\widehat{ATE}_k\right) - \frac{1}{K}\sum_{k'=1}^{K}\mathbb{E}\left(\widehat{ATE}_{k'}\right)\right)^2 \\
-&\frac{1}{\overline{n}}\frac{1}{K}\sum_{k=1}^{K}S^2_{y(1)-y(0),k}
\end{aligned}$$

Corollary 2 states that the difference between the expectations of the normalized clustered and robust variance estimators is equal to the average of $V(\eta_k(1) - \eta_k(0))$ plus the difference between the variance of the treatment effect between villages and the average variance of the treatment effect within villages divided by $\overline{n}$. It has two important implications. First, if households' potential outcomes are identically distributed, then $y_{ik}(1) - y_{ik}(0) = \tau$ for all (i, k), and $\mathbb{E}\left[K\widehat{V}_{clu}(\widehat{ATE})\right] - \mathbb{E}\left[K\widehat{V}_{rob}(\widehat{ATE})\right] = \frac{1}{K}\sum_{k=1}^{K} V\left(\eta_k(1) - \eta_k(0)\right)$. Therefore, one can test whether there are village-level shocks that affect the impact of the intervention by testing whether the two variance estimators significantly differ.

Second, consider the following assumption:

Assumption 10 *Homogeneous clustered shocks*
For all k, $\eta_k(1) = \eta_k(0)$.

Assumption 10 requires that treated and untreated households are affected similarly by the village-level shocks.[6] Under Assumptions 6 and 10, $ATE(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)) = ATE$ and $V\left(\widehat{ATE}\right) = V\left(\widehat{ATE}|\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)\right)$, so Theorems 4 and 5 imply that both $\widehat{V}_{rob}(\widehat{ATE})$ and $\widehat{V}_{clu}(\widehat{ATE})$ are conservative for $V\left(\widehat{ATE}\right)$. Corollary 2 shows that $\widehat{V}_{clu}(\widehat{ATE})$ can be less conservative than $\widehat{V}_{rob}(\widehat{ATE})$, if there is more treatment effect heterogeneity within rather than between villages. When Assumption 10 fails, Theorems 4 and 5 imply that both $\widehat{V}_{rob}(\widehat{ATE})$ and $\widehat{V}_{clu}(\widehat{ATE})$ are conservative for $\mathbb{E}\left(V\left(\widehat{ATE}\middle|\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)\right)\right)$, and Corollary 2 shows that $\widehat{V}_{clu}(\widehat{ATE})$ can be less conservative than $\widehat{V}_{rob}(\widehat{ATE})$ if the additional variance in $\widehat{ATE}$ coming from the village-level shocks is lower than the difference between the within-village variance of the treatment effect divided by $\overline{n}$ and the between-village variance of the treatment effect.

[6] It is not testable without imposing other assumptions. Under the assumption that ATE_k does not vary across k, one can test whether the $\widehat{ATE}_k$s significantly differ. If they do, that implies that $\eta_k(1) \neq \eta_k(0)$ for some k.

3.3 Large-sample results

We now derive the asymptotic distribution of $\widehat{ATE}$ considering a case where the number of villages K goes to infinity. First let:

$$AD_k = \frac{n_k}{\bar{n}}\widehat{ATE}_k,$$

and consider the following assumption:

Assumption 11 *Regularity conditions to derive the asymptotic distribution of* $\widehat{ATE}$

For some $\epsilon > 0$,

1. *For every* k, $\mathbb{E}\left(AD_k^{2+\epsilon}\right) \leq M < +\infty$, *for some* $M > 0$.
2. $\lim\limits_{K\to+\infty} \frac{1}{S_K^{2+\epsilon}} \sum_{k=1}^{K} \mathbb{E}\left[|AD_k - E(AD_k)|^{2+\epsilon}\right] = 0$, *where* $S_K^2 = \sum_{k=1}^{K} V(AD_k)$.
3. $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k\right)$, $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k^2\right)$, *and* $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k\right)^2$ *converge towards finite limits when* $K \to \infty$.

Assumption 11 contains the regularity conditions needed to apply the strong law of large numbers in Lemma 1 of Liu et al. (1988) and the Lyapunov CLT. Also let:

$$\begin{aligned}
\sigma^2 &= \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k^2\right) - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k\right)^2, \\
\sigma_+^2 &= \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k^2\right) - \left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k\right)\right)^2.
\end{aligned}$$

We show:

Theorem 6 *If Assumptions 6, 7, 8, 9, and 11 hold,*

$$
\begin{aligned}
&1. \quad \sqrt{K}\left(\widehat{ATE} - ATE\right) \xrightarrow{d} N\left(0, \sigma^2\right). \\
&2. \quad K\widehat{V}_{clu}\left(\widehat{ATE}\right) \xrightarrow{p} \sigma_+^2 \geq \sigma^2.
\end{aligned}
$$

Point 1 of Theorem 6 shows that $\widehat{ATE}$ is an asymptotically normal estimator of ATE when the number of villages goes to infinity. Point 2 shows that $K\widehat{V}_{clu}\left(\widehat{ATE}\right)$ converges to a finite upper bound of the asymptotic variance of $\widehat{ATE}$ and can be used to construct conservative confidence intervals for ATE.

3.4 Applications

3.4.1 Agricultural Decisions After Relaxing Credit And Risk Constraints

Table 4 in Karlan et al. (2014) presents the effects of having rainfall index insurance, receiving a capital grant, and having both treatments on investment decisions and value of harvest. The results are obtained using the first two years of a three-year RCT conducted in Ghana. In the first year, the authors randomly assigned households to one of four groups: the cash grant group, the insurance group, the cash grant and insurance group, and the control group. In the second year, the cash grant experiment was still present but the insurance grant experiment was replaced by an insurance pricing experiment. Insurance prices were randomized at the community level but every community also had control households without access to the insurance with the randomization being at the household level. For farmers that were offered insurance, insurance take-up is instrumented using the price offered to them (see Karlan et al., 2014).

In a re-analysis of this experiment, Rosenzweig and Udry (2019) divide communities into 11 regions, and show that returns to farmers' investments respond to the weather shocks affecting their region. In Table 3.1 below, Panel A replicates the results in Karlan et al. (2014), using heteroskedasticity-robust variance estimators. In Panel B, we instead use cluster-robust variance estimators, clustering at the level of the 11 regions indicated by Rosenzweig and Udry (2019). As there are only 11 clusters, in Panel C we present p-values computed using the wild-bootstrap test proposed in Cameron et al. (2008), and that has been shown to have good properties with a small number of large clusters, see Canay et al. (2019). Clustering at the region level does not strongly affect the results in Karlan et al. (2014). The only exception is for the outcome "value of chemicals used", for which treatment effects are less significant with the wild cluster bootstrap than with robust standard errors. Otherwise, for most of the outcomes for which we can reject $ATE(\eta(0), \eta(1)) = 0$, we can also reject $ATE = 0$.

Table 3.1: Effects in Karlan et al. (2014), without and with clustering.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Land	# of Acres	Value of	Wages Paid	Opportunity Cost	Total	Value of
	Preparation Costs	Cultivated	Chemicals Used	to Hired Labor	of Family Labor	Costs	Harvest
A: Robust SE							
Insured	25.528**	1.024**	37.904**	83.537	98.161	266.146**	104.274
	(12.064)	(0.420)	(14.854)	(59.623)	(84.349)	(134.229)	(81.198)
	[0.034]	[0.015]	[0.011]	[0.161]	[0.245]	[0.047]	[0.199]
Insured*Capital Grant	15.767	0.257	66.440***	39.760	-52.653	72.137	129.243
	(13.040)	(0.445)	(15.674)	(65.040)	(86.100)	(138.640)	(81.389)
	[0.227]	[0.563]	[0.000]	[0.541]	[0.541]	[0.603]	[0.112]
Capital Grant	15.362	0.088	55.631***	75.609	-130.562	2.438	64.822
	(13.361)	(0.480)	(17.274)	(68.914)	(92.217)	(148.553)	(89.764)
	[0.250]	[0.854]	[0.001]	[0.273]	[0.157]	[0.987]	[0.470]
B: Clustered SE							
Insured	25.528**	1.024***	37.904**	83.537	98.161	266.146***	104.274*
	(12.498)	(0.372)	(17.784)	(52.591)	(67.068)	(97.865)	(60.776)
	[0.041]	[0.006]	[0.033]	[0.112]	[0.143]	[0.007]	[0.086]
Insured*Capital Grant	15.767	0.257	66.440***	39.760	-52.653	72.137	129.243*
	(14.307)	(0.266)	(12.018)	(53.571)	(56.916)	(94.551)	(74.076)
	[0.270]	[0.332]	[0.000]	[0.458]	[0.355]	[0.445]	[0.081]
Capital Grant	15.362	0.088	55.631**	75.609	-130.562	2.438	64.822
	(15.092)	(0.504)	(25.531)	(50.493)	(114.161)	(185.320)	(122.354)
	[0.309]	[0.861]	[0.029]	[0.134]	[0.253]	[0.990]	[0.596]
C: Wild Bootstrap							
Insured	25.528*	1.024**	37.904	83.537	98.161	266.146*	104.274
	[0.069]	[0.032]	[0.141]	[0.103]	[0.299]	[0.060]	[0.106]
Insured*Capital Grant	15.767	0.257	66.440**	39.760	-52.653	72.137	129.243
	[0.336]	[0.388]	[0.022]	[0.507]	[0.393]	[0.480]	[0.115]
Capital Grant	15.362	0.088	55.631	75.609	-130.562	2.438	64.822
	[0.392]	[0.882]	[0.166]	[0.162]	[0.305]	[0.993]	[0.638]
N	2,320	2,320	2,320	2,320	2,320	2,320	2,320

Standard errors in parentheses, p-values in brackets. The results in this table are based on Table 4 from Karlan et al. (2014), in year 2 Insured is instrumented using a full set of prices. Total costs (column (6)) includes sum of chemicals, land preparatory costs (e.g., equipment rental but not labor), hired labor, and family labor (valued at gender/community/year-specific wages). Harvest value includes own-produced consumption, valued at community-specific market value. All specifications include controls for full set of sample frame and year interactions. *** p <0.01 ** p <0.05 * p <0.1.

The results of Table 3.1 are based on 2SLS regressions, while our theoretical results cover OLS ones. To alleviate this concern, we report results from reduced form regressions with only one of the instruments used by Karlan et al. (2014), the binary variable ”offered a capital grant and insurance at price 0”. We also include a full set of sample frame and year interactions as controls, as in their 2SLS regression. Results are shown in Table 3.2. Again, clustering at the region level does not change the significance of the “intention-to-treat” effects of that instrument.

Table 3.2: Reduced Form Effects in Karlan et al. (2014), without and with clustering.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Land	# of Acres	Value of	Wages Paid	Opportunity Cost	Total	Value of
	Preparation Costs	Cultivated	Chemicals Used	to Hired Labor	of Family Labor	Costs	Harvest
A: Robust SE							
K-Grant	42.186**	1.275*	110.629***	115.089	-46.286	214.860	165.269
and free insurance offered	(19.125)	(0.658)	(24.791)	(93.025)	(114.330)	(186.694)	(122.767)
	[0.027]	[0.053]	[0.000]	[0.216]	[0.686]	[0.250]	[0.178]
B: Clustered SE							
K-Grant	42.186*	1.275*	110.629***	115.089	-46.286	214.860	165.269
and free insurance offered	(20.795)	(0.584)	(23.964)	(95.827)	(110.267)	(192.092)	(150.403)
	[0.070]	[0.054]	[0.001]	[0.257]	[0.684]	[0.289]	[0.298]
C: Wild Bootstrap							
K-Grant	42.186*	1.275***	110.629***	115.089	-46.286	214.860	165.269
and free insurance offered	[0.051]	[0.006]	[0.004]	[0.268]	[0.630]	[0.384]	[0.300]
N	2,320	2,320	2,320	2,320	2,320	2,320	2,320

Standard errors in parentheses, p-values in brackets. The results in this table are based on a reduced form regression inspired by Table 4 from Karlan et al. (2014). They consist of a regression of the outcome on a dummy variable for being offered a capital grant and insurance at price 0. Total costs (column (6)) includes sum of chemicals, land preparatory costs (e.g., equipment rental but not labor), hired labor, and family labor (valued at gender/community/year-specific wages). Harvest value includes own-produced consumption, valued at community-specific market value. All specifications include controls for full set of sample frame and year interactions. *** p <0.01 ** p <0.05 * p <0.1.

We compare how our clustering method performs relative to the method proposed by

Rosenzweig and Udry (2019) to account for weather shocks. They use their method to estimate the net returns of planting-stage investments in the Ghana experiment, using the RCT treatments as instruments for investment in a 2SLS regression. Note that our results above apply to OLS regression coefficients, but we will momentarily assume they also apply to 2SLS ones, to be able to draw a comparison with the results in Rosenzweig and Udry (2019). Table 3.3 below compares three confidence intervals. The first uses the normal approximation, the estimate of returns to planting-stage investment in Table 3 Column 2 of Rosenzweig and Udry (2019), and standard errors clustered at the regional level. As there are only 11 regions, the second confidence interval uses the wild-bootstrap, clustering at the region level. The third confidence interval uses the 2.5 and 97.5 percentiles from the distribution of returns to planting stage investment in Figure 5 of Rosenzweig and Udry (2019). The confidence intervals clustered at the region level are much tighter than that in Rosenzweig and Udry (2019). This is because those confidence intervals account for different sources of variation in the estimates. Those clustered at the regional level account for the region-level shocks that occurred over the duration of the experiment, while the confidence interval in Rosenzweig and Udry (2019) accounts for the variability in rainfalls over a much longer period of time.

Table 3.3: Confidence Interval for Net Returns of Planting-Stage Investments

	95% Confidence Interval
Normal approximation with clustered SE at region level	[-64.14%,254.79%]
Wild-bootstrap with clustered SE at region level	[-97.62%, 289.10%]
Confidence interval in Rosenzweig and Udry (2019)	[-1105% , 1509%]

The results in the first two rows in this table are based on the regression in Column 2 of Table 3 of Rosenzweig and Udry (2019). The confidence interval in the first row uses the normal approximation and the estimate of returns to planting-stage investment from that regression, clustering standard errors at the region level. That in the second row uses the wild cluster bootstrap and the estimate of returns to planting-stage investment from that regression, clustering standard errors at the region level. That in the third row uses the 2.5 and 97.5 percentiles from the distribution of returns to planting stage investment in Figure 5 of Rosenzweig and Udry (2019).

3.4.2 Barriers to Household Risk Management: Evidence from India

In this section we reexamine the results in Cole et al. (2013), who conducted an experiment in India to study the effect of price and nonprice factors in the adoption of an innovative rainfall insurance product. The authors estimate the impact of the following treatments on the decision to purchase insurance: whether the household is visited by an insurance educator; whether the educator was endorsed by local agents that have close relationships with rural villages; whether the educator presented an additional education module about the financial product; and whether the visited household received a high cash reward.[7] The treatments were assigned at the household level, within each of the 37 villages participating in their experiment. Households have time after the visit to determine whether they would like to buy insurance or not, and shocks that could affect

[7] The endorsement treatment was only assigned in two-thirds of the villages.

their decision, such as weather shocks, may occur during this period.

Table 3.4 below replicates Table 5 of Cole et al. (2013). As Cole et al. (2013) find no effect for the endorsement and education treatments, we only report results for the two other treatments. We first use the robust variance estimators used by the authors, and then cluster standard errors at the village level. This RCT took place in 37 villages of two districts of Andhra Pradesh. Therefore, we are unable to cluster at a higher geographical level than these 37 villages, and we can only account for fairly disaggregated village-level shocks. Nonetheless, the results below show even these disaggregated shocks seem to alter the effect of one of the two treatments.

Table 3.4: Conditional and Unconditional Results.

Dependent Variable: Insurance Take-Up		
	Robust s.e.	Clustered s.e.
Visit	0.115***	0.115
	(0.043)	(0.089)
High reward	0.394***	0.394***
	(0.034)	(0.045)
Household controls	Yes	Yes
Village FEs	Yes	Yes
N	1047	1047
Mean of Dep Var	0.282	0.282

The results in this table are based on specification 3 of Table 5 from Cole et al. (2013). The dependent variable in the regression is an indicator for whether the household purchased an insurance policy. The treatment variables are indicators for whether the household was visited by an insurance educator; whether the educator was endorsed by an LSA; whether the educator presented the education module; and whether the visited household received a high cash reward. Household controls are the same as in Cole et al. (2013). Robust standard errors are shown in parentheses in the first column. Standard errors clustered at the village level are shown in the second column.

* p<0.10 ** p<0.05 *** p<0.01.

The first column of Table 3.4 presents results using robust standard errors. The effects of both treatments are significant, so for both of them we can reject $ATE(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)) = 0$: conditional on the village-shocks that arose during the experiment, the treatment had an effect. Assumption 10 is likely to fail in this application: receiving a visit from an educator that describes features of the insurance and answers the household's questions can affect how these households respond to village-specific economic and weather shocks arising between the visit and the time when they need to make their insurance decisions.

Therefore, using robust standard errors may not be appropriate to test $ATE = 0$, one may instead have to use clustered standard errors. With clustered errors, the second column of the table shows that the effect of the high-reward treatment is still significant. We can reject $ATE = 0$ for that treatment: its effect does not seem to be driven by the specific village-level shocks that arose during the experiment. On the other hand, the effect of the visit treatment is no longer significant with clustering. We cannot reject $ATE = 0$ for that treatment: its effect may have been driven by the village-level shocks that occurred during the experiment.

The regression in Table 3.4 has several treatments, and household-level controls. Though extending our theoretical results to regressions with several treatments and controls should be straightforward, in Table A1 in the appendix we show the results we obtain in a simplified version of the regression in Table 3.4, with only the visit treatment and without controls. Standard errors clustered at the village level are 35% larger than robust standard errors. The estimated treatment effect is significant at the 1% (resp. 5%) level with robust (resp. village-clustered) standard errors.

3.5 Conclusion

In RCTs with household-level treatment assignment and household- as well as village-level shocks affecting the potential outcomes, we show that one may use heteroskedasticity-robust or village-clustered standard errors, depending on whether one wants to draw inference on the ATE conditional on the village-level shocks, or netted out of those shocks.

Appendix A

Appendix for “A Framework for Using Value-Added in Regression”

A.1 Tables

Table A.1: Simulation Evidence of the Properties of $\widehat{\kappa}$

	OLS Current Practice (1)	OLS Corrected SEs (2)
n=900,000 and J=3,000		
SD from Monte-Carlo	2.156	2.156
Average SD of $\widehat{\kappa}$	1.265	2.100
Coverage Rate of 95% CI	0.724	0.942

Results are based on 1000 replications. The coverage rate is obtained by taking the average of an indicator of whether the true value κ_0 is in the estimated confidence interval for a 1000 replications with standard errors clustered at the teacher level. The number of students per class n_j and classes per teacher T are held constant at 30 and 10 respectively in all simulations.

Table A.2: Simulation Evidence of the Properties of $\widehat{\kappa}$

	OLS Current Practice (1)	Optimal GMM (2)
n=900,000 and J=3,000		
Variance from Monte Carlo	4.652	4.592

Results are based on 1000 replications. The number of students per class n_j and classes per teacher T are held constant at 30 and 10 respectively in all simulations.

Table A.3: Summary Statistics

	Mean (1)	S.D. (2)	Obs. (3)
A. Student Level Short-Run Variables			
Math Test Score	0.060	0.974	388,191
Class Size	21.968	3.379	388,191
Female	0.494	0.500	388,191
Lunch Eligibility	0.446	0.497	388,191
Black	0.284	0.451	388,191
Hispanic	0.053	0.223	388,191
White	0.608	0.488	388,191
English Language Learner	0.032	0.175	388,191
Special Education	0.110	0.312	388,191
B. Student Level Long-Run Outcomes			
High-School Algebra Score	0.243	0.924	303,826
Graduate High School	0.907	0.290	280,542
Plan College	0.788	0.409	272,990
Plan 4-Year College	0.417	0.493	272,987
Weighted High-School GPA	3.072	0.939	193,927
Class Rank	0.513	0.286	193,594

The sample consists of 388,191 North Carolina public schools third grade students matched to 5,266 teachers in 19,351 classrooms in the years 2000-2005.

Table A.4: Summary Statistics for VA measures

Mean of VA	0.013
S.D of VA	.177
Number of Observations	19,351

Summary statistics for the VA measures of 5,266 teachers in 19,351 classrooms in the years 2000-2005. They are estimated use within teacher variation following the procedure described in section 1.2.1. Controls include cubic polynomials in prior scores, gender, age, indicators for special education, limited English, year, lunch eligibility, ethnicity, as well as class- and school-year means of those variables

Table A.5: Estimates of Long-Run Impacts

	Algebra Score	Graduation	Plan College	Plan 4-Year College	HS GPA	Class Rank
Teacher VA	0.038	0.005	0.009	0.018	0.036	0.009
OLS Heteroskedasticity Robust SE	(0.0013)	(0.0005)	(0.0007)	(0.0008)	(0.0017)	(0.0005)
OLS Clustered SE	(0.0027)	(0.0007)	(0.0011)	(0.0015)	(0.0027)	(0.0010)
GMM Clustered SE	(0.0037)	(0.0012)	(0.0018)	(0.0022)	(0.0040)	(0.0016)
N	303,733	280,456	272,907	272,904	193,867	193,535

Teacher VA is standardized. The results in this table are obtained by a univariate regression of the residualized outcome on teacher VA, following the methodology of section 1.2.1. Standard errors are clustered at the teacher level. Controls for estimation of VA and residualization of outcome include cubic polynomials in prior scores; gender; age; indicators for special education, limited English, year, lunch eligibility, ethnicity; as well as class- and school-year means of those variables.

Table A.6: Estimates of Unconditional Sorting

Dependent Variable	Teacher VA		
	(1)	(2)	(3)
A. Regression of teacher VA measures on covariates			
Lagged Test Scores	0.010		
OLS Standard Error	(0.0012)		
Bootstrap Standard Error	(0.0018)		
Classroom Mean Lagged Test Score		0.056	
OLS Standard Error		(0.0065)	
Bootstrap Standard Error		(0.0100)	
School-Year Mean Lagged Test Score			0.041
OLS Standard Error			(0.0083)
Bootstrap Standard Error			(0.0133)
N	388,191	388,191	388,191
B. Regression of preliminary teacher VA measures on covariates			
Classroom Mean Lagged Test Score	0.073		
OLS Standard Error	(0.0063)		
GMM Standard Error	(0.0114)		
N	444,018		

Standard errors are clustered at the teacher level. Panel A presents the results of regressions of the estimated teacher VA measures on different controls. Bootstrapped standard errors are calculated by block bootstrapping the sample at the teacher level then estimating VA followed by the regressions, this is equivalent to bootstrapping the GMM system. Following Chetty et al. (2014a), the coefficients and standard errors in panel A are multiplied by 1.17 to offset shrinkage of the dependent variable. Panel B presents the results of a regression of preliminary (unshrunk VA measure) on classroom mean lagged test score following Theorem 3. OLS standard errors are the unadjusted estimators produced by statistical softwares, GMM standard errors are obtained following Theorem 3. Observations are slightly higher in panel B because teachers who only teach one year can be included when following this methodology.

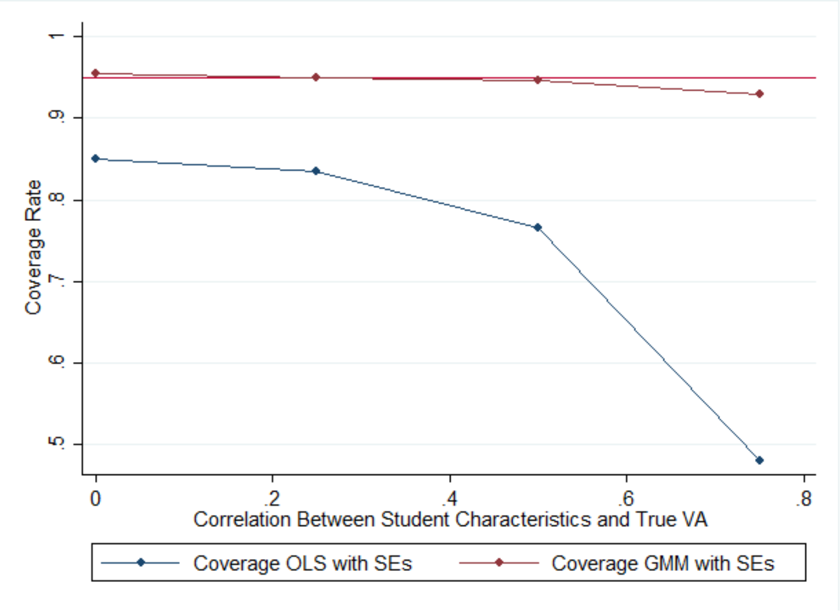
Table A.7: Correlations Between Different VA measures

	Baseline VA	VA using all variation
Baseline VA	1.00	0.965
VA using all variation	0.965	1.00

This table presents the two-way correlation coefficient between the baseline VA measures estimated using only within teacher variation, and the VA measures estimated using between and within teacher variation.

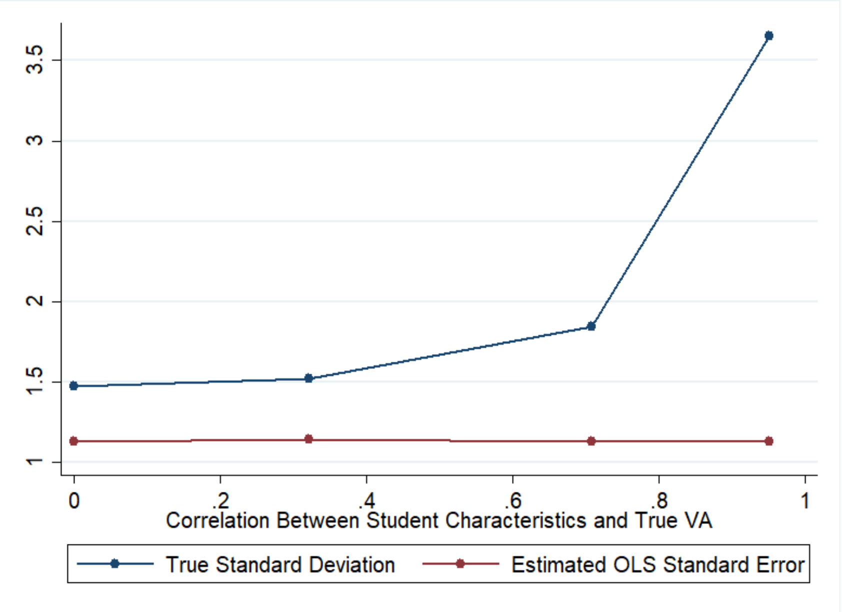
A.2 Figures

Figure A.1: Coverage Rate using OLS SEs and GMM SEs with Different Correlation Levels



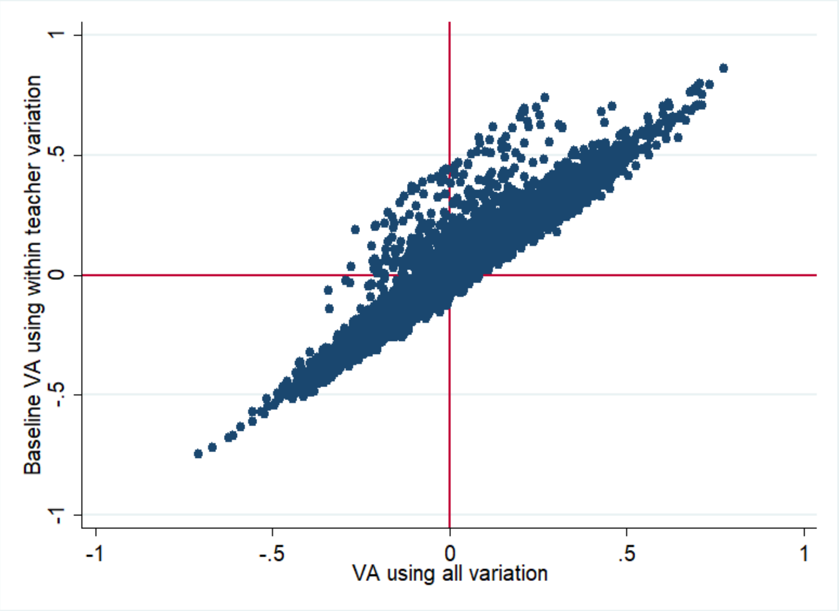
Each dot represents a coverage rate obtained by taking the average of an indicator of whether the true value κ_0 is in the estimated confidence interval for a 1000 replications with standard errors clustered at the teacher level. The number of students per class n_j and classes per teacher T are held constant at 30 and 10 respectively in all simulations, the correlation between student characteristics and true VA is increased by increasing the parameter ρ from the data generating process described in section 1.4. ρ is set to be 0, 0.25 , 0.5, and 0.75.

Figure A.2: Actual Standard Deviation of $\widehat{\kappa}$ vs Standard Error Obtained from OLS



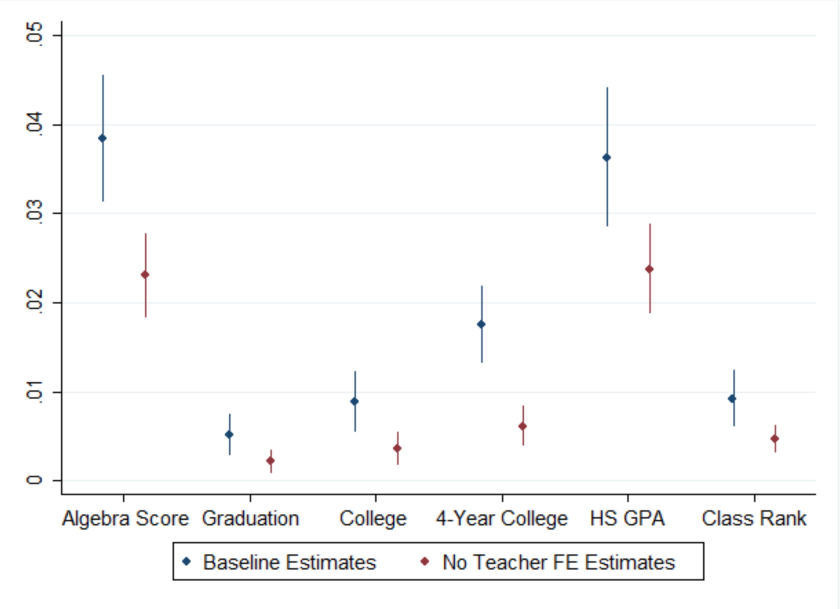
Each dot of the blue line represents the standard deviation of $\widehat{\kappa}$ obtained from a Monte-Carlo using a 1000 replications. Each dot of the red line represents the average of the standard errors estimated using OLS from 1000 replications with standard errors clustered at the teacher level. The number of students per class n_j and classes per teacher T are held constant at 30 and 10 respectively in all simulations, the correlation between student characteristics and true VA is increased by increasing the parameter ρ from the data generating process described in section 1.4. ρ is set to be 0, 0.25 , 0.5, and 0.75.

Figure A.3: Baseline VA vs VA Using All Variation



This graphs plots the baseline VA measures, constructed using an estimator of β_0 that was estimated using teacher fixed-effects, against VA measures constructed using an estimator of β_0 without fixed effects. Controls used are: cubic polynomials in prior scores; gender; age; indicators for special education, limited English, year, lunch eligibility, ethnicity; as well as class- and school-year means of those variables.

Figure A.4: Estimates of Long-Run Impacts



This graphs plots the effect of a one standard deviation increase in teacher VA on different long-run outcomes. They are obtained by a univariate regression of the residualized outcome on teacher VA, following the methodology of section 1.2.1. The estimates in blue are the estimates of κ_0 for different outcomes using the exact methodology in section 1.2.1, while the estimates in red are the estimates of κ_0 without including teacher fixed effects to estimate β_0 and β_0^Y. Standard errors are clustered at the teacher level. The standard errors for the baseline estimates are obtained by GMM, the standard errors for the other estimates are unadjusted. Controls for estimation of VA and residualization of outcome include cubic polynomials in prior scores; gender; age; indicators for special education, limited English, year, lunch eligibility, ethnicity; as well as class- and school-year means of those variables.

A.3 Heterogeneous Treatment Effects

Suppose that the potential outcome function for the adult earnings of student i is given by

$$Y_i^{pot}(\mu) = \kappa_i \mu + Y_i^{pot}(0)$$

where $Y_i^{pot}(0)$ is the same as before but κ_i is stochastic varies across students. Then the true teacher-year level residual earnings are:

$$\overline{Y}_{jt} = \kappa_{jt}\mu_{jt} + \overline{\eta}_{jt}, \tag{A.1}$$

where $\kappa_{jt} = \frac{1}{n_j}\sum_{i=1}^{n_j}\kappa_i$. Let $\boldsymbol{\kappa_j}$ be a vector stacking the κ_{jt}. I impose the following assumptions:

Assumption 12 *Heterogeneous Treatment Effects*

1. $\mathbb{E}(\boldsymbol{\kappa_j}) = \kappa^* < \infty$.
2. $\kappa_{jt} \perp\!\!\!\perp \left(\boldsymbol{\mu}_j, \boldsymbol{\epsilon_j^{(-t)}}\right)$.

Point 1 requires that the mean of the individual level effects be finite. Point 2 requires that the treatment effects in year t be independent of a teacher's true VA and of the unobserved determinants of the teacher's other students in years $s \neq t$.

Consider the linear projection of $\overline{Y}_{jt}$ from (A.1) on $\mu_{jt}^* = \sum_{k\neq t}\phi_{0k}\overline{R}_{jk}$:

$$\overline{Y}_{jt} = \kappa_0 \mu_{jt}^* + \overline{u}_{jt}, \tag{A.2}$$

where $\kappa_0 = \frac{Cov(\overline{Y}_{jt},\mu^*_{jt})}{Var(\mu^*_{jt})}$. Now to show that $\kappa_0 = \mathbb{E}(\kappa_{jt})$:

$$\begin{aligned}
\kappa_0 &= \frac{Cov(\overline{Y}_{jt},\mu^*_{jt})}{Var(\mu^*_{jt})} \\
&= \frac{Cov(\kappa_{jt}\mu_{jt}+\overline{\eta}_{jt},\mu^*_{jt})}{Var(\mu^*_{jt})} \\
&= \frac{\mathbb{E}(\kappa_{jt}\mu_{jt}\mu^*_{jt})+\mathbb{E}(\overline{\eta}_{jt}\mu^*_{jt})}{\mathbb{E}(\mu^*_{jt})} \\
&= \frac{\mathbb{E}(\kappa_{jt}\mu_{jt}\mu^*_{jt})}{\mathbb{E}(\mu^{*2}_{jt})} \\
&= \frac{\mathbb{E}(\kappa_{jt})\mathbb{E}(\mu_{jt}\mu^*_{jt})}{\mathbb{E}(\mu^{*2}_{jt})} \\
&= \mathbb{E}(\kappa_{jt})
\end{aligned}$$

where the second equality follows from (A.1). The third equality follows from the fact that μ_{jt} and μ^*_{jt} are mean zero. The fourth equality follows from Assumption 1. The fifth equality follows from Point 2 of Assumption 12. The last equality follows from the fact that $\frac{\mathbb{E}(\mu_{jt}\mu^*_{jt})}{\mathbb{E}(\mu^*_{jt})} = 1$ from Result 2.

Then we still have $\mathbb{E}\left(\boldsymbol{\phi_0'}\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)'}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}\boldsymbol{\phi_0}\right)\right) = 0$ with $\kappa_0 = \mathbb{E}(\kappa_{jt})$, then the non-optimal GMM estimator is robust to heterogeneous treatment effects under Assumption 12.

A.4 Shrinkage

Consider a simple case where value added is constant over time such that:

$$R_{it} = R_{it}^{obs} - X_{it}'\beta_0 = \mu_j + \epsilon_{it} \tag{A.3}$$

and

$$\overline{R}_{jt} = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{it} = \mu_j + \overline{\epsilon}_{jt}. \tag{A.4}$$

Consider the best linear predictor of $\overline{R}_{jt}$ using one other year $\overline{R}_{jt'}$:

$$\mu_{jt}^* = \phi_0 \overline{R}_{jt'} \tag{A.5}$$

where:

$$\phi_0 = \frac{Cov(\overline{R}_{jt}, \overline{R}_{jt'})}{Var(\overline{R}_{jt'})}. \tag{A.6}$$

Then under Assumption 1 we have:

$$\begin{aligned}\phi_0 &= \frac{Cov(\overline{R}_{jt}, \overline{R}_{jt'})}{Var(\overline{R}_{jt'})}\\ &= \frac{Cov(\mu_j + \overline{\epsilon}_{jt}, \mu_j + \overline{\epsilon}_{jt'})}{Var(\mu_j + \overline{\epsilon}_{jt'})}\\ &= \frac{Cov(\mu_j, \mu_j)}{Var(\mu_j) + Var(\overline{\epsilon}_{jt'})}\\ &= \frac{Var(\mu_j)}{Var(\mu_j) + Var(\overline{\epsilon}_{jt'})} < 1\end{aligned}$$

where the second equality follows from the fact that μ_j is uncorrelated with $\overline{\epsilon}_{jt'}$ and $\overline{\epsilon}_{jt}$ by Point 1 of Assumption 1, and the fact that $\overline{\epsilon}_{jt'}$ and $\overline{\epsilon}_{jt}$ are uncorrelated by Point 3 of Assumption 1.

This example shows that when value added is constant, μ_{jt}^* is a shrinkage estimator similar to the one proposed by Kane and Staiger (2008). One can show that the measures will still be shrunk towards the mean of zero when value added is not constant over time and more years are used.

A.5 Alternative Identification Proof

Result A.5.1 *If Assumptions 1 and 2 hold, then κ_0 is identified.*

By Points 1 and 3 of Assumption 2, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0^Y$ are identified by the coefficients on $\boldsymbol{X_j}$ in a regression of $\boldsymbol{R}_j^{obs}$ and $\boldsymbol{Y}_j^{obs}$ respectively on $\boldsymbol{X_j}$ and teacher fixed effects. Namely:

$$\begin{aligned}\boldsymbol{\beta}_0 &= \mathbb{E}(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{X}}_j)^{-1}\mathbb{E}(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{R}}_j^{obs})\\ \boldsymbol{\beta}_0^Y &= \mathbb{E}(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{X}}_j)^{-1}\mathbb{E}(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{Y}}_j^{obs})\end{aligned}$$

since the a regression on the variables using the within transform is equivalent to a regression with fixed effects.

Then starting with (1.2):

$$R_{it}^{obs} = \boldsymbol{X_{it}'\beta_0} + \mu_{jt} + \epsilon_{it},$$

and let:

$$R_{it} = R_{it}^{obs} - \boldsymbol{X_{it}'\beta_0} = \mu_{jt} + \epsilon_{it} \tag{A.7}$$

be the actual residual score, which is then collapsed to the teacher year level:

$$\overline{R}_{jt} = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{it} = \mu_{jt} + \overline{\epsilon}_{jt}, \tag{A.8}$$

we can write the best linear prediction of $\overline{R}_{jt}$ as a function of other years as:

$$\overline{R}_{jt} = \sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{j|s-t|} + \theta_{jt} \tag{A.9}$$

so that:

$$\mu_{jt} = \sum_{|s-t|\neq 0} \phi_{0|s-t|} \overline{R}_{j|s-t|} + \theta_{jt} - \overline{\epsilon}_{jt}. \tag{A.10}$$

Plugging (A.10) into (1.9) and letting $\mu^*_{jt} = \sum_{s\neq 0} \phi_{0s} \overline{R}_{js}$ yields:

$$\overline{Y}_{jt} = \kappa_0 \mu^*_{jt} + \kappa_0 \theta_{jt} - \kappa_0 \overline{\epsilon}_{jt} + \overline{\eta}_{jt}, \tag{A.11}$$

where μ^*_{jt} is identified. Therefore:

$$\begin{aligned} \frac{Cov\left(\overline{Y}_{jt}, \mu^*_{jt}\right)}{Var\left(\mu^*_{jt}\right)} &= \kappa_0 + \kappa_0 \frac{Cov\left(\theta_{jt}, \mu^*_{jt}\right)}{Var\left(\mu^*_{jt}\right)} - \kappa_0 \frac{Cov\left(\overline{\epsilon}_{jt}, \mu^*_{jt}\right)}{Var\left(\mu^*_{jt}\right)} + \frac{Cov\left(\overline{\eta}_{jt}, \mu^*_{jt}\right)}{Var\left(\mu^*_{jt}\right)} \\ &= \kappa_0 \end{aligned} \tag{A.12}$$

where the second equality holds because θ_{jt} is the error from (1.15) and the variables are mean zero, and:

$$\begin{aligned} Cov\left(\overline{\epsilon}_{jt}, \mu^*_{jt}\right) &= Cov\left(\overline{\epsilon}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|} \overline{R}_{js}\right) \\ &= Cov\left(\overline{\epsilon}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|} (\mu_{js} + \overline{\epsilon}_{js})\right) \\ &= Cov\left(\overline{\epsilon}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|} \mu_{js}\right) + Cov\left(\overline{\epsilon}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|} \overline{\epsilon}_{js}\right) \\ &= 0 \end{aligned}$$

where both terms are 0 by Assumption 1.

To show $Cov\left(\overline{\eta}_{jt}, \mu^*_{jt}\right) = 0$:

$$\begin{aligned}
Cov\left(\overline{\eta}_{jt}, \mu^*_{jt}\right) &= Cov\left(\overline{\eta}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js}\right) \\
&= Cov\left(\overline{\eta}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|}(\mu_{js} + \overline{\epsilon}_{js})\right) \\
&= Cov\left(\overline{\eta}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|}\mu_{js}\right) + Cov\left(\overline{\epsilon}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{\epsilon}_{js}\right) \\
&= 0
\end{aligned}$$

where both terms are 0 by Assumption 1.

QED.

A.6 Regularity Conditions

Assumption A.6.1 *Technical Assumptions for Consistency*

1. $(\boldsymbol{\beta_0'}, \boldsymbol{\beta_0^{Y\prime}}, \boldsymbol{\phi_0'}, \kappa_0)$ *is an element in the interior of* $\boldsymbol{\Theta}$ *and* $\boldsymbol{\Theta}$ *is a compact subset of* $\mathcal{R}^{2K+T}$.

2.

$$\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\boldsymbol{\phi},\kappa)\in\Theta}\left|\left|\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)\right|\right|\right)<\infty$$
$$\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\boldsymbol{\phi},\kappa)\in\Theta}\left|\left|\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\left(\widetilde{\boldsymbol{R}}_j-\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right)\right|\right|\right)<\infty$$
$$\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\boldsymbol{\phi},\kappa)\in\Theta}\left|\left|\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta^Y}\right)\right|\right|\right)<\infty$$
$$\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\boldsymbol{\phi},\kappa)\in\Theta}\left|\boldsymbol{\phi}'\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\left(\widetilde{\boldsymbol{Y}}_j-\kappa\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right)\right|\right)<\infty$$

3. $\left(\boldsymbol{X}_j,\boldsymbol{\mu}_j,\boldsymbol{\epsilon}_j,\boldsymbol{Y}_j^{obs},\boldsymbol{\eta}_j\right)$ *are i.i.d across* j.

where $\widetilde{R}_{jt}=\overline{R}_{jt}^{obs}-\overline{\boldsymbol{X}}_{\boldsymbol{jt}}'\boldsymbol{\beta}$, $\widetilde{\boldsymbol{R}}_j$ *is a vector stacking all* $\widetilde{R}_{jt}$ *for teacher* j, $\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}$ *is a matrix stacking* T *row vectors (of dimension* $1\times(T-1)$*) with each row (indexed by* t*) containing* $T-1$ *different* $\widetilde{R}_{jk}$ *for* $k\neq t$.

Assumption A.6.2 *Technical Assumptions for Asymptotic Normality and Consistent Variance Estimation*

1. For all $(\boldsymbol{\beta'},\boldsymbol{\beta^{Y\prime}},\boldsymbol{\phi'},\kappa)\in\mathcal{R}^{2K+T}$:

$$E\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)\left(\ddot{\boldsymbol{R}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)'\ddot{\boldsymbol{X}}_j\right)<\infty$$
$$\mathbb{E}\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}-\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right)\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}-\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right)'\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\right)<\infty$$
$$\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta^Y}\right)\left(\ddot{\boldsymbol{Y}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta^Y}\right)'\ddot{\boldsymbol{X}}_j\right)<\infty$$
$$\mathbb{E}\left(\boldsymbol{\phi}'\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\left(\boldsymbol{Y}_j-\kappa\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right)\left(\boldsymbol{Y}_j-\kappa\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right)'\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right)<\infty.$$

2. $\mathbb{E}\left|\underset{(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)\in\boldsymbol{\Theta}}{sup}\left(\boldsymbol{\phi'}\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right)\right|<\infty$ *and* $\mathbb{E}\left|\underset{(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)\in\boldsymbol{\Theta}}{sup}\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\right)\right|<\infty.$

3. $\mathbb{E}\left[\left(\boldsymbol{\phi_0'}\boldsymbol{R_j^{(-t)\prime}}\boldsymbol{R_j^{(-t)}}\boldsymbol{\phi_0}\right)\right], \mathbb{E}\left[\left(\boldsymbol{R_j^{(-t)\prime}}\boldsymbol{R_j^{(-t)}}\right)\right]$ *are invertible.*

4.

$$\begin{aligned}
&\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\boldsymbol{\phi},\kappa)\in\Theta}\left|\left|\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)\right|\right|^2\right)<\infty\\
&\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\boldsymbol{\phi},\kappa)\in\Theta}\left|\left|\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)\prime}}\left(\widetilde{\boldsymbol{R}}_j-\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}\boldsymbol{\phi}\right)\right|\right|^2\right)<\infty\\
&\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\boldsymbol{\phi},\kappa)\in\Theta}\left|\left|\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta^Y}\right)\right|\right|^2\right)<\infty\\
&\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\boldsymbol{\phi},\kappa)\in\Theta}\left|\boldsymbol{\phi'}\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)\prime}}\left(\widetilde{\boldsymbol{Y}}_j-\kappa\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}\boldsymbol{\phi}\right)\right|^2\right)<\infty
\end{aligned}$$

Assumption A.6.3 *Technical Assumptions for Optimal GMM*

1. $(\boldsymbol{\beta_0'},\boldsymbol{\beta_0^{Y\prime}},\kappa_0)$ *is an element in the interior of* $\boldsymbol{\Theta_1}$ *and* $\boldsymbol{\Theta_1}$ *is a compact subset of* $\mathcal{R}^{2K+1}$.

2.

$$\begin{aligned}
&\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\kappa)\in\Theta_1}\left|\left|\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)\right|\right|^2\right)<\infty\\
&\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\kappa)\in\Theta_1}\left|\left|\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta^Y}\right)\right|\right|^2\right)<\infty\\
&\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\kappa)\in\Theta_1}\left|\left|\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)\prime}}\left(\widetilde{\boldsymbol{Y}}_j-\kappa\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}\right)\right|\right|^2\right)<\infty
\end{aligned}$$

3. $\left(\boldsymbol{X}_j,\boldsymbol{\mu}_j,\boldsymbol{\epsilon}_j,\boldsymbol{Y}_j^{obs},\boldsymbol{\eta}_j\right)$ *are i.i.d across* j.

4. For all $(\boldsymbol{\beta}, \boldsymbol{\beta^Y}, \kappa) \in \mathcal{R}^{2K+1}$*:*

$$
\begin{gathered}
E\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)'\ddot{\boldsymbol{X}}_j\right) < \infty \\
\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta^Y}\right)\left(\ddot{\boldsymbol{Y}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta^Y}\right)'\ddot{\boldsymbol{X}}_j\right) < \infty \\
\mathbb{E}\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(\boldsymbol{-t})'}\left(\widetilde{\boldsymbol{Y}}_j - \kappa\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}\right)\left(\widetilde{\boldsymbol{Y}}_j - \kappa\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}\right)'\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(\boldsymbol{-t})}\right) < \infty.
\end{gathered}
$$

5. $\mathbb{E}\left|\underset{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\kappa)\in\boldsymbol{\Theta_1}}{sup}\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(\boldsymbol{-t})'}\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(\boldsymbol{-t})}\right)\right| < \infty.$

Assumption A.6.3 reframes Assumptions A.6.1 and A.6.2 for the new set of moments.

Assumption A.6.4 *Assumptions for Consistency, and Asymptotic Normality*

1. $(\boldsymbol{\beta_0'}, \boldsymbol{\alpha_0'})$ *is an element of the interior of* $\boldsymbol{\Theta_2}$ *and* $\boldsymbol{\Theta_2}$ *is a compact subset of* $\mathcal{R}^{K+K_D}$.

2.

$$
\begin{gathered}
\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\alpha})\in\Theta_2}\left|\left|\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)\right|\right|^2\right) < \infty \\
\mathbb{E}\left(sup_{(\boldsymbol{\beta},\boldsymbol{\alpha})\in\Theta_2}\left|\left|\boldsymbol{D_j'}\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}} - \boldsymbol{D_j}\boldsymbol{\alpha}\right)\right|\right|^2\right) < \infty
\end{gathered}
$$

3. $(\boldsymbol{X}_j, \boldsymbol{\mu}_j, \boldsymbol{\epsilon}_j, \boldsymbol{D}_j)$ *are i.i.d across j.*

4. For all $(\boldsymbol{\beta}, \boldsymbol{\alpha}) \in \mathcal{R}^{K+K_D}$*:*

$$
\begin{gathered}
E\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)'\ddot{\boldsymbol{X}}_j\right) < \infty \\
\mathbb{E}\left(\boldsymbol{D_j'}\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}} - \boldsymbol{D_j}\boldsymbol{\alpha}\right)\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}} - \boldsymbol{D_j}\boldsymbol{\alpha}\right)'\boldsymbol{D_j}\right) < \infty
\end{gathered}
$$

A.7 Consistency, Asymptotic Normality, and Consistent Variance Estimation of the GMM Estimators

Lemma A.7.1 *If Assumptions 1, 2, and A.6.1 hold, then* $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{Y}}, \widehat{\kappa}) \xrightarrow{p} (\boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)$*, where* $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{Y}}, \widehat{\kappa})$ *are the estimators obtained from a GMM minimization of the moment system of* (1.11), (1.13), (1.12), *and* (1.14) *with the identity matrix as a weighting matrix.*

Theorem A.7.1 *If Assumptions 1, 2, A.6.1, and A.6.2 hold, then*

$$\sqrt{J}\left[(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{Y}}, \widehat{\kappa}) - (\boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)\right] \rightsquigarrow N\left(0, \Omega\right).$$

where $\Omega = \widetilde{G}^{-1}\mathbb{E}[\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)']\widetilde{G}^{-1'}$ *and* $\widetilde{G} = \mathbb{E}\left[\nabla_{(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\beta^Y}, \kappa)}\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)\right]$*, and* $\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)$ *is defined in* (A.26) *in the proof.*

Let $\widehat{\Omega}$ *correspond to an estimator of* Ω*, constructed by replacing the population moments in* Ω *by averages and the parameters by the GMM estimators. Then:*

$$\widehat{\Omega} \xrightarrow{p} \Omega \tag{A.13}$$

Lemma A.7.2 *If Assumptions 1, 2, and A.6.3 hold, then*

1.

$$\sqrt{J}\left[(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{Y}}, \widehat{\kappa}) - (\boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \kappa_0)\right] \rightsquigarrow N\left(0, \Omega_1\right).$$

where $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{Y}}, \widehat{\kappa})$ *are the estimators obtained from a GMM minimization and* $\Omega_1 = \left(\widetilde{G}_1'\widetilde{G}_1\right)^{-1}\widetilde{G}_1'\mathbb{E}[\widetilde{g}_1(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \kappa_0)\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \kappa_0)']\widetilde{G}_1\left(\widetilde{G}_1'\widetilde{G}_1\right)^{-1}$*,* *and* $\widetilde{G}_1 = \mathbb{E}\left[\nabla_{(\boldsymbol{\beta},\boldsymbol{\beta^Y},\kappa)}\widetilde{g}_1(\boldsymbol{Z},\boldsymbol{\beta_0},\boldsymbol{\beta_0^Y},\kappa_0)\right]$*, where* $\widetilde{g}_1(\boldsymbol{Z},\boldsymbol{\beta_0},\boldsymbol{\beta_0^Y},\kappa_0)$ *is defined in* (A.32) *in the proof.*

2. $\widehat{\Omega}_1 \xrightarrow{p} \Omega_1$ *where* $\widehat{\Omega}_1$ *corresponds to the sample equivalent of* Ω_1 *replacing moments by sample moments and parameters by* $(\boldsymbol{\widehat{\beta}},\boldsymbol{\widehat{\beta}^Y},\widehat{\kappa})$*.*

Theorem A.7.2 *If Assumptions 1, 2, and A.6.3 hold, then*

$$\sqrt{J}\left[(\boldsymbol{\widehat{\beta}^*},\boldsymbol{\widehat{\beta}^{Y*}},\widehat{\kappa}^*)-(\boldsymbol{\beta_0},\boldsymbol{\beta_0^Y},\kappa_0)\right] \rightsquigarrow N\left(0,\Omega^*\right).$$

where $(\boldsymbol{\widehat{\beta}^*},\boldsymbol{\widehat{\beta}^{Y*}},\widehat{\kappa}^*)$ *are the estimates resulting from a GMM minimization using* $\widehat{W}^*$ *as a weighting matrix,*

$\Omega^* = \left(\widetilde{G}_1' W^* \widetilde{G}_1\right)^{-1}$*, and* $\Omega^* \leq \Omega_2$ *where* Ω_2 *is the submatrix of* Ω *that corresponds to the variance covariance matrix of* $(\boldsymbol{\widehat{\beta}},\boldsymbol{\widehat{\beta}^Y},\widehat{\kappa})$*.*

Result A.7.1 *If Assumptions 1, 2, and A.6.3 hold, then*

$$\widehat{\widetilde{g}}_1(\boldsymbol{Z},\boldsymbol{\widehat{\beta}^*},\boldsymbol{\widehat{\beta}^{Y*}},\widehat{\kappa}^*)'\widehat{W}^*\widehat{\widetilde{g}}_1(\boldsymbol{Z},\boldsymbol{\widehat{\beta}^*},\boldsymbol{\widehat{\beta}^{Y*}},\widehat{\kappa}^*) \rightsquigarrow \chi^2_{T-2}.$$

Theorem A.7.3 *If Assumptions 1, 2, 5, A.6.4 and hold, then*

$$\sqrt{J}\left[(\boldsymbol{\widehat{\beta}},\boldsymbol{\widehat{\alpha}})-(\boldsymbol{\beta_0},\boldsymbol{\alpha_0})\right] \rightsquigarrow N\left(0,\Omega_2\right).$$

where $\Omega_2 = \widetilde{G}_2^{-1}\mathbb{E}[\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta_0},\boldsymbol{\alpha_0})\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta_0},\boldsymbol{\alpha_0'})]\widetilde{G}_2^{-1'}$ *and* $\widetilde{G}_2 = \mathbb{E}\left[\nabla_{(\boldsymbol{\beta},\boldsymbol{\alpha})}\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta_0},\boldsymbol{\alpha_0})\right]$*, where* $\boldsymbol{Z_j} = \left(\boldsymbol{X_j},\boldsymbol{R_j^{obs}},\boldsymbol{D_j}\right)$*,* $\boldsymbol{Z}_2$ *stacks the* $\boldsymbol{Z_j}$*, and* $\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta_0},\boldsymbol{\alpha_0})$ *is defined in* (A.37) *in the proof.*

A.8 Variance Comparison

Theorem 1 also allows us to compare σ^2 to the variance obtained from OLS:

$$\begin{aligned}
&\sigma^2 - (G_\kappa^{-1})^2 \mathbb{E}\left(g(\boldsymbol{Z})^2\right) \\
=&(G_\kappa^{-1})^2 \\
&\left[\mathbb{E}\left(\left(g(\boldsymbol{Z}) + G_{\beta^Y}\psi_3(\boldsymbol{Z}) + G_\phi\psi_2(\boldsymbol{Z}) + G_\beta\psi_1(\boldsymbol{z}) - G_\phi M_{2\phi}^{-1} M_{2\beta}\psi_1(\boldsymbol{Z})\right)^2 - \mathbb{E}\left(\left(g(\boldsymbol{Z})^2\right)\right)\right)\right] \\
=&(G_\kappa^{-1})^2 \mathbb{E}\left((2g(\boldsymbol{Z}))\left(G_{\beta^Y}\psi_3(\boldsymbol{Z}) + G_\phi\psi_2(\boldsymbol{Z}) + G_\beta\psi_1(\boldsymbol{z}) - G_\phi M_{2\phi}^{-1} M_{2\beta}\psi_1(\boldsymbol{Z})\right)\right) \\
&+ (G_\kappa^{-1})^2 \mathbb{E}\left(\left(G_{\beta^Y}\psi_3(\boldsymbol{Z}) + G_\phi\psi_2(\boldsymbol{Z}) + G_\beta\psi_1(\boldsymbol{z}) - G_\phi M_{2\phi}^{-1} M_{2\beta}\psi_1(\boldsymbol{Z})\right)^2\right)
\end{aligned}$$

where the equality follows from factoring $a^2 - b^2$ and the linearity of the expectation operator. Given that the second term is always positive, the difference between the two variances is going to depend on the covariances between the moment used to estimate κ_0 and the other three sets moments. Then the covariances of interest are:

$$\begin{aligned}
&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j' \left(\ddot{\boldsymbol{\mu}}_{\boldsymbol{j}} + \ddot{\boldsymbol{\epsilon}}_{\boldsymbol{j}}\right)\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}} - \kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}} + \boldsymbol{\eta}_{\boldsymbol{j}}\right)' \boldsymbol{R}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}_{\boldsymbol{0}}\right) \\
&\mathbb{E}\left(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\boldsymbol{\theta}_{\boldsymbol{j}}\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}} - \kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}} + \boldsymbol{\eta}_{\boldsymbol{j}}\right)' \boldsymbol{R}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}_{\boldsymbol{0}}\right) \\
&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{\eta}}_{\boldsymbol{j}}\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}} - \kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}} + \boldsymbol{\eta}_{\boldsymbol{j}}\right)' \boldsymbol{R}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}_{\boldsymbol{0}}\right).
\end{aligned}$$

For ease of exposition of the difference between the two variances, assume that the covariances of the errors are homoskedastic with respect to $\boldsymbol{R}_{\boldsymbol{j}}^{(-t)}$ and $\ddot{\boldsymbol{X}}_{\boldsymbol{j}}$ such that:

$$
\begin{aligned}
&\mathbb{E}\left(\left(\ddot{\boldsymbol{\mu}}_{\boldsymbol{j}}+\ddot{\boldsymbol{\epsilon}}_{\boldsymbol{j}}\right)\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}}-\kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}}+\boldsymbol{\eta}_{\boldsymbol{j}}\right)'|\boldsymbol{R}_{\boldsymbol{j}}^{(-t)},\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\right)=\mathbb{E}\left(\left(\ddot{\boldsymbol{\mu}}_{\boldsymbol{j}}+\ddot{\boldsymbol{\epsilon}}_{\boldsymbol{j}}\right)\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}}-\kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}}+\boldsymbol{\eta}_{\boldsymbol{j}}\right)'\right)\\
&\mathbb{E}\left(\boldsymbol{\theta}_{\boldsymbol{j}}\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}}-\kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}}+\boldsymbol{\eta}_{\boldsymbol{j}}\right)'|\boldsymbol{R}_{\boldsymbol{j}}^{(-t)},\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\right)=\mathbb{E}\left(\boldsymbol{\theta}_{\boldsymbol{j}}\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}}-\kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}}+\boldsymbol{\eta}_{\boldsymbol{j}}\right)'\right)\\
&\mathbb{E}\left(\left(\boldsymbol{\kappa_0}\ddot{\boldsymbol{\mu}}_{\boldsymbol{j}}+\ddot{\boldsymbol{\eta}}_{\boldsymbol{j}}\right)\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}}-\kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}}+\boldsymbol{\eta}_{\boldsymbol{j}}\right)'|\boldsymbol{R}_{\boldsymbol{j}}^{(-t)},\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\right)=\mathbb{E}\left(\left(\boldsymbol{\kappa_0}\ddot{\boldsymbol{\mu}}_{\boldsymbol{j}}+\ddot{\boldsymbol{\eta}}_{\boldsymbol{j}}\right)\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}}-\kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}}+\boldsymbol{\eta}_{\boldsymbol{j}}\right)'\right)
\end{aligned}
$$

we get:

$$
\begin{aligned}
&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\mathbb{E}\left(\left(\ddot{\boldsymbol{\mu}}_{\boldsymbol{j}}+\ddot{\boldsymbol{\epsilon}}_{\boldsymbol{j}}\right)\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}}-\kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}}+\boldsymbol{\eta}_{\boldsymbol{j}}\right)'\right)\boldsymbol{R}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}_{\boldsymbol{0}}\right)\\
&\mathbb{E}\left(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)'}\mathbb{E}\left(\boldsymbol{\theta}_{\boldsymbol{j}}\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}}-\kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}}+\boldsymbol{\eta}_{\boldsymbol{j}}\right)'\right)\boldsymbol{R}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}_{\boldsymbol{0}}\right)\\
&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\mathbb{E}\left(\left(\boldsymbol{\kappa_0}\ddot{\boldsymbol{\mu}}_{\boldsymbol{j}}+\ddot{\boldsymbol{\eta}}_{\boldsymbol{j}}\right)\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}}-\kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}}+\boldsymbol{\eta}_{\boldsymbol{j}}\right)'\right)\boldsymbol{R}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}_{\boldsymbol{0}}\right).
\end{aligned}
$$

Note that the first and third set of covariances will depend on $\mathbb{E}\left(\ddot{X}_{jt}\overline{R}_{js}\right)$ for $s \neq t$. Simplifying further, it is reasonable to assume that those covariances are close to zero since they depend on the covariance between within teacher fluctuations in covariates in year t and teacher value added in years $s \neq t$, and average unobserved determinants of test scores in years $s \neq t$, $\mathbb{E}\left(\ddot{X}_{jt}\mu_{js}\right)$, and $\mathbb{E}\left(\ddot{X}_{jt}\bar{\epsilon}_{js}\right)$.

Then for the variance from OLS to be larger than σ^2 it would have to be that

$$
2G_\phi\mathbb{E}\left(\boldsymbol{R}_{\boldsymbol{j}}^{(-t)'}\mathbb{E}\left(\boldsymbol{\theta}_{\boldsymbol{j}}\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}}-\kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}}+\boldsymbol{\eta}_{\boldsymbol{j}}\right)'\right)\boldsymbol{R}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}_{\boldsymbol{0}}\right)<0
$$

and

$$
\begin{gathered}
\left|2G_\phi\mathbb{E}\left(\boldsymbol{R}_{\boldsymbol{j}}^{(-t)'}\mathbb{E}\left(\boldsymbol{\theta}_{\boldsymbol{j}}\left(\kappa_0\boldsymbol{\theta}_{\boldsymbol{j}}-\kappa_0\boldsymbol{\epsilon}_{\boldsymbol{j}}+\boldsymbol{\eta}_{\boldsymbol{j}}\right)'\right)\boldsymbol{R}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}_{\boldsymbol{0}}\right)\right|>\\
\mathbb{E}\left(\left(G_{\beta^Y}\psi_3(\boldsymbol{Z})+G_\phi\psi_2(\boldsymbol{Z})+G_\beta\psi_1(\boldsymbol{z})-G_\phi M_{2\phi}^{-1}M_{2\beta}\psi_1(\boldsymbol{Z})\right)^2\right).
\end{gathered}
$$

In other words the contribution to σ^2 of the covariances between the moments used to estimate $\boldsymbol{\phi_0}$ and κ_0 has to be large and negative enough to outweigh the contribution to σ^2 from the variances and covariances of the moments used to estimate $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_0^Y}$, and $\boldsymbol{\phi_0}$. Therefore, it is likely that the variance from OLS be smaller than σ^2 in most cases.

A.9 Proofs

A.9.1 Proof of Result 1

$$\begin{aligned}
&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right)\\
=&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{\mu}}_{\boldsymbol{j}}+\ddot{\boldsymbol{\epsilon}}_{\boldsymbol{j}}\right)\right)\\
=&0
\end{aligned}$$

where the first equality follows from (1.2), and the second equality follows from Point 3 of Assumption 2.

$$\begin{aligned}
&\mathbb{E}\left(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)'}}\left(\boldsymbol{R}_j-\boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}\boldsymbol{\phi_0}\right)\right)\\
=&\mathbb{E}\left(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)'}}\boldsymbol{\theta_j}\right)\\
=&0
\end{aligned}$$

where the first equality follows from (1.15), and the second equality follows from the fact that θ_{jt} is orthogonal to $\overline{R}_{jk}$ for $k \neq t$ by construction.

$$
\begin{aligned}
&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}_{\boldsymbol{0}}^{\boldsymbol{Y}}\right)\right)\\
=&\mathbb{E}\left(\ddot{\boldsymbol{X}}_{\boldsymbol{j}}(\ddot{\boldsymbol{\eta}}_{\boldsymbol{j}}+\kappa_0\ddot{\boldsymbol{\mu}}_{\boldsymbol{j}})\right)\\
=&0
\end{aligned}
$$

where the first equality follows from (1.9) and the second from Point 3 of Assumption 2.

$$
\begin{aligned}
&\mathbb{E}\left(\boldsymbol{\phi}_{\boldsymbol{0}}'\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\left(\boldsymbol{Y}_j-\kappa_0\boldsymbol{R}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}_{\boldsymbol{0}}\right)\right)\\
=&\mathbb{E}\left(\mu_{jt}^*\left(\boldsymbol{Y}_j-\kappa_0\mu_{jt}^*\right)\right)\\
=&\mathbb{E}\left(\mu_{jt}^*\left(\kappa_0\theta_{jt}-\kappa_0\bar{\epsilon}_{jt}+\overline{\eta}_{jt}\right)\right)\\
=&0
\end{aligned}
$$

where the second equality follow from (A.11) and the third equality follows from Assumption 1 and from the fact that θ_{jt} is orthogonal to $\overline{R}_{jk}$ for $k \neq t$ by construction.

Q.E.D

A.9.2 Proof of Result 2

Starting with $Cov\left(\overline{\eta}_{jt}, \mu^*_{jt}\right) = 0$:

$$
\begin{aligned}
& Cov\left(\overline{\eta}_{jt}, \mu^*_{jt}\right) \\
=& Cov\left(\overline{\eta}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js}\right) \\
=& Cov\left(\overline{\eta}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|}\mu_{js}\right) + Cov\left(\overline{\epsilon}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{\epsilon}_{js}\right) \\
=& 0
\end{aligned}
$$

where both terms are 0 by Assumption 1.

Now for $\frac{Cov\left(\mu_{jt}, \mu^*_{jt}\right)}{Var\left(\mu^*_{jt}\right)} = 1$:

$$
\begin{aligned}
& \frac{Cov\left(\mu_{jt}, \mu^*_{jt}\right)}{Var\left(\mu^*_{jt}\right)} \\
=& \frac{Cov\left(\sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js} + \theta_{jt} - \overline{\epsilon}_{jt}, \mu^*_{jt}\right)}{Var\left(\mu^*_{jt}\right)} \\
=& \frac{Cov\left(\sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js} + \theta_{jt} - \overline{\epsilon}_{jt}, \sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js}\right)}{Var\left(\sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js}\right)} \\
=& \frac{Cov\left(\sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js}, \sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js}\right)}{Var\left(\sum_{|s-t|\neq 0} \phi_{0|s-t|}\overline{R}_{js}\right)} \\
=& 1
\end{aligned}
$$

where the first equality follows from (A.10), the third equality follows from the fact that θ_{jt} is the error from (1.15) and by Assumption 1.

A.9.3 Proof of Lemma A.7.1

We start with the following moments:

$$\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)\right) \tag{A.14}$$

$$\mathbb{E}\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\left(\widetilde{\boldsymbol{R}}_j-\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right)\right) \tag{A.15}$$

$$\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}^{\boldsymbol{Y}}\right)\right) \tag{A.16}$$

$$\mathbb{E}\left(\boldsymbol{\phi}'\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\left(\widetilde{\boldsymbol{Y}}_j-\kappa\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right)\right) \tag{A.17}$$

where $\widetilde{R}_{jt}=\overline{R}_{jt}^{obs}-\overline{\boldsymbol{X}}_{\boldsymbol{jt}}'\boldsymbol{\beta}$, $\widetilde{\boldsymbol{R}}_j$ is a vector stacking all $\widetilde{R}_{jt}$ for teacher j, $\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}$ is a matrix stacking T row vectors (of dimension $1\times(T-1)$) with each row (indexed by t) containing $T-1$ different $\widetilde{R}_{jk}$ for $k\neq t$, and $\widetilde{Y}_{jt}=\overline{Y}_{jt}^{obs}-\overline{\boldsymbol{X}}_{\boldsymbol{jt}}'\boldsymbol{\beta}^{\boldsymbol{Y}}$ and $\widetilde{\boldsymbol{Y}}_j$ is a vector stacking them. Note that (A.14) and (A.16) make it so that β_0 and β_0^Y will be estimated using teacher fixed effects.

Let the sample equivalent of the moment conditions be:

$$\frac{1}{J}\sum_{j=1}^{J}\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right) \tag{A.18}$$

$$\frac{1}{J}\sum_{j=1}^{J}\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\left(\widetilde{\boldsymbol{R}}_j-\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right) \tag{A.19}$$

$$\frac{1}{J}\sum_{j=1}^{J}\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs}-\ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}^{\boldsymbol{Y}}\right) \tag{A.20}$$

$$\frac{1}{J}\sum_{j=1}^{J}\boldsymbol{\phi}'\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\left(\widetilde{\boldsymbol{Y}}_j-\kappa\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\phi}\right) \tag{A.21}$$

Let the GMM weighting matrix be the identity matrix.

To show the estimator is consistent, we need to check the conditions from Theorem 2.6 of Newey and McFadden (1994). Consistency of GMM when the data are i.i.d requires the following conditions:

1. The weighting matrix W is positive semi-definite
2. $W\mathbb{E}[g(\boldsymbol{X, Y^{obs}, R^{obs}}, \boldsymbol{\beta}, \boldsymbol{\beta^Y}, \boldsymbol{\phi}, \kappa)] = 0$ if and only if $(\boldsymbol{\beta}, \boldsymbol{\beta^Y}, \boldsymbol{\phi}, \kappa) = (\boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \boldsymbol{\phi_0}, \kappa_0)$, where $g(\boldsymbol{X, Y^{obs}, R^{obs}}, \boldsymbol{\beta}, \boldsymbol{\beta^Y}, \boldsymbol{\phi}, \kappa)$ is a vector stacking the moment functions.
3. $(\boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \boldsymbol{\phi_0}, \kappa_0) \in \boldsymbol{\Theta}$ and $\boldsymbol{\Theta}$ is compact.
4. $g(\boldsymbol{X, Y^{obs}, R^{obs}}, \boldsymbol{\beta}, \boldsymbol{\beta^Y}, \boldsymbol{\phi}, \kappa)$ is continuous for all $(\boldsymbol{\beta}, \boldsymbol{\beta^Y}, \boldsymbol{\phi}, \kappa) \in \boldsymbol{\Theta}$ with probability one.
5. $\mathbb{E}[sup_{(\boldsymbol{\beta}, \boldsymbol{\beta^Y}, \boldsymbol{\phi}, \kappa) \in \Theta} || g(\boldsymbol{X, Y^{obs}, R^{obs}}, \boldsymbol{\beta}, \boldsymbol{\beta^Y}, \boldsymbol{\phi}, \kappa)||] < \infty$.

The first condition is satisfied since the weighting matrix W is the identity matrix, so it is positive semi-definite.

Furthermore by Result 1 all the moments are equal to 0 when evaluated at $(\boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \boldsymbol{\phi_0}, \kappa_0)$. Therefore $\mathbb{E}[g(\boldsymbol{X, Y^{obs}, R^{obs}}, \boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \boldsymbol{\phi_0}, \kappa_0)] = 0$ and under Assumptions 1 and 2, and (1.1), (1.2), (1.7), and (1.15) we have that $(\boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \boldsymbol{\phi_0}, \kappa_0)$ are uniquely identified so that this holds if and only if $(\boldsymbol{\beta}, \boldsymbol{\beta^Y}, \boldsymbol{\phi}, \kappa) = (\boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \boldsymbol{\phi_0}, \kappa_0)$. Then the second condition holds.

The third condition holds by Point 1 of Assumption A.6.1, and the fourth condition holds by inspection.

The last condition holds by Point 2 of Assumption A.6.1 and the triangle inequality applied to the euclidian norm.

Q.E.D

A.9.4 Proof of Theorem A.7.1

Now to show asymptotic normality we start by rewriting the moments as:

$$\mathbb{E}\left(m_1(\boldsymbol{Z},\boldsymbol{\beta})\right) = E\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)\right) \tag{A.22}$$

$$\mathbb{E}\left(m_2(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\phi})\right) = \mathbb{E}\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(\boldsymbol{-t})'}\left(\widetilde{\boldsymbol{R}}_j - \widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(\boldsymbol{-t})}\boldsymbol{\phi}\right)\right) \tag{A.23}$$

$$\mathbb{E}\left(m_3(\boldsymbol{Z},\boldsymbol{\beta^Y})\right) = \mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta^Y}\right)\right) \tag{A.24}$$

$$\mathbb{E}\left(g(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)\right) = \mathbb{E}\left(\boldsymbol{\phi'}\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(\boldsymbol{-t})'}\left(\widetilde{\boldsymbol{Y}}_j - \kappa\widetilde{\boldsymbol{R}}_{\boldsymbol{j}}^{(\boldsymbol{-t})}\boldsymbol{\phi}\right)\right) \tag{A.25}$$

where $\boldsymbol{Z_j} = \left(\boldsymbol{X_j},\boldsymbol{R_j^{obs}},\boldsymbol{Y_j^{obs}}\right)$, $\boldsymbol{Z}$ stacks the $\boldsymbol{Z}_j$, and $m_1(), m_2(), m_3(), g()$ are functions. Let

$$\widetilde{g}(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa) = \left[m_1(\boldsymbol{Z},\boldsymbol{\beta})', m_2(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\phi})', m_3(\boldsymbol{Z},\boldsymbol{\beta^Y})', g(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)'\right]' \tag{A.26}$$

be a vector stacking the four functions.

A GMM estimator is asymptotically normal if it is consistent and conditions (i)-(v) of Theorem 3.4 of Newey and McFadden (1994) are satisfied.

The conditions are the following:

1. $(\boldsymbol{\beta_0},\boldsymbol{\phi_0},\boldsymbol{\beta_0^Y},\kappa_0)$ is in the interior of $\boldsymbol{\Theta}$.

2. $\widetilde{g}(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)$ is continuously differentiable in a neighborhood $\mathcal{N}$ of $(\boldsymbol{\beta_0},\boldsymbol{\phi_0},\boldsymbol{\beta_0^Y},\kappa_0)$.

3. $\mathbb{E}[\widetilde{g}(\boldsymbol{Z},\boldsymbol{\beta_0},\boldsymbol{\phi_0},\boldsymbol{\beta_0^Y},\kappa_0)] = 0$ and $\mathbb{E}[||\widetilde{g}(\boldsymbol{Z},\boldsymbol{\beta_0},\boldsymbol{\phi_0},\boldsymbol{\beta_0^Y},\kappa_0)||^2]$ is finite.

4. $\mathbb{E}\left[\underset{(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)\in\mathcal{N}}{sup}||\nabla\widetilde{g}(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)||\right] < \infty.$

5. $\widetilde{G}'\widetilde{G}$ is non singular for $\widetilde{G} = \mathbb{E}\left[\nabla_{(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)}\widetilde{g}(\boldsymbol{Z},\boldsymbol{\beta_0},\boldsymbol{\phi_0},\boldsymbol{\beta_0^Y},\kappa_0)\right]$

Under Assumptions 1, 2, and A.6.1, the GMM estimator is a consistent estimator of $(\boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)$.

Condition (i) of Theorem 3.4 of Newey and McFadden (1994) holds under Point 1 of Assumption A.6.1. Condition (ii) holds by inspection. The first part of condition (iii) is shown to hold in the consistency proof, and for the second part note that:

$$\mathbb{E}[||\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)||^2] = \mathbb{E}\left(trace\left(\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)'\right)\right).$$

Therefore for $\mathbb{E}[||\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)||^2]$ to be finite we need that:

$$\begin{aligned}
&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right) < \infty\\
&\mathbb{E}\left(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)'}}\left(\boldsymbol{R}_j - \boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}\boldsymbol{\phi_0}\right)\left(\boldsymbol{R}_j - \boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}\boldsymbol{\phi_0}\right)'\boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}\right) < \infty\\
&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{Y}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0^Y}\right)\left(\ddot{\boldsymbol{Y}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0^Y}\right)'\ddot{\boldsymbol{X}}_j\right) < \infty\\
&\mathbb{E}\left(\boldsymbol{\phi_0'}\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)'}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}\boldsymbol{\phi_0}\right)\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}\boldsymbol{\phi_0}\right)'\boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}\boldsymbol{\phi_0}\right) < \infty.
\end{aligned}$$

which holds by Point 1 of Assumption A.6.2.

For condition (iv) we need to show that: $\mathbb{E}\left[\underset{(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)\in\mathcal{N}}{sup}||\nabla\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\beta^Y}, \kappa)||\right] < \infty$

$$\begin{aligned}
&\mathbb{E}\left[\underset{(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)\in\mathcal{N}}{sup}||\nabla\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\beta^Y}, \kappa)||\right] =\\
&\mathbb{E}\left(\underset{(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)\in\mathcal{N}}{sup}\sqrt{trace\left(\nabla\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\beta^Y}, \kappa)\nabla\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\beta^Y}, \kappa)'\right)}\right)
\end{aligned}$$

by the triangle inequality applied to the Euclidean norm, a sufficient condition for the quantity above to be finite, is that the sum of absolute values of the diagonal elements be finite. So we need that:

$$\begin{aligned}
&\mathbb{E}\left|\underset{(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)\in\mathcal{N}}{sup}\left(\boldsymbol{\phi'}\boldsymbol{\widetilde{R}_j^{(-t)'}}\boldsymbol{\widetilde{R}_j^{(-t)}}\boldsymbol{\phi}\right)\right| < \infty\\
&\mathbb{E}\left|\underset{(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)\in\mathcal{N}}{sup}\left(\boldsymbol{\widetilde{R}_j^{(-t)'}}\boldsymbol{\widetilde{R}_j^{(-t)}}\right)\right| < \infty\\
&\mathbb{E}\left|\left(\boldsymbol{\ddot{X}}_j'\boldsymbol{\ddot{X}}_j\right)\right| < \infty.
\end{aligned}$$

Now by Points 1 of Assumption 2 we have $\mathbb{E}\left|\left(\boldsymbol{\ddot{X}}_j'\boldsymbol{\ddot{X}}_j\right)\right| < \infty$. The other two hold by taking the neighborhood $\mathcal{N}$ to be $\boldsymbol{\Theta}$ and then using Point 2 of Assumption A.6.2. Condition (v) is satisfied under Point 1 of Assumption 2 and Point 3 Assumption A.6.2. Furthermore we have $\widehat{W} = W = I$, such that $\left(\widetilde{G}'I\widetilde{G}\right)^{-1}\widetilde{G}' = \widetilde{G}^{-1}$, therefore the asymptotic variance of the estimator becomes:

$$\begin{aligned}
&\left(\widetilde{G}'I\widetilde{G}\right)^{-1}\widetilde{G}'IE[\widetilde{g}(\boldsymbol{Z},\boldsymbol{\beta_0},\boldsymbol{\phi_0},\boldsymbol{\beta_0^Y},\kappa_0)\widetilde{g}(\boldsymbol{Z},\boldsymbol{\beta_0},\boldsymbol{\phi_0},\boldsymbol{\beta_0^Y},\kappa_0)']I\widetilde{G}\left(\widetilde{G}'I\widetilde{G}\right)^{-1}\\
&=\widetilde{G}^{-1}E[\widetilde{g}(\boldsymbol{Z},\boldsymbol{\beta_0},\boldsymbol{\phi_0},\boldsymbol{\beta_0^Y},\kappa_0)\widetilde{g}(\boldsymbol{Z},\boldsymbol{\beta_0},\boldsymbol{\phi_0},\boldsymbol{\beta_0^Y},\kappa_0)']\widetilde{G}^{-1'}
\end{aligned}$$

Finally the consistency of $\widehat{\Omega}$ follows directly from Theorem 4.5 of Newey and McFadden (1994). We need to only check that $\widetilde{g}(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)$ is continuous in a neighborhood $\mathcal{N}$ of $(\boldsymbol{\beta_0},\boldsymbol{\phi_0},\boldsymbol{\beta_0^Y},\kappa_0)$ with probability one, which is satisfied by inspection. And a fourth moment existence condition $\mathbb{E}[sup_{(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)\in\mathcal{N}}||\widetilde{g}(\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\beta^Y},\kappa)||^2] < \infty$ which holds by Point 4 of Assumption A.6.2 and taking the neighborhood $\mathcal{N}$ to be $\boldsymbol{\Theta}$.

Q.E.D

A.9.5 Proof of Theorem 1

Let:

$$
\begin{aligned}
G_\kappa &= \mathbb{E}[\nabla_\kappa g(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)] = \mathbb{E}\left[-\left(\boldsymbol{\phi_0' R_j^{(-t)'} R_j^{(-t)} \phi_0}\right)\right] \\
G_{\beta^Y} &= \mathbb{E}[\nabla_{\beta^Y} g(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)] = \mathbb{E}\left[-\left(\boldsymbol{\phi_0' R_j^{(-t)'} X_j}\right)\right] \\
G_\phi &= \mathbb{E}[\nabla_\phi g(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)] = \mathbb{E}\left[\left(\boldsymbol{Y_j' R_j^{(-t)} - 2\kappa_0 \phi_0' R_j^{(-t)'} R_j^{(-t)}}\right)\right] \\
G_\beta &= \mathbb{E}[\nabla_\beta g(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)] = \mathbb{E}\left[-\left(\boldsymbol{Y_j' A - 2\kappa_0 \phi_0' R_j^{(-t)'} A}\right)\right] \\
M_1 &= M_3 = \mathbb{E}\left[-\left(\ddot{\boldsymbol{X}}_j' \ddot{\boldsymbol{X}}_j\right)\right] \qquad M_{2\phi} = \mathbb{E}\left[-\left(\boldsymbol{R_j^{(-t)'} R_j^{(-t)}}\right)\right] \\
M_{2\beta} &= \mathbb{E}[\nabla_\beta m_2(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0})] = \mathbb{E}\left[-\left(\boldsymbol{R_j^{(-t)'} X_j + \mathcal{X}_j^{(-t)} - \widetilde{\mathcal{X}}_j^{(-t)} - R_j^{(-t)'} A}\right)\right]
\end{aligned}
$$

where $\boldsymbol{Y}_j$ is a vector stacking $\overline{Y}_{jt}$, $R_{jt} = \overline{R}_{jt}^{obs} - \boldsymbol{\overline{X}'_{jt}\beta_0}$, $\boldsymbol{R}_j$ is a vector stacking all R_{jt} for teacher j, $\boldsymbol{R_j^{(-t)}}$ is a matrix stacking T row vectors (of dimension $1 \times (T-1)$) with each row (indexed by t) containing $T-1$ different R_{jk} for $k \neq t$. A is a $T \times K$ matrix such that each row consists of $(\phi_1 \overline{X}_{jt-1} + \phi_2 \overline{X}_{jt-2} + ...)$. $\boldsymbol{X_j^{(-t)}}$ is a $1 \times (T-1)$ block matrix stacking $T \times K$ matrices of the covariates for teacher j in all years except year t. $\mathcal{X}_j^{(-t)}$ is a $T-1 \times K$ matrix where each row consists of $\boldsymbol{R}_j'$ multiplied by a submatrix of $\boldsymbol{X_j^{(-t)}}$. $\widetilde{\mathcal{X}}_j^{(-t)}$ is a $T-1 \times K$ matrix where each row $\boldsymbol{\phi_0' R_j^{(-t)'}}$ multiplied by a submatrix of $\boldsymbol{X_j^{(-t)}}$.

Furthermore let:

$$
\begin{aligned}
\psi_1(\boldsymbol{Z}) &= -M_1^{-1} m_1(\boldsymbol{Z}, \boldsymbol{\beta_0}) \\
\psi_2(\boldsymbol{Z}) &= -M_{2\phi}^{-1} m_2(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}) \\
\psi_3(\boldsymbol{Z}) &= -M_3^{-1} m_3(\boldsymbol{Z}, \boldsymbol{\beta_0^Y}) \\
g(\boldsymbol{Z}) &= g(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0).
\end{aligned}
$$

Then we have:

by TheoremA.7.1 we have that:

$$\sqrt{J}\left[(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\beta}}^{\boldsymbol{Y}}, \widehat{\kappa}) - (\boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)\right] \rightsquigarrow$$
$$N\left(0, \widetilde{G}^{-1}\mathbb{E}[\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)']\widetilde{G}^{-1'}\right).$$

Now note that the last row of $\widetilde{G}^{-1}$ is $G_\kappa^{-1}[1 \quad G_{\beta^Y}M_3^{-1} \quad -G_\phi M_{2\phi}^{-1} \quad -G_\beta M_1^{-1} + G_\phi M_{2\phi}^{-1}M_{2\beta}M_1^{-1}]$. Multiplying that by $\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}, \boldsymbol{\beta_0^Y}, \kappa_0)$ we get:

$$\begin{aligned}
&G_\kappa^{-1}\big(g(\boldsymbol{Z}) - G_{\beta^Y}M_3^{-1}m_3(\boldsymbol{Z}, \boldsymbol{\beta_0^Y}) - G_\phi M_{2\phi}^{-1}m_2(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\phi_0}) - G_\beta M_1^{-1}m_1(\boldsymbol{Z}, \boldsymbol{\beta_0}) \\
&+ G_\phi M_{2\phi}^{-1}M_{2\beta}M_1^{-1}m_1(\boldsymbol{Z}, \boldsymbol{\beta_0})\big) \\
&= (G_\kappa^{-1})\left(g(\boldsymbol{Z}) + G_{\beta^Y}\psi_3(\boldsymbol{Z}) + G_\phi\psi_2(\boldsymbol{Z}) + G_\beta\psi_1(\boldsymbol{Z}) - G_\phi M_{2\phi}^{-1}M_{2\beta}\psi_1(\boldsymbol{Z})\right). \qquad \text{(A.27)}
\end{aligned}$$

Now note that the asymptotic variance of $\widehat{\kappa}$ would be the lower right block of the joint variance matrix. Given that the quantity in (A.27) is a scalar, we can square it to obtain the lower right block of the joint variance matrix and then:

$$\sqrt{J}(\widehat{\kappa} - \kappa_0) \rightsquigarrow N\left(0, \sigma^2\right) \qquad \text{(A.28)}$$

where $\sigma^2 = (G_\kappa^{-1})^2\mathbb{E}\left(\left(g(\boldsymbol{Z}) + G_{\beta^Y}\psi_3(\boldsymbol{Z}) + G_\phi\psi_2(\boldsymbol{Z}) + G_\beta\psi_1(\boldsymbol{Z}) - G_\phi M_{2\phi}^{-1}M_{2\beta}\psi_1(\boldsymbol{Z})\right)^2\right)$.

Q.E.D

A.9.6 Proof of Lemma A.7.2

Under Assumptions 1, 2, and A.6.3, the proof of Point 1 is similar to the proofs for Lemma A.7.1 and TheoremA.7.1. The proof of Point 2 is similar to the final part of the proof of Theorem A.7.1.

A.9.7 Proof of Theorem A.7.2

Let:

$$\mathbb{E}\left(m_1(\boldsymbol{Z},\boldsymbol{\beta_0})\right) = E\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right) \tag{A.29}$$

$$\mathbb{E}\left(m_3(\boldsymbol{Z},\boldsymbol{\beta_0^Y})\right) = \mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\mathbf{Y}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0^Y}\right)\right) \tag{A.30}$$

$$\mathbb{E}\left(g_1(\boldsymbol{Z},\boldsymbol{\beta_0,\beta_0^Y},\kappa_0)\right) = \mathbb{E}\left(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{\boldsymbol{(-t)'}}\left(\mathbf{Y}_j - \boldsymbol{a} - \kappa_0\boldsymbol{R_j}\right)\right) \tag{A.31}$$

where $m_1(), m_3(), g_1()$ are functions. Let:

$$\widetilde{g}_1(\boldsymbol{Z},\boldsymbol{\beta_0,\beta_0^Y},\kappa_0) = \left[m_1(\boldsymbol{Z},\boldsymbol{\beta_0})', m_3(\boldsymbol{Z},\boldsymbol{\beta_0^Y})', g_1(\boldsymbol{Z},\boldsymbol{\beta_0,\beta_0^Y},\kappa_0)'\right]' \tag{A.32}$$

The proof to show that: $\sqrt{J}\left[(\boldsymbol{\widehat{\beta}^*,\widehat{\beta}^{Y*}},\widehat{\kappa}^*) - (\boldsymbol{\beta_0,\beta_0^Y},\kappa_0)\right] \rightsquigarrow N\left(0,\Omega^*\right)$ is similar to the proof for Lemma A.7.1 and Theorem A.7.1. We need only that $\widehat{W^*} \xrightarrow{p} W^*$ and that W^* be invertible, these conditions are guaranteed by Lemma A.7.2, Point 1 of Assumption 2, and Points 2 and 5 of Assumption A.6.3.

$\Omega^* \leq \Omega_2$ by Theorem 3.2 of Hansen (1982).

A.9.8 Proof of Result A.7.1

The result follows from Lemmas 4.1 and 4.2 of Hansen (1982).

Note that the first two moment conditions are exactly identified. Then note that for the last moment condition we have one parameter, κ_0, and $T-1$ possible instruments such that the degrees of freedom of the χ^2 distribution will be $T-1-1=T-2$.

A.9.9 Proof of Result 4

We have established in Theorem A.7.2 that the optimal GMM estimator has an asymptotic variance of $\Omega^* = \left(\widetilde{G}_1' W^* \widetilde{G}_1\right)^{-1}$ where, $W^* = \mathbb{E}[\widetilde{g}_1(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \kappa_0)\widetilde{g}(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \kappa_0)']^{-1}$ and $\widetilde{G}_1 = \mathbb{E}\left[\nabla_{(\boldsymbol{\beta}, \boldsymbol{\beta^Y}, \kappa)}\widetilde{g}_1(\boldsymbol{Z}, \boldsymbol{\beta_0}, \boldsymbol{\beta_0^Y}, \kappa_0)\right]$. We can write our moment conditions as $\mathbb{E}[\boldsymbol{Q'u}]$ where $\boldsymbol{Q}$ is a block diagonal matrix with blocks $\ddot{\boldsymbol{X}}_{\boldsymbol{j}}, \ddot{\boldsymbol{X}}_{\boldsymbol{j}}, \boldsymbol{R}_{\boldsymbol{j}}^{(-t)}$ and $\boldsymbol{u} = \begin{pmatrix} \ddot{\boldsymbol{\epsilon}}_{\boldsymbol{j}} + \ddot{\boldsymbol{\mu}}_{\boldsymbol{j}} \\ \ddot{\boldsymbol{\eta}}_{\boldsymbol{j}} + \kappa_0 \ddot{\boldsymbol{\mu}}_{\boldsymbol{j}} \\ \boldsymbol{\eta}_{\boldsymbol{j}} - \kappa_0 \boldsymbol{\epsilon}_{\boldsymbol{j}} \end{pmatrix}$.

Note that under Assumption 4 we have:

$$\begin{aligned}
\widetilde{G}_1 &= \begin{pmatrix} -\mathbb{E}(\ddot{\boldsymbol{X}}_{\boldsymbol{j}}'\ddot{\boldsymbol{X}}_{\boldsymbol{j}}) & 0 & 0 \\ 0 & -\mathbb{E}(\ddot{\boldsymbol{X}}_{\boldsymbol{j}}'\ddot{\boldsymbol{X}}_{\boldsymbol{j}}) & 0 \\ -\mathbb{E}(\boldsymbol{B}_{\boldsymbol{j}} + \kappa_0 \overline{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)'} \boldsymbol{X}_j) & -\mathbb{E}(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)'} \boldsymbol{X}_j) & -\mathbb{E}(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)'} \boldsymbol{R}_{\boldsymbol{j}}) \end{pmatrix} \\
&= \begin{pmatrix} -\mathbb{E}(\ddot{\boldsymbol{X}}_{\boldsymbol{j}}'\ddot{\boldsymbol{X}}_{\boldsymbol{j}}) & 0 & 0 \\ 0 & -\mathbb{E}(\ddot{\boldsymbol{X}}_{\boldsymbol{j}}'\ddot{\boldsymbol{X}}_{\boldsymbol{j}}) & 0 \\ 0 & 0 & -\mathbb{E}(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)'} \boldsymbol{R}_{\boldsymbol{j}}) \end{pmatrix}
\end{aligned}$$

where $\boldsymbol{B}_{\boldsymbol{j}}$ is a matrix where each row consists of a column of $\boldsymbol{X}_{\boldsymbol{j}}^{(-t)}$ multiplied by

$(\boldsymbol{Y}_j - \kappa_0 \boldsymbol{R_j}) = (\boldsymbol{\eta}_j - \kappa_0 \boldsymbol{\epsilon}_j)$. The second equality then follows from Assumption 4 which makes all off-diagonal elements zero.

Then we can write: $\widetilde{G}_1 = -\mathbb{E}[\boldsymbol{Q'L}] = -\mathbb{E}[\boldsymbol{Q'}\mathbb{E}[\boldsymbol{L}|\boldsymbol{Q}]]$ where $\boldsymbol{L}$ is a block diagonal matrix with blocks $\ddot{\boldsymbol{X}}_{\boldsymbol{j}}, \ddot{\boldsymbol{X}}_{\boldsymbol{j}}, \boldsymbol{R}_j$ and $W^* = \mathbb{E}[\boldsymbol{Q'uu'Q}] = \mathbb{E}[\boldsymbol{Q'}\mathbb{E}[\boldsymbol{uu'}|\boldsymbol{Q}]\boldsymbol{Q}]$. Then the variance from the optimal GMM estimator will be:

$$\Omega^* = \left(\mathbb{E}[\boldsymbol{Q'}\mathbb{E}[\boldsymbol{L}|\boldsymbol{Q}]]'\mathbb{E}[\boldsymbol{Q'}\mathbb{E}[\boldsymbol{uu'}|\boldsymbol{Q}]\boldsymbol{Q}]^{-1}\mathbb{E}[\boldsymbol{Q'}\mathbb{E}[\boldsymbol{L}|\boldsymbol{Q}]]\right)^{-1}$$

and the optimal instrument following Chamberlain (1987) are:

$$\boldsymbol{Q}^* = \mathbb{E}[\boldsymbol{uu'}|\boldsymbol{Q}]^{-1}\mathbb{E}[\boldsymbol{L}|\boldsymbol{Q}] \tag{A.33}$$

and yield an asymptotic variance of

$$\Omega^* = \left(\mathbb{E}\left(\mathbb{E}[\boldsymbol{L}|\boldsymbol{Q}]'\mathbb{E}[\boldsymbol{uu'}|\boldsymbol{Q}]^{-1}\mathbb{E}[\boldsymbol{L}|\boldsymbol{Q}]\right)\right)^{-1}$$

where the moment conditions will then be:

$$\mathbb{E}[\boldsymbol{(Q^*)'u}] = \mathbb{E}[(\mathbb{E}[\boldsymbol{uu'}|\boldsymbol{Q}]^{-1}\mathbb{E}[\boldsymbol{L}|\boldsymbol{Q}])'\boldsymbol{u}] \tag{A.34}$$

where:

$$\mathbb{E}[\boldsymbol{L}|\boldsymbol{Q}] = \begin{pmatrix} \ddot{\boldsymbol{X}}_{\boldsymbol{j}} & 0 & 0 \\ 0 & \ddot{\boldsymbol{X}}_{\boldsymbol{j}} & 0 \\ 0 & 0 & \mathbb{E}[\boldsymbol{R}_j|\boldsymbol{Q}] \end{pmatrix}$$

and note that $\mathbb{E}[\boldsymbol{R}_j|\boldsymbol{Q}] = \mathbb{E}[\boldsymbol{\mu}_j|\boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}] + \mathbb{E}[\boldsymbol{\epsilon}_j|\ddot{\boldsymbol{X}}_{\boldsymbol{j}}]$ because $\boldsymbol{X}_j$ is independent of $\boldsymbol{\mu_j}$ by Assumption 4, and $\boldsymbol{\epsilon}_j$ is independent of $\boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{(-t)}}$ by point 2 of Assumption 4 and by point 3 of Assumption 1. Finally note that $\mathbb{E}[\boldsymbol{\epsilon}_j|\ddot{\boldsymbol{X}}_{\boldsymbol{j}}] = 0$ by point 3 of Assumption 4.

A.9.10 Proof of Result 5

Starting with $\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right) = 0$:

$$\begin{aligned}
&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right)\\
=&\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{\mu}}_{\boldsymbol{j}} + \ddot{\boldsymbol{\epsilon}}_{\boldsymbol{j}}\right)\right)\\
=&0
\end{aligned}$$

where the first equality follows from (1.2), and the second equality follows from Point 3 of Assumption 2. Then by Point 1 of Assumption 2, we have that $\mathbb{E}(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{X}}_j)$ is invertible, therefore $\boldsymbol{\beta_0}$ is uniquely identified.

Now moving to $\mathbb{E}\left(\boldsymbol{D_j'}\left(\boldsymbol{R}_j - \boldsymbol{D_j}\boldsymbol{\alpha_0}\right)\right) = 0$:

$$\begin{aligned}
&\mathbb{E}\left(\boldsymbol{D_j'}\left(\boldsymbol{R}_j - \boldsymbol{D_j}\boldsymbol{\alpha_0}\right)\right)\\
=&\mathbb{E}\left(\boldsymbol{D_j'}\left(\boldsymbol{\mu_j} + \boldsymbol{\epsilon_j} - \boldsymbol{D_j}\boldsymbol{\alpha_0}\right)\right)\\
=&\mathbb{E}\left(\boldsymbol{D_j'}\left(\boldsymbol{\zeta_j} + \boldsymbol{\epsilon_j}\right)\right)\\
=&0
\end{aligned}$$

where the first equality follows from (1.2), the second equality follows from (1.23). The third equality follows from the fact that ζ_{jt} is orthogonal to D_{jt} by construction and by Point 2 of Assumption 5. Then by Point 1 of Assumption 5, we have that $\mathbb{E}(\boldsymbol{D}_j'\boldsymbol{D}_j)$ is invertible, therefore $\boldsymbol{\alpha_0}$ is uniquely identified.

A.9.11 Proof of Theorem A.7.3

Note that $\widehat{W} = W = I$ and consistency of the GMM estimators follows using a similar proof to Lemma A.7.1.

Now to show asymptotic normality we start by rewriting the moments as:

$$\mathbb{E}\left(m_1(\boldsymbol{Z},\boldsymbol{\beta})\right) = E\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta}\right)\right) \tag{A.35}$$

$$\mathbb{E}\left(m_4(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\alpha})\right) = \mathbb{E}\left(\boldsymbol{D}_{\boldsymbol{j}}'\left(\widetilde{\boldsymbol{R}}_{\boldsymbol{j}} - \boldsymbol{D}_{\boldsymbol{j}}\boldsymbol{\alpha}\right)\right) \tag{A.36}$$

where $\boldsymbol{Z}_{\boldsymbol{j}} = \left(\boldsymbol{X}_{\boldsymbol{j}}, \boldsymbol{R}_{\boldsymbol{j}}^{\boldsymbol{obs}}, \boldsymbol{D}_{\boldsymbol{j}}\right)$, $\boldsymbol{Z}_2$ stacks the $\boldsymbol{Z}_j$, and $m_1(), m_4()$ are functions. Let

$$\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta},\boldsymbol{\alpha}) = [m_1(\boldsymbol{Z}_2,\boldsymbol{\beta})', m_4(\boldsymbol{Z}_2,\boldsymbol{\beta},\boldsymbol{\alpha})']' \tag{A.37}$$

be a vector stacking the two functions.

A GMM estimator is asymptotically normal if it's consistent and conditions (i)-(v) of Theorem 3.4 of Newey and McFadden (1994) are satisfied.

Condition (i) holds under Point 1 of Assumption A.6.4. Condition (ii) holds by inspection. The first part of condition (iii) holds by Result 5 and the second part: second part note that:

$$\mathbb{E}[||\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta_0},\boldsymbol{\alpha_0})||^2] = \mathbb{E}\left(trace\left(\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta_0},\boldsymbol{\alpha_0})\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta_0},\boldsymbol{\alpha_0})'\right)\right).$$

Therefore for $\mathbb{E}[||\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta_0},\boldsymbol{\alpha_0})||^2]$ to be finite we need that:

$$\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right) < \infty$$

$$\mathbb{E}\left(\boldsymbol{D}_{\boldsymbol{j}}'\left(\boldsymbol{R}_j - \boldsymbol{D}_{\boldsymbol{j}}\boldsymbol{\alpha_0}\right)\left(\boldsymbol{R}_j - \boldsymbol{D}_{\boldsymbol{j}}\boldsymbol{\alpha_0}\right)'\boldsymbol{D}_{\boldsymbol{j}}\right) < \infty$$

which holds by Point 4 of Assumption A.6.4.

For condition (iv) we need to show that: $\mathbb{E}\left[\underset{(\boldsymbol{\beta},\boldsymbol{\alpha})\in\mathcal{N}}{sup}||\nabla\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta},\boldsymbol{\alpha})||\right] < \infty$ for a neighborhood $\mathcal{N}$ around $(\boldsymbol{\beta_0},\boldsymbol{\alpha_0})$.

$$\mathbb{E}\left[\underset{(\boldsymbol{\beta},\boldsymbol{\alpha})\in\mathcal{N}}{sup}||\nabla\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta},\boldsymbol{\alpha})||\right] = \mathbb{E}\left(\underset{(\boldsymbol{\beta},\boldsymbol{\alpha})\in\mathcal{N}}{sup}\sqrt{trace\left(\nabla\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta},\boldsymbol{\alpha})\nabla\widetilde{g}_2(\boldsymbol{Z}_2,\boldsymbol{\beta},\boldsymbol{\alpha})'\right)}\right)$$

Since by the triangle inequality applied to the Euclidean norm, a sufficient condition for the quantity above to be finite, is that the sum of absolute values of the diagonal elements be finite. So we need that:

$$\mathbb{E}\left|\left(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{X}}_j\right)\right| < \infty$$
$$\mathbb{E}\left|\left(\boldsymbol{D}_j'\boldsymbol{D}_j\right)\right| < \infty.$$

these conditions are guaranteed by Point 1 of Assumption A.6.4 and Point 1 of Assumption 2.

Condition (v) is satisfied under Point 1 of Assumption 2 and Point 1 Assumption 5.

A.9.12 Proof of Theorem 3

The asymptotic variance of $\boldsymbol{\alpha}$ can be obtained from partitioned matrix inversion. Let:

$$\widetilde{G}_\alpha = -\mathbb{E}\left(\boldsymbol{D}_j'\boldsymbol{D}_j\right)$$
$$\widetilde{G}_\beta = -\mathbb{E}\left(\boldsymbol{D}_j'\boldsymbol{X}_j\right)$$
$$\widetilde{M}_1 = -\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{X}}_j\right)$$

Now note that the second row of $\widetilde{G}_2^{-1}$ is $[-\widetilde{G}_\alpha^{-1}\widetilde{G}_\beta\widetilde{M}_1^{-1}, \widetilde{G}_\alpha^{-1}]$ which can be written as $\widetilde{G}_\alpha^{-1}[-\widetilde{G}_\beta\widetilde{M}_1^{-1}, I]$. Multiplying that by $\widetilde{g}_2(\boldsymbol{Z}_2, \boldsymbol{\beta_0}, \boldsymbol{\alpha_0})$ we get:

$$\begin{aligned}
&\widetilde{G}_\alpha^{-1}\left(\boldsymbol{D_j'}\left(\boldsymbol{R}_j - \boldsymbol{D_j}\boldsymbol{\alpha_0}\right) - \widetilde{G}_\beta\widetilde{M}_1^{-1}\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right)\\
=& -\mathbb{E}\left(\boldsymbol{D}_j'\boldsymbol{D}_j\right)^{-1}\left(\boldsymbol{D_j'}\left(\boldsymbol{R}_j - \boldsymbol{D_j}\boldsymbol{\alpha_0}\right) - \mathbb{E}\left(\boldsymbol{D_j'}\boldsymbol{X_j}\right)\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{X}}_j\right)^{-1}\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right)
\end{aligned}$$

Taking the quadratic of the above with
$\Gamma = \left(\boldsymbol{D_j'}\left(\boldsymbol{R}_j - \boldsymbol{D_j}\boldsymbol{\alpha_0}\right) - \mathbb{E}\left(\boldsymbol{D_j'}\boldsymbol{X_j}\right)\mathbb{E}\left(\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{X}}_j\right)^{-1}\ddot{\boldsymbol{X}}_j'\left(\ddot{\boldsymbol{R}}_j^{obs} - \ddot{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{\beta_0}\right)\right)$ yields:

$$V_1 = \mathbb{E}(\boldsymbol{D}_j'\boldsymbol{D}_j)^{-1}\mathbb{E}\left(\Gamma\Gamma'\right)\mathbb{E}(\boldsymbol{D}_j'\boldsymbol{D}_j)^{-1'}$$

A.9.13 Algebra for Corollary 1

$$
\begin{aligned}
G_\phi &= \mathbb{E}\left[\left(\boldsymbol{Y_j'}\boldsymbol{R_j^{(-t)}} - \boldsymbol{2\kappa_0\phi_0'}\boldsymbol{R_j^{(-t)\prime}}\boldsymbol{R_j^{(-t)}}\right)\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{Y_j'} - \boldsymbol{2\kappa_0\phi_0'}\boldsymbol{R_j^{(-t)\prime}}\right)R_j^{(-t)}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{\eta_j} + \boldsymbol{\kappa_0\theta_j} - \boldsymbol{\kappa_0\epsilon_j} - \boldsymbol{\kappa_0\phi_0'}\boldsymbol{R_j^{(-t)\prime}}\right)\boldsymbol{R_j^{(-t)}}\right] \\
&= -\kappa_0\boldsymbol{\phi_0'}\mathbb{E}\left(\boldsymbol{R_j^{(-t)\prime}}\boldsymbol{R_j^{(-t)}}\right)
\end{aligned}
$$

A.9.14 Proof of Overidentification

To see that note that (1.14) is equivalent to:

$$
\begin{aligned}
&\boldsymbol{\phi_0'}\mathbb{E}\left(\boldsymbol{\overline{R}_j^{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R_j^{(-t)}}\boldsymbol{\phi_0}\right)\right) \\
=&\boldsymbol{\phi_0'}\mathbb{E}\left(\boldsymbol{\overline{R}_j^{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R_j} + \boldsymbol{\kappa_0\theta_j}\right)\right) \\
=&\boldsymbol{\phi_0'}\mathbb{E}\left(\boldsymbol{\overline{R}_j^{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R_j}\right)\right) = 0
\end{aligned}
$$

where the first equality follows from (1.15) and the second equality follows from the fact that $\boldsymbol{R_j^{(-t)}}$ and $\boldsymbol{\theta_j}$ are uncorrelated by construction. Notably, (1.14) only requires that a linear combination of $\mathbb{E}\left(\boldsymbol{\overline{R}_j^{(-t)\prime}}\left(\boldsymbol{Y}_j - \kappa_0\boldsymbol{R_j}\right)\right)$ be equal to zero. However, under Assumptions 1 and 2 we have the stronger conditions:

$$
\begin{aligned}
&\mathbb{E}\left(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\left(\boldsymbol{Y}_j - \kappa_0 \boldsymbol{R}_j\right)\right)\\
=&\mathbb{E}\left(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\boldsymbol{\eta}_{\boldsymbol{j}} - \kappa_0 \overline{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\boldsymbol{\epsilon}_{\boldsymbol{j}}\right)\\
=&0
\end{aligned}
$$

where the first equality follows from (1.3) and (1.8). The second equality follows from the fact that:

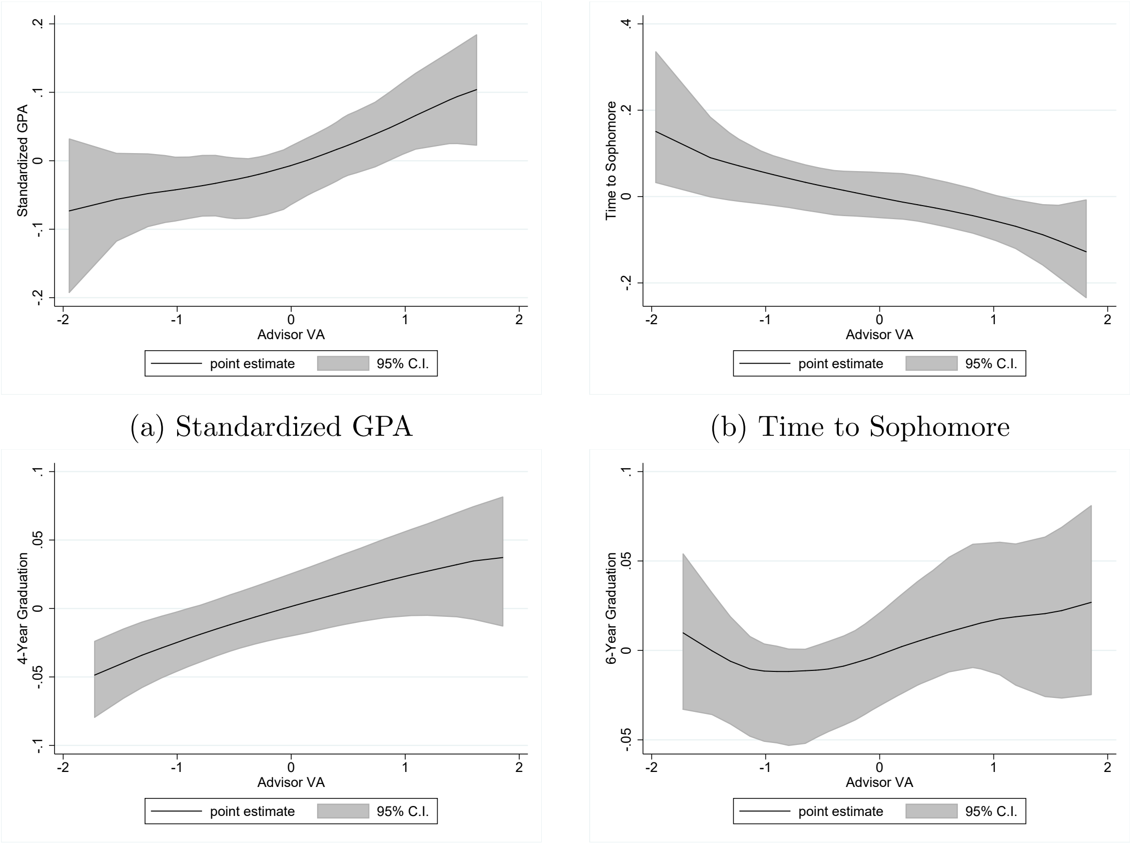
$$
\begin{aligned}
&\mathbb{E}\left(\boldsymbol{R}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\eta}_{\boldsymbol{j}}\right)\\
=&\mathbb{E}\left((\boldsymbol{\mu}_{\boldsymbol{j}}^{(-t)} + \boldsymbol{\epsilon}_{\boldsymbol{j}}^{(-t)})\boldsymbol{\eta}_{\boldsymbol{j}}\right)\\
=&\mathbb{E}\left(\boldsymbol{\mu}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\eta}_{\boldsymbol{j}}\right) + \mathbb{E}\left(\boldsymbol{\epsilon}_{\boldsymbol{j}}^{(-t)}\boldsymbol{\eta}_{\boldsymbol{j}}\right)\\
=&0
\end{aligned}
$$

where analogously to $\boldsymbol{R}_{\boldsymbol{j}}^{(-t)}$, each row in $\boldsymbol{\mu}_{\boldsymbol{j}}^{(-t)}$ stacks μ_{js} for each t with $s \neq t$ and each row in $\boldsymbol{\epsilon}_{\boldsymbol{j}}^{(-t)}$ stacks $\overline{\epsilon}_{js}$ for each t with $s \neq t$. The equalities then follow from $\mathbb{E}\left(\mu_{js}\overline{\eta}_{jt}\right) = 0$ by Assumption 1 and $\mathbb{E}\left(\overline{\epsilon}_{js}\overline{\eta}_{jt}\right) = 0$ by Assumption 1. A similar argument shows $\mathbb{E}\left(\overline{\boldsymbol{R}}_{\boldsymbol{j}}^{(-t)\prime}\boldsymbol{\epsilon}_{\boldsymbol{j}}\right)$.

Appendix B

Appendix for “Advisor Value-Added and Student Outcomes: Evidence from Randomly Assigned College Advisors”

B.1 Figures



(a) Standardized GPA

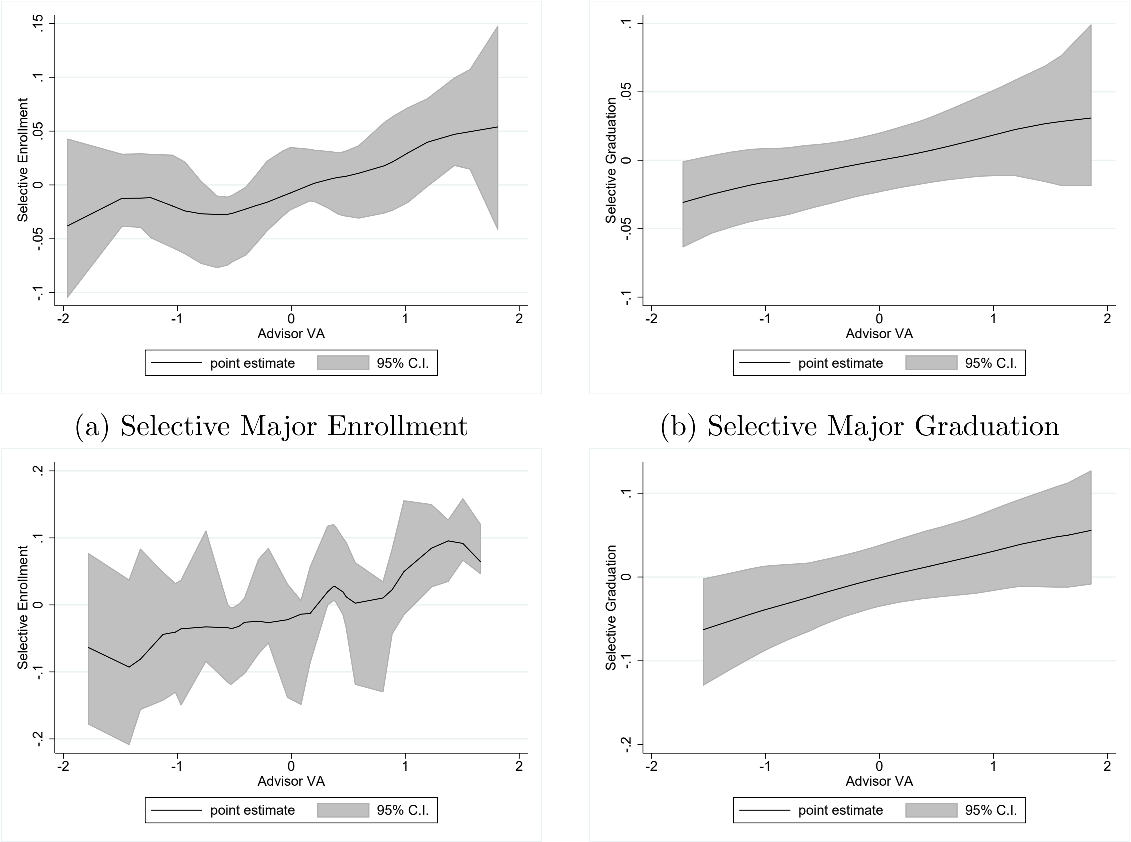
(b) Time to Sophomore

(c) 4-year Graduation

(d) 6-Year Graduation

Figure B.1: Distribution of Freshman Advisor Grade VA Effects on Academic Performance and College Completion

Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2015-2016. The sample is restricted to 2003-2004 to 2012-2013 for graduation outcomes. The figures plot a nonparametric estimate of the conditional mean of outcomes (residualized with respect to a year fixed effect) on standardized VA (residualized with respect to a year fixed effect). Specifically, we estimate a nonparametric regression of residualized outcomes on residualized standardized VA using a local linear regression. We then plot the point-estimates from the nonparametric regression with the corresponding bias-corrected confidence intervals following Calonico, Cattaneo and Farrell (2018).



(a) Selective Major Enrollment

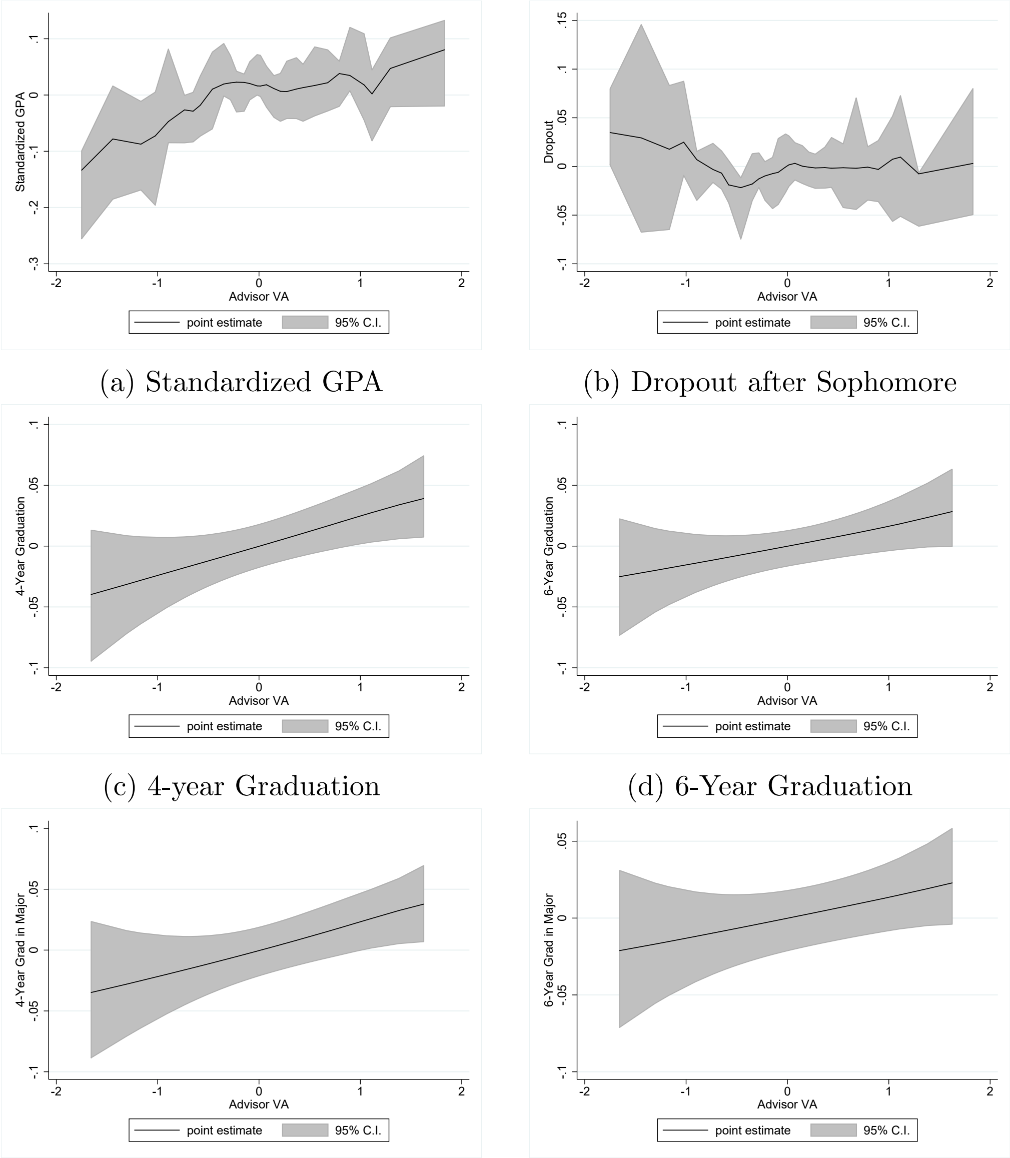
(b) Selective Major Graduation

(c) Top Students' Selective Major Enrollment

(d) Top Students' Selective Major Graduation

Figure B.2: Distribution of Freshman Advisor Grade VA Effects on Selective Major Enrollment and Graduation

Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2015-2016. The sample is restricted to 2003-2004 to 2012-2013 for graduation outcomes. The figures plot a nonparametric estimate of the conditional mean of outcomes (residualized with respect to a year fixed effect) on standardized VA (residualized with respect to a year fixed). Specifically, we estimate a nonparametric regression of residualized outcomes on residualized standardized VA using a local linear regression. We then plot the point-estimates from the nonparametric regression with the corresponding bias-corrected confidence intervals following Calonico, Cattaneo and Farrell (2018).



(a) Standardized GPA

(b) Dropout after Sophomore

(c) 4-year Graduation

(d) 6-Year Graduation

(e) 4-year In-Major Graduation

(f) 6-Year In-Major Graduation

Figure B.3: Distribution of Sophomore Advisor Grade VA Effects on Academic Performance and College Completion

The sample includes first-time enrolling sophomore students from the academic years 2003-2004 to 2015-2016. The sample is restricted to the yeas 2003-2004 to 2012-2013 for graduation outcomes. The figures plot a nonparametric estimate of the conditional mean of outcomes (residualized with respect to year and department fixed effects) on standardized VA (residualized with respect to year and department fixed effects). Specifically, we estimate a nonparametric regression of residualized outcomes on residualized standardized VA using a local linear regression. We then plot the point-estimates from the nonparametric regression with the corresponding bias-corrected confidence intervals following Calonico, Cattaneo and Farrell (2018).

B.2 Tables

Table B.1: Summary Statistics

	Mean (1)	S.D. (2)	Obs. (3)
A. Student Level Covariates			
Female	0.478	0.500	3,857
Math SAT	573	75.5	3,857
Verbal SAT	494	90.0	3,857
Legacy Status	0.202	0.402	3,857
B. Student Level Outcomes			
Freshman GPA	76.5	9.15	3,857
Become a Sophomore	0.794	0.405	3,857
Time to Sophomore	2.480	1.159	3,047
Graduate in 4 years	0.458	0.498	2,952
Graduate in 6 Years	0.575	0.494	2,952
Enroll in Selective Major	0.429	0.495	3,857
Graduate from Selective Major	0.355	0.478	2,952
C. Advisor-Year Level Characteristics			
Female	0.389	0.489	131
Science Department	0.565	0.498	131
Lecturer and Other	0.100	0.300	131
Assistant Professor	0.400	0.491	131
Associate Professor	0.221	0.417	131
Professor	0.282	0.452	131
Number of Students	31.1	7.54	131

Our main sample includes freshman students who first enrolled in AUB in the academic years 2003-2004 to 2015-2016. Data from these years comprise 38 unique advisors. Our graduation sample includes students who first enrolled in AUB in the academic years 2003-2004 to 2012-2013.

Table B.2: Estimate of Forecast Bias of Advisor Grade VA Measure

	Freshman Course Grade
Advisor VA	0.971
	(0.253)
Mean of VA	-0.005
S.D of VA	0.055
Number of Observations	39,369

Standard errors in parentheses are clustered at the advisor-year level. Regression includes year fixed effects. Freshman advisor VA is constructed using a leave-year out estimate as described in the methodology section.

Table B.3: Test of Random Assignment

	Advisor Grade VA
Math SAT	0.0004
	(0.0003)
Verbal SAT	0.0001
	(0.0003)
Female	0.0216
	(0.0320)
Legacy	-0.0354
	(0.0402)
Number of Observations	3,857
P-Value Joint Significance	0.25

Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2015-2016. Standard errors in parentheses are clustered at the advisor-year level. Regression includes year fixed effects. Advisor VA is standardized by year.

Table B.4: Random Assignment Check

	Math SAT Empirical P-Value (1)	Verbal SAT Empirical P-Value (2)
A. Test for Student Characteristics		
Kolmogorov-Smirnow test (no. failed/total tests)	0/13	0/13
χ^2 goodness of fit test (no. failed/total tests)	0/13	0/13
B. Test for Advisor Characteristics		
Advisor Grade VA	0.021	0.033
	(0.025)	(0.021)
Associate/Full Professor	-0.044	0.004
	(0.060)	(0.056)
Number of Observations	131	131

Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2015-2016. Standard errors in parentheses are clustered at the advisor level. All regressions include year fixed effects. The empirical p-value of each advisor represents the proportion of the 10,000 simulated groups of students with a summed value less than that of the observed group. Advisor VA is standardized by year. The Kolmogorov-Smirnov and χ^2 goodness of fit test results indicate the number of tests of the uniformity of the distribution of p-values that failed at the 5 percent level.

Table B.5: Effect of Advisor Grade VA on Academic Performance, Retention and College Completion

	Standardized GPA (1)	Becoming Sophomore (2)	Time to Sophomore (3)	4-Year Graduation (4)	6-Year Graduation (5)
A. No Controls					
Advisor Grade VA	0.057	0.008	-0.078	0.025	0.015
	(0.016)	(0.006)	(0.026)	(0.008)	(0.010)
B. With Controls					
Advisor Grade VA	0.048	0.007	-0.072	0.022	0.013
	(0.014)	(0.006)	(0.025)	(0.008)	(0.010)
Mean Dep Var	0.038	0.794	2.480	0.458	0.575
R^2 No Controls	0.010	0.014	0.060	0.017	0.014
R^2 with Controls	0.149	0.024	0.080	0.058	0.042
Number of Observations	3,857	3,857	3,047	2,952	2,952

Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2015-2016. The sample is restricted to 2003-2004 to 2012-2013 for graduation outcomes. Standard errors in parentheses are clustered at the advisor-year level. All regressions include year fixed effects and advisor VA is standardized by year. Controls include math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student.

Table B.6: Effect of Advisor Grade VA on Student Major Choice

	Overall Sample (1)	Non-top Students (2)	Top Students (3)	Top Male (4)	Top Female (5)
A. Selective Major					
Enrollment	0.024	0.013	0.049	0.051	0.044
	(0.008)	(0.009)	(0.011)	(0.018)	(0.020)
Graduation	0.015	0.006	0.039	0.048	0.017
	(0.009)	(0.010)	(0.015)	(0.021)	(0.025)
Mean Enrollment	0.429	0.357	0.567	0.586	0.537
Mean Graduation	0.355	0.299	0.464	0.469	0.456
B. STEM Major					
Enrollment	0.010	-0.006	0.038	0.032	0.049
	(0.007)	(0.008)	(0.014)	(0.018)	(0.022)
Graduation	0.010	-0.007	0.042	0.038	0.046
	(0.007)	(0.008)	(0.014)	(0.020)	(0.024)
Mean Enrollment	0.216	0.138	0.368	0.410	0.300
Mean Graduation	0.163	0.098	0.290	0.326	0.232
C. Business Major					
Enrollment	0.013	0.019	0.010	0.019	-0.005
	(0.006)	(0.008)	(0.010)	(0.011)	(0.019)
Graduation	0.005	0.012	-0.004	0.010	-0.029
	(0.006)	(0.007)	(0.012)	(0.014)	(0.020)
Mean Enrollment	0.212	0.219	0.190	0.175	0.238
Mean Graduation	0.191	0.200	0.174	0.143	0.224
N Enrollment	3,857	2,540	1,317	816	501
N Graduation	2,952	1,957	995	616	379

Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2015-2016. The sample is restricted to 2003-2004 to 2012-2013 for graduation outcomes. Standard errors in parentheses are clustered at the advisor-year level. All regressions include year fixed effects and advisor VA is standardized by year. Controls include math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student.

Table B.7: Effect of Various VA Skill Measures on the Corresponding Skills

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:						
	Standardized GPA			Persistence Index		
Grade VA	0.058		0.055	0.054		0.052
	(0.016)		(0.015)	(0.019)		(0.018)
Persistence VA		0.040	0.023		0.040	0.023
		(0.019)	(0.016)		(0.019)	(0.017)
Number of Observations	2,949	2,984	2,917	2,952	2,987	2,920
Panel B:						
	Standardized GPA			Selective Index		
Grade VA	0.058		0.049	0.042		0.032
	(0.016)		(0.016)	(0.019)		(0.018)
Selective VA		0.050	0.033		0.060	0.045
		(0.018)	(0.016)		(0.019)	(0.020)
Number of Observations	2,949	2,984	2,917	2,952	2,987	2,920

Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2012-2013 to be able to create graduation outcomes. Standard errors in parentheses are clustered at the advisor-year level. All regressions include year fixed effects and VA is standardized by year. Each column represents estimates from a separate regression. Controls included are math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student. The slight difference in number of observations across columns is due to missing data points for some VA measures.

Table B.8: Effect of Sophomore Advisor Grade VA on Academic Performance, Retention and College Completion

	Standardized GPA	Dropout after Sophomore	4-Year Graduation	6-Year Graduation	4-Year Graduation in Major	6-Year Graduation in Major
	(1)	(2)	(3)	(4)	(5)	(6)
A. Overall Sample						
Grade VA	0.037	-0.003	0.023	0.016	0.020	0.013
	(0.007)	(0.005)	(0.010)	(0.009)	(0.011)	(0.010)
Mean Dep Var	0.004	0.088	0.527	0.791	0.403	0.549
Number of Observations	14,055	14,055	9,120	9,120	9,120	9,120
B. STEM Majors						
Grade VA	0.041	-0.010	0.034	0.026	0.030	0.019
	(0.010)	(0.003)	(0.011)	(0.008)	(0.015)	(0.014)
Mean Dep Var	0.081	0.070	0.579	0.814	0.417	0.528
Number of Observations	7,859	7,859	4,985	4,985	4,985	4,985
C. Non-STEM Majors						
Grade VA	0.024	0.006	0.08	0.001	0.012	0.008
	(0.008)	(0.009)	(0.013)	(0.012)	(0.012)	(0.012)
Mean Dep Var	-0.092	0.109	0.466	0.765	0.386	0.574
Number of Observations	6,196	6,196	4,135	4,135	4,135	4,135

Sample includes first-time enrolling sophomore students from the academic years 2003-2004 to 2015-2016. The sample is restricted to 2003-2004 to 2012-2013 for graduation outcomes. Standard errors in parentheses are clustered at the advisor-year level. Regressions includes department and year fixed effects. Advisor VA is standardized by year. Controls include math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student. SAT scores are standardized within department and year.

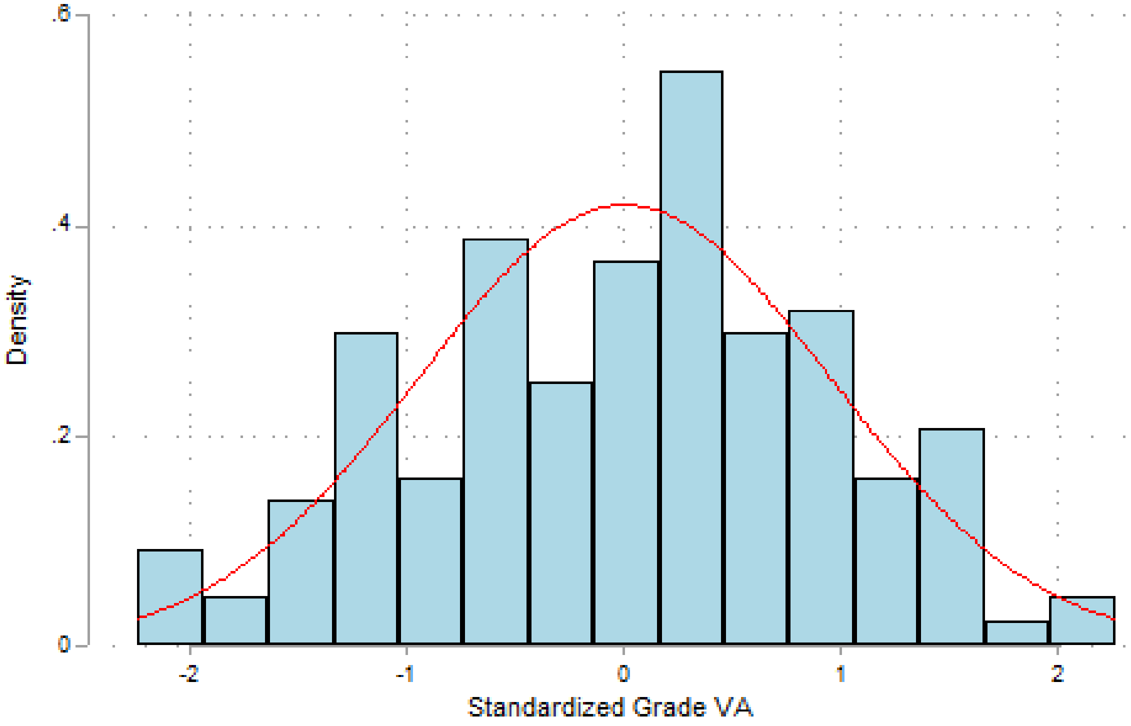
Table B.9: Effect of Advisor Grade VA on Course-Level Freshman Student Outcomes

	Take Science or Maths Course	Fail Course	Withdraw from Course
	(1)	(2)	(3)
A. First Semester			
Advisor Grade VA	-0.002	-0.009	-0.005
	(0.004)	(0.003)	(0.002)
Mean Dep. Var.	0.317	0.067	0.053
Course-Term FE	No	Yes	Yes
N	19,371	19,371	19,371
B. Second Semester			
Advisor Grade VA	-0.002	-0.009	0.000
	(0.004)	(0.003)	(0.002)
Mean Dep. Var.	0.305	0.072	0.048
Course-Term FE	No	Yes	Yes
Number of Observations	16,092	16,092	16,092

Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2015-2016. Standard errors in parentheses clustered two-ways at the advisor-year and individual level. All regressions include year fixed effects and advisor VA is standardized by year. Controls include math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student. Sample includes students from academic years 2003-2004 till 2015-2016.

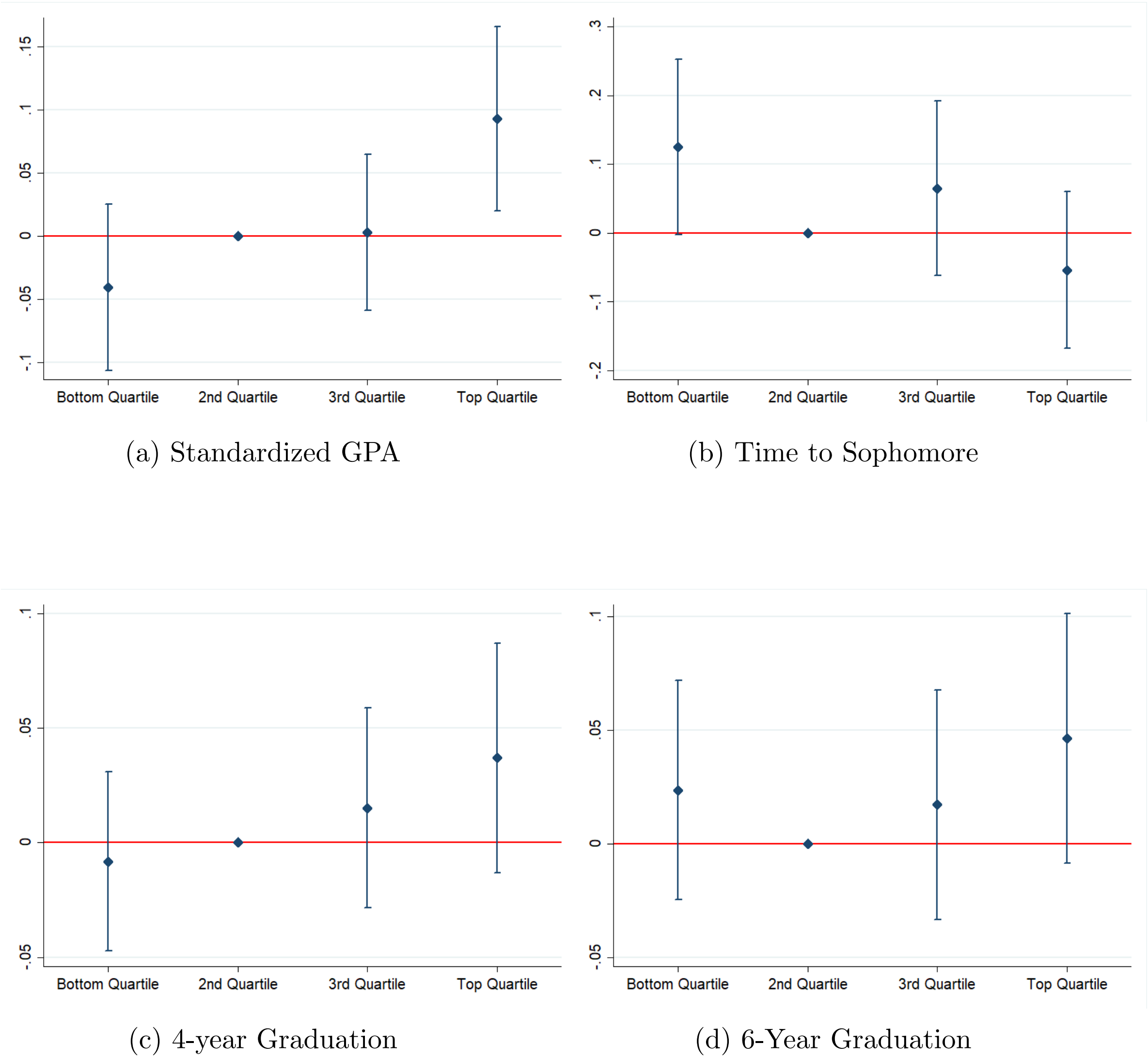
B.3 Appendix Figures

Figure A1: Standardized distribution of advisor VA measure



Notes: The above figure shows the standardized distribution of our constructed advisor value-added measure—based on student course grades. Freshman advisor VA is standardized by year and the sample includes students matched to a freshman advisor who initially enrolled at AUB from academic years 2003-2004 till 2015-2016.

Figure A2: Discrete Treatment on Freshman Academic Performance and College Completion



(a) Standardized GPA

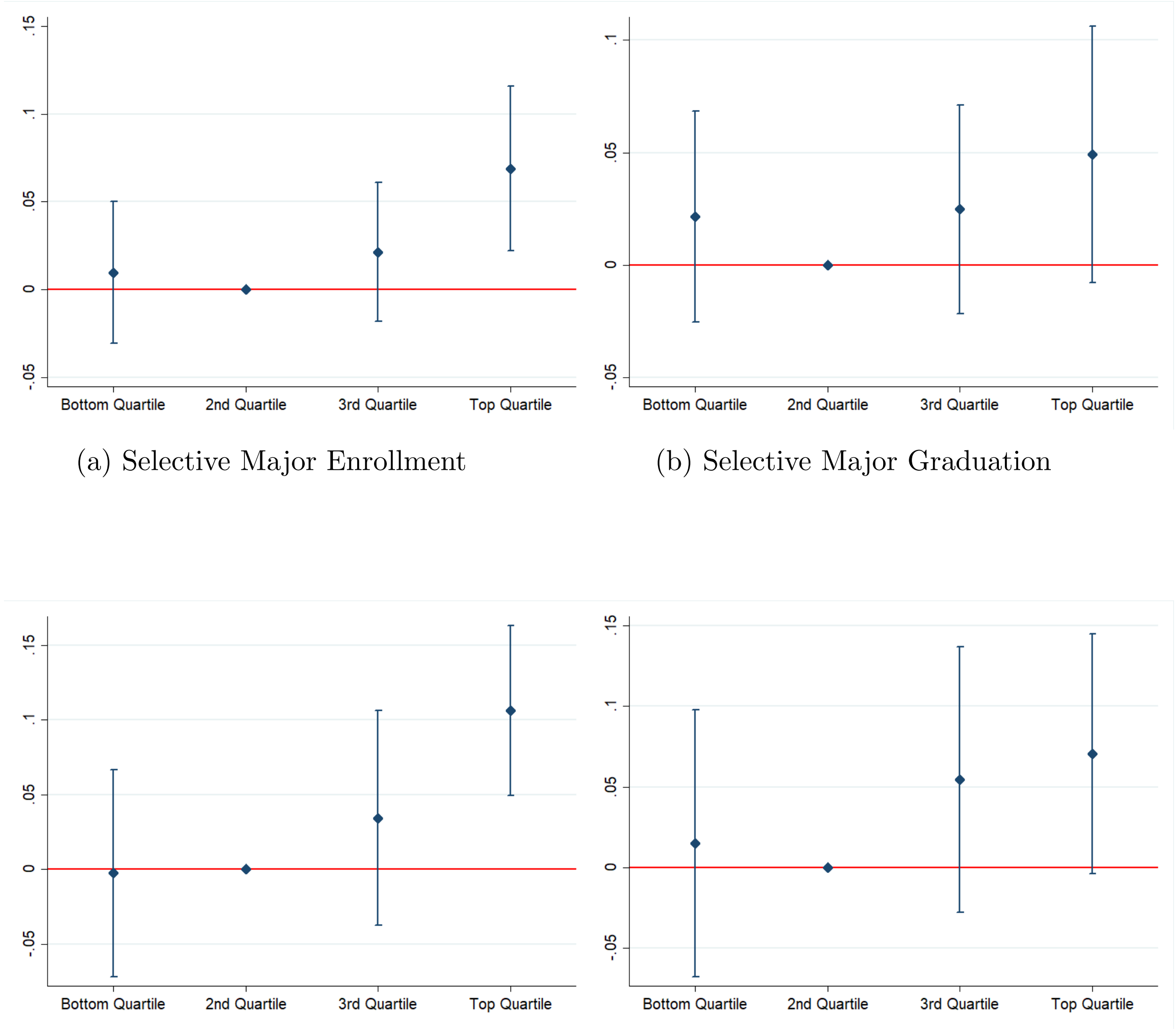
(b) Time to Sophomore

(c) 4-year Graduation

(d) 6-Year Graduation

Notes: The different panels show the impacts of being matched to freshman advisors from different quartiles of the grade VA distribution. Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2015-2016. The sample is restricted to 2003-2004 to 2012-2013 for graduation outcomes. Point estimates represent coefficients from regressions of advisor VA quartile (with the second quartile as the baseline excluded category) on student outcomes. All regression include year fixed effects and students controls. All bars represent 95% confidence intervals with standard errors clustered at the advisor-year level.

Figure A3: Discrete Treatment on Freshman Selective Major Enrollment and Graduation



(a) Selective Major Enrollment

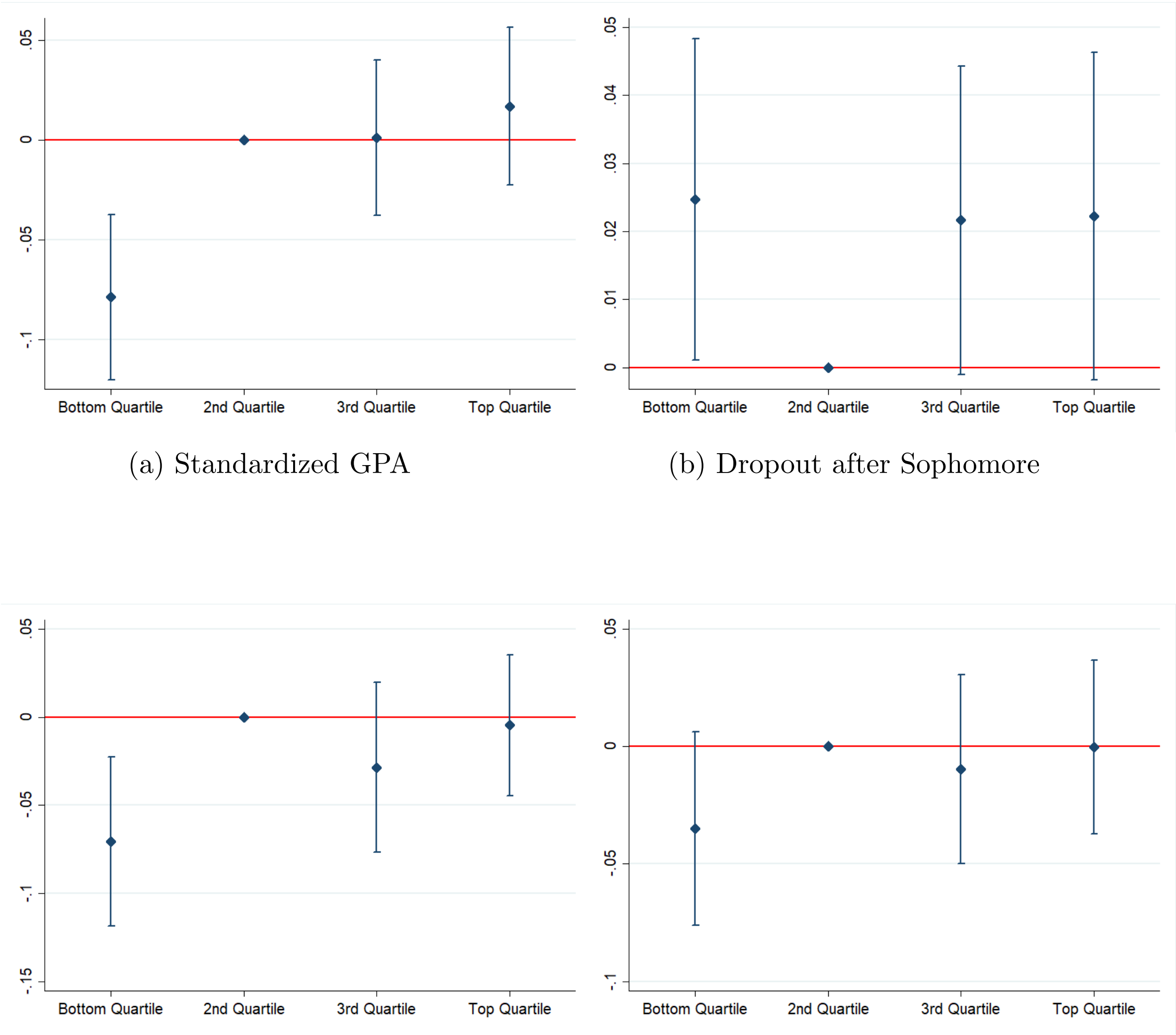
(b) Selective Major Graduation

(c) Top Students' Selective Major Enrollment

(d) Top Students' Selective Major Graduation

Notes: The different panels show the impacts of being matched to freshman advisors from different quartiles of the grade VA distribution. Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2015-2016. The sample is restricted to 2003-2004 to 2012-2013 for graduation outcomes. Point estimates represent coefficients from regressions of advisor VA quartile (with the second quartile as the baseline excluded category) on student outcomes. All regression include year fixed effects and students controls. All bars represent 95% confidence intervals with standard errors clustered at the advisor-year level.

Figure A4: Discrete Treatment on Sophomore Academic Performance and College Completion



(a) Standardized GPA

(b) Dropout after Sophomore

(c) 4-year Graduation

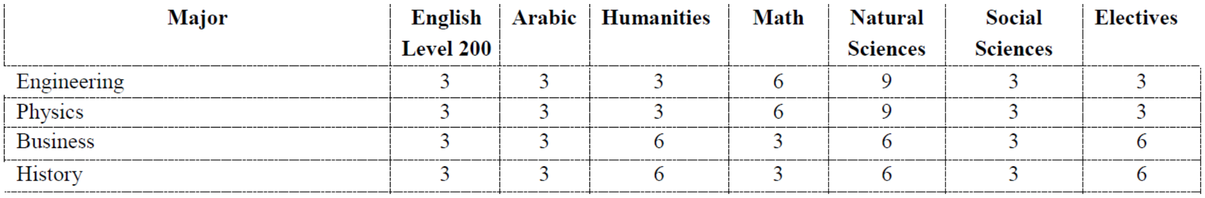
(d) 6-Year Graduation

Notes: The different panels show the impacts of being matched to sophomore advisors from different quartiles of the grade VA distribution. Sample includes first-time enrolling sophomore students from the academic years 2003-2004 to 2015-2016. The sample is restricted to 2003-2004 to 2012-2013 for graduation outcomes. Point estimates represent coefficients from regressions of advisor VA quartile (with the second quartile as the baseline excluded category) on student outcomes. All regression include year and department fixed effects and students controls. All bars represent 95% confidence intervals with standard errors clustered at the advisor-year level.

B.4 Appendix Tables

Table A1: Requirements for admission in different majors

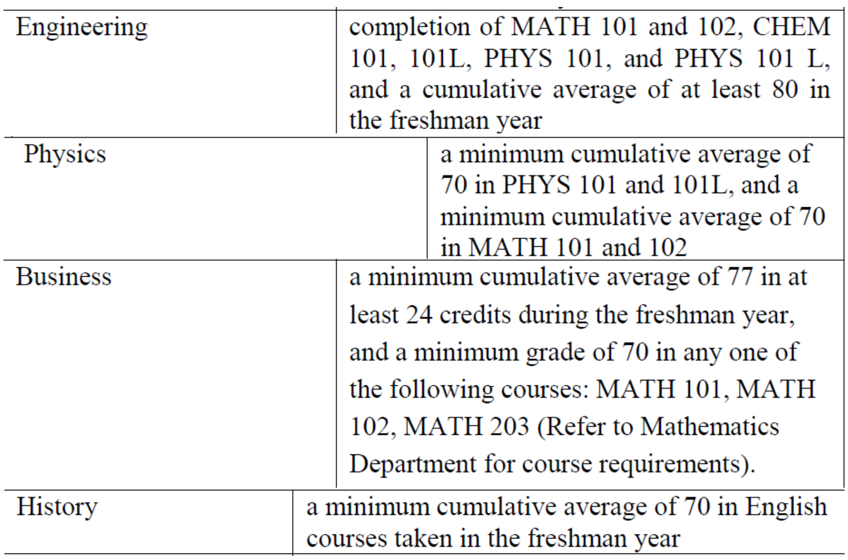
Number of credits required in each discipline by major



Major	English Level 200	Arabic	Humanities	Math	Natural Sciences	Social Sciences	Electives
Engineering	3	3	3	6	9	3	3
Physics	3	3	3	6	9	3	3
Business	3	3	6	3	6	3	6
History	3	3	6	3	6	3	6

Notes: The above table shows the number of credits that a student must pass during the freshman year within each discipline in order to be eligible for admission into engineering, physics, business and history. Each course is typically equivalent to 3 credits.

Additional course and grade requirements by major



Engineering	completion of MATH 101 and 102, CHEM 101, 101L, PHYS 101, and PHYS 101 L, and a cumulative average of at least 80 in the freshman year
Physics	a minimum cumulative average of 70 in PHYS 101 and 101L, and a minimum cumulative average of 70 in MATH 101 and 102
Business	a minimum cumulative average of 77 in at least 24 credits during the freshman year, and a minimum grade of 70 in any one of the following courses: MATH 101, MATH 102, MATH 203 (Refer to Mathematics Department for course requirements).
History	a minimum cumulative average of 70 in English courses taken in the freshman year

Notes: The above table shows specific courses and grades that students must obtain during the freshman year to be eligible for admission into engineering, physics, business and history. For example, the engineering department requires that students take Math 101 (Calculus I), Math 102 (Calculus II), CHEM 101 and 101L (General Chemistry) and PHYS 101 and 101L (Introductory Physics). By passing these courses, students receive enough credits to fulfill the math and science credit requirements for admission into engineering (the first table shows that students need 6 credits in math and 9 credits in sciences).

Table A2: Pre-major academic advising at other private 4-year colleges and universities

College/University	Advisors help students with	Meetings	Advisors are
Amherst College	defining academic goals, improving academic skills, selecting courses, exploring new areas of study and declaring a major	One-on-one meetings prior to course registration	Faculty
Duke University	selecting courses, setting academic goals, deciding on field of study, finding co-curricular opportunities	One-one meeting during orientation week and prior to course registration	Faculty or staff members
Harvard College	choosing courses, meeting degree requirements, considering concentration options, or planning for the summer	One-on-one meetings during course selection week and every 3 or 4 weeks during semester	Faculty, administrators or graduate students
Middlebury College	choosing courses and major keeping tabs on academic problems	One-on-one meetings prior to course registration	Faculty
Princeton University	setting long-term academic goals, selecting courses, discovering academic interests	One-on-one meetings each semester	Faculty
Swarthmore College	selecting courses and program of study; maintaining academic success; discuss setting goals, time management, balancing academics with other parts of life	One-on-one meetings during pre-registration period or when students have academic difficulties	Faculty, deans, administrators, or staff members
Vanderbilt University	creating course schedule; discuss academic goals and progress towards fulfilling curriculum requirements	Phone meeting prior to Fall semester and one-on-one meeting later on	Faculty
Wesleyan University	academic planning, setting long-term academic and career goals, selecting courses and program of study	One-on-one meetings	Faculty
Williams College	choosing a major and courses, setting long-term career goals; check in on students' well-being and academic progress	One-one meetings prior to each course registration period	Faculty
Yale University	selecting courses, setting academic goals, deciding on program of study	Advisors set up one-on-one meetings	Faculty, administrators or staff members

Notes: This table shows the organization of pre-major academic advising at various U.S. private 4-year colleges or universities. The information is taken from each college or university's website.

Table A3: Pre-major academic advising at other private 4-year colleges and universities (continued)

College/University	Advisors have access to students' academic records	Advisors notified of students' academic standing	Advisors approve course withdrawals
Amherst College	N/A	N/A	Yes
Duke University	N/A	N/A but students urged to talk to advisor in case of academic probation	No but students encouraged to discuss course withdrawal with advisors
Harvard College	N/A	N/A	N/A
Middlebury College	Yes	Advisors emailed when students receive course warning (i.e., expected to earn a final grade of "D" or "F")	No but students should discuss course withdrawal with advisors
Princeton University	N/A	N/A	Required approval of Residential dean, director of studies, or academic advisor
Swarthmore College	Yes	Advisors receive copies of all official correspondence concerning advisees' academic standing	No
Vanderbilt University	N/A	N/A	Yes
Wesleyan University	Yes	Advisors notified of students' Unsatisfactory Progress Report and required to schedule one-on-one meeting in that case	No
Williams College	N/A	Advisors notified of students' unsatisfactory grades	No
Yale University	N/A	N/A	No

Notes: This table reports whether pre-major advisors at various U.S. private 4-year colleges or universities perform certain tasks. Each task is shown in a different column. The information is taken from each college or university's website. Information is reported as unavailable (N/A) in case we could not find it on the corresponding university/college's website.

Table A4: Estimate of Forecast Bias of Freshman Advisor Grade VA Measure with Different Sample Splits

A. Leave Current Year and 2 Lags Out	
	Freshman Course Grade
Advisor VA	0.950*** (0.333)
N	33,981
B. Leave Current Year and 2 Leads Out	
	Freshman Course Grade
Advisor VA	1.052*** (0.301)
N	33,696
C. Random Sample Split	
	Freshman Course Grade
Advisor VA	0.811*** (0.343)
N	17,749

Notes: Standard errors in parentheses are clustered at the advisor-year level. All regressions include year fixed effects. Freshman advisor VA is constructed using a leave-year out estimate as described in the methodology section. In Panel C, the random sample split is done by randomly dropping half of the observations in a each year, estimating leave-year out VA, then checking for forecast unbiasedness using the dropped observations. *** p <0.01 ** p <0.05 * p <0.1.

Table A5: Heterogeneous Effects of Advisor Grade VA on Academic Performance and Retention

	Overall	Below Median	Above Median		
	Sample	Math SAT	Math SAT	Male	Female
	(1)	(2)	(3)	(4)	(5)
A. Standardized GPA					
No Controls	0.057***	0.042**	0.072***	0.054**	0.058***
	(0.016)	(0.017)	(0.023)	(0.024)	(0.014)
Controls	0.048***	0.041**	0.054***	0.047**	0.047***
	(0.014)	(0.016)	(0.020)	(0.023)	(0.013)
Mean Dep. Var.	0.038	-0.111	0.202	-0.094	0.182
N	3,857	2,019	1,838	2,014	1,843
B. Likelihood of Becoming Sophomore					
No Controls	0.008	0.001	0.013	0.003	0.012
	(0.006)	(0.009)	(0.008)	(0.009)	(0.008)
Controls	0.007	0.001	0.012	0.003	0.011
	(0.006)	(0.009)	(0.008)	(0.009)	(0.008)
Mean Dep. Var.	0.793	0.772	0.817	0.773	0.816
N	3,857	2,019	1,838	2,014	1,843
C. Time to Sophomore					
No Controls	-0.078***	-0.107***	-0.049	-0.062*	-0.089***
	(0.026)	(0.032)	(0.032)	(0.033)	(0.030)
Controls	-0.072***	-0.103***	-0.041	-0.056*	-0.086***
	(0.025)	(0.031)	(0.031)	(0.032)	(0.029)
Mean Dep. Var.	2.480	2.587	2.373	2.527	2.433
N	3,047	1,526	1,521	1,525	1,522

Notes: Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2015-2016. Standard errors in parentheses are clustered at the advisor-year level. All regressions include year fixed effects and advisor VA is standardized by year. Controls include math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student. *** p <0.01 ** p <0.05 * p <0.1.

Table A6: Effect of Advisor Grade VA on Student Outcomes Using Graduation Sample

	Standardized GPA (1)	Become a Sophomore (2)	Time to Sophomore (3)	Enroll in a Selective Major (3)
A. No Controls				
Advisor Grade VA	0.070***	0.007	-0.081***	0.022**
	(0.018)	(0.008)	(0.028)	(0.009)
B. With Controls				
Advisor Grade VA	0.056***	0.006	-0.075***	0.023**
	(0.016)	(0.008)	(0.026)	(0.009)
Mean Dep Var	0.035	0.776	2.575	0.434
N	2,952	2,952	2,287	2,952

Notes: Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2012-2013. Standard errors in parentheses are clustered at the advisor-year level. All regressions include year fixed effects. Advisor VA is standardized by year. Controls include math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student. Sample includes students from academic years 2003-2004 till 2012-2013. *** p <0.01 ** p <0.05 * p <0.1.

Table A7: Heterogeneous Effects of Advisor Grade VA on College Completion

	Overall Sample	Below Median Math SAT	Above Median Math SAT	Male	Female
	(1)	(2)	(3)	(4)	(5)
A. 4-Year Graduation					
No Controls	0.025***	0.024**	0.023*	0.030**	0.017
	(0.008)	(0.011)	(0.013)	(0.012)	(0.012)
Controls	0.022***	0.025**	0.020	0.028**	0.014
	(0.008)	(0.010)	(0.013)	(0.012)	(0.012)
Mean Dep. Var.	0.458	0.422	0.500	0.480	0.575
B. 6-Year Graduation					
No Controls	0.015	0.008	0.020	0.019	0.010
	(0.009)	(0.011)	(0.015)	(0.014)	(0.010)
Controls	0.013	0.010	0.018	0.018	0.008
	(0.010)	(0.011)	(0.015)	(0.014)	(0.010)
Mean Dep. Var.	0.575	0.547	0.606	0.600	0.687
N	2,952	1,551	1,401	1,551	1,401

Notes: Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2012-2013. Standard errors in parentheses clustered at the advisor-year level. All regressions include year fixed effects and advisor VA is standardized by year. Controls include math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student. *** p <0.01 ** p <0.05 * p <0.1.

Table A8: Effect of Non-Grade VA on Academic Performance, College Completion, and Major Choice

	Standardized GPA	4-Year Graduation	6-Year Graduation	Proportion of Courses Withdrawn	Proportion of Courses Failed	Enroll in Selective Major	Graduate from Selective Major	Proportion of Courses Science
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Persistence VA	0.040**	0.018**	0.014	-0.004**	-0.005	0.020**	0.024***	0.010**
	(0.019)	(0.008)	(0.009)	(0.002)	(0.004)	(0.008)	(0.008)	(0.004)
Selective VA	0.050***	0.018**	0.016*	-0.005***	-0.007*	0.024***	0.028***	0.008**
	(0.018)	(0.008)	(0.010)	(0.002)	(0.004)	(0.009)	(0.009)	(0.004)
N	2,984	2,987	2,987	2,987	2,987	2,987	2,987	2,987

Notes: Sample includes first-time enrolling freshman students from the academic years 2003-2004 to 2012-2013 to be able to create graduation outcomes. Standard errors in parentheses are clustered at the advisor-year level. All regressions include year fixed effects and VA is standardized by year. Controls included are math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student. *** p <0.01 ** p <0.05 * p <0.1.

Table A9: Maximum Likelihood Correlations Between Different VA measures

	Grade VA	Persistence VA	Selective VA
Grade VA	1.00		
Persistence VA	0.59	1.00	
Selective VA	0.60	0.78	1.00

Notes: This table presents the two-way correlation coefficient between the estimated VA on Grades, the Persistence Index, and the Selectiveness Index. Correlations were computed using the maximum likelihood approach described in section IV-E.

Table A10: Summary Statistics for Sophomore Sample

	Mean (1)	S.D. (2)	Obs. (3)
A. Student Level Covariates			
Female	0.480	0.500	14,055
Math SAT	644	72.2	14,055
Verbal SAT	530	106.2	14,055
Legacy Status	0.249	0.432	14,055
B. Student Level Outcomes			
Sophomore GPA	77.5	7.84	14,055
Dropout after Sophomore	0.088	0.283	14,055
Graduate in 4 years	0.529	0.499	9,120
Graduate in 6 Years	0.796	0.403	9,120
Graduate in 4 years in major	0.405	0.491	9,120
Graduate in 6 Years in major	0.554	0.497	9,120
C. Advisor-Year Level Characteristics			
Female	0.310	0.463	736
Science Department	0.484	0.500	736
Lecturer and Other	0.240	0.428	736
Assistant Professor	0.352	0.478	736
Associate Professor	0.174	0.379	736
Professor	0.234	0.423	736
Number of Students	19.1	19.5	736

Notes: Our main sample includes sophomore students who first enrolled in AUB in the academic years 2003-2004 to 2015-2016. Data from these years comprise 194 unique advisors. Our graduation sample includes students who first enrolled in AUB in the academic years 2003-2004 to 2012-2013. Data from these years comprise 152 unique advisors.

Table A11: Estimate of Forecast Bias of Advisor Grade VA Measure for Sophomore Sample

	Sophomore Course Grade
Advisor Grade VA	0.991***
	(0.163)
Mean of VA	0.0004
S.D of VA	0.043
N	144,093

Notes: Standard errors in parentheses are clustered at the advisor-year level. Regressions includes department and year fixed effects. Sophomore advisor VA is constructed using a leave-year out estimate as described in the methodology section. *** p <0.01 ** p <0.05 * p <0.1.

Table A12: Estimate of Forecast Bias of Sophomore Advisor Grade VA Measure with Different Sample Splits

A. Leave Current Year and 2 Lags Out

	Sophomore Course Grade
Advisor VA	1.017*** (0.164)
N	141,305

B. Leave Current Year and 2 Leads Out

	Sophomore Course Grade
Advisor VA	1.060*** (0.164)
N	139,181

C. Random Sample Split

	Sophomore Course Grade
Advisor VA	0.891*** (0.192)
N	71,939

Notes: Standard errors in parentheses are clustered at the advisor-year level. Regressions include department, and year fixed effects. Sophomore advisor VA is constructed using a leave-year out estimate as described in the methodology section. In Panel C, the random sample split is done by randomly dropping half of the observations in a each year-department, estimating leave-year out VA, then checking for forecast unbiasedness using the dropped observations. *** p <0.01 ** p <0.05 * p <0.1.

Table A13: Test of Random Assignment for Sophomore Sample

	Advisor Grade VA
Math SAT	0.010
	(0.009)
Verbal SAT	0.002
	(0.008)
Female	0.029
	(0.022)
Legacy	-0.015
	(0.020)
N	14,055
P-Value Joint Significance	0.462

Notes: Sample includes first-time enrolling sophomore students from the academic years 2003-2004 to 2015-2016. Standard errors in parentheses are clustered at the advisor-year level. Regression includes department and year fixed effects. Advisor VA is standardized by year. SAT scores are standardized within department and year. *** p <0.01 ** p <0.05 * p <0.1.

Table A14: Effect of Sophomore Advisor Persistence VA on Academic Performance, Retention and College Completion

	Standardized GPA	Dropout after Sophomore	4-Year Graduation	6-Year Graduation	4-Year Graduation in Major	6-Year Graduation in Major
	(1)	(2)	(3)	(4)	(5)	(6)
A. Overall Sample						
Persistence VA	0.003	-0.022***	0.030***	0.025***	0.035***	0.036***
	(0.007)	(0.006)	(0.009)	(0.008)	(0.010)	(0.009)
Mean Dep Var	0.006	0.085	0.527	0.791	0.403	0.549
N	8,761	8,761	8,761	8,761	8,761	8,761
B. STEM Majors						
Persistence VA	0.025*	-0.016***	0.026**	0.016*	0.037**	0.036***
	(0.013)	(0.004)	(0.011)	(0.008)	(0.015)	(0.014)
Mean Dep Var	0.083	0.067	0.579	0.814	0.417	0.528
N	4,747	4,747	4,747	4,747	4,747	4,747
C. Non-STEM Majors						
Persistence VA	0.001	-0.022**	0.038***	0.027***	0.037***	0.035***
	(0.008)	(0.009)	(0.011)	(0.010)	(0.011)	(0.011)
Mean Dep Var	-0.086	0.108	0.466	0.765	0.386	0.574
N	4,014	4,014	4,014	4,014	4,014	4,014

Notes: The sample includes first-time enrolling sophomore students from the academic years 2003-2004 to 2012-2013. Standard errors in parentheses are clustered at the advisor-year level. Regressions includes department and year fixed effects. Advisor VA is standardized by year. Controls include math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student. SAT scores are standardized within department and year.
*** p <0.01 ** p <0.05 * p <0.1.

Table A15: Effect of Various VA Skill Measures on the Corresponding Skills Sophomore Sample

	Standardized GPA			Persistence Index		
	(1)	(2)	(3)	(4)	(5)	(6)
Grade VA	0.027*** (0.009)		0.027*** (0.009)	0.038* (0.021)		0.033* (0.018)
Persistence VA		0.003 (0.007)	0.001 (0.008)		0.075*** (0.019)	0.073*** (0.018)
N	8,761	8,761	8,761	8,761	8,761	8,761

Notes: The sample includes first-time enrolling sophomore students from the academic years 2003-2004 to 2012-2013. Standard errors in parentheses are clustered at the advisor-year level. Regressions includes department and year fixed effects. Advisor VA is standardized by year. Controls include math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student. SAT scores are standardized within department and year. Each column represents estimates from a separate regression. Controls included are math and verbal SAT scores, a dummy variable for being a female, and a dummy variable for being a legacy student. *** p <0.01 ** p <0.05 * p <0.1.

Table A16: Observable Characteristics Effect on Freshman Advisor VA Measures

	Grade VA	Persistence VA	Selective VA
Professor	-0.004	0.019	0.014
	(0.018)	(0.023)	(0.020)
Associate Professor	0.006	0.023	0.013
	(0.015)	(0.024)	(0.019)
Female Advisor	0.016	0.012	0.014
	(0.012)	(0.022)	(0.017)
Science Department	0.005	0.021	0.021
	(0.010)	(0.027)	(0.022)
N	131	115	115

Notes: Sample includes academic advisors matched to first-time enrolling freshman students for academic years 2003-2004 to 2015-2016. Standard errors in parentheses are clustered at the advisor level. All regressions include year fixed effects. The number of observations drops for Persistence and Selective VA measures because they are constructed using the sample of students we can observe graduation for (2003-2004 till 2012-2013 freshman entering cohorts). *** p <0.01 ** p <0.05 * p <0.1.

Table A17: Observable Characteristics Effect on Sophomore Advisor VA Measures

	Grade VA	Persistence VA
Professor	-0.008	-0.002
	(0.006)	(0.007)
Associate Professor	-0.003	0.001
	(0.005)	(0.007)
Female Advisor	-0.004	0.008
	(0.005)	(0.007)
Science Department	0.006	0.002
	(0.004)	(0.007)
N	736	646

Notes: Sample includes academic advisors matched to first-time enrolling sophomore students for academic years 2003-2004 to 2015-2016. Standard errors in parentheses are clustered at the advisor level. All regressions include year fixed effects. The number of observations drops for Persistence VA because it is constructed using the sample of students we can observe graduation for (2003-2004 till 2012-2013). *** p <0.01 ** p <0.05 * p <0.1.

Table A18: Effect of Being Matched with a Female Rather than Male Advisor on Freshman Student Outcomes

	Standardized GPA	Becoming Sophomore	4-Year Graduation	6-Year Graduation	Selective Major Enroll	Selective Major Grad	Selective Major Enroll Top	Selective Major Grad Top
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Effect on Male Students (β_1)	-0.012	-0.023	-0.036	-0.020	-0.016	-0.009	-0.045	-0.007
	(0.037)	(0.018)	(0.023)	(0.025)	(0.022)	(0.025)	(0.035)	(0.040)
Effect on Female Students ($\beta_1 + \beta_3$)	0.077**	0.009	0.064***	0.024	0.017	0.032	0.006	0.025
	(0.032)	(0.017)	(0.024)	(0.023)	(0.024)	(0.024)	(0.049)	(0.055)
Mean Dep Var	0.038	0.794	0.458	0.575	0.429	0.355	0.567	0.464
N	3,857	3,857	2,952	2,952	3,857	2,952	1,317	995

Notes: Sample includes advisors matched to first-time enrolling freshman students from the academic years 2003-2004 to 2015-2016. The sample is restricted to 2003-2004 to 2012-2013 for graduation outcomes. Standard errors are clustered at the advisor-year level and reported in parentheses. All regressions include year fixed effects. Controls include math and verbal SAT scores and a dummy variable for being a legacy student. *** p <0.01 ** p <0.05 * p <0.1.

Table A19: Effect of Being Matched with a Female Rather than Male Advisor on Sophomore Student Outcomes

	Standardized GPA	Dropout after Sophomore	4-Year Graduation	6-Year Graduation	4-Year Graduation in Major	6-Year Graduation in Major
	(1)	(2)	(3)	(4)	(5)	(6)
Effect on Male Students (β_1)	-0.049*	-0.003	-0.043	-0.041	-0.009	-0.011
	(0.030)	(0.015)	(0.026)	(0.025)	(0.025)	(0.026)
Effect on Female Students ($\beta_1 + \beta_3$)	0.054**	-0.012	0.003	0.011	0.025	0.031
	(0.021)	(0.017)	(0.028)	(0.025)	(0.028)	(0.026)
Mean Dep Var	0.004	0.088	0.527	0.791	0.403	0.549
N	14,055	14,055	9,120	9,120	9,120	9,120

Notes: Sample includes advisors matched to first-time enrolling sophomore students from the academic years 2003-2004 to 2015-2016. The sample is restricted to 2003-2004 to 2012-2013 for graduation outcomes. Standard errors are clustered at the advisor-year level and reported in parentheses. All regressions include year and department fixed effects. Controls include math and verbal SAT scores and a dummy variable for being a legacy student. SAT scores are standardized within department and year. *** p <0.01 ** p <0.05 * p <0.1.

Appendix C

Appendix for “Clustering and External Validity in Randomized

Controlled Trials"

C.1 Appendix

C.1.1 Tables

Table A1: Conditional and Unconditional Results Of Simplified Specification from Cole et al. (2013).

Dependent Variable: Insurance Take-Up		
	Robust s.e.	Clustered s.e.
Visit	0.164***	0.164**
	(0.054)	(0.073)
Household controls	No	No
Village FEs	Yes	Yes
N	416	416

The results in this table is a simplified version of the regression of Table 5 from Cole et al. (2013). The dependent variable in the regression is an indicator for whether the household purchased an insurance policy. The treatment variable is an indicator for whether the household was visited by an insurance educator and the sample is restricted to having all other treatments equal to 0. The regression includes village fixed effects. Robust standard errors are shown in parentheses in the first column. Standard errors clustered at the village level are shown in the second column.

* p<0.10 ** p<0.05 *** p<0.01.

C.1.2 Useful Results

Under Assumption 8, one has that for all i, k

$$P(D_{ik}=1)=\frac{n_{1k}}{n_k}, \tag{C.1}$$

and for all $j \neq i$

$$\mathbb{E}(D_{ik}D_{jk})=\frac{n_{1k}(n_{1k}-1)}{n_k(n_k-1)}. \tag{C.2}$$

Lemma 1

Lemma 1 *If Assumptions 7 and 9 hold,*

$$(\boldsymbol{D}_k,(\epsilon_{ik}(0),\epsilon_{ik}(1))_{1\leq i\leq n_k}) \perp\!\!\!\perp \left(\boldsymbol{D}_{k'},(\epsilon_{ik'}(0),\epsilon_{ik'}(1))_{1\leq i\leq n_{k'}}\right)|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))$$

Proof

By Assumption 9 and Points 3 and 4 of Assumption 7,

$$(\mathbf{D}_k,(\epsilon_{ik}(0),\epsilon_{ik}(1))_{1\leq i\leq n_k}) \perp\!\!\!\perp \left(\mathbf{D}_{k'},(\epsilon_{ik'}(0),\epsilon_{ik'}(1))_{1\leq i\leq n_{k'}},\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right).$$

Then, the result follows from the fact joint independence implies conditional independence.

C.1.3 Proof of Theorem 4

Conditional Unbiasedness of $\widehat{ATE}$

$$\begin{aligned}
\mathbb{E}\left(\widehat{ATE}_k\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right) &= \mathbb{E}\left(\frac{1}{n_{1k}}\sum_i D_{ik}Y_{ik}(1)-\frac{1}{n_{0k}}\sum_i\left(1-D_{ik}\right)Y_{ik}(0)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
&= \frac{1}{n_{1k}}\sum_i\mathbb{E}\left(D_{ik}Y_{ik}(1)|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
&\quad -\frac{1}{n_{0k}}\sum_i\mathbb{E}\left(\left(1-D_{ik}\right)Y_{ik}(0)|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
&= \frac{1}{n_{1k}}\sum_i\mathbb{E}\left(D_{ik}|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\mathbb{E}\left(Y_{ik}(1)|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
&\quad -\frac{1}{n_{0k}}\sum_i\mathbb{E}\left(1-D_{ik}|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\mathbb{E}\left(Y_{ik}(0)|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
&= \frac{1}{n_{1k}}\sum_i\frac{n_{1k}}{n_k}(y_{ik}(1)+\eta_k(1))-\frac{1}{n_{0k}}\sum_i\frac{n_{0k}}{n_k}(y_{ik}(0)+\eta_k(0))\\
&= (\eta_k(1)-\eta_k(0))+\frac{1}{n_k}\sum_i\left(y_{ik}(1)-y_{ik}(0)\right).
\end{aligned}$$

The first equality holds because we observe $Y_i(1)$ for treated units and $Y_i(0)$ for untreated units, the third equality follows from Assumption 6 and Point 3 of Assumption 7, and the fourth equality follows from Assumption 6 and Points 3 and 4 of Assumption 7 and Equation (C.1).

Now $\widehat{ATE}=\frac{1}{K}\sum\limits_{k=1}^{K}\frac{n_k}{\overline{n}}\widehat{ATE}_k$, therefore :

$$
\begin{aligned}
\mathbb{E}\left(\widehat{ATE}\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right) &= \frac{1}{K}\sum_{k=1}^{K}\frac{n_k}{\overline{n}}\mathbb{E}\left(\widehat{ATE}_k\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{n_k}{\overline{n}}\left[(\eta_k(1)-\eta_k(0))+\frac{1}{n_k}\sum_i\left(y_{ik}(1)-y_{ik}(0)\right)\right]\\
&= ATE(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)).
\end{aligned}
$$

Conditional Variance of $\widehat{ATE}$

We begin by deriving the conditional variance of $\widehat{ATE}_k$.

We start with:

$$
\begin{aligned}
V\left(\widehat{ATE}_k\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right) = \quad & V\left(\mathbb{E}\left(\widehat{ATE}_k\middle|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
&+\mathbb{E}\left(V\left(\widehat{ATE}_k\middle|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right). \quad (C.3)
\end{aligned}
$$

Now begin with the first term:

$$
\begin{aligned}
& V\left(\mathbb{E}\left(\widehat{ATE}_k\middle|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
= \;& V\left(\frac{1}{n_{1k}}\sum_{i=1}^{n_k}D_{ik}y_{ik}(1)+\eta_k(1)-\frac{1}{n_{0k}}\sum_{i=1}^{n_k}(1-D_{ik})y_{ik}(0)-\eta_k(0)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
= \;& V\left(\frac{1}{n_{1k}}\sum_{i=1}^{n_k}D_{ik}y_{ik}(1)-\frac{1}{n_{0k}}\sum_{i=1}^{n_k}(1-D_{ik})y_{ik}(0)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right). \qquad (C.4)
\end{aligned}
$$

The first equality comes from the fact that:

$$\begin{aligned}
\mathbb{E}\left[D_{ik}Y_{ik}(1)|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right] &= D_{ik}\mathbb{E}\left[Y_{ik}(1)|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right]\\
&= D_{ik}\mathbb{E}\left[y_{ik}(1)+\epsilon_{ik}(1)+\eta_k(1)|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right]\\
&= D_{ik}(y_{ik}(1)+\eta_k(1)).
\end{aligned}$$

The second equality follows from Assumption 6, the third equality follows from Points 3 and 4 of Assumption 7. Similarly, one can show that $\mathbb{E}\left((1-D_{ik})Y_{ik}(0)|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right) = (1-D_{ik})(y_{ik}(0)+\eta_k(0))$.

By Point 3 of Assumption 7, $\frac{1}{n_{1k}}\sum_{i=1}^{n_k} D_{ik}y_{ik}(1) - \frac{1}{n_{0k}}\sum_{i=1}^{n_k}(1-D_{ik})y_{ik}(0) \perp\!\!\!\perp (\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))$, so:

$$\begin{aligned}
& V\left(\frac{1}{n_{1k}}\sum_{i=1}^{n_k} D_{ik}y_{ik}(1) - \frac{1}{n_{0k}}\sum_{i=1}^{n_k}(1-D_{ik})y_{ik}(0)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
= \;& V\left(\frac{1}{n_{1k}}\sum_{i=1}^{n_k} D_{ik}y_{ik}(1) - \frac{1}{n_{0k}}\sum_{i=1}^{n_k}(1-D_{ik})y_{ik}(0)\right). \qquad \text{(C.5)}
\end{aligned}$$

The right hand side of the previous equation is the variance of the estimated average treatment effect in the case of deterministic potential outcomes. Then, it follows from Equations (C.4) and (C.5) and from Neyman (1923) that:

$$V\left(\mathbb{E}\left(\widehat{ATE}_k|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right) = \frac{1}{n_{0k}}S^2_{y(0),k} + \frac{1}{n_{1k}}S^2_{y(1),k} - \frac{1}{n_k}S^2_{y(1)-y(0),k}. \qquad \text{(C.6)}$$

Moving to the second term :

$$\begin{aligned}
&\mathbb{E}\ \left(V\left(\widehat{ATE}_k\middle|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
=\ &\mathbb{E}\left(V\left(\frac{1}{n_{1k}}\sum_{i=1}^{n_k}D_{ik}\epsilon_{ik}(1)-\frac{1}{n_{0k}}\sum_{i=1}^{n_k}(1-D_{ik})\epsilon_{ik}(0)\middle|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
=\ &\mathbb{E}\left(\frac{1}{n_{1k}^2}\sum_{i=1}^{n_k}D_{ik}^2\sigma_{1i}^2+\frac{1}{n_{0k}^2}\sum_{i=1}^{n_k}\left(1-D_{ik}\right)^2\sigma_{0i}^2\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
&-E\left(\frac{1}{n_{0k}n_{1k}}Cov\left(\sum_{i=1}^{n_k}D_{ik}\epsilon_{ik}(1),\sum_{i=1}^{n_k}(1-D_{ik})\epsilon_{ik}(0)\middle|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
=\ &\frac{1}{n_{1k}^2}\sum_{i=1}^{n_k}\mathbb{E}\left(D_{ik}^2\right)\sigma_{1i}^2+\frac{1}{n_{0k}^2}\sum_{i=1}^{n_k}\mathbb{E}\left[(1-D_{ik})^2\right]\sigma_{0i}^2-\\
&\frac{1}{n_{0k}n_{1k}}\mathbb{E}\left(Cov\left(\sum_{i=1}^{n_k}D_{ik}\epsilon_{ik}(1),\sum_{i=1}^{n_k}(1-D_{ik})\epsilon_{ik}(0)\middle|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
=\ &\frac{1}{n_{1k}^2}\sum_{i=1}^{n_k}\frac{n_{1k}}{n_k}\sigma_{1i}^2+\frac{1}{n_{0k}^2}\sum_{i=1}^{n_k}\frac{n_{0k}}{n_k}\sigma_{0i}^2\\
&-\frac{1}{n_{0k}n_{1k}}\mathbb{E}\left(\sum_{i=1}^{n_k}Cov\left(D_{ik}\epsilon_{ik}(1),(1-D_{ik})\epsilon_{ik}(0)|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
=\ &\frac{1}{n_{1k}}\overline{\sigma_{1k}^2}+\frac{1}{n_{0k}}\overline{\sigma_{0k}^2}\\
&-\frac{1}{n_{0k}n_{1k}}\mathbb{E}\left(\sum_{i=1}^{n_k}D_{ik}(1-D_{ik})Cov\left(\epsilon_{ik}(1),\epsilon_{ik}(0)|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
=\ &\frac{1}{n_{1k}}\overline{\sigma_{1k}^2}+\frac{1}{n_{0k}}\overline{\sigma_{0k}^2}.
\end{aligned}\tag{C.7}$$

The first equality holds because conditional on $\mathbf{D}_k$ and $(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))$ there is no randomness in $(1-D_{ik})(y_{ik}(0)+\eta_k(0))$ and $D_{ik}(y_{ik}(1)+\eta_k(1))$. The second equality holds by Points 1, 2, 3, and 4 of Assumption 7, which imply that $V\left(\epsilon_{ik}(d)|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)=V\left(\epsilon_{ik}(d)\right)=\sigma_{dik}^2$ and $Cov(\epsilon_{ik}(d),\epsilon_{jk}(d)|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)))=Cov(\epsilon_{ik}(d),\epsilon_{jk}(d))=0$. The third equality holds by Point 3 of Assumption 7. The fourth equality holds since D_{ik} is binary therefore $D_{ik}^2=D_{ik}$, and because by Points 2, 3, and 4 of Assumption 7,

$Cov(\epsilon_{ik}(1), \epsilon_{jk}(0)|\mathbf{D}_k, (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))) = Cov(\epsilon_{ik}(1), \epsilon_{jk}(0)) = 0$. The last equality holds since $D_{ik}(1 - D_{ik}) = 0$.

Combining Equations (C.3), (C.6), and (C.7) shows that:

$$V\left(\widehat{ATE}_k \middle| (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) = \frac{1}{n_{0k}} S^2_{y(0),k} + \frac{1}{n_{1k}} S^2_{y(1),k} - \frac{1}{n} S^2_{y(1)-y(0),k} + \frac{1}{n_{1k}} \overline{\sigma^2_{1}}_k + \frac{1}{n_{0k}} \overline{\sigma^2_{0}}_k. \quad \text{(C.8)}$$

Now by Lemma 1, for all $k \neq k'$:

$$Cov\left(\widehat{ATE}_k, \widehat{ATE}_{k'} \middle| (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) = 0. \quad \text{(C.9)}$$

Finally, combining the fact that $\widehat{ATE} = \frac{1}{K}\sum_{k=1}^{K} \frac{n_k}{\overline{n}} \widehat{ATE}_k$ and Equation (C.9) we get:

$$V\left(\widehat{ATE} \middle| (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) = \frac{1}{K^2} \sum_{k=1}^{K} \left(\frac{n_k}{\overline{n}}\right)^2 V\left(\widehat{ATE}_k | ((\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right).$$

Estimating an Upper Bound for the Conditional Variance of $\widehat{ATE}$

We start by showing that

$$\mathbb{E}\left(\frac{1}{n_{1k}}\left[\frac{1}{n_{1k}-1}\sum_{i=1}^{n_k} D_{ik}\left(Y_{ik}(1) - \overline{Y}_{1k}\right)^2\right] \middle| (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) = \frac{1}{n_{1k}} S^2_{y(1),k} + \frac{1}{n_{1k}} \overline{\sigma^2_{1}}_k. \quad \text{(C.10)}$$

We have:

$$\begin{aligned}
& \frac{1}{n_{1k}}\left[\frac{1}{n_{1k}-1}\sum_{i=1}^{n_k} D_{ik}\left(Y_{ik}(1)-\overline{Y}_{1k}\right)^2\right] \\
= & \frac{1}{n_{1k}}\frac{1}{n_{1k}-1}\left[\sum_{i=1}^{n_k} D_{ik}Y_{ik}(1)^2 - n_{1k}\overline{Y}_{1k}^2\right] \\
= & \frac{1}{n_{1k}}\frac{1}{n_{1k}-1} \\
& \left[\sum_{i=1}^{n_k} D_{ik}\left(y_{ik}(1)^2 + 2\epsilon_{ik}(1)y_{ik}(1) + \epsilon_{ik}^2(1) + \eta_k^2(1) + 2y_{ik}(1)\eta_k(1) + 2\eta_k(1)\epsilon_{ik}(1)\right)\right] \\
& -\frac{1}{n_{1k}}\frac{1}{n_{1k}-1}\left[\frac{\left(\sum_{i=1}^{n_k} D_{ik}y_{ik}(1) + \sum_{i=1}^{n_k} D_{ik}\epsilon_{ik}(1) + \eta_k(1)\sum_{i=1}^{n_k} D_{ik}\right)^2}{n_{1k}}\right] \\
= & A+B+C+D+E+F,
\end{aligned}$$

with

$$\begin{aligned}
A &= \frac{1}{n_{1k}}\frac{1}{n_{1k}-1}\left[\sum_{i=1}^{n_k} D_{ik}y_{ik}(1)^2 - \frac{\left(\sum_{i=1}^{n_k} D_{ik}y_{ik}(1)\right)^2}{n_{1k}}\right], \\
B &= \frac{1}{n_{1k}}\frac{1}{n_{1k}-1}\left[\sum_{i=1}^{n_k} D_{ik}\epsilon_{ik}^2(1) - \frac{\left(\sum_{i=1}^{n_k} D_{ik}\epsilon_{ik}(1)\right)^2}{n_{1k}}\right], \\
C &= \frac{1}{n_{1k}}\frac{2}{n_{1k}-1}\left[\sum_{i=1}^{n_k} D_{ik}y_{ik}(1)\epsilon_{ik}(1) - \frac{\left(\left(\sum_{i=1}^{n_k} D_{ik}y_{ik}(1)\right)\left(\sum_{i=1}^{n_k} D_{ik}\epsilon_{ik}(1)\right)\right)}{n_{1k}}\right],
\end{aligned}$$

and

$$
\begin{aligned}
D &= \frac{1}{n_{1k}}\frac{1}{n_{1k}-1}\left[\eta_k^2(1)\sum_{i=1}^{n_k} D_{ik} - \frac{1}{n_{1k}}\left(\eta_k(1)\sum_{i=1}^{n_k} D_{ik}\right)^2\right] = 0, \\
E &= \frac{1}{n_{1k}}\frac{1}{n_{1k}-1}\left[2\eta_k(1)\sum_{i=1}^{n_k} D_{ik}y_{ik}(1) - \frac{2}{n_{1k}}\left(\sum_{i=1}^{n_k} D_{ik}y_{ik}(1)\right)\left(\eta_k(1)\sum_{i=1}^{n_k} D_{ik}\right)\right] = 0, \\
F &= \frac{1}{n_{1k}}\frac{1}{n_{1k}-1}\left[2\eta_k(1)\sum_{i=1}^{n_k} D_{ik}\epsilon_{ik}(1) - \frac{2}{n_{1k}}\left(\sum_{i=1}^{n_k} D_{ik}\epsilon_{ik}(1)\right)\left(\eta_k(1)\sum_{i=1}^{n_k} D_{ik}\right)\right] = 0.
\end{aligned}
$$

To ease notation let $\mathbb{E}^*(X) = \mathbb{E}(X|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)))$.

$$
\begin{aligned}
& \mathbb{E}^*(A) = \mathbb{E}(A) \\
= \quad & \frac{1}{n_{1k}(n_{1k}-1)} \\
& \left[\sum_{i=1}^{n_k} \mathbb{E}(D_{ik}) y_{ik}^2(1) - \frac{1}{n_{1k}} \sum_{i=1}^{n_k} \mathbb{E}(D_{ik}) y_{ik}^2(1) - \frac{1}{n_{1k}} \sum\sum_{i \neq j} \mathbb{E}(D_{ik} D_{jk}) y_{ik}(1) y_{jk}(1)\right] \\
= \quad & \frac{1}{n_{1k}(n_{1k}-1)} \\
& \left[\frac{n_{1k}}{n_k} \sum_{i=1}^{n_k} y_{ik}^2(1) - \frac{1}{n_{1k}} \sum_{i=1}^{n_k} \frac{n_{1k}}{n_k} y_{ik}^2(1) - \frac{1}{n_{1k}} \sum\sum_{i \neq j} \frac{n_{1k}}{n_k} \frac{n_{1k}-1}{n_k-1} y_{ik}(1) y_{jk}(1)\right] \\
= \quad & \frac{1}{n_{1k}(n_{1k}-1)} \left[\frac{n_{1k}-1}{n_k} \sum_{i=1}^{n_k} y_{ik}^2(1) - \frac{n_{1k}-1}{n_k(n_k-1)} \sum\sum_{i \neq j} y_{ik}(1) y_{jk}(1)\right] \\
= \quad & \frac{1}{n_{1k}(n_{1k}-1)} \\
& \left[\frac{n_{1k}-1}{n_k} \sum_{i=1}^{n_k} y_{ik}^2(1) + \frac{n_{1k}-1}{n_k(n_k-1)} \sum_{i=1}^{n_k} y_{ik}^2(1) - \frac{n_{1k}-1}{n_k(n_k-1)} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} y_{ik}(1) y_{jk}(1)\right] \\
= \quad & \frac{1}{n_{1k}(n_{1k}-1)} \\
& \left[\frac{(n_{1k}-1)(n_k-1) + (n_{1k}-1)}{n_k(n_k-1)} \sum_{i=1}^{n_k} y_{ik}^2(1) - \frac{n_k(n_{1k}-1)}{(n_k-1)} \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} \frac{y_{ik}(1)}{n_k} \frac{y_{jk}(1)}{n_k}\right] \\
= \quad & \frac{1}{n_{1k}} \left[\frac{1}{n_k-1} \sum_{i=1}^{n_k} y_{ik}^2(1) - \frac{n_k}{n_k-1} \overline{y}_k^2(1)\right] \\
= \quad & \frac{1}{n_{1k}} S_{y(1),k}^2.
\end{aligned}
$$

The first equality holds by Point 3 of Assumption 7. The third holds by Equations (C.1) and (C.2).

Moving to B,

$$\begin{aligned}
\mathbb{E}^*(B) &= \frac{1}{n_{1k}(n_{1k}-1)}\left[\sum_{i=1}^{n_k}\mathbb{E}^*(D_{ik})\mathbb{E}^*(\epsilon_{ik}^2(1)) - \frac{\mathbb{E}^*\left(\left(\sum\limits_{i=1}^{n_k} D_{ik}\epsilon_{ik}(1)\right)^2\right)}{n_{1k}}\right] \\
&= \frac{1}{n_{1k}(n_{1k}-1)}\left[\sum_{i=1}^{n_k}\mathbb{E}^*(D_{ik})\mathbb{E}^*(\epsilon_{ik}^2(1)) - \frac{1}{n_{1k}}\sum_{i=1}^{n_k}\mathbb{E}^*(D_{ik}^2)\mathbb{E}^*(\epsilon_{ik}^2(1))\right. \\
&\quad \left. -\frac{1}{n_{1k}}\sum\sum_{i\neq j}\mathbb{E}^*(D_{ik}D_{jk})\mathbb{E}^*(\epsilon_{ik}(1)\epsilon_{jk}(1))\right] \\
&= \frac{1}{n_{1k}(n_{1k}-1)}\left[\frac{n_{1k}}{n_k}\sum_{i=1}^{n_k}\sigma_{1i}^2 - \frac{1}{n_{1k}}\sum_{i=1}^{n_k}\frac{n_{1k}}{n_k}\sigma_{1i}^2\right] \\
&= \frac{1}{n_{1k}}\overline{\sigma_1^2}_k.
\end{aligned}$$

The first equality holds by Points 1 and 3 of Assumption 7. The second equality holds by Points 1 and 3 of Assumption 7. The third equality holds since $D_{ik}^2 = D_{ik}$ and by Points 1, 2, 3, and 4 of Assumption 7 as well as by Equation (C.1), which imply $\mathbb{E}^*(D_{ik}) = \mathbb{E}(D_{ik}) = \frac{n_{1k}}{n_k}$, $E^*(\epsilon_{ik}^2(1)) = E(\epsilon_{ik}^2(1)) = \sigma_{1i}^2$, and $\mathbb{E}^*(\epsilon_{ik}(1)\epsilon_{jk}(1)) = \mathbb{E}(\epsilon_{ik}(1)\epsilon_{jk}(1)) = 0$.

Finally for C:

$$\begin{aligned}
\mathbb{E}^*(C) &= \frac{1}{n_{1k}}\frac{2}{n_{1k}-1}\left[\sum_{i=1}^{n_k}\mathbb{E}^*(D_{ik})y_{ik}(1)\mathbb{E}^*(\epsilon_{ik}(1)) - \frac{\mathbb{E}^*\left(\sum\limits_{i=1}^{n_k} D_{ik}y_{ik}(1)\sum\limits_{i=1}^{n_k} D_{ik}\epsilon_{ik}(1)\right)}{n_{1k}}\right] \\
&= \frac{1}{n_{1k}}\frac{2}{n_{1k}-1} \\
&\quad \left[-\sum_{i=1}^{n_k}\mathbb{E}^*(D_{ik}^2)y_{ik}(1)\mathbb{E}^*(\epsilon_{ik}(1)) - \sum\sum_{i\neq j}\mathbb{E}^*(D_{ik}D_{jk})y_{ik}(1)\mathbb{E}^*(\epsilon_{jk}(1))\right] \\
&= 0.
\end{aligned}$$

The first equality holds by Point 3 of Assumption 7. The second and third equalities hold by Point 4 of Assumption 7, which implies $\mathbb{E}^*(\epsilon_{ik}(1)) = \mathbb{E}(\epsilon_{ik}(1)) = 0$. This completes the proof of (C.10). Using similar arguments, one can show that

$$\mathbb{E}\left(\frac{1}{n_{0k}}\left[\frac{1}{n_{0k}-1}\sum_{i=1}^{n_k}(1-D_{ik})\left(Y_{ik}(0)-\overline{Y}_{0k}\right)^2\right]\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right) = \frac{1}{n_{0k}}S^2_{y(0),k} + \frac{1}{n_{0k}}\overline{\sigma^2_{0}}_k. \tag{C.11}$$

Point 2 of the theorem and Equations (C.10) and (C.11) imply that

$$V\left(\widehat{ATE}_k|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right) \leq \mathbb{E}\left(\widehat{V}_{rob}\left(\widehat{ATE}_k\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right).$$

Finally, the result follows from Point 2 of the theorem, the definition of $\widehat{V}_{rob}\left(\widehat{ATE}\right)$, and the linearity of the conditional expectation.

C.1.4 Proof of Theorem 5

Unbiasedness of $\widehat{ATE}$

$$\begin{aligned}
\mathbb{E}\left(\widehat{ATE}\right) &= \mathbb{E}\left[\mathbb{E}\left(\widehat{ATE}|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right)\right] \\
&= \mathbb{E}\left[ATE(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right] \\
&= \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{n_k}{\overline{n}}\sum_{i=0}^{n_k}\left[(y_{ik}(1)+\eta_k(1))-(y_{ik}(0)+\eta_k(0))\right]\right] \\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{n_k}{\overline{n}}\sum_{i=0}^{n_k}\left[y_{ik}(1)-y_{ik}(0)\right] \\
&= ATE.
\end{aligned} \tag{C.12}$$

Where the first equality holds by the law of iterated expectations, and the second equality holds by Theorem 4.

Unconditional Variance of $\widehat{ATE}$

We start with:

$$V\left(\widehat{ATE}\right) = \mathbb{E}\left(V\left(\widehat{ATE}|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right)\right) + V\left(\mathbb{E}\left(\widehat{ATE}|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right)\right). \tag{C.13}$$

Now begin with the first term:

$$\begin{aligned}\mathbb{E}\left(V\left(\widehat{ATE}|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right)\right) &= \mathbb{E}\left[\frac{1}{K^2}\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2 V\left(\widehat{ATE}_k|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right)\right] \\ &= \frac{1}{K^2}\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2 V\left(\widehat{ATE}_k|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right). \end{aligned} \tag{C.14}$$

Where the first equality holds by Theorem 4, and the second equality holds because $V\left(\widehat{ATE}_k|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right)$ contains no stochastic components, as shown in Point 2 of Theorem 4.

Moving to the second term:

$$
\begin{aligned}
& V\left(\mathbb{E}\left(\widehat{ATE}|\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)\right)\right) \\
= \; & V\left(ATE(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right) \\
= \; & V\left(\frac{1}{K}\sum_{k=1}^{K}\frac{n_k}{\overline{n}}\frac{1}{n_k}\sum_{i=0}^{n_k}\left[(y_{ik}(1)+\eta_k(1))-(y_{ik}(0)+\eta_k(0))\right]\right) \\
= \; & \frac{1}{K^2}V\left(\sum_{k=1}^{K}\frac{1}{\overline{n}}\sum_{i=0}^{n_k}(y_{ik}(1)-y_{ik}(0))+(\eta_k(1)-\eta_k(0))\right) \\
= \; & \frac{1}{K^2}V\left(\sum_{k=1}^{K}\frac{n_k}{\overline{n}}(\eta_k(1)-\eta_k(0))\right) \\
= \; & \frac{1}{K^2}\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2 V\left(\eta_k(1)-\eta_k(0)\right). \qquad \text{(C.15)}
\end{aligned}
$$

The first equality holds by Theorem 4, the last holds by Assumption 9 and Point 5 of Assumption 7. The result follows from (C.13), (C.14), and (C.15), and from Point 2 of Theorem 4.

Estimating an Upper Bound for the Variance of $\widehat{ATE}$

The first inequality is trivial so we only prove the second one.

$$\begin{aligned}
\mathbb{E}\left[\widehat{V}_{clu}(\widehat{ATE})\right] &= \mathbb{E}\left[\frac{1}{K\left(K-1\right)}\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\widehat{ATE}_k - \widehat{ATE}\right)^2\right] \\
&= \frac{1}{K\left(K-1\right)}\mathbb{E}\left[\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2\widehat{ATE}_k^2 - 2\widehat{ATE}\sum_{k=1}^{K}\frac{n_k}{\overline{n}}\widehat{ATE}_k + K\widehat{ATE}^2\right] \\
&= \frac{1}{K\left(K-1\right)}\left[\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2\mathbb{E}\left(\widehat{ATE}_k^2\right) - K\mathbb{E}\left(\widehat{ATE}^2\right)\right] \\
&= \frac{1}{K\left(K-1\right)} \\
&\quad \left[\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2\left[V\left(\widehat{ATE}_k\right) + \mathbb{E}\left(\widehat{ATE}_k\right)^2\right] - K\left[V\left(\widehat{ATE}\right) + \mathbb{E}\left(\widehat{ATE}\right)^2\right]\right] \\
&= \frac{1}{K\left(K-1\right)} \\
&\quad \left[K^2V\left(\widehat{ATE}\right) + \sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2\mathbb{E}\left(\widehat{ATE}_k\right)^2 - KV\left(\widehat{ATE}\right) - K\mathbb{E}\left(\widehat{ATE}\right)^2\right] \\
&= V\left(\widehat{ATE}\right) + \frac{1}{K\left(K-1\right)}\left[\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2\mathbb{E}\left(\widehat{ATE}_k\right)^2 - K\mathbb{E}\left(\widehat{ATE}\right)^2\right]. \qquad \text{(C.16)}
\end{aligned}$$

The second and third equalities follow from algebraic manipulations and the linearity of the expectations operator, the fourth follows from the definition of a variance, and the fifth follows from $V\left(\widehat{ATE}\right) = \frac{1}{K^2}\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2 V\left(\widehat{ATE}_k\right)$.

By convexity of $x \to x^2$,

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2\mathbb{E}\left(\widehat{ATE}_k\right)^2 &\geq \left[\frac{1}{K}\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)\mathbb{E}\left(\widehat{ATE}_k\right)\right]^2 \\
\Leftrightarrow \qquad \sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)^2\mathbb{E}\left(\widehat{ATE}_k\right)^2 &\geq K\mathbb{E}\left(\widehat{ATE}\right)^2,
\end{aligned}$$

so the second term in Equation (C.16) is positive. This proves the result.

C.1.5 Proof of Corollary 2

If $n_k = \overline{n}$, it follows from Equation (C.16) that

$$\mathbb{E}\left[\widehat{V}_{clu}(\widehat{ATE})\right] = V\left(\widehat{ATE}\right) + \frac{1}{K\left(K-1\right)}\left[\sum_{k=1}^{K} E\left(\widehat{ATE}_k\right)^2 - K\mathbb{E}\left(\widehat{ATE}\right)^2\right].$$

Moreover,

$$\begin{aligned}
&\mathbb{E}\left[\widehat{V}_{rob}\left(\widehat{ATE}_k\right)\right]\\
=&\mathbb{E}\left[\mathbb{E}^*\left(\widehat{V}_{rob}\left(\widehat{ATE}_k\right)\right)\right]\\
=&\mathbb{E}\left[\mathbb{E}^*\left(\frac{1}{K^2}\sum_{k=1}^{K}\left(\frac{n_k}{\overline{n}}\right)\widehat{V}_{rob}\left(\widehat{ATE}_k\right)\right)\right]\\
=&\mathbb{E}\left[V\left(\widehat{ATE}|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right) + \frac{1}{K^2}\sum_{k=1}^{K}\frac{1}{n_k}\left(\frac{n_k}{\overline{n}}\right)^2 S^2_{y(1)-y(0),k}\right]\\
=&\mathbb{E}\left[V\left(\widehat{ATE}|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\right] + \frac{1}{K^2}\sum_{k=1}^{K}\frac{1}{n_k}S^2_{y(1)-y(0),k},
\end{aligned}$$

where the first equality holds by the law of iterated expectations, the third equality holds by Equations (C.10) and (C.11) and Point 2 of Theorem 4, and the last follows from $n_k = \overline{n}$. Combining the two preceding displays,

$$\begin{aligned}
&\mathbb{E}\left[\widehat{V}_{clu}(\widehat{ATE})\right] - \mathbb{E}\left[\widehat{V}_{rob}\left(\widehat{ATE}_k\right)\right] \\
=&V\left(\widehat{ATE}\right) - \mathbb{E}\left[V\left(\widehat{ATE}|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\right] + \\
&\frac{1}{K\left(K-1\right)}\left[\sum_{k=1}^{K} E\left(\widehat{ATE}_k\right)^2 - K\mathbb{E}\left(\widehat{ATE}\right)^2\right] - \frac{1}{K^2}\sum_{k=1}^{K}\frac{1}{n_k}S^2_{y(1)-y(0),k} \\
=&\frac{1}{K^2}\sum_{k=1}^{K} V\left(\eta_k(1)-\eta_k(0)\right) + \frac{1}{K\left(K-1\right)}\left[\sum_{k=1}^{K} E\left(\widehat{ATE}_k\right)^2 - K\mathbb{E}\left(\widehat{ATE}\right)^2\right] \\
&-\frac{1}{K^2}\sum_{k=1}^{K}\frac{1}{n_k}S^2_{y(1)-y(0),k},
\end{aligned}$$

where the second equality follows from Equations (C.13) and (C.15) and $n_k = \overline{n}$. This proves the result.

C.1.6 Proof of Theorem 6

Let:

$$\begin{aligned}
& \frac{1}{S_K/K}\left[\widehat{ATE} - ATE\right] \\
= & \frac{1}{S_K}K\left[\widehat{ATE} - \mathbb{E}\left(\widehat{ATE}\right)\right] \\
= & \frac{1}{S_K}K\left[\frac{1}{K}\sum_{k=1}^{K} AD_k - \mathbb{E}\left(\frac{1}{K}\sum_{k=1}^{K} AD_k\right)\right] \\
= & \frac{1}{S_K}\sum_{k=1}^{K}\left[AD_k - \mathbb{E}\left(AD_k\right)\right],
\end{aligned}$$

where the first equality holds by Theorem 5. Under Assumptions 6, 7, 8, 9, and 11,

$\sum_{k=1}^{K}\left[AD_k - \mathbb{E}\left(AD_k\right)\right]$ is a sum of independent mean 0 random variables with finite variance. Furthermore, by Point 3 of Assumption 11 we know that $\lim\limits_{K\to+\infty}\frac{1}{S_K^{2+\epsilon}}\sum_{k=1}^{K}\mathbb{E}\left[|AD_k - E(AD_k)|^{2+\epsilon}\right] = 0$ for some $\epsilon > 0$. Therefore by the Lyapunov CLT:

$$\frac{1}{S_K}\sum_{k=1}^{K}\left[AD_k - \mathbb{E}\left(AD_k\right)\right] \xrightarrow{d} N(0,1).$$

Now note that:

$$\begin{aligned}\frac{1}{S_K/K}\left[\widehat{ATE} - ATE\right] &= \frac{\sqrt{K}}{\sqrt{\frac{1}{K}\sum_{k=1}^{K}V\left(AD_k\right)}}\left[\widehat{ATE} - ATE\right]\\ &= \frac{\sqrt{K}}{\sqrt{KV\left(\widehat{ATE}\right)}}\left[\widehat{ATE} - ATE\right],\end{aligned}$$

so by the Slutsky Lemma and Point 3 of Assumption 11:

$$\sqrt{K}\left(\widehat{ATE} - ATE\right) \xrightarrow{d} N\left(0,\sigma^2\right).$$

Now we show that $K\widehat{V}_{clu}\left(\widehat{ATE}\right) \xrightarrow{p} \sigma_+^2 \geq \sigma^2$:

$$\begin{aligned}&\lim_{K\to\infty} K\widehat{V}_{clu}\left(\widehat{ATE}\right)\\ =\ &\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\left(AD_k - \widehat{ATE}\right)^2\\ =\ &\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}AD_k^2 - \left(\frac{1}{K}\sum_{k=1}^{K}AD_k\right)^2.\end{aligned}$$

By Assumption 9, Point 1 of Assumption 11, the strong law of large numbers in Lemma 1 of Liu et al. (1988), the fact that almost sure convergence implies convergence in probability, and Point 3 of Assumption 11,

$$\frac{1}{K}\sum_{k=1}^{K} AD_k^2 \xrightarrow{p} \lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k^2\right)$$
$$\frac{1}{K}\sum_{k=1}^{K} AD_k \xrightarrow{p} \lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k\right).$$

Therefore, by the continuous mapping theorem:

$$K\widehat{V}_{clu}\left(\widehat{ATE}\right) \xrightarrow{p} \lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k^2\right) - \left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k\right)\right)^2 = \sigma_+^2.$$

By the convexity of $x \to x^2$ we have $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k\right)^2 \geq \left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left(AD_k\right)\right)^2$, so $\sigma_+^2 \geq \sigma^2$.

C.1.7 Relaxing Assumption 6

Let $Y_{ik}(d) = f_{ikd}\left(\epsilon_{ik}(d), \eta_k(d)\right)$ for some functions $f_{ikd}(.)$. Redefine $ATE(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))$ as $\frac{1}{n}\sum_{i,k} E(Y_{ik}(1) - Y_{ik}(0)|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)))$, and ATE as $\frac{1}{n}\sum_{i,k} E(Y_{ik}(1) - Y_{ik}(0))$. It is trivial to show that $\widehat{ATE}$ is unbiased for ATE and conditionally unbiased for $E(Y_{ik}(1) - Y_{ik}(0)|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)))$, so we will focus on showing that our results regarding the variance and conditional variance of $\widehat{ATE}$ still hold.

Conditional Variance of $\widehat{ATE}$

(C.3) still holds. Starting with the first term in (C.3):

$$\begin{aligned}
&V\left(\mathbb{E}\left(\widehat{ATE}_k\middle|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\\
&=V\Bigg(\frac{1}{n_{1k}}\sum_{i=1}^{n_{1k}}D_{ik}\mathbb{E}(Y_{ik}(1)|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)))\\
&-\frac{1}{n_{0k}}\sum_{i=1}^{n_{0k}}(1-D_{ik})\mathbb{E}(Y_{ik}(0)|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1)))\Bigg|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\Bigg)
\end{aligned}$$

where the equality holds by Point 3 of Assumption 7. Now note that conditional on $(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))$, only the D_{ik}s are random inside the conditional variance operator. Then, it follows from Neyman (1923) that

$$\begin{aligned}
V\left(\mathbb{E}\left(\widehat{ATE}_k|\mathbf{D}_k,(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)\middle|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right)=&\frac{1}{n_{0k}}S^2_{\mathbb{E}^*(Y_{ik}(0)),k}+\frac{1}{n_{1k}}S^2_{\mathbb{E}^*(Y_{ik}(1)),k}\\
&-\frac{1}{n_k}S^2_{\mathbb{E}^*(Y_{ik}(1))-\mathbb{E}^*(Y_{ik}(0)),k}
\end{aligned}$$

where as before $\mathbb{E}^*(X) = \mathbb{E}(X|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)))$. Moving to the second term in (C.3),

$$\begin{aligned}
&\mathbb{E}\left(V\left(\widehat{ATE}_k \middle| \mathbf{D}_k, (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) \middle| (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) \\
&= \mathbb{E}\Bigg(\frac{1}{n_{1k}^2} V\left(\sum_{i=1}^{n_{1k}} D_{ik} Y_{ik}(1) \middle| \mathbf{D}_k, (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) \\
&+ \frac{1}{n_{0k}^2} V\left(\sum_{i=1}^{n_{1k}} (1 - D_{ik}) Y_{ik}(0) \middle| \mathbf{D}_k, (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) \Bigg| (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\Bigg) \\
&= \mathbb{E}\Bigg(\frac{1}{n_{1k}^2} \sum_{i=1}^{n_{1k}} D_{ik} V\left(Y_{ik}(1) | \boldsymbol{\eta}(0), \boldsymbol{\eta}(1)\right)\right) \\
&+ \frac{1}{n_{0k}^2} \sum_{i=1}^{n_{0k}} (1 - D_{ik}) V\left(Y_{ik}(0) | \boldsymbol{\eta}(0), \boldsymbol{\eta}(1)\right) \Bigg| (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\Bigg) \\
&= \frac{1}{n_{1k}^2} \sum_{i=1}^{n_{1k}} \mathbb{E}^*\left(D_{ik}\right) V\left(Y_{ik}(1) | (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) + \frac{1}{n_{0k}^2} \sum_{i=1}^{n_{0k}} \mathbb{E}^*\left(1 - D_{ik}\right) V\left(Y_{ik}(0) | (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) \\
&= \frac{1}{n_{1k}} \sum_{i=1}^{n_{1k}} \frac{1}{n_k} V\left(Y_{ik}(1) | (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) + \frac{1}{n_{0k}} \sum_{i=1}^{n_{0k}} \frac{1}{n_k} V\left(Y_{ik}(0) | (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right)
\end{aligned}$$

where the second equality holds because $V\left(Y_{ik}(d) | \mathbf{D}_k, \boldsymbol{\eta}(0), \boldsymbol{\eta}(1)\right) = V\left(Y_{ik}(d) | \boldsymbol{\eta}(0), \boldsymbol{\eta}(1)\right)$ by Point 3 of Assumption 7, and
$cov\left(Y_{ik}(d), Y_{jk}(d) | \mathbf{D}_k, \boldsymbol{\eta}(0), \boldsymbol{\eta}(1)\right) = cov\left(Y_{ik}(d), Y_{jk}(d) | \boldsymbol{\eta}(0), \boldsymbol{\eta}(1)\right) = 0$, by Point 3 of Assumption 7 and because $\epsilon_{ik}(d) \perp\!\!\!\perp \epsilon_{jk}(d) | \boldsymbol{\eta}(0), \boldsymbol{\eta}(1)$ by Points 2 and 4 of Assumption 7. The third equality holds because $V\left(Y_{ik}(1) | (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right)$ and $V\left(Y_{ik}(0) | (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right)$ are functions of $(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))$. The fourth equality holds because $\mathbb{E}^*\left(D_{ik}\right) = \mathbb{E}\left(D_{ik}\right)$ by Point 3 of Assumption 7. Finally, the first equality holds because:

$$\begin{aligned}
&\frac{1}{n_{1k} n_{0k}} Cov\left(\sum_{i=1}^{n_k} D_{ik} Y_{ik}(1), \sum_{i=1}^{n_k} (1 - D_{ik}) Y_{ik}(0) \middle| \mathbf{D}_k, (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) \\
&= \frac{1}{n_{1k} n_{0k}} \sum_{i=1}^{n_k} D_{ik}(1 - D_{ik}) Cov\left(Y_{ik}(1), Y_{ik}(0) | \mathbf{D}_k, (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) = 0.
\end{aligned}$$

The first equality holds because

$Cov(Y_{ik}(d), Y_{jk}(d)|\mathbf{D}_k, (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))) = Cov(Y_{ik}(d), Y_{jk}(d)|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))) = 0$ by Points 2, 3 and 4 of Assumption 7, the second equality holds because $D_{ik}(1 - D_{ik}) = 0$. Therefore,

$$
\begin{aligned}
& V\left(\widehat{ATE}_k \middle| (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) \\
& = \frac{1}{n_{0k}} S^2_{\mathbb{E}^*(Y_{ik}(0)),k} + \frac{1}{n_{1k}} S^2_{\mathbb{E}^*(Y_{ik}(1)),k} - \frac{1}{n_k} S^2_{\mathbb{E}^*(Y_{ik}(1)) - \mathbb{E}^*(Y_{ik}(0)),k} \\
& + \frac{1}{n_{1k}} \sum_{i=1}^{n_{1k}} \frac{1}{n_k} V\left(Y_{ik}(1)|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) + \frac{1}{n_{0k}} \sum_{i=1}^{n_{0k}} \frac{1}{n_k} V\left(Y_{ik}(0)|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) \qquad \text{(C.17)}
\end{aligned}
$$

Finally, by Lemma 1,

$$
V\left(\widehat{ATE} \middle| (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) = \frac{1}{K^2} \sum_{k=1}^{K} \left(\frac{n_k}{\overline{n}}\right)^2 V\left(\widehat{ATE}_k | ((\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right). \qquad \text{(C.18)}
$$

Estimating an Upper Bound for the Conditional Variance of $\widehat{ATE}$

$$
\begin{aligned}
& \frac{1}{n_{1k}} \mathbb{E}^* \left[\frac{1}{n_{1k} - 1} \sum_{i=1}^{n_k} D_{ik} \left(Y_{ik}(1) - \overline{Y}_{1k}\right)^2 \right] \\
& = \frac{1}{n_{1k}} \frac{1}{n_{1k} - 1} \mathbb{E}^* \left[\sum_{i=1}^{n_k} D_{ik} Y_{ik}(1)^2 - n_{1k} \overline{Y}^2_{1k} \right] \\
& = \frac{1}{n_{1k}} \frac{1}{n_{1k} - 1} \left[\sum_{i=1}^{n_k} \mathbb{E}^*(D_{ik} Y_{ik}(1)^2) - n_{1k} \mathbb{E}^*(\overline{Y}^2_{1k}) \right] \\
& = \frac{1}{n_{1k}} \frac{1}{n_{1k} - 1} \\
& \left[\sum_{i=1}^{n_k} \frac{n_{1k}}{n_k} \mathbb{E}^*(Y_{ik}(1)^2) - \frac{1}{n_{1k}} \sum_{i=1}^{n_k} \mathbb{E}^*(D^2_{ik} Y^2_{ik}(1)) - \frac{1}{n_{1k}} \sum\sum_{i \neq j} \mathbb{E}^*(D_{ik} D_{jk} Y_{ik}(1) Y_{jk}(1)) \right] \\
& = \frac{1}{n_{1k}} \frac{1}{n_{1k} - 1} \\
& \left[\sum_{i=1}^{n_k} \frac{n_{1k}}{n_k} \mathbb{E}^*(Y_{ik}(1)^2) - \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{E}^*(Y^2_{ik}(1)) - \frac{n_{1k} - 1}{n_k(n_k - 1)} \sum\sum_{i \neq j} \mathbb{E}^*(Y_{ik}(1)) \mathbb{E}^*(Y_{jk}(1)) \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n_{1k}} \frac{1}{n_{1k}-1} \\
&\Bigg[\sum_{i=1}^{n_k} \frac{n_{1k}-1}{n_k} [V\left(Y_{ik}(1)|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) + \mathbb{E}^*(Y_{ik}(1))^2] \\
&- \frac{n_{1k}-1}{n_k(n_k-1)} \sum\sum_{i \neq j} \mathbb{E}^*(Y_{ik}(1)) \mathbb{E}^*(Y_{jk}(1)) \Bigg] \\
&= \frac{1}{n_{1k}} \frac{1}{n_{1k}-1} \left[\sum_{i=1}^{n_k} \frac{n_{1k}-1}{n_k} V\left(Y_{ik}(1)|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) \right. \\
&+ \frac{n_{1k}-1}{n_k} \sum_{i=1}^{n_k} \mathbb{E}^*(Y_{ik}(1))^2 + \frac{n_{1k}-1}{n_k(n_k-1)} \sum_{i} \mathbb{E}^*(Y_{ik}(1))^2 \\
&\left. - \frac{n_{1k}-1}{n_k(n_k-1)} \sum\sum_{i,j} \mathbb{E}^*(Y_{ik}(1)) \mathbb{E}^*(Y_{jk}(1)) \right] \\
&= \frac{1}{n_{1k}} \left[\frac{1}{n_k} \sum_{i=1}^{n_k} V\left(Y_{ik}(1)|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right) \right] \\
&+ \frac{1}{n_{1k}} \frac{1}{n_k-1} \left[\sum_{i=1}^{n_k} \mathbb{E}^*(Y_{ik}(1))^2 - n_k \overline{\mathbb{E}^*(Y_{ik}(1))}^2 \right] \\
&= \frac{1}{n_{1k}} S^2_{\mathbb{E}^*(Y_{ik}(1)),k} + \frac{1}{n_{1k}} \sum_{i=1}^{n_{1k}} \frac{1}{n_k} V\left(Y_{ik}(1)|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right). \qquad \text{(C.19)}
\end{aligned}
$$

The third equality holds by Point 3 of Assumption 7 and by Equation (C.1). The fourth equality holds because treatment is binary, by Point 3 of Assumption 7, by Points 2 and 4 of Assumption 7, and by Equations (C.1) and (C.2). The fifth equality holds by the definition of a conditional variance. The remaining equalities hold by algebraic manipulations.

Using similar arguments, one can show:

$$
\begin{aligned}
&\mathbb{E}\left(\frac{1}{n_{0k}} \left[\frac{1}{n_{0k}-1} \sum_{i=1}^{n_k} (1 - D_{ik}) \left(Y_{ik}(0) - \overline{Y}_{0k}\right)^2 \right] \Bigg| (\boldsymbol{\eta}(0), \boldsymbol{\eta}(1)) \right) \\
=& \frac{1}{n_{0k}} S^2_{\mathbb{E}^*(Y_{ik}(0)),k} + \frac{1}{n_{0k}} \sum_{i=1}^{n_{0k}} \frac{1}{n_k} V\left(Y_{ik}(0)|(\boldsymbol{\eta}(0), \boldsymbol{\eta}(1))\right). \qquad \text{(C.20)}
\end{aligned}
$$

Equations (C.17), (C.19), and (C.20) show that $V\left(\widehat{ATE}_k|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right) \leq \mathbb{E}^*\left(\widehat{V}_{rob}\left(\widehat{ATE}_k\right)\right)$. Then, it follows from Equation (C.18) and the definition of $\widehat{V}_{rob}\left(\widehat{ATE}_k\right)$ that

$$V\left(\widehat{ATE}|(\boldsymbol{\eta}(0),\boldsymbol{\eta}(1))\right) \leq \mathbb{E}^*\left(\widehat{V}_{rob}\left(\widehat{ATE}\right)\right).$$

Estimating an Upper Bound for the Variance of $\widehat{ATE}$

The proof in Section 6.3.3 does not make use of Assumption 6 so the result is already proven.

Bibliography

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. (2017). When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research.

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296.

Abdulkadiroglu, A., Parag, P., Schellenberg, J., and Walters, C. (2020). Do parents value school effectiveness? *American Economic Review*, 110(5):1502–1539.

Angrist, J., Autor, D., and Pallais, A. (2020). Marginal effects of merit aid for low-income students. Technical report, NBER Working Paper No. 27834.

Angrist, J., Lang, D., and Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 1(1):136–63.

Angrist, J., Oreopoulos, P., and Williams, T. (2014). When opportunity knocks, who answers? new evidence on college achievement awards. *Journal of Human Resources*, 49(3):572–610.

Angrist, J. D., Hull, P. D., Pathak, P. A., and Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, 132(2):871–919.

Arcidiacono, P., Kinsler, J., and Price, J. (2017). Productivity spillovers in team production: Evidence from professional basketball. *Journal of Labor Economics*, 35(1):191–225.

Avery, C., Howell, J. S., and Page, L. (2014). *A Review of the role of college counseling, coaching, and mentoring on students' postsecondary outcomes*. College Board Research Brief.

Barr, A. (2016). Enlist or enroll: Credit constraints, college aid, and the military enlistment margin. *Economics of Education Review*, 51:61–78.

Barr, A. and Castleman, B. (2018). An engine of economic opportunity: Intensive advising, college success, and social mobility. Texas A&M Working Paper,.

Barr, A. and Castleman, B. (2019). Exploring variation in college counselor effectiveness. *AEA Papers and Proceedings*, 109:227–231.

Bertrand, M. and Schoar, A. (2003). Managing with style: The effect of managers on firm policies. *The Quarterly journal of economics*, 118(4):1169–1208.

Bettinger, E., Gurantz, O., Kawano, L., and Sacerdote, B. (2016). The long run impacts of merit aid: Evidence from california's cal grant. Technical report, NBER Working Paper No. 22347.

Bettinger, E. P. and Baker, R. B. (2014). The effects of student coaching: An evaluation of a randomized experiment in student advising. *Educational Evaluation and Policy Analysis*, 36(1):3–19.

Bettinger, E. P., Long, B. T., Oreopoulos, P., and Sanbonmatsu, L. (2012). The role of application assistance and information in college decisions: Results from the h&r block fafsa experiment. *The Quarterly Journal of Economics*, 127(3):1205–1242.

Beuermann, D., Jackson, C. K., Navarro-Sola, L., and Pardo, F. (2020). What is a good school, and can parents tell? evidence on the multidimensionality of school output. *National Bureau of Economic Research, No. w25342*.

Bird, K. A., Castleman, B. L., Denning, J. T., Goodman, J., Lamberton, C., and Rosinger, K. O. (2021). Nudg- ing at scale: Experimental evidence from fafsa completion campaigns. *Journal of Economic Behavior & Organization*, 183:105–128.

Blau, F. D., Currie, J. M., Croson, R. T. A., and Ginther, D. K. (2010). Can mentoring help female assistant professors? interim results from a randomized trial. *American Economic Review Papers & Proceedings*, 100(2):348–352.

Bo, H. and Galiani, S. (2019). Assessing external validity. Technical report, National Bureau of Economic Research.

Board, C. (2011). How four-year colleges and universities organize themselves to promote student persistence: The emerging national picture. Technical report.

Bound, J., Lovenheim, M. F., and Turner, S. (2010). Why have college completion rates declined? an analysis of changing student preparation and collegiate resources. *American Economic Journal: Applied Economics*, 2(3):129–157.

Buckles, K. (2019). Fixing the leaky pipeline: Strategies for making economics work for women at every stage. *Journal of Economic Perspectives*, 33(1):43–60.

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2018). Inference under covariate adaptive randomization. *Journal of the American Statistical Association, forthcoming.*, 113(524):1784–1796.

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, 10(4):1747–1785.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2012). Robust inference with multiway clustering. *Journal of Business & Economic Statistics.*

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2):317–372.

Canaan, S., Deeb, A., and Mouganie, P. (2021). Advisor value-added and student outcomes: Evidence from randomly assigned college advisors. *American Economic Journal: Economic Policy.*

Canaan, S. and Mouganie, P. (2018). Returns to education quality for low-skilled students: Evidence from a discontinuity. *Journal of Labor Economics*, 36(2):395–436.

Canaan, S. and Mouganie, P. (2022). *The Impact of advisor gender on female students' STEM enrollment and persistence.* Journal of Human Resources.

Canay, I. A., Santos, A., and Shaikh, A. M. (2019). The wild bootstrap with a "small" number of "large" clusters.

Carrell, S. E., Page, M. E., and West, J. E. (2010). Sex nd science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics*, 125(3):1101–1144.

Carrell, S. E. and Sacerdote, B. I. (2017). Why do college-going interventions work? *American Economic Journal: Applied Economics*, 9(3):124–151.

Carrell, S. E. and West, J. E. (2010). Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3):409–432.

Castleman, B. and Goodman, J. (2018). Intensive college counseling and the enrollment and persistence of low-income students. *Education Finance and Policy*, 13(1):19–41.

Castleman, B. L. and Long, B. T. (2016). Looking beyond enrollment: The causal effect of need-based grants on college access, persistence, and graduation. *Journal of Labor Economics*, 34(4):1023–1073.

Castleman, B. L., Page, L. C., and Schooley, K. (2014). “the forgotten summer: Does the offer of college counseling after high school mitigate summer melt among college-intending, low-income high school graduates?”. *Journal of Policy Analysis and Management*, 33(2):320–344.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of econometrics*, 34(3):305–334.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review*, 104(9):2633–79.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2017). Measuring the impacts of teachers: Reply. *American Economic Review*, 107(6):1685–1717.

Clearinghouse, N. S. (2018). *First-year persistence and re- tention. Snapshot report.*

Cole, S., Giné, X., Tobacman, J., Topalova, P., Townsend, R., and Vickery, J. (2013). Barriers to household risk management: Evidence from india. *American Economic Journal: Applied Economics*, 5(1):104–35.

Currie, J. and Zhang, J. (2021). Doing more with less: Predicting primary care provider effectiveness. Technical report, National Bureau of Economic Research.

Davezies, L., D'Haultfoeuille, X., and Guyonvarch, Y. (2019). Empirical process results for exchangeable arrays. *arXiv preprint arXiv:1906.11293*.

Dehejia, R., Pop-Eleches, C., and Samii, C. (2019). From local to global: External validity an a fertility natural experiment. *Journal of Business & Economic Statistics*, (just-accepted):1–48.

Deming, D. J. (2017). Increasing college completion with a federal higher education matching grant. *The Hamilton Project, Policy Proposal*, pages 2017–03.

Deming, D. J. and Walters, C. R. (2017). The impact of price caps and spending cuts on us postsecondary attainment. Technical report, NBER Working Paper No. 23736.

Denning, J. T. and Turley, P. (2017). Was that smart? institutional financial incentives and field of study. *Journal of Human Resources*, 52(1):152–186.

Dobronyi, C. R., Oreopoulos, P., and Petronijevic, U. (2019). Goal setting, academic reminders, and college success: A large-scale field experiment. *Journal of Research on Educational Effectiveness*, 12(1):38–66.

Donner, A. and Klar, N. (2000). Design and analysis of cluster randomization trials in health research.

Dynarski, S. (2003). Does aid matter? measuring the effect of student aid on college attendance and completion. *American Economic Review*, 93(1):279–288.

Eicker, F. et al. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics*, 34(2):447–456.

Evans, B. J. (2017). Smart money: Do financial incentives encourage college students to study science? *Education Finance and Policy*, 12(3):342–368.

Gilraine, M., Gu, J., and McMillan, R. (2020). A new method for estimating teacher value-added. Working Paper 27094, National Bureau of Economic Research.

Guarino, C. M. and Borden, V. M. (2017). Faculty service loads and gender: Are women taking care of the academic family? *Research in Higher Education*, 58(6):672–694.

Gurantz, O., Pender, M., Mabel, Z., Larson, C., and Bettinger, E. (2020). Virtual advising for high-achieving high school students. *Economics of Education Review*, 75(10197):4.

Hahn, J., Kuersteiner, G., and Mazzocco, M. (2020). Estimation with aggregate shocks. *The Review of Economic Studies*, 87(3):1365–1398.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Hastings, J. S., Neilson, C. A., and Zimmerman, S. D. (2013). Are some degrees worth more than others? evidence from college admission cutoffs in chile. Technical report, NBER Working Paper No. 19241.

Huber, J. A. and Miller, M. A. (2011). Chapter 11- advisor job responsibilities – four year institutions. *2011 NACADA National Survey of Academic Advising.*

Huber, P. J. et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, 126(5):2072–2107.

Jackson, C. K., Rockoff, J. E., and Staiger, D. O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics*, 26(1):801–825.

Jacob, B. A., Lefgren, L., and Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human resources*, 45(4):915–943.

Kane, T. J., Rockoff, J. E., and Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education*, 27:615–631.

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Karlan, D., Osei, R., Osei-Akoto, I., and Udry, C. (2014). Agricultural decisions after relaxing credit and risk constraints. *The Quarterly Journal of Economics*, 129(2):597–652.

Kirkeboen, L. J., Leuven, E., and Mogstad, M. (2016). Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, 131(3):1057–1111.

Koedel, C., Mihaly, K., and Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education*, 47:180–195.

Kot, F. C. (2014). The impact of centralized advising on first-year academic performance and second-year enrollment behavior. *Research in higher education*, 55(6):527–563.

Lazear, E. P., Shaw, K. L., and Stanton, C. T. (2015). The value of bosses. *Journal of Labor Economics*, 33(4):823–861.

Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Liu, J. and Loeb, S. (2021). Engaging teachers measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources*, 56(2):343–379.

Liu, R. Y. et al. (1988). Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16(4):1696–1708.

Malcom, S. and Feder, M. (2016). *Barriers and opportunities for 2-year and 4-year STEM degrees: Systemic change to support students' diverse pathways.* National Academies Press, Washington, DC.

Menzel, K. (2018). Bootstrap with cluster-dependence in two or more dimensions. *ArXiv eprints, New York University.*

Mulhern, C. (2019). Beyond teachers: Estimating individual guidance counselors' effects on educational attainment.

Murray, D. M. et al. (1998). *Design and analysis of group-randomized trials*, volume 29. Oxford University Press, USA.

NCES (2018). *Digest of education statistics.* National Center for Education Statistics.

Newey, K. and McFadden, D. (1994). Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, pages 2112–2245.

Newey, W. K. (1984). A method of moments interpretation of sequential estimators. *Economics Letters*, 14(2-3):201–206.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. translated in Statistical Science 5(4), 465-472, 1990.

of Advisors on Science, P. C. and Technology (2012). *"Engage and excel: producing one million additional college graduates with degrees in science, technology, engineering, and mathematics."*. Report to the President. Executive Office of the President. Washington, DC.

Opper, I. M. (2019). Does helping john help sue? evidence of spillovers in education. *American Economic Review*, 109(3):1080–1115.

Oreopoulos, P. and Petronijevic, U. (2013). Making college worth it: A review of the returns to higher education. *The Future of children*, 23(1):41–65.

Oreopoulos, P. and Petronijevic, U. (2019). The remarkable unresponsiveness of college students to nudging and what we can learn from it. Technical report, NBER Working Paper No. 26059.

Pagan, A. (1986). Two stage and related estimators and their applications. *The Review of Economic Studies*, 53(4):517–538.

Patterson, R. W., Pope, N. G., and Feudo, A. (2022). *Timing is everything: Evidence from college major decisions.* Journal of Human Resources.

Phillips, M. and Sarah, J. R. (2019). "does virtual advising increase college enrollment? evidence from a random assignment college access field experiment.". Technical report, NBER Working Paper No. 26509.

Porter, C. and Serra, D. (2020). Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics*, 12(3):226–254.

Riddell, C. and Riddell, C. (2020). Interpreting experimental evidence in the presence of post-randomization events: A re-assessment of the self sufficiency project. *Journal of Labor Economics*.

Rose, E., Schellenberg, J., and Shem-Tov, Y. (2019). The effects of teacher quality on criminal behavior.

Rosenzweig, M. R. and Udry, C. (2019). External Validity in a Stochastic World: Evidence from Low-Income Countries. *The Review of Economic Studies*.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, 4(4):537–571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1):175–214.

Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic Review*, 107(6):1656–84.

Scrivener, S. and Weiss, M. J. (2009). *More guidance, better results? Three-year effects of an enhanced student services program at two community colleges*. MDRC, New York.

Sjoquist, D. L. and Winters, J. V. (2015). State merit aid programs and college major: A focus on stem. *Journal of Labor Economics*, 33(4):973–1006.

Staiger, D. O. and Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, 24(3):97–118.

Stange, K. (2015). Differential pricing in undergraduate education: Effects on degree production by field. *Journal of Policy Analysis and Management*, 34(1):107–135.

Sullivan, Z., Castleman, B., and Bettinger, E. (2019). College advising at a national scale: Experimental evidence from the collegepoint initiative. *Annenberg Ed Working Paper*, pages 19–123.

Tinto, V. (2010). From theory to action: Exploring the institutional conditions for student retention. In *In Higher education: Handbook of Theory and Research: , Dordrecht*, pages 51–89.

Weiss, M. J., Ratledge, A., Sommo, C., and Gupta, H. (2019). Supporting community college students from start to degree completion: Long-term evidence from a randomized trial of cuny's asap. *American Economic Journal: Applied Economics*, 11(3):253–97.

White, H. et al. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *econometrica*, 48(4):817–838.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review*, 93(2):133–138.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.