The Predictive Validity of Information From Clinical Practice Lessons: Experimental Evidence From Argentina

The Predictive Validity of Information from Clinical Practice Lessons: Experimental Evidence from Argentina

Alejandro J. Ganimian

Richard J. Murnane

{David J. Deming

Felipe Barrera-Osorio}

A Thesis Presented to the Faculty

of the Graduate School of Education of Harvard University

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Education

2015

Dedications

I dedicate this thesis to my family and friends, who have supported me unconditionally for the past six years. I also dedicate it to my many mentors at the Harvard Graduate School of Education (HGSE), who have been incredibly generous with their time, advice, and opportunities: Felipe Barrera-Osorio, Dave Deming, Andrew Ho, Tom Kane, Susan Moore Johnson, Dan Koretz, Dick Murnane, Emiliana Vegas, and Marty West.

Acknowledgements

This thesis arises from joint work with Andrew D. Ho and Mariana Alfonso. As a result of this collaboration, I use "we" throughout this thesis. We gratefully acknowledge the funding provided by the Inter-American Development Bank (IDB) and the Harvard Graduate School of Education (HGSE) for this project. We thank the staff members of *Enseñá por Argentina* (ExA) who contributed to the design and implementation of this project. We are especially grateful to Pedro Cosme Cavallo, Fabiola Fares, Pablo Princz, and Fernando Viola. We also thank Elena Arias, Felipe Barrera-Osorio, María Emilia Buchannan, Susana Claro, Miguel Costa, David Deming, Yyannú Cruz-Aguayo, Fabiola Fares, Ariel Fiszbein, Oscar Ghillione, María Gil Izquierdo, Carlos Guedes, Heather Hill, Kendra Hennig, Brian Jacob, Tom Kane, Wendy Kopp, John Krieg, Octavio Lizama, Bárbara Loza, Juan Pablo Martínez, Andrew McEachin, Dick Murnane, Katherine Onorato, Pablo Princz, Jeff Puryear, Alonso Sánchez, Ismael Sanz, Norbert Schady, Ernesto Schargrodsky, Andy Sokatch, Ana Tejedor de la Mata, Emiliana Vegas, Fernando Viola, the participants of the APPAM, LACEA, AEFP, and SREE annual meetings, and of the seminars at the Universidad Camilo José Cela, Fundación Sociedad y Educación, Universidad Autónoma de Madrid, and the Inter-American Development Bank for their input at various stages of this project. The usual disclaimers apply. The pre-analysis plan for this study can be accessed at: http://bit.ly/1e9spsk.

Table of Contents

Abstract

A growing number of teacher preparation programs require trainees to practice teaching. Yet, there is almost no evidence on whether the performance of individuals during clinical practice lessons predicts how they fare once they enter the school system.

We address this question by taking advantage of the fact that an alternative pathway into teaching in Argentina requires admitted applicants to complete two weeks of clinical practice. We collect information both during clinical practice and the school year. During clinical practice, we measure the performance of teaching trainees using classroom observations and student surveys. During the school year, we measure their performance using classroom observations, student surveys, and principal surveys.

We find that the overall performance of trainees during clinical practice predicts their overall performance during the school year, but this prediction is only statistically under certain model specifications. The performance of these individuals during clinical practice predicts their ratings on classroom observations during the school year. This relationship remains statistically significant even when we account for how trainees fare on the application and selection processes of the alternative pathway.

We also find that the performance of trainees on a brief demonstration lesson, delivered during the selection process, predicts their performance on classroom observations during the school year. The predictive effect is smaller than that of clinical practice lessons, but it raises the question of whether the additional effort required to collect information during clinical practice is worth the improved predictive validity.

1. Introduction

A growing number of teacher preparation programs require trainees to practice teaching. In the United States, there are over 30 "teacher residency" programs in which individuals with a bachelor's degree simultaneously complete coursework and have a supervised fieldwork experience of at least one year before being hired by the school system (Silva, McKie, Knechtel, Gleason, & Makowsky, 2014). Additionally, there are 36 alternative pathways into teaching across the world. In these programs, college graduates complete two weeks of workshops on pedagogy and leadership and participate in two weeks of "clinical practice," teaching a group of volunteer students and receiving feedback from instructional coaches before they start working in schools (Glazerman, Mayer, & Decker, 2006).

This requirement is becoming increasingly popular partly because observing an individual teaching is expected to yield information about his or her instructional skills that is not captured by other types of assessments typically conducted during teacher training (e.g., written and oral exams).[1] This theory of action is highly intuitive, but there is surprisingly little evidence on whether it works as expected.

There are several reasons why what individuals do while they are practicing teaching might not be indicative of how they will teach once they are working in schools. In practice lessons, teachers experiment with new approaches, possibly failing the first time they attempt a new task, and/or constantly learning and adjusting their practice. It is

[1] Additionally, practice teaching offers an opportunity for trainees to anticipate some of the challenges that they will face in their first few years in the profession and accelerate their learning. It might also dissuade individuals who are not serious about entering teaching, although that is hardly ever its main goal.

not clear that their performance during this trial-and-error period should predict how they fare once they figured out what classroom strategies work best for them. If practice lessons predict how individuals perform during the school year, however, we could use this information to make decisions about teacher training, allocation, and support.

In this paper, we examine whether the performance of teaching trainees during clinical practice predicts their performance during the school year. We take advantage of the fact that an alternative pathway into teaching in Argentina requires admitted applicants to go through two weeks of clinical practice and we collect additional information both during clinical practice and the school year. During clinical practice, we measure the performance of teaching trainees using classroom observations and student surveys. During the school year, we measure their performance using classroom observations, student surveys, and principal surveys.

We find that the average score that individuals receive across both instruments administered during clinical practice predicts their average score across the three instruments administered during the school year. The predictive effect is large: for every standard deviation unit in clinical practice scores, an individual performs on average .611 standard deviations better during the school year. Yet, this relationship is only statistically significant when we include covariates.

The performance of teaching trainees during clinical practice also predicts their performance on classroom observations conducted during the school year: for every standard deviation unit in clinical practice scores, an individual performs on average .72 standard deviations better on these observations. This relationship is statistically

significant even when we account for how individuals fare on the application and selection processes of the alternative pathway.

We also find that the performance of trainees on a brief demonstration lesson, delivered during the selection process, predicts their performance on classroom observations during the school year. For every standard deviation in the demonstration lesson score, a corps member performs on average .348 standard deviations better on the classroom observations. This raises the question of whether the additional effort required to collect information during clinical practice is worth the improved predictive validity.

The paper is structured as follows. Section 2 reviews prior research. Section 3 describes the experiment we conducted. Section 4 introduces the sample and variables in our analysis. Section 5 presents the empirical strategy. Section 6 reports the results. Finally, Section 7 discusses the policy implications.

2. Prior Research

Over the past two decades, many studies have shown that good teaching matters. The earlier studies in this literature found that some teachers had students who consistently performed better than expected, while others had students who consistently performed below expectations (Aaronson, Barrow, & Sander, 2007; Hanushek & Rivkin, 2010b; Jacob & Lefgren, 2008; Kane, Rockoff, & Staiger, 2008; Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004).[2]

[2] These studies use "value-added models" that estimate teaching effectiveness based on a two-step process. First, they predict how the students assigned to a teacher are expected to perform on a standardized test based on their prior achievement, socio-economic status, race or ethnicity, and/or peer group. Then, they calculate how much better or worse than expected these students actually performed.

Recent studies found that these differences in the performance of students on standardized tests across teachers remained when teachers were randomly assigned to schools (Glazerman, Protik, Teh, Bruch, & Max, 2013) and to classrooms within schools (Chetty et al., 2011; Kane & Staiger, 2008). Some studies found that these differences were also related to other metrics of teacher performance (e.g., classroom observations and student surveys) (Kane, McCaffrey, Miller, & Staiger, 2013), and to students' long-run outcomes (Chetty, Friedman, & Rockoff, 2014).

The importance of good teaching has fueled four decades of research into whether it is possible to predict which teachers will consistently raise student achievement using information collected at the time of teacher selection. Early studies examined the predictive validity of the information that school districts collect about their teachers. Yet, nearly every metric found to be predictive of differences in student achievement gains across teachers in one study was later found not to be predictive in another study—including teacher certification, licensure exams, graduate degrees, and college selectivity (Clotfelter, Ladd, & Vigdor, 2007; Darling-Hammond, Berry, & Thoreson, 2001; Darling-Hammond, Holtzman, Gatlin, & Heilig, 2005; Goldhaber & Brewer, 2000; Hanushek, Kain, O'Brien, & Rivkin, 2005; Hanushek & Rivkin, 2006; Kane et al., 2008).

In recent studies, researchers administered their own instruments to measure the academic ability, socio-emotional skills, and subject-matter pedagogical knowledge of entrants into teaching (Bastian, 2013; Dobbie, 2011; Duckworth, Quinn, & Seligman, 2009; Gitomer, Phelps, Weren, Howell, & Croft, 2014; Hill, Kapitula, & Umland, 2011; Metzler & Woessmann, 2012). By themselves, each of these metrics explained a relatively small share of differences in student achievement gains across teachers. Yet,

when these instruments were used together, they predicted moderate to large differences (Rockoff, Jacob, Kane, & Staiger, 2011).

Research on the predictive validity of teacher selection metrics, however, remains limited on at least three fronts. First, most studies explore whether new instruments can predict differences in student achievement gains across teachers, but they seldom examine whether these instruments predict these differences after accounting for teachers' performance on application and/or selection processes already in place.[3] This is an important limitation in the literature because the instruments that school systems use are typically cheaper and easier to administer than the ones developed by researchers. Additionally, reforming teacher selection systems is highly contentious.

Second, nearly all of the studies in this literature focus on identifying effective teachers on subjects that are tested on consecutive grades—mostly, math and reading. This research offers little guidance to assess the predictive validity of teacher selection metrics in most countries other than the United States (especially, developing countries), which lack such testing infrastructure. It also has limited implications for the selection of teachers in non-tested grades, even in countries with extensive student assessments.

Prior research has found that, once teachers enter the school system, differences in student achievement gains across teachers correlate with several instruments that can be administered across subjects, such as classroom observations (Grossman et al., 2013; Kane & Staiger, 2012; Kane, Taylor, Tyler, & Wooten, 2011; Taylor & Tyler, 2012),

[3] Rockoff et al. (2011) is an exception. The authors compared the predictive validity of instruments designed to measure teachers' cognitive ability, subject-specific pedagogical knowledge, personality traits, and beliefs about self-efficacy with the predictive validity of administrative information available to a school district and teachers' performance on a commercial teacher selection instrument.

principal surveys (Rockoff, Staiger, Kane, & Taylor, 2012), student surveys (Kane & Staiger, 2011), and subjective evaluations (Rockoff & Speroni, 2011). Yet, no study has assessed the predictive validity of these metrics *before* teachers start working in schools.[4]

Finally, individuals are rarely randomly assigned to raters or classrooms in studies of the predictive validity of teacher selection metrics. This allows for the possibility that any relationships between these metrics and subsequent differences in student achievement gains across teachers are attributable to more able teachers being assigned to advantaged students, and less able teachers being assigned to disadvantaged students.[5]

3. Experiment

The question this study aims to answer is: does the performance of teaching trainees during clinical practice predict their performance during the school year? Specifically, we are interested in whether the performance of these individuals during clinical practice adds information that can improve our predictions of their performance during the school year and that is not captured by other metrics that are already available. To address this question, we take advantage of the fact that an alternative pathway into teaching in Argentina requires its admits to go through two weeks of clinical practice.

---

[4] Predicting teachers' performance on these instruments early in their careers might be challenging because teachers' capacity to raise student achievement typically increases in their first few years of service (Atteberry, Loeb, & Wyckoff, 2013; Boyd, Lankford, Loeb, Rockoff, & Wyckoff, 2008; Harris & Sass, 2011; Papay & Kraft, 2011). This is especially true for teachers who undergo clinical practice (Papay, West, Fullerton, & Kane, 2012).

[5] Kane et al. (2013) and Araujo, Carneiro, Cruz-Aguayo, & Schady (2014) assessed the predictive validity of multiple metrics of teaching effectiveness using random assignment in the U.S. and Ecuador, respectively. Yet, these studies were conducted once teachers had already entered the profession.

Our experiment differs from previous studies in three important aspects. First, we observe how trainees perform on the application and selection processes of their alternative pathway. Therefore, we can examine whether their performance during clinical practice adds predictive information not captured in these previous stages. Second, we administer multiple instruments to measure trainees' performance during clinical practice and the school year that can be applied to all teachers—not just teachers of subjects tested in adjacent grades.[6] Third, individuals are randomly assigned to raters at all stages of our study, and they are also randomly assigned to students during clinical practice. Thus, we can reduce the possibility that any relationship that we observe between clinical practice and school year metrics is due to selection bias.[7]

Context

Enseñá por Argentina (ExA) is a non-profit founded in 2009 that recruits college graduates with disciplinary majors (economics, engineering, psychology, etc.) to teach in hard-to-staff schools for two years.[8] Its dual mission is to provide low-income students with effective teachers and to transform its corps members into leaders for education reform, regardless of whether they stay in education after their two-year commitment. ExA is an adaptation of Teach for America (TFA) in the United States. It follows largely

[6] This was particularly important in the context of our study, since Argentina's national student assessment occurs once every three years, results are disseminated two years after its administration, and the National Education Law prohibits the disaggregation of its results at the student, teacher, or school level.

[7] We cannot, however, eliminate this possibility because assignment of individuals to schools and classrooms during the school year is not random. We return to this issue in the empirical strategy section.

[8] ExA places most of its teachers in high schools, but some work at the pre-school and elementary levels.

TFA's strategies to recruit, select, and train its corps members. ExA, TFA, and 34 similar organizations around the world are part of the Teach for All (TFALL) network.

As Table 1 shows, an individual goes through four stages from the moment that he or she applies to the program to when he or she graduates from it. First, ExA invites all those interested in the program to submit an online application and it reviews these applications (stage 1).[9] Then, ExA calls back the candidates with the most promising applications to participate in an "assessment center", which includes two parts. The first part includes a one-on-one interview. The second part includes a demonstration lesson, a written exercise, a critical thinking assessment, and a group discussion (stage 2).[10] Next, ExA offers the best performers in the assessment center admission to the program and it requires that they attend a "summer training institute", which includes two weeks of workshops and two weeks of clinical practice teaching math, English, or Spanish to a

[9] Applicants who score above pre-determined cutoffs for all three rubrics administered during stage 1 automatically move on to stage 2. Applicants who score above the cutoffs for two of these rubrics are considered by a committee that decides whether they move on. Applicants who score below two or more of these cutoffs are automatically disqualified.

[10] All candidates participate in the second part, regardless of their performance on the first part. Applicants who score above cutoffs for both rubrics administered during part one and all five rubrics administered during part two are automatically offered a spot on the program and move on to stage 3. Applicants who score above the cutoffs for most of these rubrics are considered by a committee that decides whether they move on. Applicants who score below two or more of these cutoffs are automatically disqualified. The critical thinking assessment includes 15 multiple-choice questions that require applicants to analyze data, draw cause-and-effect connections, and draw conclusions from both of these processes.

group of volunteer students (stage 3).[11] Finally, it places these corps members in schools, where they teach for two years (stage 4).[12]

<Table 1>

Instruments

As Table 1 indicates, in stage 1, ExA uses structured rubrics to score applicants. These rubrics measure three competencies: (a) accomplishment (i.e., whether an applicant has achieved ambitious, measurable results in academics); (b) leadership (i.e., his or her experience leading others in extracurricular activities and/or jobs); and (c) perseverance (i.e., whether he/she overcomes obstacles with determination).[13]

In stage 2, ExA uses another set of rubrics. The rubrics for the first part of the assessment center measure two competencies: (a) organization (i.e., whether an applicant plans well and manages responsibilities successfully); and (b) communication (i.e., whether he/she expresses his/her ideas effectively).[14] The rubrics for the second part

---

[11] The two weeks of workshops focus on: (a) organizational culture; (b) leadership; (c) the social context of schools; and (d) pedagogy (e.g., lesson planning, assessments, and classroom management). These workshops count towards a teacher certification program in which all corps members must enroll. Corps members continue participating in this program at night during their two-year commitment. Once they finish their two-year commitment, they receive an official teaching certificate.

[12] All of these stages are compulsory for all corps members; opting out is not allowed.

[13] In 2014, ExA also assessed alignment (i.e., whether an applicant shares ExA's mission and vision). Rubrics are protected by confidentiality agreements between TFALL and ExA, so we cannot disclose them. Previous studies in the U.S. found the scores that corps members received on these rubrics were predictive of small differences in student achievement gains (Bastian, 2013; Dobbie, 2011).

[14] ExA also scores applicants on critical thinking (i.e., whether they are good problem solvers), but it does not use this score to decide which applicants are offered a spot in the program.

measure five competencies: (a) openness to feedback (i.e., whether an applicant listens to criticism and reacts adequately); (b) alignment (i.e., whether he/she shares ExA's mission and vision); and three competencies that have already been measured in different ways in previous stages: (c) leadership; (d) perseverance; and (e) communication.[15]

The demonstration lesson at the assessment center is particularly interesting to us because it is the first time that ExA observes its corps members teaching. There are three reasons why demonstration lessons are different from clinical practice lessons. First, they occur only once and last five minutes. Second, they are assessed by one rater, an ExA staffer who may or may not have prior teaching experience. Third, they do not include students.[16] Instead, ExA breaks applicants into small groups and asks them to "act like students," asking question whenever a fellow applicant is teaching. The rubric that assesses these lessons measures five competencies: (a) planning; (b) organization; (c) student engagement; (d) communication; and (e) listening and answering skills.[17]

Until 2014, ExA did not collect any information during stage 3. In 2014, we asked ExA to administer classroom observations and student surveys to measure corps members' performance.[18] The classroom observation protocol used during clinical practice assesses six competencies: (a) whether a corps member presents content clearly; (b) whether he/she checks that students understand the material; (c) whether he/she

---

[15] ExA also scores applicants on respect for diversity (i.e., whether they have high expectations for low-income students), but it does not use this score to decide which applicants are offered a spot in the program.

[16] Applicants are asked to prepare this lesson in advance and are given a template for a lesson plan.

[17] Applicants do not know the content of any of the selection rubrics, either before or after they apply.

[18] The 2014 cohort was introduced to all rubrics during the first two weeks of the summer training institute. All corps members in this cohort received copies of the rubrics. They were not administered in 2013.

effectively manages student behavior; (d) whether he/she consistently implements classroom procedures; (e) whether he/she creates an environment conducive to learning; and (f) whether he/she convinces students that their effort is important for their success.[19]

The student surveys administered during clinical practice assess seven competencies: (a) care (i.e., the ability of a corps member to show concern and commitment for his or her students); (b) confer (i.e., his/her capacity to invite students' ideas and promote discussion); (c) captivate (i.e., whether he/she inspires students' interest); (d) clarify (i.e., his/her ability to cultivate students' understanding and help them overcome confusion); (e) consolidate (i.e., his/her capacity to integrate students' ideas and check for their understanding); (f) challenge (i.e., whether he/she demands rigor); and (g) control (i.e., his/her ability to sustain order, respect, and focus).[20]

Until 2014, ExA did not collect any information during stage 4. In 2014, we asked ExA to administer classroom observations, student surveys, and principal surveys to measure corps members' performance.[21] The classroom observation protocol used during the school year assesses eight competencies. The first six are those assessed during clinical practice. The last two are only measured at this instance: (a) whether a

---

[19] The protocol is available at: http://bit.ly/1BcUzgR. It draws on five protocols in the U.S.: (a) the Classroom Assessment Scoring System; (b) the Framework for Teaching; (c) Teaching As Leadership; and the non-discipline specific parts of (d) the Protocol for Language Arts Teaching Observation; and (e) Mathematical Quality of Instruction.

[20] The survey is a Spanish translation of the abridged version of the Tripod survey. It is available here: http://bit.ly/1o1g9UW (elementary) and here: http://bit.ly/1qg71Yr (secondary).

[21] The 2014 cohort was introduced to all rubrics in the first two weeks of the summer training institute. The 2013 cohort was informed that they would be used in 2014. Both cohorts received copies of all rubrics.

corps member plans for the year and for every lesson; and (b) whether he/she prepares activities for students to practice what they learn.[22]

The student survey administered during the school year is the same as the one administered during clinical practice. The survey of principals asks school administrators to rate corps members on fourteen competencies. The first nine are the same as those assessed on the classroom observations. The last five are only measured through these surveys: (a) whether a corps member assesses the progress of his or her students; (b) whether he/she analyzes the results of the student assessments he/she administers; (c) whether he/she tries to identify the student behaviors that are associated with student achievement; (d) whether he/she tries to identify his/her own behaviors that are associated with student achievement; (e) whether he/she actively explores ways to improve classroom instruction; (f) whether he/she constantly adjusts his/her practice based on the data he/she collects.[23]

Raters, Assignment, and Frequency

As Table 1 shows, in stages 1 and 2, applicants are evaluated by ExA staff members, who act as raters using structured rubrics. Staff members from *all* areas in the organization (e.g., recruiting, fundraising, communications, etc.) act as raters. Raters differ in their time availability, so they are randomly assigned to applicants conditional on their availability (i.e., raters with more time have more "slots" in the lottery that

[22] The protocol is available at: http://bit.ly/1wnxWHN.

[23] The survey is available at: http://bit.ly/1uTUT67. Principals were asked to assign ratings regardless of whether they had observed corps members in the classroom.

randomly assigns applicants and hence end up reviewing more cases). Each rater measures an applicant's performance only once.

In stage 3, ExA groups corps members into "teaching teams" of two or three and randomly assigns them to a coach and to a group of volunteer students.[24] Coaches are ExA staff members with teacher training and experience. Corps members take turns teaching each day (e.g., corps member A teaches on Monday, corps member B on Tuesday, etc.) When a corps member teaches, his/her coach and peers in the teaching team observe him/her and complete a classroom observation protocol individually. ExA first calculates how many teaching teams it can form based on the number of corps members that it has; then, it randomly assigns corps members, coaches, and students to each of these teaching teams.[25]

The number of classroom observations per corps member depends on whether he or she was assigned to a teaching team of two or three people. In the two weeks of clinical practice, there are seven days in which corps members teach lessons individually.[26] Corps members assigned to teaching teams of two teach at least three and at most four times and have observations scores from their one other peer and coach (i.e., between six and eight observation scores in total). Corps members assigned to teaching teams of three teach at least two and at most three times and have observation scores from their two other peers and coach (i.e., between six and nine scores in total).

---

[24] ExA recruits volunteer students from the schools where it places its corps members.

[25] Each teaching team is devoted to one subject: English, Spanish, or math.

[26] There are 10 potential teaching days, but the first and last days are devoted to diagnostic and final exams, and the second to last day is devoted to a wrap-up taught jointly by all members of a teaching team.

Each group of students completes one survey about each of the corps members in the teaching team to which they were assigned after they have seen that corps member teach for two lessons (i.e., the minimum number of lessons taught by all corps members).

In stage 4, each corps member is assigned to a mentor who conducts all of the classroom observations for that corps member during the school year. Mentors are ExA staff members with teacher training and experience and many individuals who act as coaches are then hired as mentors.[27] For logistical reasons, instead of randomly assigning individual corps members to mentors, ExA first groups corps members by their geographic location and then randomly assigns these groups to mentors. Thus, random assignment is conditional on the geographic location of the corps members (i.e., mentors assigned to a region with more corps members end up observing more lessons). Each mentor is supposed to observe each corps member five times during the year.

Corps members are not randomly assigned to schools or students. ExA assigns corps members to schools by matching their college majors with teaching vacancies (e.g., math majors teach math, biology majors teach science, etc.) and principals assign corps members to classrooms at their own discretion. However, corps members teach multiple schools and classrooms, so ExA selects two schools and two classrooms at random to administer the principal and student surveys, respectively. The same principals and students complete these surveys twice per year.

4. Data

Sample

[27] Mentors do not work for schools. They work for ExA and visit corps members in their classrooms.

We have data on the 822 individuals who applied to ExA in 2013 and the 1,420 who did so in 2014. However, in our analyses, we focus on the 24 individuals who became corps members in 2013 and the 32 who did so in 2014.

As expected, the observable characteristics of applicants who were admitted to ExA differ from those of applicants who were not admitted. Table 2 shows the means and standard deviations of demographic, academic, and professional variables for individuals at each stage of ExA's pipeline by cohort: (a) those whose online applications were reviewed (columns 1 and 5); (b) those who participated in the assessment center (columns 2 and 6); (c) those who taught during clinical practice (columns 3 and 7); and (d) those who taught during the school year (columns 4 and 8).[28]

<Table 2>

Data availability differs by cohort. We observe: (a) the information from the online application and the stage 1 scores of corps members in both cohorts; (b) the stage 2 scores of corps members in both cohorts; (c) the stage 3 scores of corps members in the 2014 cohort; and (d) the stage 4 scores of corps members in both cohorts.[29]

Out of the 56 corps members from both cohorts, we do not observe two corps members during the school year because they taught at after school programs, where none of our instruments could be administered. Of the remaining 54, we have scores from student surveys for 52 corps members, scores from classroom observations for 54

[28] The individuals who *participated* at each stage differ from those who were *invited* to participate at each stage because there is considerable attrition throughout ExA's selection, training, and placement processes. Table A.1 in Appendix A shows the corresponding figures for this second group.

[29] For the 2013 cohort, this was the second year of teaching; for the 2014 cohort, it was the first.

corps members, and scores from principal surveys for 51 corps members. Forty-seven corps members have scores for all three instruments.[30]

Variables

The key variables in our analyses are the average scores of corps members at each stage in Table 1. In stage 1, we focus on $STAGE1$, an individual's score on the online application (the average of three competencies assessed at this stage).[31] Figure A.1 in Appendix A shows the distribution of these scores by cohort and admission status. As expected, the distributions of scores of the 2013 and 2014 cohorts are similar and corps members perform better than non-admitted applicants.[32]

In stage 2, we focus on $STAGE2\_FPT$, an applicant's score on the first part of the assessment center (the average of two scores), $STAGE2\_SPT$, his/her score on the second part (the average of five scores), and $STAGE2$, the average of these two scores. The distributions of these scores are shown in Figure A.2-Figure A.4. These distributions are similar for the 2013 and 2014 cohorts and corps members outperform non-admitted applicants. We also focus on $STAGE2\_DEMO$, a corps member's score on the demonstration lesson. The distribution of these scores is shown in Figure A.5.[33]

---

[30] As Table A.2 indicates, corps members with and without scores from each of these instruments differed in some observable characteristics, but no difference is statistically significant.

[31] To make stage 1 scores comparable across years, we average the scores for the three competencies assessed on both years. Our results do not change if we include all competencies assessed each year.

[32] There are many applicants who had relatively high scores, but were not admitted into ExA. These applicants scored above the cutoffs for two out of the three rubrics administered at this stage, their applications were reviewed by a committee, and the committee decided not to admit them.

[33] Again, some applicants with high scores were not admitted into ExA for the same reasons as in stage 1.

In stage 3, we focus on $STAGE3\_OBS$, an applicant's score on the classroom observations (the average of six scores),[34] $STAGE3\_STU$, his/her score on the student surveys (the average of seven scores), and $STAGE3$, the average of these two scores. The distributions of these scores are in Figure A.6-Figure A.8. Average scores span a very limited range (between one or two points in a five-point scale). This is not atypical based on administrations of similar instruments in other settings.[35]

In stage 4, we focus on $STAGE4\_OBS$, an applicant's score on the classroom observations (the average of nine scores),[36] $STAGE4\_STU$, his/her score on the student surveys (the average of seven scores),[37] $STAGE4\_PRI$, his/her score on the principal surveys (the average of fourteen scores),[38] and $STAGE4$, the average of these three scores.[39] Figure A.9-Figure A.12 show the distributions of these scores by cohort. There are no clear differences in performance across cohorts, but the scores on the principal surveys for both cohorts show some evidence of "ceiling effects" (i.e., a considerable percentage of observations concentrated at the maximum score).

---

[34] We calculate the average score for each corps member using the scores on his/her first two lessons assigned by his/her coach and one random peer.

[35] Araujo et al. (2014) report that 90 percent of teachers in their study were assigned scores of three or four on a one-to-four scale on the CLASS observation protocol.

[36] All corps members have five observations per year, so we average over these five scores.

[37] Surveys are administered in two classrooms twice a year, so we average over classrooms and occasions.

[38] Surveys are administered in two schools twice a year, so we average over both schools and occasions.

[39] We only calculate the stage 4 average score for the 47 corps members with non-missing data for classroom observations, student surveys, and principal surveys so that this score always has the same interpretation. Yet, none of our results change if we include all corps members.

We standardize these scores to interpret our results in terms of corps member-level standard deviations of the average scores at each stage. First we average the scores for each instrument at the corps member level. Then, we standardize the corps member-level scores for each instrument separately. Next, we average the scores for all instruments administered at each stage to obtain a corps member's score for that stage. And finally, we standardize these averaged scores.[40] This final standardization allows us to interpret effects in corps member-level standard deviations of the average scores.[41]

In standardizing the scores for each instrument, we use the mean and standard deviation of corps members from both cohorts. We standardize with respect to corps members (rather than to all applicants) because the number of applicants at each stage differs, but more importantly, because we are interested in understanding whether clinical practice allows us to predict differences in performance among corps members. We standardize with respect to both cohorts (instead of standardizing with respect to each cohort) to preserve meaningful differences in means and variances across cohorts and to ensure that the standardized scores have the same interpretation across cohorts.[42]

5. Empirical Strategy

[40] For example, for stage 4, we first average all classroom observations, student surveys, and principal surveys at the corps member level; then we standardize the scores for each of these three instruments separately; then we calculate a simple average of the three standardized scores; and finally, we standardize this average to obtain the standardized score for stage 4.

[41] If we did not conduct this final standardization, our effect sizes would be expressed in terms of standard deviations of the instruments that we administered.

[42] The only exception is stage 3 scores, which are only available for the 2014 cohort. For these scores, we use the mean and standard deviation for the 2014 cohort of corps members alone.

We want to know whether the performance of teaching trainees during clinical practice adds information that helps us predict their performance during the school year that is not captured by metrics already available. Thus, first we explore whether the scores that corps members receive on ExA's application and selection processes (i.e., stages 1 and 2) predict their scores during the school year (i.e., stage 4). Then, we ask whether their scores from clinical practice (i.e., stage 3) predict their scores from the school year, before and after we account for their scores on stages 1 and 2.

Raw Correlations

We begin by calculating the raw correlations between all variables described in the previous section. If corps members had been randomly assigned individually to raters and students at all stages and measurement error were negligible, these correlations would offer unbiased estimates of the magnitude of the linear relationships across scores in all stages in the population. Yet, random assignment was conducted within blocks and these correlations do not account for these blocks.

Disattenuated Regressions

The performance of corps members in stages 1 through 4 is measured with error, which attenuates correlations across scores in these stages. Therefore, we also use structural equation modeling (SEM) to run disattenuated regressions of stage 4 scores on scores from previous stages that take the reliability of all of these scores into account.[43] We do this as follows. Each applicant's score on each stage is a composite of scores on

[43] We run disattenuated *regressions* instead of *correlations* because the main goal of our paper is predicting corps members' performance during the school year, rather than to estimate the theoretical relationship between the scores assigned in different stages.

two or more competencies or instruments. For example, the stage 1 score is a composite of the scores on the leadership, perseverance, and accomplishment rubrics at that stage. Similarly, the stage 4 score is a composite of the scores on the classroom observations, principal surveys, and student surveys. Thus, we can use the scores from these competencies and instruments to disattenuate the regression of stage 4 on stage 1 scores.

Average Predictive Effects

We account for the way in which corps members were randomly assigned to raters at all stages, and to students during clinical practice, by fitting the following model:

$$Y_i^{POST} = \gamma_k^{POST} + \alpha_j^{PRE} + \lambda_t + \beta X_i^{PRE} + \epsilon_{ijkt} \quad (1)$$

where Y_i^{POST} is the score assigned to corps member i in stage 4, X_i^{PRE} is the score assigned to that corps member during a previous stage (i.e., 1, 2, or 3), γ_k^{POST} are fixed effects for the randomization blocks in stage 4,[44] α_j^{PRE} are fixed effects for the blocks on the previous stage,[45] λ_t are cohort fixed effects,[46] and ϵ_{ijkt} is the error term. We fit this model separately for stages 1, 2, and 3. The coefficient of interest is β, the predictive effect of the score on the prior stage in corps member-level standard deviations.[47]

The fixed effects account for the way in which corps members were randomly assigned to raters in stages 1 through 4, the way in which they were randomly assigned to students in stage 3, and differences in predictive effects across cohorts. Yet, they do not

[44] These are geographical regions.

[45] For stages 1 and 2, these are rater fixed effects; for stage 3, these are teaching team fixed effects. Our main results do not change if we use subjects (rather than teaching teams) as our fixed effects.

[46] We cannot include cohort fixed effects when we estimate the average predictive effect of stage 3 scores because we only observe these scores for the 2013 cohort.

[47] We estimate Huber-White robust standard errors in this and all other models presented in this section.

account for the way in which corps members were assigned to schools and classrooms in stage 4. If corps members who performed well on stages 1 through 3 were assigned to the most challenging schools or students within those schools, our estimate of β will be biased downwards. Conversely, if the best-performing corps members were assigned to the best schools and/or classrooms, β will be biased upwards.[48]

Heterogeneity by Performance Quantile

In model (1), we assume that the relationship between corps members' scores from stages 1, 2, or 3 and their scores from stage 4 is linear. Yet, it is possible that the scores from stages 1 through 3 are better at predicting stage 4 scores for individuals who performed particularly well or poorly in these previous stages. To test for non-linear patterns in predictive effects, we fit the following model:

$$Y_i^{POST} = \gamma_k^{POST} + \alpha_j^{PRE} + \lambda_t + \beta X_i^{PRE} + \phi (X_i^{PRE})^2 + \epsilon_{ijkt} \quad (2)$$

where $(X_i^{PRE})^2$ is the square of the score assigned to each corps member in a previous stage and everything else is defined as above.[49] We fit this model separately for stages 1, 2, and 3. The coefficient of interest is ϕ. If it is statistically significant, it indicates that there is a non-linear relationship between the scores from stages 1, 2, or 3 and stage 4.

[48] Note, however, that for our estimates to be biased, both the ExA staff and school principals should be able to accurately predict corps members' performance on stage 4. It is not enough that corps members are not randomly assigned to schools and/or classrooms. They have to be assigned to schools and classrooms based on factors that correlate with their performance on stage 4.

[49] We experimented with different functional forms that allow for a decreasing positive slope. Our main results do not change if we use a different functional form with this property.

Heterogeneity by Timing of Performance Information

In models (1) and (2), the scores from stage 4 are averages over multiple instances of measurement. As we have already discussed, each corps member has scores from five classroom observations conducted throughout the school year, two student surveys conducted with the same two classrooms twice during the year, and two surveys completed by the same two principals twice during that year. It is possible that the scores from stages 1, 2, or 3 are better at predicting scores from stage 4 that were assigned earlier or later in the year.[50] To address this possibility, we fit the following model:

$$Y_{in}^{POST} = \gamma_k^{POST} + \alpha_j^{PRE} + \lambda_t + \beta_1 X_i^{PRE} + \beta_2 S_i^{POST} + \beta_3 (X_i^{PRE} * S_i^{POST}) + \epsilon_{ijktn} \quad (3)$$

In this model, each corps member has two measures of stage 4 performance: one from the first six months of the school year and one from the last six months. Y_{in}^{POST} is the score of corps member i during stage 4 at time n,[51] S_i^{POST} is a dummy indicating whether the stage 4 score was assigned in the last six months of the year, and $X_i^{PRE} * S_i^{POST}$ is the interaction between the score on a prior stage and the dummy. β_3 indicates whether the scores from stages 1, 2, or 3 are more (or less) predictive of the scores from stage 4 if the latter were assigned in the last six months of the school year.

[50] For example, principals might not be able to observe corps members teaching by the first time that they have to complete the survey. Consequently, their ratings from the beginning of the year might offer a less accurate picture of corps members' skills than those from the end of the year. This would make it easier for scores on stages 1 through 3 to predict stage 4 scores assigned at the end of the year.

[51] The raters (observers, students, principals) remain the same throughout the school year, which is why we do not include the n subscript in the fixed effects for randomization blocks.

Heterogeneity by Cohort

In models (1) through (3), we estimate the predictive effects of the scores from stages 1 and 2 across cohorts. We also test whether the predictive effects for these two stages are different across cohorts by fitting the following model:[52]

$$Y_i^{POST} = \gamma_k^{POST} + \alpha_j^{PRE} + \lambda_t + \beta_1 X_i^{PRE} + \beta_2 (X_i^{PRE} * \lambda_t) + \epsilon_{ijkt} \quad (4)$$

where λ_t is a dummy indicating whether a corps member belongs to the 2014 cohort and $X_i^{PRE} * \lambda_t$ is the interaction between the score on a previous stage and that dummy. β_2 indicates whether the score from stage 1 or 2 is more (or less) predictive of the score from stage 4 for corps members in the 2014 cohort.

6. Results

Raw Correlations

Table A.3 shows the pairwise correlations across all the metrics described in Section 4.[53] As the table shows, the correlations across scores assigned in different stages are typically low (correlation coefficients are always below .4) and they are not statistically significant. This holds true even for scores on the same instruments—or variations of the same instrument—assigned in different stages (e.g., classroom observations and student surveys in stages 3 and 4). The correlations across scores assigned in the same stage are also low. There is, however, an important exception: the scores on classroom observations from stage 4 are moderately correlated with the scores on student and principal surveys from that stage (the coefficients are .406 and .654, respectively). These correlations are statistically significant at the 1% level.

[52] Recall that we do not observe stage 3 scores for the 2013 cohort.

[53] We also calculated the rank correlations across all metrics and obtained very similar results.

Disattenuated Regressions

Figure A.13-Figure A.15 show the results from the disattenuated regressions of stage 4 scores on scores from previous stages. As Figure A.13 indicates, once we account for measurement error, there is a *negative* relationship between stage 1 and stage 4 scores. For each unit in the stage 1 score, an individual performs .390 standard deviations worse on stage 4.[54] Figure A.14 shows that there is a positive relationship between the scores from stages 2 and 4. For each unit in the stage 2 score, a corps member scores .325 standard deviations better on stage 4. There is also a positive relationship between stage 3 and stage 4 scores. For each unit in the stage 3 score, an individual performs .826 standard deviations better on the stage 4 score. None of the relationships across latent factors, however, are statistically significant.

Predictive Validity of the Online Application Scores

Table 3 shows the results from our estimation of the average predictive effects of stage 1 scores using equation (1).[55] We find that stage 1 scores have *negative* predictive effects, which is consistent with Figure A.13. In column (4), once we introduce cohort fixed effects and rater fixed effects for both stages, for every standard deviation in the stage 1 scores, a corps member performs on average .421 standard deviations worse in stage 4. This result is statistically significant at the 5% level, and it maintains its

[54] The scores for all stages (1 through 4) are latent (unobserved) factors in this framework.

[55] Table A.4 includes the results for the 2014 cohort. As a benchmark, Table A.5 also includes the predictive effects of descriptive characteristics of both cohorts. We fit the same regressions as in Table 3 with additive metrics of performance on stages 1 and 4 (instead of averaging over the scores for different competencies and instruments) and obtain very similar results.

magnitude and statistical significance after we introduce controls in column (5). In fact, based on the 95% confidence intervals, we can discard *any* positive predictive effects.[56]

<Table 3>

It is not entirely clear why stage 1 scores are negatively related to stage 4 scores. Perhaps the least likely explanation is measurement error. In theory, it is possible that the individuals who performed best on stage 1 and became corps members did so because of positive transitory variance (stemming, for example, from being well-rested on the day that they wrote the online application, or from having worked or studied at a place that is instantly recognizable to the application reviewers). If this was the case, when corps members were assessed during stage 4, they reverted back to the mean.[57] Yet, the rubrics for stage 1 raters are highly prescriptive and leave far less room for discretion—and thus, for transitory variance—than the rubrics for the other stages. And, as Figure A.13 shows, once we account for error, the relationship between stage 1 and stage 4 scores is negative (i.e., error is attenuating this negative relationship).

There are two more plausible explanations that are not mutually exclusive. One is that there may be a tradeoff between some competencies that make a good leader (which feature more prominently in stage 1 rubrics) and those that make a good teacher (which

[56] There is also a negative relationship between a corps member's score on stage 1 and his or her score on the student and principal surveys administered in stage 4, but it is marginally statistically significant.

[57] There is a large literature that documents the effects of transitory variance on rankings (Barrera-Osorio & Ganimian, 2015; Chay, McEwan, & Urquiola, 2003; Kane & Staiger, 2001, 2002). If the high performance of corps members were due entirely to positive transitory variance, the correlation between changes in the same scores would be -0.5, which is similar to the coefficients in Table 3. However, in that table we are estimating relationships between levels in different scores.

are given more attention in later stages).[58] We consider this possibility by fitting model (1) separately for each competency assessed in stage 1. Table A.6 shows the predictive effects of accomplishment (Panel A), leadership (Panel B), and perseverance (Panel C). As the table shows, the scores on all competencies assessed in stage 1 are negatively related to stage 4 scores. In fact, in the case of leadership, the negative relationship is statistically significant at the 10% before and after we add the controls (columns 4-5).

Another potential explanation is that the best performers in stage 1 were assigned to the most challenging schools and/or classrooms.[59] We do not have data on schools or students to assess this hypothesis directly. However, we graph the slopes from regressions that only include the rater fixed effects from stage 1 or the geographic region fixed effects from stage 4. Figure A.16 confirms that the relationship between stage 1 and stage 4 scores is negative when we only include stage 1 rater fixed effects.[60] However, Figure A.17 shows this relationship is (slightly) positive when we only include stage 4 rater fixed effects. This is consistent with a "compensatory" non-random assignment of corps members to schools and classrooms.

---

[58] For example, to be a good (first year) teacher, an individual may need to stick to the classroom practices that he or she knows best, while a good leader may feel compelled to experiment with new strategies, even if this means that he or she will often struggle.

[59] This is more likely than it may seem. The competencies assessed in stage 1 are those that are observable from an online application. Thus, they lend themselves most easily to non-random assignment of corps members to schools and classrooms.

[60] Note that some of our rater fixed effects include one corps member. We re-ran our regressions and graphs including corps members who were evaluated by a rater who also evaluated another corps member. Our results are similar and are available upon request.

Our results indicate that there is a negative relationship between stage 1 and stage 4 scores among corps members. They do not say anything, however, about the relationship between these two scores among applicants. As Figure A.1 shows, corps members performed better and varied less in their stage 1 scores than non-admits. It is possible that individuals who were disqualified based on their (low) stage 1 scores would have been ineffective teachers if we had observed them during stage 4.

We consider whether the relationship between stage 1 and stage 4 scores is non-linear by plotting this relationship with kernel-weighted local polynomial smoothing. Figure A.18 suggests that we should use a functional form that allows for a decreasing positive slope, so we fit model (2). Yet, Table A.7 shows the quadratic term is only statistically significant when we try to predict corps members' scores on the student surveys from stage 4.

We do not find evidence that the predictive validity of stage 1 scores varies by timing of effectiveness information or cohort using equations (3) and (4) (see Table A.8-Table A.9). The standard errors are too large to estimate these effects precisely.

Predictive Validity of the Assessment Center Scores

Table 4 shows the average predictive effects of stage 2 scores using equation (1).[61] Panel A uses the average score in the first and second parts of the assessment center. Panel B uses the scores from the first part and Panel C from the second part.[62]

<Table 4>

[61] We fit the same regressions as in Table 4 with additive metrics of performance on stages 2 and 4 (instead of averaging over the scores for different competencies and instruments) and obtain very similar results.

[62] Table A.10 includes the results for the 2014 cohort.

As in Figure A.14, the relationship between stage 2 and stage 4 scores is positive. However, it is estimated imprecisely, and we cannot reject the existence of moderate to large predictive effects. Panel B indicates that the relationship between the scores from the first part of the assessment center and those from stage 4 is estimated consistently around zero. In column (4), we can discard predictive effects larger than .417 standard deviations. Panel C indicates that the relationship between the scores on the second part of the assessment center and those from stage 4 is also around zero.[63] In column (4), we can discard predictive effects larger than .327 standard deviations. In all panels, the results for individual instruments (columns 6-8) are imprecisely estimated.[64]

We consider whether the relationship between stage 2 and stage 4 scores is non-linear by plotting this relationship. Then, we fit model (2). Figure A.19 offers no evidence of a non-linear trend and Table A.11 confirms that none exists.[65]

We fit models (3) and (4) and find little evidence that the predictive effects of stage 2 scores vary by timing of effectiveness information or cohort. As Table A.12-Table A.13 indicate, the standard errors are too large to estimate these effects precisely.[66]

Again, our results say nothing about whether stage 2 scores would predict differences among the entire applicant pool, for the same reasons as above. As Figure

---

[63] In all panels, the controls in column (5) change the magnitude of the coefficient considerably, offering evidence of non-random assignment of corps members to schools and classrooms in stage 4.

[64] We also explore the relationship between stage 4 scores and the stage 2 scores on each competency assessed in the first and second part of the assessment center in Table A.14 and Table A.15, respectively.

[65] Only the quadratic term of the regression predicting classroom observations is statistically significant.

[66] The interaction between the stage 2 score and the cohort dummy is only statistically significant when predicting the scores that corps members received on the student surveys at stage 4.

A.2 shows, corps members performed better on stage 2 than non-admits. Thus, it is possible that individuals who were disqualified due to their (low) stage 2 scores would have been ineffective teachers if we had observed them during stage 4.

The first and second parts of the assessment center evaluate many competencies. We fit model (1) using the stage 2 scores from each competency to examine whether any of them are predictive of stage 4 performance. As Table A.14 and Table A.15 show, none of the scores on these competencies have a positive and statistically relationship with stage 4 scores once we account for the randomization blocks in stages 2 and 4.

The competencies evaluated during the assessment center cut across different activities, including a critical thinking assessment, a written exercise, and a demonstration lesson. We also fit model (1) using the stage 2 scores from each of these activities.[67] As Table A.16 shows, the scores on these three activities do not predict the average scores from stage 4.

Notably, however, the scores that corps members received on their demonstration lessons during the assessment center predict their performance on the classroom observations during the school year. On average, for every standard deviation in the demonstration lesson, a corps member performs .348 standard deviations better on the classroom observations in stage 4. This relationship is statistically significant at the 1% level. This is particularly interesting, given that demonstration lessons are much shorter than clinical practice lessons, they occur only one time, they are assessed by only one rater, and they do not require the presence of students in the classroom.

[67] We do not observe the scores for the one-on-one interview, so we do not include it here.

Table A.17 examines the relationship between stage 4 scores and each competency assessed during the demonstration lessons at stage 2: planning (Panel A), organization (Panel B), student engagement (Panel C), listening and answering (Panel D), and communication (Panel E). As Panel A shows, there is a positive and statistically significant relationship between the score on planning and stage 4 scores. As column (4) indicates, for each standard deviation unit on the planning score, a corps member performs on average .285 standard deviations *better* on stage 4. This relationship is statistically significant at the 5% level, and it becomes larger and maintains its statistical significance when we add controls. In fact, the scores on this competency also predict corps members' scores on the student surveys and classroom observations conducted during stage 4 (the coefficients are .304 and .315, and they are statistically significant at the 10 and 1% levels, respectively). No other competency shows a consistently positive relationship with stage 4 performance.

Predictive Validity of the Clinical Practice Scores

Table 5 displays the average predictive effects of stage 3 scores using equation (1).[68] Panel A uses the average score in the student surveys and classroom observations as the main predictor, Panel B uses the scores in the student surveys, and Panel C uses the scores in the classroom observations.

<Table 5>

In Panel A, the relationship between stage 3 and stage 4 scores is positive, consistent with what we found in Figure A.20, but it is imprecisely estimated. The

[68] We fit the same regressions as in Table 5 with additive metrics of performance on stages 3 and 4 (instead of averaging over the scores for different instruments) and obtain very similar results.

predictive effect of stage 3 scores is large (.611 standard deviations), but only statistically significant (at the 5% level) after we introduce the controls in column (4). As column (6) indicates, however, there is a positive and statistically significant relationship between stage 3 and stage 4 scores on the classroom observations. The predictive effect is large (.72 standard deviations) and statistically significant at the 1% level. In Panels B and C, the relationships between student surveys and classroom observations at stage 3 and the stage 4 scores are imprecisely estimated.

We consider whether the relationship between stage 3 and stage 4 scores is non-linear by plotting this relationship. Figure A.20 shows clear evidence of a decreasing positive slope, so we fit model (2) with a quadratic term.[69] As Table A.18 shows, this term is statistically significant, confirming the non-linear pattern. As Table A.19 indicates, we cannot precisely determine whether the predictive effects of stage 3 scores vary by the timing of effectiveness information.

Finally, we consider whether stage 3 scores add information not captured in stages 1 and 2 that can help predict which corps members will perform better on stage 4. Table 6 shows the results from our estimation.

<Table 6>

We find some evidence that the scores from clinical practice lessons add value. The relationship between stage 3 and stage 4 scores is positive but imprecisely estimated once we account for stage 1 and stage 2 scores. Once again, the predictive effect of stage

[69] This figure also raises the question of whether the predictive effects that we observe are due to the outlier who performed more than two standard deviations below the mean during stage 3. We re-run all regressions excluding this observation and found similar coefficients and levels of statistical significance, indicating that our results are not driven by this observation. Results are available upon request.

3 scores is large (.75 standard deviations), but only statistically significant (at the 5% level) after we introduce the controls in column (4). Stage 3 scores are also predictive of classroom observations in stage 4: for every standard deviation unit in stage 3 scores, a corps member performs .72 standard deviations better on these observations, and this coefficient is statistically significant at the 1% level. The magnitude of this coefficient is similar to the one from Table 5 that does not account for stage 1 and stage 2 scores.

In Table 6, the information on corps members' performance before clinical practice is aggregated at the stage level. This need not be the case. Some competencies may be more helpful in making distinctions between corps members than others, but because they are lumped together with other competencies assessed on the same stage, they have limited influence over the average score for that stage.

To address this possibility, we conduct a principal component analysis of all of the competencies assessed in stages 1 and 2. The resulting eigenvalues are shown in Table A.20. We follow standard convention by keeping the principal components with eigenvalues greater than 1 (i.e., those with variances that are larger than the variance of any indicator on its own). Based on the competencies that are given more importance on each of the principal components, they appear to measure: (a) corps members' ability to communicate with others (component 1); (b) their drive and determination (component 2); (c) their level of discipline and achievement (component 3); and (d) their ability to

work with others (component 4).[70] Together, these components explain 63% of the variance in the scores for all competencies assessed in stages 1 and 2.

We use the four principal components as controls in the regression of stage 4 scores on stage 3 scores.[71] As Table 7 shows, the results are similar to those in Table 6. The predictive effect of stage 3 scores is large (.732 standard deviations), but only statistically significant (at the 5% level) after we introduce the controls in column (4). Stage 3 scores are also predictive of classroom observations in stage 4: for every standard deviation unit in stage 3 scores, a corps member performs .743 standard deviations better on these observations, and this coefficient is statistically significant at the 1% level.

<Table 7>

In Table 7, we maximize the variation in corps members' performance before clinical practice. Yet, another way to assess the added value of clinical practice is to give more importance to competencies assessed in stages 1 and 2 that are more predictive of stage 4 performance, and less importance to those that are less predictive. We do so by running an unconditional regression of stage 4 scores on the full set of competencies assessed in stages 1 and 2 and creating a variable that uses the (normalized) coefficients in this regression as weights in a weighted average of all these competencies.[72]

We use the new composite as a control in regressions of stage 4 scores on stage 3 scores and find a similar pattern. As Table 8 shows, the predictive effect of stage 3

[70] It is common practice to assign names to the principal components. The names that we assigned are based on our interpretation of the relative importance given to each competency and they are only meant to help us interpret our results. Other interpretations are certainly possible.

[71] We also tried adding them one by one as controls and found very similar results.

[72] We reverse-coded the scores of the competencies that were negatively related to stage 4 scores.

scores is still large (.612 standard deviations), but only statistically significant (at the 5% level) after we introduce the controls in column (4). Stage 3 scores are also predictive of classroom observations in stage 4: for every standard deviation unit in stage 3 scores, a corps member performs .687 standard deviations better on these observations, and this coefficient is statistically significant at the 1% level.

<Table 8>

Finally, we also run the same model including the score from the demonstration lesson as the only control. As Table 9 indicates, when we include this control, we do not see a statistically significant relationship between stage 3 scores and stage 4 scores, even when we include covariates in the regression (column 4). The coefficient on the demonstration lesson score is large (.819 standard deviations) and statistically significant at the 5% level. However, stage 3 scores continue to predict stage 4 scores on classroom observations (column 6): for every standard deviation in stage 3 scores, a corps member performs .69 standard deviations better on these observations, and this coefficient is statistically significant at the 1% level.

<Table 9>

7. Discussion

In this paper, we explore whether the performance of admits to an alternative pathway into teaching during clinical practice predicts their performance during the school year. Performance during clinical practice is measured by student surveys and classroom observations, and performance during the school year is measured by student surveys, classroom observations, and principal surveys. We explore whether the

performance of individuals during clinical practice adds information not previously captured by their performance on the application and selection processes of the program.

We start by examining whether the application and selection metrics, by themselves, predict performance during the school year. We find that the application metrics are *negatively* associated with school year performance among program admits. This is an interesting puzzle and we consider three potential explanations. First, we argue that the negative relationship is not caused due to measurement error, since our disattenuated regressions show that, when we account for error during the application process and the school year, this negative relationship remains. Second, we consider the possibility that there may exist a tradeoff between teaching and leadership skills, which feature more prominently on the application rubrics. Third, we contend that the application variables may be used to assign the "best" corps members to the most challenging schools and classrooms. Admittedly, however, we cannot either discard or confirm any of these possibilities.

By contrast, we find that that the selection metrics are positively associated with school year performance, and this is true even when we account for measurement error. However, we lack sufficient statistical power to estimate this relationship precisely. A brief demonstration lesson delivered during the selection process, however, has a positive and statistically significant relationship to performance during the school year. This finding is intuitive, yet surprising. It is intuitive because it lends empirical support to the belief that observing an individual teaching before he or she begins working in schools should yield some useful information about that individual's instructional skills. Yet, it is surprising because these demonstration lessons look very little like an actual lesson.

Finally, we explore whether the performance of individuals during clinical practice predicts their performance during the school year. We find that how individuals perform during clinical practice is positively related to how they perform during the school year, although this relationship is only statistically significant when we introduce controls. We also find that an individual's performance during clinical practice has a positive and statistically significant relationship to his or her performance on classroom observations conducted during the school year. This predictive effect is large (about twice the size of the predictive effect of demonstration lessons) and it is robust to a number of ways of accounting for prior performance.

These are encouraging findings and they suggest that observing someone teaching with multiple occasions, instruments, and raters can add valuable information about their instructional skills. Yet, it is important to put them into perspective.

There are at least four reasons why clinical practice lessons should predict school year performance, even when we account for the performance of individuals on the application and selection processes. First, we administered similar instruments during clinical practice and the school year. Thus, there could be a positive and statistically significant relationship between clinical practice and the school year due to sheer "method effects." Second, unlike the application and selection metrics, the clinical practice metrics were only administered among admits to the alternative pathway. Therefore, it would be reasonable to expect there to be more variation in these metrics than in the application and selection metrics, which would increase our chances of detecting a statistically significant positive relationship if one exists.

Third, unlike the application and selection instruments, the ones administered during clinical practice were low-stakes. Thus, there were no incentives for individuals to compete to obtain the best score (and hence, to reduce the variance) on these metrics. This would also improve our chances of detecting a statistically significant relationship. Finally, more time elapses between the application and selection processes and the school year than between clinical practice and the school year. If the teaching skills of individuals are evolving, more proximal measures ought to be more predictive.

Note that these four factors would lead the performance of individuals during clinical practice to improve our predictions of school year performance *even if* the latent skills measured during the application, selection, and training processes were the same (which is not the case). In light of these factors, the relationship between information collected during clinical practice and the school year is less clear than we would expect. This raises the question of whether the cost (in time, resources, and effort) of collecting information during clinical practice is worth the information that it adds.

Our small sample size limits our ability to differentiate between metrics that are unrelated to school year performance and metrics with small predictive effects. We are currently in the process of incorporating information from the 2015 school year. We hope this new information will improve our ability to make these distinctions.

Our findings so far suggest that alternative pathways that already collect information during clinical practice using similar instruments can capitalize on this information to improve their predictions about the performance of their teaching trainees. These predictions could prove useful to make relatively low-stakes decisions about

teaching trainees, such as whether they should be assigned to less challenging schools and classrooms, or whether they should receive intensive induction or mentoring.

This is admittedly a very specific policy implication, but an important one nonetheless. Currently, new teachers are often assigned to challenging schools and classrooms, where they struggle considerably (Boyd, Lankford, Loeb, Ronfeldt, & Wyckoff, 2011; Grissom, Loeb, & Nakashima, 2013; Hanushek & Rivkin, 2010a; Loeb, Kalogrides, & Béteille, 2012; Ronfeldt, Lankford, Loeb, & Wyckoff, 2013). Mentors or coaches are either assigned unsystematically, or they are assigned to all new teachers, neither of which is efficient (Glazerman et al., 2010; Glazerman & Seifullah, 2010).

Further research should explore whether information collected during clinical practice could and should be used for high-stakes decisions. If alternative pathways started using performance on clinical practice lessons as a final screening mechanism, individuals are likely to adjust their behavior accordingly (e.g., by working harder during clinical practice, or focusing on the practices assessed by their observers and students). These changes in behavior could reduce the variation in clinical practice performance, making it harder to predict differences in school year performance. It could also distort the relationship between performance in clinical practice and the school year if the least skilled applicants make the greatest effort to game the system. We believe that this is an important question, but one that ought to be addressed separately.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25*, 95-135.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2014). A Helping Hand? Teacher Quality and Learning Outcomes in Kindergarten. Washington, DC: Inter-American Development Bank.

Atteberry, A., Loeb, S., & Wyckoff, J. (2013). Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness *NBER Working Paper No. 19096*. Cambridge, MA: National Bureau of Economic Research (NBER).

Barrera-Osorio, F., & Ganimian, A. J. (2015). The Implications of Volatility in School Test Scores for Accountability Policies in Pakistan. Cambridge, MA: Harvard Graduate School of Education.

Bastian, K. C. (2013). Do Teachers' Non-Cognitive Skills and Traits Predict Effectiveness and Evaluation Ratings? Chapel Hill, NC: University of North Carolina at Chapel Hill.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). The Narrowing Gap in New York City Teacher Qualifications and its Implications for Student Achievement in High‐Poverty Schools. *Journal of Policy Analysis and Management, 27*, 793-818.

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2011). The Role of Teacher Quality in Retention and Hiring: Using Applications to Transfer to Uncover Preferences of Teachers and Schools. *Journal of Policy Analysis and Management, 30*, 88-110.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2003). The Central Role of Noise in Evaluating Interventions that Use Test Scores to Rank Schools. *American Economic Review, 95*, 1237-1258.

Chetty, R., Friedman, J., & Rockoff, J. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review, 104*, 2633-2679.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR. *Quarterly Journal of Economics, 126*, 1593-1660.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). *How and Why do Teacher Credentials Matter for Student Achievement?* NBER Working Paper No. 12828. National Bureau of Economic Research (NBER). Cambridge, MA.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does Teacher Certification Matter? Evaluating the Evidence. *Educational Evaluation and Policy Analysis, 23*, 57-77. doi: 10.2307/3594159

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005). Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America, and Teacher Effectiveness. *education policy analysis archives, 13*, n42.

Dobbie, W. (2011). Teacher Characteristics and Student Achievement: Evidence from Teach For America. Cambridge, MA: Harvard University.

Duckworth, A. L., Quinn, P. D., & Seligman, M. E. (2009). Positive Predictors of Teacher Effectiveness. *The Journal of Positive Psychology, 4*, 540-547.

Gitomer, D. H., Phelps, G., Weren, B. H., Howell, H., & Croft, A. J. (2014). Evidence on the Validity of Content Knowledge for Teaching Assessments. In T. J. Kane, K. A. Kerr & R. C. Pianta (Eds.), *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*. San Francisco, CA: Jossey-Bass.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010). Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative Routes to Teaching: The Impacts of Teach for America on Student Achievement and Other Outcomes. *Journal of Policy Analysis and Management, 25*, 75-96.

Glazerman, S., Protik, A., Teh, B.-r., Bruch, J., & Max, J. (2013). Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment: Mathematica Policy Research.

Glazerman, S., & Seifullah, A. (2010). An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year Two Impact Report. Washington, DC: Mathematica Policy Research.

Goldhaber, D. D., & Brewer, D. J. (2000). Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement. *Educational Evaluation and Policy Analysis, 22*, 129-145.

Grissom, J. A., Loeb, S., & Nakashima, N. (2013). Strategic Involuntary Teacher Transfers and Teacher Performance: Examining Equity and Efficiency *NBER Working Paper No. 19108*. Cambridge, MA: National Bureau of Economic Research (NBER).

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2013). Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores. *American Journal of Education, 119*, 445-470.

Hanushek, E., & Rivkin, S. G. (2010a). Constrained Job Matching: Does Teacher Job Search Harm Disadvantaged Urban Schools? *NBER Working Paper No. 15816*. Cambridge, MA: National Bureau of Economic Research (NBER).

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). The Market for Teacher Quality *NBER Working Paper 11154*. Cambridge, MA: National Bureau of Economic Research (NBER).

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher Quality. *Handbook of the Economics of Education, 2*, 1051-1078.

---. (2010b). Constrained Job Matching: Does Teacher Job Search Harm Disadvantaged Urban Schools? *NBER Working Paper No. 15816*. Cambridge, MA: National Bureau of Economic Research (NBER).

Harris, D. N., & Sass, T. R. (2011). Teacher Training, Teacher Quality and Student Achievement. *Journal of Public Economics, 95*, 798-812.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A Validity Argument Approach to Evaluating Teacher Value-Added Scores. *American Educational Research Journal, 48*, 794-831.

Jacob, B. A., & Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics, 26*, 101-136.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City. *Economics of Education Review, 27*, 615-631.

Kane, T. J., & Staiger, D. O. (2001). Improving School Accountability Measures *NBER Working Paper No. 8156*. Cambridge, MA: National Bureau of Economic Research (NBER).

---. (2002). The Promise and Pitfalls of Using Imprecise School Accountability Measures. *The Journal of Economic Perspectives, 16*, 91-114.

Kane, T. J., & Staiger, D. O. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation *NBER Working Paper No. 14607*. Cambridge, MA: National Bureau of Economic Research (NBER).

Kane, T. J., & Staiger, D. O. (2011). Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teachers: Combining High-Quality Observations with Student Surveys and Achievement Gains. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources, 46*, 587-613.

Loeb, S., Kalogrides, D., & Béteille, T. (2012). Effective Schools: Teacher Hiring, Assignment, Development, and Retention. *Education Finance and Policy, 7*, 269-304.

Metzler, J., & Woessmann, L. (2012). The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation. *Journal of Development Economics, 99*, 486-496.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis, 26*, 237-257.

Papay, J. P., & Kraft, M. A. (2011). Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Wvidence on Long-Term Career Growth. Providence, RI: Brown University.

Papay, J. P., West, M. R., Fullerton, J. B., & Kane, T. J. (2012). Does an Urban Teacher Residency Increase Student Achievement? Early Evidence From Boston. *Educational Evaluation and Policy Analysis, 34*, 413-434.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica, 73*, 417-458.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review*, 247-252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy, 6*, 43-74.

Rockoff, J. E., & Speroni, C. (2011). Subjective and Objective Evaluations of Teacher Effectiveness: Evidence from New York City. *Labour economics, 18*, 687-696.

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools. *American Economic Review, 102*, 3184-3213.

Ronfeldt, M., Lankford, H., Loeb, S., & Wyckoff, J. (2013). How Teacher Turnover Harms Student Achievement. *American Educational Research Journal, 50*, 4-36.

Silva, T., McKie, A., Knechtel, V., Gleason, P., & Makowsky, L. (2014). Teaching Residency Programs: A Multisite Look at a New Model to Prepare Teachers for High-Need Schools. Princeton, NJ: Mathematica Policy Research.

Taylor, E. S., & Tyler, J. H. (2012). The Effect of Evaluation on Teacher Performance. *American Economic Review, 102*, 3628-3651.

Table 1: ExA's Corps Member Pipeline

	Stage 1: Application review (February to September)	**Stage 2:** Assessment center (October to November)	**Stage 3:** Summer training institute (January)	**Stage 4:** School year (February to December)
Activity	ExA invites online applications to the program and reviews these applications.	ExA invites individuals with the best applications to participate in: • Group discussion • One-on-one interview • Demonstration lesson • Written exercise • Critical thinking assessment	ExA offers individuals who performed best in the assessment a spot in the program and requires that they participate in: • Workshops • Clinical practice	ExA places corps members in schools, where they begin teaching.
Instrument	• Selection rubric	• Selection rubrics	• Clasroom observation protocol • Student surveys	• Clasroom observation protocol • Student surveys • Principal surveys
Raters	• ExA's staff members	• ExA's staff members	• Classroom observations: – Coaches – Teaching team • Surveys: – Students	• Classroom observations: – Mentors • Surveys: – Students – Principals
Assignment	• Applicants are randomly assigned to raters.	• Applicants are randomly assigned to raters.	• Corps members are randomly assigned to coaches and their teaching team. • Corps members are randomly assigned to students.	• Corps members are randomly assigned to mentors. • Corps members are not randomly assigned to principals or students.
Frequency	• Once per applicant.	• Once per applicant.	• Corps members in teaching teams of two are observed teaching 3-4 times. • Corps members in teaching teams of three are observed teaching 2-3 times. • Students complete one survey per corps member.	• Corps members are observed teaching 5 times. • Two groups of students complete two surveys per corps member. • Two principals complete two surveys per corps member.
Sample	• 2013 and 2014 cohorts.	• 2013 and 2014 cohorts.	• 2014 cohort only.	• 2013 and 2014 cohorts.

Table 2: Individuals Participating at Each Stage of ExA's Pipeline, 2013-2014

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	2013				2014			
	Stage 1	Stage 2	Stage 3	Stage 4	Stage 1	Stage 2	Stage 3	Stage 4
Female	.71	.63***	-	.71	.73	.71	.63	.63*
	(.45)	(.49)		(.46)	(.45)	(.46)	(.49)	(.49)
Prov. of B.A.	.44	.36**	-	.29	.42	.36**	.47	.47
	(.5)	(.48)		(.46)	(.49)	(.48)	(.51)	(.51)
City of B.A.	.53	.62***	-	.71*	.53	.62***	.5	.5
	(.5)	(.49)		(.46)	(.5)	(.49)	(.51)	(.51)
GPA (out of 10)	7.4	7.55**	-	7.61	7.51	7.73***	7.47	7.47
	(.91)	(.89)		(.82)	(.94)	(.87)	(.88)	(.88)
STEM major	.13	.13	-	.13	.09	.08**	.22**	.22**
	(.33)	(.34)		(.34)	(.29)	(.27)	(.42)	(.42)
Education major	.05	.03	-	0	.03	.03	0	0
	(.21)	(.17)		(0)	(.17)	(.18)	(0)	(0)
Graduated	.82	.85	-	.71	.86	.87*	.75*	.75*
	(.38)	(.36)		(.46)	(.35)	(.33)	(.44)	(.44)
Applied before	.04	.05	-	.04	.08	.1	.13	.13
	(.19)	(.23)		(.2)	(.27)	(.3)	(.34)	(.34)
N	821	168		24	1412	211	32	32

Notes: Prov. of B.A. refers to the Province of Buenos Aires and City of B.A. refers to the City of Buenos Aires. This table displays the means and standard deviations (between parentheses) of a set of variables collected during ExA's online application for individuals who *participated* at each stage of ExA's selection, training, and placement pipeline (i.e., who have a non-missing score for that stage). Stars indicate the levels of statistical significance for t-tests comparing the means of individuals who participated and who did not participate at each stage: * p<.10, ** p<.05, and *** p<.01. There are no figures in column (3) because we do not observe which corps members in the 2013 cohort participated in the summer training institute.

Table 3: Relationship between the Scores from the Online Application and School Year, 2013-2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4	(6) Stage 4 (student surveys)	(7) Stage 4 (classroom observations)	(8) Stage 4 (principal surveys)
Stage 1	-0.001	-0.567*	-0.431**	-0.421**	-0.461**	-0.423*	-0.165	-0.284*
	(0.194)	(0.295)	(0.195)	(0.189)	(0.217)	(0.211)	(0.204)	(0.143)
Stage 1 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
	47	47	47	47	47	52	54	51

Notes: This table displays the coefficients from regressions of stage 4 scores on stage 1 scores and their associated robust standard errors (between parentheses). Stars indicate the levels of statistical significance of each coefficient: * p<.10, ** p<.05, and *** p<.01.

Table 4: Relationship between the Scores from the Assessment Center and School Year, 2013-2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4	(6) Stage 4 (student surveys)	(7) Stage 4 (classroom observations)	(8) Stage 4 (principal surveys)
Panel A.								
Stage 2	0.060	0.219	0.176	0.201	0.419	0.123	0.293	-0.050
	(0.109)	(0.258)	(0.286)	(0.342)	(0.340)	(0.258)	(0.276)	(0.245)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51
Panel B.								
Stage 2 (first part)	0.127	0.078	0.041	0.078	0.207	-0.028	0.056	-0.082
	(0.118)	(0.150)	(0.133)	(0.167)	(0.171)	(0.184)	(0.136)	(0.184)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51
Panel C.								
Stage 2 (second part)	-0.082	0.050	-0.012	0.002	0.077	-0.082	0.106	-0.136
	(0.120)	(0.206)	(0.163)	(0.160)	(0.169)	(0.190)	(0.123)	(0.147)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51

Notes: This table displays the coefficients from regressions of stage 4 scores on stage 2 scores and their associated robust standard errors (between parentheses). Panel A uses the average scores of the first and second parts as the main predictor, Panel B uses the scores from the first part, and Panel C uses the scores from the second part. Stars indicate the levels of statistical significance of each coefficient: * p<.10, ** p<.05, and *** p<.01.

Table 5: Relationship between the Scores from Clinical Practice and the School Year, 2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4 (student surveys)	(6) Stage 4 (classroom observations)	(7) Stage 4 (principal surveys)
Panel A.							
Stage 3	0.265	0.319	0.409	0.611**	0.272	0.720***	0.537
	(0.258)	(0.300)	(0.307)	(0.236)	(0.431)	(0.198)	(0.434)
Stage 3 FEs?		Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes
Controls?				Yes			
N	26	26	26	26	29	31	29
Panel B.							
Stage 3 (student surveys)	0.294	0.203	0.057	0.198	0.238	0.280	0.109
	(0.237)	(0.302)	(0.329)	(0.314)	(0.272)	(0.350)	(0.355)
Stage 3 FEs?		Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes
Controls?				Yes			
N	26	26	26	26	29	31	29
Panel C.							
Stage 3 (classroom observations)	0.102	0.107	0.107	0.005	0.090	0.206	0.245
	(0.220)	(0.174)	(0.174)	(0.170)	(0.197)	(0.204)	(0.256)
Stage 3 FEs?		Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes
Controls?				Yes			
N	26	26	26	26	29	31	29

Notes: This table displays the coefficients from regressions of stage 4 scores on stage 3 scores and their associated robust standard errors (between parentheses). Panel A uses the average scores of the student surveys and classroom observations as the main predictor, Panel B uses the scores from the student surveys, and Panel C uses the scores from the classroom observations. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table 6: Relationship between the Scores from Clinical Practice and the School Year, Accounting for Prior Performance, 2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4 (student surveys)	(6) Stage 4 (classroom observations)	(7) Stage 4 (principal surveys)
Stage 3	0.264	0.398	0.416	0.750**	0.317	0.741***	0.531
	(0.239)	(0.322)	(0.243)	(0.268)	(0.371)	(0.215)	(0.404)
Stage 2	-0.031	-0.158	-0.012	-0.354	-0.548	-0.117	0.458***
	(0.148)	(0.429)	(0.382)	(0.249)	(0.326)	(0.191)	(0.144)
Stage 1	0.036	-0.043	-0.325	0.200	-0.004	-0.027	-0.659***
	(0.291)	(0.270)	(0.253)	(0.317)	(0.180)	(0.244)	(0.177)
Stage 3 FEs?		Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes
Controls?				Yes			
N	26	26	26	26	29	31	29

Notes: This table displays the coefficients from regressions of stage 4 scores on stage 3 scores, accounting for stage 1 and 2 scores, and their associated robust standard errors (between parentheses). Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table 7: Relationship between the Scores from Clinical Practice and the School Year, Accounting for Prior Performance (using Principal Component Analysis), 2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4 (student surveys)	(6) Stage 4 (classroom observations)	(7) Stage 4 (principal surveys)
Stage 3	0.322	0.622*	0.638	0.732**	0.483	0.743***	0.050
	(0.243)	(0.298)	(0.556)	(0.217)	(0.401)	(0.193)	(0.288)
Comp. 1	-0.114	-0.267	-0.301	-0.263**	-0.365**	-0.139	0.107
	(0.099)	(0.336)	(0.261)	(0.096)	(0.149)	(0.170)	(0.174)
Comp. 2	0.070	0.123	0.055	-0.058	0.181	0.122	-0.379**
	(0.163)	(0.214)	(0.247)	(0.085)	(0.179)	(0.169)	(0.128)
Comp. 3	0.222	0.067	0.083	0.750***	-0.170	0.114	0.403*
	(0.187)	(0.335)	(0.262)	(0.185)	(0.137)	(0.137)	(0.189)
Comp. 4	0.102	0.158	-0.132	0.235	0.294	-0.045	-0.460***
	(0.210)	(0.262)	(0.300)	(0.136)	(0.245)	(0.194)	(0.148)
Stage 3 FEs?		Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes
Controls?				Yes			
N	26	26	26	26	29	31	29

Notes: This table displays the coefficients from regressions of stage 4 scores on stage 3 scores, accounting for principal components of competencies assessed in stage 1 and 2, and their associated robust standard errors (between parentheses). Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table 8: Relationship between the Scores from Clinical Practice and the School Year, Accounting for Prior Performance (using a Regression-Weighted Composite), 2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4 (student surveys)	(6) Stage 4 (classroom observations)	(7) Stage 4 (principal surveys)
Stage 3	0.224	0.355	0.441	0.612**	0.332	0.687***	0.481
	(0.216)	(0.294)	(0.261)	(0.265)	(0.383)	(0.205)	(0.391)
Composite (regression weights)	0.875	0.817	0.602	1.546*	0.801	0.401	0.678
	(0.650)	(1.014)	(0.772)	(0.734)	(1.175)	(0.565)	(0.880)
Stage 3 FEs?		Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes
Controls?				Yes			
N	26	26	26	26	29	31	29

Notes: This table displays the coefficients from regressions of stage 4 scores on stage 3 scores, accounting for a (regression) weighted composite of competencies assessed in stage 1 and 2, and their associated robust standard errors (between parentheses). Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table 9: Relationship between the Scores from Clinical Practice and the School Year, Accounting for Performance on Stage 2 Demonstration Lessons, 2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4 (student surveys)	(6) Stage 4 (classroom observations)	(7) Stage 4 (principal surveys)
Stage 3	0.262	0.211	0.292	0.412	0.272	0.690***	0.416
	(0.261)	(0.362)	(0.307)	(0.229)	(0.454)	(0.214)	(0.434)
Stage 2 (demo lesson)	0.188	0.317	0.379	0.819**	0.010	0.116	0.332
	(0.210)	(0.365)	(0.220)	(0.284)	(0.340)	(0.178)	(0.305)
Stage 3 FEs?		Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes
Controls?				Yes			
N	26	26	26	26	29	31	29

Notes: This table displays the coefficients from regressions of stage 4 scores on stage 3 scores, accounting for stage 1 and 2 scores, and their associated robust standard errors (between parentheses). Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Appendix A. Additional Tables and Figures

Table A.1: Individuals Invited to Participate at Each Stage of ExA's Pipeline, 2013-2014

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	2013				2014			
	Stage 1	Stage 2	Stage 3	Stage 4	Stage 1	Stage 2	Stage 3	Stage 4
Female	.71	.67*	.64	.63	.73	.72	.7	.64
	(.45)	(.47)	(.48)	(.49)	(.45)	(.45)	(.46)	(.49)
Prov. of B.A.	.44	.39**	.34	.34	.42	.38**	.36	.48
	(.5)	(.49)	(.48)	(.48)	(.49)	(.48)	(.48)	(.51)
City of B.A.	.53	.58**	.64	.66	.52	.58***	.61	.48
	(.5)	(.49)	(.48)	(.48)	(.5)	(.49)	(.49)	(.51)
GPA (out of 10)	7.4	7.51***	7.55	7.63	7.51	7.68***	7.42	7.49
	(.91)	(.88)	(.92)	(.83)	(.94)	(.85)	(.83)	(.87)
STEM major	.13	.17***	.14	.11	.09	.11*	.18**	.21**
	(.33)	(.38)	(.35)	(.32)	(.29)	(.32)	(.39)	(.42)
Education major	.05	.04	0	0	.03	.04	.02	0
	(.21)	(.2)	(0)	(0)	(.17)	(.19)	(.12)	(0)
Graduated	.82	.85*	.78	.77	.86	.86	.79	.76
	(.38)	(.35)	(.42)	(.43)	(.35)	(.35)	(.41)	(.44)
Applied before	.04	.05	.08	.09	.08	.09	.06	.12
	(.19)	(.22)	(.27)	(.28)	(.27)	(.28)	(.24)	(.33)
N	822	326	50	35	1420	439	66	33

Notes: This table displays the means and standard deviations (between parentheses) of a set of variables collected during ExA's online application for individuals who *participated* at each stage of ExA's selection, training, and placement pipeline (i.e., who have a non-missing score for that stage). Stars indicate the levels of statistical significance for t-tests comparing the means of individuals who participated and who did not participate at each stage: * p<.10, ** p<.05, and *** p<.01.

Table A.2: Individuals With and Without Scores from the School Year, 2013-2014

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Total score		Student survey		Classroom obs.		Principal survey	
	Yes	No	Yes	No	Yes	No	Yes	No
Female	.62	.89	.63	1	.65	1	.65	.8
	(.49)	(.33)	(.49)	(0)	(.48)	(0)	(.48)	(.45)
Lives in Prov. B.A.	.38	.44	.4	.25	.41	0	.37	.6
	(.49)	(.53)	(.5)	(.5)	(.5)	(0)	(.49)	(.55)
Lives in City of B.A.	.6	.56	.58	.75	.57	1	.61	.4
	(.5)	(.53)	(.5)	(.5)	(.5)	(0)	(.49)	(.55)
GPA	7.51	7.6	7.52	7.59	7.51	8	7.52	7.62
	(.82)	(1.02)	(.84)	(1.04)	(.86)	(0)	(.83)	(1.13)
STEM major	.17	.22	.17	.25	.19	0	.18	.2
	(.38)	(.44)	(.38)	(.5)	(.39)	(0)	(.39)	(.45)
Education major	0	0	0	0	0	0	0	0
	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0)
Graduated	.72	.78	.73	.75	.72	1	.73	.8
	(.45)	(.44)	(.45)	(.5)	(.45)	(0)	(.45)	(.45)
Applied before	.11	0	.1	0	.09	0	.1	0
	(.31)	(0)	(.3)	(0)	(.29)	(0)	(.3)	(0)
N	47	9	52	4	54	2	51	5

Notes: This table displays the means and standard deviations (between parentheses) of a set of variables collected during ExA's online application for individuals with and without missing stage 4 scores. Stars indicate the levels of statistical significance for t-tests comparing the means of individuals with and without missing stage 4 scores: * p<.10, ** p<.05, and *** p<.01.

Table A.3: Pairwise Correlations between the Scores from All Stages of ExA's Pipeline, 2013-2014

	$STAGE1$	$STAGE2$	$STAGE2\_FPT$	$STAGE2\_SPT$	$STAGE3$	$STAGE3\_OBS$	$STAGE3\_STU$	$STAGE4$	$STAGE4\_OBS$	$STAGE4\_STU$	$STAGE4\_PRI$
$STAGE1$	1										
$STAGE2$	0.257	1									
$STAGE2\_FPT$	0.184	0.849***	1								
$STAGE2\_SPT$	0.210	0.619***	0.110	1							
$STAGE3$	0.326	0.223	0.170	0.189	1						
$STAGE3\_OBS$	0.079	0.223	0.234	0.095	0.679***	1					
$STAGE3\_STU$	0.363*	0.080	-0.003	0.161	0.679***	-0.078	1				
$STAGE4$	-0.001	0.061	0.131	-0.083	0.261	0.100	0.272	1			
$STAGE4\_OBS$	0.048	0.120	0.184	-0.047	0.336	0.307	0.149	0.839***	1		
$STAGE4\_STU$	0.012	-0.105	-0.026	-0.162	0.206	0.101	0.178	0.740***	0.406**	1	
$STAGE4\_PRI$	-0.106	-0.026	0.025	-0.086	0.151	0.018	0.198	0.759***	0.654***	0.267	1

Notes: This table shows the pairwise correlations between all of the standardized versions of the variables described in Section 4. Stars indicate the levels of statistical significance of each correlation coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.4: Relationship between the Scores from the Online Application and School Year, 2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4 (student surveys)	(5) Stage 4 (classroom observations)	(6) Stage 4 (principal surveys)
Stage 1	0.129	-0.057	-0.395	-0.395	0.300	-0.236
	(0.355)	(0.429)	(0.620)	(0.577)	(0.202)	(0.331)
Stage 1 FEs?		Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes
	26	26	26	29	31	29

Notes: This table displays the coefficients from regressions of stage 4 scores on stage 1 scores and their associated robust standard errors (between parentheses). These regressions include only the 2014 cohort of ExA's corps members. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.5: Relationship between Corps Members' Characteristics and the School Year Scores, 2013-2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4 (student surveys)	(4) Stage 4 (classroom observations)	(5) Stage 4 (principal surveys)
Female	-0.195	-0.085	0.019	-0.235	-0.493
	(0.295)	(0.305)	(0.290)	(0.297)	(0.297)
Lives in City of B.A.	-0.540*	-0.638**	-0.700**	-0.183	-0.358
	(0.295)	(0.314)	(0.271)	(0.296)	(0.250)
GPA	0.284	0.424*	0.280	0.188	0.296
	(0.237)	(0.237)	(0.253)	(0.173)	(0.221)
STEM major	-0.161	-0.196	-0.037	-0.123	-0.620
	(0.422)	(0.476)	(0.423)	(0.417)	(0.460)
Graduated	0.073	-0.060	-0.420	-0.004	0.083
	(0.371)	(0.348)	(0.306)	(0.371)	(0.321)
Applied before	-0.267	-0.421	-0.334	-0.059	-0.021
	(0.301)	(0.308)	(0.341)	(0.328)	(0.411)
Constant		-2.635	-1.372	-1.123	-1.636
		(1.751)	(1.869)	(1.281)	(1.635)
	47	47	52	54	51

Notes: This table displays the coefficients from regressions of stage 4 scores on corps members' characteristics and their associated robust standard errors (between parentheses). Column (1) includes the results from bivariate regressions (the constants are omitted). Columns (2) through (5) include the results from multivariate regressions. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.6: Relationship between the Scores from the Online Application and School Year by Competency, 2013-2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4	(6) Stage 4 (student surveys)	(7) Stage 4 (classroom observations)	(8) Stage 4 (principal surveys)
Panel A.								
Stage 1 (accomplishment)	0.222 (0.191)	-0.046 (0.301)	-0.240 (0.249)	-0.177 (0.269)	-0.593 (0.692)	-0.309 (0.275)	-0.204 (0.146)	-0.203 (0.237)
Stage 1 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	47	52	54
Panel B.								
Stage 1 (leadership)	-0.146 (0.142)	-0.390** (0.140)	-0.236 (0.152)	-0.273* (0.153)	-0.297* (0.146)	-0.270 (0.183)	-0.103 (0.157)	-0.088 (0.152)
Stage 1 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	47	52	54
Panel C.								
Stage 1 (perseverance)	-0.083 (0.124)	-0.331 (0.329)	-0.225 (0.226)	-0.226 (0.188)	-0.170 (0.214)	-0.129 (0.222)	0.019 (0.221)	-0.319** (0.151)
Stage 1 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	47	52	54

Notes: This table displays the coefficients from regressions of stage 4 scores on each criterion assessed during stage 1: accomplishment (Panel A), leadership (Panel B), and perseverance (Panel C). Their associated robust standard errors are shown between parentheses. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.7: Relationship between the Scores from the Online Application and School Year by Initial Performance, 2013-2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4 (student surveys)	(6) Stage 4 (classroom observations)	(7) Stage 4 (principal surveys)
Stage 1	-0.001	-0.572*	-0.441**	-0.436**	-0.394**	-0.163	-0.294**
	(0.179)	(0.299)	(0.192)	(0.169)	(0.164)	(0.186)	(0.140)
Stage 1 (squared)	-0.165	0.040	0.137	0.252	0.347**	0.245	0.037
	(0.175)	(0.193)	(0.117)	(0.147)	(0.155)	(0.182)	(0.152)
Stage 1 FEs?		Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes
N	47	47	47	47	52	54	51

Notes: This table displays the coefficients from regressions of stage 4 scores on the stage 1 scores and the stage 1 scores squared and their associated robust standard errors (between parentheses). Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.8: Relationship between the Scores from the Online Application and School Year, by Instrument and Round of Data Collection, 2013-2014

	(1)	(2)	(3)	(4)
Panel A. Stage 4 (student surveys)				
Stage 1	0.108	-0.348**	-0.393***	-0.391***
	(0.169)	(0.150)	(0.145)	(0.147)
End of year	-0.006	0.146	0.138	0.137
	(0.208)	(0.155)	(0.156)	(0.157)
Stage 1*End of year	-0.105	0.003	0.008	0.009
	(0.273)	(0.144)	(0.139)	(0.143)
Stage 1 FEs?		Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes
Cohort FEs?				Yes
N	92	92	92	92
Panel B. Stage 4 (classroom observations)				
Stage 1	0.008	-0.339*	-0.202	-0.208
	(0.152)	(0.202)	(0.163)	(0.163)
End of year	0.005	0.006	-0.003	-0.004
	(0.196)	(0.164)	(0.144)	(0.145)
Stage 1*End of year	0.113	0.090	0.086	0.087
	(0.204)	(0.167)	(0.143)	(0.144)
Stage 1 FEs?		Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes
Cohort FEs?				Yes
N	106	106	106	106
Panel C. Stage 4 (principal surveys)				
Stage 1	-0.137	-0.457**	-0.289*	-0.313*
	(0.153)	(0.227)	(0.168)	(0.163)
End of year	-0.015	-0.103	-0.044	-0.043
	(0.222)	(0.200)	(0.179)	(0.178)
Stage 1*End of year	0.096	0.246	0.146	0.153
	(0.202)	(0.256)	(0.222)	(0.214)
Stage 1 FEs?		Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes
Cohort FEs?				Yes
N	82	82	82	82

Notes: This table displays the coefficients from regressions of stage 4 scores for stage 1 scores, a dummy for whether the stage 4 score is from the end of the school year, and its interaction with the stage 1 score. Robust standard errors are between parentheses. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.9: Relationship between the Scores from the Online Application and School Year, by Instrument and Cohort, 2013-2014

	(1)	(2)	(3)
Panel A. Stage 4 (student surveys)			
Stage 1	-0.106	-0.294	-0.376
	(0.193)	(0.215)	(0.277)
2014 cohort	-0.191	0.088	0.105
	(0.266)	(0.458)	(0.445)
Stage 1*2014 cohort	0.203	-0.292	-0.168
	(0.354)	(0.443)	(0.568)
Stage 1 FEs?		Yes	Yes
Stage 4 FEs?			Yes
N	52	52	52
Panel B. Stage 4 (classroom observations)			
Stage 1	-0.116	-0.584	-0.273
	(0.159)	(0.363)	(0.275)
2014 cohort	-0.150	0.009	-0.017
	(0.278)	(0.580)	(0.485)
Stage 1*2014 cohort	0.307	1.097**	0.419
	(0.299)	(0.437)	(0.440)
Stage 1 FEs?		Yes	Yes
Stage 4 FEs?			Yes
N	54	54	54
Panel C. Stage 4 (principal surveys)			
Stage 1	-0.193	-0.511	-0.267
	(0.162)	(0.321)	(0.164)
2014 cohort	-0.135	-0.259	-0.207
	(0.274)	(0.514)	(0.418)
Stage 1*2014 cohort	0.143	0.321	-0.054
	(0.241)	(0.478)	(0.369)
Stage 1 FEs?		Yes	Yes
Stage 4 FEs?			Yes
N	51	51	51

Notes: This table displays the coefficients from regressions of stage 4 scores for stage 1 scores, a dummy for whether the stage 4 score is from the end of the school year, and its interaction with the stage 1 score. Robust standard errors are between parentheses. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.10: Relationship between the Scores from the Assessment Center and School Year, 2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(5) Stage 4 (student surveys)	(6) Stage 4 (classroom observations)	(7) Stage 4 (principal surveys)
Panel A.						
Stage 2	0.056	-0.266	-0.286	-0.363	0.241	-0.462
	(0.135)	(0.913)	(1.109)	(0.200)	(0.738)	(1.046)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes
N	26	26	26	29	31	29
Panel B.						
Stage 2 (first part)	0.213	0.118	0.099	-0.318	-0.044	0.126
	(0.178)	(0.428)	(0.383)	(0.478)	(0.272)	(0.292)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes
N	26	26	26	29	31	29
Panel C.						
Stage 2 (second part)	-0.218	-0.171	-0.218	-0.404	-0.025	-0.152
	(0.134)	(0.252)	(0.330)	(0.348)	(0.232)	(0.206)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes
N	26	26	26	29	31	29
Panel C.						
Stage 2 (demo lesson)	0.259	0.246	0.301	0.142	0.370**	0.279
	(0.198)	(0.251)	(0.230)	(0.245)	(0.139)	(0.201)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes
N	26	26	26	29	31	29

Notes: This table displays the coefficients from regressions of stage 4 scores on stage 2 scores and their associated robust standard errors (between parentheses). Panel A uses the average scores of the first and second parts as the main predictor, Panel B uses the scores from the first part, and Panel C uses the scores from the second part. These regressions include only the 2014 cohort of ExA's corps members. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.11: Relationship between the Scores from the Assessment Center and School Year by Initial Performance, 2013-2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4 (student surveys)	(6) Stage 4 (classroom observations)	(7) Stage 4 (principal surveys)
Stage 2	0.089	0.221	0.144	0.148	0.144	0.179	-0.069
	(0.110)	(0.259)	(0.305)	(0.364)	(0.264)	(0.266)	(0.227)
Stage 2 (squared)	0.159*	0.043	-0.190	-0.185	0.070	-0.393**	-0.069
	(0.094)	(0.314)	(0.494)	(0.543)	(0.225)	(0.158)	(0.405)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes
N	47	47	47	47	52	54	51

Notes: This table displays the coefficients from regressions of stage 4 scores on the stage 2 scores and the stage 2 scores squared and their associated robust standard errors (between parentheses). Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.12: Relationship between the Scores from the Assessment Center and School Year, by Instrument and Round of Data Collection, 2013-2014

	(1)	(2)	(3)	(4)
Panel A. Stage 4 (student surveys)				
Stage 2	-0.145	0.014	0.051	0.103
	(0.150)	(0.123)	(0.130)	(0.144)
End of year	-0.002	0.144	0.141	0.157
	(0.209)	(0.121)	(0.116)	(0.117)
Stage 2*End of year	0.221	0.045	0.017	0.022
	(0.200)	(0.124)	(0.126)	(0.124)
Stage 2 FEs?		Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes
Cohort FEs?				Yes
N	92	92	92	92
Panel B. Stage 4 (classroom observations)				
Stage 2	0.008	0.177	0.107	0.197
	(0.125)	(0.172)	(0.166)	(0.170)
End of year	0.0001	-0.018	-0.029	-0.028
	(0.195)	(0.153)	(0.146)	(0.143)
Stage 2*End of year	0.119	0.119	0.116	0.116
	(0.174)	(0.146)	(0.149)	(0.142)
Stage 2 FEs?		Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes
Cohort FEs?				Yes
N	106	106	106	106
Panel C. Stage 4 (principal surveys)				
Stage 2	0.069	0.158	0.096	0.039
	(0.153)	(0.185)	(0.177)	(0.189)
End of year	-0.000	-0.174	-0.118	-0.133
	(0.225)	(0.199)	(0.213)	(0.215)
Stage 2*End of year	-0.091	-0.128	-0.090	-0.090
	(0.206)	(0.132)	(0.138)	(0.140)
Stage 2 FEs?		Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes
Cohort FEs?				Yes
N	82	82	82	82

Notes: This table displays the coefficients from regressions of stage 4 scores for stage 2 scores, a dummy for whether the stage 4 score is from the end of the school year, and its interaction with the stage 2 score. Robust standard errors are between parentheses. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.13: Relationship between the Scores from the Assessment Center and School Year, by Instrument and Cohort, 2013-2014

	(1)	(2)	(3)
Panel A. Stage 4 (student surveys)			
Stage 2	0.060	0.321	0.439
	(0.203)	(0.241)	(0.247)
2014 cohort	-0.182	-0.451	-0.232
	(0.275)	(0.445)	(0.385)
Stage 2*2014 cohort	-0.241	-0.444	-0.644**
	(0.279)	(0.358)	(0.271)
Stage 2 FEs?		Yes	Yes
Stage 4 FEs?			Yes
N	52	52	52
Panel B. Stage 4 (classroom observations)			
Stage 2	0.202	0.513*	0.315
	(0.186)	(0.264)	(0.380)
2014 cohort	-0.200	-0.516	-0.542
	(0.282)	(0.502)	(0.660)
Stage 2*2014 cohort	-0.117	-0.232	-0.044
	(0.234)	(0.445)	(0.576)
Stage 2 FEs?		Yes	Yes
Stage 4 FEs?			Yes
N	54	54	54
Panel C. Stage 4 (principal surveys)			
Stage 2	-0.211	0.248	0.108
	(0.200)	(0.234)	(0.255)
2014 cohort	-0.054	0.528	0.872
	(0.301)	(1.012)	(0.812)
Stage 2*2014 cohort	0.329	-0.462	-0.408
	(0.282)	(0.550)	(0.580)
Stage 2 FEs?		Yes	Yes
Stage 4 FEs?			Yes
N	51	51	51

Notes: This table displays the coefficients from regressions of stage 4 scores for stage 2 scores, a dummy for whether the stage 4 score is from the end of the school year, and its interaction with the stage 2 score. Robust standard errors are between parentheses. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.14: Relationship between the Scores from the First Part of the Assessment Center and School Year by Competency, 2013-2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4	(6) Stage 4 (student surveys)	(7) Stage 4 (classroom observations)	(8) Stage 4 (principal surveys)
Panel A.								
Stage 2 (organization)	0.055 (0.116)	0.201 (0.294)	0.222 (0.321)	0.371 (0.450)	0.545 (0.463)	0.144 (0.260)	0.201 (0.346)	0.153 (0.339)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51
Panel B.								
Stage 2 (communication)	0.173 (0.150)	0.102 (0.281)	0.060 (0.274)	0.063 (0.303)	0.175 (0.281)	0.100 (0.212)	0.132 (0.250)	-0.079 (0.219)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51

Notes: This table displays the coefficients from regressions of stage 4 scores on each criterion assessed in the first part of stage 2: organization (Panel A) and communication (Panel B). Their associated robust standard errors are shown between parentheses. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.15: Relationship between the Scores from the Second Part of the Assessment Center and School Year by Competency, 2013-2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4	(6) Stage 4 (student surveys)	(7) Stage 4 (classroom observations)	(8) Stage 4 (principal surveys)
Panel A.								
Stage 2	0.057	0.426**	0.408	0.411	0.381	0.317	0.351	0.043
(leadership)	(0.135)	(0.186)	(0.267)	(0.296)	(0.375)	(0.212)	(0.284)	(0.182)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes	Yes	Yes	Yes
N	47	47	47	47	47	52	54	51
Panel B.								
Stage 2	-0.208	-0.122	-0.059	-0.068	0.044	-0.008	-0.023	-0.147
(perseverance)	(0.125)	(0.238)	(0.215)	(0.222)	(0.313)	(0.196)	(0.252)	(0.176)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes	Yes	Yes	Yes
N	47	47	47	47	47	52	54	51
Panel C.								
Stage 2	-0.057	-0.008	-0.049	-0.058	0.185	-0.122	0.197	-0.070
(communication)	(0.132)	(0.634)	(0.470)	(0.533)	(0.952)	(0.249)	(0.350)	(0.268)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes	Yes	Yes	Yes
N	47	47	47	47	47	52	54	51
Panel D.								
Stage 2	-0.094	-0.073	-0.036	-0.042	0.116	-0.183	0.059	-0.013
(openness to feedback)	(0.153)	(0.230)	(0.259)	(0.301)	(0.616)	(0.179)	(0.280)	(0.206)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes	Yes	Yes	Yes
N	47	47	47	47	47	52	54	51
Panel D.								
Stage 2	0.083	0.129	0.110	0.110	0.156	0.015	0.189	-0.004
(alignment)	(0.148)	(0.286)	(0.209)	(0.219)	(0.329)	(0.193)	(0.182)	(0.134)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes	Yes	Yes	Yes
N	47	47	47	47	47	52	54	51

Notes: This table displays the coefficients from regressions of stage 4 scores on each criterion assessed in the individual activities of stage 2: leadership (Panel A), perseverance (Panel B), communication (Panel C), and openness to feedback (Panel D). Their associated robust standard errors are shown between parentheses. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.16: Relationship between the Scores from Selected Activities of the Assessment Center and School Year, 2013-2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4	(6) Stage 4 (student surveys)	(7) Stage 4 (classroom observations)	(8) Stage 4 (principal surveys)
Panel A.								
Stage 2 (demo lesson)	0.258*	0.258	0.213	0.222	0.199	0.139	0.348***	0.200
	(0.139)	(0.166)	(0.134)	(0.142)	(0.151)	(0.170)	(0.115)	(0.174)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51
Panel B.								
Stage 2 (critical thinking)	0.051	-0.073	-0.109	-0.102	0.040	-0.061	0.027	-0.035
	(0.148)	(0.164)	(0.167)	(0.164)	(0.184)	(0.187)	(0.129)	(0.135)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51
Panel C.								
Stage 2 (written exercise)	-0.012	-0.093	-0.097	-0.083	0.031	-0.156	-0.042	-0.139
	(0.124)	(0.169)	(0.154)	(0.155)	(0.149)	(0.183)	(0.137)	(0.133)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51

Notes: This table displays the coefficients from regressions of stage 4 scores on the scores from the individual activities at stage 2 and their associated robust standard errors (between parentheses). Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.17: Relationship between the Scores from the Demonstration Lesson at the Assessment Center and School Year by Competency, 2013-2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4	(5) Stage 4	(6) Stage 4 (student surveys)	(7) Stage 4 (classroom observations)	(8) Stage 4 (principal surveys)
Panel A.								
Stage 2 (planning)	0.235 (0.147)	0.222 (0.166)	0.283* (0.139)	0.285** (0.138)	0.332** (0.136)	0.304* (0.166)	0.315*** (0.111)	0.086 (0.135)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51
Panel B.								
Stage 2 (organization)	0.300** (0.148)	0.263 (0.175)	0.223 (0.176)	0.219 (0.175)	0.222 (0.172)	0.155 (0.216)	0.244* (0.132)	0.111 (0.148)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51
Panel C.								
Stage 2 (student engagement)	0.111 (0.133)	0.075 (0.167)	0.056 (0.153)	0.073 (0.155)	0.112 (0.149)	-0.029 (0.152)	0.180 (0.112)	0.130 (0.135)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51
Panel D.								
Stage 2 (listening and answering)	0.147 (0.146)	0.254 (0.156)	0.144 (0.166)	0.142 (0.167)	-0.035 (0.174)	0.088 (0.155)	0.151 (0.151)	0.226 (0.149)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51
Panel E.								
Stage 2 (communication)	0.101 (0.143)	0.055 (0.181)	0.041 (0.160)	0.051 (0.163)	0.045 (0.157)	0.018 (0.168)	0.253* (0.127)	0.069 (0.173)
Stage 2 FEs?		Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs?				Yes	Yes	Yes	Yes	Yes
Controls?					Yes			
N	47	47	47	47	47	52	54	51

Notes: This table displays the coefficients from regressions of stage 4 scores on each of the competencies assessed in the demonstration at stage 2 and their associated robust standard errors (between parentheses). Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.18: Relationship between the Scores from Clinical Practice and the School Year by Initial Performance, 2014

	(1) Stage 4	(2) Stage 4	(3) Stage 4	(4) Stage 4 (student surveys)	(5) Stage 4 (classroom observations)	(6) Stage 4 (principal surveys)
Stage 3	0.154	0.274	0.439	0.305	0.762***	0.526
	(0.186)	(0.266)	(0.250)	(0.372)	(0.197)	(0.439)
Stage 3 (squared)	-0.374**	-0.295*	-0.311***	-0.473**	-0.253*	0.060
	(0.138)	(0.166)	(0.094)	(0.163)	(0.132)	(0.133)
Stage 3 FEs?		Yes	Yes	Yes	Yes	Yes
Stage 4 FEs?			Yes	Yes	Yes	Yes
N	26	26	26	29	31	29

Notes: This table displays the coefficients from regressions of stage 4 scores on dummies for corps members who were in the bottom 10% (Panel A) or top 10% (Panel B) of the stage 2 scores distribution and their associated robust standard errors (between parentheses). These regressions include only the 2014 cohort of ExA's corps members. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.19: Relationship between the Scores from Clinical Practice and the School Year, by Instrument and Round of Data Collection, 2014

	(1)	(2)	(3)
Panel A. Stage 4 (student surveys)			
Stage 3	0.242	0.398	0.279
	(0.229)	(0.263)	(0.297)
End of year	0.161	0.185	0.171
	(0.305)	(0.231)	(0.224)
Stage 3*End of year	0.014	-0.033	-0.009
	(0.369)	(0.260)	(0.259)
Stage 3 FEs?		Yes	Yes
Stage 4 FEs?			Yes
N	54	54	54
Panel B. Stage 4 (classroom observations)			
Stage 3	0.338**	0.420*	0.630***
	(0.159)	(0.216)	(0.194)
End of year	0.157	0.151	0.145
	(0.251)	(0.243)	(0.224)
Stage 3*End of year	-0.110	-0.102	-0.095
	(0.220)	(0.216)	(0.214)
Stage 3 FEs?		Yes	Yes
Stage 4 FEs?			Yes
N	61	61	61
Panel C. Stage 4 (principal surveys)			
Stage 3	0.138	0.407*	0.520*
	(0.172)	(0.240)	(0.275)
End of year	-0.051	-0.075	-0.087
	(0.315)	(0.325)	(0.300)
Stage 3*End of year	0.067	0.019	0.028
	(0.306)	(0.297)	(0.276)
Stage 3 FEs?		Yes	Yes
Stage 4 FEs?			Yes
N	49	49	49

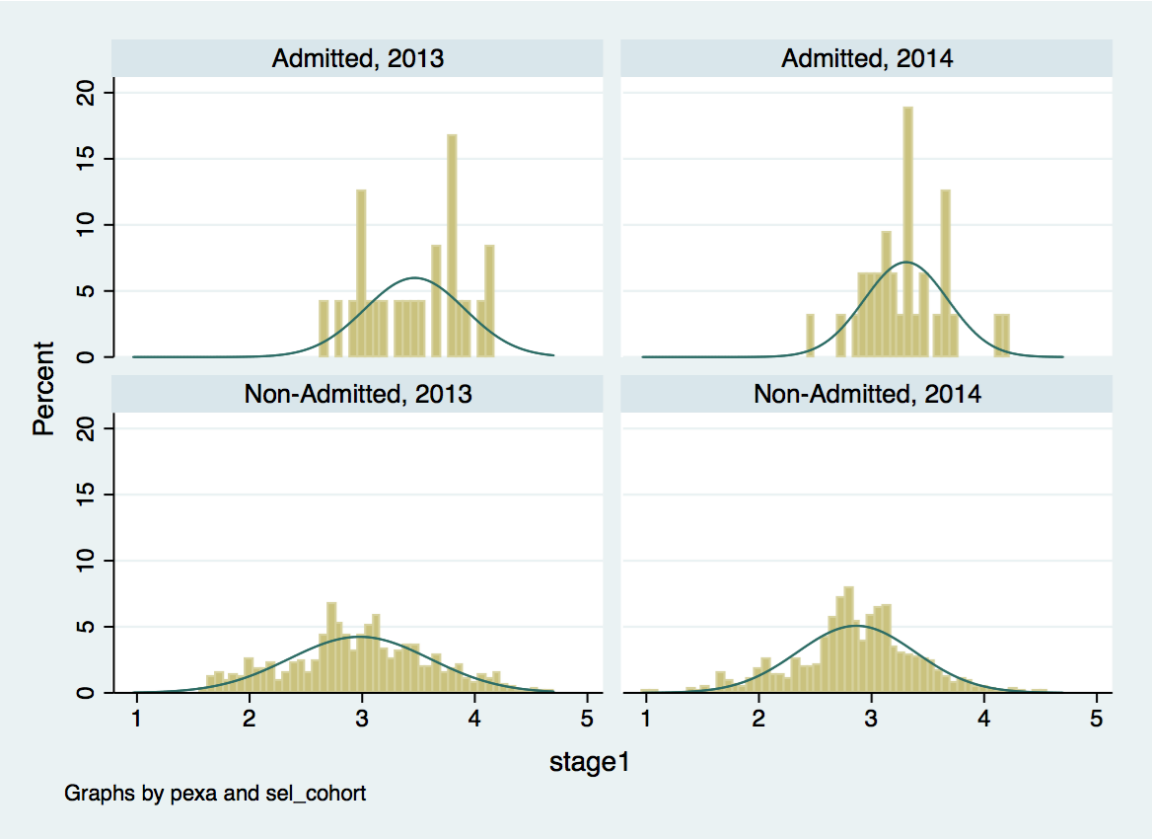
Notes: This table displays the coefficients from regressions of stage 4 scores for stage 3 scores, a dummy for whether the stage 4 score is from the end of the school year, and its interaction with the stage 3 score. Robust standard errors are between parentheses. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Table A.20: Principal Component Analysis of Competencies Assessed in Stages 1 and 2, 2014

	Eigenvalue	Difference	Proportion	Cumulative
Component 1	2.04188	.37205	0.2042	0.2042
Component 2	1.66983	.229801	0.1670	0.3712
Component 3	1.44003	.33745	0.1440	0.5152
Component 4	1.10258	.105627	0.1103	0.6254
Component 5	.996953	.152312	0.0997	0.7251
Component 6	.844641	.0816291	0.0845	0.8096
Component 7	.763012	.212833	0.0763	0.8859
Component 8	.550179	.140783	0.0550	0.9409
Component 9	.409396	.2279	0.0409	0.9819
Component 10	.181496	-	0.0181	1

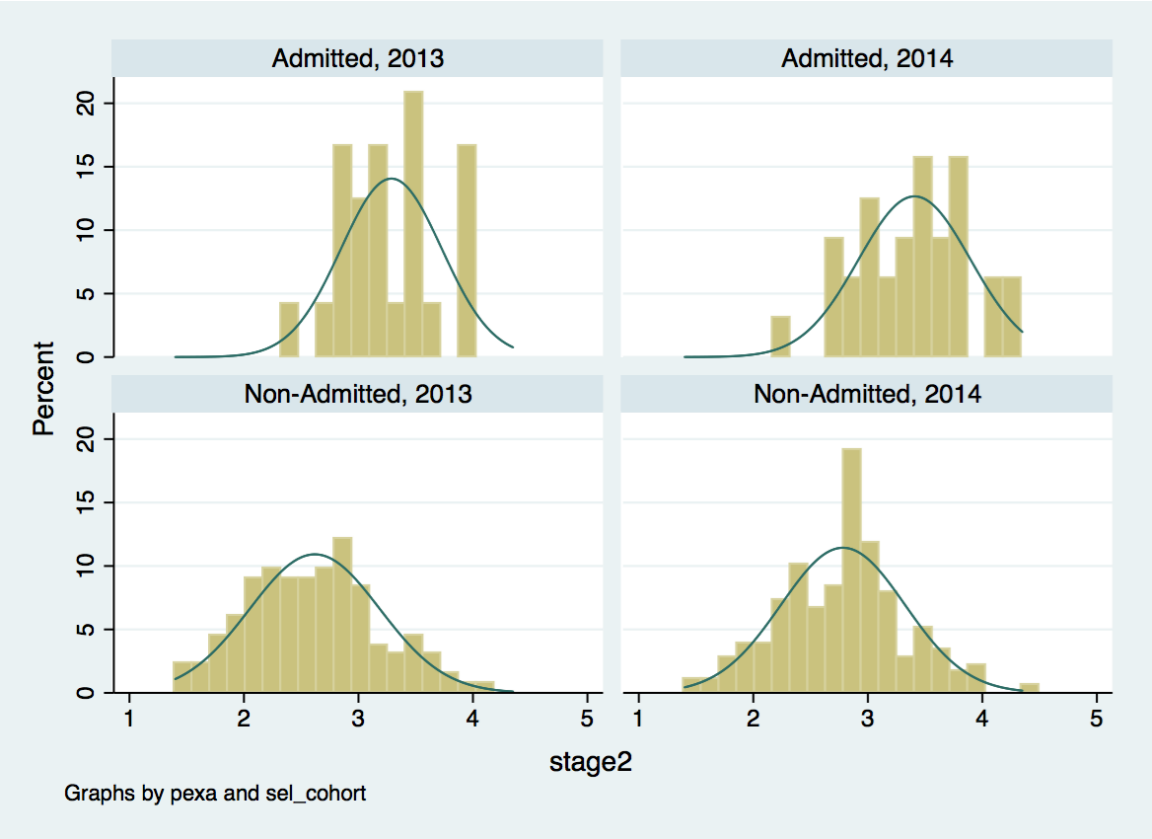
Notes: This table shows the results from a principal component analysis of all 10 competencies assessed in stages 1 and 2 for the 32 corps members with clinical practice scores.

Figure A.1: Online Application Scores, by Cohort and Admission Status



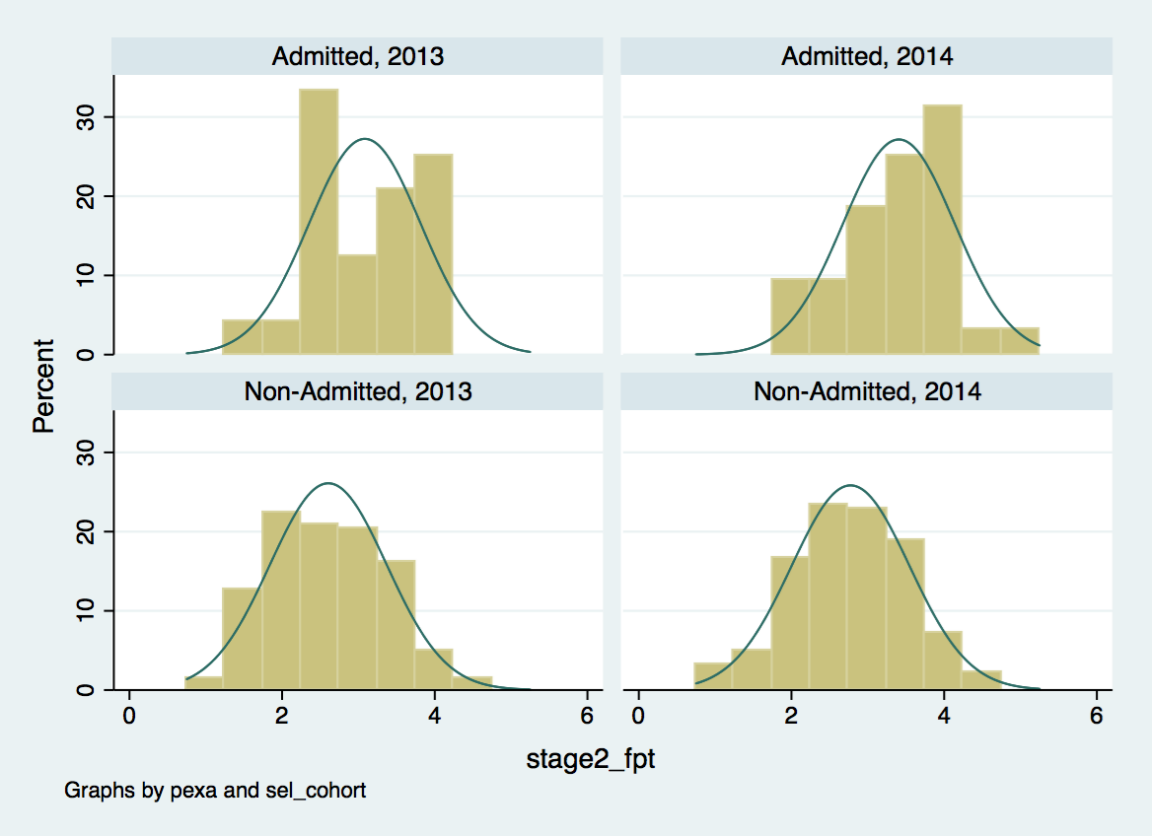
Notes: These histograms display the distribution of Stage 1 scores by cohort and admission status, with superimposed normal densities. The first column shows the distributions for the 2013 cohort and the second column shows the distributions for the 2014 cohort. The first row shows the distribution for ExA's corps members and the second row shows the distribution for all applicants. This graph only includes individuals with non-missing scores for the three of the competencies assessed in the online application.

Figure A.2: Assessment Center Scores, by Cohort and Admission Status



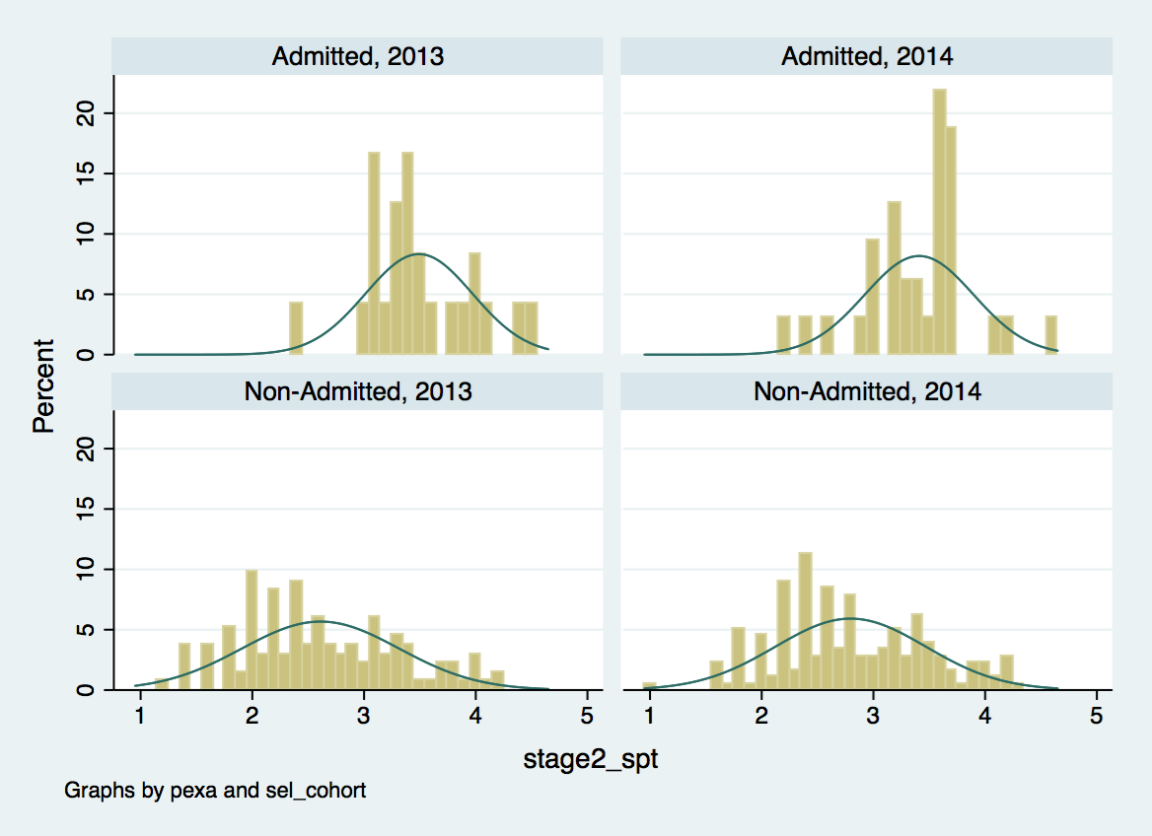
Notes: These histograms display the distribution of Stage 2 scores, by cohort and admission status, with superimposed normal densities. The first column shows the distributions for the 2013 cohort and the second column shows the distributions for the 2014 cohort. The first row shows the distribution for ExA's corps members and the second row shows the distribution for all applicants. This graph only includes individuals with non-missing scores for the three competencies assessed in the individual activities.

Figure A.3: Scores from the First Part of the Assessment Center, by Cohort and Admission Status



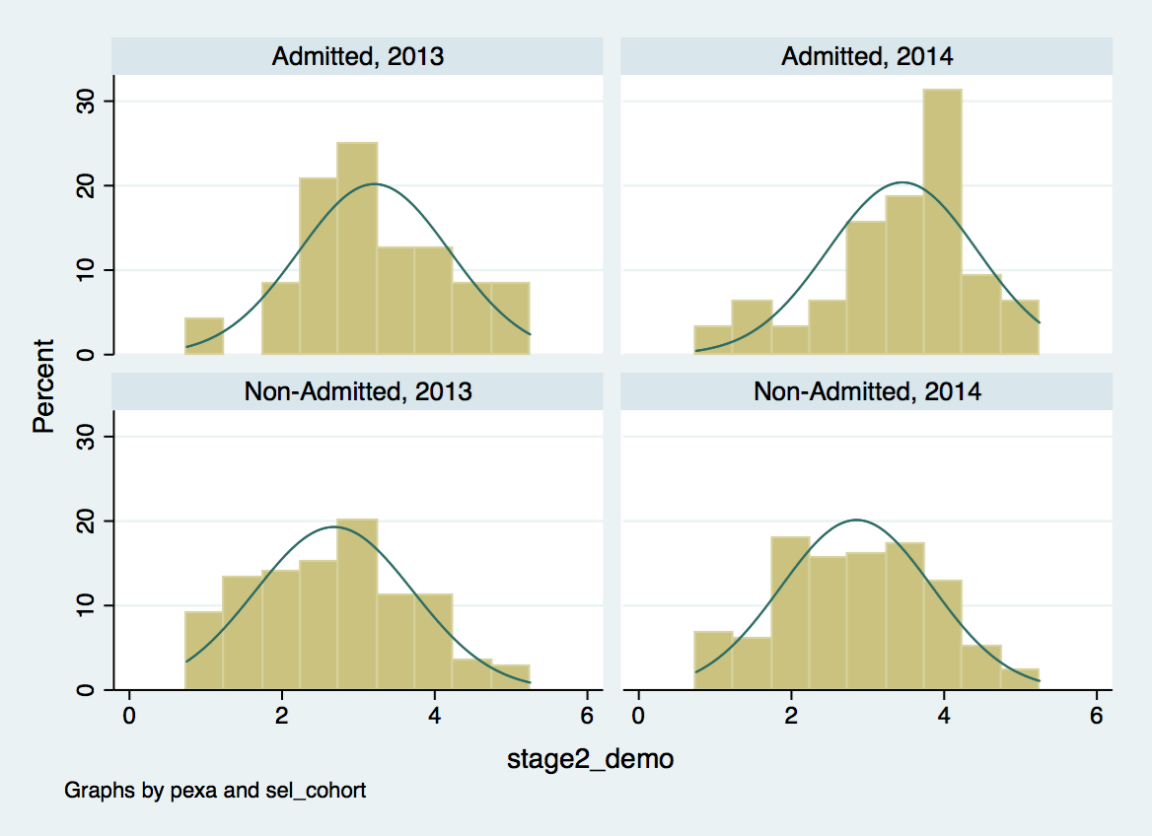
Notes: These histograms display the distribution of Stage 2 scores in the first part of the assessment center, by cohort and admission status, with superimposed normal densities. The first column shows the distributions for the 2013 cohort and the second column shows the distributions for the 2014 cohort. The first row shows the distribution for ExA's corps members and the second row shows the distribution for all applicants. This graph only includes individuals with non-missing scores for the two competencies assessed in the first part.

Figure A.4: Scores from the Second Part of the Assessment Center, by Cohort and Admission Status



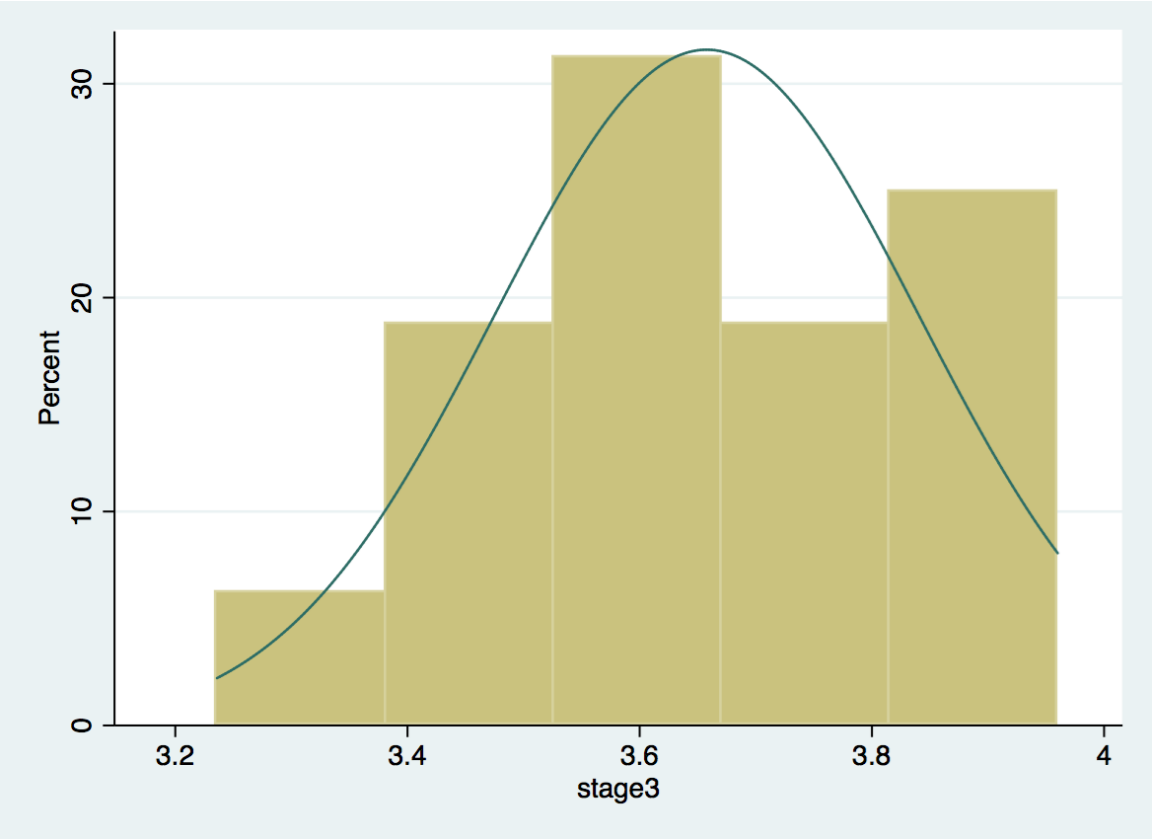
Notes: These histograms display the distribution of Stage 2 scores in the second part of the assessment center, by cohort and admission status, with superimposed normal densities. The first column shows the distributions for the 2013 cohort and the second column shows the distributions for the 2014 cohort. The first row shows the distribution for ExA's corps members and the second row shows the distribution for all applicants. This graph only includes individuals with non-missing scores for the five competencies assessed in the individual activities.

Figure A.5: Scores from the Demonstration Lesson at the Assessment Center, by Cohort and Admission Status



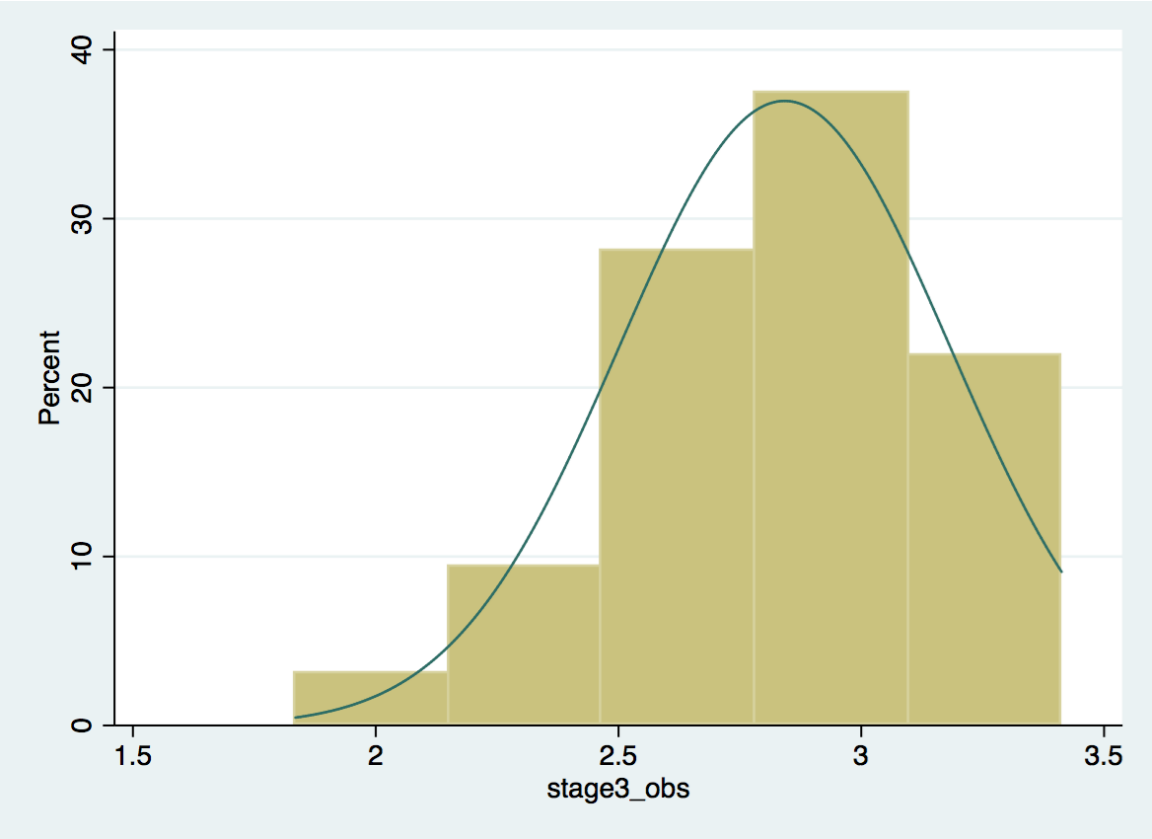
Notes: These histograms display the distribution of Stage 2 scores in the demonstration lesson, by cohort and admission status, with superimposed normal densities. The first column shows the distributions for the 2013 cohort and the second column shows the distributions for the 2014 cohort. The first row shows the distribution for ExA's corps members and the second row shows the distribution for all applicants. This graph only includes individuals with non-missing scores for the five competencies assessed in the demonstration lessons.

Figure A.6: Clinical Practice Scores, 2014



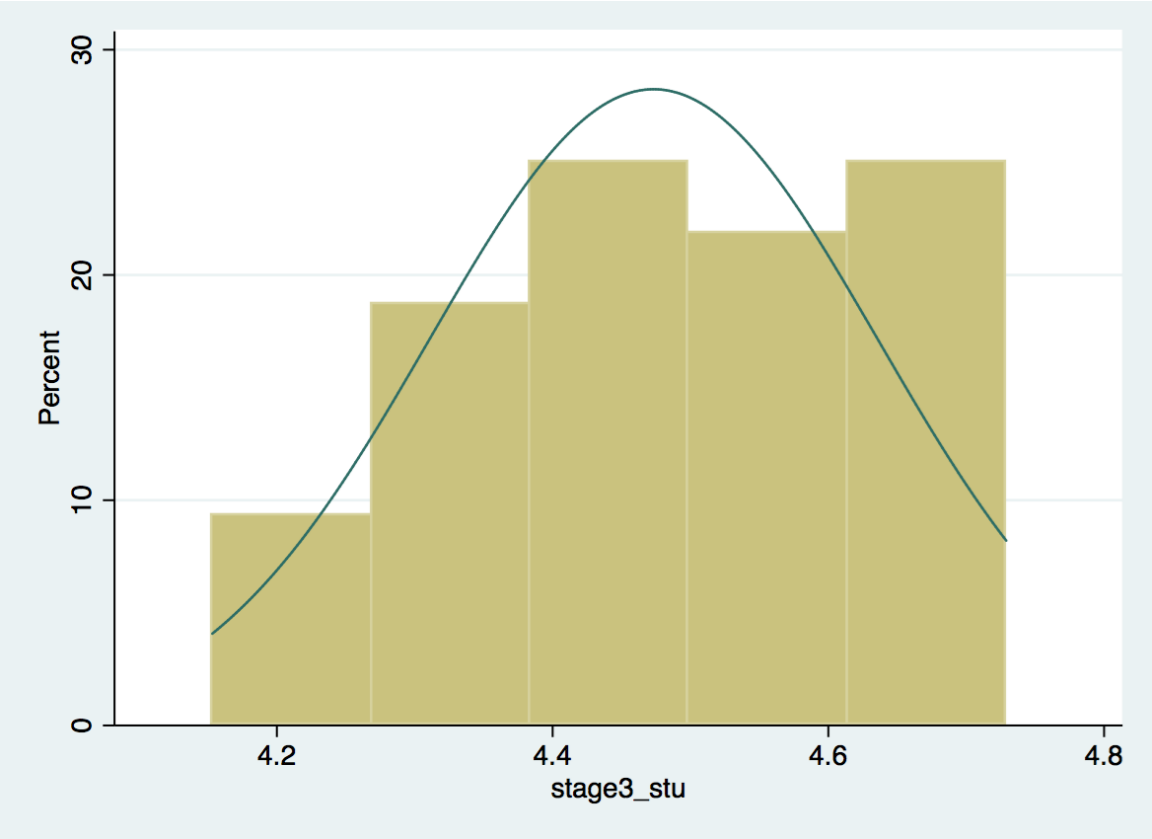
Notes: This histogram displays the distribution of Stage 3 scores, with a superimposed normal density.

Figure A.7: Scores from the Classroom Observations at Clinical Practice, 2014



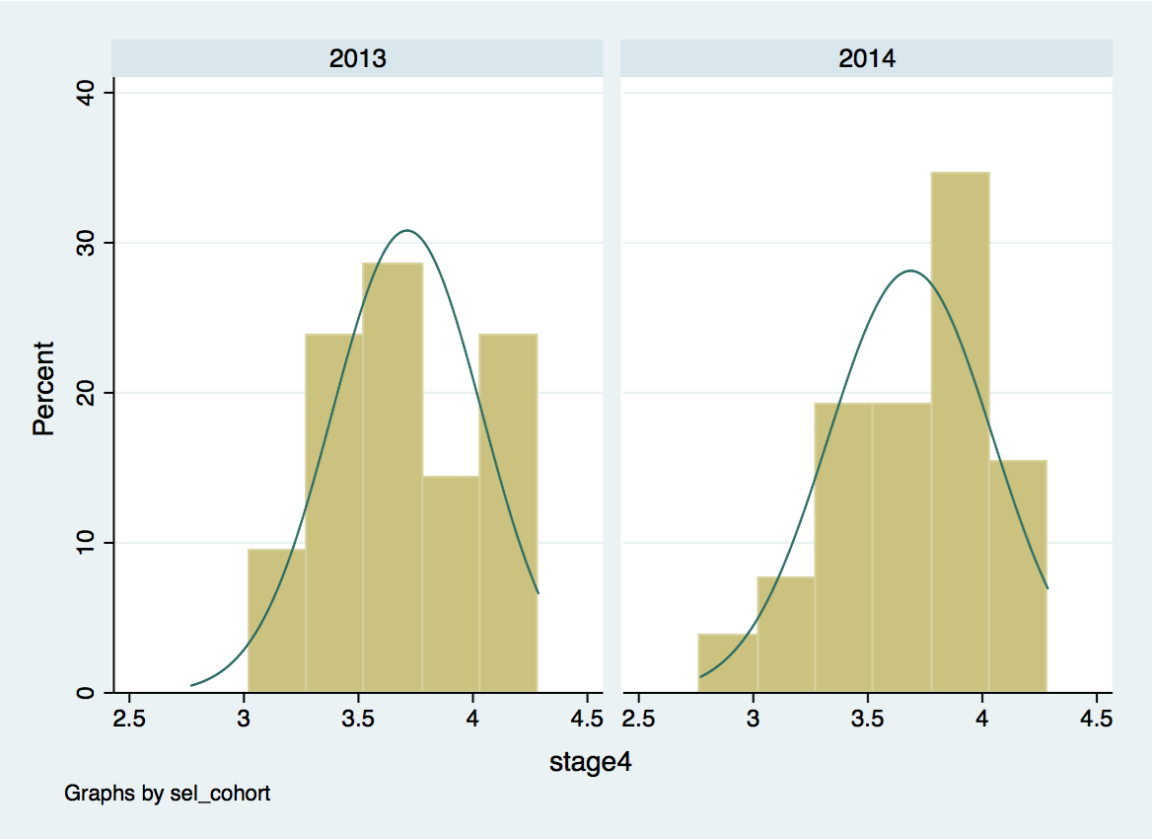
Notes: These histograms display the distribution of Stage 3 scores in the classroom observations, with a superimposed normal density.

Figure A.8: Scores from the Student Surveys at Clinical Practice, 2014



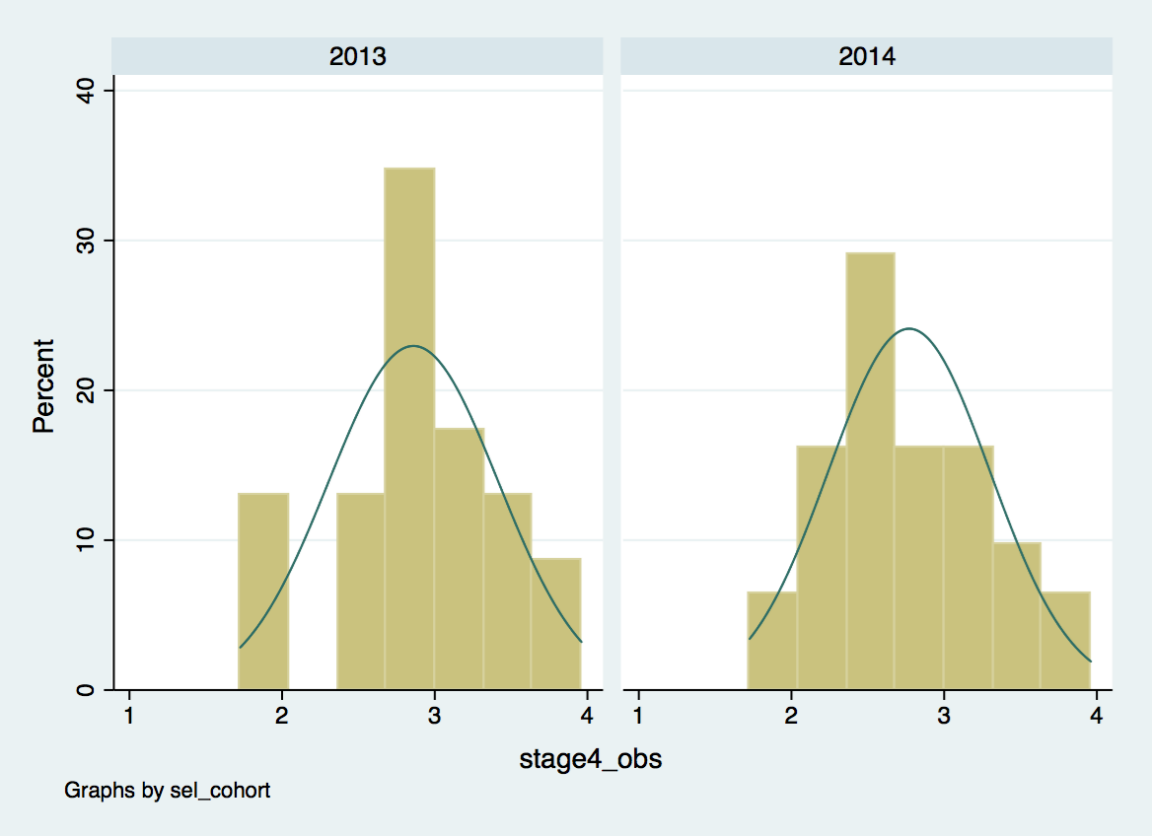
Notes: This histogram displays the distribution of Stage 3 scores in the student surveys, with a superimposed normal density.

Figure A.9: School Year Scores by Cohort, 2013-2014



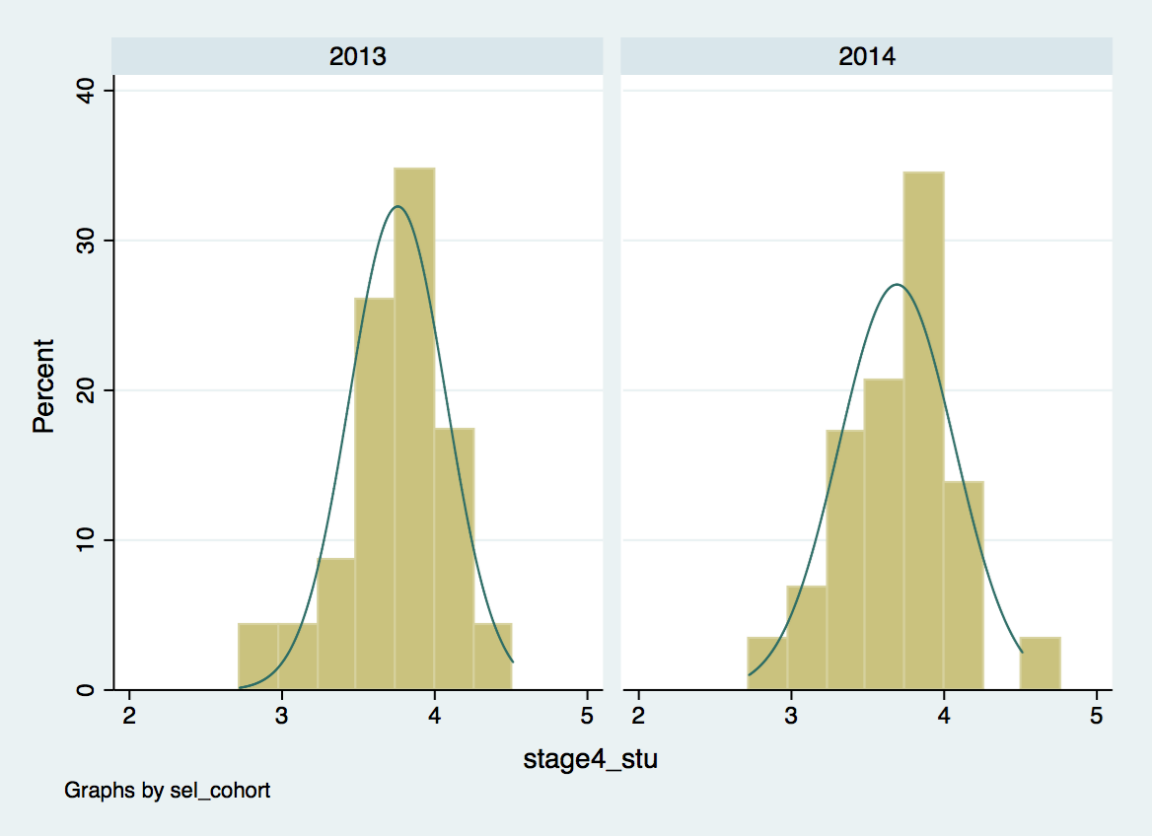
Notes: These histograms displays the distributions of Stage 4 scores by cohort, with superimposed normal densities. The first panel shows the distributions for the 2013 cohort and the second panel shows the distributions for the 2014 cohort. This graph only includes individuals with at least one classroom observation, a student survey, *and* a principal survey during the school year.

Figure A.10: Scores from the Classroom Observations during the School Year by Cohort, 2013-2014



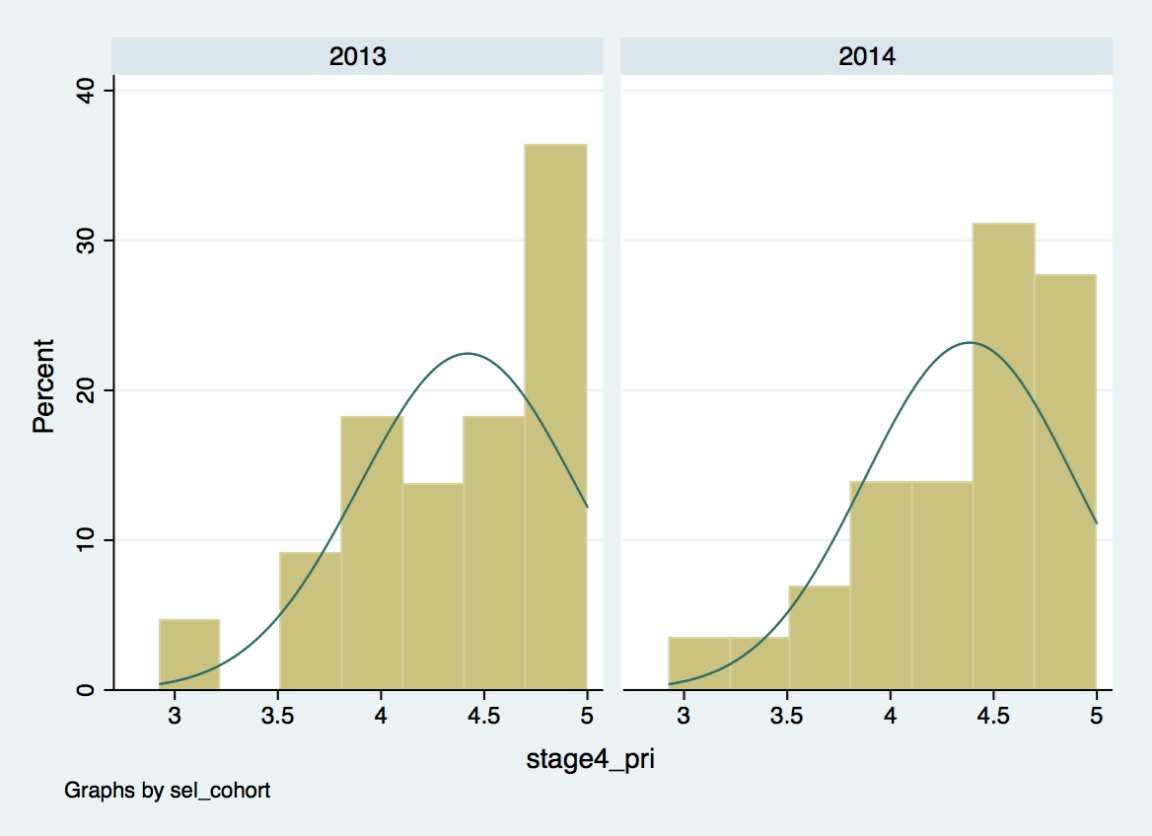
Notes: These histograms displays the distributions of Stage 4 scores in classroom observations by cohort, with superimposed normal densities. The first panel shows the distributions for the 2013 cohort and the second panel shows the distributions for the 2014 cohort. This graph only includes individuals with at least one classroom observation during the school year.

Figure A.11: Scores from the Student Surveys during the School Year by Cohort, 2013-2014



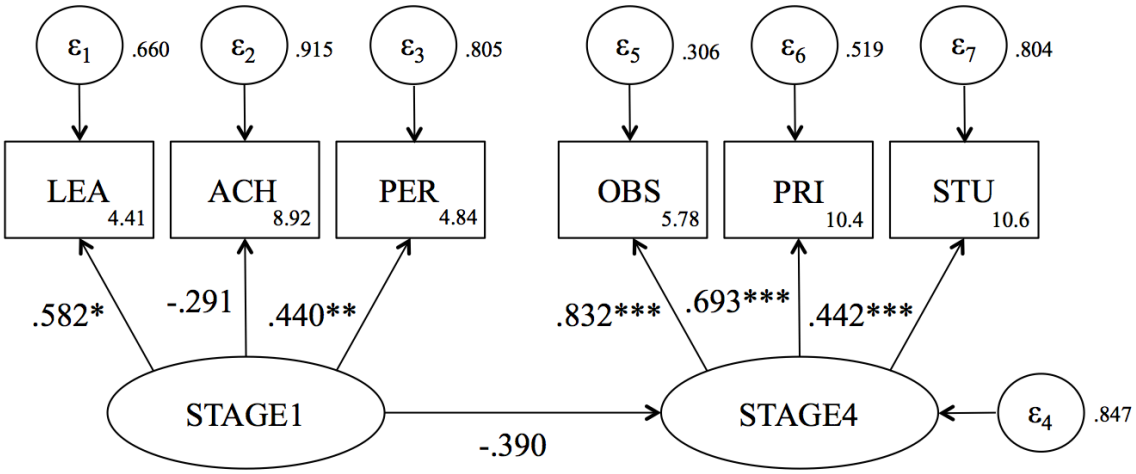
Notes: These histograms displays the distributions of Stage 4 scores in student surveys by cohort, with superimposed normal densities. The first panel shows the distributions for the 2013 cohort and the second panel shows the distributions for the 2014 cohort. This graph only includes individuals with at least a student survey during the school year.

Figure A.12: Scores from the Principal Surveys during the School Year by Cohort, 2013-2014



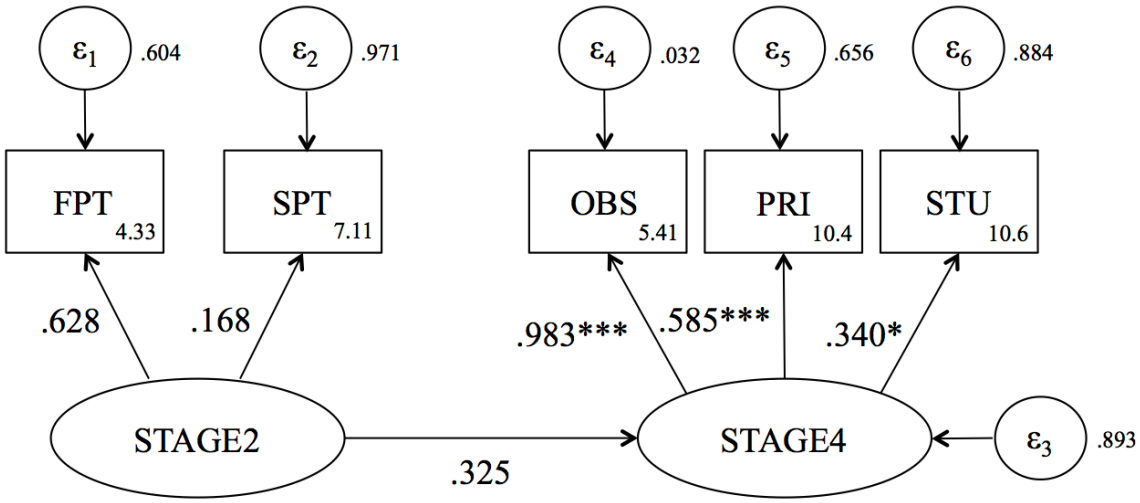
Notes: These histograms displays the distributions of Stage 4 scores in principal surveys by cohort, with superimposed normal densities. The first panel shows the distributions for the 2013 cohort and the second panel shows the distributions for the 2014 cohort. This graph only includes individuals with at least a principal survey during the school year.

Figure A.13: Disattenuated Regression of School Year Scores on Online Application Scores, 2013-2014



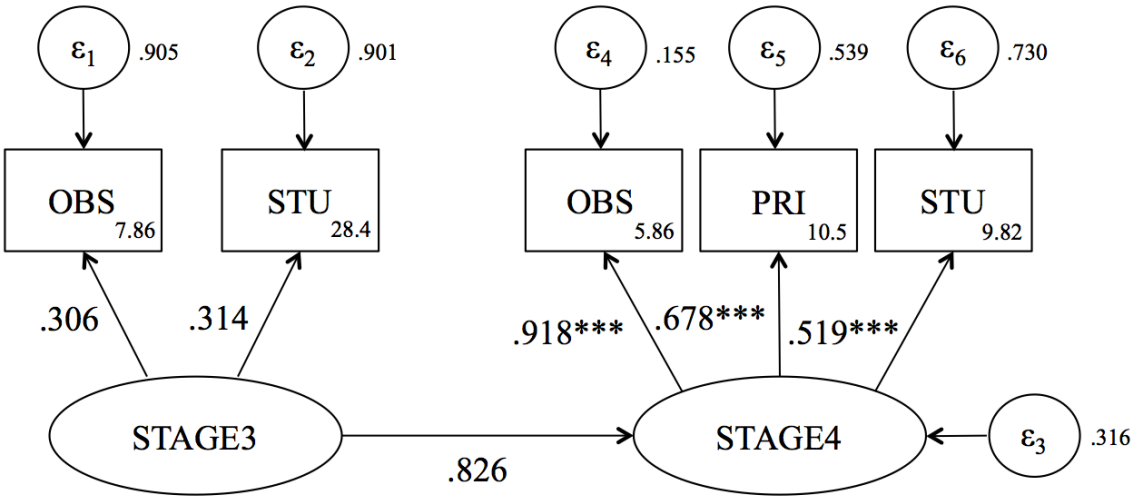
Notes: This path diagram shows the results from a structural equation model that predicts stage 4 scores with stage 1 scores using the scores that make both composites to disattenuate the regression. Stage 1 scores are disattenuated using the scores on the leadership, perseverance, and achievement rubrics administered at that stage. Stage 4 scores are disattenuated using the scores from the student surveys, principal surveys, and classroom observations at that stage. The test statistic for the likelihood ratio test of this model versus the fully saturated model was $\chi^2(8) = 5986$. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Figure A.14: Disattenuated Correlation between the Assessment Center and School Year Scores, 2013-2014



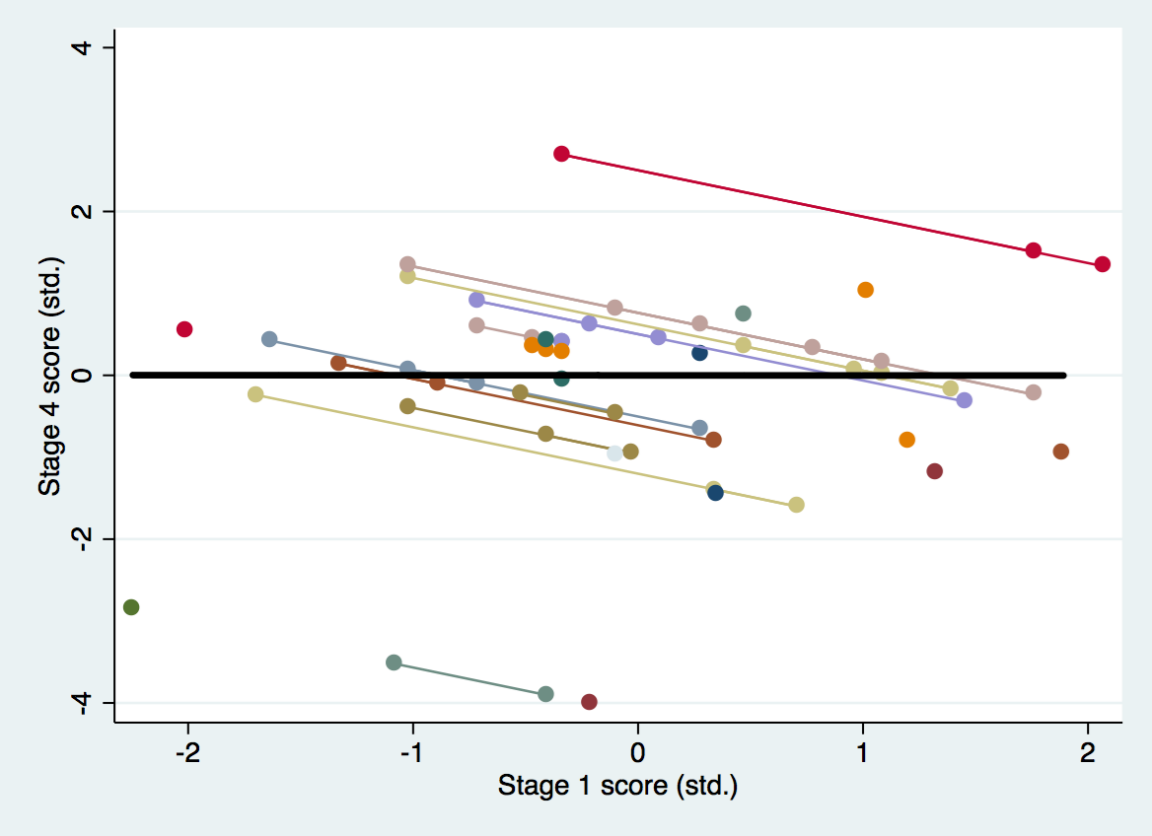
Notes: This path diagram shows the results from a structural equation model that predicts stage 4 scores with stage 2 scores using the scores that make both composites to disattenuate the regression. Stage 2 scores are disattenuated using the scores on the first and second parts of the assessment center. Stage 4 scores are disattenuated using the scores from the student surveys, principal surveys, and classroom observations at that stage. The test statistic for the likelihood ratio test of this model versus the fully saturated model was $\chi^2(4) = 5997$. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Figure A.15: Disattenuated Correlation between the Clinical Practice and School Year Scores, 2013-2014



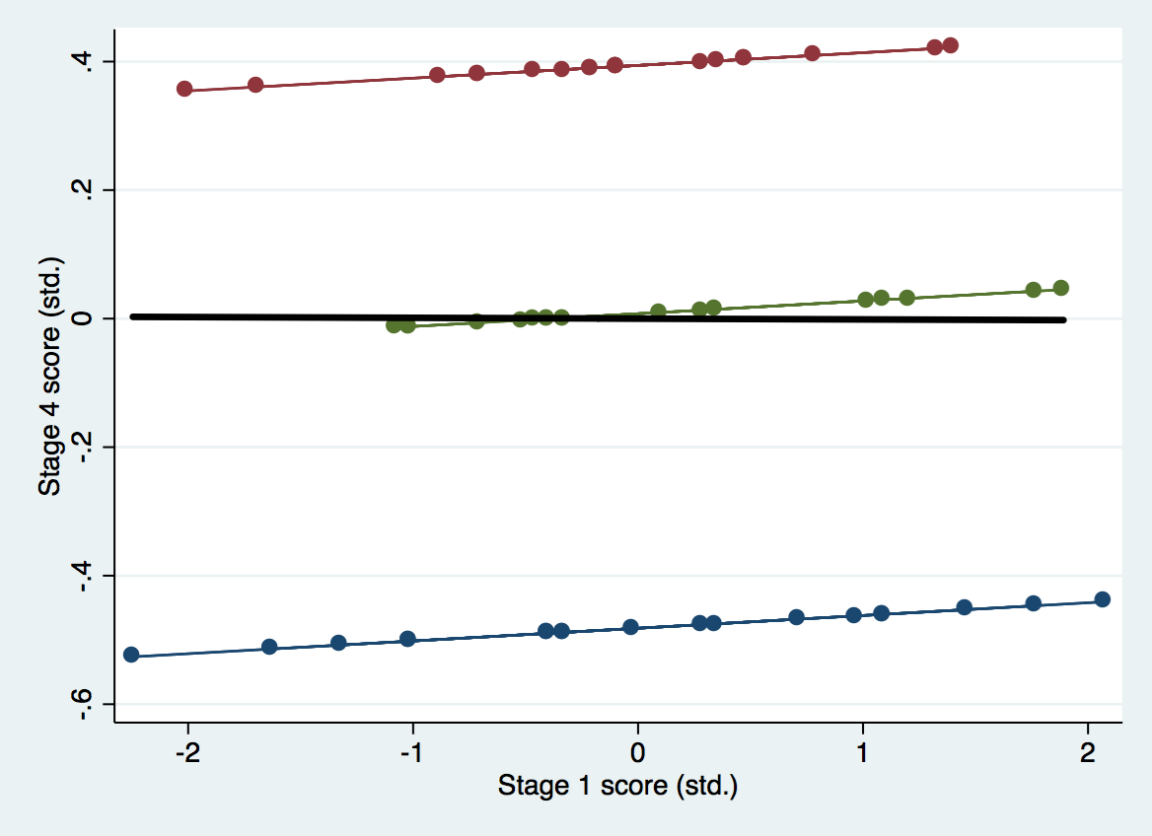
Notes: This path diagram shows the results from a structural equation model that predicts stage 4 scores with stage 2 scores using the scores that make both composites to disattenuate the regression. Stage 2 scores are disattenuated using the scores on the first and second parts of the assessment center. Stage 4 scores are disattenuated using the scores from the student surveys, principal surveys, and classroom observations at that stage. The test statistic for the likelihood ratio test of this model versus the fully saturated model was $\chi^2(4) = 6721$. Stars indicate the levels of statistical significance of each coefficient: * significant at 10% level, ** at 5%, *** at 1%.

Figure A.16: Relationship between Stage 1 and Stage 4 Scores within Stage 1 Raters, 2013-2014



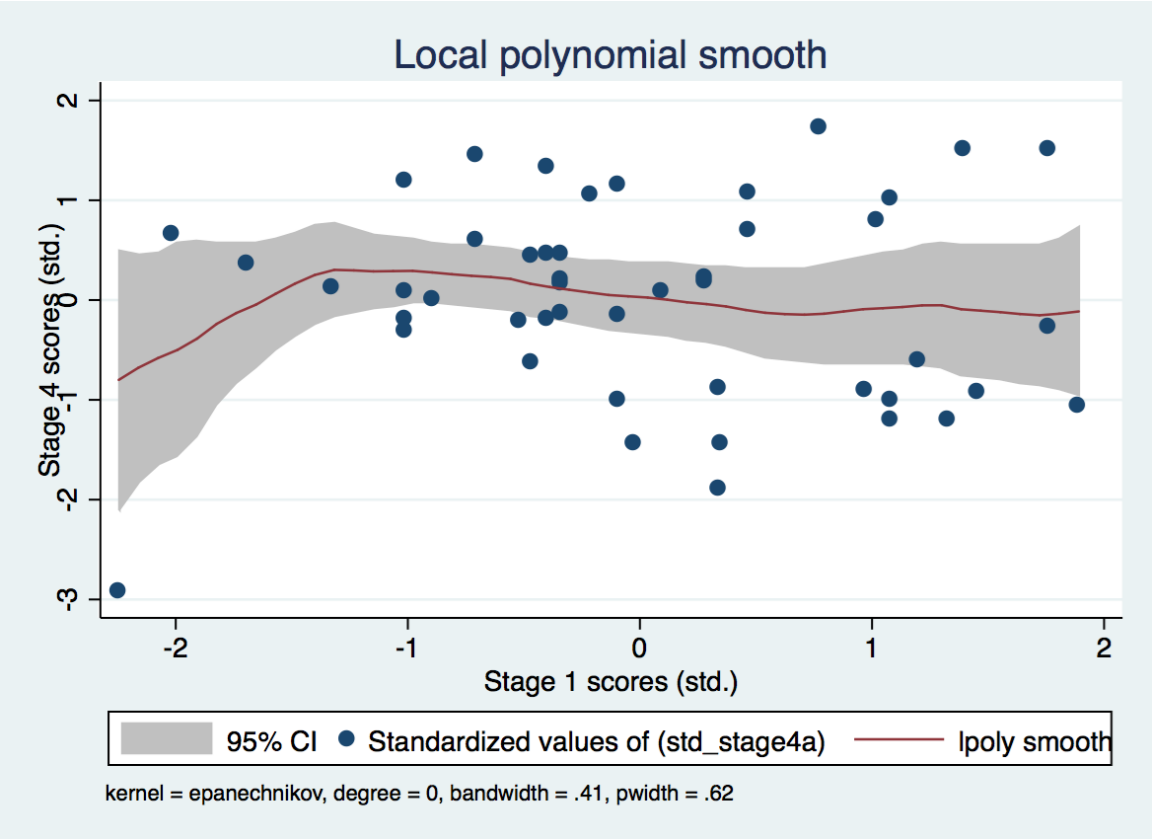
Notes: The thin lines show the relationship between stage 1 and stage 4 scores within stage 1 raters. The thick black line shows the relationship in the sample without any fixed effects.

Figure A.17: Relationship between Stage 1 and Stage 4 Scores within Stage 4 Raters, 2013-2014



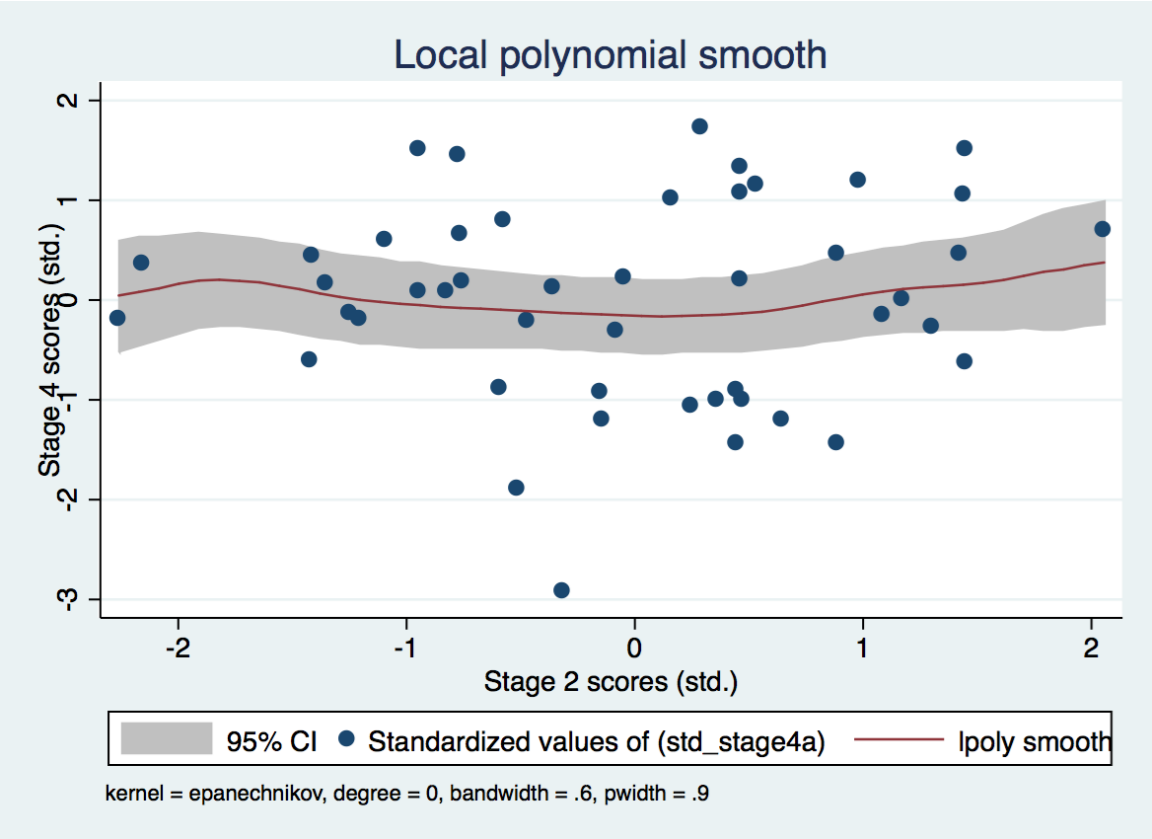
Notes: The thin lines show the relationship between stage 1 and stage 4 scores within stage 1 raters. The thick black line shows the relationship in the sample without any fixed effects.

Figure A.18: Relationship between Stage 1 and Stage 4 Scores, 2013-2014



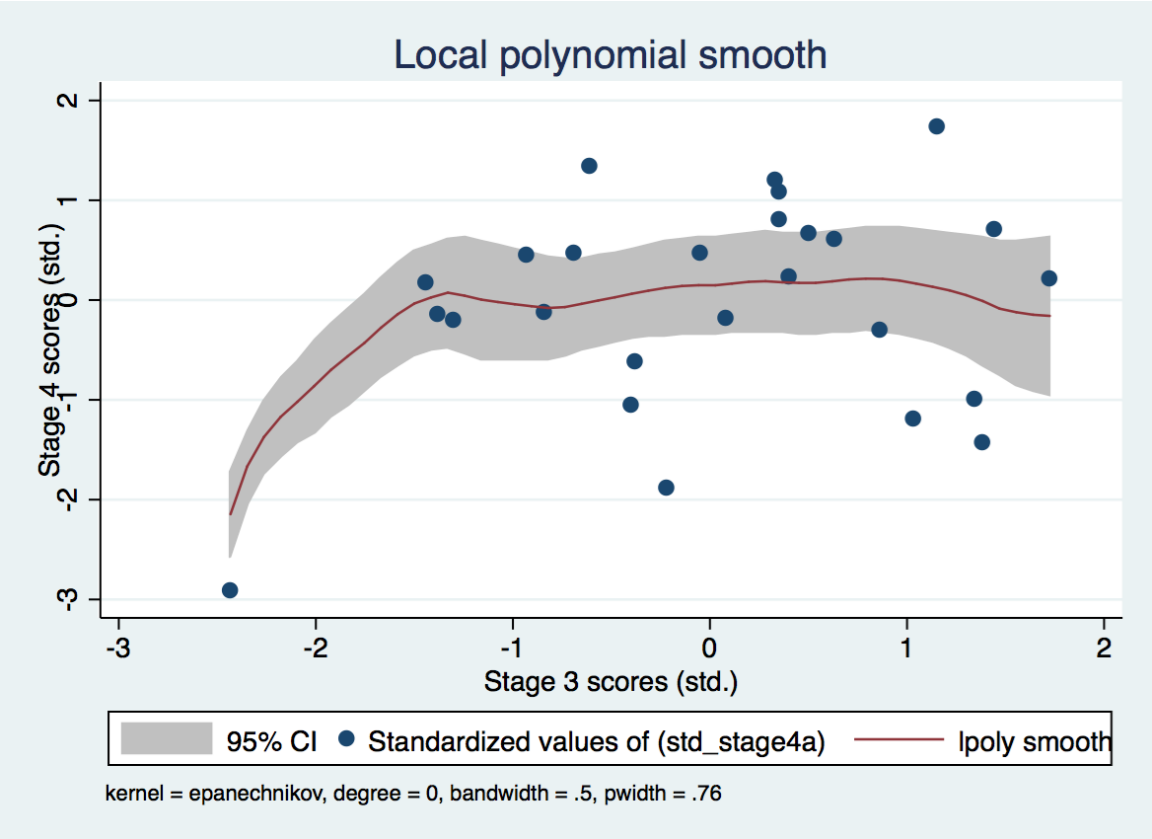
Notes: This graph plots the relationship between the standardized versions of stage 1 and stage 4 scores. The relationship between the two is shown using a kernel-weighted local polynomial smoothing.

Figure A.19: Relationship between Stage 2 and Stage 4 Scores, 2013-2014



Notes: This graph plots the relationship between the standardized versions of stage 2 and stage 4 scores. The relationship between the two is shown using a kernel-weighted local polynomial smoothing.

Figure A.20: Relationship between Stage 3 and Stage 4 Scores, 2014



Notes: This graph plots the relationship between the standardized versions of stage 3 and stage 4 scores. The relationship between the two is shown using a kernel-weighted local polynomial smoothing.

VITA

Alejandro J. Ganimian

2004-2006	Georgetown University Washington, DC	B.S.F.S. May 2006
2006-2007	University of Cambridge Cambridge, UK	M.Phil May 2007
2006-ongoing	Co-founder, *Educar y Crecer* Buenos Aires, Argentina	
2007-2009	Program Associate, Partnership for Educational Revitalization in the Americas Washington, DC	
2009-ongoing	Co-founder, *Enseñá por Argentina* Buenos Aires, Argentina	
2009-2015	Doctoral Candidate, Quantitative Policy Analysis in Education Harvard Graduate School of Education	
2010-2012	Consultant, The World Bank Washington, DC	
2011-2015	Doctoral Fellow, Multidisciplinary Program in Inequality and Social Policy Harvard Kennedy School	
2012	Consultant, Bill & Melinda Gates Foundation Cambridge, MA	
2013-2014	Consultant, Inter-American Development Bank Washington, DC	
2015	Consultant, American Institutes for Research Cambridge, MA	
2015	Consultant, *Grupo de Análisis para el Desarrollo* Cambridge, MA	
2015-ongoing	Postdoctoral Fellow, Abdul Latif Jameel Poverty Action Lab New Delhi, India	