Who Should Teach: Optimizing the Selection of Teacher Candidates at Relay Graduate School

University of Virginia

Kylie Anglin

May 3, 2018

Contents

Chapter 1

Executive Summary

Problem Statement

Decades of research have shown that a school's most valuable resource is its teachers. More than any curriculum, innovative technology, or set of standards, it is teachers who have the greatest impact on a student's future. Given the importance of teacher quality, schools, districts and teacher recruitment organizations would like to be able to predict which teacher candidates are most likely to become effective (or ineffective) teachers. Unfortunately, researchers have been unable to confidently predict which teacher candidates will be successful in the classroom, and employers do not fare much better. I address this prediction problem within the context of Relay Graduate School.

Options

I offer three options regarding Relay's application process. The first option is for Relay to let the current application process continue undisturbed, with the same format, the same rubric weighting system, and the same discretion given to reviewers. The second option is to use the same materials, but rely on a deterministic data-driven approach for acceptance which is wholly dependent on rubric scores. Under this option the rubric weighting system would be altered so that total application scores are better predictive of Gateway scores. The third option is to extend the application process by adding a research-backed and validated teacher pre-screener.

Recommendation

Based on ability to predict effective teaching, costs, and replicability, I recommend that Relay extends their application process by including the Haberman Star Teacher Pre-Screener. I estimate that this option would increase the the mean teacher resident evaluation by .79 percentage points (a small number, but an amount that is likely to have very real consequences for student learning).

Chapter 2

Problem Statement

Decades of research have shown that a school's most valuable resource is its teachers. More than any curriculum, innovative technology, or set of standards, it is teachers who have the greatest impact on a student's future. An effective teacher increases the likelihood that a child attends college, the selectivity of their college, the quality of neighborhood the child will live in as an adult and their future salary (Chetty, Friedman, and Rockoff 2011). A one-standard deviation increase in teacher quality for a single grade will increase a child's cumulative lifetime income by $39,000 on average. No other classroom-level intervention can boast these figures.

Unfortunately, teachers vary widely in their ability to increase student achievement. A one standard deviation increase in teacher performance is associated with a 0.13 standard deviation increase in reading test scores and 0.17 standard deviation in math. These results imply that having a math teacher at the 25th percentile compared to the 75th would increase test scores by nearly a fifth of a standard deviation (Hanushek and Rivkin 2012). In other words, differences even in the middle of the teacher quality distribution have real effects on student learning.

Given the importance of teacher quality, schools, districts and teacher recruitment organizations would like to be able to predict which teacher candidates are most likely to become effective (or ineffective) teachers. If districts and agencies could identify future successful teachers at the hiring stage, they could reduce the negative effects of ineffective teachers and could better place effective teachers where they are most needed. Yet, predicting teacher effectiveness at the hiring stage is not easy.

A large literature base has attempted to identify the characteristics of future effective teachers. While the literature has brought important insights to light it has also exemplified a problem: **researchers have been unable to confidently predict which teacher candidates will be successful in the classroom, and**

employers do not fare much better (Jacob et al. 2016; Rockoff, Jacob, and Kane 2011). I address this prediction problem within the context of Relay Graduate School. Specifically, I offer three options for how Relay's application process may be designed to best identify candidates who are likely to become effective teachers.

Chapter 3

Background

3.1 Introduction to Relay Graduate School

Relay Graduate School is a non-profit graduate school which offers Masters of Arts in Teaching for aspiring teachers. There are currently 15 Relay Graduate School Campuses: Baton Rouge, Chicago, Connecticut, Dallas-Fort Worth, Delaware, Denver, Houston, Memphis, Nashville, New Orleans, New York, Newark, Philadelphia, San Antonio, and Washington D.C. Each campus has a partnership with a number of local schools (many of which are charter schools) where graduate students serve as teachers-in-residence for one year and as full-time lead teachers thereafter. In the first year of the two-year program, graduate students balance course work with a gradual transition into a teaching residency. In the second year, students are employed as fully-licensed teachers, but continue taking courses and complete their master's thesis. The Relay curriculum is delivered 60% in person and 40% online and focuses on four core areas: subject knowledge, classroom culture, developing personal connections with students and families, and lesson planning.

Students enter Relay Graduate School through one of two paths. Before the 2017-18 school year, all residents were identified by local school districts in a recruitment and selection process separate from Relay. Relay has attempted to centralize and improve the recruitment and selection process through the creation of a second entry point - the Teacher Pathways program. In the Teacher Pathways program, a prospective teacher applies directly to Relay Graduate school. Relay then places the student in residency and employment in one of the partner schools. Relay Graduate School has full discretion over who will be admitted to the Teacher Pathways program. Because of this, Relay has a strong commitment to designing an application process that optimally identifies applicants who are likely to become effective

teachers.

3.1.1 Current Application Process

Currently, the Pathways selection process consists of three rounds: an online application form, a video interview, and a one-on-one interview. As you can see in Tables 3.1 and 3.2, applicants are evaluated on a variety of skills and characteristics which are described and weighted in a rubric. While reviewers are advised to take the rubric weighting system into account, there is no cut-score which determines whether a candidate moves to the next stage.

In the online application, the candidate submits a resume, an academic transcript and short essay responses. An application reviewer then evaluates the candidate on a four-point scale with respect to the thirteen weighted criteria. Note that while Relay weights academic preparation more heavily than other qualities, in combination, personality qualities like grit and respect make up the majority of a candidate's application round score. This past year, approximately 70% of candidates were invited to move forward past the online application.

Criteria	Weight	Transcript	Resume	Essay
GPA	3	X		
Content Knowledge & Academic Preparedness	3	X	X	
Critical Thinking	1			X
Commitment to the Mission	2		X	X
Grit	2		X	X
Feedback & Growth Mindset	2			X
Respect and Humility	2			X
Communication	1		X	X
Cultural Responsiveness	1			
Professionalism	2			X
Commitment to Teaching & Service	1		X	X
Initiative and Achievement Orientation	1			
Passion or Excitement (bonus)	1			X

Table 3.1: Online Application Criteria and Weights

Candidates who are not rejected move on to a video interview. The video interview consists of three parts: traditional interview questions, a sample teaching

Criteria	Weight	Interview	Sample Teach	Reflection
Critical Thinking	2	X		X
Cultural Responsiveness	1	X		
Respect and Humility	2	X		X
Grit	2	X	X	
Growth Mindset	3	X		X
Content Knowledge and Academic Preparedness	1		X	
Initiative and Achievement Orientation	1		X	
Bonus: Communication Skills	1			
Bonus: Passion and Excitement	1			

Table 3.2: Video Criteria and Weights

session, and a reflection session where the candidate talks-aloud about their performance on the teaching sample. Table 3.2 displays the weight given to each quality and the portion of the video interview where reviewers are told to look for evidence of the quality. Note that growth mindset is the most heavily weighted component at this stage. Last year, this was the least selective round of the application process - approximately 80% of video interviewees were invited to move forward to the final round.

The final round of the application process is a one-on-one interview which involves traditional interview questions and role play scenarios. The criteria for one-on-one interviews are exactly the same as for the online application, but, in this round, all criteria are weighted equally. Last year, around 60% of applicants who participated in the one-on-one interview were invited to enroll in Relay Graduate school and begin their teaching residency.

Relay's application process stands in contrast to the process of most districts. First, it is much more involved than common practice. Between resumes, essays, interview questions, role plays, and sample teaching sessions, Relay gathers far more extensive information on their candidates than most districts. Many districts do not ask for a writing sample, other than a cover letter (just 25% of Pennsylvanic districts and 60% of New York Districts) and an even smaller proportion use sample lessons. A study of hiring practices across five states found that just 15% of districts required sample lessons from teacher candidates (Liu and Johnson 2006). Further, Relay's emphasis on measures of academic achievement is not common practice - most district's hiring policies do not significantly preference high achieving college graduates (Ballou 1996).

3.2 Introduction to the Data

Relay provided data on the first cohort of Teacher Pathways applicants. This includes both background characteristics like undergraduate major and GPA as well as rubric scores on each of the criteria listed in Tables 3.1 and 3.2. Because data is available on both rejected and accepted candidates, this analysis is partially free from the selection problem of only observing current teachers, a problem which previous studies have faced including Clotfelter, Ladd, and Vigdor (2007) and Goldhaber, Gratz, and Theobald(2017).

Relay also provided data on residents' first and second evaluations. Evaluations were conducted by Relay faculty and are meant to capture residents' performance as classroom teachers in residence. Of course, this data is only available for accepted Pathways applicants. The first evaluation, called Gateway 1, was conducted across the 16 districts in the span of two months, September and October of 2017. Gateway 1 includes seven rubric metrics: attendance, timeliness, preparation, follow-through, engagement, feedback, and mindsets. Gateway 2 was conducted in November and December of 2017 and also includes eight rubric metrics: communication, expectations, appropriate corrective actions, warm/demanding tone, actively engaged students, lesson plan submissions, use of the first few minutes of class time, and professionalism. Generally speaking, Gateway 1 measures items related to professionalism and Gateway 2 measures a resident's ability to instruct and manage behavior. The data also indicates whether the resident is a Pathway's participant or a traditional resident, whether the candidate passed the Gateway, and how many attempts the candidate needed to pass each Gateway.

Finally, Relay provided survey data on their resident teachers. The survey contains very general demographic information (like whether or not the resident is of a historically underrepresented ethnicity) and a number of questions relating to grit and classroom competency. Analysis of survey data can be found in Appendix A.

In this report, I analyze application and evaluation data to better optimize the selection of teacher candidates. Table 3.3 shows the number of observations for each data type broken down by the candidate's entry into Relay Graduate School (i.e. through Pathways or through district identification). As the table indicates, there is currently application and evaluation data on 150 residents.

Scores on a well-designed application process should be a strong predictor of teacher effectiveness. This is the goal of an application process, to predict with limited information which candidates are likely to be successful. To analyze the extent to which this is the case in Relay's application process, I analyze whether rubric scores are significant predictors of the Gateway evaluation performance. Ideally,

Entry	Application	Gateway 1	Gateway 2	Survey
Rejected Pathways Applicant	290	0	0	0
Accepted Pathways Applicant	160	136	150	145
Non-Pathways Resident	0	441	462	481
Total	450	577	612	626

Table 3.3: Sample Size by Data Type

this analysis would use application data to predict student test scores and principal evaluations. At this point in Relay's development, however, no Teacher Pathways residents have become full-time teacher and, therefore, no student test scores of principal evaluations are available. In this report, I treat the Gateway data as a proxy for teacher effectiveness. Though residents are not full-time teachers, the Gateway is similar to the observational evaluations conducted on full-time teachers across many districts.

3.3 Direction of this Report

In this report, I offer three policy options to Relay Graduate School. Each options addresses the predictive validity of Relay's application process in order to identify candidates who are likely to be successful in the classroom. To guide the policy options, Chapter 4 outlines insights gleaned from the literature on which traits are likely to predict effective teaching and table 4.1 presents an informal meta-analysis of the findings in table form. Chapter 5 describes the policy options and the criteria I use to identify the best option. Chapter 6 uses application and evaluation data to evaluate the options and in the final section, I present an outcome matrix which summarizes each policy option according to the criteria listed.

Chapter 4

Literature Review

Teaching quality is defined and measured using both practice-oriented and achievement-oriented evaluations. In a practice-oriented approach, a teacher is evaluated through observation under the assumption that we understand which teacher practices lead to positive outcomes for students. The outcome of a practice-oriented evaluation is usually an aggregated score from a rubric. In an achievement-oriented approach, researchers attempt to directly estimate the effect of teachers on student test scores, most commonly through a value-added measure (VAM). VAMs compare test scores across classrooms, controlling for previous test scores and student characteristics. Highly-effective teachers will have students who score higher than expected given their previous year's scores. The outcome of an achievement-oriented evaluation is usually an estimated increase in student test scores attributed to the teacher, commonly reported in standard deviation units. There is growing evidences of a significant correlation between subjective evaluations of teacher practice and objective measures of teacher performance through VAMs (Harris and Sass 2009), but because both measures likely capture some amount of student learning, the best evaluations are a combination of both approaches.

After determining the outcome of interest, school leaders face another challenge - with limited information, they seek to predict which teaching candidates and potential hires will have high VAM scores or score highly on an observation rubric. At this stage of the process, stakeholders do not necessarily need to know what *causes* a person to become a successful teacher, only whether a certain characteristic makes it more likely that the person will be effective. In other words, it is not necessary that the correlation coefficients between an application traits and evaluation scores provide an unbiased causal estimate. For example, college selectivity is a strong predictor of effective teaching perhaps because selective college institutions prepare

their teaching candidates well - thus causing more effective teaching. Alternatively, attending a selective college institution may simply be correlated with motivation which is the true driver of successful teaching. In this case, employers do not need to tease apart causes and correlations. A reliable correlation, even if confounded, can still be useful for teacher hiring (Kleinberg et al. 2015).

Despite avoiding the challenge of causal inference, researchers have not yet developed a strong link between information available at hire and effective teaching. This is not for lack of trying, but largely due to the issue of small samples and limited data availability, both in terms of measures of teacher quality and informative teacher characteristics (Jacob et al. 2016; Rockoff, Jacob, and Kane 2011). Still, the current literature can guide stakeholders towards the applicant characteristics that should be prioritized in the hiring process. I summarize these findings in the sections below and in Table 4.1.

4.0.1 Certification

Though most districts rely on traditional teacher licensure (through a state-approved teacher education institution) as an indicator of preparedness (NCES 2006), the research suggests more subtle distinctions between licensure routes are necessary. Teachers with standard certifications typically outperform those with emergency licenses (Clotfelter, Ladd, and Vigdor 2007), but some alternative certification programs have been found to be highly effective.

Of all certification routes, including standard certification, status as a Teacher for America corps members has the strongest association with effective teaching (Kane, Rockoff, and Staiger 2008; Decker, Mayer, and Glazerman 2004). Students assigned to Teach for America corps members score around 0.15 standard deviations higher in math and 0.04 standard deviations higher in reading compared to peers assigned to traditionally certified teachers. I interpret the effectiveness of Teach for America recruits as a teacher selection success story. Teach for America provides minimal training but is highly selective, drawing applicants from the most selective universities and offering positions only to a fraction of those who apply (Dobbie 2011). Unfortunately, the success of alternative programs likely depends on a large high-achieving pool of applicants. New York City Teaching Fellows, for example, draws from a more limited pool of candidates and has not seen the same success as Teach for America (Kane, Rockoff, and Staiger 2008).

These findings imply that Relay graduates are not destined to become ineffective or effective teachers by virtue of their alternative certification status. Certification status alone provides little information on a teacher's future effectiveness.

4.0.2 Certification Tests

Certification tests are theoretically promising predictors of effective teaching - they are designed to measure a candidate's knowledge and skills related to pedagogy and/or content. Two recent rigorous studies have tested the predictive validity of certification tests. In the first study, researchers found no evidence that teachers who pass the New York State Teacher Certification Exam on the first attempt are more effective in the classroom (Rockoff, Jacob, and Kane 2011). However, the second study found a small statistically significant relationship between teacher quality and the North Carolina Elementary Education exam scores (Clotfelter, Ladd, and Vigdor 2007). Teachers who scored two standard deviations above average on the exam increased student test scores by 0.07 standard deviations. Differences in findings for the New York and North Carolina exams may be due to varying predictive validity of the tests, sample sizes, or simply the limited variation in the New York dataset (i.e. an indicator of passing rather than continuous scores). Note that each of these studies were only able to evaluate the teaching of candidates who eventually pass the licensure tests and enter the teacher workforce. Certification tests may prevent the least prepared candidates from entering the field, but this function would not be reflected in the data.

4.0.3 Academic Achievement

Teachers who have graduated from highly-selective institutions are more effective on average than those from the least selective colleges (Clotfelter, Ladd, and Vigdor 2007). Other indicators of academic preparedness like SAT scores and GPA are similarly promising predictors (Clotfelter, Ladd, and Vigdor 2007; Jacob et al. 2016). I hypothesize two routes by which academic achievement could be strongly associated with teaching success: undergraduate education may provide needed pedagogical and content knowledge, or prior academic success may be associated with personality traits and characteristics which are important for teaching. Interestingly, intelligence tests which attempt to measure cognitive ability without the usual confounders of academic success (like access to resources and motivation) are not strong predictors of successful teaching (Rockoff, Jacob, and Kane 2011). Further confusing the matter, there is little evidence that graduate education is predictive of better teaching (Rockoff, Jacob, and Kane 2011), and in fact the correlation between graduate degree attainment and effectiveness may even be negative (Clotfelter, Ladd, and Vigdor 2007; Rivkin, Hanushek, and Kain 2005).

Though we don't understand the mechanism by which college selectivity, GPA, and SAT scores are predictive of effective teaching, academic achievement is one of

the strongest predictors available.

4.0.4 Personal Beliefs and Personality Traits

A large body of research establishes a link between teacher personality and teacher behavior, but considerably less links these traits to student success. Interestingly, the most rigorous study to date (in terms of measurement, sample size, and methods) found that personality traits like extraversion and levels of conscientiousness are predictive of observational teacher-evaluations, but not of teacher VAMs - possibly indicating the subjective nature of practice-based evaluations (Rockoff, Jacob, and Kane 2011). Of all the personality traits, a teacher's confidence in their ability to promote student learning, termed self-efficacy, is most predictive of student achievement. This finding is backed by extensive research that establishes a correlation between teacher self-efficacy and teacher behavior (Moulding, Stewart, and Dunmeyer 2014; Gibson and Dembo 1984).

There are two widely-used commercially available instruments designed to measure beliefs and values indicative of future success in the classroom - the Haberman PreScreener and the Gallup TeacherInsight Assessment. Researchers have documented a correlation between the Haberman instrument and teacher effectiveness. Teachers who score one standard deviation higher on the Haberman increase math achievement by 0.022 standard deviations (Rockoff, Jacob, and Kane 2011) and score 0.291 standard deviations higher on practice-oriented evaluations (Jacob et al. 2016). No such correlation has been found with the TeacherInsight assessment (Evans 2016).

4.0.5 Systematic Ratings of and Portfolios

There are two widely-used teacher assessments that are more involved than the previously mentioned exams - the Gallup Teacher Perceiver Interview and the edTPA. Gallup's Interview can be administered by district leaders and has a systematic scoring system. These interview scores are modestly predictive of principal evaluations (Young and Delli 2002), but the edTPA seems to garner greater attention because of its theoretical backing (Darling-Hammond 2012). The edTPA is a portfolio-based assessment that scores teacher candidates on videotaped lessons, lesson plans, and student work samples. Though the edTPA is costly (candidates pay $300 to take the assessment) and time consuming for candidates who often spend months preparing portfolios, it is be a moderate indicator of ineffective teachering - teachers who fail the edTPA tend to be considerably less effective at teaching reading, but the same results don't hold for math (Goldhaber, Cowan, and Theobald 2017).

4.0.6 Holistic Measures of Applicants

While a singular teacher characteristic may not have significant predictive capability on its own, systematically combining multiple factors can provide more information. For example, Rockoff and colleagues combined measures of college selectivity, SAT scores, certification test scores, intelligence test scores and math content knowledge to create a composite measure of cognitive ability which is modestly predictive of effective teaching. Similarly, they combine measures of extraversion, conscientiousness, efficacy and Haberman scores to form a measure of non-cognitive ability which is associated with a small though insignificant increase in math scores (Rockoff, Jacob, and Kane 2011). When Rockoff and colleagues make use of all available data, they are able to explain 12 percent of the expected variance in teacher effectiveness, a small overall number, but substantial in the context of the literature.

Findings in Spokane Public School District are perhaps even more encouraging. A study of Spokane Public Schools hiring process found that a one-standard deviation increase in reviewer assessments of candidates was associated with a 0.064 standard deviation increase in student math scores (Goldhaber, Grout, and Huntington-Klein 2017) The authors note that while the assessment is subjective, it is by no means ad-hoc. Thus, a systematic and thoughtful assessment of teacher candidates can be a key lever towards improving the quality of the teacher workforce.

Table 4.1: Summary of the Literature

Quality	Paper	Context	Subject	Effect	Sample Size
Provisional License					
	Clotfelter, Ladd, and Vigdor 2007	North Carolina	Math	- - -	>1,000,000 Students
	Clotfelter, Ladd, and Vigdor 2007	North Carolina	Reading	- - -	>1,000,000 Students
	Boyd et al. 2006	New York City	Combined	- - -	≈1,000, 000 Students
Teach for America					
	Kane, Rockoff, and Staiger 2008	New York City	Math	+++	>600,000 Students
	Kane, Rockoff, and Staiger 2008	New York City	Reading	X	>600,000 Students
	Decker, Mayer, and Glazerman 2004	Five Cities	Math	+++	≈2000 Teachers
Teaching Fellowships					
	Kane, Rockoff, and Staiger 2008	New York City	Math	X	>600,000 Students
	Kane, Rockoff, and Staiger 2008	New York City	Reading	- - -	>600,000 Students
Certification Test Score					
	Clotfelter, Ladd, and Vigdor 2007	North Carolina	Math	++	>1,000,000 Students
	Clotfelter, Ladd, and Vigdor 2007	North Carolina	Reading	++	>1,000,000 Students
Competitive Undergrad					
	Clotfelter, Ladd, and Vigdor 2007	North Carolina	Math	X	>1,000,000 Students
	Clotfelter, Ladd, and Vigdor 2007	North Carolina	Reading	+	>1,000,000 Students
	Jacob et al. 2016	DC	Combined	+++	≈1500 Teachers
	Aaronson, Barrow, and Sander 2007	Chicago	Math	X	≈600 Students
Degree in Subject Area					
	Betts, Zau, and Rice 2003	San Diego	Reading	X	≈5000 Teachers
	Betts, Zau, and Rice 2003	San Diego	Math	X	≈5000 Teachers
	Aaronson, Barrow, and Sander 2007	Chicago	Math	X	≈600 Teachers
Degree in Education					
	Aaronson, Barrow, and Sander 2007	Chicago	Math	X	≈600 Teachers
GPA					
	Kane, Rockoff, and Staiger 2008	New York City	Reading	X	>600,000 Students
	Kane, Rockoff, and Staiger 2008	New York City	Math	X	>600,000 Students
	Jacob et al. 2016	DC	Combined	+++	≈1500 Teachers
SAT					

	Kane, Rockoff, and Staiger 2008	New York City	Reading	X	>600,000 Students
	Kane, Rockoff, and Staiger 2008	New York City	Math	X	>600,000 Students
	Jacob et al. 2016	DC	Combined	+++	≈1500 teachers
Graduate Degree					
	Clotfelter, Ladd, and Vigdor 2007	North Carolina	Math	- -	>1,000,000 Students
	Clotfelter, Ladd, and Vigdor 2007	North Carolina	Reading	X	>1,000,000 Students
	Rivkin, Hanushek, and Kain 2005	Texas	Math	X	>80,000 Students
	Rivkin, Hanushek, and Kain 2005	Texas	Reading	X	>80,000 Students
	Aaronson, Barrow, and Sander 2007	Chicago	Math	X	≈600 Teachers
	Jacob et al. 2016	DC	Combined	+++	≈1500 teachers
Resume Screening	Goldhaber, Grout, and Huntington-Klein 2017	Washington	Math	++	>15,000 Students
	Goldhaber, Grout, and Huntington-Klein 2017	Washington	Reading	X	>15,000 Students
Application Essay					
	Jacob et al. 2016	DC	Combined	+++	≈1500 teachers
Interview					
	Jacob et al. 2016	DC	Combined	+++	≈1500 teachers
Mock Teaching Lesson					
	Jacob et al. 2016	DC	Combined	++	≈1500 teachers
Haberman					
	Jacob et al. 2016	DC	Combined	+++	≈1500 teachers
	Rockoff, Jacob, and Kane 2011	New York City	Math	+	≈300 Teachers
edTPA					
	Goldhaber, Cowan, and Theobald 2017	Washington	Math	+	≈200 Teachers
	Goldhaber, Cowan, and Theobald 2017	Washington	Reading	X	≈200 Teachers
Math Knowledge for Teaching					
	Rockoff, Jacob, and Kane 2011	New York City	Math	+	≈300 Teachers

The effect column shows the correlation between student test scores and the teacher characteristic. +++ = positive and significant at p=.01, ++ = positive and significant at p=.05, + = positive and significant at p=.1, - - - = negative and significant at p=.01, - - = negative and significant at p=.05, - = negative and significant at p = .1. When a paper presents multiple specifications, I showthe effects from the author's preferred specification or, if not available, the specification with the greatest number of controls. Only estimates where the authors at least control for previous year's test scores are included in this model.

Chapter 5

Evaluative Criteria

Using findings from the literature, three policy options were developed for improving Relay's application process. Of these three options, I base my recommendation on the following criteria:

1. *Ability to Predict Successful Teaching* The central goal of this report is to answer the following question: To what extent can the average skills of Relay residents be improved through changes in the application process? Relay hopes to identify candidates will become effective classroom teachers. To this end, the application process should be predictive of effective teaching as measured by the Gateway - Relay's evaluation of effective teaching.

   I estimate the degree of alignment between the application process and the Gateway using regressions of Gateways scores on application scores. Formally, estimates of predictive validity will be reported using the following regression:

   $$Gateway = \beta_0 + \beta_1 \text{TotalApplicationScore} + \epsilon$$

   where Gateway is a candidate's average score across the two Gateway evaluations and total application score represents a candidate's average assessment score across the three application rounds. The coefficient on Total Application Score can be interpreted as the expected increase in Gateway scores given a one standard deviation increase in application score. If application criteria are well-aligned with the skills teachers employ in the classroom, then application scores and Gateway evaluations scores should rise and fall together.

   The predictive ability of application scores depends not only on the expected magnitude of the coefficient, but also on the reliability and precision of the

estimates. For this reason, I report the mean coefficient on Total Application score across 1000 bootstrapped random samples. These random samples reflect that applicant pools will not always look like the current sample - there will be some random variation and the bootstrapped estimates reflect this.

2. *Replicability* I also consider the extent to which Relay will be able to replicate the application process in upcoming years and through any changes in personnel. If an option depends heavily on characteristics of current reviewers, Relay will face challenges should they decide to scale the Pathways program or if there is significant reviewer turnover. To ensure the sustainability of the application process, reviewers should be able to teach the process to new reviewers in upcoming years.

3. *Cost* I evaluate costs both in terms of one-time expenditures required to implement a new hiring policy and in terms of the ongoing costs required to continue the policy each year. I only consider costs which would be incurred by Relay Graduate school within three years, because these are the costs that are of the most immediate interest to the organization. Total costs of implementation will be estimated by adding start-up costs to three years of ongoing costs. All payroll costs are estimated using the national median salary calculated from job postings on Indeed.com. In this report, administrative assistants are estimated to have salaries of $28,000 ($15 an hour), admissions counselors are estimated to have salaries of $36,000 ($19 an hour), research assistants are estimated to have salaries of $25,000 ($13 an hour), and web developers are estimated to have salaries of $79,000 ($42/hour). No options requires additional hiring, but per hour labor costs represent the opportunity cost of an option. Any time spent changing the application process is time that an employee could have spent meeting other goals. Other than opportunity costs, personnel costs are assumed to be constant across years and options.

    - Start-up Costs (Year 1) Start-up costs concern any initial purchases or administrative efforts which only occur once in three years.
    - Ongoing Costs (Years 1-3) Ongoing costs recur each year the policy is implemented.

Chapter 6

Policy Options

I offer three options regarding Relay's application process. The first option is for Relay to let the current application process continue undisturbed, with the same format, the same rubric weighting system, and the same discretion given to reviewers. The second option is to use the same materials, but rely on a deterministic data-driven approach for acceptance which is wholly dependent on rubric scores. Under this option the rubric weighting system would be altered so that total application scores are better predictive of Gateway scores. The third option is to continue giving reviewers the leeway to accept or reject candidates, but to extend the application process by adding a research-backed and validated teacher pre-screener.

6.1 Option 1: Let the current application process continue undisturbed.

As noted in Section 3.1.1, Relay's Teacher Pathways program currently combines the use of traditional credentials like GPA and undergraduate degree with performance in interviews and in sample lessons. Based on these factors, candidates are evaluated on a rubric within each round (outlined in section 3.1.1). However, because there is no strict application score cutoff, reviewers have the final say on whether or not a candidate moves forward.

6.1.1 Ability to Predict Successful Teaching

Though rubric scores do not determine candidate acceptance or rejection, they can at least partially explain reviewer decisions. Table 6.1 shows the results of a logistic

regression of the rubric subscores which are most predictive of acceptance (chosen through stepwise selection procedures). Together, these subscores explain 55% of variation in acceptance decisions by reviewers. From these results, I calculate the probability that a candidate is accepted and use this to predict Gateway scores. We would hope that a high probability of acceptance is associated with higher Gateway evaluation scores as this would suggest that the screening is serving the intended purpose.

I find that a one standard deviation increase in the probability of acceptance (here, a 40% difference in probability) is associated with a 1.1 percentage point increase in mean Gateway score - less than one rubric point across the two Gateways (see Table 6.2).

However, as an estimate of reviewers ability to predict successful teaching, this number is certainly biased to some degree. As noted earlier, rubric scores explain 55% of variation in probability of acceptance. Other factors must affect the the decision making process. The unexplained variation in probability of acceptance is due to both within and between reviewer variation. Between-reviewer variation - whether because some reviewers are more lenient or because different reviewers implicitly weight rubric items differently - will not bias the estimate so long as reviewers are randomly assigned to applicant cases.

Within-reviewer variation unexplained by rubric scores is more problematic because it implies that reviewers consider criteria which are not faithfully represented in the rubric. If these unobserved qualities are positively correlated with Gateway scores, then the above estimate will underestimate the true correlation between probability of acceptance and the Gateway.

I probe the degree of within-reviewer variation by including reviewer fixed effects in a logistic regression of likelihood of acceptance on rubric subscores. Controlling for assigned reviewer increases the pseudo R^2 to .61, implying that 39% of unexplained variation in probability of acceptance is due to within-reviewer variation. Thus, decisions are likely being made based in part on decisions not represented in the rubric.

The natural next question is whether these unobserved qualities are predictive of Gateway scores. I am able to partially investigate this question using an indicator of reviewer confidence. When a reviewer moves a candidate forward to the next application round, they also note if they are confident in their decision. Controlling for rubric scores, this confidence indicator is significantly correlated with the Gateway, implying that the unobserved criteria used by reviewers is aligned with Gateway evaluations. A one-standard deviation increase in probability of confidence is associated with a .5 percentage point increase in mean Gateway score. If you accept this as the

best estimate of the additional ability of reviewers to predict successful teaching, the total effect size for this option is 1.6 percentage points.

Predictive Validity Estimate: A one-standard deviation increase in probability of acceptance by reviewers is associated with an estimated 1.6 percentage point increase in Gateway scores.

6.1.2 Replicability

Success with this option depends heavily on the current Relay reviewers. Because of the discretion given to reviewers, the current application process relies on reviewers having a good intuition for which candidates are likely to become effective teachers. This intuition is likely very difficult to teach or even to evaluate. Though I do not present results in this report (both because of limited sample sizes and to protect the privacy of reviewers), there is variation in Gateway performance by reviewers conditional on rubric scores, implying that some reviewers have a better intuition for candidates that are likely to be successful than others. Thus, intuition for strong candidates is unlikely a universal skill for reviewers. If there is significant turnover or if Relay's Pathway program expands, it will be difficult to teach this skill to new employees.

Replicability: Low

6.1.3 Cost

In estimating costs, I treat the current option as a baseline - the cost of other options are reported in comparison to current ongoing costs. Though I don't report a monetary amount, the baseline includes items that are common across options like personnel, office space, and professional development.

Cost Estimate: N/A

	Coef./se
Accepted Pathways Applicant	
GPA	.519
	(.33)
Professionalism	−.523*
	(.22)
Grit	−.449
	(.35)
Communication	.223
	(1.27)
$Communication^2$	.028
	(.23)
Commitment to Mission	−2.641*
	(1.28)
$CommitmenttoMission^2$	.877**
	(.28)
Professionalism	−2.899
	(2.79)
$Professionalism^2$	.744
	(.55)
Growth Mindset	.548
	(2.88)
$GrowthMindset^2$	.118
	(.51)
Initiative	−3.069*
	(1.31)
$Initiative^2$	.744*
	(.31)
Grit	.001
	(2.46)
$Grit^2$	.105
	(.44)
Passion	−.709
	(.49)
Communication	.531
	(1.20)
$Communication^2$	.091
	(.25)
Content Knowledge	.603
	(.35)
Q2 Grit	3.378
	(2.09)
$Q2Grit^2$	−.816
	(.43)
Q2 Respect and Humility	−3.416
	(2.22)
$Q2RespectandHumility^2$	.806
	(.43)
Q3 Growth Mindset	−.592
	(1.53)
$Q3GrowthMindset^2$	.170
	(.28)
Constant	.957
	(1.35)
Adj. R^2	.551
No. of cases	250

Signifcance: * for p<.05, ** for p<.01, and *** for p<.001 Horizontal lines separate application phases. Subsets of subscores chosen through stepwise selection.

Table 6.1: Logistic Regression of Acceptance on Subscores

	Coef./se
Pr. Accepted	1.116
	(1.00)
Constant	90.925***
	(.87)
Adj. R^2	.009
No. of cases	140
Mean Bootstrapped Coef.	1.094

Significance: * for p<.05, ** for p<.01, and *** for p<.001

Table 6.2: Regression of Gateway % on Probability of Acceptance

6.2 Option 2: Use a data-driven approach to determine acceptance in each round.

The second option is to rely on round scores to determine which candidates should move forward in each round. This option intends to be make the selection process more data-driven and so it also requires altering Relay's rubric weighting system so that the greatest weight is given to portions of the application process which are best able to predict strong Gateway scores.

6.2.1 Ability to Predict Successful Teaching

Tables 6.3, 6.4, and 6.5 display the ability of rubric subscores to predict evaluation scores. The current rubric weighting system underweights some items that are significantly predictive of Gateway scores and overweights other items which are not significant predictors. For example, in the online application round, communication is a significant predictor of Gateway scores, while commitment to the mission has no relationship with evaluations. Yet, commitment to the mission is weighted twice as much as communication. With this option, subscores will be re-weighted so that final scores in each round are better predictive of Gateway performance.

The canonical method of weighting sub-scores would be to use the regression coefficients as weights since regression models weight variables in terms of importance in predicting the outcome (as long as all variables are on the same scale). However, this approach does not appropriately deal with the problem of correlations between subscores themselves. When all subscores are included in the same regression model,

many coefficients are negative simply because they are negatively correlated with other variables.

The better option would be to reduce the number of variables in the model so that each variable adds new information. To this end, I use principal component analysis to crease new variables (called components) by fitting straight lines to the data points in such a way that minimizes the distance from the data to the line. Here, the straight line is a weighted combination of subscores. The first component is the best fit to the data and each additional component is fit to the errors of the previous model. Thus, each component is uncorrelated with other components, limiting redundant information.

Principal component analysis should uncovers some interesting relationships in the application rounds and makes better use of variance than using the coefficients in a linear regression. See Appendix B for more information of how principal components were estimated and how they may be interpreted.

After created principal components, I estimated the following regression for each application round:

$$GatewayScores = \beta_0 + \beta_1 \text{PCA1} + \beta_2 \text{PCA2} + \beta_3 \text{PCA3} + \epsilon$$

where the coefficients on the principal components become the weight in the following linear combination:

$$Weight = \beta_1 \text{PCA1} + \beta_2 \text{PCA2} + \beta_3 \text{PCA3}$$

This method weights subscores both by their variability (the amount of information they contain) and their ability to predict Gateway scores.

Regression of Gateway scores on Principal components can be viewed in Table 6.6. Note that because the principal components can include both positive and negative loadings, a negative coefficient on a component does not mean that all of the variables are negatively correlated with the Gateway. Still, this method negatively weights some items. In order to not penalize applicants for high scores on any portion of the rubric, I adjusted all negative weights to zero.

Final weights (and the loadings which determined the weights) can be viewed in Tables 6.7, 6.9, and 6.8. In the online application round cultural responsiveness, passion, and professionalism were weighted as zero and so will not be used in the new total online application score. In the video round, cultural responsiveness, grit, and growth mindset were all dropped from the final score. And in the interview round communication and passion were dropped from the final score.

To estimate the ability of this weighting system to predict effective teaching, I calculate candidates' new mean round score. I find a standard deviation increase

	coef	se	t-stat	p-value
Commitment to Mission	0.30	0.69	0.43	0.664
Communication	1.34	0.60	2.23	0.027
Content Knowledge	1.21	0.59	2.04	0.043
Critical Thinking	0.61	0.60	1.02	0.308
Cultural Responsiveness	-0.17	0.63	-0.28	1.217
Grit	-0.05	0.65	-0.07	1.056
Growth Mindset	0.79	0.69	1.16	0.248
Initiative	-0.32	0.37	-0.87	1.612
Passion	1.03	0.33	3.10	0.002
Professionalism	0.86	0.34	2.53	0.013
Respect and Humility	-0.05	0.70	-0.06	1.051
Commitment to Teaching and Service	-0.48	0.62	-0.77	1.559
GPA	0.92	0.59	1.55	0.122

Table 6.3: Summary of Regression Analyses of Combined Gateway on Online Application Subscores

	coef	se	t-stat	p-value
Bonus: Communication Skills	1.23	0.50	2.48	0.014
Sample Teach: Content Knowledge and Academic Preparedness	1.34	0.47	2.84	0.005
Q1 Critical Thinking	1.02	0.48	2.14	0.034
Reflection: Critical Thinking	1.26	0.49	2.56	0.012
Q1 Cultural Responsiveness	0.78	0.50	1.55	0.123
Q2 Grit	0.41	0.55	0.75	0.456
Q3 Grit	0.85	0.47	1.84	0.068
Q2 Growth Mindset	0.67	0.51	1.33	0.184
Q3 Growth Mindset	1.03	0.47	2.22	0.028
Reflection: Growth Mindset	0.96	0.50	1.90	0.060
Sample Teach: Initiative and Achievement Orientation	1.25	0.49	2.57	0.011
Bonus: Passion and Excitement	0.87	0.47	1.85	0.066
Sample Teach: Professionalism	1.12	0.46	2.41	0.017
Q2 Respect and Humility	0.65	0.52	1.26	0.211
Reflection: Respect and Humility	1.04	0.50	2.06	0.041

Table 6.4: Summary of Regression Analyses of Combined Gateway on Video Subscores

	coef	se	t-stat	p-value
Commitment to Mission	0.42	0.48	0.88	0.382
Communication	0.11	0.38	0.29	0.769
Content Knowledge	0.74	0.49	1.51	0.134
Critical Thinking	0.62	0.51	1.23	0.222
Cultural Responsiveness	0.50	0.51	0.98	0.330
Grit	0.43	0.52	0.82	0.414
Growth Mindset	0.56	0.50	1.11	0.269
Initiative	1.02	0.46	2.22	0.028
Passion	0.20	0.38	0.53	0.598
Professionalism	0.84	0.51	1.65	0.101
Respect and Humility	0.29	0.48	0.60	0.549
Commitment to Teaching and Service	0.62	0.46	1.35	0.180
Commitment to Teaching and Service	0.00	0.00	0.00	0.000
Commitment to Teaching and Service	0.00	0.00	0.00	0.000
Commitment to Teaching and Service	0.00	0.00	0.00	0.000

Table 6.5: Summary of Regression Analyses of Combined Gateway on 1:1 Subscores

in the new total application score (i.e. the sum of the online, video, and in-person round scores) is associated with 1.9 percentage point increase in Gateway scores. Table 6.10 presents these results.

Predictive Validity Estimate: A one-standard deviation increase in combined round scores is associated with an estimated 1.9 percentage point increase in Gateway scores.

6.2.2 Replicability

This option is highly replicable as long as rubric scores carry the same meaning across reviewers. In other words, for any given candidate there should be reliable agreement across reviewers on what the candidates round score should be. Currently, there is more variation in round scores by reviewers than we would expect by chance, implying that some reviewers systematically rate candidates higher than others. Under the current application process, it is acceptable that rubric carry different meanings across reviewers because reviewers are using their own scoring system to make decisions. If scores are deterministic, though, there must be between-reviewer reliability. This is certainly possible to develop but would require additional training (which

	Online Application Coef./se	Video Coef./se	Interview Coef./se
Scores for component 1	.036		
	(.09)		
Scores for component 2	.197		
	(.14)		
Scores for component 3	–.103		
	(.13)		
Scores for component 1		.131	
		(.07)	
Scores for component 2		–.369**	
		(.14)	
Scores for component 3		.270	
		(.17)	
Scores for component 1			.152*
			(.07)
Scores for component 2			–.104
			(.13)
Scores for component 3			–.108
			(.15)
Constant	30.295***	30.362***	30.277***
	(.18)	(.17)	(.19)
Adj. R^2	.021	.089	.041
No. of cases	150	136	129

Signifcance: * for p<.05, ** for p<.01, and *** for p<.001

Table 6.6: Regression of Gateway Scores on Principal Components

Subscore	Comp1	Comp2	Comp3	New Weight
Commitment to the Mission	0.3285	-0.1211	-0.2055	0.90
Communication	0.3328	0.0956	-0.0019	3.09
Content Knowledge and Academic Preparedness	0.1504	0.6341	0.0702	12.28
Critical Thinking	0.3145	0.0725	-0.1699	4.29
Cultural Responsiveness	0.3178	-0.1277	-0.0174	0.00
Grit	0.2905	-0.1136	-0.1375	0.22
Feedback and Growth Mindset	0.3382	-0.0481	-0.1072	1.36
Initiative and Achievement Orientation	0.1835	0.0834	0.0898	1.37
Passion or Excitement (bonus)	0.2702	-0.1243	0.433	0.00
Professionalism	0.1828	0.0408	0.8054	0.00
Respect and Humility	0.3473	-0.0912	-0.0853	0.32
Commitment to Teaching	0.3048	-0.1109	-0.1573	0.52
GPA	0.0796	0.7019	-0.1306	15.42

Table 6.7: Online Application PCA Loadings and New Weight

will be calculated in the costs section).

Replicability: High.

6.2.3 Cost

There are two primary costs to implementing this options. First, reviewers will need to spend time ensuring that their scores are in concordance with one another. For the sake of analysis, I assume interrater reliability training will take about two days. For example, the MyTeaching Interrater Reliability Certification training for teachers takes approximately 14 hours in total. The MyTeaching certification trains teachers to evaluate student portfolios and is meant to teach evaluators to measure students skills, knowledge, and attitudes. We might expect that a similar analysis of teacher candidates' skills, knowledge, and attitudes would take the same amount of time. The second significant cost to this option is retraining reviewers on scoring candidates under the new weighting system.

Subscore	Comp1	Comp2	Comp3	New Weight
Bonus: Communication Skills	0.2779	-0.0962	0.2137	12.97
Sample Teach: Content Knowledge	0.2556	-0.3393	-0.0139	15.51
Q1 Critical Thinking	0.2511	0.0508	0.484	14.49
Reflection: Critical Thinking	0.2878	-0.2777	-0.196	8.74
Q1 Cultural Responsiveness	0.2533	0.1647	0.4547	9.52
Q2 Grit	0.2335	0.3807	-0.221	0.00
Q3 Grit	0.2233	0.2577	0.1426	0.00
Q2 Growth Mindset	0.2541	0.3278	-0.3172	0.00
Q3 Growth Mindset	0.2453	0.2761	0.1394	0.00
Reflection: Growth Mindset	0.2791	-0.1432	-0.3185	0.35
Sample Teach: Initiative and Achievement	0.2345	-0.335	0.074	17.45
Bonus: Passion and Excitement	0.2836	-0.0335	0.148	8.96
Sample Teach: Professionalism	0.2669	-0.2642	0.015	13.66
Q2 Respect and Humility	0.2434	0.3696	-0.2517	0.00
Reflection: Respect and Humility	0.2727	-0.1831	-0.3051	2.10

Table 6.8: Video PCA Loadings

Subscore	Comp1	Comp2	Comp3	New Weight
Commitment to the Mission	0.3105	-0.0899	0.0519	5.10
Communication	0.2611	0.593	0.1155	0.00
Content Knowledge and Academic Preparedness	0.2723	0.0317	-0.5473	9.71
Critical Thinking	0.3388	-0.0948	-0.3424	9.83
Cultural Responsiveness	0.2499	-0.2479	0.3687	2.41
Grit	0.2456	-0.2397	0.3274	2.70
Feedback and Growth Mindset	0.2998	-0.1018	0.3273	2.09
Initiative and Achievement Orientation	0.3085	0.0824	-0.3327	7.41
Passion or Excitement (bonus)	0.2596	0.5976	0.1878	0.00
Professionalism	0.2412	-0.3537	-0.1042	8.48
Respect and Humility	0.3357	-0.0497	0.2061	3.39
Commitment to Teaching	0.3168	-0.0969	-0.1327	7.26

Table 6.9: Interview PCA Loadings and New Weight

	Coef./se
Standardized Application Score	1.821***
	(.50)
Constant	91.484***
	(.54)
Adj. R^2	.087
No. of cases	140
Mean Bootstrapped Coef.	1.875

Significance: * for p<.05, ** for p<.01, and *** for p<.001

Table 6.10: Regression of Gateway % on Weighted Application Score

Item	Description	Cost
Start-up		
Create new process for calculating rubric scores	20 hours, $15/hr, 1 administrative assistant	$300
Initial training on new process	5 hours, $19/hr, 44 reviewers	$4180
	Start-up Total	$4480
Ongoing		
Syncing sessions for inter-rater reliability	5 hours, $19/hr, 44 reviewers	$11,704
Updating the rubric as new cohort data is gathered	50 hours, $13/hr, 1 research assistant	$650
	Ongoing Yearly Total	$12,354
	Total Cost	$41,542

Table 6.11: Option 2 Costs

Cost Estimate: $41,542

6.3 Option 3: Include the Haberman Star Prescreener in the Online Application Round

The third option is to replace current measures of personality traits in the online application round with the Haberman Star Teacher Prescreener. The Haberman Prescreener is a 50-item multiple choice assessment which is intended to evaluate

the knowledge and beliefs needed to teach lower-income students. The Prescreener was developed by interviewing highly effective teachers and questions are designed to measure those traits in teacher candidates.

The Haberman PreScreener is designed to measure many of the same characteristics which are currently included in the application round, for example ability to work with at-risk students and the ability to persist through challenges in the classroom. Rubric components which would be dropped and replaced with Haberman scores are commitment to the mission, cultural responsiveness, grit, growth mindset, initiative, passion, respect and humility, and commitment to teaching and service.

Under this option, the video and interview rounds will remain unchanged, but the online application round of decisions will be made using a score cutoff.

6.3.1 Ability to Predict Successful Teaching

A study of schools in Washington D.C. found that a one-standard deviation increase in an applicant's score on the Haberman was associated with a .16 standard deviation increase in subjective principal evaluations (Jacob et al. 2016).

For reference, one Gateway standard deviation is equal to 6.16 percentage points. If we assume that Haberman scores would have a similar relationship with Gateway evaluations, then a one-standard deviation increase in Haberman scores is likely to be associated with a .98 percentage point increase in the Gateway.

This option does not only rely on the Haberman to predict successful teach. An estimate of predictive validity should also include the predictive power of GPA, content knowledge, critical thinking, communication and professionalism subscores in the online application round and the probability of acceptance in the video and final interview round. (Remember, this option combines subjective reviewer decisions in the video and interview round with a data-driven approach in the first round.) A single standard deviation increase in online-application round scores without personality traits (including only communication, critical thinking, content and academic preparation, GPA, and professionalism) is associated with a .39 percentage point increase in the Gateway evaluation. Together, a one standard deviation in online application scores and probability of acceptance in the video and interview round is associated with a 1.9 percentage point increase in the Gateway.

An upperbound estimate of the predictive validity of this option would be to combine the 1.9 percentage point increase in the Gateway with the expected increase due to the Haberman. However, an estimate of the impact on student learning of including Haberman scores depends on the amount of *new* information contained in the Haberman. The same DC study discussed earlier presented pairwise correlations

of Haberman scores with other application components in DC. They found a .21 correlation with SAT scores, .2 correlation with GPA, .19 correlation with college selectivity, .12 correlation with interview performance, and .01 correlation with mock teaching performance. These correlations suggest that there is indeed new information to be found in Haberman scores. If we conservatively assume that 50% of variation in the Haberman can be explained by other application components, then adding the Haberman would result in a .49 increase in predicative ability, bring the estimated impact of the option to 2.39 percentage points.

Predictive Validity Estimate: A one-standard deviation increase in application round scores (including the Haberman measures) and a one-standard deviation increase in probability of acceptance by reviewers in the video and one-on-one round are together associated with an estimated 2.39 percentage point increase in Gateway scores.

6.3.2 Replicability

This option combines objective and subjective measures to determine candidate selection. The data-driven application round is highly replicable, but video and interview rounds rely on reviewer's intuition for which candidates are likely to become strong teachers. As discussed in section 6.1, this is a challenge for replicability. Combining this challenge with the high replicability of the application round, I conclude this option is moderately replicable.

Replicability: Moderate.

6.3.3 Cost

There are two primary costs of incorporating the Haberman into the application process. First, the test costs $20 per applicant. Second, the labor and time spent incorporating the Haberman into the online application process will also be a significant cost. Note that while substituting the Haberman for parts of the online application rubric may save some time in the application process, resumes and essays will still need to be read so I expect the time saved would be negligible.

Item	Description	Cost
Start-up		
Integrate Haberman test into online application website	10 hours, $42/hr, 1 web developer	$420
Create new process for translating Haberman scores into rubric scores	10 hours, $19/hr, team of 3 lead reviewers	$570
Initial training on new process	5 hours, $19/hr, 44 reviewers	$4180
	Start-up Total	$5170
Ongoing		
Haberman Assessments	$20 per test, 450 applicants	$9,000
	Ongoing Yearly Total	$9,000
	Total Cost	$32,170

Table 6.12: Option 3 Costs

Cost Estimate: $32,170

Chapter 7

Recommendation

Table 7.1 presents relative strengths and weaknesses of each option. Option 1 has the lowest cost, Option 2 has the highest replicability, and Option 3 has the highest predictive ability. Given Option 3's ability to predict effective teaching, **I recommend that Relay expands their application process to include the Haberman Star Prescreener**. A .79 percentage point difference in the Gateway may have very real consequences for student learning.

Using a back of the envelope calculation, I estimate that this increase would translate into a .03 standard deviation increase in student test scores and $360 per student per teacher in increased lifetime earnings. Harris and Sass estimate that a one-standard deviation increase in principal ratings translates to .2 student test score standard deviations (Harris and Sass 2009). If we assume that Gateway evaluations perform similarly to principal evaluations, we can use this same translation to better address this impact on student learning. I use an average estimate of increased student lifetime earnings based on two of the prominently cited studies on this topic (Hanushek and Rivkin 2012; Chetty, Friedman, and Rockoff 2011). My best guess estimate using this average is a discounted $12,000 of student lifetime earnings per one standard deviation of teacher improvement. For a classroom of 20 students and an estimated 150 teacher candidates, $360 in lifetime earnings per student per teacher per year far outweighs the costs.

7.1 Implementation

Should Relay choose to implement Option 3, I recommend the following regarding implementation:

- The website should be redesigned so that the Haberman becomes a seamless part of the application process.
- Online application round reviewers should be blind to Haberman results when scoring applicants on other portions of the application. This will ensure that Haberman results do not bias reviewer's evaluations of other criteria.
- Because my best estimate of the predictive validity of the Haberman and other application application criteria suggests a .49 percentage point increase in Gateway scores per Haberman standard deviation and a .39 percentage point increase per application standard deviation, I suggest that roughly half of an applicant's online application score should come from their rubric ratings and the other half from the quartile ranking of Haberman test takers. The new score could allot 20 points to communication, critical thinking, content and academic preparation, GPA, and professionalism, and 5 points per quartile ranking on the Haberman.
- The first year of implementation, it will be difficult to empirically choose an appropriate cut score for the online application round. This year, the average applicant who moved past the first round scored 14 out 20 points on communication, critical thinking, content and academic preparation, GPA, and professionalism. I recommend that Relay plan to accept online application round applicants who score above 24 out of 40 (averaging 3.5 per rubric subscore and 50th percentile on the Haberman). However, this cutoff should be adjusted if Relay finds that either too many or too few applicants are meeting this criteria.
- The rest of the application process should continue as usual to minimize the cost of unneeded changes.

Option	Ability to Predict Successful Teaching (in Haberman Percentage Points)	Replicability	Cost for Three Years of Implementation
Option 1: Let the current application process continue undisturbed.	N/A	Low	N/A
Option 2: Use a data-driven approach to determine acceptance in each round.	+.3 percentage points	High	$44,050
Option 3: Include the Haberman Star Prescreener in the Online Application Round.	+.79 percentage points	Moderate	$34,678

Table 7.1: Outcomes Matrix

Appendix A

Survey Analysis

Relay provided survey data on 626 teacher residents - 145 who entered through the Pathways program. The survey asked eight questions related to the concept of grit and twelve questions related to classroom competency. (See tables A.6 and A.7 for a list of questions classified as grit-related and questions classified as classroom competency.) At 92%, The response rate was quite high. Just 52 residents are missing from the survey data, of which, 15 are Pathways entrants and 10 are partner recruits.

A.1 Demographics

Just over 70% of surveyed residents are classified as historically under-represented graduate students (HUGS), a term that includes all ethnicities outside of white, non-Hispanic residents and Asian residents. These residents are represented in both Pathways entry points and traditional entry points, though a greater proportion of Pathways residents are HUGS, see Table A.1.

Currently, 70 residents are at-risk of dismissal due to financial obligations to Relay. Of those at risk, 17 are Pathways residents and 53 come from other entry points. So, while the number of at-risk Pathways residents is less than traditional residents, Pathways residents are more likely to be at risk due to financial obligations. See Tables A.2 and A.3.

Under-represented	Non-Pathways		Teacher Pathways Pathways		Total	
	Freq	col %	Freq	col %	Freq	col %
No	114.0	28.4	27.0	20.6	141.0	26.5
Yes	288.0	71.6	104.0	79.4	392.0	73.5
Total	402.0	100.0	131.0	100.0	533.0	100.0

Table A.1: Under-represented Ethnicity by Pathway

At Risk 12.20	Non-Pathways		Teacher Pathways Pathways		Total	
	Freq	col %	Freq	col %	Freq	col %
No	396.0	88.2	101.0	85.6	497.0	87.7
Yes	53.0	11.8	17.0	14.4	70.0	12.3
Total	449.0	100.0	118.0	100.0	567.0	100.0

Table A.2: At Risk Due to Financial Obligations by Pathway

At Risk 12.20	No		Under-represented Yes		Total	
	Freq	col %	Freq	col %	Freq	col %
No	114.0	91.9	291.0	87.9	405.0	89.0
Yes	10.0	8.1	40.0	12.1	50.0	11.0
Total	124.0	100.0	331.0	100.0	455.0	100.0

Table A.3: At Risk Due to Financial Obligations by Ethnicity

A.2 Intentions to Teach

Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
1%	5%	7%	18%	30%	40%

Table A.4: As of this moment, I intend to be a classroom teacher 5 years from now.

The majority of residents either agree or strongly agree that they intend to be a teacher five years from now (see Table A.4).

There is also a strong positive relationship between a resident's intention to teach long-term and their mean response to grit-related questions (see Table A.5). This fits with the intuition that teaching requires an ability to bounce back from setbacks. There is also a significant positive relationship between a resident's intention to teach and their mean response to classroom competency questions.

Interestingly, there is no significant relationship between a resident's intentions to teach and their Gateway scores. This could imply that *confidence* in classroom competency is more important for retention than competence as measured by evaluations. Note, though, that we should be cautious in generalizing from a resident's current intentions to teach and their future retention. It is also important to note that while these correlations are statistically significant, they are not large in magnitude.

	Mean Grittiness Coef./se	Classroom Competency Coef./se	Gateway 1 Coef./se	Gateway 2 Coef./se
Intend to Teach	.085***	.166***	–.019	–.104
	(.02)	(.04)	(.07)	(.13)
cons	3.238***	6.196***	29.889***	31.878***
	(.10)	(.20)	(.35)	(.64)
Adj. R^2	.034	.035	.000	.004
No. of cases	481	478	463	191

Significance: * for p<.05, ** for p<.01, and *** for p<.001. Grit questions have been rescaled so that a higher score implies more grittiness.

Table A.5: Regression Survey Responses and Gateway Scores on Intention to Teach

A.3 Grit

Residents responded to grit-related questions on a 5-point scale ranging from "Very much like me" to "Not at all like me", but responses have been recoded so that higher scores represent more gritty answers.

Means and standard deviations for grit-related questions may be found in Table A.6. Residents strongly believe that they are hard workers and diligent. However, they sometimes struggle to maintain focus and to remain encouraged in the face of setbacks. The modal response to these questions was "Somewhat like me".

	. mean	sd
New ideas and projects sometimes district me from previous ones.	3.17	0.97
I have been obsessed with an idea for a short time but lost interest.	3.37	1.02
I often set a goal but later choose to pursue a different one.	3.48	0.98
I have difficulty focusing on projects that take more than a few months.	3.43	1.08
Setbacks don't discourage me.	3.27	1.08
I am a hard worker.	4.54	0.68
I am diligent.	4.31	0.74

Table A.6: Grit-Related Survey Questions

A.4 Classroom Competency

Residents responded to classroom competency questions on a 9-point scale ranging from "None at all" to "A great deal". Questions generally took the form of "How much can you do to..." where the subject may be calming a disruptive student or providing an alternate explanation.

Means and standard deviations for grit-related questions may be found in Table A.7. Generally speaking, residents are confident in their classroom abilities. The modal response on all questions was "Quite a bit". Drilling down a bit more, residents seem to be more confident in their ability to motivate and manage the behavior of students than they are in their instruction practices and in assisting families.

	mean	sd
...prevent and respond to disruptive behavior in the classroom?	7.18	1.44
...motivate students who show low interest in school work?	7.10	1.45
...calm a student who is disruptive or noisy?	6.93	1.42
...help your students value learning?	7.36	1.43
...craft good questions for your students?	6.82	1.50
...get children to follow classroom rules?	7.16	1.36
...motivate students who show low interest in school work?	7.44	1.31
...establish a classroom management system with each group of students?	7.03	1.52
...use a variety of assessment strategies?	6.48	1.61
...provide an alternative explanation when students are confused?	7.05	1.38
...assist families in helping their children do well in school?	6.66	1.59
...implement alternative strategies in your classroom?	6.62	1.50

Table A.7: Classroom Competency Survey Questions: How much can you do to...

A.5 Principal Component Analysis

Measurements on survey questions are highly correlated with one-another, so looking at each question separately provides redundant information. For example, a resident who scores themselves highly on their ability to establish a classroom management system is also likely to score themselves highly on their ability to respond to disruptive behavior. This covariance may be because these questions capture a latent concept, like confidence, which explains responses to many questions.

To simplify information and investigate covariance, I use principal component analysis on classroom competence and grit questions.

For classroom competency questions, principal component analysis showed that two components explain over 66% of the variation in classroom competency questions, which each of the other components explain less than 6% of the remaining variation. Table A.8 shows the loadings (or weights) that go into each component. A

The first component has positive loadings of roughly equal size for all variables. We can think this component as overall confidence in teaching ability. The second component has positive loadings on instruction questions and negative loadings on questions relating to behavior management. This component differentiates between whether a resident views themselves as better at instruction or behavior management.

Question	Comp 1	Comp 2
...prevent and respond to disruptive behavior in the classroom?	0.2894	-0.3001
...motivate students who show low interest in school work?	0.3002	-0.2186
...calm a student who is disruptive or noisy?	0.2912	-0.2814
...help your students value learning?	0.3145	-0.2077
...craft good questions for your students?	0.2589	0.2293
...get children to follow classroom rules	0.3129	-0.2098
...motivate students who show low interest in school work?	0.3063	-0.1589
...establish a classroom management system with each group of students?	0.3051	-0.1486
...use a variety of assessment strategies?	0.2751	0.4187
...provide an alternative explanation when students are confused?	0.2465	0.4978
...assist families in helping their children do well in school?	0.2616	0.2392
...implement alternative strategies in your classroom?	0.2928	0.3399

Table A.8: Classroom Competency Survey Questions PCA Loadings: How much can you do to...

Question	Comp 1	Comp 2	Comp 3
New ideas and projects sometimes district me from previous ones.	0.4039	-0.1222	0.0034
I have been obsessed with an idea for a short time but lost interest.	0.4389	-0.2567	0.1576
I often set a goal but later choose to pursue a different one.	0.4331	-0.3216	0.1291
I have difficulty focusing on projects that take more than a few months.	0.4442	-0.2113	0.0995
Setbacks don't discourage me.	0.0477	0.4738	0.8775
I am a hard worker.	0.3377	0.5503	-0.3197
I am diligent.	0.3782	0.4937	-0.2764

Table A.9: Grit-related Survey Questions PCA Loadings

Between the loadings and the scree plot, we can conclude that most of the variance in classroom competency questions is a measure of the same construct - likely best understood as general confidence.

Principal component analysis of grit-related questions does not simplify the information as easily as for classroom-competency questions. The first component explains just under 40% of the variation, the second just under 20%, and the third, fourth, and fifth roughly 10%.

Nonetheless, the loadings do provide some interesting information (see Table A.9). The first component has positive loadings for all questions andis best thought of as overall self-reported grit. The second component differentiates between whether a resident scores themselves higher on focus-related questions compared to other questions. The third component seems to differentiate between action-related questions and the two questions that ask about personality (i.e. I am a hard-worker and I am diligent).

A.6 Correlation between Survey Questions and Gateway Scores

We may be interested in examining the relationship between survey responses and resident scores on the Gateway. In particular, a strong correlation between a resident's evaluation of their teaching and Relay's formal evaluation system would lend some credence to using survey responses as indicators of teaching quality. Generally speaking, regressions of Gateway scores and survey questions do not uncover a strong predictive relationship. Even when models are statistically significant, they are not often large in magnitude.

Table A.10 shows regressions of total Gateway scores on the two strongest principal components for classroom competency questions. Neither component explains a significant amount of variation in Gateway 1 scores or Gateway 2 scores. (Though the first component (overall confidence) trends towards significance with Gateway 2.) Similarly, principal components for grit do not explain any significant variation in Gateway 1 or Gateway 2 scores (see Table A.11).

Looking at sub-scores provides a little more context. Tables A.12, A.13, A.14, and A.15 show regressions of Gateway sub-scores on classroom competency survey questions. Of all Gateway 1 sub-scores, classroom competency questions best predict the overall subscore and engagement subscore. Five percent of variation in these subscores can be explained by classroom competency questions. Of all Gateway 2 subscores, classroom competency questions best predict aa resident's ability to convey

	Gateway 1 Coef./se	Gateway 2 Coef./se
Scores for component 1	.019	.109
	(.03)	(.06)
Scores for component 2	.018	–.098
	(.07)	(.14)
cons	29.797***	31.277***
	(.08)	(.15)
Adj. R^2	.001	.021
No. of cases	447	182

Significance: * for p<.05, ** for p<.01, and *** for p<.001.

Table A.10: Regression of Gateway scores on Principal Components for Classroom Competency Questions

	Gateway 1 Coef./se	Gateway 2 Coef./se
Scores for component 1	–.015	.118
	(.05)	(.09)
Scores for component 2	.062	.084
	(.07)	(.13)
Scores for component 3	–.005	.023
	(.09)	(.16)
cons	29.830***	31.352***
	(.08)	(.15)
Adj. R^2	.002	.012
No. of cases	451	183

Significance: * for p<.05, ** for p<.01, and *** for p<.001.

Table A.11: Regression of Gateway scores on Principal Components for Grit Questions

a warm but demanding personality. Twelve percent of variation in this subscore can be explained by survey responses. In particular, warm/demanding personality has a significant positive relationship with a resident's self-reported ability to calm disruptive students and ability to establish a classroom management system.

The small magnitude of correlation between self-reported competency and Gateway evaluations deserves some thought. There are a couple reasons we may not see a strong correlation. One hypothesis is that resident self-evaluations may be uncovering a a facet of teacher efficacy not captured by the Gateway. If this were true, we would want to optimize the selection of teacher candidates using survey-responses as well as Gateway scores. Alternatively, confidence in teaching ability may not be strongly related to actual teaching ability. It is easy to imagine this if we consider that effective teachers may hold themselves to higher standards. If this were the case, we would not want to predict survey responses as a means of predicting efficacy.

As a compromise, I report relationships between survey responses and application sub-scores in the appendix below, but I will not use survey responses to determine how application sub-scores should be weighted.

	attendance Coef./se	timeliness Coef./se	preparation Coef./se	followthrough Coef./se
...prevent and respond to disruptive behavior in the classroom?	−.002	−.006	.004	.015
	(.01)	(.02)	(.02)	(.02)
...motivate students who show low interest in school work?	.004	.000	.017	.020
	(.01)	(.02)	(.02)	(.02)
...calm a student who is disruptive or noisy?	.000	−.017	−.024	−.043*
	(.01)	(.02)	(.02)	(.02)
...help your students value learning?	.008	.007	−.018	−.016
	(.01)	(.02)	(.02)	(.02)
...craft good questions for your students?	.017	−.005	.016	.014
	(.01)	(.01)	(.01)	(.02)
...get children to follow classroom rules?	−.013	.000	−.010	−.023
	(.01)	(.02)	(.02)	(.02)
...motivate students who show low interest in school work?	−.002	−.018	.010	−.011
	(.01)	(.02)	(.02)	(.02)
...establish a classroom management system with each group of students?	.003	.044**	.021	.044*
	(.01)	(.01)	(.02)	(.02)
...use a variety of assessment strategies?	−.000	.005	.004	.007
	(.01)	(.01)	(.02)	(.02)
...provide an alternative explanation when students are confused?	−.010	.005	.005	−.004
	(.01)	(.01)	(.02)	(.02)
...assist families in helping their children do well in school?	−.017	.004	.010	.026
	(.01)	(.01)	(.01)	(.01)
...implement alternative strategies in your classroom?	.004	−.024	−.017	−.013
	(.01)	(.02)	(.02)	(.02)
cons	3.905***	3.770***	3.579***	3.591***
	(.08)	(.10)	(.11)	(.12)
Adj. R^2	.018	.027	.018	.038
No. of cases	447	447	447	447
F Statistic	.671	1.011	.681	1.417
Prob > F	.780	.437	.770	.155

Significance: * for p<.05, ** for p<.01, and *** for p<.001. All regressions are run with each of the classroom competency survey questions included.

Table A.12: Regression of Gateway 1 Scores on Survey Questions: How much can you do to...

	engagement Coef./se	feedback Coef./se	mindsets Coef./se	overall Coef./se
...prevent and respond to disruptive behavior in the classroom?	.001	–.009	.027	–.001
	(.02)	(.01)	(.02)	(.02)
...motivate students who show low interest in school work?	.008	.008	–.025	.031
	(.02)	(.01)	(.02)	(.02)
...calm a student who is disruptive or noisy?	–.034*	–.017	–.019	–.029
	(.02)	(.01)	(.02)	(.02)
...help your students value learning?	.026	.027	.015	–.017
	(.02)	(.02)	(.02)	(.02)
...craft good questions for your students?	.008	–.002	.025	.020
	(.01)	(.01)	(.01)	(.01)
...get children to follow classroom rules?	–.033	.003	.013	–.009
	(.02)	(.02)	(.02)	(.02)
...motivate students who show low interest in school work?	.003	.005	.004	–.001
	(.02)	(.02)	(.02)	(.02)
...establish a classroom management system with each group of students?	.009	.006	–.003	.015
	(.02)	(.01)	(.02)	(.02)
...use a variety of assessment strategies?	.043**	.020	–.004	.043**
	(.01)	(.01)	(.02)	(.01)
...provide an alternative explanation when students are confused?	–.023	–.002	.000	–.048**
	(.01)	(.01)	(.02)	(.02)
...assist families in helping their children do well in school?	–.017	–.022*	–.028	–.010
	(.01)	(.01)	(.01)	(.01)
...implement alternative strategies in your classroom?	–.010	–.019	.011	.014
	(.02)	(.01)	(.02)	(.02)
cons	3.819***	3.713***	3.538***	3.665***
	(.10)	(.09)	(.12)	(.11)
Adj. R^2	.055	.034	.032	.063
No. of cases	447	447	447	447
F Statistic	2.092	1.280	1.200	2.428
Prob > F	.016	.227	.280	.005

Significance: * for p<.05, ** for p<.01, and *** for p<.001. All regressions are run with each of the classroom competency survey questions included.

Table A.13: Regression of Gateway 1 Scores on Survey Questions Cont.: How much can you do to...

	setsexpectations Coef./se	reinforceexpectations Coef./se	corrective Coef./se	warmdemand Coef./se
...prevent and respond to disruptive behavior in the classroom?	.008	.025	.012	−.003
	(.02)	(.02)	(.02)	(.01)
...motivate students who show low interest in school work?	.051**	.038	.023	.037*
	(.02)	(.02)	(.02)	(.02)
...calm a student who is disruptive or noisy?	−.022	.006	.009	.018
	(.02)	(.02)	(.02)	(.02)
...help your students value learning?	−.001	−.021	−.015	−.012
	(.02)	(.02)	(.02)	(.02)
...craft good questions for your students?	−.002	−.034*	−.018	−.016
	(.01)	(.02)	(.02)	(.01)
...get children to follow classroom rules?	.008	−.004	−.015	.015
	(.02)	(.02)	(.02)	(.02)
...motivate students who show low interest in school work?	.030	−.001	.019	−.036*
	(.02)	(.02)	(.02)	(.02)
...establish a classroom management system with each group of students?	.016	.010	.038*	.041**
	(.02)	(.02)	(.02)	(.01)
...use a variety of assessment strategies?	.013	.001	−.009	.020
	(.01)	(.02)	(.02)	(.01)
...provide an alternative explanation when students are confused?	.002	.020	.011	−.021
	(.01)	(.02)	(.02)	(.01)
...assist families in helping their children do well in school?	−.019	.000	.012	−.028*
	(.01)	(.02)	(.01)	(.01)
...implement alternative strategies in your classroom?	−.021	−.002	−.008	.030*
	(.02)	(.02)	(.02)	(.02)
cons	3.053***	3.212***	3.015***	3.240***
	(.10)	(.13)	(.12)	(.10)
Adj. R^2	.099	.035	.055	.123
No. of cases	434	434	434	434
F Statistic	3.847	1.275	2.046	4.927
Prob > F	.000	.230	.019	.000

Significance: * for p<.05, ** for p<.01, and *** for p<.001. All regressions are run with each of the classroom competency survey questions included.

Table A.14: Regression of Gateway 2 Scores on Survey Questions: How much can you do to...

	engage	lesson	firstfew	profession	overall
	Coef./se	Coef./se	Coef./se	Coef./se	Coef./se
...prevent and respond to disruptive behavior in the classroom?	.008	.002	−.017	.001	.002
	(.02)	(.02)	(.03)	(.02)	(.01)
...motivate students who show low interest in school work?	.017	.020	.061*	.029	.029*
	(.02)	(.02)	(.03)	(.02)	(.01)
...calm a student who is disruptive or noisy?	−.002	−.018	−.026	−.021	−.013
	(.02)	(.02)	(.03)	(.02)	(.01)
...help your students value learning?	−.017	−.029	−.011	−.014	−.003
	(.02)	(.03)	(.03)	(.02)	(.02)
...craft good questions for your students?	.007	.038*	.049*	.006	−.008
	(.01)	(.02)	(.02)	(.01)	(.01)
...get children to follow classroom rules?	.015	−.008	.025	−.012	.028
	(.02)	(.03)	(.03)	(.02)	(.02)
...motivate students who show low interest in school work?	.015	.003	−.013	−.002	−.015
	(.02)	(.03)	(.03)	(.02)	(.02)
...establish a classroom management system with each group of students?	.005	.012	.030	.021	.013
	(.02)	(.02)	(.02)	(.02)	(.01)
...use a variety of assessment strategies?	−.016	−.022	.005	.019	.016
	(.01)	(.02)	(.02)	(.01)	(.01)
...provide an alternative explanation when students are confused?	.012	.013	−.027	−.007	−.010
	(.02)	(.02)	(.02)	(.02)	(.01)
...assist families in helping their children do well in school?	−.014	−.016	−.009	−.014	−.023*
	(.01)	(.02)	(.02)	(.01)	(.01)
...implement alternative strategies in your classroom?	.006	.020	−.006	−.025	.018
	(.02)	(.02)	(.03)	(.02)	(.01)
cons	3.203***	3.427***	3.119***	3.823***	3.313***
	(.11)	(.15)	(.18)	(.10)	(.09)
Adj. R^2	.027	.026	.087	.031	.068
No. of cases	434	350	268	432	433
F Statistic	.976	.764	2.023	1.130	2.561
Prob > F	.471	.687	.023	.333	.003

Significance: * for p<.05, ** for p<.01, and *** for p<.001. All regressions are run with each of the classroom competency survey questions included.

Table A.15: Regression of Gateway 2 Scores on Survey Questions Cont.: How much can you do to...

A.7 Correlation between Survey Questions and Application Scores

As seen in tables A.16, A.17, and A.18, there is not a strong correlation between survey responses and application sub-scores. In each round, rubric sub-scores do not explain a significant amount of variation survey responses (as measured by the F-statistic from regressions of survey components on application sub-scores). However, there are some interesting exceptions.

Of all the models, the closest to significant is the relationship between online application scores and the second classroom competency principal component (see Table A.16, Classroom PC 2), which differentiates between residents who believe they are better at instruction and residents who believe they are better at behavior management. Drilling down, we see that residents who scored higher on communication and commitment to teaching were more likely to report that they were better at instruction than behavior management. On the other hand, those who scored highly on respect and commitment to the mission were more likely to believe that they were better at behavior management.

In the video round, *high* growth mindset scores and *low* respect and humility scores were both predictive of a resident believing themselves to be better at instruction than behavior management. Additionally, high self-reported grit was strongly predictive of Growth Mindset, but only as measured by Question 2.

In the in-person interview, communication is again predictive of classroom confidence. There is also a significant *negative* relationship between confidence and critical thinking scores. In other words, those who scored higher on critical thinking were more likely to score themselves lower on classroom competency questions. We might interpret this in one of two ways. Either critical thinking does not impact classroom competence, or, more likely, critical thinkers might notice more where their teaching falls short of their standards.

Ultimately, how we interpret relationships between survey responses and application scores depends on how we interpret relationship between the survey and Gateway scores. Given the small correlation between Gateway scores and survey scores, I'm hesitant to believe the that survey scores have a substantial amount to tell us about how to form the Relay application process.

	Classroom PC 1 Coef./se	Classroom PC 2 Coef./se	Grit PC 1 Coef./se	Grit PC 2 Coef./se	Grit PC 3 Coef./se
Commitment to Mission	–.212	–.817**	.232	–.277	.046
	(.59)	(.29)	(.34)	(.28)	(.25)
Communication	–.716	.924***	–.308	.416	–.056
	(.51)	(.25)	(.30)	(.25)	(.22)
Content Knowledge	–.073	–.073	–.173	.025	–.074
	(.33)	(.16)	(.20)	(.17)	(.15)
Critical Thinking	.764	–.289	.181	.067	.220
	(.44)	(.22)	(.27)	(.22)	(.20)
Cultural Responsiveness	–.287	.242	–.447	–.144	.098
	(.49)	(.24)	(.29)	(.24)	(.22)
Grit	.135	.298	–.091	.133	–.285
	(.40)	(.20)	(.23)	(.19)	(.17)
Growth Mindset	.044	–.286	–.030	–.184	.115
	(.47)	(.23)	(.29)	(.24)	(.21)
Initiative	.195	–.045	.024	.077	.068
	(.17)	(.09)	(.10)	(.08)	(.08)
Passion	–.406	–.010	–.022	.130	.104
	(.28)	(.14)	(.17)	(.14)	(.13)
Professionalism	.620*	–.066	.041	.096	–.116
	(.28)	(.14)	(.16)	(.14)	(.12)
Respect and Humility	–.763	–.649*	.198	–.335	–.325
	(.60)	(.30)	(.36)	(.30)	(.27)
Commitment to Teaching and Service	.743	.599*	.229	.085	.313
	(.46)	(.23)	(.26)	(.22)	(.19)
GPA	–.150	.118	.186	–.134	–.010
	(.32)	(.16)	(.19)	(.16)	(.14)
cons	.359	.342	–.022	–.003	–.276
	(1.05)	(.52)	(.61)	(.51)	(.46)
Adj. R^2	.128	.193	.054	.081	.097
No. of cases	118	118	120	120	120
F Statistic	1.172	1.910	.461	.718	.877
Prob > F	.311	.037	.941	.742	.579

Significance: * for p<.05, ** for p<.01, and *** for p<.001. All regressions are run with each of the classroom competency survey questions included.

Table A.16: Regression of Survey Components on Online Application Scores

	Classroom PC 1 Coef./se	Classroom PC 2 Coef./se	Grit PC 1 Coef./se	Grit PC 2 Coef./se	Grit PC 3 Coef./se
Bonus: Communication Skills	−.398	.297	−.109	−.056	.250
	(.59)	(.30)	(.33)	(.28)	(.25)
Sample Teach: Content Knowledge and Academic Preparedness	−.313	−.293	−.097	.089	−.180
	(.46)	(.23)	(.26)	(.23)	(.20)
Q1 Critical Thinking	.197	.211	−.055	−.050	.295
	(.47)	(.24)	(.26)	(.23)	(.20)
Reflection: Critical Thinking	.212	.337	.014	.076	−.079
	(.57)	(.29)	(.32)	(.28)	(.24)
Q1 Cultural Responsiveness	.135	−.310	.035	.091	−.040
	(.52)	(.26)	(.28)	(.25)	(.21)
Q2 Grit	.009	.278	−.218	−.003	−.282
	(.46)	(.23)	(.26)	(.23)	(.20)
Q3 Grit	−.922	−.222	−.013	.049	−.473*
	(.53)	(.27)	(.28)	(.25)	(.21)
Q2 Growth Mindset	.270	−.044	.975**	.056	.198
	(.56)	(.28)	(.31)	(.27)	(.23)
Q3 Growth Mindset	.899	.106	.132	.001	.383
	(.54)	(.27)	(.29)	(.25)	(.22)
Reflection: Growth Mindset	.480	.802*	−.444	−.115	−.082
	(.68)	(.34)	(.39)	(.34)	(.29)
Sample Teach: Initiative and Achievement Orientation	.204	.172	−.336	−.073	−.145
	(.47)	(.24)	(.26)	(.22)	(.19)
Bonus: Passion and Excitement	.013	−.268	−.168	−.400	.230
	(.51)	(.26)	(.29)	(.25)	(.22)
Sample Teach: Professionalism	−.920	−.355	.134	.231	−.035
	(.53)	(.26)	(.29)	(.25)	(.22)
Q2 Respect and Humility	.123	−.096	−.370	−.097	−.028
	(.60)	(.30)	(.35)	(.31)	(.27)
Reflection: Respect and Humility	.107	−.591	.323	.103	−.007
	(.61)	(.31)	(.34)	(.29)	(.26)
cons	.173	.216	.690	.248	−.091
	(.67)	(.34)	(.37)	(.33)	(.28)
Adj. R^2	.092	.142	.129	.049	.130
No. of cases	115	115	117	117	117
F Statistic	.672	1.092	1.000	.349	1.005
Prob > F	.806	.374	.461	.988	.457

Significance: * for p<.05, ** for p<.01, and *** for p<.001. All regressions are run with each of the classroom competency survey questions included.

	Classroom PC 1 Coef./se	Classroom PC 2 Coef./se	Grit PC 1 Coef./se	Grit PC 2 Coef./se	Grit PC 3 Coef./se
Commitment to Mission	.155	–.449	–.182	.227	–.197
	(.47)	(.25)	(.27)	(.23)	(.22)
Communication	.994*	–.159	.350	.303	–.073
	(.38)	(.20)	(.22)	(.19)	(.18)
Content Knowledge	.094	.207	–.090	–.344	–.259
	(.37)	(.20)	(.21)	(.18)	(.17)
Critical Thinking	–1.395**	–.444	–.290	.155	.248
	(.47)	(.25)	(.27)	(.23)	(.22)
Cultural Responsiveness	.149	.050	.188	–.013	.107
	(.40)	(.21)	(.22)	(.19)	(.18)
Grit	.296	–.053	.210	.190	–.029
	(.41)	(.22)	(.24)	(.20)	(.19)
Growth Mindset	–.179	–.043	–.510	–.183	.079
	(.54)	(.29)	(.31)	(.26)	(.25)
Initiative	.207	–.019	.186	–.051	–.033
	(.35)	(.18)	(.20)	(.17)	(.16)
Passion	–.614	.202	–.351	–.074	.110
	(.38)	(.20)	(.22)	(.19)	(.18)
Professionalism	–.260	.142	–.244	.049	–.083
	(.40)	(.21)	(.22)	(.19)	(.18)
Respect and Humility	.091	.069	.608*	–.182	–.072
	(.53)	(.28)	(.30)	(.25)	(.24)
Commitment to Teaching and Service	.337	.417	.216	–.100	.224
	(.51)	(.27)	(.30)	(.25)	(.24)
cons	.460	.207	–.134	.042	–.038
	(.84)	(.45)	(.45)	(.38)	(.36)
Adj. R^2	.147	.092	.120	.109	.051
No. of cases	106	106	108	108	108
F Statistic	1.336	.784	1.082	.971	.424
Prob > F	.212	.665	.384	.482	.951

Significance: * for p<.05, ** for p<.01, and *** for p<.001. All regressions are run with each of the classroom competency survey questions included.

Table A.18: Regression of Survey Components on In-person Scores

Appendix B

Description of Principal Component Analysis for Application Round sub-scores

Images B.1a, B.1b, and B.1c show the eigenvalues for principal components calculated within each round. I retained components with eigenvalues above one - three components for each round.

In the online application round, the first component weights all sub-scores other than GPA roughly equally. The second component focuses in on academic factors like GPA and content knowledge and the third component places the most weight on professionalism and passion.

In the video round, the first component weights all sub-scores roughly equally. This component can be thought of as the overall score. The second component appears to differentiate between skill questions and personality questions, weighting the personality questions like grit positively. The third component is more difficult to interpret, but it puts the greatest weight on critical thinking and cultural responsiveness.

In the interview round, the first component again weights all sub-scores roughly equally and can be thought of as an overall score. The second component places the greatest weight on communication and passion and the third component places the greatest weight on personality factors - cultural responsiveness, grit, and growth mindset.

Principal components/correlation — Number of obs = 440
Number of comp. = 13
Trace = 13
Rotation: (unrotated = principal) — Rho = 1.0000

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	6.07031	4.62489	0.4669	0.4669
Comp2	1.44542	.423985	0.1112	0.5781
Comp3	1.02143	.19208	0.0786	0.6567
Comp4	.829353	.198957	0.0638	0.7205
Comp5	.630396	.0965735	0.0485	0.7690
Comp6	.533823	.0616155	0.0411	0.8101
Comp7	.472207	.0249395	0.0363	0.8464
Comp8	.447268	.0515259	0.0344	0.8808
Comp9	.395742	.0244855	0.0304	0.9112
Comp10	.371257	.0822256	0.0286	0.9398
Comp11	.289031	.0221579	0.0222	0.9620
Comp12	.266873	.0399893	0.0205	0.9825
Comp13	.226884	.	0.0175	1.0000

(a) Online Application Round Eigenvalues

Principal components/correlation — Number of obs = 308
Number of comp. = 15
Trace = 15
Rotation: (unrotated = principal) — Rho = 1.0000

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	6.63937	5.05745	0.4426	0.4426
Comp2	1.58192	.498416	0.1055	0.5481
Comp3	1.08351	.12013	0.0722	0.6203
Comp4	.963376	.175579	0.0642	0.6845
Comp5	.787797	.23842	0.0525	0.7371
Comp6	.549377	.0497676	0.0366	0.7737
Comp7	.49961	.0281058	0.0333	0.8070
Comp8	.471504	.0238612	0.0314	0.8384
Comp9	.447643	.0332226	0.0298	0.8683
Comp10	.41442	.048344	0.0276	0.8959
Comp11	.366076	.0171297	0.0244	0.9203
Comp12	.348947	.0269757	0.0233	0.9436
Comp13	.321971	.0458477	0.0215	0.9650
Comp14	.276123	.0277657	0.0184	0.9834
Comp15	.248357	.	0.0166	1.0000

(b) Video Round Eigenvalues

Principal components/correlation — Number of obs = 231
Number of comp. = 12
Trace = 12
Rotation: (unrotated = principal) — Rho = 1.0000

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	5.29897	3.88102	0.4416	0.4416
Comp2	1.41796	.30242	0.1182	0.5597
Comp3	1.11554	.178566	0.0930	0.6527
Comp4	.93697	.28403	0.0781	0.7308
Comp5	.65294	.0804208	0.0544	0.7852
Comp6	.57252	.0701198	0.0477	0.8329
Comp7	.5024	.0623527	0.0419	0.8748
Comp8	.440047	.0664099	0.0367	0.9114
Comp9	.373637	.0739889	0.0311	0.9426
Comp10	.299648	.0385586	0.0250	0.9676
Comp11	.26109	.132807	0.0218	0.9893
Comp12	.128283	.	0.0107	1.0000

(c) Video Round Eigenvalues

Bibliography

Aaronson, Daniel, Lisa Barrow, and William Sander (2007). "Teachers and Student Achievement in the Chicago Public High Schools." In: *Journal of Labor Economics* 25.1, pp. 95–135. ISSN: 0734-306X. DOI: 10.1086/508733. URL: http://www.journals.uchicago.edu/doi/10.1086/508733.

Ballou, Dale (1996). "Do Public Schools Hire the Best Applicants?" In: *The Quarterly Journal of Economics* 111.1, pp. 97–133. ISSN: 0033-5533, 1531-4650. DOI: 10.2307/2946659.

Betts, Julian, Andrew Zau, and Lorien Rice (2003). *Determinants of Student Achievement: New Evidence from San Diego*. Tech. rep., p. 174. URL: http://www.oecd-ilibrary.org/social-issues-migration-health/international-migration-outlook-2011%7B%5C_%7Dmigr%7B%5C_%7Doutlook-2011-en.

Boyd, Donald et al. (2006). "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement." In: *Education Finance and Policy* 1.2, pp. 176–216. ISSN: 1557-3060. DOI: 10.1162/edfp.2006.1.2.176. URL: http://www.mitpressjournals.org/doi/10.1162/edfp.2006.1.2.176.

Chetty, Raj, John Friedman, and Jonah Rockoff (2011). "The Long-term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." In: *American Economic Review* 104.9, p. 94. ISSN: 00028282. DOI: 10.1257/aer.104.9.2633. arXiv: arXiv:1011.1669v3.

Clotfelter, Charles T, Helen F Ladd, and Jacob L Vigdor (2007). "How and Why Do Teacher Credentials Matter for Student Achievement?"

Darling-Hammond, Linda (2012). "Teaching and the change wars: The professionalism hypothesis." In: *Leading Professional Practice in Education*, pp. 124–136. ISBN: 9781473915152. DOI: 10.4135/9781473915152.n12.

Decker, Paul T, Daniel P Mayer, and Steven Glazerman (2004). "The Effects of Teach For America on Students: Findings from a National Evaluation." In: *Mathematica Policy Research* 609, pp. 1–82.

Dobbie, Will (2011). "Teacher Characteristics and Student Achievement: Evidence from Teach For America." In: *Journal of Urban Economics* 57.2, pp. 302–319. ISSN: 00941190. DOI: 10.1016/j.jue.2004.11.001.

Evans, Kerri E (2016). "The Role of Teacher TeacherInsight Scores and Teacher Demographic Characteristics in the Identification of Effective Teachers: Using Student Performance as a Validation Tool." PhD thesis. Baker University.

Gibson, Sherri and Myron H. Dembo (1984). "Teacher efficacy: A construct validation." In: *Journal of Educational Psychology* 76.4, pp. 569–582. ISSN: 00220663. DOI: 10.1037/0022-0663.76.4.569.

Goldhaber, Dan, James Cowan, and Roddy Theobald (2017). "Evaluating Prospective Teachers: Testing the Predictive Validity of the edTPA." In: *Journal of Teacher Education* 68.4, pp. 377–393. ISSN: 00224871. DOI: 10.1177/0022487117702582.

Goldhaber, Dan, Cyrus Grout, and Nick Huntington-Klein (2017). "Screen Twice, Cut Once: Assessing the Predictive Validity of Applicant Selection Tools." In: *Education Finance and Policy* 12.2, pp. 197–223. ISSN: 1557-3060. DOI: 10.1162/EDFP_a_00200. URL: http://www.mitpressjournals.org/doi/10.1162/EDFP%7B%5C_%7Da%7B%5C_%7D00200.

Hanushek, Eric A. and Steven G. Rivkin (2012). "The Distribution of Teacher Quality and Implications for Policy." In: *Annual Review of Economics* 4.1, pp. 131–157. ISSN: 1941-1383. DOI: 10.1146/annurev-economics-080511-111001. URL: http://www.annualreviews.org/doi/10.1146/annurev-economics-080511-111001.

Harris, Douglas and Tim Sass (2009). "What Makes for a Good Teacher and Who Can Tell?" In: *Nacional center for Analysis of Longitudinal Data in Education Research* September, p. 30. ISSN: 1098-6596. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3. URL: http://www.urban.org/url.cfm?ID=1001431.

Jacob, Brian et al. (2016). "Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools." URL: http://www.nber.org/papers/w22054.

Kane, Thomas, Jonah Rockoff, and Douglas Staiger (2008). "What does certification tell us about teacher effectiveness? Evidence from New York City." In: *Economics of Education Review* 27.6, pp. 615–631. ISSN: 02727757. DOI: 10.1016/j.econedurev.2007.05.005. arXiv: arXiv:1011.1669v3.

Kleinberg, Jon et al. (2015). "Prediction Policy Problems." In: *American Economic Review* 105.5, pp. 491–495. ISSN: 0002-8282. DOI: 10.1257/aer.p20151023. arXiv: 15334406. URL: http://pubs.aeaweb.org/doi/10.1257/aer.p20151023.

Liu, Edward and Susan Moore Johnson (2006). "New Teachers' Experiences of Hiring: Late, Rushed, and Information-poor." In: *Educational Administration Quarterly* 42.3, pp. 324–360. ISSN: 0013161X. DOI: 10.1177/0013161X05282610.

Moulding, Louise R., Penée W. Stewart, and Megan L. Dunmeyer (2014). "Preservice teachers' sense of efficacy: Relationship to academic ability, student teaching placement characteristics, and mentor support." In: *Teaching and Teacher Education* 41, pp. 60–66. ISSN: 0742051X. DOI: 10.1016/j.tate.2014.03.007. URL: http://dx.doi.org/10.1016/j.tate.2014.03.007.

NCES (2006). "2003-04 Schools and Staffing Survey."

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain (2005). "Teachers, Schools, and Academic Achievement." In: *Econometrica* 73.2, pp. 417–458. ISSN: 0012-9682. DOI: 10.1111/j.1468-0262.2005.00584.x. URL: http://doi.wiley.com/10.1111/j.1468-0262.2005.00584.x.

Rockoff, Jonah, Brian Jacob, and Thomas Kane (2011). "Can you recognize an effective teacher when you recruit one?" In: *Association for Education Finance and Policy* 6.1, pp. 43–74.

Young, I. Phillip and Dane A. Delli (2002). "The validity of the teacher perceiver interview for predicting performance of classroom teachers." In: *Educational Administration Quarterly* 38.5, pp. 586–612. ISSN: 0013161X. DOI: 10.1177/0013161X02239640.