Improving the Signal Quality of Grades

Adam Chilton, Peter Joy, Kyle Rozema, and James Thomas*

December 29, 2022

We investigate how improving the signal quality of grades could enhance the matching of students to selective opportunities that are awarded early in academic programs. To do so, we develop methods to measure the signal quality of grades and to estimate the impact of changes to university policies on the identification of exceptional students for these opportunities. We focus on law schools, a setting where students are awarded important academic and professional opportunities after just one year of a three-year program. Using transcript data from a top law school over a 40-year period, we document large gains in identifying exceptional students if changes were made to certain personnel, course, and grading policies. Our findings provide motivation and a blueprint for how universities could leverage their internal records to ensure that fewer exceptional students miss out on selective opportunities.

JEL codes: I20, I23, J20

Keywords: Job market signaling, university faculty, employer learning, higher education.

* Chilton: University of Chicago Law School, e-mail: adamchilton@uchicago.edu. Joy: Washington University School of Law, e-mail: joy@wustl.edu. Rozema: Washington University School of Law, e-mail: kylerozema@wustl.edu. Thomas: Federal Trade Commission, email: jthomas2@ftc.gov. For helpful comments, we are grateful to workshop participants at the American Law and Economics Association Annual Conference, Conference on Empirical Legal Studies, Cornell Law School, Duke Law School, the University of Chicago Law School, the University of Southern California Law School, Society of Benefit Cost Analysis Annual Conference, and the Federal Trade Commission. We thank the law school administrators who generously agreed to make this data available for our research. The views expressed in this article are those of the authors and not necessarily those of the Federal Trade Commission or any individual Commissioner.

1. Introduction

Grades provide a signal about students' abilities and achievement (e.g., Spence, 1973; Stiglitz, 1975). Although the signal from any individual grade may be noisy, the grades earned throughout an entire academic program usually offer a more reliable picture of a student's potential (e.g., Becker, 1982; Grant, 2007; Ostrovsky and Schwarz, 2010). However, many selective opportunities are awarded early in academic programs, which means that a relatively small number of grades can have a profound influence on which students receive them.

The costs of awarding selective opportunities early in academic programs have been prominently documented in the case of unraveling markets for post-graduation employment (e.g., Roth and Xing, 1994; Kagel and Roth, 2000). Unraveling occurs when employers race to hire students earlier and earlier, causing decisions to be made with a limited set of grades and leading to costly mismatches (e.g., Fredriksson, Hensvik, and Skans, 2018). Efforts to combat unraveling have focused on market-wide re-organizations to prevent early hiring, including the National Resident Matching Program for new physicians (Roth, 1984; Niederle and Roth, 2003). Beyond post-graduation employment, many other selective opportunities are awarded early in academic programs, including participation in competitive activities, admission into selective majors, and transferring to other universities. The effects of receiving these opportunities can be large. For example, Bleemer and Mehta (2021) find that students who receive grades barely above the threshold needed to major in economics earned 45 percent higher annual early-career wages than those just below the threshold. Unlike the efforts to combat unraveling in labor markets, however, we are unaware of research on how to improve the matching process for these early opportunities.

We investigate how improving the signal quality of grades could enhance the matching of students to selective opportunities. We begin with a simple conceptual framework where employers would like to hire students based on fixed legal abilities that are incrementally revealed through grades.[1] Employers would like to hire students based on grades at graduation; however, hiring decisions are made early in the academic program when only a subset of grades are observed. Our guiding framework abstracts from strategic choices of study effort; however, we

[1] Although our framework assumes that students have a time-invariant ability, we explain in Section 3.1 that our framework is robust to changes in student ability over the course of law school in ways that preserves initial student rankings. However, if there are changes to student ability over the course of law school in ways that do not preserve initial student ranking, it suggests our approach is conservative in the sense that it would strengthen the case against relying on first year grades for awarding selective opportunities. Similarly, our approach is also robust to rank-preserving human capital investments.

investigate this issue empirically and find that changes in effort are not influencing our findings.[2] This framework, where a fixed ability is partially revealed by a series of noisy signals, is consistent with an extensive literature focused on incomplete information in human capital settings (e.g., Farber and Gibbons, 1996; Altonji and Pierret, 2001; Arcidiacono, 2004; Thomas, 2019). Under this framework, we show that it is possible to measure the signal quality of grades, and we develop methods to estimate whether changes to university policies that improve the quality of those signals could aid the identification of exceptional students early in academic programs.

We focus on a setting where early grades determine important academic and professional opportunities: law schools. In this setting, it has been widely documented how the first year of grades out of a three year program determine who receives many selective opportunities, including participation in student journals, summer internships, post-graduate employment at top law firms, and post-graduate service as judicial law clerks (e.g., Li and Rosen, 1998; Kagel and Roth, 2000; Ostrovsky and Schwarz, 2010; Engel, 2018; Ambuehl and Groves 2020). Despite widespread acknowledgement of this problem by academics (Avery et al., 2007) and law firm partners charged with hiring (Ginsburg and Wolf, 2003), repeated attempts to reorganize this market have collapsed as individual judges and law firms continue to "jump the gun" (Roth and Xing, 1994). For example, there have been repeated attempts to reorganize the market for hiring law students as judicial law clerks, but they have all failed (Haruvy, Roth, and Ünver, 2006).

Beyond these substantive reasons for focusing on law schools, this setting also offers several methodological advantages. First, the first year law school curriculum has been stable over time, which allows for comparisons between the signal quality of grades for the same courses across time and professors. Second, first year students do not choose their courses or professors, which eliminates concerns of student self-selection into particular classes and allows us to exploit random assignment to classes (e.g., Golsteyn et al., 2021). Third, law professors grade exams themselves rather than utilizing teaching assistants, which removes a layer of noise that is present in many other departments. Fourth, law school grades in the first year are usually based on a single, blind-graded exam, which ensures grades are based on students' written exams and professors' evaluation of them. Finally, grades in the first year law school curriculum usually must meet a mandatory mean requirement (they can, however, have different variances).

We develop a measure of the signal quality of grades and methods to estimate the impact

[2] Section 5.4 addresses problems posed by changes in student effort in response to first-year grades.

of changes to university policies on the identification of exceptional students.[3] We specifically measure the signal quality of grades at the classroom-level as the correlation between those grades and students' GPAs in all second and third year classes.[4] Using this measure, we exploit plausibly exogenous variation in professors' teaching assignments to estimate differences in the classroom-level signal quality by professors and course policies. We then apply these estimates in a series of simulations to estimate how the share of missing exceptional students would differ under alternative university policies that are not directly observed in the data.

As we demonstrate, this approach to measuring the signal quality of grades is robust to several important aspects of the institutional setting. For instance, one potential concern is that the students may only make it into the top of the class at the end of law school because the truly exceptional students change their behavior after hiring decisions are made after the first year. Investigating this concern, we find that our measure is largely unaffected by potential changes in student effort after the first year.[5] Additionally, another potential concern is that professors who assign grades that have a low correlation with other courses may be measuring alternative dimensions of student ability, but we present evidence suggesting that this is not a major concern because law school grades primarily measure a single-dimension of student ability. In other validation exercises, we also find that our measure identifies professors who consistently assign low or high-noise grades and does not punish professors who assign more dispersed grades.[6]

To illustrate the value of our approach, we use transcript data from a top ranked American law school over a 40 year period to investigate how three kinds of policy changes would influence the share of "missing exception students"—that is, students who exceeded a particular class rank

[3] Designating students with terms like "exceptional" or "gifted" often runs into criticism (e.g., Kamenetz, 2015). The term "exceptional" has been used, however, in a number of contexts to identify specific sets of students. For instance, "twice-exceptional students" is a term "used to describe gifted children who have the characteristics of gifted students with the potential for high achievement and give evidence of one or more disabilities" (National Association for Gifted Children, 2021). We use the term "exceptional student" to refer to exceptional on only one dimension: final law school grades. Of course, there are many ways that students can be exceptional that are not captured by their grades, and there may be costs of more precise grades in terms of equity and culture. We acknowledge that these and other considerations are important. This article seeks to estimate the benefits of more precise grades so they can be weighed against these costs.

[4] We use the term "classroom" to refer to a specific course taught by a specific professor in a specific semester (e.g., Professor Doe's Torts class in the fall of 2010) and the term "course" to refer to classes on the same subject (e.g., Torts).

[5] For instance, Section 5.4 documents that the correlation by semester does not vary greatly, and this is true for all the different groups of exceptional students. This suggests that behavioral changes after the first year are not affecting our measure of the signal quality of grades.

[6] For instance, Section 5.5 documents that when professors who assign noisy grades have the same student in multiple classes, the grades the professor assigns to the same student across classes are less correlated than the correlation of the student's grades assigned by professors who assign more precise grades.

threshold at graduation but did not exceed that same threshold early in their academic program.[7]

First, we estimate the effects of changing personnel management policies. We find that changing the set of professors teaching first year classes would have a meaningful impact on the share of missing exceptional students after the first year of law school. For example, we find that replacing the noisiest quarter of professors with the least noisy professors would decrease the share of missing exceptional students in the top 10 percent by 16 percent.[8] This magnitude is equivalent to 45 percent of the decrease in missing exceptional students if the labor market for new lawyers were reorganized to ensure hiring decisions were based on two years of law school grades. We also find that the failure to ensure that students are taught by professors of similar signal quality can have considerable distributional consequences.

Second, we estimate the effects of changing aspects of the first year curriculum. We find that some widely discussed curriculum reforms—like reducing class sizes—would do little to reduce the share of missing exceptional students, but that some other reforms—like increasing the number of assessments—could meaningfully reduce it. For example, we find replacing each semester long 4-credit class in the first year with two 2-credit classes would decrease the share of missing exceptional students in the top 10 percent by 15 percent. This magnitude is equivalent to 29 percent of the decrease in missing exceptional students if the labor market for new lawyers were reorganized to ensure hiring decisions were based on two years of law school grades.

Third, we estimate the effects of several kinds of changes to grading systems. We begin by assessing whether the number of distinctions available on a grading scale affects the identification of exceptional students. We find that moving from 30 grade distinctions (which is used at the law school we study) to 5 grade distinctions (which is used at several top schools) would increase the share of missing exceptional students in the top 10 percent by 27 percent. We also assess the impact of moving to pass-fail grading during emergencies, such as during the COVID-19 pandemic. We find that even if emergencies dramatically lower the signal quality of traditional grades, those grades may still provide valuable information not offered by pass-fail assessment.

[7] For example, a student who finishes in the top 10 percent of their class by our definition is an exceptional student; if that student is not in the top 10 percent of their class after their first year, then they would be "missed." We focus on this outcome because the costs of awarding early opportunities and the costs of unraveling are disproportionately borne by these students. This measure thus offers a concrete way to express the share of students who could benefit from increased focus on the signal quality of grades.

[8] In Section 5.6, we show that our results are robust to measuring the costs of noisy grades without using cutoffs that identify the students whose grades fall above a specific cutoff.

These results suggest that the implications for identifying exceptional students should be a first order concern for law schools considering changing their grading policies.

Our findings provide motivation and a blueprint for how universities could leverage their internal records and policies to ensure that fewer exceptional students miss out on selective opportunities. Although we focus on the labor market for law students, the methods we develop could be used by any academic department to assess their personnel management, curriculum, and grading policies. For example, many colleges and universities admit undergraduate students into selective majors based on their first few semesters of grades. These admissions decisions would be more accurate if greater emphasis was placed on measuring and improving the signal quality of undergraduate grades. Additionally, many academic departments currently rely heavily on student evaluations as a measure of professor effectiveness in the classroom, but they should also consider the signal quality of professors' grades as a measure of their teaching effectiveness. That said, future research should also explore the relationship between a professors' ability to effectively help their students improve their human capital and the signal quality of grades. This is because academic departments should not make personnel decisions solely based on professors' ability to assign grades with high signal quality. Instead, academic departments should also try to ensure that they prioritize identifying instructors that are effective educators more generally.

This article contributes to several academic literatures, including on the signaling potential of grades (e.g., Stiglitz, 1975; Arcidacono, 2004; Zafar, 2011; Stinebrickner and Stinebrickner, 2014; Kuehn and Moss, 2019; Thomas, 2019; Hvidman and Sievertsen, 2021), how to improve the matching of students to employment opportunities (e.g., Roth and Zing, 1994; Li and Rosen, 1998), how grading methods can impact student assessment (Negin, 1981; Schwarcz and Farganis, 2017; Colker et al., 2017), and law school administration (e.g., Clarke, 2014; Simkovic and McIntyre, 2014; McIntyre and Simkovic, 2017; Ho and Kelman, 2014; Ginsburg and Miles, 2015; Chilton, Masur, and Rozema, 2021).

This article proceeds as follows. Section 2 provides background and describes the data. Section 3 introduces a measure of the signal quality of grades and our methods for estimating the impact of changes to university policies on the identification of exceptional students. Section 4 reports results. Section 5 reports a series of exercises that test the validity of our measure of the signal quality of grades. Section 6 concludes.

2. Background and Data

2.1. Law School Grading

At American law schools, the primary degree program is typically three years long. During the first year, students primarily take a set of required courses—including Civil Procedure, Constitutional Law, Contracts, Criminal Law, Torts, Property, and Legal Writing—with assigned professors. During the second and third year, students primarily select their own courses. Although this curriculum is fairly standard, there is considerable heterogeneity in the way these courses are graded across law schools. Many law schools use A to F grades (with pluses and minuses), some law schools use pass-fail grades, and a few law schools have bespoke grading systems. Moreover, within a school, there can be differences in the grading policies across courses. For instance, certain courses may be pass-fail or have different required distributions.

Regardless of the grading system, a range of selective opportunities are extended to law students in large part based on their grades (Ginsburg and Wolf, 2003; Engel, 2018). For instance, law review membership is an important credential valued by employers that is often largely determined based on first year grades.[9] Additionally, there is a sizable market in students transferring to higher ranked law schools, which largely occurs after students complete their first year. Finally, although post-graduation employment opportunities were extended to students based on their final grades through the 1950s, over time the market unraveled so that many of these selective opportunities are now extended to students after just one year of law school (Roth, 2012). For example, of all students graduating in 2021 that were hired as law firm associates, 96 percent were summer associates after their second year.[10]

Of course, even though there is evidence that law school grades are associated with future career opportunities and success (Sander and Bambauer, 2012), it does not mean that grades are destiny. But it does suggest the value of improving the matching of students to early opportunities. That said, because unraveling is a collective action problem, solving this problem directly would likely require coordinated, industry wide reforms. But these kinds of sweeping reforms have not

[9] Research has suggested that law review membership is associated with greater probabilities of obtaining judicial clerkships (Samida, 2004), higher salaries for associates at law firms (Rebitzer and Taylor, 1995), and better placements in the academic job market (Merritt and Reskin, 1997).

[10] This is an implied estimate from National Association for Law Placement (2022) that divides (1) the number of summer associates who accepted full time offers by (2) the number of summer associates who accepted full time offers plus the number of third year law students hired (which is estimated by the number of offers made times the acceptance rate).

occurred decades after the market for new lawyers has been widely acknowledged to have unraveled and they also risk violating federal antitrust laws.[11] This suggests the need to find other ways to improve the matching of students to selective opportunities.

2.2. Institutional Setting and Data

To illustrate how better grading could improve the matching of students to early opportunities, we use data from a top-20 ranked American law school over a 40 year period.[12] The curriculum at this school is consistent with the curriculum at other major American law schools. For instance, the students are required to take six doctrinal courses and a legal writing class during their first year.[13] The first year students are randomly assigned to three sections that are demographically balanced. Students take four of the six doctrinal first year courses with all the members of their section. The remaining two doctrinal courses are taught with the section randomly split in half to create a "small class" format. Depending on teaching needs and professor schedules, the six doctrinal classes are offered in either the fall and spring semester of the first year.[14] After the first year, the only requirements are for students to take a legal ethics course and to satisfy a writing requirement by taking a seminar. The grading scale used ranges from a low of 2.50 (F) to a high of 4.30 (A+) at intervals of 0.06 GPA points.[15]

We specifically have transcript data of JD candidates who attended this law school from 1979 to 2019. We make two sample restrictions to our data. First, we restrict our sample to full time students. Second, we restrict our sample to students who started and then graduated from this law school within 3 years (by doing so, we exclude all transfer students and students who did not complete their degree on the normal schedule). After these restrictions, our sample includes 8,486 students.[16] For these students, we have data at the student-classroom level. In total, we have

[11] For instance, the National Resident Matching Program for physicians required Congress to pass an antitrust exception.

[12] Our data is considerably more extensive than other recent studies of law school assessment. For instance, Ho and Kelman (2014) study the effect of law school class size on the gender gap in student performance using transcript data from students at Stanford Law School from 2001 to 2012; Schwarcz and Farganis (2017) study the impact of individualized feedback on future grades using transcript data from students at the University of Minnesota Law School from 2011 to 2015; and Birdsall, Gershenson, and Zuinga (2020) study the effect of study-faculty demographics on student performance using transcript data from students at one anonymous top-100 law school during a 10 year period.

[13] Figure A1 in the Appendix reports the required first year courses for each entering cohort.

[14] Table A1 in the Appendix reports the number of classrooms in the fall and spring semesters for each first year course.

[15] This law school changed its grading system during our sample period. However, this change was just a scale change and therefore grades in the two systems are exactly convertible through the conversion in Table A2 in the Appendix.

[16] Figure A2 in the Appendix reports the number of students in each cohort in our sample.

data on 242,695 grades at the student-class level for 10,069 individual classrooms.[17]

2.3. Defining Exceptional Students

To quantify the problems associated with awarding selecting opportunities without complete grade information, we begin by defining thresholds based on class rank. Once we have identified students who are exceptional based on various thresholds of graduation rank, we then determine which of these students would not have been classified as exceptional if they had been evaluated based on class rank earlier in law school. We call these "missing exceptional students" and our main analysis focuses on how various policies would reduce the share of exceptional students who are "missing" when screening occurs earlier in law school.

We focus on five categories of exceptional students.[18] First, we focus on the top 33 percent at graduation because this is roughly the cutoff for students who graduate cum laude. Second, we focus on the top 20 percent at graduation because it roughly corresponds to the share of students in our sample that receive top law firm jobs. Third, we focus on the top 10 percent because this is the group of students who effectively automatically make law review and because this is usually the cutoff for students who get federal district court clerkships. Fourth, we focus on the top 5 percent because this is roughly the cutoff for top federal circuit court clerkships based on the historical share of students getting these jobs in our sample. Finally, we focus on the top 1 percent because these are the students who are likely to receive clerkships with the most selective judges.

2.4. Stability of Grades Across Semesters

To provide context for our results on the extent to which improving the signal quality of grades could increase identification of exceptional students, we first use our data to explore the stability of grades across semesters.

First, Figure 1 assess the stability of grades over time. At a basic level, improving the signal quality of grades means making grades more stable over time. An initial question is thus the extent that grades are already stable across semesters. To assess this, Panel A of Figure 1 reports a scatterplot of first year grades and upper level grades. We find that first year grades explain 80

[17] For first year classes, the data includes 1,354 individual classrooms, consisting of 68 different first year professors. Figure A3 in the Appendix reports the time span that each professor taught at the law school.

[18] When defining thresholds based on class rank, the thresholds that are associated with different student outcomes likely differ across law schools. For example, circuit court judges and top firms likely pull deep into the class at the highest ranked law schools, but they likely pull less deep into the class at lower ranked law schools.

percent of the variation in upper level grades. Panel B of Figure 1 reports a Sankey diagram of the first year and upper level grades by quartile. For students in the top quartile of the class after the first year, 68 percent remain in the top quartile by graduation, 24 percent drop to the 2nd quartile, 7 percent drop to the 3rd quartile, and 1 percent drop to the bottom quartile.

Second, Figure 2 investigates how the share of missing exceptional students would decrease if employers were to wait for more grades before making hiring decisions. To do so, we calculate the share of students who would graduate with a given class rank who were not at the class rank after each semester. The figure reveals that each semester of delay before hiring would meaningfully reduce the share of missing exceptional students. For instance, the share of missing students in the top 10 percent decreases from 24 percent after 2 semesters to 16 percent after 3 semesters and to 12 percent after 4 semesters; the share of missing students in the top 1 percent decreases from 40 percent after 2 semesters to 27 percent after 3 semesters and to 22 percent after 4 semesters.[19] Additionally, although the results in Figure 2 are aggregated across the 40 year sample, the share of missing students has been stable over time (see Appendix Figure A5).

3. Methods

Our empirical approach proceeds in three steps: measuring the signal quality of grades, estimating how various policies might affect signal quality, and simulating how changing these policies would reduce the share of missing exceptional students.

3.1. Measuring Signal Quality

To begin, suppose student *i* has latent legal ability α_i, which captures all fixed attributes that affect law school performance.[20] Employers would like to hire students based on α_i; however, because this is not directly observed, they rely on earned grades as a noisy signal of α_i. Suppose the grade g_{isc} earned by student *i* in semester *s* in classroom *c* is a function of student ability α_i and idiosyncratic grading noise η_{ic} according to Equation (1).

$$g_{ics} = \alpha_i + \eta_{ic} \tag{1}$$

Classes are taken in subsequent semesters, so ability is partially revealed by a series of noisy

[19] Figure A4 in the Appendix reports the share of missing exceptional students separately for each first year course.

[20] In particular, α_i includes student *i*'s general willingness to spend time and effort studying. If student *i* studies particularly hard in class *c*, this will be subsumed into η_{ic}.

signals over time. Let the set of all classes that students take in the academic program over all semesters be partitioned into a set of first year classes {1L} and a set of upper level classes {UL}. Employers would like to hire students based on GPA in all classes, denoted by $\bar{g}_i$. However, because hiring decisions are made early in the academic program when employers only observe the set of first year grades, students are screened based on first year GPA, denoted by $\bar{g}_i^{1L}$.

As such, the match quality of hiring depends on the quality of first year grades $\bar{g}_i^{1L}$ as a signal for overall grades $\bar{g}_i$. To the extent that idiosyncratic grading noise η_{ic} in Equation (1) varies across professors, the curriculum, and grading systems, law school administrators can improve the matching process by choosing a set of policies $\boldsymbol{P}$ that minimize the sum of squared errors between first year GPA and final GPA according to Equation (2).

$$\min_{\boldsymbol{P}} \sum_i \left(\bar{g}_i(\boldsymbol{P}) - \bar{g}_i^{1L}(\boldsymbol{P})\right)^2 \tag{2}$$

With this setup in mind, we are interested in measuring the signal quality of grades in first year classes and estimating the effects of policies on them. To measure the signal quality of grades in first year classes, we use the correlation between grades in classroom c and GPA across all upper level classes, denoted by $\bar{g}_i^{UL}$,[21] according to Equation (3).

$$I_c = \text{Corr}\left(g_{ics}, \bar{g}_i^{UL}\right) \tag{3}$$

As we document below, the measure in Equation (3) has a number of desirable properties for measuring the signal quality of grades. For example, Section 5 shows that the measure is robust to heterogeneous professor preferences for tight or disperse grading distributions, whereas other intuitive measures mechanically label classes with disperse grades as noisier. Section 5 also reports a series of validation exercises showing, among other things, that correlation with upper level GPA is a reliable measure of signal quality, the measure is able to identify professors who consistently assign low or high-noise grades, the measure does not punish classes for evaluating another dimension of student achievement, and there is no evidence that student effort changes over time

[21] Note that we use correlation with upper level GPA rather than leave-one-out GPA. In addition to a technical reason related to our simulation approach discussed below, there are two practical reasons for preferring the correlation with upper level GPA. First, leave-one-out correlation partially measures how well a particular first year grade explains variation in other first year grades. This is less intuitive because employers already observe all first year grades so it is unclear why they would care about explaining variation in first year grades. Second, correlation with upper level GPA is easier for administrators to compute because it does not require constructing separate leave-one-out GPAs for each first year class.

in our sample. Additionally, our framework assumes that students have a time-invariant ability, but our framework is robust to assuming that student ability changes over the course of law school in a way that preserves initial student rankings (see Appendix B1).[22] But if student ability instead changes over the course of law school in a way that does not preserves initial student rankings, this only strengthens the case against relying on first year grades for awarding selective opportunities. This is because ability changes are effectively additional noise in upper level grades that cannot be predicted by first year grades. However, first year grades that are more correlated with upper level GPA will still be more informative about final performance than less correlated first year grades. Therefore, the correlation measure would still be a useful measure of signal quality.[23]

Intuitively, I_c measures the extent to which grades in a particular first year class explain future performance in all second and third year courses. To illustrate, Figure 3 provides an example of how the same set of forty students may be graded in hypothetical Torts and Contracts classrooms. For both classrooms, the distribution of grades is identical, there is one box for students S1 to S40, and the boxes are numbered based on the student's ranking in the distribution of upper level grades. There is one key difference between the classrooms: the grades the students received in Torts are more aligned with their upper level grades. The correlation between Torts grades and upper level grades is 0.70—indicating relatively precise grades—and the correlation between Contracts grades and upper level grades is 0.40—indicating relatively noisy grades. This means that grades in the Torts classroom provide more information than the grades in the Contracts class about how students will perform in upper level courses.

Figure 4 reports the distribution of the correlation measure from our actual data. The solid line reports the distribution across all classroom observations and the dashed line reports the distribution of professor averages. The mean correlations across classrooms and professors are 0.66 and 0.67, suggesting that average first year grades are relatively informative. However, both distributions have long left tails suggesting that there are some individual classes with noisy grades

[22] Similarly, rank-preserving human capital investments also do not affect the results.
[23] It is important to note that changes in student effort in response to first-year grades pose potential problems for our approach. We address this issue in Section 5.4.

and some professors who assign noisy grades on average.

3.2. Estimating How Policies Affect Signal Quality

After computing the correlation measure at the classroom level, the second step of our approach is to estimate how various policies would affect the signal quality of grades as measured by these correlations. The purpose is to understand how individual policies affect the sum of squared errors between final GPA and first year GPA as described in Equation (2). In some cases, there is insufficient variation to credibly identify the effects of specific professor or course attributes. In those cases, we rely on descriptive comparisons of correlations across classes and admit the limitations of such comparisons. In other cases, described in detail below, we estimate the effects of specific attributes by estimating Equation (4) for classroom c (e.g., Torts by Professor Doe in Fall 2010), course subject s (e.g., Torts), professor p (e.g., Professor Doe), and semester-year t (e.g., Fall 2010), where X_c is some characteristic of a classroom, γ_t are semester-year fixed effects, and ϕ_{ps} are professor-course fixed effects.

$$\mathrm{I}_c = \alpha + \beta\, \mathrm{X}_c + \gamma_t + \phi_{ps} + \epsilon_c \qquad (4)$$

The coefficient of interest is β, which estimates the differences in signal quality between some characteristic of the classroom, whether it be related to the professor or the course. For example, we estimate the effects of class size on signal quality exploiting the fact that two of the six doctrinal courses each year are taught in a small class format with half the usual number of students. Because the courses offered in the small class format is based on teaching needs that vary from year to year, we observe professors teaching the same courses in both small and large class formats. This allows us to use an indicator for "small" in place of X_c in Equation (4). In this case, β represents the effect of the small class format on the correlation signal quality measure using within professor-course variation in class size.

3.3. Simulating How Policy Changes Would Affect Missing Students

The methods described above identify the effects of various policies on the signal quality of grades at the classroom level. To evaluate how changes in policies affect the matching process for entire cohorts, our third step is to simulate the effects of cohort-wide policies on the share of missing exceptional students across an entire cohort.

To do so, we simulate individual first year course grades and upper level GPAs for many

synthetic cohorts of students. In these simulations, we draw classroom-specific correlations from distributions that differ to reflect counterfactual scenarios.[24] For example, for an analysis of the benefits of replacing large classes with small classes, we draw classroom-specific correlations from the observed distribution and from a counterfactual distribution that is shifted to reflect the estimated difference in signal quality between large and small classes. The magnitudes of these shifts come from estimates of β in Equation (4).

We then use these simulated grades to compute missing exceptional student. Varying the correlations between first year grades and upper level grades directly affects the probability that exceptional students are missed. If the correlation between first year grades and upper level grades is equal to one, all students in the top, say, 10 percent based on final grades will also be in the top 10 percent based on first year grades. Conversely, if the correlation between first year grades and upper level grades is zero, the share of missing exceptional students in the top 10 percent approaches 90 percent—the share that would be missing from the top 10 percent by pure chance.

Before continuing, two additional clarifications about our approach are necessary. First, although we use correlation with upper level GPA as our measure of signal quality, our missing exceptional student statistics always use final GPA at graduation to classify students as exceptional. Second, because graduation rank is still a noisy measure of unobserved student ability, some students will still be misclassified at graduation. Because of this, hiring after graduation should not be treated as a full information scenario; instead, it should be considered the most informative scenario given only information on law school performance.

4. Results

We now use our data and methods to explore the potential gains in identifying exceptional students from changes to personnel, the first year curriculum, and grading systems.

4.1. Changes to Personnel Management

We first explore the extent to which changes to personnel management policies could improve the signal quality of grades. Figure 5 reports the classroom level correlation for different

[24] Appendix B provides details on our simulation procedure. The specific simulation procedure for each of the baseline simulations is different for each counterfactual. As described in Appendix B, we model the correlation structure and choose the specific way we ran the simulations for the different counterfactuals so that the average absolute value of the difference between the simulated and actual share of missing exceptional students is less than a 5 percent deviation of the actual share of missing exceptional students. But note that there are still differences between the baseline results in the simulations and in the observed data.

first year law professors separately by course.[25] Within a panel, a professor is shown by a horizontal line representing their minimum and maximum correlation and the points are the different classroom level correlations. These results reveal meaningful differences between professors. The average professors in a course have a 42 percent higher correlation than the noisiest professors in the same course (0.62 compared to 0.44). Moreover, the least noisy professors in a course have a 67 percent higher correlation than the noisiest professors in the same course (0.72 compared to 0.44). We conduct four simulations using these estimates and the methods described above.

Range of Professors. Using the set of professor correlations from Figure 5, Panel A of Figure 6 compares a setting where all first year classes are taught by the noisiest professors who teach that course to one where all first year classes are taught by the least noisy professors who teach that course. Correlations are drawn randomly from all observations of these professors teaching these courses to account for within-professor variation in signal quality (see Appendix B2 for details). To provide a baseline, the light gray shaded region represents the share of missing students after one and two years of law school observed in the data and as reported in Figure 2 (we also provide shaded region in Figures 8 and 9). We find that the share of missing students in the top 10 percent would be 47 percent lower if all classes were taught by the least noisy professors compared to if they were taught by the noisiest professors. The gains would be larger in absolute terms for students in the top 5 and top 1 percent, but the gains are similar in relative terms. The magnitude of this is equivalent to 92 percent of gains from delaying hiring by a year.

Replacing Professors. One potential reform would be replacing noisy professors with less noisy professors. To simulate this scenario, we assume three types of professors for each first year course: ones that have the minimum, median, and maximum professor-level mean correlation. We then simulate replacing minimum correlation professors with maximum correlation professors for each doctrinal course (excludes legal writing). Panel B of Figure 6 reports these results. We find that replacing the worst professors in the doctrinal classes would decrease the share of missing students in the top 10 percent by 16 percent. This magnitude is

[25] We restrict each professor-course to those where a professor has taught the class at least 5 times. Additionally, Table A3 in the Appendix shows that there are few professors who teach multiple first year classes, and Figure A6 in the Appendix reports the relationship between a professor's correlations between the first year courses they teach for the few professors who teach multiple first year courses. As a result, although there are technically enough bridging observations to separately identify professor and course effects, this decomposition would be based on a small and unusual set of bridging observations and thus would lack credibility. Therefore, we take a more basic approach and assess differences between professors within different first year classes.

equivalent to 32 percent of the gains from delaying hiring by a year.

Coaching Professors. Another potential reform would be providing professors with coaching on their grading. Across a range of domains, there is evidence that simply providing feedback can produce considerable improvements in performance (e.g., Devine, 2012; de Haan and Linde, 2018). This is especially true in areas where individuals otherwise do not get prompt feedback on their performance (Thaler and Sunstein, 2009). For instance, there is evidence that racial bias in NBA referees' foul calls disappeared after they were notified of the issue (Pope, Price, and Wolfers, 2018) and that developing rating systems for hospitals led to improvements in medical care (Hibbard, Stockard, and Tusler, 2003; Propper et al., 2010).

Although we do not have direct evidence of how much professors could improve from coaching, we do have evidence that professors improve their grading over time. Regressing an indicator variable for first time teaching a class on professor-course fixed effects, we find that the correlation is 0.04 lower the first time a professor teaches a class ($p<0.1$). Although we cannot observe why professors improved or whether feedback would yield improvements of this magnitude, we take this as evidence that such improvements are feasible. To estimate the potential benefits of coaching, we therefore assume that every professor could increase their correlation by the same amount it increases after the first time teaching a class. Panel C of Figure 6 reports these results. We find that increasing all classroom correlations by 0.04 would reduce the share of missing students in top 10 percent by 10 percent. The magnitude of this is equivalent to 28 percent of the gains from delaying hiring by a year.

Our research also suggests that students would benefit from coaching on taking law school exams. We collected data on the order that all first year law school exams were administered for the law school we study since 2001. There is considerable variation in the order of the exams, and that variation is exogenous with respect to the students because the schedule is set based on the schedules of professors teaching in that first year section. We are therefore able to identify how exam order affects the signal quality of grades that students receive. We find that students receive grades that have a 0.04 lower correlation on the first exam they take during law school than the correlation for all subsequent exams ($p<0.1$). This suggests that signal quality of grades could be improved if students were given more coaching in advance of their first high-stakes evaluation.

Balancing Sections. One potential reform would be balancing the signal quality of professors across the sections that the first year students are divided into. If one section is taught

disproportionately by noisy professors, the distribution of first year grades for the top students in this section would be compressed because the greater noise prevents the best students from consistently earning the best grades. However, unbalanced sections would also yield fewer missing exceptional students in precise sections. In aggregate, these opposing effects might be expected to counterbalance one another.

To investigate this empirically, we assume there are three sections of students in every cohort. Each student is randomly assigned to one section, and the students in each section take all seven classes together. This implies there are 21 first year professors for every cohort. To analyze the effects of balancing sections, we assume the signal qualities of these 21 professors are drawn from all observations of the professors with the minimum, median, and maximum correlation from Figure 5. In the unbalanced scenario, one section draws all correlations from the least noisy professor, one section draws all correlations from the median professor, and one section draws correlations from the noisiest professor. In the balanced scenario, these 21 distributions are assigned to sections to balance the mean and standard deviation of professor correlation across sections (see Appendix B2 for details).

Panel D of Figure 6 reports these results. We find that the difference in the share of missing exceptional students between balanced and unbalanced sections is very small. However, we do find that balancing sections would promote equity. Figure 7 reports the share of missing students separately by section, and reveals that the share of missing exceptional students is substantially higher in the noisy section and lower in the precise section. For instance, the share of missing students in the top 10 percent is 31 percent for the section with the noisiest professors, 22 percent for the section with average professors, and 18 percent for the section with the least noisy professors. As such, although balancing sections would have limited effects of overall signal quality, it may still be worth considering given the noticeable impact on equity.

4.2. Changes to the First Year Curriculum

We next estimate the extent that changes to three aspects of the curriculum of required first year courses affect the share of missing exceptional students.

Small Classes. A large literature investigates the effects of class size on student evaluation and performance (e.g., Angrist and Lavy, 1999; Hoxby, 2000; Urquiola and Verhoogen, 2009; Ho and Kelman, 2014). To investigate the relationship between the signal quality of grades and class size, we exploit quasi-random variation in class size. As discussed above, the law school we study

assigns students to one small doctrinal class each semester in the first year. Professors are assigned to small classes based on teaching needs. Even though a small class comes with fewer exams to grade, professors for these classes are required to have a midterm assessment and offer feedback to each student. This generates plausibly exogenous year-to-year variation in whether a professor is teaching a first year course as a small class. We regress the classroom-level correlation on an indicator for small class and professor-course fixed effects. Intuitively, this estimates how the signal quality of grades varies when the same professor teaches the same course in small and large sizes. The results reported in Column 1 of Table 1 reveal that professors have a 0.025 higher correlation when they teach a course as a small class compared to teaching it as a large class. However, because the small class format also has required midterm assessments and additional feedback, the correct interpretation of these estimates is that they are the combined effects of small class sizes and additional assessment.

Based on that estimate, Panel A of Figure 8 simulates the share of missing exceptional students if all first year classes were small or large. We find that the share of missing students in the top 10 percent would be 5 percent lower if all classes were small classes compared to if all classes were large classes. While this is an improvement equivalent to 14 percent of the gains from delaying hiring by a year, it is a modest effect for a policy that doubles first year staffing needs.

Splitting Classes. We next examine the difference in the signal quality of grades for courses of different credit hours. Even if classes for fewer credits produce noisier grades, a larger set of slightly lower signal quality grades is better than the aggregate signal of a smaller set of better signal quality grades. To explore this, we first estimate the difference in signal quality between 1-2 credit courses and 3-4 credit courses by regressing the classroom level correlation on professor fixed effects. Because there are not 1-2 credit classes in the first year other than Legal Writing, we pool all course observations for this analysis and use a different correlation measure that leaves out only the class in question rather than leaving out all first year classes as in the other correlation measure. The results are reported in Column 2 of Table 1.[26] We find suggestive evidence that 1-2 credit classes produce grades that have a 0.022 lower correlation than 3-4 credit classes.

Based on that estimate, Panel B of Figure 8 simulates how splitting 4-credit first year classes

[26] We cannot include professor-course fixed effects, but we can include professor fixed effects so that identifying variation is the same professors teaching both 1-2 credit classes and 3-4 credit classes. However, given that 1-2 credit classes could differ from 3-4 credit classes in unobserved ways other than the number of credits, we view these results as descriptive.

into two 2-credit classes would alter the share of missing exceptional students identified after the first year. We find that splitting classes would decrease the share of missing students in the top 10 percent by 15 percent. The magnitude of this is equivalent to 30 percent of the gains from delaying hiring by a year. If the school could find a way to split classes without decreasing the signal quality of individual classes, the intervention would have even larger effects decreasing the share of missing students in the top 10 percent by 23 percent. Altogether, this suggests that law schools could identify exceptional students earlier if they were to find ways to increase the total number of assessments, such as dividing a semester-long course into two parts and awarding separate grades at the end of each part of the course. Or, as an extreme policy, this suggests a benefit of moving from a semester system to a quarter system (e.g., Bostwick, Fischer, and Lang, 2021).[27]

Reweighting Classes. We next investigate whether there are differences in the signal quality of grades for different first year courses. To begin, we regress classroom level correlation on course fixed effects and semester-year fixed effects. Column 3 of Table 1 reports the difference between the most positive course fixed effect (Civil Procedure) and the most negative course fixed effect (Contracts). This indicates that Civil Procedure classes have 0.118 higher correlations than Contracts classes on average, controlling for semester-year effects.

Given the relatively high stability of the first year law school courses over time, it may be unrealistic to assume that a course would be replaced for the purpose of increasing the signal quality of grades. However, law schools could change the relative credits for certain courses.[28] We therefore assess the potential gains from changing the first year curriculum to increase the number of credits for the least noisy courses and decrease the number of credits for other courses. For this simulation, we assume that the two least noisy first year courses (Civil Procedure and Property) become two 3-credit classes, for a total of 6 credits for each course, and that the four other non-legal writing first year courses become 3-credit classes. Panel C of Figure 8 reports these results. We find that share of missing students in the top 10 percent would decrease by 13 percent in a world where first year courses had their credits changed to reflect the relative signal quality of their

[27] The gains from switching to 2-credit classes could be achieved either by actually splitting 4-credit classes into two 2-credit classes, or by switching from a semester system to a quarter system (which is used at a handful of law schools). Moreover, if the number of assessments is related to the signal quality of a grade in a class, then the fact that law schools courses are graded by a single final exam would imply that there may be gains by requiring semester length courses to conduct both midterm and final exams.

[28] At the school we study, each first year doctrinal course is 4 credits, but this is not true at all law schools.

grades in this way. The magnitude of this is equivalent to 25 percent of the gains from delaying hiring by a year.

4.3. Changes to Grading Systems

Some grading systems may be better at reducing the share of missing exceptional students after the first year of law school.[29] We assess this possibility by simulating two possible changes to grade systems for first year classes. For these exercises, we use a modified simulation framework because, as discussed below, we can impose counterfactuals directly on the microdata.

Grading Scale Distinctions. Law schools vary considerably in the number of grade distinctions that they use. For example, the law school we study has 30 grade distinctions (ranging from 2.5 to 4.3 in steps of 0.06), but Harvard Law School has just five distinctions (Honors, High Pass, Pass, Low Pass, and Fail). To assess the impact of changing the number of grades distinctions, we simulate the effects of reducing the number of grade distinctions from 30 to 20, 10, and 5. The 10 distinctions correspond to a single letter grade reported next to the GPA in our law school's grading system which helps for interpretation (F, D, C, C+, B-, B, B+, A-, A, and A+), and the 5 distinctions correspond to the letter value grades (F, D, C, B, A).[30]

Panel A of Figure 9 reports these results. We find that moving to much fewer grade distinctions would meaningfully increase the share of missing exceptional students. For example, moving from 30 grade distinctions to 5 grade distinctions would increase the share of missing exceptional students in the top 10 percent by 27 percent and in the top 1 percent by 65 percent. Given that this difference is even larger than the gains from delaying hiring by a year, law school administrators should carefully evaluate the number of grade distinctions in their grading system.

Grading During Emergencies. Schools may temporarily change their grading systems during emergencies. Most notably, due to the COVID-19 pandemic, all top ranked law schools moved to some form of pass-fail grading in the spring of 2020. This decision was driven by the recognition that emergencies can warrant a temporary pass-fail system to avoid evaluating students during challenging times. But there are also costs to moving to a pass-fail system. For instance, after the move to pass-fail grading during the spring of 2020, employers subsequently evaluated

[29] A literature has explored the relationship between grading policies and the signal quality of grades (e.g., Sabot and Wakenman-Linn, 1991). For instance, research has suggested that grade inflation can be used as a strategy to "help mediocre students" at the expense of its good students (Chan, Hao, and Suen, 2007).

[30] For these simulations, we use the grade conversions reported in Table A4 of the Appendix and assign the grade that would have been received under different counterfactuals.

first students based on their Fall 2019 grades.

However, the value of a temporary pass-fail system cannot be assessed relative to a situation that would produce normal grades. This is because the grades produced during emergencies could either be "noisy" if they are not as accurate of a measure of the students' performance as the grades would have been if the semester had not been disrupted, and they could be "biased" if some students are disproportionally affected by the emergency.[31]

To assess pass-fail systems in times of emergency, we run simulations where fall semester grades come directly from the data but spring semester grades are either omitted (pass-fail scenario) or contain noise and bias. We model noise by taking the students' actual second semester grades but simulate adding noise to those grades from a normal distribution of mean zero. We vary the standard deviation of the distribution of random numbers. We model bias by having some students be systematically disadvantaged.[32] To do this, we take random shares of students and remove a fixed amount (0.3) from their second semester GPA.[33] We vary the percent of students who have the fixed 0.3 GPA points removed from them.

Panel B of Figure 9 reports these results. The reported counterfactuals model both noise and bias at the same time but at different levels. The low, high, and very high level of noise has a standard deviation of noise of 0.05, 0.10, and 0.4 GPA points. The low, high, and very high level of bias add 0.3 GPA points to 10 percent, 20 percent, and 50 percent of students. This is admittedly just one of many possible ways to model the noise and bias we are interested in assessing. However, we find consistent results when using other specifications.[34]

If an emergency only increased noise, the average grade for a class would have to be 0.3

[31] The optimal Bayesian solution to noisier spring semester grades would be to continue assigning grades but give spring grades less weight when calculating GPA. While theoretically optimal, this faces practical limitations: Most notably, the optimal weight on spring grades would depend on the noise of these grades, which could only be estimated after the spring semester. Adding uncertainty about how grades during an emergency would be weighted would likely exacerbate the already stressful emergency environment. For this reason, we focus on simulations that reflect the more practical pass-fail solution that was implemented by many law schools.

[32] Another possible source of bias is cheating. Here, some students may instead be disproportionally "advantaged" by choosing to collude or cheat in some other way. We do not estimate the possible effects of increased cheating, but the analysis would follow the same structure as the way we estimated the potential impact for students who are disproportionally negatively affected. That said, many students already have many take home finals, and there is good reason to think that cheating will be more difficult this semester than normal (e.g., students are isolated from their peers and thus more isolated from co-conspirators for cheating on exams).

[33] Deducting 0.3 GPA points is admittedly arbitrary, and the actual average GPA deduction could be higher or lower.

[34] Figure A7 in the Appendix reports the results while using a continuous range of noisy and bias. Theoretically, if administrators knew the amount of noise and bias that would be in grades produced this semester, they could use the results in Figure A7 to help decide between either moving forward with grades or switching to a pass-fail system.

GPA points off from the true grade before a pass-fail system would produce fewer missing exceptional students in the top 10 percent. If an emergency only imposed bias on some students, it would take roughly 45 percent of students being negatively biased by 0.3 GPA points before a pass-fail system would produce fewer missing exceptional students in the top 10 percent. Overall, the results provide evidence that emergencies would need to dramatically lower the signal quality of grades before pass-fail assessment would produce fewer missing exceptional students.

5. Validation

The results in Section 4 illustrate how our approach could be used to improve the identification of exceptional students early in academic programs. These results, however, are contingent on our approach offering valid way to quantify the signal quality of grades. We thus conduct a series of tests that assess the validity our measure of the signal quality of grades.

5.1. Grade Dispersion

One advantage of correlation as a measure of signal quality is that it is independent of differences in grade dispersion (see Appendix B1 for a proof). This is a key feature because our setting had required means but allowed professors to assign grades with different variances.

To assess whether this holds in practice, we compare our correlation measure to two alternative measures of signal quality. First, we compare our correlation measure to an alternative measure based on the mean absolute difference between student grades in class c and student GPAs across all classes other than class c.[35] The mean absolute difference measure is intuitive in that it captures how much grades in class c generally deviate from other grades. Second, we compare our correlation measure to coefficients from classroom-specific regressions of student GPAs across all classes other than class c on student grades in class c.[36] This alternative measure uses a regression framework to analyze the extent to which variation in class c grades explains variation in grades in all other classes.

[35] Specifically, the alternative measure is:

$$MAD_c = \frac{1}{N}\sum_{i=1}^{N} abs(g_{ic} - \bar{g}_i)$$

[36] Specifically, the alternative measure for class c is β_c from the following regression:

$$\bar{g}_i^{\sim c} = \beta_c g_{ics} + \xi_{ics}$$

where $\bar{g}_i^{\sim c}$ is student i's GPA across all classes other than class c.

Figure 10 reports binscatters that compares our measure with these two alternative measures. In the binscatters, the measures of signal value of grades are on the x-axis and the standard deviation of grades at the classroom level are on the y-axis. Panel A reveals that our correlation measure is consistent across different levels of grade dispersion, but Panels B and C reveal that the absolute difference measure and the coefficient measure mechanically classifies classes with higher standard deviations as noisier. These illustrates that our correlation measure does not mechanically label classes with dispersed grades as inherently noisier.

5.2. Professor-Level Analysis

Many of the potential policy reforms we considered in Section 4 focused on identifying low and high noise professors. These investigations implicitly assumed that professor-level variation had a meaningful impact on overall signal quality, and that professors signal quality is consistent over time. We now directly explore both issues.

First, we directly investigate the relative importance of professors on signal quality through a simple variation decomposition. To do so, we regress the classroom-level correlations on professor, cohort, and course fixed effects. The results in Table 2 reveal that the professor fixed effects explain 28 percent of the variation in classroom-level correlations, whereas course fixed effects explain only 8 percent. This suggests that professor-level variation is not the only source of variation in grades' signal quality, but that it is a substantial component of the variation. That said, with our data it is impossible to test whether the across professor variation is due to differences in teaching style, exam writing, grading method, or other factors. Future research should explore the drivers of variation in signal quality across professors.

Second, we assess whether professors' signal qualities are consistent over time. To investigate this, we first consider whether professors assign grades with similar signal quality when they teach the same course in different semesters. Specifically, for every observation of a professor teaching a particular course, we compute the average correlation across all other observations of that professor teaching the same course. We then regress the correlation from the original course on this professor-course leave out mean.[37] Panel A of Figure 11 and Column 1 of Table 3 report the results. The strong positive relationship reveals that professor that have a high correlation in

[37] For all regressions in Table 3, we correct standard errors by clustering by professor. Moreover, because the independent variable in this regression and the other regressions reported in Part 5 is an estimate, we bootstrap standard errors.

one classroom generally have high correlations when teaching the same course.

To investigate this further, we next consider whether professors assign grades with similar signal quality when they teach different courses. In practice, the other courses taught by professors are almost always upper level courses, because professors rarely teach multiple first year classes (see Appendix Table A3). As a measure of the signal quality of an upper level classroom, we use the correlation between grades in a particular upper level classroom and the average GPA in all other classes except that classroom. We then average these correlations across all upper level courses taught by a particular professor to yield a measure of that professor's signal quality in upper level classes. Finally, we regress the correlation from the original first year course on these upper level average correlations. Panel A of Figure 11 and Column 2 of Table 2 reports the results. The strong positive relationship reveals that professors who have high correlation in a first year course also generally have higher correlations in upper level courses.

5.3. Comparison to Pre-Law School Student Achievement

To investigate whether our measure is likely capturing some degree of general aptitude, we next examine whether higher signal quality grades are also more closely related to measures of pre-law school student achievement. For this, we employ data on students' LSAT scores and college GPAs. Specifically, for every first year classroom, we compute the correlation between students' grades in that classroom and either the students' LSAT scores or college GPAs. This yields an alternative measure of signal quality—grades that are highly correlated with these measures of pre-law school student achievement.

To compare these two measures, we first regress the alternative measure (i.e., correlations with LSAT scores or college GPAs) on our original measure (i.e., correlations with upper level GPA). Panels B and C of Figure 11 and Columns 3 and 4 of Table 3 report the results. These results reveal that grades that are more correlated with upper level GPAs also are more correlated with LSAT scores and college GPAs. However, the R-squared statistics for these regressions are quite small. This implies that there are plenty of classrooms with grades that are highly correlated with upper level GPA but loosely correlated with pre-law school achievement, and vice versa.

To explore this issue further, we next regress upper level GPA on our data on pre-law school student achievement. Columns 1, 2, and 3 of Table 4 report the results. These results reveal that LSAT scores explain 14 percent of the variation in upper level GPAs, college GPAs explain 13 percent of the variation, and together they explain 20 percent of the variation. Columns 4 and

5 of Table 4 add first year grades to these regressions. The results reveal that adding first year grades to the regression explains an additional 49 to 60 percent of the variation of upper level GPAs.[38]

Taken together, these results provide evidence for the need to improve the signal quality of grades. If our measure were nearly identical to pre-law measures of achievement, then employers could just predict upper level performance with LSAT scores and college GPA, making the signal quality of first year grades less relevant.[39] But because first year grades provide considerable independent information about eventual achievement, they are worthy of attention by employers and thus also policymakers.

5.4. Changes in Effort

Our framework assumes that student effort is either constant, and thus captured by α_i, or idiosyncratic, and thus captured by η_{ic}. One potential concern with this assumption is that the students may only make it into the top of the class at the end of law school because students change their behavior after the first year. For example, the most exceptional students could be strategically working hard in their first year to get a prestigious clerkship but then decrease their effort after obtaining selective opportunities. Relatedly, some strong students could have responsibilities during their second or third year—like law review membership—that cause their effort on classes to decrease.

Although it is possible that students change their behavior to some degree, we do not expect behavioral changes after the first year to be large.[40] This is in part because there are several incentives for top students to continue working hard after their first year, including that graduation honors are based on all grades and hiring for the most elite clerkships can take place after graduation. Furthermore, Panel B of Figure 1 revealed that most top students after the first year stayed in the top of their class at graduation, meaning that the majority of law students are not

[38] As a further analysis, Appendix Table A5 regresses first year grades—individually and overall—on LSAT scores and college GPAs, revealing that these measures explain between 10 and 20 percent of the variation in first year grades.

[39] Technically, correlation is not transitive. As such, in this scenario, it is still possible that LSAT scores and college GPA are poor predictors of upper level GPA. However, non-transitive correlation is empirically rare, especially in cases such as this where all variables relate to the same underlying factor—in this case, student ability.

[40] Research in other settings has found that even grading changes that alter employer behavior do not necessarily change student behavior (Hansen, Hvidman, and Sieversten, 2021).

strategically only working early in law school to obtain a job.

To further assess the potential effects of changes in effort, we investigate the signal quality of grades in different semesters. To do so, we calculate the correlation of grades in one semester with grades in all other semesters for the different thresholds of exceptional students based on first year grades. For example, for students in the top 10 percent after the first year, we calculate the mean grade in the first semester and the mean grade in all other semesters and estimate the correlation between them. If changes in effort after the first year artificially create the appearance of missing exceptional students, then upper level grades of the top students after the first year would not be as informative as their grades were in the first year. Panel D of Figure 11 reports these results. We find that the correlation by semester does not vary greatly (from a low of 0.76 in the last semester to a high of 0.85 in the third semester), and this is true for all the different groups of exceptional students. This suggests that behavioral changes after the first year have limited effects on our measure of signal quality.

Another way to think about this potential concern with our measure is that lower stakes in upper level classes (or other changes after the first year) could make upper level grades fundamentally different from first year grades. For example, suppose students who excel under high stakes are very different from those who stand out under low stakes. Because many employers are presumably more interested in hiring those who excel under high stakes, first year grades may be more relevant to hiring decisions than upper level grades. If this were the case, then correlation with upper level GPA would not be a desirable measure of signal quality because upper level grades themselves would not be that relevant.

To investigate this, we examine the extent that first year grades that best predict success in (supposedly lower stakes) upper level classes also best predict performance in other (supposedly higher stakes) first year classes. Specifically, for every first year class, we compute the correlation between grades in that class and average grades across all other first year classes. This yields an alternative metric of signal quality that measures the capacity of grades to predict performance in high stakes courses. It is an inherently noisier measure because there are fewer first year classes to average over, but it is not potentially confounded by decreases in effort after the first year. We then regress this alternative measure on our preferred correlation measure. Panel E of Figure 11 and Column 5 of Table 3 reports the results. The strong positive relationship implies that the same first year grades that best predict success in upper level classes also best predict performance in

other first year classes. As such, this provides additional strong evidence that potential changes after the first year have limited effects on our measure of signal quality.

5.5. Multiple Dimensions of Achievement

It may be the case that grades in some first year classes have a low correlation with upper level GPA because they are measuring a less evaluated aspect of student achievement. For example, suppose most classes primarily evaluate legal analysis skills but some classes mostly evaluate legal writing skills. Even if a class perfectly identifies the best legal writers, its grades may appear noisy according to our measure because they may have little correlation with the majority of classes. Because a class that measures an uncommon skill well may be more informative than another class that does an average job evaluating the same skills as most other courses, it is important that a measure of the signal quality of grades does not punish classes for evaluating another dimension of student achievement. We evaluate this possibility in three ways.

First, if low correlation professors were measuring a less evaluated aspect of student achievement, grades from low correlation professors may be more similar to one another than to grades from high correlation professors. To test this possibility, we consider pairs of first year professors who have taught the same students. For each professor pair, we calculate the between-professor correlation in student grades for the shared students. We then evaluate how these pairwise correlations compare by professors' overall signal quality. If low correlation professors are measuring a less evaluated skill, pairwise correlations should be higher when both professors are in the bottom quartile. Panel A of Figure 12 plots the distributions of pairwise correlations for (1) pairs where the other professor is also in the bottom quartile, and (2) pairs where the other professor is not in the bottom quartile. The results show that pairwise correlations between two bottom-quartile professors are stochastically dominated by correlations between one bottom quartile professor and other professors. This implies that grades assigned by bottom quartile professors are actually more similar to grades assigned by top three quartile professors than to grades assigned by other bottom quartile professors. These results suggest that low correlation professors are not just measuring a less evaluated aspect of student achievement.

Second, although these results reveal that low correlation professors are not measuring the same less evaluated aspect of student achievement, low correlation professors could all be

evaluating different overlooked aspects of achievement.[41] This possibility can be analyzed by examining cases where a professor teaches the same student multiple times. For this analysis, we restrict to professor-students where a professor taught a student in a first year class and at least one upper level class.[42] We then compute within-professor-student correlations for each professor.[43] We then assess the relationship between this professor level within-professor-student correlation with the professor level average correlation with upper level GPA.[44] Panel B of Figure 12 reports the results. If professors with low overall correlation were consistently measuring an uncommon skill, these within-professor-student correlations should be similar for professors with low and high overall correlations. But the opposite is true. Column 6 of Table 3 regresses these within-professor-student correlations on the professor's mean correlation. Here, the coefficient is approximately 1, implying that grades that better predict overall upper level performance also better predict future grades assigned by the same professor.

Third, if different skills were evaluated differently across classes, we might observe systematic differences in grades across different first year courses. For example, if grades in Constitutional Law and Property are more related to writing skills than other first year courses, these grades would be more correlated to each other than other courses. To assess whether different first year courses evaluate skills differently, we perform a Principal Component Analysis ("PCA") of all doctrinal first year grades.[45] Panel C of Figure 12 reports the eigenvalue for each

[41] It is somewhat difficult to imagine employers keeping track of how different grades evaluate many different skills; as such, in this scenario, it may still make sense to reduce the number of low correlation classes in the first year.

[42] Students who do well in the first year typically also do well in upper level classes (see Figure 1), so there will be a positive correlation between these grades. However, if low correlation professors are consistently picking out a different dimension of achievement, then within professor-student correlations should be similar for low and high correlation professors.

[43] Figure A8 in the Appendix reports a similar analysis that focuses on the relationship between upper level grades and first year grades for professors of different mean correlation. In particular, it reports a binscatter of the relationship between upper level grades and first year grades, separately by professors' average correlation in first year classes. The figure provides further evidence that low correlation professors are worse at predicting their future assessment of the same students than high correlation professors.

[44] We restrict to professors who have had at least 50 students taking one of their upper level classes to yield within professor-student correlations that are reliably estimated.

[45] PCA is a standard method for assessing the dimensionality of data and for collapsing a large number of correlated variables into a smaller number of orthogonal components preserving as much information from the larger set as possible. The idea here is to generate linear combinations of first year courses to generate fewer components that reflect different dimensions of achievement, where each component is unrelated to the next. PCA is possible here because, unlike above where all students do not have first year classes by all professors, each student takes all first year courses.

additional component in what is known as a scree plot.[46] If all courses weight skills approximately equally, a PCA will find that variation in grades can be mostly explained by a single component. Conversely, if some courses were assessing a different skill, a PCA would extract at least two significant components. Here, there is a large break in the eigenvalue after the first component, and each additional component has an eigenvalue below 1. These findings suggest the meaningful variation in the data is one-dimensional,[47] suggesting that all first year courses evaluate approximately the same set of skills. This provides further evidence that low correlation courses are not just measuring less evaluated aspects of student achievement.

5.6. Absolute Differences in Student Rank

Finally, measuring the share of missing exceptional students is a way to quantify the economic importance of improving the signal quality of grades. However, this method has a notable limitation: because it only depends on the number of students who cross from below a class rank threshold early in law school to above that threshold at graduation, the measure does not reflect how far students move to cross this threshold. If all missing exceptional students were only slightly below a threshold early in law school, their movement above that threshold may not be that important. Yet, if missing exceptional students are making large movements, this would suggest they really are missing out on selective opportunities that are awarded early in law school. To address this limitation, we also use average absolute rank change as an alternative measure of signal quality. Specifically, we compute the absolute difference between each student's class rank at graduation and class rank early in law school and average this across all students. Although this measure does not rely on crossing rank thresholds, it yields results that are qualitatively similar to results using the missing exceptional student metric. Appendix Table A6 reports results using this alternative measure and reveals results that are consistent with our primary measure.

[46] Eigenvalues for each component reflect the share of total variance explained by each component and the standard Kaiser Criterion says that components with eigenvalues greater than one capture meaningful variation while components with eigenvalues less than one contain insignificant information (Kaiser, 1960).

[47] With any PCA, the components are a weighting of the independent variables. It is thus possible to generate a "loading" of components, which can be interpreted as the optimal weighting of each of the first year classes if all of the course grades were to be represented by a single component. We find an arguably narrow range of weighting, from a loading of 0.36 for Contracts to a loading of 0.41 for Civil Procedure. Whereas Panel F provides evidence that student achievement is one dimensional, this range of loadings is evidence that each of the first classes have similar abilities to explain that dimension.

6. Conclusion

Many of the most important selective academic and professional opportunities for law students are awarded based on grades from just one year of a three year program. As a result, many exceptional students miss out on career changing opportunities. Using transcript data from one law school over a 40 year period, we document large gains in identifying exceptional students if law schools were to change certain personnel, course, and grading policies. Our findings provide motivation and a blueprint for how law schools could leverage their internal records to ensure that fewer exceptional students miss out on selective opportunities.

However, there are several important caveats about these findings. First, we use data from a single law school, and it is possible that some of our findings do not generalize to other law schools. As such, other schools should replicate our analysis with their own data to determine how they can improve the signal quality of their grades. Second, improving the signal quality of grades may have distributional consequences that are undesirable. One advantage of noisy grades is that they obscure differences between students, and some schools may prefer to retain noisy grades for this reason alone. Third, the costs and benefits of improving the signal quality of grades may not be the same for all law schools. Some schools, for example, may decide that the additional costs to students and faculty of increased attention to grading, such as dividing 4-credit courses into two 2-credit courses, are not worth the benefit they provide in identifying exceptional students. Fourth, there are many policies that may improve the signal quality of grades—for instance, the format of the final exam—that we were unable to assess with our data. Finally, improving the signal quality of grades should not be the only goal in student assessment. Efforts should also be taken to ensure that grades accurately reflect students' mastery of a material in an individual course.

With those caveats in mind, there are at least four reasons that academic departments should use the tools we have developed to improve the identification of exceptional students. First, improving the signal quality of grades is important as a matter of fairness between students. There will always be some amount of randomness in the grade a student receives. But if some students are taught by a higher share of professors that assign grades with low signal quality, they may be predictably less likely to receive selective opportunities. Second, noisy grades may negatively change student behavior because they weaken the relationship between effort and outcome (e.g., Becker, 1982). Improving the signal quality of grades can thus increase students'

incentives to apply themselves during academic programs. Third, if an academic department evaluated their students more accurately, employers may be more likely to hire students from that department. In this way, improving the signal quality of grades may not just reshuffle which students within a particular program receive opportunities, but it could actually increase the employment opportunities for all students in a department. Finally, improving the signal quality of their grades can help improve the matching of students and jobs (e.g., Roth, 1984; Roth and Xing, 1994). When students receive career opportunities based on incomplete grades, improving the signal quality of grades leads to students being hired into jobs that better match their abilities and potential. This improved matching could in turn have direct effects on social welfare.

References

Altonji, Joseph G. and Charles R. Pierret. 2001. "Employer Learning and Statistical Discrimination." *Quarterly Journal of Economics*, 116(1): 313-350.

Ambuehl, Sandro and Vivienne Groves. 2020. "Unraveling Over Time." *Games and Economic Behavior*, 121: 252-264.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics,* 114(2): 533-575.

Arcidiacono, Peter. 2004. "Ability Sorting and the Returns to College Major." *Journal of Econometrics,* 121(1): 343-375.

Avery, Christopher, Christine Jolls, Richard A. Posner, and Alvin E. Roth. 2007. "The New Market for Federal Judicial Law Clerks." *University of Chicago Law Review*, 74(2): 447-486.

Becker, William E., Jr. 1982. "The Educational Process and Student Achievement Given Uncertainty in Measurement." *American Economic Review*, 72(1): 229-236.

Birdsall, Chris, Seth Gershenson, and Raymond Zuinga. 2020. "The Effects of Demographic Mismatch in an Elite Professional School Setting." *Education Finance and Policy,* 15(3): 457-486.

Bleemer, Zachary and Aashish Mehta. 2021. "Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major." *American Economics Journal: Applied Economics,* 14(2): 1-22.

Bostwick, Valerie, Stefanie Fischer, and Matthew Lang. 2021. "Semesters or Quarters? The Effect of the Academic Calendar on Postsecondary Student Outcomes." *American Economic Journal: Economic Policy,* 14(1): 40-80.

Chan, William, Hao Li, and Wing Suen. 2007. "A Signaling Theory of Grade Inflation." *International Economic Review,* 48(3): 1065-1090.

Chilton, Adam, Jonathan Masur, and Kyle Rozema. 2021. "Rethinking Law School Tenure Standards." *Journal of Legal Studies*, 50(1): 1-34.

Clark, Jessica L. 2014. "Grades Matter; Legal Writing Grades Matter Most." *Mississippi College Law Review*, 32(3): 375-418.

Colker, Ruth, Ellem Deason, Deborah Merritt, Abigail Shoben, Monte Smith. 2017. "Formative Assessments: A Law School Case Study." *University of Detroit Mercy Law Review*, 94(3): 387-428.

de Haan, Thomas and Jona Linde. 2018. "'Good Nudge Lullaby': Choice Architecture and Default Bias Reinforcement." *Economic Journal,* 128(610): 1180-1206.

Devine, Patricia G., Patrick S. Forscher, Anthony J. Austin, and William T.L. Cox. 2012. "Long-term Reduction in Implicit Race Bias: A Prejudice Habit-Breaking Intervention." *Journal of Experimental Social Psychology,* 48(6): 1267-1278.

Engel, Samuel P. 2018. "The Economics of Law School: Employment Prospects and Market Inefficiencies." *Mississippi Law Journal,* 87(4): 501-576.

Farber, Henry S. and Robert Gibbons. 1996. "Learning and Wage Dynamics." *Quarterly Journal of Economics*, 111(4): 1007-1047.

Fredriksson, Peter, Lena Hensvik, and Oskar Nordstrom Skans. 2018. "Mismatch of Talent: Evidence on Match Quality, Entry Wages, and Job Mobility." *American Economic Review*, 108(11): 3303-3338.

Ginsburg, Tom and Jeffrey A. Wolf. 2003. "The Market for Elite Law Firm Associates." *Florida State University Law Review,* 31: 909-963.

Ginsburg, Tom and Thomas J. Miles. 2015. "The Teaching /Research Trade-Off in Law: Data from the Right Tail." *Evaluation Review*, 39(1): 46-81.

Golsteyn, Bart H. H., Arjan Non, and Ulf Zölitz. 2021. "The Impact of Peer Personality on Academic Achievement." *Journal of Political Economy* 129(4): 1052-1099.

Grant, Darren. 2007. "Grades as Information." *Economics of Education Review* 26(2): 201-214.

Hansen, Anne Toft, Ulrik Hvidman, and Hans Henrik Sievertsen. 2021. "Grades and Employer Learning." *IZA Working Paper No. 14200*

Haruvy, Ernan, Alvin E. Roth, and M. Utku Ünver. 2006. "The Dynamics of Law Clerk Matching: An Experimental and Computational Investigation of Proposals for Reform of the Market." *Journal of Economic Dynamics and Control,* 30(3): 457-486.

Hibbard, Judith H., Jean Stocard, and Martin Tusler. 2003. "Does Publicizing Hospital Performance Stimulate Quality Improvement Efforts." *Health Affairs*, 22 (2): 84-94.

Ho, Daniel and Mark G. Kelman. 2014. "Does Class Size Affect the Gender Gap? A Natural Experiment in Law." *Journal of Legal Studies,* 43(2): 291-321.

Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of Economics*, 115(4): 1239-1285.

Hvidman, Ulrik, and Hans Henrik Sievertsen. 2021. "High-Stakes Grades and Student Behavior." *Journal of Human Resources*, 56(3): 821-849.

Kaiser, Henry. 1960. "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement*, 20(1): 141-151.

Kagel, John H., and Alvin E. Roth. 2000. "The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment." *Quarterly Journal of Economics,* 115(1): 201-235.

Kamenetz, Anya. 2015. "Who Are The 'Gifted And Talented' And What Do They Need?" *National Public Radio*. https://www.npr.org/sections/ed/2015/09/28/443193523/who-are-the-gifted-and-talented-and-what-do-they-need

Kuehn, Robert R. and David R. Moss. 2019. "A Study of the Relationship between Law School Coursework and Bar Exam Outcomes." *Journal of Legal Education*, 68(3): 624-650.

Li, Hao, and Sherwin Rosen. 1998. "Unraveling in Matching Markets." *American Economic Review,* 88(3): 371-387.

Merritt, Deborah and Barbara F. Reskin. 1997. "Sex, Race, and Credentials: The Truth about Affirmative Action in Law Faculty Hiring." *Columbia Law Review*, 97(2): 199-311.

McIntyre, Frank and Michael Simkovic. 2017. "Timing Law School." *Journal of Empirical Legal Studies,* 14(2): 258-300.

National Association for Law Placement. 2022. Perspectives on Law Student Recruiting, 2020-21.

National Association for Gifted Children. 2021. "Twice Exceptional Students." https://www.nagc.org/resources-publications/resources-parents/twice-exceptional-students

Negin, Gary A. 1981. "The Effects of Test Frequency in a First-Year Torts Course." *Journal of Legal Education,* 31(3): 673-376.

Niederle, Muriel, and Alvin E. Roth. 2003. "Unraveling Reduces Mobility in a Labor Market: Gastroenterology with and without a Centralized Match." *Journal of Political Economy,* 111(6): 1342-1352.

Ostrovsky, Michael and Michael Schwarz. 2010. "Information Disclosure and Unraveling in Matching Markets." *American Economic Journal: Microeconomics,* 2(1): 34-63.

Pope, Devin G., Joseph Price, and Justin Wolfers. 2018. "Awareness Reduces Racial Bias." *Management Science,* 64(11): 4988-4995.

Propper, Carol, Matt Sutton, Carolyn Whitnall, and Frank Windmeijer. 2010. "Incentives and Targets in Hospital Care: Evidence from a Natural Experiment." *Journal of Public Economics*, 94(23): 318-355.

Rebitzer, James B. and Lowell J. Taylor. 1995. "Efficiency Wages and Employment Rents; The Employer-Size Wage Effect in the Job Market for Lawyers." *Journal of Labor Economics,* 13(4): 678-708.

Roth, Alvin E. 1984. "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory." *Journal of Political Economy,* 92(6): 991-1016.

Roth, Alvin E. 2012. "Marketplace Institutions Related to the Timing of Transactions: Reply to Priest." *Journal of Labor Economics,* 30(2): 479-494.

Roth, Alvin E., and Xiaolin Xing. 1994. "Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions." *American Economic Review,* 84(4): 992-1044.

Sander, Richard and Jane Bambauer. 2012. "The Secret of My Success: How Status, Eliteness, and School Performance Shape Legal Careers." *Journal of Empirical Legal Studies,* 9(4): 893-930.

Sabot, Richard, and John Wakeman-Linn. 1991. "Grade Inflation and Course Choice." *Journal of Economic Perspectives*, 5 (1): 159-170.

Samida, Dexter. 2004. "The Value of Law Review Membership." *University of Chicago Law Review,* 71(4): 1721-1748.

Schwarcz, Daniel, and Dion Farganis. 2017. "The Impact of Individualized Feedback on Law Student Performance." *Journal of Legal Education*, 67(1): 139-175.

Simkovic, Michael and Frank McIntyre. 2014. "The Economic Value of a Law Degree." *Journal of Legal Studies*, 43(2): 249-289.

Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics*, 87(3): 355-374.

Stinebrickner, Todd, and Ralph Stinebrickner. 2012. "Learning about Academic Ability and the College Dropout Decision." *Journal of Labor Economics,* 30(4): 707-748.

Stiglitz, Joseph. 1975. "The Theory of 'Screening,' Education, and the Distribution of Income." *American Economic Review*, 65(3): 283-300.

Thaler, Richard and Sunstein Cass. 2009. *Nudge: Improving Decisions about Health, Wealth and Happiness*. London: Penguin.

Thomas, James. 2019. "The Signal Quality of Grades Across Academic Fields." *Journal of Applied Econometrics,* 34(4): 566-587.

Urquiola, Miguel, and Eric Verhoogen. 2009. "Class-Size Caps, Sorting, and the Regression-Discontinuity Design." *American Economic Review*, 99(1): 179-215.

Zafar, Basit. 2011. "How Do College Students Form Expectations?" *Journal of Labor Economics,* 29(2): 301-348.