The Effects of Reputational Information

Stuck in Place? A Field Experiment on the Effects of Reputational Information on Student Evaluations

James Chu, *Stanford University*
Guirong Li, *Henan University*
Prashant Loyalka, *Stanford University*
Chengfang Liu, *Peking University*
Leonardo Rosa, *Stanford University*
Yanyan Li, *Henan University*

Studies suggest that students' prior performance can shape subsequent teacher evaluations, but the magnitude of reputational effects and their implications for educational inequality remain unclear. Existing scholarship presents two major perspectives that exist in tension: do teachers primarily use reputational information as a temporary signal that is subsequently updated in response to actual student performance? Or do teachers primarily use reputational information as a filter that biases perception of subsequent evidence, thus crystallizing student reputations and keeping previously poor-performing students stuck in place? In a field experiment, we recruited a random sample of 832 junior high school teachers from the second-most populous province of China to grade a sequence of four essays written by the same student, and we randomly assign both the academic reputation of the student and the quality of the essays produced. We find that (1) reputational information influences how teachers grade, (2) teachers rely on negative information more heavily than positive information, and (3) negative reputations are crystallized by a single behavioral confirmation. These results suggest that students can escape their prior reputations, but to do so, they must contradict them immediately, with a single confirmation sufficient to crystallize a negative reputation.

To what extent does reputational information shape subsequent evaluations of student performance? When do students' reputations crystallize, leading teachers to apply them even in the face of disconfirming evidence? If students' past performance shapes how they are subsequently evaluated, then students perceived

We would like to thank Jeremy Freese, David Grusky, Michelle Jackson, David Pedulla, Scott Rozelle, and Robb Willer for helpful comments on earlier versions of the manuscript. James Chu was supported by the National Science Foundation Graduate Research Fellowship under grant number DGE-114747. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation. *Please address correspondence to James Chu, Stanford University, E-mail: jchu1225@stanford.edu; or Guirong Li, Henan University, E-mail: guirong1965@163.com.*

Social Forces 98(4) 1578–1612, June 2020
doi: 10.1093/sf/soz097
Advance Access publication on 10 July 2019

to be low performers may become "stuck in place," limited by negative reputations regarding their educational capacities and performance (Brophy and Good 1974; Rosenthal and Rubin 1978). Of particular concern are *crystallized reputations*, where teachers and other evaluators rely so heavily on prior performance that they discount contrary evidence. Because such reputations are resistant or even immune to updating, students have little incentive to improve or continue to do well. Moreover, to the extent that initial performance is based on chance (Bol, de Vaan, and van de Rijt 2018), crystallized reputations entrench the position of a fortunate few who happen to have favorable initial performance while penalizing equally-talented students with fewer opportunities to succeed (DiPrete and Eirich 2006; Salganik and Watts 2008; van de Rijt et al. 2014).

There remains ambiguity over the power of reputational information to shape subsequent evaluations, both in educational settings and more generally. One perspective is that reputational information serves as a *temporary signal*, which evaluators use to make initial judgments but jettison as they have more opportunities to observe performance (Kollock 1998; Podolny 2005; Smith, Jussim, and Eccles 1999). A competing perspective is that reputational information affects subsequent evaluation by acting as a *filter*, biasing teacher perceptions of subsequent evidence (Correll and Benard 2006; Knobloch-Westerwick and Meng 2009; Nickerson 1998). If reputational information is used more as a temporary signal, the power of reputational information is limited: it is unlikely to crystallize and will diminish in influence in subsequent evaluations. If reputational information is used more like a filter, reputations are likely to crystallize: their effects persist even when teachers are exposed to student performance that contradicts their expectations.

These competing intuitions remain unresolved in part because, with exception to a handful of observational studies (e.g., Kelly 2008), little empirical work has directly examined how the past performance of students might affect subsequent evaluation. Although existing experimental work does show that evaluations are shaped by teacher expectations, these studies focus on teacher evaluations at a single point in time, leaving unknown how the expectations of teachers change as they observe additional student performance. Finally, existing work tends to hold student performance constant to observe how teachers grade when encountering different student identities. Without manipulations of underlying performance, it is difficult to assess how evaluators respond to confirmatory versus contradictory evidence.

In this study, we present field experimental results that enable us to observe how evaluators react when faced with new information on student performance. In our study, teachers are asked to grade multiple essays written by the same student. We randomly assign both the student's reputation (as someone who performed well or poorly in the past) and the quality of essays produced, allowing us to observe how teachers respond to confirmatory or contradictory evidence. We also include a control group where the prior performance of the student is not revealed. Our experiment is conducted among 832 junior high teachers in China—a context where the sharing of prior student grades is both customary and public. A novel feature of our study is that enumerators traveled to

approximately 300 schools across the entire province to recruit teachers and conduct our experiment. As such, not only did teachers believe they were grading real student essays, they were doing so in a setting that approximates day-to-day grading.

Our results show that reputational information shapes subsequent teacher judgment, and reputational information functions more (on average) as a signal than a filter, such that students are able to escape from their reputations when given opportunities to demonstrate their performance. However, our results suggest two critical revisions to this perspective. First, teachers rely on negative reputational information more heavily than positive ones, such that students who are known as having low prior performance are more likely to be stuck in place. Second, negative reputations that receive initial behavioral confirmation—a student who is presumed to perform poorly produces a poorly written essay—become crystallized: teacher reliance on reputational information remains unchanged even after grading essays that contradict this expectation. This implies that students can escape from their reputations. However, with initial behavioral confirmation, a reputation becomes more persistent and likely to keep students stuck in place. In this sense, one of our major contributions is to suggest how existing perspectives are cases of a more contingent model: one where individuals rely heavily on reputations after initially putting them to the test.

Reputational Information: A Temporary Signal or Filter?

People regularly use reputational information[1] to decide whom to trust (Feinberg et al. 2012), what products to purchase (Diekmann et al. 2014), or whose ideas to cite (Merton 1968). Perhaps because reputational information reduces the time and energy required for individuals to make evaluations (Zahra and George 2002), teachers also rely on how well students have done in the past to evaluate their present performance (Allen 2005; Kelly 2008). Despite consensus that reputational information can shape subsequent evaluation, social scientists remain divided over the magnitude of reputational effects and their implications for educational inequality (Botelho and Abraham 2017; Jussim and Harber 2005). One line of scholarship asserts that teacher reliance on reputational information sets in place self-fulfilling prophecies (Rosenthal and Jacobson 1968). This prior research focuses on how assessments of performance are shaped by teacher expectations about student caste (Hanna and Linden 2012), immigrant status (Sprietsma 2013), or race (Botelho, Madeira, and Rangel 2015). Outside the classroom context, employers' expectations of performance are often based on stereotypes. Existing laboratory and field experiments show that expectations of better or worse performance often lead to biased judgment as employers evaluate candidates in a way that is consistent with their expectations (Correll, Benard, and Paik 2007; Pager 2003; Pager and Shepherd 2008; Pedulla 2016).

However, recent work raises questions about the power of reputational information to create self-fulfilling dynamics (Elashoff and Snow 1971; van de Rijt 2019; Wineburg 1987). Even if teachers rely on their expectations to make

initial judgments, teachers update their expectations as they have more opportunities to observe student performance (Swann and Ely 1984). Thus, even if a teacher is initially biased when grading a student with a reputation for low performance, the strength of this expectation might diminish with each subsequent evaluation. Moreover, evaluations may be biased by but need not confirm expectations. For instance, expectations may lead to subsequent disconfirming evaluations, such as when evaluators who expect low performance judge subsequent performances more leniently: "good work is evaluated much more favorably when it follows poorer quality work (Guskey and Bailey 2001: 34)." These possibilities collectively challenge the power of reputational information to lead to self-fulfilling prophecies.

Underlying this debate is a theoretical divide over how reputational information shapes subsequent evaluation. One perspective treats reputational information as a *temporary signal* that helps individuals make judgments (Kollock 1998; Milinksi, Semmann, and Krambeck 2002; Podolny 2010). Each time evidence is inconsistent with reputational information, individuals adjust their expectations to fit the evidence (Chamley 2004; Gilboa and Schmeidler 1993). Because they update their expectations according to new evidence each time, individuals will rely less on reputations when given sufficient information. Assuming teachers have multiple opportunities to assess student performance, inequalities are less likely to become entrenched, and self-fulfilling prophecies are far less likely to occur.

This view is implicit in theories of statistical discrimination, where evaluators use signals like race or nationality to make judgments about job candidates (Altonji and Pierret 2001) or green cards (Rissing and Castilla 2014) but jettison these signals when better evidence is available. Even if teachers assign grades that do not reflect underlying student performance initially, this is only an *informational* bias that will resolve as they receive more opportunities to observe student performance. Thus, the solution to biased assessment is to give evaluators more information, such as giving them additional opportunities to assess performance.

A competing perspective is that reputational information *filters* subsequent evidence. This perspective suggests that reputational information shapes the evidence that teachers perceive (Knobloch-Westerwick and Meng 2009; Nickerson 1998). Put differently, reputational information colors the very process of judgment itself, creating *cognitive* biases that cannot be addressed with more information (Correll and Benard 2006). For instance, teachers may operate under confirmation bias. Rather than updating their expectations, individuals may selectively focus on mistakes or problems when grading a student with a reputation for low performance, thus confirming this reputational information (e.g., Darley and Gross 1983).

These two perspectives imply divergent implications for educational inequality. If reputational information serves as a filter, reputations are liable to crystallize: the influence of reputational information remains constant even when individuals receive disconfirming evidence (Rabin and Schrag 1999). If true, reliance on reputational information could be keeping students stuck in place, and

the use of reputations in schooling systems may entrench differences in performance among students. By contrast, if reputational information serves as a temporary signal, students will be able to overcome inaccurate reputations by demonstrating contrary performance. In this case, a student can redeem a negative reputation given sufficient opportunities to demonstrate exemplary performance.

Although these two perspectives imply divergent implications, existing work has not adequately established which perspective most closely aligns with how teachers rely on reputational information to make evaluations. More importantly, these perspectives are not mutually exclusive. It is possible that reputational information acts more like a filter in some situations and more like a temporary signal in other situations. This point underscores the need to further identify circumstances where teachers rely on reputational information more as filters versus temporary signals.

Empirical Context: School Grading and Reputational Information in China

Our study examines the effects of reputational information on the most common form of school assessment: grades. The grades that students receive in school have far-reaching implications (Fleming 1999; Wormeli 2006: 90). Grades are a measure of a student's academic merit, both in absolute terms and in relation to peers (Marsh 1986; Marzano 2006). Students, parents, teachers, and school administrators rely on grades to allocate opportunities and effort (Guskey and Bailey 2001; Harlen and James 1997). Grades are also used to justify ostensibly merit-based allocations of life-altering opportunities like scholarships and college admissions. If inaccuracies in teacher grading persist, school systems may unfairly and incorrectly track students and misallocate financial aid (Callahan 2005; Hoxby 2007). Moreover, teachers may misallocate their efforts, which may in turn affect long-term student outcomes as varied as college attendance, future salaries, and teenage pregnancy (Chetty et al. 2014). Indeed, the importance of grades extends outside of school: absent other clear signals of ability, employers rely on applicant grades to make hiring and wage decisions (e.g., Daley and Green 2014).

We study how reputational information shapes how teachers grade in China, the largest public education system in the world (by enrollments). In general, education contexts across different countries have converged in structure (Meyer and Ramirez 2009), and the practice of student grading is similar across most country contexts: a primary duty of teachers is to grade student assignments, with the general expectation that they do so in a fair and unbiased manner (Baker 2014). Almost all education systems aggregate grades into grade point averages, a well-established device that communicates the prior performance of students. Similar to the education systems of many developed and developing countries (Clark 1986; Shavit and Blossfeld 1993), policymakers, school administrators, teachers, parents, and students in China rely on grades to make a host

of educational decisions. For example, school administrators track students into classes by prior grades: top scoring students are placed together in elite classes and given the best teachers and resources (Postiglione 2015).

Although these shared institutional arrangements and uses of grading make generalization from the Chinese context possible, we also acknowledge several features of Chinese schooling that set it apart from contexts like the United States. Although teachers in all schooling systems likely draw on reputational information to make grading decisions, the Chinese context is one where reputational information is public and well-established. As students transition across school years, administrators give teachers a roster with the accompanying grades of students. After major assignments or examinations, teachers post student grades publicly (Liu, Ross, and Kelly 2000). The public nature of academic reputations simplifies its experimental operationalization, as the sharing of student grades is customary. However, the high-stakes nature of grading and routinized usage of reputational information likely affects how teachers rely on reputational information, which we consider more fully below in our discussion of the external validity of the study.

Data and Methods

Sampling

Our field experiment was conducted in the second-most populous province in China. Henan province has a population of over 94 million (as of the 2010 Census), and it was ranked 22 out of 31 Chinese provinces in terms of GDP per capita (National Bureau of Statistics 2015). The province is becoming increasingly urbanized, but 55 percent of Henan's population lives in rural areas (Henan Statistical Yearbook 2015). In 2015, there were 285,946 full-time teachers across 4,565 junior high schools (Ministry of Education 2016).

Within this province in Central China, we sampled 298 junior high schools from across 94 counties (there are 158 counties in total). We then randomly selected 886 seventh and eighth grade language arts teachers in these schools. A small percentage of randomly selected teachers (n = 54 or 6 percent) did not participate in the experiment: 42 or 78 percent were not present or sick on the days in which the study was conducted in schools; 4 or 7 percent declined to participate; and 8 or 15 percent began the study but left all grading forms blank.[2] This yields a total sample of 832 teachers (mean age = 36.9 years; proportion female = 0.676—see Table 1 for more details of this sample). Each consenting teacher was then asked to grade a series of essays and fill out a questionnaire.

Experimental Design

A series of pilot studies were first conducted to ensure the grading exercise would be understandable and routine. For instance, we adapted a grading rubric from national scoring standards for standardized tests, such that the rubric would appear standard for junior high school teachers. In addition to

Table 1. Descriptive Statistics of the Variables Used in the Study

Variables	(1) Mean	(2) SD	(3) Min	(4) Max
Outcomes				
Total score for essay B, Standardized	77.26	13.19	30	100
Total score for essay C, Standardized	81.14	11.66	38	100
Total score for essay D, Standardized	85.69	10.98	29	100
Treatments				
Positive Reputational Information—High Prior Grades	0.333	0.472	0	1
Negative Reputational Information—Low Prior Grades	0.332	0.471	0	1
Low-quality Essay B	0.507	0.500	0	1
Background Variables				
1. Baseline Essay Score (Standardized)	0.0948	0.950	−4.751	1.622
2. Female Teacher (1 = yes)	0.676	0.468	0	1
3a. Experience 0–2 years	0.056	0.230	0	1
3b. Experience 3–5 years	0.122	0.327	0	1
3c. Experience 6–12 years	0.107	0.310	0	1
3d. Experience 13–20 years	0.367	0.482	0	1
3e. Experience +21 years	0.348	0.477	0	1
4. Went to college (1 = yes)	0.629	0.483	0	1
5. Majored in Chinese language (1 = yes)	0.594	0.491	0	1
6. Urban residential registration (1 = yes)	0.767	0.423	0	1
7. Grew up and works within same county (1 = yes)	0.895	0.306	0	1

Source: Authors' Survey.
Notes: Observations here are of language arts primary school teachers, N = 832. For comparison, 80.2 percent of teachers across China had a college degree, and 53.5 percent were female (Ministry of Education of the People's Republic of China 2016). Note that official government statistics include teachers of all subjects.

routinizing the essay grading protocol, the pilot helped us select appropriate essays. Specifically, we asked teachers to grade (blind to information about the writer) a set of candidate essays during the pilot. This created a distribution of essay scores that allowed us to choose essays that were neither right nor left censored (too high or low quality). This information was also used to identify "high" (top 25th percentile), "low" (bottom 25th percentile), and "average" (50th percentile) quality essays.

After we incorporated feedback from the pilot study, we asked participating teachers to grade a set of four essays according to a standardized rubric. The

four essays (from henceforth, Essays A, B, C, and D) were selected from a series of essays written by anonymous seventh grade students as part of a standardized exam. The length of each essay was approximately one page or 700 Chinese characters.

All teachers first graded the same Essay A, an essay of "average" (50th percentile) quality. There was no reputational information attached to Essay A, and thus Essay A provides baseline estimates for the strictness or lenience of individual teacher grading, which we use as a control variable. After teachers finished grading Essay A, enumerators revealed that this essay was a solitary practice exercise to ensure that teachers understood the directions. Teachers were then told that subsequent essays B, C, and D were written by a junior high school student at the teacher's school. For these essays, we manipulated the reputational information associated with the essay writer (low prior grades, high prior grades, or no indication of prior grades) via a cover sheet. The cover sheet for all teachers noted that "the following essays were written by a first semester seventh grade student at your school." Positive or negative reputational information was manipulated by adding the following sentence: "In a previous assessment at this school, this student had grades in the bottom (top) 25th percentile." This cover sheet was repeated for each essay to remind teachers about the student's reputation.

We also manipulated essay quality for Essay B by assigning either a "high" (top 25th percentile) or "low" (bottom 25th percentile) quality essay. The quality of Essays C and D were not manipulated and remain average quality (50th percentile) essays. Although teachers initially understood that they were grading the essay of a real student, enumerators debriefed teachers after the study in accordance with an approved IRB protocol, explaining that the exercise was for research purposes only and would not affect the grades of actual students.

In sum, our experimental interventions were randomly assigned per a 2×3 matrix (Table 2). The columns indicate teacher assignment to receive negative, positive, or no reputational information, and the rows indicate assignments to grade a high- or low- quality Essay B. Balance tests suggest that randomization successfully created comparable groups of teachers across each treatment condition. Appendix Table 1a and 1b present balance statistics across the treatment conditions for 7 covariates (we will discuss these covariates in the variables section below). In the case of randomization in terms of reputational information, we found that none of the covariates tested were significantly different at the 10 percent level (Table 3a). In the case of essay quality, we again found that none of the covariates were significantly different at the 10 percent level (Table 3b).

Tests also suggest that our experimental manipulations were perceived as customary. Approximately 93 percent of teachers said they used the same standards in their day-to-day grading in a post-experiment questionnaire (Appendix Table 2). More importantly, the responses did not differ between teachers who were randomly assigned to receive prior grade information and those who graded blindly. Indeed, teachers who saw information about student reputations were *more* likely to state that the exercise was similar to day to day grading. For instance, when positive student reputations were available, the proportion of

Table 2. Treatment Assignment Matrix

	Positive Reputational information (Prior Performance at Top 25th Percentile)	Negative Reputational Information (Prior Performance at Bottom 25th Percentile)	No reputational information
High quality essay (Essay B at Top 25th percentile)	139	139	136
Low quality essay (Essay B at Bottom 25th percentile)	141	140	145

Notes: Numbers in each cell refer to number of teachers in each treatment group. Numbers are not equal because of variations in randomization.

teachers who believed the essay topics encountered were similar to day to day grading increased by 6.6 percentage points ($p = 0.098$—versus the control group that did not receive information about student reputations). The same was true when teachers were able to rely on negative student reputations to grade (9.7 percentage points—$p = 0.017$).

Of course, an important consideration in any experimental research is reactivity. In our case, asking teachers to grade essays may itself have shaped teacher recollection of their day-to-day grading practices, or the presence of out-of-town enumerators may have led teachers to answer the questionnaire in a way they thought best matched our research interests. Although we cannot claim to have addressed all such possibilities, certain features of our experiment may reduce the collective effect of reactivity. All enumerators were recruited from a local university and could speak the local dialect. The essays were written by real students from the same educational context. Teachers were asked to help grade the essays as part of a broader study about teacher training, with clear protocols to debrief teachers about this deception only at the end of the study. Pilot studies helped us streamline our protocol, such as giving teachers familiar grading rubrics.

Measures

The outcome variables of interest are the grades that teachers gave to Essays B, C, and D according to the standard rubric. The rubric required teachers to grade the essays along three dimensions: content (up to 40 points), language (up to 40 points), and structure (up to 20 points). Teachers were told that the scores for the three dimensions would be summed to create a total score ranging from 0 to 100 points, which is also the outcome of interest for this study.

After teachers finished grading the essays, enumerators asked them to fill out a questionnaire. Through the questionnaire, we collected information on teacher

Table 3. Main Effects of Reputational Information on Subsequent Evaluation

Dependent Var: Total Grade of Each Essay	(1) Essay B Score (SDs)	(2) Essay B Raw Score	(3) Essay C Score (SDs)	(4) Essay C Raw Score	(5) Essay D Score (SDs)	(6) Essay D Raw Score
Positive Reputational Information (Essay Writer Was Top 25th Percentile)	0.086 (0.079)	1.078 (0.980)	0.082 (0.082)	0.885 (0.883)	−0.045 (0.074)	−0.490 (0.802)
Negative Reputational Information (Essay Writer Was Bottom 25th Percentile)	−0.345*** (0.090)	−4.297*** (1.120)	−0.294*** (0.085)	−3.166*** (0.917)	−0.177** (0.079)	−1.905** (0.856)
Baseline Essay Score	0.404*** (0.039)	5.034*** (0.489)	0.429*** (0.041)	4.617*** (0.441)	0.285*** (0.039)	3.074*** (0.416)
Constant	−0.004 (0.060)	77.766*** (0.744)	0.004 (0.057)	81.453*** (0.610)	0.009 (0.056)	86.224*** (0.600)
R-squared	0.153	0.153	0.155	0.155	0.070	0.070
F-test comparing top 25 to bottom 25	27.47		17.06		2.357	
Prob > *F*	3.03e-07		4.71e-05		0.126	

Notes:

1. Standard errors adjusted for clustering at the school level in parentheses. $N = 832$ participants.
2. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

personal and professional background, such as whether the teacher was female or not (1 = female), whether the teacher was from an urban or rural area (1 = urban), whether the teacher attended primary school in the same county he or she now teaches (1 = yes), whether the teacher attended any college (1 = yes), whether the teacher's major was Chinese language arts (1 = yes), and teaching experience (broken into categories to capture any nonlinearities in how experience relates to teacher quality—Rivkin, Hanushek, and Kain 2005). As shown above, these variables were used to test for balance across treatment arms.

Statistical Approach

We examine how varying student reputations affect teacher grading using the ordinary least squares regression model below:

$$Y_i = \alpha + \beta_1 P_i + \beta_2 N_i + \beta_3 Q_i + \gamma X_i + \varepsilon_i \quad (1)$$

where Y_i is the grading outcome for teacher i for Essays B, C, or D; P_i is a dummy variable indicating that teacher i is grading the essay of a student with ostensibly positive prior performance; N_i is a dummy variable indicating that teacher i is grading the essay of a student with ostensibly negative prior performance; Q_i is a dummy variable indicating that Essay B was of high quality (75^{th} percentile) versus low quality (25^{th} percentile); and X_i is the baseline grade from Essay A. The coefficients of interest are β_1 and β_2, which correspond to the effect of positive or negative reputational information, relative to blind grading. Finally, all standard errors are adjusted to account for clustering at the school level.

To examine whether reputational information is relied upon primarily as a filter or temporary signal, we assess how β_1 and β_2 change as teachers grade additional essays. To do so, we reformulate the basic model above by pooling essays B, C, and D. By pooling observations, we can include indicator variables for essays B, C, and D as well as interaction terms of these indicator variables with the reputational information treatment dummies. We have evidence that reliance on reputational information is changing if the interaction terms are statistically significant. Finally, to test how contradictory or confirmatory evidence may change teacher reliance on reputational information, we add interaction terms between Q_i, P_i, and N_i. For instance, if the coefficient on the interaction term between Q_i (high quality essay) and P_i (high prior grades) is statistically significant, we have evidence that confirmatory evidence changes teacher grading beyond high prior grades or essay quality alone.

An important feature of this statistical approach is that it was strictly based on a pre-analysis plan. Recent research suggests that experimental results fail to replicate more than expected by chance and identifies two key reasons for the low levels of replicability (Camerer et al. 2018; Freese and Peterson 2017). First, experimental research may fall prey to the "file-drawer effect," where researchers present only statistically significant results. Since five of every 100 regressions will yield statistically significant results at the 5 percent level, the analyses and regressions that yield non-significant results should also be presented,

rather than left in a file-drawer. Second, experiments may have limited replicability due to the "problem of forking paths." Datasets can be analyzed in multiple ways, and researchers can find statistically significant results by minor adjustments in how they analyze a dataset (choosing different "forking paths"). Low replicability occurs if the particular way a researcher analyzes the data leads to a statistically significant result, but other approaches do not (Gelman and Loken 2013).

In light of these concerns, we registered a pre-analysis plan detailing our primary hypotheses and analytical methods.[3] This was done prior to analyzing the outcomes in our data. The usage of pre-registration does not solve all challenges relating to replicability (Olken 2015). However, it is a useful tool and a current standard for addressing replicability concerns (Nosek et al. 2018). Our experiment is less prone to the file-drawer effect, as we commit to presenting all analyses rather than only the ones that were statistically significant. Analyses that are identified as under-powered and exploratory in the pre-analysis plan are also labeled as such in the manuscript. Moreover, our experiment is less prone to the problem of forking paths, as we commit to a plan to analyze the dataset in advance, rather than adjusting our analysis to find a statistically significant result. Prior to running our experiment, we also conducted power calculations to ensure that the number of participants in the experiment were sufficient to find statistically significant results on the dimensions that we hypothesized in advance. Any handling or filtering of the data was registered in advance such that we could not return and make adjustments after running results.

We note one discrepancy in sample size between our current analyses and our pre-analysis plan. At the time we pre-registered our analyses, we believed that we had a sample of 840 teachers, but our effective sample was only 832. This is because eight teachers consented to the experiment but left us with blank grading sheets. We categorized these non-responses as teachers who declined to participate.

Results

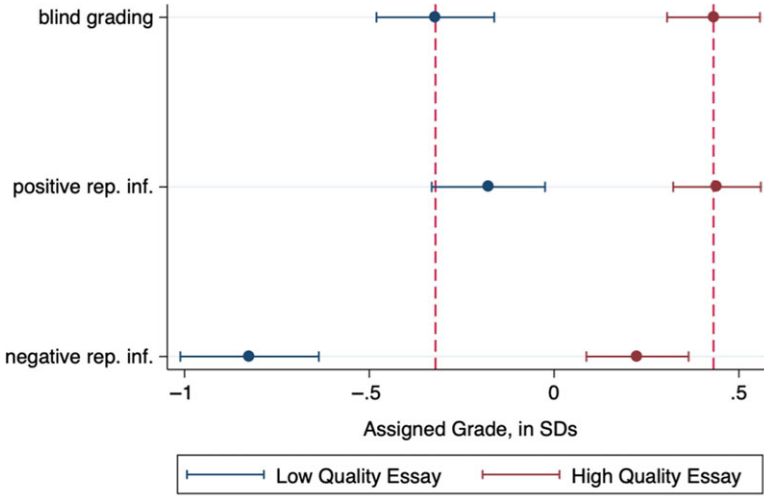
Initial Effects of Reputational Information

Reputational information shapes how teachers initially grade essays (Table 3). On Essay B, negative reputational information causes teachers to penalize students by 0.345 standard deviations (SDs) ($p < 0.001$—Table 3, Column 1, Row 2). In terms of raw scores, this is equivalent to a decline of 4.3 points or half a letter grade. When a student is labeled as having high prior grades (positive reputational information), teachers grade their essays with a 0.086 SD boost. However, this latter effect is not statistically significant ($p = 0.295$—Column 1, Row 1).[4] Moreover, the magnitude of the 0.345 SD penalty is larger than 0.086 SD boost ($p < 0.001$). As such, this result suggests that negative reputational information shapes teacher evaluation more powerfully than positive reputational information.

Teachers also correctly distinguish quality differences between essays, validating that the "high" or "low" quality essays identified in our pilot studies were also perceived as such among the teachers in our sample. High quality essays were graded 0.808 SDs better than low quality essays (Appendix Table 3, Column 1). In raw terms, low quality essays were graded at the 33rd percentile (with 95 percent confidence intervals crossing the 25th percentile), and high-quality essays were graded at the 64th percentile (also with 95 percent confidence intervals crossing the 75th percentile).[5]

How does encountering confirmatory versus contradictory evidence change how teachers initially rely on reputational information? Figure 1 illustrates how teachers graded high and low-quality essays across the reputational information groups (these results can also be found in table form in Appendix Table 3, Column 2). Blind grading is placed at the top of the figure and serves as the baseline for comparison. For clarity, we use dotted vertical lines to indicate the mean grades assigned to low and high-quality essays, respectively. The figure reiterates our earlier finding that negative reputational information shapes teacher judgment more powerfully than positive reputational information, highlighting that the asymmetry is further driven by the interaction between reputational information and type of performance: negative reputations are cumulative with confirmatory performance, with confirmatory evidence increasing the penalty that teachers apply to students with negative reputations. In our experiment, a student with poor prior performance (had grades in the bottom 25th percentile in a previous assessment) produces confirmatory evidence by writing a low-quality (25th percentile) essay. In this confirmatory case, teachers penalize the essay by

Figure 1. Reliance on Reputational Information, across Confirmatory and Contradictory Evidence.



Notes: error bars correspond to 95% confidence intervals.
All standard errors adjusted for clustering at the school level.
Dotted lines are of blind grading (control) group for either high (top 25th percentile) or low-quality (bottom 25th percentile) essays.

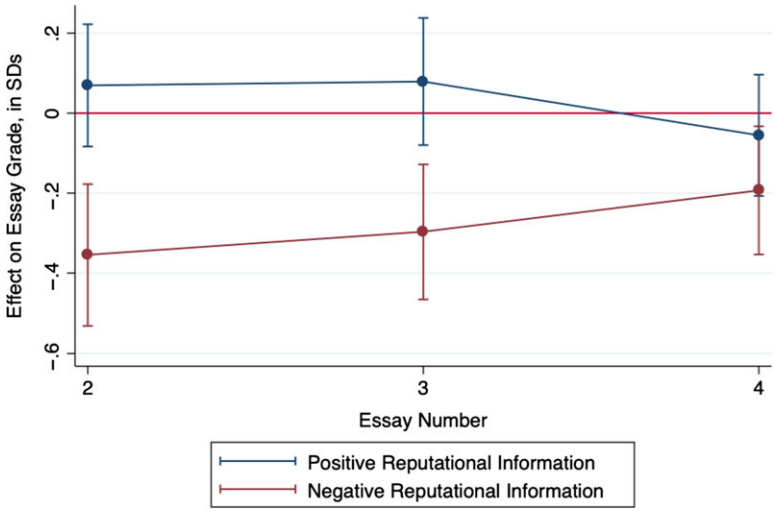
0.502 SDs ($p < 0.001$). By contrast, when negative reputational information is contradicted by good performance, teachers only penalize such a student by 0.202 SDs ($p = 0.057$). The difference between these two estimates is statistically significant ($p = 0.045$).

As with our average results, positive reputational information appears to have a weaker effect on teacher evaluation, even when we decompose these effects by confirmatory or contradictory evidence. When positive reputational information interacts with low quality performance, teachers assign a minor boost of 0.137 SDs, but this effect is not statistically significant ($p = 0.187$—Figure 1). When positive reputational information is confirmed by good performance, teachers grade as if blind to prior grades ($p = 0.911$).

Effects of Reputational Information on Subsequent Essays

Central to our study is how reliance on reputational information changes as teachers grade subsequent essays. Figure 2 illustrates the change in effects across all three essays by plotting the effects (along with 95 percent confidence intervals) from Table 3. Effects are statistically significant at the 5 percent level if the intervals exclude the horizontal line, which indicates a zero effect. The Figure demonstrates that, on average, reliance on reputational information diminishes in magnitude over subsequent essays. Students with low prior grades receive a penalty across all three essays but this effect diminishes as teachers grade subsequent student essays. The penalty after grading three essays halves from 0.345 SDs to 0.177 SDs (a statistically significant change, $p = 0.095$). Labeling a student as having high prior grades continues to have no effect across subsequent essays (the confidence intervals always cross zero). This suggests

Figure 2. Main Effects of Reputational Information, across Essays.



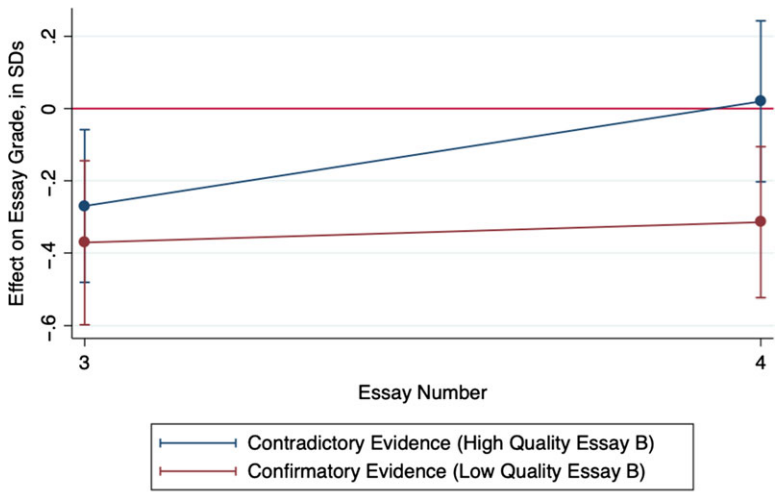
Notes: error bars correspond to 95% confidence intervals.
All standard errors adjusted for clustering at the school level.

that, on average, teachers rely on reputational information as temporary signals, updating their expectations as they observe student performance.

These average results mask heterogeneity by whether teachers initially encountered contradictory or confirmatory evidence. To explore this heterogeneity, Appendix Table 4 presents the three-way interaction effects between essay number, the quality of essay B, and the reputational information treatment condition. This model decomposes the separate effects of each treatment (high- or low-quality essay; high, low, or null reputational information) and interactions among treatments, across each essay. Although the interaction terms in this model lack statistical power because they test for *additional* rather than average effects, Appendix Table 4 allows us to isolate our estimates of interest. For instance, to isolate the effect of seeing contradictory evidence on how teachers grade essay D, we take a linear combination of the overall effect of negative reputations when essay B is high quality with the interaction of negative reputations and the indicator variable for essay D. To calculate the effect of seeing confirmatory evidence (a low-quality essay B), we further combine the coefficients from (a) the overall interaction term between negative reputations and low-quality essay B and (b) the three-way interaction term between seeing a low-quality essay B, the indicator for essay D, and negative reputations. For ease of interpretation, Figure 3 plots these coefficients to show how reliance on negative reputational information changes between essays C and D, for teachers encountering confirmatory versus contradictory evidence.

Figure 3 shows that the penalty for confirmatory evidence (negative reputational information and low-quality Essay B) fails to diminish in magnitude as teachers have more opportunities to grade. Even though essays C and D are

Figure 3. Effect of Negative Reputational Information over Essays C and D.



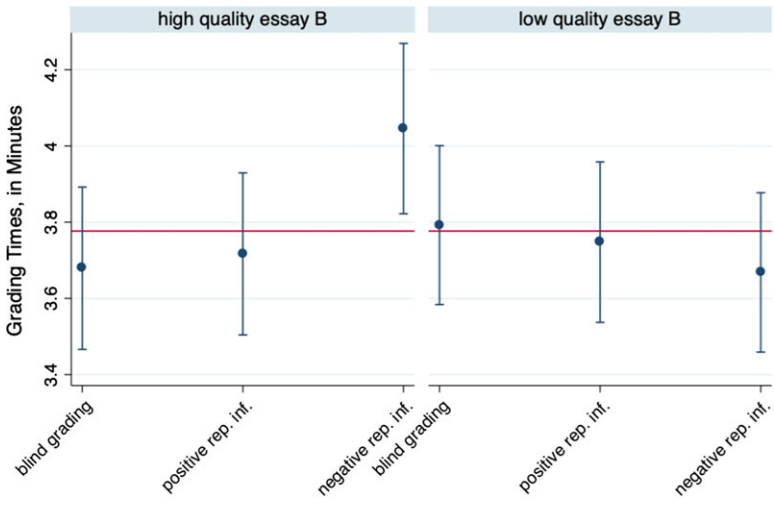
Notes: error bars correspond to 95% confidence intervals.
All standard errors adjusted for clustering at the school level.
Coefficients and errors calculated from linear combinations of coefficients from Appendix Table 4.
Effects of positive reputational information (high prior grades) not shown to enhance clarity.

identical and of average quality, having a negative reputation causes students to incur a penalty of 0.325 SDs even by Essay D (a difference that is statistically different from zero, $p = 0.002$). The fact that the penalty persists without decline suggests that negative reputations crystallize if given initial behavioral confirmation in Essay B. By contrast, teachers reduce their reliance on reputational information if exposed initially to contradictory evidence: if Essay B disconfirmed reputational information, the grades that teachers assign by Essay D are no longer distinguishable from blind grading ($p = 0.808$). Because the confidence intervals overlap between the profiles for confirmatory and contradictory evidence, we also test if the difference between these parameter estimates is statistically significant. Results from a Wald test lead us to reject the null hypothesis that the difference between these profiles is zero ($F = 4.94$, $p = 0.0269$), implying that the difference is statistically significant.

Potential Mechanisms and Robustness Checks

One mechanism that may explain this difference in updating is that teachers pay longer attention to evidence that is initially contradictory and thus avoid taking reputations for granted. As an exploratory test of this proposed mechanism, we examine average grading time by each treatment group in Figure 4 (also displayed in Appendix Table 5). The horizontal line indicates the average amount of time (across all treatment groups) it took a teacher to grade each essay (3.78 minutes). The key takeaway is that teachers who thought the essay writer had low prior grades but wrote a high-quality essay (contradictory evidence) took 4.04 minutes to grade the essay—an increase of approximately 16 seconds compared to the average. We reject the null hypothesis that the grading time for

Figure 4. Teachers Spend More Time Grading when Encountering Contradictory Evidence.



Note: Horizontal line indicates the average amount of time a teacher spent grading each essay (average across all treatment groups).

teachers who encountered contradictory evidence is the same as blind grading (two-tailed t-test = 2.25, p = 0.025).

Of course, a statistically significant increase may not be substantively significant. How large is a 16 second increase in grading time, and is this sufficient for a teacher to begin updating his or her expectations? To more meaningfully interpret this amount of time, we consider how long teachers take to grade essays in other contexts. For instance, in the formal context of a university, graders spend approximately three minutes assessing essays with approximately 350–500 words (Holstein 1983). Another estimate for how long a grader takes to reliably assess performance is two minutes (Diederich 1974). Moreover, although 16 seconds appears miniscule, it is a 9 percent increase on the approximately 180 seconds (3 minutes) that teachers spent grading each essay. Indeed, we conjecture that a 9 percent increase in grading time sufficient for teachers to begin updating their expectations. Nevertheless, we acknowledge that this is speculative and, as we note below, encourage future work to examine the amount of exposure to contradictory evidence needed before teachers jettison reliance on reputational information.

Discussion

Our results imply that teachers initially rely on reputational information to evaluate performance. In general, reputational information operates as a temporary signal: teachers rely less on reputational information as they have additional opportunities to evaluate student performance. However, negative reputations appear to shape teacher evaluations more powerfully than positive ones, and initial behavioral confirmation appears to crystallize negative reputations, solidifying a presumption of negative performance despite subsequent disconfirming evidence. As such, teachers do not consistently update when encountering information about student performance, neither do teachers always filter new information through the lens of their prior expectations. Instead of supporting a uniform view of reputations as temporary signals or filters, our results point to a more contingent model that depends on the initial evidence presented.

Our interpretation of the theoretical implications of these results should be tempered by recognizing certain limitations. First and foremost, the experiment was conducted among teachers in a particular historical moment and context, and the high-stakes, exam-oriented educational context shapes how teachers rely on reputational information when assigning grades. On the one hand, it could be that the high-stakes nature of grades increases teachers' sense of responsibility to ensure that grades are assigned accurately. If true, this would imply teachers rely less on reputational information in this context, making this study a conservative estimate of the effects of reputational information. On the other hand, the fact that students are publicly ranked in terms of their academic performance means that teachers may be more willing to rely on reputational information when grading. For instance, teachers may seek to minimize the probability that the grades they assign will be questioned by students, parents, colleagues, or administrators. Especially in exam-oriented education systems

where grades are high-stakes, teachers may be concerned about criticisms that require them to justify the grades they assign. To the extent that this is true, it is possible that teachers assign grades consistent with reputational information to help them "cover their assessment" from future criticism. If reliance on reputational information is indeed used to disarm criticisms, we acknowledge that the magnitude of effects observed in this study might well constitute an upper-bound on how educators use reputational information.

Moreover, certain features of exam-oriented education systems may explain increased teacher reliance on negative over positive reputational information. One possibility, for example, is that teachers in China traditionally believe that identifying student mistakes improves educational outcomes (Brown and Gao 2015). High-stakes, exam-oriented education systems prioritize the assessment of memory, repetition, and imitation, and students may indeed benefit from feedback on errors and mistakes (Bell 2016: 107). Moreover, some research suggests that the effort of Chinese students is less associated with their perceptions of self-esteem than educational contexts like that of the United States (Hau and Ho 2012), where negative feedback from teachers is believed to impede learning by reducing students' self-esteem and subsequent effort (Ames 1990; Stipek and Gralinski 1996). In addition to factors that increase reliance on negative reputational information, instructors in exam-oriented systems may also actively reduce reliance on positive reputational information. Teachers and parents in China, for instance, traditionally eschew positive encouragement, often with the assumption that excessive encouragement will reduce student motivation to work hard (Sargent 2011).

A related limitation to external validity is that the experiment observes how teachers grade essays, a task that is relatively subjective. Would reputational information still affect teacher evaluations if they were grading subjects with more objective criteria, such as a multiple-choice mathematics examination? We conjecture that teacher reliance on reputational information would be reduced, but prior work suggests that even evaluations of multiple-choice mathematics tests remain biased by expectations: Hanna and Linden (2012) observed grading discrimination even in mathematics tests in India. Even when asked to grade identical multiple-choice tests, teachers refrained from giving partial credit to lower-caste students. This led to a penalty of 0.03 to 0.08 standard deviations (SDs) relative to higher-caste students.

Aside from limitations in external validity, the present study also raises new questions that are worthy of future research. First, our distinction between filters (which bias perception of subsequent evidence) and temporary signals (which do not) does not imply mutual exclusion of these two explanations. How people actually use reputational information likely falls in a spectrum between these two ideal types. In other words, our claim should be interpreted as a matter of degree: on average, teachers in our context use reputational information more like a signal than a filter, except when encountering confirmatory, negative reputational information. A related point is that crystallized reputations imply persistence but not permanence. Reliance on reputations as filters implies biased perceptions of subsequent evidence, and excepting extreme cases, it is plausible

that an extraordinary student performance will break through. And it is also plausible that a teacher, seeing the same student perform well above expectations for an entire year, changes his or her mind.

If this is true, how many opportunities are needed, and how contradictory does subsequent evidence have to be to break teachers out of using reputational information? Reliance on confirmed negative reputations might diminish when teachers are given a sufficiently large number of opportunities to assess performance, perhaps over the course of a school year. Similarly, unless teachers rely on reputational information completely as a filter for subsequent evidence, there will likely be a threshold for when contradictory evidence will break through. If a student with a reputation for producing low-quality work produces something of the highest caliber, the extraordinary nature of this product is likely to cause teachers to break away from reliance on crystallized reputations. The present study suggests that producing something average is unlikely to cut through a crystallized reputation, but this does not resolve where the threshold would be. Our study does not vary essay quality on a continuous scale and thus we are unable to answer this question definitively; however, we believe this is an important question for future research.

We also invite future work to further explore why negative reputational information has a stronger influence on teacher evaluation than positive reputational information. The asymmetry of effects for negative and positive information has been found in other settings, such as situations where employers recall and rely more strongly on negative prior experiences to determine whom they will hire (Leung 2017; Pager and Karafin 2009; Rozin and Royzman 2001). In these contexts, this asymmetry is generally explained by noting that the losses from a bad hire outweigh the benefits from a good hire. Although this loss aversion principle is well-established in hiring contexts (Kahneman, Knetsch, and Thaler 1991), it is difficult to imagine how it would apply to grading, as there is no clear "loss" that teachers experience when grading essays.

As an avenue for future research, one preliminary conjecture is that the reputational information we assigned was *interpreted* as asymmetric even if it was mathematically symmetric. The reputational information was communicated using students' percentile rank in the school because a student percentile rank of 50 is mathematically the "average" percentile rank of students, while a B student may be average in some cases while A- might be average in other contexts. Nevertheless, teachers might *still* have interpreted signals like 75th percentile as asymmetric from 25th percentile, even when they were mathematically symmetric. For instance, this could occur if teachers categorize a 25th percentile student as a "lost cause." If it is true that teachers interpreted information about "25th percentile" more negatively than they interpreted "75th percentile" positively, this would be one explanation for why teachers in this context relied on negative reputations more heavily than positive information.

With these limitations and suggestions for future work in mind, it is interesting to note how the basic findings of this study are consistent with findings from different settings. The findings from this study suggest that teachers rely on reputational information as temporary signals, except when these reputations receive

behavioral confirmation. The idea that confirmatory signals allow evaluators to act upon negative preconceptions is found in other contexts as well. For instance, in studying when decision-makers apply sanctions on welfare applicants, Schram et al. (2009) show how "discrediting markers," such as prior failures to comply with welfare regulations, enable decision-makers to act upon their negative stereotypes of Latina and African-American clients.

To the extent that the patterns observed in this study hold more broadly, they have significant implications both for theory and practice. For sociologists of education and inequality, these results clarify the relative power of reputational information and the reality of student performance (Henshel 1982; Jussim and Harber 2005). The idea that expectations change reality was established more than 40 years ago (Rosenthal and Jacobson 1968), and, as noted above, recent work continues to show how teachers rely on their prior expectations when grading (Botelho, Madeira, and Rangel 2015; Hanna and Linden 2012; Sprietsma 2013). These studies, however, privilege measurements of bias without a concomitant measure of how these underlying prior expectations are updated, making it difficult to isolate whether these biases are self-correcting or self-fulfilling. By theorizing and experimentally testing how reputational information affects subsequent evaluation, this study suggests that teachers generally rely on reputational information temporarily, except when behavioral confirmation leads to the crystallization of negative reputations. These patterns imply that students in educational settings like China can generally recover from having negative reputations and are unlikely to be stuck in place. However, they must challenge negative reputations rapidly to do so.

Our study also complements extensive research examining how structural changes like ability-grouping or grade retention might keep students stuck in place (Andrew 2014; Eder 1981; Hallinan 1994; Hattie 2008; Karlson 2015; Oakes 1985). By focusing on within-person evaluative processes and when they are updated in response to student performance, our findings imply that tracking or other institutionalized changes in resource distributions are problematic not merely because they shape the resources that students obtain. Instead, such groupings are problematic also because they also keep teachers from being exposed to evidence that updates their expectations, further increasing the chances that negative reputations become crystallized. In the school setting, this could occur if students with negative reputations are tracked or given assignments that are easier than other students. Such students may find it more challenging to escape negative reputations.

Policymakers are generally invested in ensuring students perform to the best of their potential rather than being stuck in place by their reputations. A plausible policy implication of the study would be to conduct grading blindly. In practice, this is difficult in classrooms teachers might be able to identify students based on the topics they write about, their handwriting, and a series of other personal identifiers. Teachers should also have the opportunity to provide tailored, formative assessment for students by accounting for their prior performance (Fisher and Frey 2007) and providing tailored feedback encourages further learning (Harlen and James 1997).

Considered more broadly, the use of reputational information in grading has both benefits and costs. As noted at the outset, reputations reduce the amount of time teachers need to grade, and existing sociological work has demonstrated how actors make faster judgments (Raub and Weesie 1990). School districts may seek to increase the amount of prior assessment information shared with teachers precisely for this purpose. Indeed, one reason why Chinese classrooms post grades publicly is to create a reputational system that ostensibly motivates students to work harder. While these may be legitimate reasons to institute reputational systems, they must be weighed against the possibility that, under some circumstances, reputational effects overpower student performance and create inequalities in the classroom that are not merit-based.

One policy strategy would be to increase opportunities for students to reset their reputations. For instance, the results from this field experiment highlight the importance of timing in interventions. Programs designed to reduce achievement gaps may have differential effects based on when they are implemented. Assuming that teachers share reputational information about students across school years, the best time to update such expectations is at the beginning of a new school year. Thus, beyond their immediate effects, interventions like summer remedial programs (Banerjee et al. 2016) may help low performing students catch up with their higher performing peers while also helping to recalibrate teacher expectations.

Conclusion

Whether in educational settings or more generally, individuals regularly rely on reputational information to evaluate one another. If used as temporary signals, reputational information can help teachers reduce time and energy spent evaluating quality. However, when used as filters, reputations are likely to crystallize and entrench inequalities, making it difficult for students to escape the presumption of negative performance. By leveraging unique data from a field experiment, this experiment examines how teacher reliance on reputational information changes over the course of grading multiple essays. Finding evidence that reliance on reputational information generally diminishes, this article begins to resolve whether the metaphor of signal or filter better reflects how teachers use reputational information. By demonstrating how behavioral confirmation is necessary for the crystallization of reputations, the article also makes important inroads to understanding *when* teachers rely on reputational information as a signal or filter. Taken together, these theoretical insights clarify the effects of reputational information on subsequent teacher evaluations, with implications for how reputational information shapes patterns of inequality more generally.

Notes

1. We use the term "reputational information" in lieu of "reputation," which is commonly deployed to refer to different underlying concepts: prominence or generalized awareness, an expectation of some behavior based on past demonstrations, or

generalized favorability (Bromberg and Fine 2012; Lange, Lee, and Dai 2011). In this paper, we focus on the second of these definitions and use the term reputational information to enhance clarity: it is an "expectation of some behavior or behaviors based on past demonstrations of those same behaviors" (Podolny 2005: 14).

2. We lack information on teachers who were in our sample but declined to participate, and to the extent that their grading practices may differ from those who chose to participate, our results only represent the subset of teachers who were present and willing to participate. However, the primary reason for non-participation was not related to treatment assignment, including being sick or out of town.
3. The pre-analysis plan can be accessed on the following website: Authors. 2016. "Accuracy and Updating: How Labels and Initial Performance Affect Teacher Grading." August 20. AEA RCT Registry. (https://www.socialscienceregistry.org/trials/1498)"
4. We do not conclude that positive reputational information has no effect. First, the magnitude of the effects suggest that teachers give a boost to students with positive reputations, even if this is not statistically significant. Second, our experiment benchmarks prior performance at the 25th and 75th percentile, and it is plausible that the boost would be observed if positive reputational information were indicated by the top 10th or 5th percentile.
5. We pre-registered an analysis to examine whether the initial performance of the student alone (i.e., independent of reputational information) would have spillover effects on subsequent grading. Our hypothesis was that high or low initial essay quality would inform the teacher's expectations directly, thus affecting subsequent essay grading. To comply fully with our pre-analysis plan we display these results in Appendix Table 5. We find that a single high- or low-quality performance on an essay is insufficient to influence how teachers grade subsequent essays.

About the Authors

James Chu is a sociology PhD Candidate at Stanford University. His research explores how status and reputations shape patterns of inequality and stratification, especially in educational systems.

Guirong Li is a Professor at the School of Education, Henan University. Her research focuses on program and policy evaluation and the economics of education. She is the Director of the International Center for Action Research on Education (ICARE) and conducts large-scale empirical research in central China to improve education quality and reduce educational inequality.

Prashant Loyalka is an Assistant Professor at the Graduate School of Education at Stanford University. His research focuses on examining and addressing inequalities in the education of youth as well as on understanding and improving the quality of education received by youth in various countries including China, India, and Russia. As part of his work, he conducts large-scale evaluations of educational programs and policies that seek to improve student outcomes.

Chengfang Liu is Associate Professor at the China Center for Agricultural Policy, School of Advanced Agricultural Science, Peking University. She works on the economics of poverty reduction by focusing on the issues of education, health, migration, off-farm employment and infrastructure in China.

Leonardo Rosa is a doctoral candidate in the Economics of Education program at the Graduate School of Education in Stanford. He graduated from University of Sao Paulo in 2015 with a MA degree in Economics and from Pontifical University Catholic of Sao Paulo in 2011 with a BA in Economics.

Yanyan Li is a Center Research Fellow at the International Center for Action Research on Education (ICARE) and an Assistant Professor at the School of Education, Henan University. Her research focuses on reducing the educational disadvantages of students by evaluating and improving the quality of educational programs in rural China.

References

Allen, James D. 2005. "Grades as Valid Measures of Academic Achievement of Classroom Learning." *The Clearing House: A Journal of Educational Strategies, Issues and Ideas* 78(5):218–223.

Altonji, Joseph G., and Pierret Charles R. 2001. "Employer learning and statistical discrimination." *The Quarterly Journal of Economics* 116(1):313–350.

Ames, Carole. 1990. "Motivation: What Teachers Need to Know." *Teachers College Record* 91(3): 409–421.

Andrew, M. 2014. "The Scarring Effects of Primary-Grade Retention? A Study of Cumulative Advantage in the Educational Career." *Social Forces* 93(2):653–85.

Baker, David. 2014. *The Schooled Society: The Educational Transformation of Global Culture*. Stanford, CA: Stanford University Press.

Banerjee, Abhijit, Banerji Rukmini, Berry James, Duflo Esther, Kannan Harini, Mukherji Shobhini, Shotland Marc, and Walton Michael. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of 'Teaching at the Right Level' in India." National Bureau of Economic Research Working Papers Series: No. 22746.

Bell, Daniel A. 2016. *The China Model: Political Meritocracy and the Limits of Democracy*. Princeton: Princeton University Press.

Bol, Thijs, de Vaan Mathijs, and van de Rijt Arnout. 2018. "The Matthew Effect in Science Funding." *Proceedings of the National Academy of Sciences* 115(19):4887–4890.

Botelho, Tristan L., and Abraham Mabel. 2017. "Pursuing Quality: How Search Costs and Uncertainty Magnify Gender-Based Double Standards in a Multistage Evaluation Process." *Administrative Science Quarterly* 62(4):698–730.

Botelho, Fernando, Madeira Ricardo A., and Rangel Marcos A. 2015. "Racial Discrimination in Grading: Evidence from Brazil." *American Economic Journal: Applied Economics* 7(4):37–52.

Bromberg, Minna, and Fine Gary Alan. 2012. "Resurrecting the Red: Pete Seeger and the Purification of Difficult Reputations." *Social Forces* 80(4):91–111.

Brophy, Jere E., and Good Thomas L. 1974. *Teacher-Student Relationships: Causes and Consequences*. Holt, Rinehart & Winston.

Brown, Gavin T. L., and Gao Lingbiao. 2015. "Chinese Teachers' Conceptions of Assessment for and of Learning: Six Competing and Complementary Purposes" edited by Kris Gritter. *Cogent Education* 2(1).

Callahan, Rebecca M. 2005. "Tracking and high school English learners: Limiting opportunity to learn." *American Educational Research Journal* 42(2):305–328.

Camerer, Colin F., et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* 2(9):637–44.

Chamley, Christophe. 2004. *Rational Herds: Economic Models of Social Learning.*. Cambridge, UK: Cambridge University Press.

Chetty, Raj, Friedman John N, and Rockoff Jonah E. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9): 2633–79.

Clark, Burton R. 1986. *The Higher Education System: Academic Organization in Cross-national Perspective*. Berkeley, CA: University of California Press.

Correll, Shelley J., and Benard Stephen. 2006. "Biased Estimators? Comparing Status and Statistical Theories of Gender Discrimination." *Advances in Group Processes* 23:89–116.

Correll, Shelley J., Benard Stephen, and Paik In. 2007. "Getting a Job: Is There a Motherhood Penalty?" *American Journal of Sociology* 112(5):1297–338.

Daley, B., and Green B. 2014. "Market Signaling with Grades." *Journal of Economic Theory* 151:114–145.

Darley, John M., and Gross Paget H. 1983. "A Hypothesis-Confirming Bias in Labeling Effects." *Journal of Personality and Social Psychology* 44(1):20.

Diederich, Paul B. 1974. *Measuring Growth in English*. Urbana, IL: National Council of Teachers of English.

Diekmann, Andreas, Jann Ben, Przepiorka Wojtek, and Wehrli Stefan. 2014. "Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets." *American Sociological Review* 79(1): 65–85.

DiPrete, Thomas A., and Eirich Gregory M. 2006. "Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments." *Annual Review of Sociology* 32:271–297.

Eder, Donna. 1981. "Ability Grouping as a Self-fulfilling Prophecy: A Micro-analysis of Teacher-student Interaction." *Sociology of Education* 54(3):151–162.

Elashoff, Janet D., and Snow Richard E. 1971. *Pygmalion Reconsidered*. Worthington, OH: Charles A. Jones.

Feinberg, Matthew, Willer Robb, Stellar Jennifer, and Keltner Dacher. 2012. "The Virtues of Gossip: Reputational Information Sharing as Prosocial Behavior." *Journal of Personality and Social Psychology* 102(5):1015.

Fisher, Douglas, and Frey Nancy. 2007. *Checking for Understanding: Formative Assessment Techniques for Your Classroom*, 2nd ed. Alexandria, VA: Association for Supervision and Curriculum Development.

Freese, Jeremy, and Peterson David. 2017. "Replication in Social Science." *Annual Review of Sociology* 43(1):147–65.

Fleming, Neil D. 1999. "Biases in marking students' written work: quality." In *Assessment Matters in Higher Education: Choosing and Using Diverse Approaches*, edited by Brown S., and Glasner A., Pp83–92. Philadelphia, PA: Open University Press.

Gilboa, Itzhak, and Schmeidler David. 1993. "Updating Ambiguous Beliefs." *Journal of Economic Theory* 59(1):33–49.

Gelman, Andrew, and Loken Eric. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "p-Hacking" and the Research Hypothesis Was Posited Ahead of Time." Open Science Foundation Working Paper (https://osf.io/n3axs/)

Guskey, Thomas R., and Bailey Jane M. 2001. *Developing Grading and Reporting Systems for Student Learning*. Thousand Oaks, CA: Corwin Press.

Hallinan, Maureen T. 1994. "Tracking: From Theory to Practice." *Sociology of Education* 67(2):79–84.

Hanna, Rema N., and Linden Leigh L. 2012. "Discrimination in Grading." *American Economic Journal: Economic Policy* 4(4):146–168.

Harlen, Wynne, and James Mary. 1997. "Assessment and learning: differences and relationships between formative and summative assessment." *Assessment in Education* 4(3):365–379.

Hattie, John. 2008. *Visible Learning: A Synthesis of Over 800 Meta-analyses Relating to Achievement*. New York: Routledge.

Hau, Kit Tai, and Ho Irene T. 2012. "Chinese Students' Motivation and Achievement." In *Oxford Handbook of Chinese Psychology*, 187–204. Oxford: Oxford University Press.

2015. *Henan Statistical Yearbook*. Zhengzhou: Henan Municipal Bureau of Statistics, China Statistics Press.

Henshel, Richard L. 1982. "The Boundary of the Self-fulfilling Prophecy and the Dilemma of Social Prediction." *British Journal of Sociology* 33(4):511–528.

Holstein, J. A. 1983. "Grading Practices: The Construction and use of background knowledge in evaluative decision-making." *Human Studies* 6(1):377–392.

Hoxby, Caroline M. 2007. *College choices: The economics of where to go, when to go, and how to pay for it*. Chicago: University of Chicago Press.

Jussim, Lee, and Harber Kent D. 2005. "Teacher Expectations and Self-fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies." *Personality and Social Psychology Review* 9(2): 131–155.

Kahneman, Daniel, Knetsch Jack L., and Thaler Richard H. 1991. "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias." *Journal of Economic Perspectives* 5(1):193–206.

Karlson, Kristian Bernt. 2015. "Expectations on Track? High School Tracking and Adolescent Educational Expectations." *Social Forces* 94(1):115–41.

Kelly, S. 2008. "What Types of Students' Effort Are Rewarded with High Marks?" *Sociology of Education* 81(1):32–52.

Knobloch-Westerwick, Silvia, and Meng Jingbo. 2009. "Looking the Other Way: Selective Exposure to Attitude-Consistent and Counter-Attitudinal Political Information." *Communication Research* 36(3): 426–448.

Lange, Donald, Peggy M. Lee, and Dai Ye. 2011. "Organizational Reputation: A Review." *Journal of Management* 37(1):153–184.

Leung, Ming D. 2017. "Learning to Hire? Hiring as a Dynamic Experiential Learning Process in an Online Market for Contract Labor." *Management Science* 64(12): 5651–5668.

Liu, Judith, Ross Heidi A., and Kelly Donald P. 2000. *The Ethnographic Eye: Interpretive Studies of Education in China*. Taylor & Francis.

Marsh, Herbert W. 1986. "Verbal and Math Self-concepts: An Internal/External Frame of Reference Model." *American Educational Research Journal* 23(1):129–149.

Marzano, Robert J. 2006. *Classroom Assessment & Grading that Work*. Alexandria, VA: Association for Supervision and Curriculum Development.

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science (New York, N.Y.)* 159(3810):56–63.

Meyer, John W., and Ramirez Francisco O. 2009. "The World Institutionalization of Education." In *Discourse Formation in Comparative Education*, edited by Schriewer Jürgen, Pp111–132. Frankfurt: Peter Lang.

Milinski, Manfred, Semmann Dirk, and Krambeck Hans-Jürgen. 2002. "Reputation Helps Solve the 'Tragedy of the Commons'." *Nature* 415(6870):424–426.

Ministry of Education of the People's Republic of China. 2016. "Number of Full-time Teachers in Junior Secondary Schools by Educational Attainment and Professional Rank (Total)." Retrieved July 17, 2017 (http://www.moe.edu.cn/s78/A03/moe_560/jytjsj_2015/2015_gd/201610/t20161019_285616.html).

National Bureau of Statistics. 2015. *China Statistical Yearbook*. Beijing, China: China Statistical Publishing House.

Nickerson, Raymond S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology* 2(2):175.

Nosek, Brian A., Ebersole Charles R., DeHaven Alexander C., and Mellor David T. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences of the United States of America* 115(11):2600–2606.

Oakes, Jeannie. 1985. *Keeping Track*. New Haven: Yale University Press.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-analysis Plans." *Journal of Economic Perspectives* 29(3):61–80.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108(5):937–975.

Pager, Devah, and Shepherd Hana. 2008. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annual Review of Sociology* 34:181–209.

Pager, Devah, and Karafin Diana. 2009. "Bayesian Bigot? Statistical Discrimination, Stereotypes, and Employer Decision Making." *The Annals of the American Academy of Political and Social Science* 621 (1):70–93.

Pedulla, David S. 2016. "Penalized or Protected? Gender and the Consequences of Nonstandard and Mismatched Employment Histories." *American Sociological Review* 81(2):262–289.

Podolny, J. M. 2005. *Status Signals: A Sociological Theory of Market Competition*. Princeton: Princeton University Press.

Postiglione, Gerald A. 2015. *Education and Social Change in China: Inequality in a Market Economy*. New York: Routledge.

Rabin, Matthew, and Schrag Joel L. 1999. "First Impressions Matter: A Model of Confirmatory Bias." *The Quarterly Journal of Economics* 114(1):37–82.

Raub, Werner, and Weesie Jeroen. 1990. "Reputation and Efficiency in Social Interactions: An Example of Network Effects." *American Journal of Sociology* 96(3):626–654.

Rissing, Ben A., and Castilla Emilio J. 2014. "House of Green Cards: Statistical or Preference-based Inequality in the Employment of Foreign Nationals." *American Sociological Review* 79(6):1226–1255.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73(2):417–458.

Rosenthal, Robert., and Jacobson Lenore. 1968. "Pygmalion in the Classroom." *The Urban Review* 3(1): 16–20.

Rosenthal, Robert, and Rubin Donald B. 1978. "Issues in Summarizing the First 345 Studies of Interpersonal Expectancy Effects." *Behavioral and Brain Sciences* 1(3):410–415.

Rozin, Paul, and Royzman Edward B. 2001. "Negativity Bias, Negativity Dominance, and Contagion." *Personality and Social Psychology Review* 5(4):296–320.

Salganik, Matthew J., and Watts Duncan J. 2008. "Leading the Herd Astray: An Experimental Study of Self-fulfilling Prophecies in an Artificial Cultural Market." *Social Psychology Quarterly* 71(4):338–355.

Sargent, Tanja Carmel. 2011. "New Curriculum Reform Implementation and the Transformation of Educational Beliefs, Practices, and Structures in Gansu Province." *Chinese Education & Society* 44(6): 47–72.

Schram, Sanford F., Soss Joe, Fording Richard C., and Houser Linda. 2009. "Deciding to Discipline: Race, Choice, and Punishment at the Frontlines of Welfare Reform." *American Sociological Review* 74(3): 398–422.

Shavit, Yossi, and Blossfeld Hans-Peter. 1993. *Persistent Inequality: Changing Educational Attainment in Thirteen Countries*. Boulder, CO: Westview Press.

Stipek, Deborah, and Gralinski J. Heidi. 1996. "Children's Beliefs about Intelligence and School Performance." *Journal of Educational Psychology* 88(3):397–407.

Kollock, Peter. 1998. "Social Dilemmas: The Anatomy of Cooperation." *Annual Review of Sociology* 24(1): 183–214.

Smith, Alison E., Jussim Lee, and Eccles Jacquelynne. 1999. "Do Self-fulfilling Prophecies Accumulate, Dissipate, or Remain Stable Over Time?" *Journal of Personality and Social Psychology* 77(3):548.

Sprietsma, Maresa. 2013. "Discrimination in Grading: Experimental Evidence from Primary School Teachers." *Empirical Economics* 45(1):523–538.

Swann, William B., and Ely Robin J. 1984. "A Battle of Wills: Self-verification Versus Behavioral Confirmation." *Journal of Personality and Social Psychology* 46(6):1287.

van de Rijt, Arnout, Kang Soong Moon, Restivo Michael, and Patil Akshay. 2014. "Field Experiments of Success-Breeds-Success Dynamics." *Proceedings of the National Academy of Sciences* 111(19): 6934–6939.

van de Rijt, Arnout. 2019. "Self-Correcting Dynamics in Social Influence Processes." *American Journal of Sociology* 124(5):1468–95.

Wineburg, Samuel S. 1987. "The Self-fulfillment of the Self-fulfilling Prophecy." *Educational Researcher* 16(9):28–37.

Wormeli, Rick. 2006. *Fair Isn't Always Equal: Assessing & Grading in the Differentiated Classroom*. Portland, ME: Stenhouse Publishers.

Zahra, Shaker A., and George Gerard. 2002. "Absorptive Capacity: A Review, Reconceptualization, and Extension." *Academy of Management Review* 27(2):185–203.

Appendix Table 1a. Test of Reputational Information Treatment Balance (across 7 Covariates)—Pooled Estimates

	Experience of Teacher										
	(1)	(2)	(3a)	(3b)	(3c)	(3d)	(3e)	(4)	(5)	(6)	(7)
Variables	Baseline score	Female	0–2 years	3–5 years	6–12 years	13–20 years	21+ years	College (yes/no)	Chinese major (yes/no)	Urban hukou (yes/no)	Local resident (yes/no)
Positive Reputational Information (1 = yes)	0.123	0.040	0.007	0.032	−0.052	−0.008	0.029	0.025	−0.007	−0.004	0.021
	(0.077)	(0.040)	(0.028)	(0.026)	(0.041)	(0.041)	(0.041)	(0.042)	(0.036)	(0.026)	(0.019)
Negative Reputational Information (1 = yes)	0.069	−0.046	0.022	0.037	−0.006	−0.024	−0.017	0.015	0.013	−0.030	−0.029
	(0.077)	(0.040)	(0.028)	(0.026)	(0.041)	(0.041)	(0.041)	(0.042)	(0.036)	(0.026)	(0.019)
Constant	0.043	0.680***	0.112***	0.083***	0.386***	0.361***	0.622***	0.579***	0.766***	0.906***	0.058***
	(0.054)	(0.028)	(0.020)	(0.018)	(0.029)	(0.029)	(0.029)	(0.030)	(0.025)	(0.018)	(0.014)
R-squared	0.003	0.006	0.001	0.003	0.002	0.000	0.002	0.000	0.000	0.002	0.008
Group1-Group2	−0.0540	−0.0854	0.0154	0.00446	0.0459	−0.0156	−0.0460	−0.0101	0.0200	−0.0261	−0.0502
p-value of diff	0.483	0.0317	0.581	0.865	0.263	0.702	0.264	0.809	0.578	0.317	0.00986

Notes:

1. Number of observations for each regression = 832. Each model is a separate linear regression of the covariate on two treatment indicators (reference group is no label).
2. Robust standard errors in parentheses.
3. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Variables are (1) standardized score of essay A; (2) whether the teacher is female or not (1 = female); (3a-3e) teacher experience; (4) whether the teacher attended any college (1 = yes); (5) whether the teacher's major was Chinese (1 = yes); (6) whether the teacher is registered as an urban resident (1 = yes); and (7) whether the teacher attended primary school in the same county he or she now teaches (1 = yes)

Appendix Table 1b. Test of Essay Quality Treatment Balance (across 7 Covariates)—Pooled Estimates

	Experience of Teacher										
	(1)	(2)	(3a)	(3b)	(3c)	(3d)	(3e)	(4)	(5)	(6)	(7)
Variables	Baseline score	Female	0–2 years	3–5 years	6–12 years	13–20 years	21+ years	College (yes/no)	Chinese major (yes/no)	Urban hukou (yes/no)	Local resident (yes/no)
Essay B is Low Quality (1 = yes)	0.052	0.009	−0.012	0.005	0.021	−0.002	0.038	0.030	0.027	−0.022	0.052
	(0.063)	(0.032)	(0.023)	(0.021)	(0.033)	(0.033)	(0.034)	(0.034)	(0.029)	(0.021)	(0.063)
Constant	0.080*	0.673***	0.128***	0.103***	0.356***	0.351***	0.607***	0.577***	0.754***	0.907***	0.080*
	(0.045)	(0.023)	(0.016)	(0.015)	(0.024)	(0.024)	(0.024)	(0.024)	(0.021)	(0.015)	(0.045)
R-squared	0.001	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.001	0.001	0.001

Notes:

1. Number of observations for each regression = 832. Each model is a separate linear regression of the covariate on one treatment indicator (reference group is grading high quality essay).
2. Robust standard errors in parentheses.
3. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Variables are (1) standardized score of essay A; (2) whether the teacher is female or not (1 = female); (3a-3e) teacher experience; (4) whether the teacher attended any college (1 = yes); (5) whether the teacher's major was Chinese (1 = yes); (6) whether the teacher is registered as an urban resident (1 = yes); and (7) whether the teacher attended primary school in the same county he or she now teaches (1 = yes)

Appendix Table 2. Logistic Regression Predicting Treatment Effects on Teacher Perceptions of Grading Exercise

Outcome:	(1) "I used the same grading standard here as with usual grading"	(2) "Essay topics were extremely similar or identical to usual grading"
Treatments (Compared to Blind Grading):		
Positive Reputational Information (Essay Writer Was Top 25th Percentile)	−0.192 (0.331)	0.298* (0.181)
Negative Reputational Information (Essay Writer Was Bottom 25th Percentile)	0.096 (0.364)	0.428** (0.172)
Baseline Essay Score (Standardized)	0.246* (0.133)	0.204** (0.086)
Constant	2.625*** (0.243)	−0.836*** (0.126)
Mean of Blind Grading Group	0.932	0.305
Margins for Positive Reputational Information Group (at the mean):	−0.013 (0.022)	0.066* (0.040)
Margins for Negative Reputational Information Group (at the mean):	0.004 (0.021)	0.097** (0.040)

Notes:

1. Standard errors adjusted for clustering at the school level in parentheses. N = 832 participants.
2. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix Table 3. Main and Interaction Effects of Grading High or Low-Quality Essay B

Dependent Var: Total Grade of Each Essay	(1) Essay B Score (SDs)	(2) Essay B Score (SDs)
Essay B Is Low Quality	−0.808*** (0.062)	−0.767*** (0.096)
Positive Reputational Information		−0.010 (0.089)
Negative Reputational Information		−0.212** (0.098)
Pos. Rep. Inf. x Low Quality Essay B		0.159 (0.136)
Neg. Rep. Inf. x Low Quality Essay B		−0.286* (0.154)
Baseline Essay Score (Standardized)	0.418*** (0.035)	0.425*** (0.035)
Constant	0.321*** (0.045)	0.373*** (0.035)
R-squared	0.153	0.327

Notes:

1. Standard errors adjusted for clustering at the school level in parentheses. N = 832 participants.
2. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix Table 4. Full Interaction Terms in Pooled Model

	Essay Score (SDs)	Interpretive Notes
Overall Effect of Treatments		
1a. Negative Reputation	−0.0354 (0.0525)	*Effects are relative to blind grading group/ high quality essay B. These terms indicate overall treatment effects* e.g., *how teachers graded Essay A, and coefficients should be null as treatments not active in Essay A.*
1b. Essay B Is Low Quality	−0.00994 (0.0368)	
1c. Essay B Low Quality x Negative Reputation	0.104 (0.0648)	
Effect of Low-Quality Essay B on Subsequent Essays		
2a. Low Quality Essay B x Essay B	−0.676*** (0.0820)	*Captures whether teachers distinguish low from high quality essay B in the absence of reputational information and lingering effects on seeing a low-quality essay B on subsequent essays C and D.*
2b. Low Quality Essay B x Essay C	0.0816 (0.0901)	
2c. Low Quality Essay B x Essay D	0.216** (0.0989)	
Effect of Negative Reputations on Subsequent Essays when Teachers see Contradictory Evidence		
3a. Negative Reputation x Essay B	−0.145 (0.107)	*Captures the effect of seeing negative reputations on high-quality essay B, on subsequent essay grading.*
3b. Negative Reputation x Essay C	−0.232* (0.122)	
3c. Negative Reputation x Essay D	0.0726 (0.130)	
***Additional* Effect of Negative Reputations on Subsequent Essays when Teachers see Confirmatory Evidence (Low-Quality Essay x Negative Reputation)**		
4a. Low Q. Essay B x Negative Rep. x Essay B	−0.494*** (0.156)	*Captures the additional effects from the interaction of low-quality essay B and negative reputational information*
4b. Low Q. Essay B x Negative Rep. x Essay C	−0.210 (0.176)	
4c. Low Q. Essay B x Negative Rep. x Essay D	−0.455** (0.181)	

(*Continued*)

Appendix Table 4. ***continued***

	Essay Score (SDs)	Interpretive Notes
Interaction Term for High Reputational Information	YES	nets out effects of positive reputations, allowing us to focus on interpretation on effect of negative reputations
Essay Number Fixed Effects	YES	nets out differences in essay quality between essays A, B, C, and D
Constant	0.0394 (0.0374)	
Observations	3,328	
R-squared	0.288	

Notes:

1. Standard errors adjusted for clustering at the school level in parentheses.
2. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
3. Reliance on negative reputational information when teachers encounter **contradictory** information for any essay is calculated by the following linear combination: *Negative Reputation x Essay # + Average Effect of Negative Reputation*. In the case of Essay D: 0.0726 + −0.0354 = 0.0372.
4. When teachers encounter **confirmatory** information, further sum all terms where low quality essay B interacts with negative reputation: *Low Q. Essay B x Negative Rep. x Essay # + Negative Reputation x Essay # + Average Effect of Negative Reputation + Average Effect of Negative Reputation x Low Q. Essay B*. In the case of Essay D: −0.455 + 0.0726 + 0.104 + −0.0354 = −0.314.

Appendix Table 5. The Effect of Low-Quality Initial Essay on Subsequent Grading

Variables	(1) Essay B Score (SDs)	(2) Essay B Raw Score	(3) Essay C Score (SDs)	(4) Essay C Raw Score	(5) Essay D Score (SDs)	(6) Essay D Raw Score
Low-quality Essay	−0.808*** (0.062)	−10.07*** (0.771)	0.043 (0.071)	0.466 (0.767)	0.103 (0.072)	1.111 (0.772)
Baseline Essay Score (Standardized)	0.418*** (0.036)	5.213*** (0.446)	0.430*** (0.042)	4.625*** (0.454)	0.282*** (0.039)	3.041*** (0.420)
Constant	0.321*** (0.039)	81.826*** (0.488)	−0.088* (0.050)	80.460*** (0.538)	−0.117** (0.053)	84.865*** (0.568)
Observations	832	832	832	832	832	832
R-squared	0.269	0.269	0.132	0.132	0.067	0.067

Notes:

1. Standard errors adjusted for clustering at the school level in parentheses. N = 832 participants.
2. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix Table 6. Negative Binomial Model of Treatment Effects on Essay Grading Time (in Minutes)

Dependent Var: Number of Minutes Spent Grading	(1)	(2)
Full Treatment Interactions		
Positive Reputational Information	−0.004 (0.041)	0.013 (0.042)
Negative Reputational Information	0.085** (0.041)	0.095** (0.041)
Low Quality Essay B	0.024 (0.041)	0.030 (0.041)
Negative Reputation x Low Quality Essay B	−0.011 (0.058)	−0.022 (0.058)
Negative Reputation x High Quality Essay B	−0.129** (0.057)	−0.128** (0.057)
Essay Fixed Effects (Reference Group = Essay B)		
Essay C	−0.061** (0.029)	−0.060** (0.028)
Essay D	−0.090*** (0.029)	−0.089*** (0.029)
Teacher Characteristics Controlled?	NO	YES
Alpha	−2.579*** (0.101)	−2.620*** (0.104)

Notes:

1. Standard errors adjusted for clustering at the school level in parentheses. N = 2,474 essays.
2. *** p < 0.01, ** p < 0.05, * p < 0.1.