I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12684

Does Increased Teacher Accountability Decrease Leniency in Grading?

Patrick A. Puhani
Philip Yang

OCTOBER 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12684

Does Increased Teacher Accountability Decrease Leniency in Grading?

Patrick A. Puhani
Leibniz Universität Hannover, CReAM, University College London, SEW, University of St. Gallen and IZA

Philip Yang
Eberhard Karls Universität Tübingen, LEAD Graduate School & Research Network

OCTOBER 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Does Increased Teacher Accountability Decrease Leniency in Grading?*

Because accountability may improve the comparability that is compromised by lenient grading, we compare exit exam outcomes in the same schools before and after a policy change that increased teacher accountability by anchoring grading scales. In particular, using a large administrative dataset of 364,445 exit exam outcomes for 72,889 students, we assess the effect of introducing centralized scoring standards into schools with higher and lower quality peer groups. We find that implementation of these standards increases scoring differences between the two school types by about 25 percent.

JEL Classification: H83, I20, I28

Keywords: subjective performance evaluation, rating standards, policy reform, transparency

Corresponding author:
Patrick A. Puhani
Leibniz Universität Hannover
Institut für Arbeitsökonomik
Königsworther Platz 1
30167 Hannover
Germany

E-mail: puhani@aoek.uni-hannover.de

* Part of this research was funded by the Fritz Thyssen Foundation. This research would not have been possible without the onsite data access provided by the Ministry of Culture and Education of the State of Hesse (Hessisches Kultusministerium) in cooperation with the Research Data Center (Forschungsdatenzentrum) of the Statistical Office of the State of Hesse (Hessisches Statistisches Landesamt). We thank two anonymous referees Manuel Boos, Thomas Cornelissen, Marc Deutschmann, Christian Dustmann, Peter Gottfried, Claudia Leiterholt, Alexander Richter, Uta Schönberg, and Klaus F. Zimmermann as well as seminar participants at CReAM, University College London and at the International Economic Association World Congress for helpful comments.

1 Introduction

Although fairness in evaluating student performance on exit exams is crucial for their subsequent access to higher education, training, and career paths, a significant body of literature documents a tendency fosr ratings to be biased upward when conditions allow a large degree of subjectivity (Prendergast and Topel, 1993; Kane et al., 1995; Moers, 2005; Spence and Keeping, 2011; Gollman and Bhatia, 2012). For instance, raters exhibit more leniency as a function of their social proximity to subordinates, out of a simple desire to avoid the confrontation costs of their decisions being challenged (Bernardin, Cooke, and Villanova, 2000), or when ratings also reflect on their own performance (Spence and Keeping, 2011). Rating bias is particularly relevant in the educational setting when teacher discretion facilitates lenient grading practices such as bunching at critical thresholds (see Diamond and Persson, 2016 for Sweden; Dee et al., 2019, for the US). Borcan et al. (2017) even document that a more transparent camera recoding system for high school exit exams in Romania produced larger declines in exam results for poor than for non-poor students.

To address this leniency problem, the education literature mainly considers two strategies for enhancing teacher accountability:[1] external grading and centralization. Whereas the former may eliminate any opportunity to rescore or bunch results when stakes are high (Dee et al., 2019), the latter limits the potential for lenient grading by centralized grading standards, which may in turn even improve the students' subsequent performance in both education and the labor market. For instance, not only is student performance higher in schools with centralized (vs. decentralized) grading systems (Bishop, 1997; Bishop et al., 2000), but states with centralized high school exit exams enjoy greater student learning and

[1] Because of its usefulness in agency relationships (Frink and Klimoski, 1998; Lerner and Tetlock, 1999), "accountability" has received considerable attention from organizational researchers and been studied in various environments, including that of performance ratings (Klimoski and Inks, 1990; Mero and Motowidlo, 1995). Mero and Motowidlo (1995), for example, conceptualizing accountability as the threat of justification, show that undergraduate evaluators rate the simulated performance of videotaped subordinates more accurately when expecting to later justify their ratings to the researchers. Other experiments using undergraduate raters likewise find improved task and context attentiveness in the performance rating process under the accountability condition (Frink and Ferris, 1995; Brtek and Motowidlo, 2002.

teacher effort (Jürges, Büchel, and Schneider, 2005; Jürges and Schneider, 2010), as well as better school grades associated with higher labor market returns (Schwerdt and Woessmann, 2015; Backes-Gellner and Veen, 2008). Nevertheless, little is known about the direct impact of implementing centralized exit exams on teacher grading.

The novel contribution of the present analysis is its determination of the ways in which centralized standard or external grading systems can improve teacher accountability and their association with students' high school exit exam results. The empirical setting is schools in the German state of Hesse (which includes the city of Frankfurt) following the introduction of centralized written exams for all exit-year high school students that comprises three written and two oral sections. Prior to this centralization reform, all teachers involved in the exit exams could submit a set of three questions from which pool the state Ministry of Education would select a subset for the subject exam at that particular school. Following the reform, the ministry not only began setting its own unique written exit exams but issued scoring guidelines for each written subject to be administered in every school in Hesse, thereby anchoring the grading scale and enhancing teacher accountability through improved rating comparability. This reform did not, however, centralize the rating standards for the oral exams, which enables us to conduct a difference-in-differences comparison between treated and untreated exit exams pre and post reform. All through our observation period, that is both before and after the reform, the ministry implemented a form of external grading by having the written tests in a subset of subjects evaluated by a second examiner from another school and the exam score determined by averaging both teachers' scores.

Our first analysis, based on 364,445 written and oral exit exam scores from 72,889 students over 5 years, assesses whether improved teacher accountability via centralized performance standards has affected the average written exit exam performance, irrespective of peer group quality. According to the results, in general, centralization reform has not widened the gap between treated written and untreated oral exams, meaning that it has not changed absolute grading standards. Yet it is also rational to assume that in schools with less proficient students, where teachers have necessarily had to be more lenient in order to award a top grade to at least one student in the classroom, exam centralization might have

diminished such leniency. Conversely, in classrooms of very high achievers, teachers may have been even stricter under the decentralized system than under the new centralized standards. Under either condition, centralization may widen the gap in average written exit exam scores between schools with more and those with less proficient students. Although our analysis confirms this assumption to be true, when we evaluate the centralization effect on untreated oral exams, we find tentative evidence for a *decreasing* gap in oral exam scores between students from schools with higher versus lower quality peers. This observation implies a "John Henry effect" (Krueger, 1999) by which teachers use oral exam scores to compensate for the post-centralization decrease in leniency on written exam scores. In a second set of tests, despite the absence of any exploitable policy reform, we compare exit exam scores in the subjects and years for which the Ministry of Education did or did not mandate external examiners. These analyses reveal that the reform effect on the higher versus lower quality peer gap for treated written exams is about the same size as the external examiner effect on written exam scores.

The remainder of the paper is structured as follows: Section 2 details the policy reform aimed at increasing teacher accountability by centrally standardizing the scoring of written (but not oral) high school exit exams in the German state of Hesse and describes the administrative data used to estimate the reform effects. Section 3 then outlines the empirical approach and reports the estimation results, after which Section 4 concludes the paper.

2 Policy Reform and Administrative Data Source

2.1 The centralization of high school exit exams in Hesse, Germany: A reform to increase performance rating accountability in teachers

The rater accountability reform analyzed here is the introduction of centrally standardized written (but not oral) exit exams in high schools in the state of Hesse with the aim of increasing exam score objectivity and decreasing subjectivity. In Germany, high schools rank students applying for entry into the mostly publicly funded university system using a grade point average (GPA) based on coursework

from the previous two school years (about two thirds of the GPA) and points received in a set of exit exams (another third).[2] The reform was implemented beginning in the 2006/07 school year when Hesse switched from a decentralized system of the state Ministry of Education selecting questions for each school from a teacher-submitted pool to a centralized system in which it sets the same exam for *all* schools across the state. By enforcing centralized standards, this new system increases accountability and result comparability across all schools and teachers.

Nevertheless, because this centralization affects only the written exit exams, it has no direct influence on oral exit exams or past performance evaluations (grades in the previous two school years), which also count toward the overall GPA. In fact, other than introducing centralized questions for the written exit exams, the reform has made no changes to the exit exam structure—which comprises three (now centralized) written tests and two oral exams compiled by course teachers at the respective schools—or to the scoring of these exams by local schools. The major difference is that the centrally set exams are subject to a standardized scoring scheme designed to increase accountability.

Nor has the reform affected another accountability feature in place over our entire (pre- and post-reform) observation period; namely, the selective external scoring of written exit exams in (usually) two subjects determined each school year before the exam period begins (Kultusministerium Hessen, 2015). Even pre reform, the ministry set the teaching curricula and chose exam questions from sets submitted by local teachers, who also served as external scorers (and thus a checks and balance device) for other local schools. Hence, the procedure for written exit exams, even when locally set, still included certain accountability features.

In addition to assessing the general centralization reform effect on scoring by higher or lower performing schools, we also differentiate three major subject groups: (i) natural sciences, technology, and

[2] In our case, a total of 840 points is possible in the overall rating: 300 on the exit exams, 210 on the previous two years of coursework in two major subjects picked by the student, and 330 on the previous two years of coursework in other student-selected subjects. However, because a change in weights occurs three years after introduction of the centralized exit exams, from school year 2009/10 onward, the total points increase to 900: 300 on the exit exams, and 240 and 360, respectively, on the two years of coursework in the two major and other selected subjects. This change slightly decreases the exit exam weights with respect to past performance from 36 to 33 percent.

math (STEM), (ii) social sciences, and (iii) language arts (including German as well as foreign languages). The rationale is that because these subject areas are characterized by differences in performance transparency, they may not lend themselves equally to the enforcement of accountability standards, which could result in different scoring standardization effects on leniency in each area. We also explicitly study the reform effect on math and German, because unlike the other student-selected subject areas of which two must be covered in the written exams, these subjects are requirements for the exit exam either in written or oral form.[3]

2.2 Source of Administrative Data

Relative to the data used in most previous studies, our data set, which encompasses the administrative records of all high school exit exams in Hesse for the 2005/06 (pre-reform) and 2007/2008 to 2010/11 (post-reform) school years, contains a large number of observations. In particular, we use individual level student exit exam and previous two-year performance data for the universe of students and schools across the state. This comprehensive data set is taken from the administrative teacher and student data base (*Lehrer- und Schülerdatenbank*, LUSD), which contains information on all students, teachers, courses, and exit exams in all Hessian schools. Because these LUSD data formally begin with the 2007/08 (post-reform) school year, for our pre-reform data, the Ministry of Education provided us with exam result data for 2005/06, the only such data set available because of a server change in the actual year of system implementation (2006/2007). Unfortunately, however, these 2005/2006 data files include common identifiers for schools but not for students, which prevents us from linking individual student demographics to individual exit exam results and forces us to employ only school-level demographic controls.

Hence, in Table A1, we list the number of observations for high school exit exams in the 2005/06 and 2007/2008 to 2010/11 school years, including the number of students and schools from either the

[3] The five exit exams must cover three subject groups: (i) STEM (ii) social sciences, and (iii) language arts, music, and art. The written exams are set on the two major subjects selected by the student for the two-year coursework phase in senior high school. The third written exam must cover at least two of the above subject groups, and the five exit exams together must include math and German, either written or oral.

main data sources files (see column (2)) or the school census database, which we use to fill in the data made unavailable by the server change (column (1)). Comparing columns (1) and (2) reveals that up until 2007/08 (inclusive), only three quarters of the schools have exam result entries in the main administrative exit exam database, compared to a share higher than 95 percent from school year 2008/09 onward. In columns (3) and (4), we restrict the sample to schools observed over the entire study period that have no missing entries for either past (two-year) performance or exit exam scores, which with yearly variations eliminates about 20 percent of the students. To hold unobserved school characteristics constant (e.g., students, teachers, parental socioeconomic background) and explore the reform-induced within-school variations in ratings over time, we further limit ourselves to the same set of schools across our entire study period. As expected, the number of exam results is five times the number of students.

By reporting the means for our selected variables for both the post and pre reform period (Table A2), we show that, as expected, in the post-reform period of 2007/08 through 2010/11, 60 percent of the exit exams (i.e., the three written tests out of the total five) are centrally administered. As regards student sociodemographics (whose means can be merged with those of the exit exam outcomes on a school level), about 43 percent are male and about 6.5 percent have a foreign passport, less than 2 percent of them from Turkey and slightly over 1 percent from a Western industrialized country. To differentiate the schools into two groups based on peer group quality, we employ only the centralized written exam data for the post-reform years 2007/08 through 2010/11 and define schools with a higher (lower) performing peer group as those that exhibit above (below) average performance on these exams during this period. Our key hypothesis is that in the presence of a teacher tendency to be lenient, our treatment (written exam centralization) should increase the written exam performance gap between these two school types.

3 Empirical Strategy and Estimation Results

As Fig. 1 shows, before written exam centralization, the gap in average scores between the two school types increased from one-fifth of a standard deviation to over one-quarter in the 2007/08 school year (our first post-reform observation point) and remained at this level in subsequent years. We thus

observe an immediate and persistent increase in inequality after teacher accountability increased due to centralization. The difference in the unaffected oral scores, in contrast, remains fairly constant, oscillating around 0.15 standard deviations. Nonetheless, although the figure demonstrates that centralization may widen the scoring gap between higher and lower performing schools by decreasing leniency in the latter, it also exhibits a slight upward trend in outcomes, suggesting that the treatment may not have halted general grade inflation. However, it is impossible to determine whether grade inflation is due to less challenging exit exams over time or whether the general rise in average student scores may stem from other influences, such as advances in teaching methods that have improved student learning or perceptions of increased competition for college places and jobs that are motivating students to study harder.

These Fig. 1 results, however, although based on the same set of schools over the study period, are merely raw comparisons that do not control for potential changes in student composition over time. We thus now perform regression analyses in which the difference-in-differences (DID) estimators are implemented by the interaction effects of higher performing schools and post reform (dummy) variable indicators. The identifying assumption of the DID estimator is that in the absence of treatment (written exam centralization), higher and lower performing schools would have experienced common trends. This assumption could have been violated, however, by compositional (or other) changes affecting the two school sets differently. Because testing this common trends assumption is difficult given our single pre reform period, we address potential compositional changes by also controlling for past performance on the two-years of pre-exit exam coursework. We also include school-level controls for sociodemographics (gender, citizenship, school size), as well as school fixed effects and a dummy for the post reform period.

We first test whether centralization has affected absolute standards by changing scores on the (treated) written exams compared to the (untreated) oral exams, although because centralization involves common standards, it is unclear whether we should expect any change in the average required level. We therefore use the equation below to first estimate the *average* change effected by centralization and then measure the *relative* change between schools of different peer group quality:

$$y_{ist} = \tau(written)_{ist}(post - reform)_t + \gamma_1(written)_{ist} + \gamma_2(post - reform)_t + \beta x_{ist} + \mu_s + \varepsilon_{ist}$$

where *y* denotes the exit exam results for student *i* in school *s* at time *t*, *written* and *post-reform* are dummy variables for written (versus oral) exit exams and the post-reform period (a single dummy for post-reform in the interaction term as well as non-interacted separate year dummies), respectively, and *x* represents the school and sociodemographic control variables (measured on the school level), as well as the student's past coursework performance (measured on the individual level). The treatment effect in this DID model is the coefficient τ of the interaction effect of the dummy variables *written* and *post-reform*. The equation also includes school fixed effects μ_s to control for time-constant unobserved school-specific effects.

According to these DID estimations of the centralization effect on the written versus oral exit exam gap (see Table 1), the point estimate on the interaction of the treatment group indicator (written exit exams) and the post-reform period time dummy variable is a mere, and statistically insignificant, -0.006 standard deviations. Hence, centralization seems not to have altered the absolute standards of the written exit exams. Likewise, separate estimates for the three different subject groups (STEM, social sciences, and language arts), only for language arts do we find a negative and statistically significant (at the 5 percent level) point estimate of -0.038 standard deviations with a standard error of 0.012 standard deviations. The point estimates for the math and German requirements are also statistically insignificant, being both close to zero (last two columns of Table 2), which provides barely any evidence for a centralization effect on absolute scoring outcomes.

Next, to test our main hypothesis that written exam centralization increases the exam outcome gap between higher and lower performing schools, we implement the same type of DID identification strategy as in Section 2.1 but in a regression context. To do so, we use the following estimating equation:

$$y_{ist} = \tau(higher)_{is}(post - reform)_t + \gamma_1(higher)_{is} + \gamma_2(post - reform)_t + \beta x_{ist} + \mu_s + \varepsilon_{ist}$$

where *y* denotes the exam results, *higher* and *post-reform* are dummy variables for higher performing schools and the post-reform period (a single dummy for post-reform in the interaction term as well as

non-interacted separate year dummies), respectively, and *x* denotes the control variables. As before, the treatment effect is the coefficient τ of the interaction effect of the dummy variables *higher* and *post-reform*, and school fixed effects μ_s control for time-constant unobserved school-specific effects.[4]

For the treated written exit exams (top half of Table 2), we expect a positive DID estimate representing a widening gap between higher and lower performing schools as centralization reduces rating leniency through increased accountability. In fact, the estimated coefficient for the written exit ratings overall is 0.046 standard deviations, statistically significant at the 1 percent level, which suggests that increased accountability has indeed widened the gap in written exit exam scores by about a quarter of the initial gap between the two school types. As regards the different subject groups, the increase in the scoring gap is positive and statistically significant for STEM subjects and social sciences, with point estimates of 0.066 and 0.063, respectively (significant at the 1 percent level), but not for language arts, with a statistically insignificant point estimate of 0.005 standard deviations. On the other hand, at 0.037 standard deviations, the point estimate for the German requirement, which does not lend itself easily to accountability enforcement, is only slightly smaller than the 0.045 standard deviations for the math requirement, whose higher performance transparency is assumed to facilitate accountability. However, the former is only statistically significant at the 10 percent level, whereas the latter more precisely estimated and statistically significant at the 1 percent level.

The analogous estimates for the oral exit exams, as anticipated based on their lack of any direct reform effect, are statistically insignificant, with a point estimate of -0.017[5] (bottom half of Table 2), with none of the estimates for the STEM, social science, and math subject areas being statistically significant from zero. For the language arts, however, as well as for the German requirement, the estimates are negative at -0.46 and -0.045, respectively, both significant at the 5 percent level. Taken at face value,

[4] Because the dummy for higher quality peer groups is collinear with the fixed effects, we drop it in the fixed effect regressions. It is also worth noting that the school fixed effects specification controls for potential time-constant peer effects arising from the socioeconomic composition of students in a given school (Eisenkopf et al., 2015).

[5] We also perform separate tests of the effects for each post-reform year, which yield a positive and statistically significant (at the 1 percent level) coefficient for all the (treated) written exams in each post-reform year, but a statistically insignificant coefficient for the (untreated) oral exams except in 2008/09 and 2010/11, where it is negative and significant at the 10 percent level.

these negative estimates imply that written exam centralization has led to a *decrease* in the gap between higher and lower school performance on the oral exams, although these latter have not themselves been affected by centralization. One explanation consistent with these estimates is that teachers evaluating oral language arts exams, for which we expect more scoring leeway than for written exams or exams in subjects such as math, increase their leniency to compensate for the leniency reduction induced by written exam centralization (i.e., a possible John Henry effect; Krueger, 1999). In general, however, our regressions support the hypothesis that centralizing written exit exams increases the performance gap on these exams between higher and lower performing schools.

Although it would also be interesting to determine whether the reform has differential effects based on sociodemographic group, our inability to consider sociodemographics on an individual level (see Section 2.1) prevents us from directly testing whether students with non-Western or Turkish citizenship, or male versus female students, are differently affected than their peers. Instead, we run separate regressions to assess the effects of written exam centralization on students in schools with above (below) average shares of these groups. We find that centralization has resulted in a 0.032 and 0.043 standard deviation relative decrease in standardized scores for students in schools with above average shares of non-Western and Turkish students, respectively, but a 0.029 standard deviation relative increase in schools with above average shares of male students. When interpreting these results, however, it is important to note that schools with above average shares of noncitizens or males may also have citizen and female students with a different (unobserved) sociodemographic background than schools with below average shares of these characteristics. Hence, the Table 3 results cannot simply be interpreted as evidence of leniency toward non-Western and Turkish students or discrimination against male students pre centralization. Rather, non-Western students may be more likely to attend schools with generally disadvantaged students and teachers who are generally lenient toward both disadvantaged German and disadvantaged non-German citizens. Analogously, albeit less intuitively obvious, students in schools with a higher share of females might have more lenient teachers, for example if parents in these schools care more about the education of their children and put more pressure on teachers to be lenient. Unfortunately,

because most regional information is anonymized, the available data do not allow further investigation into this issue. However, when we carry out the analyses for non-Western students separately for the mathematics and German requirement (available in the Online Appendix), the results for mathematics are fairly large at -0.083 and 0.121 standard deviations for non-Western vs. Western and male vs. female students, respectively (standard errors in both cases at 0.018), whereas the estimates for German are insignificant with point estimates close to zero at 0.017 and -0.013 standard deviations for non-Western vs. Western and male vs. female students, respectively (standard error in both cases at 0.024). It thus seems that the effects are mostly driven by mathematics, a subject that lends itself more easily to accountability enforcement.

Another difference in accountability is revealed by the external ratings scheme in place for written exit exams both pre and post reform, under which both a local *and* an outside teacher scores select subject exams. Because the exit outcome for these subjects is the average rating of the two examiners, our data set provides no separate scores for each. Hence, in Table 4, we report regressions that, absent any reform, pool school types and time periods and instead compare the scores for written exams that are affected or unaffected by external examination. These regressions are specifically designed to test the hypothesis that factors which increase transparency and improve measurement (e.g., external examiners) also decrease leniency, implying that the additional scoring serves as a type of increased accountability different from that generated by exam centralization. In fact, all the results reported in Table 4 suggest that external examination decreases the ratings by about 0.056 standard deviations, an effect statistically significant at the 1 percent level and of similar absolute magnitude to the centralization-induced increase in the rating gap between high and low peer group quality schools. This estimate, as the different columns show, is robust to the inclusion of year dummies and past performance on two years of coursework.

4 Conclusions

In this paper, we use the increased rating system accountability generated by the centralization of written school exit exams in the German state of Hesse to estimate this reform's effect on relative rater leniency in schools with higher versus lower peer group quality. Because of the availability of administrative data, we are able to estimate rather precise effects, even in school fixed effect regressions. Overall, we find that the reform-induced increase in accountability has led to a decrease in leniency as reflected by the approximately 25 percent increase in pre-reform ratings gap between higher and lower performing schools. Comparing reform effects for the three different subject groups further reveals that centralization widens this performance gap on STEM and social science but not on language arts, implying that a lack of performance transparency may weaken attempts to increase accountability. For language arts, in contrast, we find that centralization has *decreased* the gap between the two school types on oral exams, which remain unaffected by centralization. We thus observe a John Henry effect (Krueger, 1999) of teachers in lower performing schools compensating for a forced decrease in leniency on written exams with an increased leniency on oral exams. As regards the use of external examiners as an additional means of increasing transparency, the exam scores produced by the local-external rater combination are 0.056 standard deviations lower than the scores given by local examiners only, this estimate is similar in size to the effect estimated for increased accountability through centralization.

References

Backes-Gellner, U., Veen, S. 2008. The consequences of central examinations on educational quality standards and labour market outcomes. *Oxford Review of Education* 34, 569-588.

Bernardin, H.J., Cooke, D.K., Villanova, P. 2000. Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology* 85, 232-234.

Bishop, J.H. 1997. The effect of national standards and curriculum-based exams on achievement. American Economic Review, Papers and Proceedings 87. 260-264.

Bishop, J.H., Moriarty, J.Y., Mane, F. 2000. Diplomas for learning, not seat time: The impacts of New York regents examinations. *Economics of Education Review* 19, 333–349.

Borcan, O., Lindahl, M., Mitrut, A. 2017. Fighting corruption in education: What works and who benefits?. *American Economic Journal: Economic Policy* 9, 180–209.

Brtek, M.D., Motowidlo, S.J. 2002. Effects of procedure and outcome accountability on interview validity. *Journal of Applied Psychology* 87, 185-191.

Dee, T.S., Dobbie, W., Jacob, B.A., Rockoff, J. 2019. The causes and consequences of test score manipulation: Evidence from the New York regents examinations. *American Economic Journal: Applied Economics,* 11, 382–423.

Diamond, R., Persson, P. 2016. The long-term consequences of teacher discretion in grading of high-stakes tests, NBER Working Paper No. 22207, Cambridge, MA.

Eisenkopf, G., Hessami, Z., Fischbacher, U., Ursprung, H.W. 2015. Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland. *Journal of Economic Behavior Organization* 115, 123-143.

Frink, D.D., Ferris, G.R. 1999. The moderating effects of accountability on the conscientiousness-performance relationship. *Journal of Business and Psychology* 13(4), 515-524.

Frink, D., Klimoski, R. 1998. Toward a theory of accountability in organizations and human resources management. *Research in Personnel and Human Resources Management* 16, 1-51.

Jürges, H., Büchel, F., Schneider, K. 2005. The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany. *Journal of the European Economic Association* 3, 1134-1155.

Jürges, H., Schneider, K. 2010. Central exit examinations increase performance … but take the fun out of Mathematics. *Journal of Population Economics* 23, 497-517.

Kane, J.S., Bernardin, H.J., Villanova, P., Peyrefitte, J. 1995. Stability of rater leniency: Three studies. *Academy of Management Journal* 38, 1036-1051.

Klimoski, R., Inks, L. 1990. Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes* 48, 70-88.

Krueger, A.B. 1999. Experimental estimates of education production function. *Quarterly Journal of Economics* 114, 497-532.

Kultusministerium Hessen. 2015. Durchführungsbestimmungen zum Landesabitur 2016, Erlass vom 21. Mai 2015, Ministry of Education of the State of Hesse, Wiesbaden, retrieved August 4, 2016, from its web page: https://kultusministerium.hessen.de/sites/default/files/media/hkm/la16-durchfuehrungsbestimmungen_0.pdf.

Lerner, J.S., Tetlock, P.E. 1999. Accounting for the effects of accountability. *Psychological Bulletin* 125, 255-275.

Mero, N.P., Motowidlo, S.J. 1995. Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psycholo*gy 80, 517-524.

Moers, F. 2005. Discretion and bias in performance evaluation: The impact of diversity and subjectivity. *Accounting, Organizations and Society* 30, 67–80.

Prendergast, C., Topel, R. 1993. Discretion and bias in performance evaluation, *European Economic Review* 37. 355-365.

Schlenker, B.R., Britt, T.W., Pennington, J., Murphy, R., Doherty, K. 1994. The triangle of model responsibility. *Psychological Review* 101, 632-652.

Schwerdt, G., Woessmann, L. 2015. *The information value of central school exams*. CESifo Working Paper No. 5404, Munich.

Spence, J.R., Keeping, L. 2011. Conscious rating distorting in performance appraisal: A review, commentary, and proposed framework for research. *Human Resource Management Review* 21, 85-95.

Tables and Figures

Table 1
Centralization and scoring outcomes

		Subject groups			Requirements	
Exit exam outcomes	All	STEM	Social sciences	Language arts	Math	German
DID (Post-reform*Written)	-0.006	0.003	-0.008	-0.038***	0.010	-0.007
	(0.007)	(0.014)	(0.013)	(0.012)	(0.019)	(0.015)
Observations	364,445	123,319	82,127	116,756	72,897	72,848
R-squared	0.429	0.470	0.453	0.447	0.467	0.424

Notes: This table reports the DID estimates of the reform effect on the difference between written (treated) and oral (untreated) exam outcomes. These differences are estimated for both the set of all exam outcomes ("All") and for the subject groups, including separate estimations for the math and German requirements. All regressions also include student gender and citizenship composition as school-level sociodemographics, school enrollment for the exit exam cohort, school fixed effects, year controls, a control dummy for written exams, past performance, and a constant. Robust standard errors are clustered at the individual level (with each student taking five exams: two written and two oral). *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.
Source: Administrative Teacher and Student Data Set for the State of Hesse 2005/06-2010/11 (*Lehrer- und Schülerdatenbank*, LUSD).

Table 2

Centralization and marking outcomes in schools of different peer group quality: Upper versus lower half of post-reform exam results

		Subject groups			Requirements	
Exit exam outcomes	All	STEM	Social sciences	Language arts	Math	German
Treated performance ratings (written)						
DID (Post-reform*Upper half peer group	0.046***	0.066***	0.063***	0.005	0.045***	0.037*
quality school)	(0.008)	(0.013)	(0.022)	(0.015)	(0.017)	(0.028)
Observations	218,667	97,233	31,643	72,205	57,510	31,581
R-squared	0.442	0.481	0.472	0.473	0.496	0.446
Untreated performance ratings (oral)						
DID (Post-reform* Upper half peer group	-0.017	-0.022	0.014	-0.046**	0.001	-0.045**
quality schools)	(0.011)	(0.025)	(0.018)	(0.020)	(0.034)	(0.021)
Observations	145,778	26,086	50,484	44,551	15,387	41,267
R-squared	0.404	0.423	0.443	0.423	0.404	0.416

Notes: The table shows DID estimates of the reform effect on the difference in exam outcomes between schools in the upper and lower half of post-reform exam results. These differences are estimated for both the set of all exam outcomes (column "All") and the subject groups, including separate estimations for the math and German requirements. All regressions hold constant the effect of student gender and citizenship composition as school-level sociodemographics, school enrollment for the exit exam cohort, school fixed effects, year controls, a control dummy for upper half peer group quality schools, and past performance. The estimate for higher quality peer group schools is dropped due to collinearity with the school fixed effects. Robust standard errors are clustered at the individual level (with each student taking five exams: two written and two oral). *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.
Source: Administrative Teacher and Student Data Set for the State of Hesse 2005/06-2010/11 (*Lehrer- und Schülerdatenbank*, LUSD).

Table 3

Centralization and scoring outcomes in schools with different sociodemographic composition

Exit exam outcomes	Non-Western citizens	Turkish citizens	Male
Treated performance ratings (written)			
DID (Post-reform*Upper half share of)	-0.032***	-0.043***	0.029***
sociodemographic group in school	(0.009)	(0.009)	(0.009)
Observations	218,667	218,667	218,667
R-squared	0.442	0.442	0.442
Untreated performance ratings (oral)			
DID (Post-reform* Upper half share of	0.008	-0.002	-0.004
sociodemographic group in school)	(0.012)	(0.013)	(0.012)
Observations	145,778	145,778	145,778
R-squared	0.404	0.404	0.404

Notes: The table reports the DID estimates of the reform effect on the difference in exam outcomes between schools with the upper and lower half of student shares of non-Western citizens (24 vs. 11 percent, respectively), Turkish citizens (3.9 vs. 0.5 percent, respectively), and male students (50 vs. 37 percent, respectively). All regressions hold constant the effect of student gender and citizenship composition as school-level sociodemographics, school enrollment for the exit exam cohort, school fixed effects, year controls, a control dummy for the upper half share of the respective sociodemographic group in the school, and past performance. Robust standard errors are clustered at the individual level (with each student taking five exams: two written and two). *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.
Source: Administrative Teacher and Student Data Set for the State of Hesse 2005/06-2010/11 (*Lehrer- und Schülerdatenbank*, LUSD).

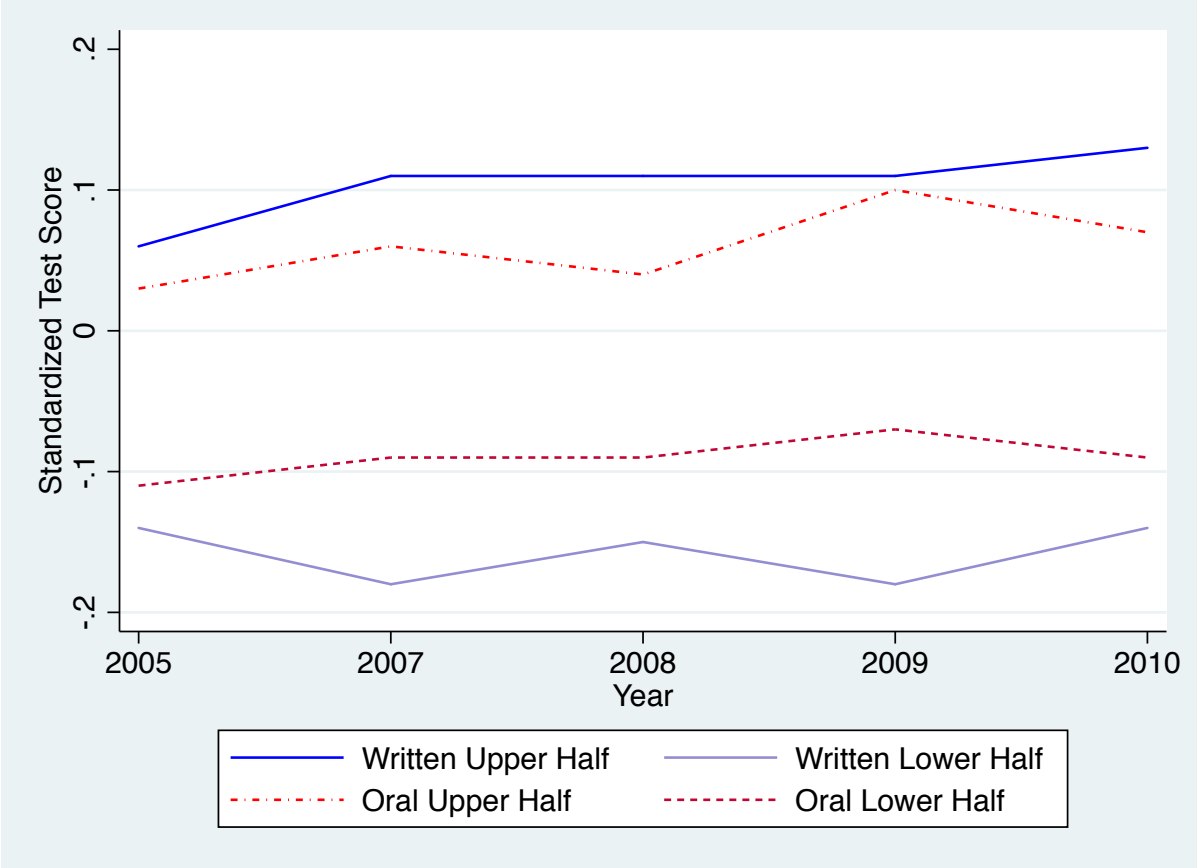
Table 4

External rating and scoring outcomes

Exit exam outcomes	Model 1	Model 2	Model 3
External	-0.056***	-0.055***	-0.056***
	(0.006)	(0.006)	(0.005)
School characteristics	Yes	Yes	Yes
Subject dummies	Yes	Yes	Yes
Year dummies	-	Yes	Yes
Past performance rating	-	-	Yes
Constant	0.394***	0.393***	0.196***
	(0.061)	(0.061)	(0.033)
Observations	218,667	218,667	218,667
R-squared	0.061	0.061	0.488

Notes: The table reports the estimates from a linear regression model of the effect of external evaluation on scoring outcomes. Model 1 estimates this effect while holding school and subject fixed-effects constant, Model 2 adds in year dummies, and Model 3 includes information on past coursework performance. Robust standard errors are clustered at the individual level (with each student taking five exams: two written and two oral). *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.
Source: Administrative Teacher and Student Data Set for the State of Hesse 2005/06-2010/11 (*Lehrer- und Schülerdatenbank,* LUSD)

Figure 1: Exit exam outcomes dependent on school type (upper vs. lower half of post-reform exam results).



Note: Because of a server change, data are available only for the 2005/06, 2007/08, 2008/09, 2009/10, and 2010/11 school years, all designated by the first year only.
Source: Administrative Teacher and Student Data Set for the State of Hesse 2005/06-2010/11 (*Lehrer- und Schülerdatenbank,* LUSD).

Appendix

Table A1

Sample selection

No. in sample	(1)	(2)	(3)	(4)	(5)
Students					
2005-06	17,421	15,569	13,553	13,040	12,982
2007-08	18,835	15,522	14,746	14,025	14,006
2008-09	19,594	19,072	15,460	14,822	14,803
2009-10	20,625	20,102	16,329	15,421	15,396
2010-11	21,076	20,588	16,554	15,730	15,702
Exit exams					
2005-06	-	77,829	67,757	65,200	64,910
2007-08	-	76,970	73,102	70,125	70,030
2008-09	-	90,708	76,668	74,110	74,015
2009-10	-	97,830	80,893	77,105	76,980
2010-11	-	100,248	81,930	78,650	78,510
Schools					
2005-06	203	175	145	145	145
2007-08	203	155	145	145	145
2008-09	203	195	145	145	145
2009-10	204	198	145	145	145
2010-11	208	198	145	145	145

Notes: This table lists the number of exit year high school students, high school exit exams, and schools in the sample. Column (1) gives the approximate number of students and schools from the school census, which we also use to calculate the total number of high schools in the state of Hesse. Column (2) reports the number of students, exit exams taken, and schools from the administrative teacher and student data set affected by the server change (the LUSD). Column (3) limits the observations to schools with available data in every year. Columns (4) restricts the students to only those no missing data on exit exam scores. Column (5) further excludes students with missing information on past performance ratings.
Source: Administrative Teacher and Student Data Set for the State of Hesse 2005/06-2010/11 (*Lehrer- und Schülerdatenbank*, LUSD)

Table A2

Select summary statistics: pre and post reform

Label	Variable description	Post-reform (N = 299.535)		Pre-reform (N = 64.910)	
		M	*SD*	*M*	*SD*
Exam outcomes					
Outcomes	Standardized exit scores	0.005	1.002	-0.024	0.993
Treated	Share of exams treated by the reform	0.600	0.490	0.600	0.490
School characteristics					
Male	Share of male students in schools	0.434	0.093	0.430	0.103
Foreign	Share of non-German students	0.065	0.063	0.068	0.066
Turkey	Share of students from Turkey	0.018	0.022	0.014	0.019
Western	Share of students from Western countries	0.012	0.016	0.013	0.019
Size	Number of students per school/1000	1.179	0.386	1.113	0.382

Source: Administrative Teacher and Student Data Set for the State of Hesse 2007/08-2011/12 *(Lehrer- und Schülerdatenbank,* LUSD).