WORKING PAPERS IN ECONOMICS

No. 10/17

LEROY ANDERSLAND

THE EXTENT OF BIAS IN GRADING

Department of Economics
UNIVERSITY OF BERGEN

The Extent of Bias in Grading

Leroy Andersland[†]

This version: 30 August 2017

Abstract

Do biased perceptions and behaviors affect teachers' assessment of students? To investigate this question, a number of studies use data on two different scores for the same individuals: one non-blind score based on classroom tests assessed by the student's own teacher and one blind test score based on a national exam marked externally and anonymously. In the absence of bias in teachers' assessments, it is argued, there should not be significant differences in the gaps in blind and non-blind scores between different groups. This article present a parsimonious econometric framework that distills out the assumptions necessary to identify group bias in teachers' assessment from such a comparison of blind and non-blind scores. This framework lays the foundation for our empirical analysis, where data from the Norwegian school system are employed to estimate and interpret differences between non-blind and blind assessments. The results show that the relationship between the subject ability and non-blind results tends to be different from the relationship between subject ability and blind results. Evidence of this is found both when grades are recorded when teachers grade the same test and when they grade based on different assessments that are meant to test the same skill. The difference between non-blind and blind will therefore be a function of the skill tested. This leads to different estimates of the group bias when holding ability fixed.

Keywords: Discrimination; bias; human capital; test scores
JEL codes: D80; D63; J15; J16; J24

[†] Department of Economics, University of Bergen, 5020 Bergen, Norway; leroy.andersland@econ.uib.no

1 Introduction

Economists and policymakers are keenly interested in the existence and importance of stereotyping and discrimination by schoolteachers. One question receiving particular attention is whether gender-biased perceptions and behaviors affect teachers' evaluation of students. To answer this question, a number of studies compare teachers' average marking of boys and girls in a classroom exam assessed by the student's own teacher (non-blind scores) to the respective means in a nationally set exam marked externally and anonymously (blind scores). This approach was pioneered in Lavy's (2008) study of gender bias in Israel, and subsequently, it has been applied to data from many other countries (see, for example, Lindahl, 2007; Cornwell, Mustard, & Van Parys, 2013; Burgess & Greaves, 2013).[1] These studies report significant differences across groups in blind and non-blind test scores, and interpret these differences as evidence of stereotyping or discrimination by teachers.

The goal of this paper is to assess whether and in what situations systematic differences between non-blind and blind assessment across groups can be interpreted as evidence of stereotyping or discrimination by teachers. We focus on two types of data generating processes of the blind and non-blind scores. The first type occurs when the student's own teacher and an external examiner are marking the *same* test. As in most previous studies, the second is a data-generating process in which the student's own teacher and an external teacher are marking *different* tests that are meant to measure the student's knowledge of the same material. We present a parsimonious econometric framework that shows, for each data-generating process, the assumptions under which one can draw causal inferences about bias in teachers' assessment from a comparison of blind and non-blind test

[1] Differences between non-blind and blind assessment across groups have been used to measure discrimination or stereotypes in several other settings (see, for example, Blank, 1991; Goldin & Rouse, 2009). An alternative approach to measuring discrimination or stereotyping is to randomly assign certain characteristics (e.g., gender) to students' exam scripts (Hanna & Linden, 2009; Sprietsma, 2013) or job applications (Bertrand & Mullainathan, 2004).

scores. This framework lays the groundwork for our empirical analysis, where data from the Norwegian school system is employed to estimate and interpret differences between non-blind and blind assessment of students.

Importantly for our analysis, the Norwegian data offer information on two sets of blind and non-blind scores. One set of scores is generated by assessment of the same test by examiners that do not know the identity of the student and the student's own teacher. The other set of scores comes from assessment on different tests (testing the student's knowledge of the same material) by external examiners and the student's own teacher. As in previous studies, the results show that the scores of boys and girls differ significantly in the non-blind classroom assessments marked by the student's own teacher as compared to the scores in a nationally set exam marked remotely and anonymously by an external examiner. If data from two evaluations of the same test are used, a similar difference appears, though it is not statistically significant. A possible explanation for a potential difference between the two types of data is that females tend to perform better than boys in classroom tests assessed by their own teacher as compared to nationally set exams marked by an external examiner. Another is that female students are better at a potential skill only tested in teacher assessment compared to boys. The result shows that the relationship between subject ability and non-blind grades is different from the relationship between subject ability and blind grades. This is found even when teachers grade the same exam. This leads to different estimates of the group bias when holding ability fixed.

The remainder of the paper proceeds as follows. The next section provides background on the Norwegian school system, discusses how exams are set and assessed, and describes our data. Section 3 presents the econometric framework, laying out the possible sources of differences in blind and non-blind test scores. Section 4 describes and discusses our findings, and the final section offers some concluding remarks.

2 Institutional Background and Data

This analysis employs data comparing exams that are graded externally and anonymous with local teacher evaluations. These records will be referred to as the administrative data. In addition, we have been given access to files from experiments in two different areas comparing the same test graded anonymous and by the students' teacher. These records are called the non-administrative data. This section gives an overview of the education system, focusing on the importance of the tests, how grading is undertaken, and a data and variable description.

2.1 The Norwegian Education System

The Norwegian pre-college education system consists of primary school (level 1-7), lower secondary school (level 8-10), and upper secondary school (level 11-13). Both primary and lower secondary schools are compulsory. The majority of students attend a public institution, and even private institutions are funded and regulated by the Ministry of Education and Research. There are generally no tuition fees.

Norwegian municipalities operate primary and lower secondary schools. At the primary school level, all students are allocated to schools based on fixed school catchment areas within municipalities. With the exception of some religious schools and schools using specialized pedagogic principles, parents are not able to choose the schools to which their children are sent (except by moving neighborhoods). There is a direct link between elementary school attendance and attendance at middle or lower secondary schools (ages 13–16/grades 8–10), in that elementary schools feed directly into lower secondary schools. In many cases, primary and lower secondary schools are also integrated. At the end of middle school, students are evaluated both non-anonymously by their teachers for most subjects taught in school, and in addition anonymously and externally in one to two central exit

exams.

At the end of 10th grade, students apply for upper secondary school. The high schools have two main tracks, vocational and academic. They are administered at the county level (above the level of municipalities) and are not mandatory in Norway, although, since the early 1990s, everybody graduating from middle schools is guaranteed a slot in high school.

Admissions procedures differ across counties for upper secondary schools. In most counties, students can freely choose schools, but in others, children are allocated to schools based on well-defined catchment areas, or high school zones. In both regions we focus on, students are free to choose schools within their regions. This means that middle school grades are important for intake to schools and tracks where there is competition.

At the middle school level, the final Gradepoint is based on teacher evaluations of in-school performance, as well as central exams. The Gradepoint summarizes student performance at school, and is used for track and school placement later. Both oral and written performance are assessed in some subjects, and both oral and written exams are given. Our data show that, in the period 2000–2010, students had on average 14.0 teacher-given grades and 1.37 written exam grades and 1.0 oral exam grades in middle school.[2] In middle school, the Gradepoint consists of the average grade times 10, where all topics (exams and in-school assessments) have a grade between 1 and 6. A new Gradepoint is calculated at the end of high school, and is the average grade on all high school exams and subjects times 10. In addition to Gradepoints, points are given according to specific criteria to make up the final measure that determines school and track placement at post-secondary education. It is also common to attach high school certificates showing grades to job applications.

[2] In the same period, students had on average 22.4 teacher-given grades and 6.5 exams (oral and written) at academic track high schools.

2.2 Grading

Grading principles are set by the *Education Act* (*Opplæringslova*). It is stated that teacher course evaluations shall be based on to what degree students have achieved the competence goals, stated by the subject-specific and nationally set learning goals. For most subjects, the final teacher evaluation grade is set based on achieved competence in the late spring each year. Notably, it is specifically stated that student behavior (*orden og oppførsel*) is not to be reflected in grading, and (of course) that student background should not count in grading. Effort is allowed to be included in grading in gymnastics. Teacher course assessment grades are set before the grading of exams. Normally, schools have a local test called *Tentamen* near the end of each semester in middle school. It is an important part of the teacher's final evaluation.

Students do not have an exam in each subject. At the end of middle school, students are drawn to take Norwegian, Math, or English written exams. Students drawn to Norwegian perform two exams, one for each official written language (Bokmål and Nynorsk). The written exam is nationally prepared and corrected by two sensors that are external to the school and who do not know the identity of the student. Students are also drawn to perform an oral exam in any subject, which is administered at the local level. The exams are part of the evaluation of the students' achieved competences in a subject according to the centrally set learning goals. The learning goals have an oral component in some subjects. We focus on teacher assessments in two subjects, Math and written Norwegian, where the oral component does not matter. Thus, teacher assessments and the national exams we use are supposed to test the same skills.

For two regions, Rogaland and Bergen, we have the non-administrative datasets. In the spring of 2015, the school authorities in the municipality of Bergen conducted an experiment on all students at middle schools in Bergen. For the high-stakes *Tentamen* at the

end of 10th grade, an additional teacher graded the tests anonymously, in addition to the students' teacher in that topic. All students take the *Tentamen* in Math, English, and Norwegian, but it varies by class in which subject an additional teacher graded the test anonymously. We have information on the gender of the students, as well as whether they are immigrants or not for a part of this sample. The teacher that was to grade the test anonymously was another teacher at the same school. Therefore, it is likely that all teachers knew that this experiment took place.

For Rogaland, we have a similar dataset at the high school level. Here, a student's tests were graded both by teachers at the same school and by an external group of examiners elected by the county-level school authorities. The test is a locally administered end-of-year exam. In contrast to *Tentamen*, the grade on this exam appears as a separate grade on the students' certificates and counts in calculating their Gradepoints. In addition, students' names do not appear on their exams. This is different from centrally administered exams in that the students' local teachers participate in both making and grading the exam.[3] In addition to the external group of examiners that grade the tests anonymously for the blind evaluation, two teachers grade the locally administered exams for the non-blind evaluation, one of which is the student's teacher. The other is a teacher external to the school. The procedure in the first year, 2010, was that 6 schools were drawn to provide 10 exams each (at random) and submitted to the school authorities. Then a group of external examiners were chosen to grade the exams. Half of the tests this year were in Mathematics and half were in Norwegian. In 2012 and 2013, the experiment was followed up and extended to include more schools. In these years, the schools were also randomly drawn. We do not have any observable characteristics of the students for this dataset. Nevertheless, we can still examine the pattern in non-blind and blind scores, and compare it to the results from in the administrative data.

[3] In Math one part of the exam is made at the county level, and one part on the school level. In Norwegian the whole exam is made on the county level.

The experiment in Bergen was performed at the middle school level, while the Rogaland experiment was performed at the high school level. The comparable administrative data is at the same school level. As discussed in this section, there are some differences between scores within the non-blind and blind definitions. Table 1 summarizes institutional details about the grader, number of graders, etc.

[Table 1]

One thing to note from Table 1 is that in the data from Rogaland, the blind grade in the administrative data is the same as the non-blind in the non-administrative dataset. For the administrative data, the score on the local exam is defined as blind since the name does not appear on the test, while the teacher knows the student's identity for course assessments. In the non-administrative data, the same score on the local exam is non-blind since the student's teacher grades the exams, while external examiners give the blind scores.

Since 2012, the standard has been that exams in Norwegian are written on a computer, while exams are written on both paper and digitally in Mathematics. For the experiment conducted in Bergen, the Norwegian tests are written digitally, while the Mathematics tests are written on paper.

Failing a course assessment or an exam (local or external) in middle school, the student will still be able to attend high school, but it may have consequences for the student's options regarding track and school placement. If the student fails a compulsory course assessment or compulsory exam in high school, he or she will not be able to complete the education in that track. Failing the *Tentamen* does not have any direct consequences other than being negative for the course assessment.

2.2.1 Variable Definitions

For the administrative dataset, we are able to match middle and high school grades to

register-based files. Students are defined as having a low socio-economic-status (SES) if none of their parents have completed college/university and the father has earnings below the 50th percentile in the income distribution of fathers in the sample. Students are categorized as immigrants if they have one or two parents born in a non-Western country. The register-based files also provide information on the student's gender. The grade files provide information on which school the student attended for each year the grade is registered.

We have split the administrative data into three main samples. The first is a sample of grades given to students at the same schools, years, subjects, and level as in the experiments. The second is a sample of grades from all middle schools/high schools in Bergen/Rogaland from the same years, subjects, and level, while the third contains grades from these areas given in the period 2008–2015 for the same subjects and levels. The grades in the non-administrative data are from 2015 in the Bergen experiment, while grades are from 2010, 2012, and 2013 in the Rogaland experiment.

School administrators supplied data from the experiments directly to us. In the Bergen experiment, we were able to derive grades, school, year, gender, class, subject, and a personal identifier. For the Rogaland experiment, in addition to grades, we have information on the school, subject, and year.

3 Setup

3.1 Notation and Modeling

3.1.1 Data-Generating Processes

The data are from two different data-generating processes.[4] The first is the non-administrative data, which are based on the experiments that assigned the same test in Bergen to be graded by different examiners. Here, we can observe student i at school (or class) s

[4] The model is explained in terms of the institutional setup for Bergen, since it is here we compare non-administrative and administrative group coefficients. Differences in the setup for Rogaland are presented in the institutional details and data section, and will be discussed in the results section.

taking only one test. The grade student i receives from her teacher is Y_i^n (non-blind grading result), whereas the one from the other grader is Y_i^b (blind grading result of the same exam). Therefore, we define the grade difference using non-administrative data as Δ_i^e, where

$$\Delta_i^e = Y_i^n - Y_i^b.$$

The second set of records is the administrative data. For this data-generating process, we can observe student i at school (or class) s now taking two different tests: a blind exam and a teacher assessment at her own school, which is graded by her own teacher. The grade student i receives from her teacher is $\tilde{Y}_i^n$, whereas the grade from the external graders is Y_i^b. Therefore, we define the grade difference using administrative data as Δ_i^o, where

$$\Delta_i^o = \tilde{Y}_i^n - Y_i^b.$$

3.1.2 Grades Given by Students' Own Teacher in Non-Administrative Data

Let us assume that Y_{is}^n , the grade given by the teacher in the non-administrative data, can be written as

$$Y_{is}^n = t(X_{is}) + \varrho\theta_{is} + \varepsilon_{is}^n.$$

The function $t(\cdot)$ expresses how the teacher at school (class) s affects student i's grade. We assume that $t(\cdot)$ has the following functional form:

$$t(X_{is}) = t(G_{is}, \kappa_{is}, \bar{\theta}_{is}) = \alpha + \beta G_{is} + \gamma\kappa_{is} + (1 - \varrho)\bar{\theta}_{is}.$$

The function $t(\cdot)$ is the **biased grading function**, or simply, **bias**. It describes how teachers bias grades according to student characteristics. The variable X_{is} is a vector that contains G_{is}, which are some observable characteristics to the researcher and the teacher. κ_{is} represents student behavior in class, and $\bar{\theta}_{is}$ is a compound of other information about the students that the teacher uses to grade. In particular, $\bar{\theta}_{is}$ is, for example, other student abilities/behavior, grades in other subjects, or previous grades. κ_{is} and $\bar{\theta}_{is}$ are not necessarily observable to the researcher. The variable θ_{is} is the true ability being measured. The parameter ϱ reflects the

relationship (mapping) of that ability to the score given by the teacher. ε_{is}^{n} is an error. The parameter α captures grade inflation, β captures discrimination in favor of groups of students with observable characteristics G, and γ and $(1-\varrho)$ capture the effect of components that are unobservable to us but that are used by the teacher when grading exams.

There are two components of ε_{is}^{n}. The first one, ξ_{is}^{n}, is specific to the grader when assigning a grade to student i. The second one is a component reflecting the student's idiosyncrasy, ϵ_{is}, which is not related to the grader. For example, ϵ_{is} may be any deviation (luck, not feeling well on the day of the internal exam, etc.) that makes the student's grade not reflect exactly his or her level of ability θ_{is}. Thus,

$$\varepsilon_{is}^{n} = \xi_{is}^{n} + \epsilon_{is}.$$

For those reasons, we rewrite the previous equation for Y_{is}^{n} as

$$Y_{is}^{n} = \alpha + \beta G_{is} + \gamma \kappa_{is} + (1-\varrho)\bar{\theta}_{is} + \varrho\theta_{is} + \xi_{is}^{n} + \epsilon_{is} \quad (1).$$

3.1.3 Grades Given by Students' Own Teachers in Administrative Data

Let us assume that $\tilde{Y}_{is}^{n}$, the grade given by the teacher in the administrative data, can be written as

$$\tilde{Y}_{is}^{n} = t(X_{is}) + \tilde{t}(X_{is}) + \rho\theta_{is} + (1-\rho)\tilde{\theta}_{is} + \tilde{\varepsilon}_{is}^{n}.$$

$t(X_{is})$ is the biased grading function in the administrative data. The parameter ρ reflects the relationship of ability θ_{is} to the score given by the teacher. $\tilde{\theta}_{is}$ measures a compound of other abilities that are captured by the teacher grade in administrative data. Note that an important difference here from the non-administrative data is that we do not include this term in the biased grading function. This is because, in the administrative data, the two different tests can actually measure different subject skills. The function $\tilde{t}(X_{is})$ explains why some students perform relatively better under in-class tests graded by the teacher. Finally, $\tilde{\varepsilon}_{is}^{n}$ is some variation, containing, for example, an error that is due to the grader, $\tilde{\xi}_{is}^{n}$, and another coming

from the student, $\tilde{\epsilon}_{is}$, as he or she may have different performance at another time due to various causes. That is,

$$\tilde{\varepsilon}_{is}^{n} = \tilde{\xi}_{is}^{n} + \tilde{\epsilon}_{is}.$$

Let $t(X_{is})$ be

$$t(X_{is}) = t(G_{is}, \kappa_{is}) = \alpha + \beta G_{is} + \gamma \kappa_{is}.$$

We write $\tilde{t}(X_{is})$ as

$$\tilde{t}(X_{is}) = \tilde{\alpha} + \tilde{\beta} G_{is} + \tilde{\gamma} \kappa_{is}.$$

$\tilde{Y}_{is}^{n}$ is then

$$\tilde{Y}_{is}^{n} = \alpha + \beta G_{is} + \gamma \kappa_{is} + \tilde{\alpha} + \tilde{\beta} G_{is} + \tilde{\gamma} \kappa_{is} + (1-\rho)\tilde{\theta}_{is} + \rho \theta_{is} + \tilde{\xi}_{is}^{n} + \tilde{\epsilon}_{is} \quad (2).$$

3.1.4 Grades Given by External Reviewers

The grade given on the exam from the external grader is Y_{is}^{b}.

$$Y_{is}^{b} = \theta_{is} + \varepsilon_{is}^{b}.$$

We then write

$$\varepsilon_{is}^{b} = \xi_{is}^{b} + \epsilon_{is},$$

where ξ_{is}^{b} is the measurement error that is specific to the external evaluator when assigning a grade to student i and ϵ_{is} is the same term that explains deviations between grades and skills that appeared as a component of ε_{is}^{n}. We therefore rewrite the equation for Y_{is}^{b} as

$$Y_{is}^{b} = \theta_{is} + \xi_{is}^{b} + \epsilon_{is} \quad (3).$$

3.2 Parameters of Interest

The biased grading function $t(\cdot)$ is unknown and is the main object of interest. We want to learn how teachers distort grades. For example, as in Lavy (2008), do teachers favor girls? Or is it another reason for this difference, as suggested in Hinnerich, Hoglin, and Johanneson (2011).

Identification of the parameters α, β, and γ is not feasible without imposing some untestable assumptions. For example, we do not observe κ_{is}, $\tilde{\theta}_{is}$, or $\bar{\theta}_{is}$, which may be arbitrarily correlated with G. However, the relevance for explaining outcome differences between groups of separating out the effect of those variables is not clear, as all can have an effect on future outcomes. In what follows, we show what can be identified from the non-administrative data we have available. We also show what under different assumptions can be identified by administrative data. The main threat to identifying relevant bias in the administrative data would be the function $\tilde{t}(X_{is})$ and the difference between $\tilde{\theta}_{is}$ and $\bar{\theta}_{is}$. If some students perform better at in-class exams, or if the teacher assessments and national exams actually tests different skills, then this should not be characterized as bias.

3.3 Identification Using the Non-Administrative Data

3.3.1 Identification Using Non-Administrative Data: $\varrho = 1$

We have that the variable that measures differences in grades, Δ_i^e, in the non-administrative data can be written as

$$\Delta_{is}^e = Y_{is}^n - Y_{is}^b = t(X_{is}) + \tau_{is} \quad (4),$$

where

$$\xi_{is}^n - \xi_{is}^b = \tau_{is}$$

captures differences in error terms coming from the fact that grades are given by two different people (teacher, ξ_{is}^n, and external reviewer, ξ_{is}^b) for the same exam. We assume that the error τ_{is} is idiosyncratic and not related to any of the other variables on the right-hand side. The differences in grades are equal to $t(\cdot)$ plus the unobserved component τ_{is}:

$$\Delta_{is}^e = \alpha + \beta G_{is} + \gamma \kappa_{is} + \tau_{is}.$$

Identification of the parameters α, β, and γ is not possible without further assumptions because G and κ_{is} are arbitrarily correlated. In this case, the unobserved

component τ_{is} is uncorrelated to the function $t(X_{is})$ and is not the reason the structural parameters of the biased grading function are not identified. Even though we cannot identify α and β, we can identify the parameters of the regression of Δ_{is}^{e} on G:[5] [6]

$$\bar{\beta} = \frac{Cov(\Delta^{e}, G)}{Var(G)} = \beta + \gamma \frac{Cov(\kappa, G)}{Var(G)} \qquad (5)$$

The parameter $\bar{\beta}$ can be interpreted as the total effect of a given characteristic G on the differences in grades. For example, suppose that teachers do not favor girls ($\beta = 0$), but that girls are typically better-behaved in class than boys and that teachers reward girls for their behavior. Thus, $Cov(\kappa, G)$ and γ are both positive. In that case, $\bar{\beta}$ is positive even though β equals zero. Nevertheless, given that G and κ are correlated, any intervention that tries to minimize bias in grading will necessarily be a policy whose overall effect will be measured in terms of $\bar{\beta}$, not β. The intercept $\bar{\alpha}$ can be written as

$$\bar{\alpha} = E(\Delta^{e}) - \bar{\beta}E(G) = \alpha - (\beta + \gamma \frac{Cov(\kappa, G)}{Var(G)})E(G) \qquad (6).$$

Again, the mean bias, α, is not identifiable, but the parameter that will be used to measure the effectiveness of bias on outcomes is not α, but $\bar{\alpha}$.

3.3.2 Identification Using Non-Administrative Data: $\varrho \neq 1$

The difference is then equal to the function of interest, $t(\cdot)$, a function of the skills being measured. Moreover, the unobserved component τ is:

$$\Delta_{is}^{e} = \alpha + \beta G_{is} + \gamma \kappa_{is} + (\varrho - 1)\theta_{is} + (1 - \varrho)\bar{\theta}_{is} + \tau_{is} \qquad (7).$$

Even though we cannot identify α and β, we can identify the parameters estimated by a regression of Δ_{e} on G:

[5] When G is a vector, the usual matrix notation has to be employed. We present the simple regression algebra just to facilitate the exposition of the argument.

[6] The notation in this analysis is based on having a sample of the full population. Since we only have samples in the empirical section, the more precise notation would specify that the expressions are probability limits of estimators.

$$\bar{\beta} = \frac{Cov(\Delta^e, G)}{Var(G)} = \frac{Cov\big(\alpha + \beta G_{is} + \gamma\kappa_{is} + (\varrho - 1)\theta_{is} + (1 - \varrho)\overline{\theta}_{is} + \tau_{is}, G\big)}{Var(G)}$$

$$\bar{\beta} = \frac{Cov(\Delta^e, G)}{Var(G)} = \beta + \gamma\frac{Cov(\kappa, G)}{Var(G)} + (\varrho - 1)\frac{Cov(\theta, G)}{Var(G)} + (1 - \varrho)\frac{Cov\big(\overline{\theta}, G\big)}{Var(G)} \quad (8).$$

The parameter $\bar{\beta}$ will then consist of gender bias, differences in behavior correlated with gender, and a function of how gender is correlated with the different skills and information that teachers use in setting grades. Importantly, note that, in this case, it is not obvious that $\bar{\beta}$ is the only parameter of interest for evaluating how the total amount of bias affects student outcomes. In particular, it is interesting to know, for a given ability in a subject, the total amount of bias one group receives compared to another. The alternative parameter of interest would be:

$$\bar{\bar{\beta}} = \frac{Cov(\Delta^e, G|\theta)}{Var(G)} = \frac{Cov\big(\alpha + \beta G_{is} + \gamma\kappa_{is} + (\varrho - 1)\theta_{is} + (1 - \varrho)\overline{\theta}_{is} + \tau_{is}, G|\theta\big)}{Var(G)} \quad (9).$$

One way of obtaining an estimate of this would be to insert θ into the right-hand side of Equation (7), using Y_{is}^b:

$$Y_{is}^b = \theta_{is} + \xi_{is}^b + \epsilon_{is}$$

$$\theta_{is} = Y_{is}^b - \xi_{is}^b - \epsilon_{is}$$

Inserting into Equation (7):

$$\Delta_{is}^e = \alpha + \beta G_{is} + \gamma\kappa_{is} + (\varrho - 1)Y_{is}^b + (1 - \varrho)\overline{\theta}_{is} + \xi_{is}^n - \varrho\xi_{is}^b - (\varrho - 1)\epsilon_{is} \quad (10).$$

Because the errors in $-\varrho\xi_{is}^b - (\varrho - 1)\epsilon_{is}$ are correlated with Y_{is}^b, a regression of Δ_{is}^e on G_{is} and Y_{is}^b would not yield the parameter of interest:

$$\ddot{\beta} = \frac{Cov\left(\Delta^e, G|Y^b\right)}{Var(G)} \neq \bar{\bar{\beta}} = \frac{Cov(\Delta^e, G|\theta)}{Var(G)} \quad (11).$$

A solution to this problem is to obtain an unbiased estimate of $(\varrho - 1)$ and fix this parameter

in the estimation of Equation (7).

3.4 Using Administrative Data

For the administrative setting, differences in grades can now also be explained by differences in test-type specific performance and differences in the skills that the assessments measure. The grade difference can now be written as:

$$\Delta_i^o = \tilde{Y}_i^n - Y_i^b$$

$$= t(X_{is}) + \tilde{t}(X_{is}) + (\rho - 1)\theta_{is} + (1 - \rho)\tilde{\theta}_{is} + \tilde{\varepsilon}_{is}^n - \varepsilon_{is}^b$$

$$= \alpha + \tilde{\alpha} + \left(\beta + \tilde{\beta}\right)G_{is} + (\gamma + \tilde{\gamma})\kappa_{is} + (\rho - 1)\theta_{is} + (1 - \rho)\tilde{\theta}_{is} + \tilde{\xi}_{is}^n - \xi_{is}^b + \tilde{\epsilon}_{is} - \epsilon_{is}$$

In this case it is important to notice that:

- Although $t(X_{is})$ and $\tilde{t}(X_{is})$ are functions of observable (G) and unobservable κ, they have different interpretations. So, a general function $g(X_{is}) = t(X_{is}) + \tilde{t}(X_{is})$ is not measuring biased grading, but the biased grading effect over the fact that some groups of students (e.g., females) perform relatively better under in-class exams than under external exams. Therefore, we cannot necessarily claim that $g(X_{is})$ is biased grading.
- If ρ is different from 1, the grade difference is a function of the competence level of the skill being evaluated. The closer to 1, the smaller the effect of subject-specific ability on the grade difference. Thus, for ρ < 1, and as in the non-administrative setting, differences in grades Δ_i^o will depend directly on skills being measured. In contrast to the non-administrative setting, the reason for ρ < 1 is not only different grading practices between external and internal teachers, but could also be due to tests measuring different subject skills.
- Unlike in the non-administrative setting, $\tilde{\epsilon}_{is} \neq \epsilon_{is}$, as these two objects come from different exams and luck or feeling ill on an exam day may differ across days.

In what follows, we impose some assumptions that allow us to identify parameters related to the biased grading function using administrative data.

3.4.1 Identification Using Administrative Data: $\rho = 1$

We have that

$$\Delta_i^o = \tilde{Y}_i^n - Y_i^b = g(X_{is}) + \tilde{\tau}_{is} \quad (12)$$

$$= \alpha + \tilde{\alpha} + \left(\beta + \tilde{\beta}\right)G_{is} + (\gamma + \tilde{\gamma})\kappa_{is} + \tilde{\tau}_{is},$$

where

$$\tilde{\tau}_{is} = \tilde{\xi}_{is}^n - \xi_{is}^b + \tilde{\epsilon}_{is} - \epsilon_{is}.$$

As with non-administrative data, we assume that $\tilde{\tau}_{is}$ is idiosyncratic, and that $\tilde{\tau}_{is}$ and X are independent. Thus, one can identify the coefficients of a regression of Δ_i^o on G, exactly as in Equations (5) and (6). The key difference here is that the interpretation of these coefficients would be different, since $\tilde{t}(\cdot)$ is not null. Specifically, they will reflect both bias and differences coming from different test types. Note that both types may explain outcome differences between groups. However, this combined effect could rather be described as the effect of grading schemes rather than bias.

3.4.2 Identification Using Administrative Data: $\rho = 1$ and $\tilde{\boldsymbol{t}}(\cdot) = 0$

We have that

$$\Delta_i^o = \tilde{Y}_i^n - Y_i^b = t(X_{is}) + \tilde{\tau}_{is} \qquad (13).$$

In this case, the parameters of equations Equations (5) and (6) could be identified.

3.4.3 Identification Using Administrative Data: $\rho \neq 1$ and $\tilde{\boldsymbol{t}}(\cdot) = 0$

We have that

$$\Delta_i^o = \tilde{Y}_i^n - Y_i^b = \alpha + \beta G_{is} + \gamma\kappa_{is} + (\rho - 1)\theta_{is} + (1 - \rho)\widetilde{\theta}_{is} + \tilde{\xi}_{is}^n - \xi_{is}^b + \tilde{\epsilon}_{is} - \epsilon_{is} \quad (14).$$

Even though we cannot identify α and β, we can identify the parameters of the regression of Δ_i^o on G:

$$\bar{\beta}' = \frac{Cov(\Delta^o, G)}{Var(G)} = \beta' + \gamma' \frac{Cov(\kappa, G)}{Var(G)} + (\rho - 1)\frac{Cov(\theta, G)}{Var(G)} + (1 - \rho)\frac{Cov(\tilde{\theta}, G)}{Var(G)} \quad (15)$$

The parameter $\bar{\beta}'$ will then consist of gender bias, differences in behavior correlated with gender, and a function of how gender is correlated with the different skills and information that teachers use in setting grades. Again, it is not obvious that $\bar{\beta}'$ is the only parameter of interest for evaluating how the total amount of bias affects student outcomes. In particular, it is interesting to know, for a given ability in a subject, the total amount of bias one group receives compared to another. An alternative parameter of interest would be:

$$\bar{\bar{\beta}}' = \frac{Cov(\Delta^o, G|\theta)}{Var(G)} \quad (16)$$

A way to estimate total amount of bias conditional on subject-specific ability is to use the blind score:

$$Y_{is}^b = \theta_{is} + \xi_{is}^b + \epsilon_{is}$$

$$\theta_{is} = Y_{is}^b - \xi_{is}^b - \epsilon_{is}$$

Inserting into Equation (14):

$$\Delta_{is}^o = \alpha + \beta G_{is} + \gamma \kappa_{is} + (\rho - 1)Y_{is}^b + (1 - \rho)\tilde{\theta}_{is} + \tilde{\xi}_{is}^n + \tilde{\epsilon}_{is} - \rho(\xi_{is}^b - \epsilon_{is}) \quad (17)$$

Because the errors in $-\rho(\xi_{is}^b - \epsilon_{is})$ are correlated with Y_{is}^b, a regression of Δ_{is}^o on G_{is} and Y_{is}^b would not yield the parameter of interest:

$$\ddot{\beta}' = \frac{Cov\left(\Delta^e, G|Y^b\right)}{Var(G)} \neq \bar{\bar{\beta}}' = \frac{Cov(\Delta^e, G|\theta)}{Var(G)} \quad (18).$$

A solution to this problem is to obtain an unbiased estimate of $(\rho - 1)$ and fix this parameter in the estimation of Equation (17).

3.5 Comparing Non-Administrative with Administrative Data

Under certain assumptions, administrative data may not be useful for testing for the existence of biased grading. A potential reason for that has to do with the fact that blindly and non-blindly graded exams may differ because these are two different tests. Thus, it is likely that the abilities being measured may be different ($\tilde{\theta} \neq \bar{\theta}$), or that the systematic reaction to the exam may be different ($\tilde{t}(X) \neq 0$). These factors are the main potential reasons resulting from administrative data but do not necessarily identify the same objects as results from non-administrative data. The next section will provide evidence on the difference in estimates produced when using data based on the same exam and data based on different assessments meant to test the same skills.

4 Results

4.1 Descriptive Statistics

Table 2 presents summary statistics of blind and non-blind grades for both non-administrative and administrative datasets. Grades are reported for Math and Norwegian, separately. We also report proportions of students by gender, immigration status, and SES.

[Table 2]

Each column presents the different samples used in the analysis. Column (1) shows summary statistics for the Bergen experiment, while Column (2) shows statistics for middle school grades for the same schools, year, subjects, and level in the administrative data. Columns (3) and (4) show administrative middle school grades for all students in Bergen the same year, subject, and level, and in the period 2008–2015, respectively. In Column (5), statistics from the Rogaland experiment are reported, while Columns (6), (7), and (8) report statistics from administrative samples that include the same schools, years, subjects, and levels as in the experiment; all schools in Rogaland the same years, subjects, and levels; and

these grades in Rogaland recorded in the period 2008–2015, respectively.

The total number of observations from the experiment in Bergen is 99. Most students take the *Tentamen* in Norwegian, Math, and English, but only one test for each student was selected for re-grading. For the administrative sample, all students are drawn to perform a national exam in Norwegian, Math, or English. The number of observations is fairly similar to that in the experiment, 105, which is reasonable given the similar system of all students being exposed to anonymous grading in one subject. There are relatively more recordings in Norwegian in the Bergen experiment. For our estimates to be unaffected by the proportion of exams in a particular subject, inverse proportion subject weights are used in the empirical specifications. In the Rogaland experiment, the experiment was carried out by drawing a sample of exams from each school. There are thus more observations for the same schools from the administrative data than in the experiment.

The averages of Math grades are lower than the averages of Norwegian grades, and average blind grades are lower than average non-blind grades. Standard deviations are lower in Norwegian than in Math, but there is not any pattern in the differences in standard deviations between blind and non-blind grades. This does not suggest a leniency bias or centrality bias (Landy & Farr, 1980; Prendergast, 1999). However, if non-blind includes more or different attributes than blind, a centrality bias based on subject-specific ability might not appear as lower standard deviations in non-blind. This is because the different attributes included in non-blind may lead to additional variation in this variable.

Table 3 reports summary statistics of the grade difference between the non-blind and blind grades. Depending on the type of data being used, the grade difference could be a sum of several terms and does not necessarily reflect only the teachers' biased grading, or bias. In administrative data, as discussed in the previous section, differences in grades could be because of teachers' biased grading, that students perform better at one type of exam, or that

non-blind and blind grades relate differently to the subject-specific skill. In the non-administrative data, in addition to noise, differences in grades are a sum of teachers' biased grading and that non-blind and blind map differently onto the subject-specific skill. Table 3 presents summary statistics of the grade difference both when aggregating subjects and by subject. Weighted delta (grade difference) is computed by using inverse proportion subject weights. A standardized measure of the difference is constructed by dividing by the standard deviation of the blind exam grade.

[Table 3]

In every subject, the average grade difference is positive. In the non-administrative data from Bergen, the difference is 0.17 standard deviations (SD) of blind exam, whereas in the corresponding administrative data, it is almost three times larger (0.42 SD). In contrast, the non-administrative differences in Rogaland are about twice as large as the difference in the corresponding administrative data. Average differences are smallest in Math, both in absolute terms and relative to variation in blind grades. According to our model, there are several possible explanations for this. The parameters of the biased grading function, $t(X)$, may be different in different subjects. This could, for example, be because there are different types of teachers. Disparities in grade differences across subjects could be explained by the fact that there are differences in student performance across test types in the two subjects. For example, students perform relatively better at in-class exams compared to external exams in Norwegian, compared to Math. Lastly, the students' teachers could weigh skills the students are better at more than external teachers, or in addition, for the administrative data, students are better at the subject skills tested in the non-blind test that are not tested in the blind test. In our model, this would mean that ρ and ϱ differ across subjects, or $\tilde{\theta}$ and $\bar{\theta}$ differ across subjects. A general pattern to notice is that examiners external to the school seem to lead to lower grading. The grade difference is higher in Bergen administrative than non-

administrative, while the reverse is true for Rogaland. At the same time, the external graders to the school are grading the blind in the administrative data for Bergen, and in the non-administrative for Rogaland.

4.2 Comparing Estimates of Bias in Administrative and Non-Administrative Data

Table 4 focuses on the non-administrative data from Bergen and compares them to the administrative data from Bergen. The Bergen experiment is particularly interesting because a variable for the observable characteristic *gender* is available, which is used to estimate a coefficient that can be compared to the coefficient obtained from the administrative data from the same year, schools, level, and subject.

[Table 4]

The table presents results from regressions with grade difference as the dependent variable. According to our model, the parameters shown in Equations (5) and (6) would be the correct expressions for the population regression coefficients on group dummies under the assumption that the non-blind and blind relationship to subject skill is the same ($\rho = 1$ and $\varrho = 1$). In addition, for the administrative data, students do not perform differently under different types of tests ($\tilde{t}(X) = 0$).

First, Column (1) shows the results including only subject dummies on the right hand side. Since within transformation on the all of the binary variables used in Table 4 have been performed, the intercept reflects the weighted average grade difference. Adding an indicator variable for gender, Column (2) shows that the gender coefficient is close to 0, with a standard deviation of 0.098. Column (3) displays results when school-interacted fixed effects are included. This increases the gender coefficient to 0.12, though it stays statistically insignificant. Columns (4), (5), and (6) show the same specifications performed on a sample of students from the same schools, years, level, and subjects, using administrative data. The

weighted average grade difference is much larger in the administrative data, which may be due to the blind graders being external to the schools. Alternatively, students may perform better on in-class exams, or better at the skills tested by in-class exams, but this is not what was suggested by the data from Rogaland shown in the descriptive statistics of grade differences. An explanation for the higher standard errors in the results in the administrative data is that the variation in student performance across different tests is included as unexplained variation. Results in Columns (1), (2), and (3) and (4), (5), and (6) show that we are not able to reject the null hypothesis of no gender bias in either the administrative or non-administrative data, respectively. The evidence does not suggest that the explanation for the positive gender coefficient in administrative data is that females perform better at in-class tests than external exams compared to males, or that females are better at the skills tested by the in-class tests. Three points are important to note about the non-administrative data for Bergen. First, the low sample size makes it difficult to make any precise statements on the size of gender bias. Thus, the true gender coefficient derived from the comparable sample in the administrative and non-administrative may actually be different. Second, the setup for the Bergen trial make it possible that all teachers knew about the fact that the grading where to be audited. This is different from normal grading of students. Lastly, only two schools were available, and grading in these schools can be different from a representative sample of schools.

In Column (7), we include all students in Bergen. The intercept is relatively similar as in results from the two experiment schools, suggesting similar grading in these schools and the rest of Bergen. Column (8) shows a significant coefficient at the 5% significance level of 0.086, which is close to the estimate with school-interacted fixed effects in the non-administrative data. Including school-interacted fixed effects in Column (9) increases the coefficient to 0.098.

According to our model that describes the content of blind and non-blind grades, there could be several reasons for finding non-blind-blind grade differences that are different across groups in the administrative data. Teacher-biased grading, different performance across test types, and two tests measuring different skills can all be potential explanations. Our results do not suggest that different performance across test types, or that the two test measure different skills, explain the findings for the gender coefficient in the administrative data. However, data limitations make us carful to conclude about the existence of gender bias only relying on the non-administrative sample from Bergen.

4.3 The Relationship between Non-Blind, Blind, and Subject Ability

It is important to determine the relationship between non-blind and subject ability, and blind and subject ability. If these relationships are unequal, it has consequences for the interpretation of grade differences. The slope parameters estimated in Table 4 would then be more correctly described by Equations (8) and (15). For example, grade differences between groups could arise just because the two groups are at different ability levels in the subject (Burgess & Greaves, 2013). There could be several reasons for why the relationship to subject ability differs between the two grades. Tests could measure different skills, or graders are looking at different skills when grading. It is also possible that teachers that know the student avoid giving the student a failing grade, while, for an external grader, it is easier to fail a student.

One approach to evaluate this relationship is to rearrange the variables in our model by inserting Y_{is}^{b} for θ_{is}, as shown in Equations (10) and (17). Blind scores are measuring ability with an error. Therefore, Y_{is}^{b} is correlated with the unexplained part in this equation. Because of this, a simple regression of the grade difference on blind score does not reveal ρ or ϱ, but with a classical measurement error in blind yields a negatively biased estimate of

these parameters.

There are two main ways to investigate the importance of measurement error in blind score. First, one could use lagged blind scores as an instrument. The main problem with this procedure is that teachers and the student generally have information on previous and other exam grades the student receives. It is therefore possible that lagged blind grades have a separate impact on non-blind grades. In our model, this is reflected in the terms $\bar{\theta}$ in the non-administrative data and $\tilde{\theta}$ in the administrative data. The other method is to use a regression of the grade difference on grouped average blind score (Deaton, 1985). In our case, the natural way to group students is by school. Table 5 shows these estimations for data from Bergen.

[Table 5]

The table shows results from regressions of the grade difference on individual blind score and school blind score for administrative data in Columns (1)–(4), and for non-administrative data in Columns (5)–(8). Recordings from more years, 2008–2015, are included in the administrative sample to increase precision. The difference between the specification used in Columns (1) and (2) is that school-interacted fixed effects are included in Column (2). This lowers the coefficient on blind scores from –0.25 to –0.22 and suggests that the negative relationship between the grade difference and blind score is partly explained by school factors. The difference between the specifications in Columns (3) and (4) is that blind score-interacted fixed effects are included in Column (4). Column (3) shows a negative relationship of –0.35 between the grade difference and school average blind.[7] This relationship cannot be explained by classical measurement errors in blind score since this is a precise estimate of the school-level ability. There could, however, be other school-level

[7] The regression is performed at the individual level, while school blind is the school average blind for the subject the individual recording is measured in. This specification simply allows for appropriate weighting by school size, while at the same time including subject weights in the regression. School averages are calculated without own recording.

factors that contribute to the negative relationship between group differences and school blind. For example, schools with higher average blind scores are less lenient in non-blind grading. Including blind score-interacted fixed effects reduces the size of the school blind coefficient by 0.21, suggesting that a substantial portion of the negative coefficient is not due to school-level factors. These results are in line with the explanation that a large part of the negative coefficient on blind score is due to non-blind and blind grades mapping differently onto subject ability in the administrative data.

A reason for the negative relationship between the grade difference and school blind could be that non-blind and blind tests in the administrative data actually measure different subject skills. Even though Table 4 did not indicate it, this could lead to the gender coefficient reflecting that females perform better at the subject skills tested in non-blind, but not in blind. Columns (5) and (6) show that the coefficient is only –0.11 using the non-administrative data, indicating that the non-blind and blind grade is more likely measuring the same ability. Still, it also suggests that teachers that know the student, and teachers that do not know the student, grade differently even though they grade the same test. Possible explanations are that the student's teacher knows the identity of the student, the student's class behavior, previous grades, and grades in other subjects. Columns (7) and (8) provide negative, but much less precise, estimates of the relationship between grade difference and school blind. Due to large standard errors, the difference between Column (7) and (8) tells us little, but since both non-blind and blind graders are internal to the school, school-level influences are less likely to be responsible for the negative relationship.

[Table 6]

Table 6 provides results for the separate experiment performed in Rogaland and a comparable administrative dataset. Individual observable characteristics are not available for the non-administrative data, but there are more individual recordings, and recordings from

more schools than from the Bergen experiment. Columns (1) and (2) show a negative relationship between grade differences and blind scores of –0.27 and –0.28, respectively. Interestingly, the coefficient increases when including school-interacted fixed effects, suggesting that school-level factors do not contribute to a negative relationship between the grade difference and blind. Column (3) reveals a negative and statistically significant coefficient on school blind of –0.17, which disappears when including blind-interacted fixed effects in Column (4). Note that the student's teacher also grades the locally administered exam, which is used as a blind score for the administrative sample from Rogaland. There is, however, another teacher that also grades the exam, who has less prior information on the student and is external to the school.

As discussed, the negative relationship between the grade difference and blind, indicated by the results provided in Columns (1)–(4), may indicate that non-blind and blind measure different skills. Columns (5) and (6) explore the relationship using the non-administrative data. Column (5) shows a coefficient of –0.17, while the estimate in Column (6) is –0.18. Again, these results do not suggest that schools with higher blind scores are less lenient for the sample of schools from Rogaland. Regressing the grade difference on school blind provides a negative and statistically significant coefficient at the 5% significance level of –0.13 in Column (7). This coefficient increases to 0.05 when including blind-interacted fixed effects. The difference between Columns (7) and (8) is 0.17, which is identical to the coefficient in Column (5).

The results provided in Tables 5 and 6 make it possible to determine the size of 1-ρ and 1-ϱ. Generally, we find a negative relationship when regressing the grade difference on blind. This negative relationship is somewhat smaller in the non-administrative datasets. We also find a similar negative relationship when regressing grade difference on school blind, something that is not explained by measurement error in blind. For the schools from

Rogaland, we do not see any signs that schools with better students are less lenient. Because school-level factors seem to be less of a concern in the Rogaland sample, we use the indicated impact of measurement error in the sample from Rogaland to determine the parameters in Bergen. Using the point estimates of coefficients, the impact of measurement error in blind in administrative data is –0.10 in administrative and –0.04 in non-administrative. Based on results with school-interacted fixed effects from Bergen, this indicates a $1-\rho$ of –0.12 and $1-\varrho$ of –0.05.[8]

Table 5 and 6 provide estimates for $1-\varrho$ from both Bergen and Rogaland. For the non-administrative data from Rogaland, the estimate is larger in magnitude. A possible explanation for this is that the blind evaluator is external to the school in this experiment, leading to a different mapping of grades. In addition, exams were randomly drawn from schools chosen by county level administrative personnel. These features correspond more to the field experiment conduced in Hinnerich et al. (2011), and suggest that the gender bias holding ability fixed is different from the gender bias estimated in that paper.

4.4 Alternative Parameter of Interest

The previous section examined the relationship between non-blind, blind, and subject ability, and found convincing evidence that $1-\rho<0$ and $1-\varrho<0$. This changes the interpretation of the coefficient from a regression of grade difference on the group dummy to now also including a term that reflects the ability difference between groups. In our model, the coefficient is characterized by Equation (8) for non-administrative data and Equation (15) for administrative data. Group bias will now arise if the groups have different subject abilities.

[8] We use the difference between Column (1) and (3) in Table 6 to get an estimate for the impact of measurement error in the administrative data $(-0.27-(-0.17))=-0.10$. The difference between Column (5) and (7) gives an estimate for the impact in non-administrative data $((-0.17-(-0.13))=-0.04$. To get estimates of $1-\rho$ and $1-\varrho$, these numbers are subtracted from coefficients in Column (2) and (6) in Table 5: $(-0.22-(-0.10))=-0.12$ and $(-0.09-(-0.04))=-0.05$.

An alternative parameter of interest reflects group bias holding ability constant. For example, this parameter describes the amount of bias a female can expect to get compared to a male of equal ability. In the terms of the model, this parameter is defined in Equations (9) and (16). A way to retrieve an estimate of this parameter is to add blind score as a right-side variable, as shown in Equations (10) and (17). However, this is not feasible using ordinary least squared regression since the model is unidentified, because the blind score is correlated with the unexplained part. Therefore, we use the fact that the last section provided credible estimates of $1-\rho$ and $1-\varrho$, and estimate parameters using constrained least squared estimation, fixing the coefficient of Y_{is}^{b} to the specific values.

[Table 7]

Columns (1)–(3) show results from the non-administrative sample, Columns (4)–(6) for the same schools using the administrative sample, and Columns (7)–(9) all recordings from Bergen in 2015. Column (1) shows the result from an OLS regression of the grade difference on a gender coefficient, shown earlier in Table 4. Column (2) shows results from a constrained least squared estimation, where, in addition to having the gender dummy on the right side, the blind score is included as described in Equation (10). The results confirm that fixing the coefficient on blind to be 0 with this specification gives the same results as an OLS regression of the grade difference on the gender dummy shown in Column (1). Column (3) displays results when fixing $1-\varrho$ to –0.05. The gender coefficient increases to 0.15, but is still not statistically significant.[9] Columns (4)–(6) repeat this procedure for administrative data from the same schools. Column (6) uses the estimate of $1-\rho$ obtained for the administrative data of –0.12. Also here, the gender coefficient increases but is still not significant. In Columns (7)–(9) the gender coefficient is significant for all specifications, and the point estimate of the gender coefficient is very similar to that obtained from the non-

[9] Note that this procedure does not account for uncertainty regarding the size of the fixed parameters.

administrative data.

4.5 Other Observable Characteristics

This analysis has compared estimates of gender bias using two different data-generating processes—one where the student's teacher and a teacher that does not know the identity of the student grade the same test, and where the student's teacher performs a final course evaluation and two external examiners grade a final course exam. The results did not confirm that estimates of gender bias were different in the administrative data, not suggesting that the explanation for the positive gender coefficient found using administrative data is because females perform better at in-class exams or tests measuring different skills. This analysis therefore proceeds to look at other observable determinants of the grade difference using administrative data. Given that estimates of gender bias in the non-administrative bias were similar to the coefficient obtained from the administrative data, there is no reason to mistrust estimates from administrative data based on other student characteristics. In addition, we provided evidence that bias depends on ability. In Table 8, we examined how other observable characteristics are related to grade differences using the administrative data, fixing the coefficient on blind to specific values in the model described in Equation (17). Since the purpose no longer is to compare bias estimates for the same school, year, and course as in the Bergen experiment, we use a sample of recordings from all schools in Bergen for 2008–2015.

[Table 8]

Columns (1)–(3) show the coefficient on the gender dummy with subject- and cohort-interacted fixed effects, and subject-, cohort-, and school-interacted fixed effects. The results are similar to the findings previously shown for 2015. The coefficient changes marginally when moving from Column (1) to Column (2) when adding the school-interacted fixed effect. Middle school attendance is determined by catchment area. This means that the gender

balance should be unrelated to school characteristics, since the proportion of females is the same in different types of catchment areas. This may explain why including school-interacted fixed effects only has a small impact on the coefficient. When moving from Column (2) to Column (3), the coefficient on blind score is fixed to −0.12. Holding subject ability fixed significantly increases the size of the coefficient. As we have discussed, this is because the non-blind relationship to subject ability is different than the relationship between blind and subject ability, and females have different ability levels than males.

Columns (4)–(6) repeat the procedure, but jointly include additional observable characteristics. Column (4) suggests that immigrants are positively rewarded by teachers compared to non-immigrants. This is in line with findings in Lindahl (2007) and Falch and Naper (2013). Adding school-interacted fixed effects leaves the estimate unchanged. Column (6) fixes the coefficient on blind score to −0.12, and the coefficient decreases to 0.01 and becomes insignificant. This suggests that, when holding subject ability fixed, immigrants do not receive a positive amount of bias compared to non-immigrants in Bergen.

Column (4) suggests that low-SES students receive a negative amount of bias compared to non-low-SES students. The coefficient becomes larger in magnitude and statistically significant at the 1% significant level when including school-interacted fixed effects. The results suggest that low-SES students are overrepresented in areas where schools are more lenient. Column (6) shows that holding subject ability fixed more than doubles the estimate of the amount of negative bias that low-SES students receive.

Table 8 shows that the estimates of the total amount of group bias, and the total amount of group bias conditioning on ability, parameters described in Equations (15) and (18), may provide widely different estimates of the size of discrimination. According to the econometric model we specify, since we find that $1 - \rho < 0$, estimates of bias that do not take into account subject ability indicates that the bias in favor of the group with lower

abilities is larger than when holding subject ability constant.

5 Conclusion

Several studies use data where teachers that know the identity of the students, and teachers that do not, grade students' tests. Systematic differences in grading between these teachers could then be attributed to biased grading. This paper develop an econometric framework that clarifies underlying reasons for differences in grading between teachers that know the students and teachers that do not. In our model, blind scores include subject-specific ability and measurement errors. Furthermore, the model describe that non-blind grades may contain more information than only the subject-specific ability. In addition to subject-specific ability, non-blind includes teacher biased grading according to observable student characteristics of the teacher, the information the teacher has on previous grades and grades in other subjects, and measurement errors. In the administrative data, non-blind may also contain information on subject-specific ability not tested in blind, and the relative performance of students in the non-blind test situations compared to the blind test situation. Our model points to two important issues. First, if administrative non-blind includes more subject-specific ability than blind, or if some students perform better at a specific test type, then using administrative data may not yield an appropriate measure of the total amount of bias one group receives compared to another. Differences across groups can therefore more correctly be ascribed to the effect of test type/grading scheme. Second, if non-blind and blind map differently onto subject-specific skills, the non-blind-blind grade difference is a function of skill. Therefore, differences in grading between the two groups can be a result of different skill levels. This could happen using both administrative and non-administrative data. In addition to developing the econometric framework, this paper compare estimates of bias for comparable administrative and non-administrative data. The results are not able to show that the estimate

of the amount of bias females receive compared to males is different when using the two data types. Note that data limitations restrict our conclusion based on this specific dataset. Furthermore, the analysis shows that the relationship between subject-specific ability and non-blind is not equal to the relationship between subject ability and blind. The consequence of this is that subject ability level should be accounted for when estimating the group bias parameter holding the ability level constant.

References

Bertrand, M. & Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *The American Economic Review*, 94 (4), pp.991–1013.

Blank, R. M. 1991. The effects of double-blind versus single-blind reviewing: experimental evidence from the *American Economic Review*. *The American Economic Review*, 81 (5), pp.1041–1067.

Burgess, S. & Greaves, E. 2013. Test scores, subjective assessment, and stereotyping of ethnic minorities. *Journal of Labor Economics*, 31 (3), pp.535–576.

Cornwell, C., Mustard, D. B., & Van Parys, J. 2013. Noncognitive skills and the gender disparities in test scores and teacher assessments: evidence from primary school. *Journal of Human Resources*, 48 (1), pp.236–264.

Deaton, A. 1895. Panel data from time series of cross-sections. *Journal of Econometrics*, 30 (1), pp.109–126.

Falch, T. & Naper, L. R. 2013. Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review*, 36, pp.12–25.

Goldin, C. & Rouse, C. 2000. Orchestrating impartiality: the impact of "blind" auditions on female musicians. *The American Economic Review*, 90 (4), pp.715–741.

Hanna, R. N. & Linden, L. L. 2000. Discrimination in grading. *American Economic Journal: Economic Policy*, 4 (4), pp.146–168.

Hinnerich, B. T., Höglin, E., & Johannesson, M. 2011. Are boys discriminated against in Swedish high schools? *Economics of Education Review*, 30 (4), pp.682–690.

Hinnerich, B. T., Höglin, E., & Johannesson, M. 2015. Discrimination against students with foreign backgrounds: evidence from grading in Swedish public high schools. *Education Economics*, 23 (6), pp.660–676.

Landy, F. J., and Farr, J. L. 1980. Performance rating. *Psychological Bulletin*, 87 (1), pp.72–107.

Lavy, V. 2008. Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, 92 (10–11), pp.2083–2105.

Lindahl, E. 2007. Comparing teachers' assessments and national test results: evidence from Sweden. IFAU Institute for Evaluation of Labour Market and Education Policy, Uppsala.

Prendergast, C. 1999. The provision of incentives in firms. *Journal of Economic Literature*, 37 (1), pp.7–63.

Sprietsma, M. 2013. Discrimination in grading: experimental evidence from primary school teachers. *Empirical Economics*, 45 (1), pp.523–538.

Van Ewijk, R. 2011. Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review*, 30 (5), pp.1045–1058.

Table 1: Institutional details - grading

Region	Dataset	Variable definition	Grader	# graders	Name on test	External/local to school	Test type
Bergen	Administrative	Non-blind	Students' teacher	1	Yes	Local	Course assessment
		Blind	External teachers	2	No	External	National exam
	Non-administrative	Non-blind	Students' teacher	1	Yes	Local	Local test (Tentamen)
		Blind	Another teacher	1	No	Local	Local test (Tentamen)
Rogaland	Administrative	Non-blind	Students' teacher	1	Yes	Local	Course assessment
		Blind	Students' teacher/external teacher	2	No	Local and external	Local exam
	Non-administrative	Non-blind	Students' teacher/external teacher	2	No	Local and external	Local exam
		Blind	External teachers	2	No	External	Local exam

Notes: The table summarize institutional details about the grades used in the analysis. Scores from Bergen are at middle-school level, while scores from Rogaland are at the high school level.

Table 2: Descriptives - Grades

	Bergen				Rogaland			
	Non-adm.	Adm.			Non-adm.	Adm.		
		Same schools	Bergen	08-15		Same schools	Rogaland	08-15
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math								
Non-blind average	3.18	3.21	3.66	3.69	3.26	3.00	3.08	3.14
Non-blind sd	1.10	1.20	1.22	1.18	1.27	1.39	1.40	1.37
Blind average	3.08	2.72	3.04	3.25	2.98	2.98	3.07	3.16
Blind sd	1.13	1.24	1.29	1.22	1.32	1.25	1.29	1.27
# Math	39	67	1024	8747	135	649	782	1922
Norwegian								
Non-blind average	3.60	3.95	3.97	3.97	3.16	3.50	3.56	3.52
Non-blind sd	1.01	0.93	1.00	1.00	0.93	1.01	0.99	0.98
Blind average	3.37	3.45	3.56	3.58	2.86	3.21	3.32	3.30
Blind sd	0.97	1.01	1.06	1.02	0.90	0.96	0.96	0.99
# Norwegian	60	38	662	4481	148	536	665	1309
Female	0.43	0.44	0.50	0.49	.	0.40	0.43	0.44
Ses	.	0.22	0.20	0.24	.	0.29	0.29	0.28
Immigrant	0.09	0.08	0.11	0.10	.	0.06	0.05	0.05
# All	99	105	1686	13228	283	1185	1447	3231
Schools	2	2	28	28	15	15	29	29

Notes: Non-blind are grades given by the students' teacher, while blind are grades given by other examinators. Descriptives are on the student level. Columns (1)-(3) consist of students in Bergen exiting middle school in the year 2015 (cohort 1999). Column (4) consists of students in Bergen exiting middle school in the period 2008-2018. Grades are given at the end of the last year of middle school. Columns (5)-(7) consist of students in Rogaland taking high school courses in the years 2010, 2012 and 2013. Column (8) consists of students in Rogaland taking courses in the period 2008-2015. Grades are given at the first and second level of high school. In the non-administrative data, non-blind and blind grades are evaluations of the same test for each student. The test in Bergen is the Tentamen, a locally administered written test. The test in Rogaland is the locally administered end-of-year exam. In the administrative data, the non-blind grade is a teacher evaluation of the students' performance in the course, while the Blind grade is a grade given on a test at the end of the year. Non-blind grades are set before the blind grades are set in the administrative data. In the administrative data from Bergen, the blind evaluation is performed anonomously by two external examinators. In the administrative data from Rogaland, the blind evaluation is set by an examinator together with the students' teachers. This is the locally administered end-of-year exam also used in the experiment. In both Bergen and Rogaland, grades are recorded in the same subject and level.

Table 3: Descriptives - Delta

	Bergen						Rogaland					
	Non-administrative			Administrative			Non-administrative			Administrative		
	W. delta	Norwegian	Math	W. Delta	Norwegian	Math	W. Delta	Norwegian	Math	W. Delta	Norwegian	Math
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Average delta	0.18	0.23	0.10	0.50	0.50	0.49	0.29	0.30	0.28	0.17	0.29	0.01
SD blind	1.06	0.97	1.13	1.18	1.01	1.24	1.13	0.90	1.32	1.10	0.96	1.25
Average delta/blind SD	0.17	0.24	0.09	0.42	0.50	0.40	0.26	0.33	0.21	0.15	0.30	0.01
Average blind	3.22	3.37	3.08	3.08	3.45	2.72	2.92	2.86	2.98	3.11	3.21	2.98
N	99	60	39	105	38	67	283	148	135	1185	536	649

Notes: The table shows descriptives of the grade difference (Δ_i). Results from Bergen are shown in columns (1)-(6), while results from Rogaland are shown in columns (7)-(12). Columns (1)-(3) and (7)-(9) show descriptives for non-administrative data, while columns (4)-(6) and (10)-(12) show descriptives for administrative data for the same schools, subject, level, and year as in experiment. Recordings in Bergen are from 2015, while recordings for Rogaland are from 2010, 2012, and 2013.

Table 4: Comparing non-administrative and administrative - Bergen

Dependent variable: Grade difference

	Non-administrative			Administrative					
				Same schools			Bergen		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Female		0.009	0.121		0.077	-0.004		0.086**	0.098***
		(0.098)	(0.100)		(0.164)	(0.145)		(0.040)	(0.035)
Intercept	0.182***	0.182***	0.182***	0.495***	0.495***	0.495***	0.537***	0.537***	0.537***
	(0.050)	(0.051)	(0.049)	(0.067)	(0.067)	(0.061)	(0.018)	(0.018)	(0.016)
R2	0.018	0.019	0.113	0.000	0.003	0.202	0.019	0.022	0.254
Adj. R2	0.008	-0.002	0.075	-0.010	-0.017	0.170	0.018	0.021	0.229
N	99	99	99	105	105	105	1686	1686	1686
N Math	39	39	39	67	67	67	1024	1024	1024
N Nor	60	60	60	38	38	38	662	662	662
Female blind	3.609	3.609	3.609	3.286	3.286	3.286	3.571	3.571	3.571
Male blind	2.924	2.924	2.924	2.917	2.917	2.917	3.080	3.080	3.080
Fixed effects									
Subject	x	x		x	x		x	x	
Subject*School			x			x			x

Notes: Results from the non-administrative data from Bergen is reported in columns (1)-(3), while results from the same schools using administrative data are reported in columns (4)-(6). Columns (7)-(9) use administrative recordings from all middle schools in Bergen. Grade differences are recorded in Norwegian and Math. Only recordings from 2015 are included in both the non-administrative and the administrative datasets. Each observation is weighted by the inverse of the proportion of recordings in that subject. The subject weighted average blind grade by gender is shown. The two lowest rows indicate demeaned variables included. The gender variable is also demeaned. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 5: Delta on blind - Bergen

	Administrative				Non-administrative			
Dependent variable: Grade difference (delta)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Blind	-0.25***	-0.22***			-0.11**	-0.09**		
	(0.01)	(0.01)			(0.04)	(0.04)		
School-blind			-0.35***	-0.14***			-0.60**	-0.68**
			(0.03)	(0.02)			(0.27)	(0.28)
Intercept	1.27***	1.17***	1.60***	0.89***	0.52***	0.46***	2.13**	2.38**
	(0.02)	(0.02)	(0.09)	(0.08)	(0.15)	(0.14)	(0.91)	(0.92)
r2	0.16	0.29	0.04	0.20	0.07	0.13	0.07	0.17
Adj. r2	0.16	0.27	0.04	0.19	0.05	0.10	0.05	0.09
N	13228	13228	13228	13228	99	99	99	99
Fixed effects								
Subject*Cohort	x		x		x		x	
Subject*Cohort*School		x				x		
Subject*Cohort*Blind				x				x

Notes: The dependent variable is the grade difference (delta) measured at the individual level. Blind grade is individual blind grade. Only Math and Norwegian recordings are included. School-blind is the school average blind score calculated without the students' own individual blind score for each subject. School-blind to a student drawn in Math is the school average blind grade in Math of all other students drawn in Math at the same school. The regression is weighting each observation with the inverse of the proportion of that subject being recorded. Three lowest rows indicate demeaned variables included. The administrative sample consists of recordings measured in middle school 2008 - 2015, while Non-administrative sample is only for the year 2015. Heteroscedasticity robust standard errors reported. * p<0.10, ** p<0.05, *** p<0.01

Table 6: Delta on blind - Rogaland

	Administrative				Non-administrative			
Dependent variable: Grade difference (delta)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Blind	-0.27***	-0.28***			-0.17***	-0.18***		
	(0.01)	(0.01)			(0.03)	(0.04)		
School Blind			-0.17***	0.02			-0.13**	0.05
			(0.04)	(0.04)			(0.06)	(0.07)
Intercept	0.95***	0.98***	0.62***	0.00	0.79***	0.81***	0.67***	0.13
	(0.04)	(0.05)	(0.14)	(0.14)	(0.09)	(0.11)	(0.18)	(0.21)
r2	0.15	0.35	0.04	0.19	0.11	0.22	0.03	0.22
Adj. r2	0.14	0.25	0.04	0.17	0.09	0.13	0.01	0.13
N	3231	3231	3231	3231	283	283	283	283
Fixed effects								
Subject*Cohort	x		x		x		x	
Subject*Cohort*School		x				x		
Subject*Cohort*Blind				x				x

Notes: The dependent variable is delta (Δ_i) measured at the individual level. Blind grade are individual blind grade. Only Math and Norwegian recordings included. School-blind is school average blind grade calculated without own individual blind grade for each subject. School-blind to a student drawn in Math is school average blind grade in Math of all other students drawn in Math at the same school. Regressions are weighting each observation with the inverse of the proportion of that subject being recorded. The three lowest rows indicate the demeaned variables included. Non-administrative and administrative samples consist of recordings measured in vocational track high schools for 2010, 2012, and 2013 in Rogaland. Heteroscedasticity robust standard errors reported. * p<0.10, ** p<0.05, *** p<0.01

Table 7: Gender bias - fixing ability

Dependent variable: Grade difference									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Girls	0.121	0.121	0.153	-0.003	-0.003	0.042	0.098***	0.098***	0.158***
	(0.100)	(0.100)	(0.097)	(0.135)	(0.135)	(0.128)	(0.035)	(0.035)	(0.033)
Blind		0.000	-0.050		0.000	-0.120		0.000	-0.120
		(.)	(.)		(.)	(.)		(.)	(.)
Intercept	0.182***	0.182***	0.344***	0.495***	0.495***	0.853***	0.537***	0.537***	0.926***
	(0.049)	(0.049)	(0.048)	(0.061)	(0.061)	(0.058)	(0.016)	(0.016)	(0.015)
N	99	99	99	105	105	105	1686	1686	1686
Subject*Cohort*School	x	x	x	x	x	x	x	x	x

Notes: Results from the non-administrative data from Bergen is reported in columns (1)-(3), while results from the same schools using administrative data are reported in Columns (4)-(6). columns (7)-(9) use administrative recordings from all middle schools in Bergen. The columns (2)-(3), (5)-(6), and (8)-(9) show results from a constrained least squareds estimation where the coefficient of blind grade is fixed. Grade differences are recorded in Norwegian and Math. Only recordings from 2015 are included in both the non-administrative and the administrative datasets. Each observation is weighted by the inverse of the proportion of recordings in that subject. Heteroscedasticity robust standard errors reported.. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 8: Group bias

Dependent variable: Grade difference						
	(1)	(2)	(3)	(4)	(5)	(6)
Girls	0.06***	0.07***	0.11***	0.06***	0.08***	0.11***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Immigrant				0.06***	0.06***	0.01
				(0.02)	(0.02)	(0.02)
Low-SES				-0.03*	-0.05***	-0.12***
				(0.02)	(0.02)	(0.02)
Blind	0.00	0.00	-0.132	0.00	0.00	-0.12
	(.)	(.)	(.)	(.)	(.)	(.)
Intercept	0.42***	0.42***	0.86***	0.42***	0.42***	0.86***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
N	13228	13228	13228	13228	13228	13228
Subject*Cohort	x			x		
Subject*Cohort*School		x	x		x	x

Notes: Results use administrative recordings from all middle schools in Bergen recorded in the period 2008-2015. The table shows results from a constrained least squareds estimation where the coefficient of blind is fixed. Grade differences are recorded in Norwegian and Math. Only recordings from 2015 are included in both the non-administrative and the administrative datasets. Each observation is weighted by the inverse of the proportion of recordings in that subject. Heteroscedasticity robust standard errors reported.. * p<0.10, ** p<0.05, *** p<0.01

Department of Economics
University of Bergen
PO BOX 7800
5020 Bergen
Visitor address: Fosswinckels gate 14
Phone: +47 5558 9200
www.uib.no/econ/