Pre-Analysis Plan

Ragnar Hjellset Alne, NIFU
Eyo I. Herstad, Department of Economics, Ohio State University
Rune Borgan Reiling, NIFU*

November 23, 2023

Abstract

We present causal evidence regarding the impact of a student's ethnicity and gender on teacher evaluations in middle school. In this study, we engage teachers to assess written assignments for full classes in two subjects: mathematics and Norwegian language. By randomly assigning names that signal both gender and ethnicity to the assignments within teachers, we investigate whether students with equal ability but different names receive equal grades, on average. While our primary focus is on the average effects of student gender and ethnicity, we also conduct an extensive analysis of heterogeneity, considering factors such as ability, student gender, teacher experience, and the gender of the teacher. This document outlines our analysis plan, including our primary specifications of interest.

Keywords: Discrimination
JEL Classification: J71

*Corresponding author: rune.borgan.reiling@nifu.no. Phone: (+47) 997 72 589.

1 Introduction

Student grades in schools are supposed to be a reflection of the students' real ability in the subject, and commonly decide which students go to which high school, higher education and job. As such, from a fairness perspective, it is important that student grades reflect the the student's ability in the subject. A student's background, including ethnicity and gender, should not affect the teachers' evaluation of the student.

To identify whether students of different backgrounds are treated fairly, we recruit real teachers to evaluate real student assignments. The assignments are randomly assigned student names, which are intended to signal the students' ethnicity and gender. The identification strategy is motivated by traditional CV-experiments, where job-applicants are randomly assigned names to study whether applicants with different names are treated unequally in job application processes. In these experiments, the ethnicity and gender of the job applicant are traditionally signaled through a random assignment of names (See e.g., Midtbøen, 2015; Midtbøen, 2016; Ahmad, 2020). Our study also builds upon a small literature using comparable experiments to investigate discrimination in grading (see e.g., Hanna and Linden, 2012; Sprietsma, 2013; Van Ewijk, 2011).

2 Experimental design

2.1 Teacher evaluations

To study the effect of students' ethnicity and gender on teacher evaluations, we randomly assign student names to actual assignments to investigate whether the name, as an indicator of ethnicity and gender, causally influences evaluations of the students' abilities. The student names are easily visible to the participating teachers at the beginning of the student assignment, as well as where the grades are assigned.

Teachers are recruited through collaboration with multiple schools that assist in distributing the assignments and the questionnaire to them. All assessed assignments are completed by actual students, and each teacher is compensated economically for evaluating a complete

class' assignments. The assignment should prevent teachers from discerning the gender and ethnicity of the original student/author, and both the grading and post-grading questionnaires are distributed electronically via Qualtrics. In the questionnaire, both the order of the assignments and the names on the assignments are randomized. The teacher will initially assign a grade, which can be adjusted in the final overview of grades for all the assignments. Therefore, we will have both the initial grade and the grades set after adjustments.

2.2 Control groups

In the study, we establish a few control groups. The first control group involves teacher evaluations of assignments without any access to students' names. This blind evaluation aims to assess students' abilities impartially and is incorporated to enable us to estimate the impact of gender and ethnicity in comparison to this untreated control group.

Our second control group consists of assignments with native names, and it is employed as a control group in regression analyses where we estimate the influence of ethnicity. This allows us to assess the impact of ethnicity in comparison to both a blind control group and a control group with Norwegian names.

In our examination of gender effects, we utilize male and female names for both native and immigrant students. Female names serve as the control group when estimating the impact of having a male name on the assignment. We therefore evaluate the gender effect for all assignments collectively, as well as separately for assignments with Norwegian names and those with non-native names. This approach allows us to assess whether there is variation in gender bias between native and non-native names. The student names are listed in Appendix A.

2.3 Outcome 1: Student grades

Our primary outcome is the assessment of students' ability levels by teachers. We collect this data through our questionnaire, in which teachers evaluate written assignments for an entire class using the standard Norwegian grading scale. The Norwegian grading scale spans from

1 to 6, with 6 representing the highest grade and 1 signifying a failed assignment.

We collect student grades for both a Mathematics and a Norwegian language assignment. This approach allows us to gain insights into the variations in biased grading within and between subjects. Previous research, for instance, has indicated that bias tends to be minimal for high-ability students in Mathematics (Alne and Herstad, 2020).

2.4 Control variables: Demographics and classroom characteristics

We collect demographic data through a questionnaire in which teachers provide information on various demographic factors, including:

- Gender of the teacher
- Age of the teacher
- Years of teaching experience
- Whether the teacher is a first or second generation immigrant
- Municipality of the school they work

We also ask the teachers about their own objective classroom characteristics, including:

- Number of students in their class and school
- Number of students with immigrant background in their class and school
- Number of male students in their class and school
- Number of students with parents of higher education in their class and school

Finally we ask about a few subjective classroom-characteristics of the teacher's class and school, including:

- Noise in their own class and school
- The focus of the students in their own class and school

All the questions listed above are placed at the end of the questionnaire to prevent teachers from becoming aware of the study's purpose before they complete the assignment grading. This information is collected for the purpose of conducting exploratory analyses and examining whether these characteristics are associated with the observed bias.

3 Control variables

When the background characteristics are used as controls in the regression, they will be coded as follows:

- Grades will be standardized to have a mean of zero and a standard deviation of one in the primary specification. We will also estimate regressions with non-standardized grades as the outcome variable.
- Gender will be coded as a male dummy variable for both teachers and students.
- Work experience will be coded as the number of years of teaching experience, and we will also use it as a binary variable if the teacher has more more than the median years of teaching experience.
- Ethnicity will be coded as a dummy variable to indicate whether the name signals a non-native background.
- Ability will be coded as a continuous variable, but will also be converted into an ordinal variable, to represent the Norwegian grade scale, such that the heterogeneity in bias across blind ability can be estimated non-parametrically. We will allow the teachers to use a '+' or a '-' to indicate whether it is a strong or a weak grade, which will be coded as +/- 0.25 to the grade.

4 Inference

In our data analysis, we employ cluster-robust standard errors at the teacher level to account for correlations between observations within each teacher.

5 Setting and sample size

We recruit teachers through the schools that we engage in the project. Teachers will receive invitations to participate in a study aimed at exploring variations in grading practices. Our goal is to recruit approximately 150 teachers for the project, with an anticipated participation of over 100 teachers who will each grade approximately 30 assignments. This amounts to a total of roughly 3,000 graded assignments. Around 1/5 of these assignments will be evaluated without student names.

5.1 Power analysis

With 3,000 observations, we have a statistical power of 0.8 to detect an effect size of .004 (Cohen's f2) at a significance level (α) of 0.05. Each teacher grades around 30 assignments, and we will cluster these observations due to correlations between each teacher's evaluations. However, previous studies, such as Hanna and Linden (2012), Sprietsma (2013), and Van Ewijk (2011), have reported an average effect size of Cohen's f2 = .37, which is significantly larger than our calculated threshold of .004. Our analysis suggests that our experiment should possess sufficient statistical power.

6 Hypotheses

To understand the level and heterogeneity in both ethnic and gender biases in grading, we have formulated a set of hypotheses. These hypotheses are informed by the existing literature on gender and ethnic bias, particularly drawing from Lavy (2008). Of interest for Norway is Falch and Naper (2013) and Alne and Herstad (2020), which is using Norwegian registry data in the identification of gender bias. Falch and Naper (2013) finds grading biases against boys in Norwegian middle schools. Similarly, Alne and Herstad (2020) finds a significant bias against male students in Norwegian high-schools, where the bias varies across both subjects and the students' ability level. Our hypotheses are given as follows:

6.1 Hypotheses regarding ethnicity

Hypothesis 1 and 2 is based primarily on the results in Hinnerich, Höglin, and Johannesson (2015), Sprietsma (2013) and Van Ewijk (2011). See also Neumark (2018) for a review of ethnic discrimination in labor settings and Shea (2022) for recent work on the subject. Specifically, Ba, Knox, Mummolo, and Rivera (2021) showed that ethnic and gender bias can depend on the ethnicity and gender of the decision maker.

Hypothesis 1. *On average, assignments with foreign names, which serve as a signal of a student's foreign ethnicity, are expected to receive less favorable evaluations from teachers compared to assignments without any name, which act as the control group.*

Hypothesis 2. *Assignments with foreign names, signifying a student's foreign ethnicity, are expected to receive, on average, less favorable evaluations from teachers compared to assignments with native names, which serve as the control group*

6.2 Hypotheses regarding gender

Hypotheses 3, 4, and 5 are primarily based on the findings of Alne and Herstad (2020), Falch and Naper (2013), and Lavy (2008).

Hypothesis 3. *Assignments with male names, signifying a student's gender, are expected to receive less favorable evaluations from teachers on average compared to assignments without any name, which serve as the control group.*

Hypothesis 4. *Assignments with male names, indicating a student's gender, are expected to receive, on average, less favorable evaluations from teachers compared to assignments with female names, which serve as the control group.*

Hypothesis 5. *Assignments with a female name, indicating a student's gender, are expected to, on average, receive more favorable evaluations from teachers compared to assignments without any name, which serve as the control group.*

6.3 Hypotheses regarding male gender and ethnicity

Hypotheses 6, 7, and 8 are primarily based on the findings of Hinnerich et al. (2015), Sprietsma (2013), Van Ewijk (2011), Alne and Herstad (2020), Falch and Naper (2013), and Lavy (2008).

Hypothesis 6. *Assignments with male foreign names, signifying both a student's foreign ethnicity and gender, are expected to receive, on average, less favorable evaluations from teachers compared to assignments without any name, which serve as the control group.*

Hypothesis 7. *Assignments with a male foreign name, signifying both a student's foreign ethnicity and gender, are expected to receive, on average, less favorable evaluations from teachers compared to assignments with a female foreign name.*

Hypothesis 8. *Assignments with a male foreign name, signifying both a student's foreign ethnicity and gender, are expected to receive, on average, less favorable evaluations from teachers compared to assignments with a male native name.*

6.4 Hypotheses regarding female gender and ethnicity

Hypotheses 9 and 10 are primarily based on the findings of Hinnerich et al. (2015), Sprietsma (2013), and Van Ewijk (2011).

Hypothesis 9. *Assignments with a female foreign name, signifying both a student's foreign ethnicity and gender, are expected to receive, on average, less favorable evaluations from teachers compared to assignments with a female native name.*

Hypothesis 10. *Assignments with a female foreign name, signifying both a student's foreign ethnicity and gender, are expected to receive, on average, less favorable evaluations from teachers compared to assignments without any name, which serve as the control group.*

6.5 Hypotheses regarding heterogeneity: Ability

Hypothesis 11 is primarily based on the findings of Alne and Herstad (2020).

Hypothesis 11. *Assignments that receive a low grade from the original teacher or have, on average, lower scores in the blind evaluation are expected to receive, on average, larger ethnic and gender bias across all specifications.*

6.6 Hypotheses regarding heterogeneity: Subjects

Hypothesis 12 is primarily based on the findings of Alne and Herstad (2020).

Hypothesis 12. *Assignments in Mathematics are expected to, on average, receive a lower level of ethnic and gender bias compared to assignments in Norwegian.*

6.7 Hypotheses regarding heterogeneity: Gender of the teacher

The increasing proportion of female teachers is often cited as a contributing factor to the observed reversal of gender achievement gaps in education over the last few decades. Mechanisms such as the diminishing presence of male role models, a shift towards a more feminine school environment, and gender bias can all contribute to gender-related differences in achievement. While some studies suggest that girls benefit from gender-based grading bias, others indicate that the occurrence of such bias may depend on the gender of the teacher. Hypothesis 13 primarily draws on the findings of Cho (2012).

Hypothesis 13. *On average, teachers are expected to exhibit greater ethnic and gender biases across all specifications against students with different ethnicity and gender than themselves, compared to teachers with the same ethnicity and gender as the student.*

6.8 Hypotheses regarding heterogeneity: Teacher experience

Hypothesis 14. *Experienced teachers are expected to exhibit, on average, lower ethnic and gender bias across all specifications.*

7 Effect of foreign ethnicity and gender

In our analysis, we aim to gain insights into the impact of foreign, native, male, and female names on teacher evaluations in a holistic sense. We investigate the extent to which there is variation in teachers' average biases across these groups and whether there is heterogeneity in teacher biases within these groups based on dimensions such as student ability, student gender and ethnicity, teacher experience, and the gender of the teacher. Additionally, we examine assignments in both Mathematics and Norwegian language to assess the presence of bias heterogeneity across different subjects.

7.1 Primary specification

In our primary specification, we examine the impact of foreign and male names on students' average grades by estimating the following equation:

$$Y_{ij} = \alpha_j + \beta T_{ij} + \epsilon_{ij}, \tag{1}$$

where

- Y_{ij}: A variable representing teacher j's evaluation of assignment i
- T_{ij}: An indicator that equals 1 if assignment i is in the treatment group and 0 if it is in the control group for teacher j when evaluating hypotheses 1 to 12.
- α_j: Teacher fixed effects, as the student names are randomized within teachers.
- ϵ_{ij} represents the error term. For all specifications, we employ clustered standard errors at the teacher level to account for correlations between observations within a teacher.

For all specifications, we will also test our hypothesis without using teacher fixed effects, α_j. In addition, we will include a version where we replace α_j by a vector of control variables, including those mentioned in Section 2.4.[1]

[1]For the regression in Equation 1, this sensitivity test is represented by the following regression:

Consistent with Hypotheses 1 to 4 and 6 to 10, we anticipate rejecting the null hypothesis that $\beta = 0$ in favor of $\beta < 0$ when T_{ij} is an indicator for having a male/foreign name. We expect that assignments with foreign/male names will receive unfavorable evaluations from their teachers.

Consistent with Hypothesis 5, we anticipate rejecting the null hypothesis that $\beta = 0$ in favor of $\beta > 0$ when T_{ij} is an indicator for having a female name. We expect that assignments with female names will, on average, receive favorable evaluations from their teachers.

Following Hypothesis 12, we anticipate that the absolute value of β will be greater in regressions using teacher evaluations in Norwegian language as the outcome variable, compared to the absolute value of β in regressions using evaluations in Mathematics as the outcome variable. In some regressions, we will jointly estimate the coefficients to compare effect sizes across hypotheses.

8 Heterogeneous effects within gender and ethnicity

As suggested by Alne and Herstad (2020), teacher biases exhibit variations based on observable characteristics. We aim to estimate whether teacher biases differ across the students' ability level, the teachers' gender, the teachers' work experience, and the students' gender and ethnicity. Previous studies have not extensively explored the variation in bias across ability, teacher experience, and the teachers' gender. Nevertheless, we seek to estimate whether bias varies along these dimensions.

8.1 Ability within ethnicity and gender

$$Y_{ij} = \alpha_j + \beta T_{ij} + \gamma_A Ability_i + \beta_A Ability_i x T_{ij} + \epsilon_{ij} \tag{3}$$

where $Ability_i$ is a variable taking values between one and six, representing the average

$$Y_{ij} = \alpha_0 + \beta T_{ij} + X_j \gamma^X + \epsilon_{ij}, \tag{2}$$

We will conduct similar sensitivity tests for other regressions as well. As a result, we will estimate two versions of most regressions: one with α_j, as a measure of the average strictness of the teacher, and another specification that includes covariates to enhance statistical power. Equation 2 will also be estimated without a vector of control variables.

grade assigned by teachers who do not see a student's name when grading assignments. To account for heterogeneity in blind ability, we will also convert this continuous average into an ordinal variable, similar to the grade scale. Our primary focus is on the coefficient β_A, which provides an estimate of the differential effect of having a foreign/male name versus a native/female/no name for different ability levels.[2] In this specification, T_{ij} is equal to 1 if the student has a foreign/male name, and 0 if the student has a native/female/no name.

Consistent with Hypothesis 11, we anticipate rejecting the null hypothesis that $\beta_A = 0$ in favor of $\beta_A < 0$. We expect that students whose assignments receive a lower average grade in blind evaluations are likely to experience, on average, larger ethnic and gender biases across all specifications, as bias is expected to have a negative relationship with ability.

8.2 The teachers' gender within the students' gender

$$Y_{ij} = \alpha_j + \beta T_{ij} + \gamma_S \text{SameGender}_j + \beta_S \text{SameGender}_j x T_{ij} + \epsilon_{ij} \tag{5}$$

where $SameGender_j$ takes the value 1 if the assignment is graded by a teacher who shares the same gender as student i and 0 if it's graded by a teacher of a different gender than student i, . Our primary focus here is the coefficient β_S, which provides an estimate of the differential effect of having a same gender grader versus a different gender grader for assignments of students with a given gender. In this specification, T_{ij} equals 1 if the student has been assigned a name signaling same gender, and 0 if the student has been assigned no name, or a different gendered name.

Consistent with Hypothesis 13, we anticipate rejecting the null hypothesis that $\beta_S = 0$ in favor of $\beta_S > 0$. We expect that assignments graded by same gender teachers will, on average, experience smaller gender biases across all specifications. We do not have a enough power to explore this hypothesis for teachers and students with the same ethnic background.

[2] In the non-parametric case, the parameters of interest are denoted as β_g in:

$$Y_{ij} = \alpha_j + \sum_{g=2}^{6} \beta_g (T_{ij} * \mathbf{I}\{\text{j has ability g}\}) + \epsilon_{ij}, \tag{4}$$

where g denotes the ordinal variable generated from the average grade received by each assignment in a blind grading environment.

8.3 The teachers' work experience within the students' ethnicity and gender

$$Y_{ij} = \alpha_j + \beta T_{ij} + \beta_W \text{Workexperience}_j \times T_{ij} + \epsilon_{ij} \tag{6}$$

where Workexperience_j is either the number of years the grader has worked as a teacher or a binary variable that equals 1 if the teacher has more than 10 years of teaching experience. Our primary focus in this context is the coefficient β_W, which provides an estimate of the differential effect of the teacher's work experience on gender and ethnic bias. In this specification, T_{ij} is equal to 1 if the student has been assigned a male/ethnic name, and 0 if the student has been assigned no name, a female, or a native name.

In line with Hypothesis 14, we expect to reject the null that $\beta_W = 0$ in favor of $|\beta + \beta_W| < |\beta|$. We expect that assignments, which are graded by teachers with long experience, will on average receive a lower ethnic and gender bias across all specifications.

9 Test of signaling of gender and ethnicity within assignments

$$Y_{ij} = \alpha_j + \delta_i + \beta T_{ij} + \epsilon_{ij}, \tag{7}$$

As a robustness check, we include the regression specified in Equation 7, where we introduce δ_i representing fixed effects for the original assignment number or original student fixed effects. This specification essentially employs a two-way fixed-effects difference-in-differences approach, which assesses whether there are any indications of gender and ethnicity signals in the original assignment and the extent to which this influences our estimated bias

10 Robustness specification heterogeneity

To analyze whether there are certain assignments that contribute to the estimated bias to a greater extent than others, we estimate the regression in Equation 8.

$$Y_{ij} = \alpha_j + \beta_i T_{ij} + \epsilon_{ij} \tag{8}$$

In Equation 8 we interact our treatment indicator with an indicator for each original assignment to assess the extent to which each original assignment contributes to our estimated average effect. The inclusion of the regression in Equation 8 serves the primary purpose of gaining insights into the distribution of the estimated bias. Additionally, it allows us to qualitatively understand the characteristics of the assignments that appear to significantly influence the estimated averages and, thus, provides insight into the mechanisms underlying the estimated bias.

11 Implicit Association Bias

In the project, our intention is to conduct a test of the graders' implicit association bias following standardized procedures. We will then utilize this test to explore the mechanisms behind the estimated bias.

References

Ahmad, A. (2020). Ethnic discrimination against second-generation immigrants in hiring: empirical evidence from a correspondence test. *European societies 22*(5), 659–681.

Alne, R. and E. I. Herstad (2020). Heterogeneity in educational gender biases and the effect on labor market outcomes. *Available at SSRN 3479209*.

Ba, B. A., D. Knox, J. Mummolo, and R. Rivera (2021). The role of officer race and gender in police-civilian interactions in chicago. *Science 371*(6530), 696–702.

Cho, I. (2012). The effect of teacher–student gender matching: Evidence from oecd countries. *Economics of Education Review 31*(3), 54–67.

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

Hanna, R. N. and L. L. Linden (2012). Discrimination in grading. *American Economic Journal: Economic Policy 4*(4), 146–168.

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

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), 2083–2105.

Midtbøen, A. H. (2015). The context of employment discrimination: interpreting the findings of a field experiment. *The British journal of sociology 66*(1), 193–214.

Midtbøen, A. H. (2016). Discrimination of the second generation: Evidence from a field experiment in norway. *Journal of International Migration and Integration 17*, 253–272.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature 56*(3), 799–866.

Shea, J. (2022). Testing for racial bias in police traffic searches. *University of Illinois, Champaign Urbana, USA*.

Sprietsma, M. (2013). Discrimination in grading: Experimental evidence from primary school

teachers. *Empirical economics 45*, 523–538.

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

A Names

- Immigrant females:
  - Faduma (Is the Somali version of the Arabic name, Fatima.)
  - Fawzia (Common Somali name)
  - Ayesha (Nr. 1 in Pakistan)
  - Noor (Nr. 12 Pakistan)
  - Aljicja (Nr. 9 Poland)
  - Valentyna (Nr. 6 Ukraine)
  - Althea (Nr. 3 Philippines)
- Immigrant males:
  - Mohamed (Number 1 in Somalia)
  - Abdi (Number 2 in Somalia)
  - Muhammad (Number 1 i Pakistan)
  - Abdul (Nr. 3 in Pakistan)
  - Szymon (Nr. 4 in Poland)
  - Vladymyr (Nr.2 in Ukraine)
  - Rodrigo (Nr. 94 in Chile)
- Native males:
  - Emil (Number 2 in Norway in 2012)
  - Mathias (Number 3 in Norway in 2012)
  - Jonas (Number 4 in Norway in 2012)
  - Alexander (Number 5 in Norway in 2012)
  - William (Number 6 in Norway in 2012)

  - Oskar (Number 7 in Norway in 2012)
  - Magnus (Number 8 in Norway in 2012)

- Native females:
  - Nora (Number 1 in Norway in 2012)
  - Emma (Number 2 in Norway in 2012)
  - Sofie (Number 3 in Norway in 2012)
  - Linnea (Number 4 in Norway in 2012)
  - Emilie (Number 6 in Norway in 2012)
  - Ingrid (Number 7 in Norway in 2012)
  - Tea (Number 8 in Norway in 2012)

- Blind control:
  - No names assigned