The Importance of External Assessments: High School Math and Gender Gaps in STEM Degrees[1]

Simon Burgess, Daniel Sloth Hauberg, Beatrice Schindler Rangvid, & Hans Henrik Sievertsen[2]

April 2022

Abstract

We exploit the random allocation to a semi-external assessment in Math (SEAM) at the end of high school in Denmark to test the effect of SEAM on subsequent enrollment and graduation in post-secondary education. We find that SEAM in high school reduces the gender gap in graduation from post-secondary STEM degrees, and we discuss possible mechanisms. Our results show that cancelling external assessments, as was temporarily implemented in many regions during the COVID-19 pandemic, may impact gender differences in human capital accumulation in the long run.

Highlights:

- During the COVID-19 pandemic many countries cancelled exams.
- We exploit the random allocation to Math exams at end of high school in Denmark.
- Math exams reduce gender differences in graduation from STEM degrees.
- Replacing exams with teacher assessments may have effects on human capital accumulation.

JEL

I20, I24

---

[1] We thank Thomas Dee, Ellen Greaves, Ulrik Hvidman, Richard Murphy, Stefania Simion, Sarah Smith, Anna Vignoles, Miriam Wüst, two anonymous referees, and workshop participants at the IZA Workshop on the Economics of Education and the Joint IZA & Jacobs Center Workshop: Consequences of COVID-19 for Child and Youth Development for helpful comments.

[2] Burgess: University of Bristol and IZA; Rangvid: VIVE; Sievertsen: University of Bristol, VIVE, and IZA. Corresponding Author: Hans Henrik Sievertsen, University of Bristol, School of Economics, Priory Road Complex, Bristol BS8 1TU. E-mail: h.h.sievertsen@bristol.ac.uk.

1 Introduction

Gender gaps persist in educational attainment. Although women have overtaken men in college completion rates in most developed countries[3], women are still less likely to enroll and graduate in STEM (Science, technology, engineering, and Math) degrees (OECD 2017, Joensen and Nielsen 2016). STEM degrees in turn often lead on to STEM careers, which are typically among the highest paid, so lower female participation in STEM has implications for gender earnings inequality.

In this paper we study how the assignment to a semi-external assessment (SEAM) at the end of high school affects gender differences in enrollment and graduation in post-secondary education. Our motivation to study the role of assessments is twofold. Firstly, one potential explanation for lower STEM enrollment is a gender difference in students' own beliefs about their ability in these subjects, and principally in Math.[4] A key source for students' belief formation is the professional opinion of people they work with all the time – their teachers (Dee 2015; Gershenson, Holt, and Papageorge 2016). Evidence from several countries suggests that these internal teacher assessments might suffer from stereotype biases (Burgess and Greaves, 2013; Falch and Naper, 2013; Rangvid, 2015). Increasing the share of external assessments in terms of a SEAM might therefore affect students' beliefs and choices. Secondly, during the COVID-19 lockdown, countries around the world cancelled external and blind exams and replaced them with teacher assessments. This study provides conservative estimates of the potential long-run consequences for human capital accumulation as we study the effect of a semi-external and non-blind exam.

We use an extraordinary context in which students are randomly assigned to have an oral exam with an external examiner present in a subset of the subjects they are taking. As an example, consider two students, Alice and Carol, who both attended Math and Danish in high school. However, at the end of high school, Alice is randomly allocated to an oral exam in Math, but not in Danish, and Carol is randomly allocated to an oral exam in Danish, but not in Math. Because Alice (Carol) did not attend an oral exam in Danish (Math), she will receive a teacher assessed mark in that unit instead. Note that the teacher assessment will not reveal new information to the

[3] https://data.oecd.org/eduatt/population-with-tertiary-education.htm. Available by gender.

[4] Other leading ideas include academic skills and preparation, family background and expectations, tastes, or preferences such as those related to pecuniary payoffs or the work environment.

student as the teacher assessment is identical to the teacher assessment which the students receive in any case. If Alice and Carol face a stereotype bias in teacher assessments in Math, the new feedback from being randomly allocated to an oral exam in Math (=SEAM), might update Alice's beliefs about her Math abilities and change her subsequent educational choices.

In the illustration above students are either assigned to an oral exam with an external examiner present in Math (Alice) or in Danish (Carol), but in practice they are assigned to several oral exams in any of the subjects attended. We define the treatment we study as being assigned to the oral exam in Math and the counter-factual being assigned to an oral exam in any other subject. We consider this specific treatment as semi-external because it provides the students with some external assessment, additional to the assessment of their own teacher. Given the ample evidence of biases in teacher assessments, the exam lottery might play a non-trivial role in shaping individuals' beliefs.

Using data on the universe of graduates from Danish high schools in the years 2003 to 2007, we exploit the random assignment to SEAM to explore the likelihood of graduating from post-secondary education after high school. We find that for a female student, being assigned the SEAM is a lottery win. The overall baseline gender gap in graduating from a (*Math-requiring*) STEM degree is substantial: 4.6 percent of girls graduated from such a degree, and 10.1 percent of boys. With an oral exam in Math this gap is reduced by 1.2 percentage points, or more than a fifth. The effect is stronger still for graduating from a (*Math-demanding)* STEM degree, where the gender gap is reduced by more than a half. Here the overall baseline gap is 1.1 percentage points, between 1.9% of girls and 3% of boys. Among students assigned to SEAM, the gap falls to 0.4 percentage points.

This study contributes to the literature in two ways. First, we contribute to the literature on gender differences in enrollment and graduation from STEM degrees. Joenson & Nielsen (2016) show how lowering the costs of attending the advanced math track in high school induced girls to take more math and led to higher earnings later in life. Terrier (2020) shows that girls who benefit from teacher favoritism are more likely to select a science track in high school. Carlana (2019) shows that teacher gender stereotypes induce girls to underperform in math. Together these studies suggest that assessments and high school math can affect educational choice and lifetime earnings. Second, our work adds to the much smaller literature on the medium-term consequences of teacher

discretion in high stakes assessments. Diamond and Persson (2016) and Dee et al. (2019) document significant manipulation of test scores around discrete grade cutoffs in Sweden and the US and find that inflating a score increases educational attainment.

The paper is organised as follows. Section 2 provides a brief description of the institutional setting. Section 3 presents the research design and the data. Section 4 presents the results. Section 5 offers some wider conclusions from the results.

2 Institutional Setting

2.1 The Danish School System

For the individuals in our sample, compulsory schooling started in August of the calendar year they turned seven years old and ended after nine years. Having completed compulsory schooling, approximately 55 percent of a cohort continue to high school (either general academic or vocational), 25 percent continue in vocational education and training and the remaining 20 percent either join the labour force or attend other educational options.[5]

The main objective of the high school programs is to prepare students for higher education, and both the academic and vocational high schools provide the same access to higher education. The high school programs all consist of a range of subjects that can be studied on three levels. Level A, the most advanced level, spans all three years; level B spans two years of high school, and level C units one year.[6]

At the end of the last high-school year, students are given teacher grades in all subjects and sit exams in a subset of subjects. The overall composite GPA is the mean of two intermediate averages. The first is the average of teacher assessments. The second intermediate score is the average across national exams administered by the Ministry of Education, exams marked by

[5] There is also a two-year high school program (called "HF"). This program requires students to have attended the optional tenth grade. In this study, we do not consider students in this program. See Ministry of Higher Education and Science (2016) for more information on the Danish education system.

[6] Our description of the organisation of the Danish high school is based on the "High school law" (in Danish: "Bekendtgørelse af lov om gymnasiet m.v.") of October 8, 2003 available here: https://www.retsinformation.dk/eli/lta/2003/833.

independent (i.e., external) examiners. For subjects not assessed by exam, the teacher assessment will carry double weight in the GPA calculations.

Access to post-secondary programs is almost exclusively determined by the high school GPA. After completing high school, all students who wish to enroll in a post-secondary program apply through a centralized system with a prioritized list of educational programs. The programs set the number of available slots and the course requirements (for example, economics at the University of Copenhagen requires A-level in Math and in Danish and B-level in history and in English). All students who fulfil the course requirements for the program are ranked according to their overall high school GPA, and the students ranked within the capacity constraints are given an offer. The high school GPA is thus particularly important for the students who wish to continue in post-secondary schooling.[7]

2.2 The Danish High School Exam Lottery

All students must sit a written examination for each of their A-level subjects, except for the subject History where there is no written examination. A particularity of the Danish high school system is that another four exams are determined by a lottery. For each student, the lottery randomizes them to an oral exam with an external examiner present in four of the subjects they attended. Specifically concerning the focus of this study: students taking A-level Math must all sit the *written* Math exam, while some students are also assigned an *oral* Math exam with an external examiner present (the others are assigned to an oral exam in another subject). The treatment in our analysis is being assigned to an oral Math exam and thus receiving *additional* (semi-) external information on ability in Math. The performance in the oral Math exam is evaluated jointly by the student's teacher and an external examiner and the exam is therefore considered a more objective assessment than the own-teacher score alone. Also, while the teacher assessment is given for course work during a full

[7] For more information, see https://ufm.dk/en/education/admission-and-guidance/how-to-apply-for-a-higher-education-programme-in-denmark-1.

term, the exam assesses a specific performance at an exam. This might leave less room for discretion in grading.[8]

3 Research Design

3.1 Empirical Strategy

To estimate the effect of the external Math exam on girls' and boys' subsequent education, we estimate the following equation using Ordinary Least Squares:

$$y_i = \beta_0 + \beta_1 SEAM_i + \beta_2 Female_i + \beta_3 Female_i \times SEAM_i + \boldsymbol{\beta}'\mathbf{X}_i + e_i$$

where y is our outcome of interest, for example enrollment in higher education. The variable SEAM is a $(0, 1)$ dummy for being assigned to an oral exam in Math. $\mathbf{X}$ is a vector of controls including cohort and high school fixed effects, high school program fixed effects (see section 3.2 for a description of the high school program), as well as indicators for parental educational level (separately for mother and father), and parental income decile. We cluster the standard errors at the program level.

3.2 Data

We use detailed high school records from the administrative high-school database from Statistics Denmark on the full population of students (83,880) who graduated from a three-year academic high school program in the years 2003 to 2007, as well as students (34,584) who graduated from a vocational high school in the years 2004 to 2007.[9] Data for academic high schools are not available before 2003 and data for vocational high schools are not available before 2004. We limit our sample to 2007, to allow us to track students for ten years after high school. Moreover, major reforms in 2008 changed the curriculum and the grading scale. The high school records are linked to administrative records from Statistics Denmark using a unique personal identifier. We thereby

[8] For further details about the exam lottery, see the law of "High school examination" (in Danish: "Bekendtgørelse om studentereksamen og om højere forberedelseseksamen" of October 20, 2003 available here: https://www.retsinformation.dk/eli/lta/2003/850, as well as Ministry of Children and Education (2020).

[9] Note that these numbers are after we delete 57 individuals with missing information about gender and 560 observations who study a unique high school program which means that there is no variation after conditioning on the fixed effects.

obtain information about gender, parental income, and parental education. We also link the individual to education registers to obtain information on higher education enrollment and graduation.

The treatment variable, SEAM, takes the value of one if the student was assigned to an oral exam in A-level math and zero otherwise. We consider five outcome variables: First, if SEAM causes a higher high school GPA it will give access to more university programs as described in section 2.1. To test for such a "mechanical" relationship, we use the overall GPA as the dependent variable. Second, we create a variable that takes the value of one if the individual enrolled in higher education within the ten years after completing high school, and zero otherwise. Third, we create a variable that takes the value of one if the individual graduated with a higher education degree within the ten years after completing high school, and zero otherwise. Fourth, as we expect SEAM to affect beliefs about Math skills, we test the effect on completing a degree *requiring* A-level Math using a variable that takes the value of one if everyone who graduated from that degree completed A-level Math, and zero otherwise. These degrees includes subjects such as engineering, nanoscience, math, physics, and statistics. Fifth, to measure the effect on completing a Math *demanding* degree, we create a variable that takes the value of one if the average written high school Math exam mark of the graduates are in the top ten percentile of all degrees. Examples of subjects in this group include natural science, technical science, health science, and Math.

To make the students as comparable as possible, we condition on program fixed effects. At enrollment to high school students choose between the science and humanities track, and within these tracks students select their own portfolio of subjects and levels, subject to some restrictions. One restriction is that students on the science track have to attend at least one science subject at the highest level (A-level) and students on the humanities track have to study one foreign language at the highest level (A-level). The students typically attend 15 of about 168 subject times level combinations. Controlling for $\approx 15^{168}$ program fixed effects is not feasible. Instead, we create 2727 program fixed effects in terms of the combination of subject and level among the 20 most common subjects and levels (see Appendix Table A1). Furthermore we control for cohort and high school institution fixed effects, as well as for the income decile of the parents (indicators), and the highest educational degree completed by each parent (indicators). Missing values are set to zero and we

separately include variables that takes the value of one if respectively parental income or education are missing. The characteristics of the parents are measured in the year of the students' graduation.

Table 1 shows summary statistics for our sample. We observe that 58 percent of the high school graduates are girls; however, conditioning on attending A-level Math reduced the female share to 47 percent. We observe that among the students attending A-level Math, 61 percent are treated with SEAM. While 91 percent of the A-level Math sample enroll in a higher education and 82 percent complete a degree within ten years after high school, only 13 percent complete a degree requiring A-level math in high school, and only five percent complete a degree where the average written high school Math exam mark among the graduates was in the highest 10 percent. For the analysis, we restrict the sample to the students 48,165 who attended A-level Math. By doing so, we avoid endogeneity issues related to selection into programs.

Table 1: Summary Statistics

	All		Math sample	
	Mean	SD	Mean	SD
A. Background				
Female	0.58	0.49	0.47	0.5
At least one parent with a university degree	0.19	0.39	0.24	0.43
Parental income (1,000 EUR 2015 level)	37.62	48.84	39.43	62.39
B. Assessments				
Math exam	0.25	0.43	0.61	0.49
GPA	8.24	0.99	8.50	0.97
B. Post-secondary schooling				
Enrolled within 10y	0.86	0.35	0.91	0.29
Graduated within 10y	0.76	0.43	0.82	0.38
Graduated from maths req. degree within 10y	0.07	0.25	0.13	0.34
Graduated from maths dem. degree within 10y	0.02	0.15	0.05	0.22
Observations	118464		48165	

Notes: The Math sample consists of all students who attended A-level Math in high school. Parental variables are measured in the year of high school graduation.

Table 2 shows the difference between exam assessments and teacher assessments in our setting. In line with a stereotype pattern, we observe that relative to boys, girls are more generously assessed by their own teacher in Danish compared to Math. There is little difference in the grade given by their own teacher and in the oral exam in Danish and Math for girls. However, boys on average receive a much higher (lower) mark by their teacher in Math (Danish) than they receive in the oral exam. One potential explanation for why we do not observe gender differences in teacher assessments of the written performance is that teacher assessment of oral performance is based on classroom participation, while the written assessment typically is based on written assignments and in-class tests. The latter is therefore potentially more objective.

Table 2: Grading patterns: Exam assessment-Teacher assessment by gender and subject

		Overall		Danish		Math	
Girls		Mean	SD	Mean	SD	Mean	SD
	Overall	-0.18	1.30	-0.29	1.28	-0.31	1.38
	Oral	-0.02	1.31	0.04	1.34	-0.02	1.42
	Written	-0.44	1.23	-0.48	1.21	-0.48	1.33
Boys							
	Overall	-0.08	1.36	-0.07	1.37	-0.33	1.46
	Oral	0.01	1.37	0.18	1.42	-0.30	1.53
	Written	-0.24	1.33	-0.22	1.32	-0.35	1.41

Notes: The table shows the mean difference between the exam assessment and the teacher assessment as well as the standard deviation of this difference. The sample is the same as in the main analysis. A negative value thus suggests that the teacher gives a higher grade than the exam assessment.

4 Results

4.1 Validity of the research design

To assess the identifying assumption that allocation to SEAM is indeed as-good-as random as stated by the Ministry of Education, we regress the SEAM variable on the full set of covariates, conditional on the program, high school, and cohort fixed effects. The left panel in Figure 1 shows the t-statistics for the 35 variables included in our main specification and Table A2 in the online appendix reports the coefficients and standard errors. If allocation to SEAM is as-good-as-random, we would expect the t-statistics for the included covariates to be small. Indeed, as Figure 1 shows, only one t-statistic exceeds the -1.96 to 1.96 interval suggesting significance at the 5 percent level. Furthermore, we fail to reject null hypothesis that all coefficients are zero, with a p-value of 0.47,

suggesting that the identifying assumption is satisfied. The right panel in Figure 1 shows results from a specification where we also include the written teacher assessment in math, which is also not predictive of SEAM allocation.

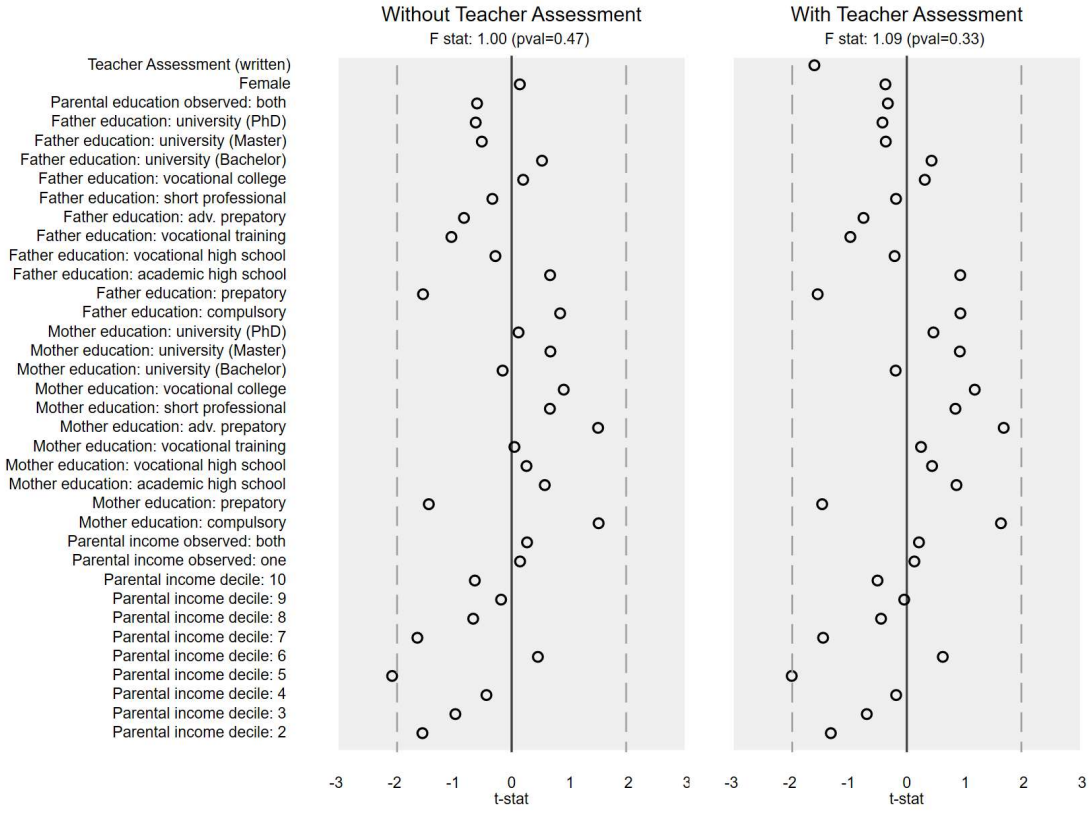


Figure 1: Covariate Balance

Notes: This figure shows the t-statistics from regressing the SEAM variable on all variables listed above, as well as program, high school, and cohort fixed effects. The left panel shows results from a specification without the written teacher assessment in math, as in the main specification. The right panel shows the results from a specification including the written teacher assessment in math. Table A2 in the online appendix lists all coefficients and standard errors. The F-statistic and p-value are for a test of the null-hypothesis that all variables listed above are zero jointly.

4.2 Main results

In Table 3 we show the main results. Column (1) shows a very small effect of SEAM on overall GPA for boys (-0.026 grade points), and no effect for girls (0.004 grade points). This is in line with the patterns shown in Table 2: While girls on average receive the same mark by their teacher

as in the oral exam, boys receive a mark that is 0.30 grade points lower in the exam compared to the teacher assessment. Columns (2) and (3) provide evidence that SEAM increases the lead of female over male students in enrollment and graduation rates: ten years after high school 84 percent of the untreated girls have completed a degree and 80 percent of the untreated boys have completed a degree. This gap increases by 1.5 percentage points if girls and boys are assigned to SEAM.

Table 3: Regression results: the effect of SEAM on subsequent education

	Overall			Math degrees	
	GPA	Enroll	Graduate	Req.	Dem.
	(1)	(2)	(3)	(4)	(5)
Female	0.204***	0.009	0.049***	-0.049***	-0.005
	(0.018)	(0.006)	(0.007)	(0.006)	(0.004)
SEAM	-0.026**	-0.003	-0.007	-0.005	-0.002
	(0.012)	(0.004)	(0.005)	(0.004)	(0.003)
SEAM X Female	0.030*	0.010**	0.015**	0.012**	0.007*
	(0.017)	(0.005)	(0.007)	(0.006)	(0.004)
SEAM+SEAM X Female	0.004	0.006	0.008	0.007	0.005
P-val	0.768	0.092	0.120	0.049	0.040
MDV Female	8.593	0.906	0.842	0.089	0.040
MDV Male	8.432	0.905	0.800	0.165	0.057
Observations	48165	48165	48165	48165	48165

Notes: SEAM+SEAM X female shows the sum of the coefficient on the variable SEAM and the coefficient on the interaction between SEAM and female. P-val shows the p-value for the test of the null hypothesis that the sum of the coefficient on SEAM and the coefficient on SEAM interacted with female is zero. MDV is the mean of the dependent variable for untreated students. All models are estimated with the full set of background controls, cohort, high school and program fixed effects. Background controls include indicators for parental income decile, and indicators for parental educational level (separate for each parent). Standard errors clustered on the high school program level in parenthesis. Asterisks indicate significance at the following levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

In columns (4) and (5) we show results for graduation from Math-intensive higher education degrees. Column (4) shows that 8.9 percent of untreated girls graduated from a Math *required* degree, compared to 16.5 percent of boys. The effect of SEAM is again positive for female students, reducing this gap by 1.2 percentage point, or about a sixth. As shown in column (5), there are also substantial gender gaps in mean graduation rates for Math *demanding* degrees: 4.0 percent of untreated girls have graduated from such a degree compared to 5.7 percent among

untreated boys. Strikingly, the results in column (5) show that when girls and boys both are assigned to SEAM in Math reduces the gap by 0.7 percentages points, corresponding to more than 40 percent.

Appendix Figure A1 shows results for 16 different specifications including a specification using a probit instead of a linear probability model (for all binary dependent variables), controlling for prior attainment, and excluding controls. Overall, the point estimates are consistent across specifications, and our preferred specification, used in Table 3, is not an outlier.

Due to data limitations, we must remain somewhat agnostic about why assignment to SEAM affects long-run outcomes. However, we will briefly discuss three potential mechanisms. First, as column (1) in Table 3 suggests a small effect of SEAM on the overall GPA there could be a mechanical effect of SEAM by simply giving access to more education programs through a higher GPA. However, the GPA on the transcripts and higher education enrollment GPA cutoffs are defined with one decimal place, and both the level effect of SEAM and the interaction with female are less than one-third of 0.1. The results thus suggest that there is limited scope for a purely "mechanical" effect where effects on subsequent education are driven by having access to more degrees.

Second, SEAM might induce students to invest more in studying Math, which in turn might increase their interest in pursuing a STEM career. However, as students typically are assigned to SEAM in May and attend the exam in June, there is limited scope for such a "human capital" effect. In Table A3 in the online appendix we show that being assigned to SEAM does not affect the performance in the blind written math exam. However, we acknowledge that this is not an ideal test of the human capital channel, because the time between the SEAM announcement and the written exam is even shorter than the time between the SEAM announcement and the oral exam.

Third, SEAM might affect the students' beliefs about their Math skills. Providing students with an additional "second opinion", the view of an external assessor on their Math performance, may influence their own estimation of their ability in Math, and so lead them to adjust their future plans, university course choices and effort-allocation strategies. Given that the results (Table 3 and Table A3 in the online appendix) suggest a limited role for the first two channels, we believe that changes in beliefs might be an important driver of the identified effects.

We consider heterogeneity by student background and high school type in Table 4. While there is in general little evidence of heterogeneous effects, it is worth pointing out a few cases where effects are larger and more precisely estimated.

Table 4: Regression results: heterogeneous effects of SEAM on subsequent education

	Overall GPA	Enroll	Graduate	Math degrees Req.	Math degrees Dem.
	(1)	(2)	(3)	(4)	(5)
A. By parental income					
SEAM	-0.040**	0.001	-0.012	-0.011	-0.003
	(0.019)	(0.008)	(0.011)	(0.007)	(0.004)
SEAM X female	0.023	0.007	0.018	0.012	0.004
	(0.025)	(0.009)	(0.012)	(0.008)	(0.005)
SEAM X income >p(50)	0.021	-0.008	0.009	0.009	0.003
	(0.027)	(0.010)	(0.013)	(0.009)	(0.005)
SEAM X income >p(50) X female	0.015	0.003	-0.004	0.001	0.006
	(0.037)	(0.012)	(0.014)	(0.011)	(0.007)
Sum (<p50)	-0.017	0.009	0.005	0.002	0.001
P-value (<p50)	0.335	0.180	0.524	0.731	0.804
Sum (>p50)	0.020	0.005	0.011	0.012	0.010
P-value (>p50)	0.302	0.358	0.076	0.012	0.012
B. By parental education (university degree)					
SEAM	-0.031**	-0.008	-0.011	-0.003	-0.001
	(0.015)	(0.005)	(0.007)	(0.005)	(0.003)
SEAM X female	0.047**	0.020***	0.025***	0.007	0.006
	(0.019)	(0.006)	(0.009)	(0.007)	(0.005)
SEAM X university	0.007	0.008	0.002	-0.003	-0.007
	(0.028)	(0.007)	(0.010)	(0.010)	(0.008)
SEAM X university X female	-0.031	-0.023**	-0.018	0.028*	0.016
	(0.039)	(0.010)	(0.014)	(0.015)	(0.011)
Sum (no university)	0.015	0.012	0.014	0.004	0.005
P-value (no university)	0.291	0.015	0.028	0.431	0.084
Sum (university)	-0.008	-0.004	-0.002	0.028	0.013
P-value (university)	0.759	0.461	0.762	0.007	0.079

Continued on the next page

Table 4 continued

	Overall GPA (1)	Enroll (2)	Graduate (3)	Math degrees Req. (4)	Math degrees Dem. (5)
A. By attainment (above median grade in the written Mathematics exam)					
SEAM	-0.020	-0.001	0.000	-0.008	0.002
	(0.014)	(0.006)	(0.009)	(0.006)	(0.002)
SEAM X female	0.018	0.009	0.011	0.017**	-0.002
	(0.020)	(0.009)	(0.010)	(0.007)	(0.004)
SEAM X math >p(50)	-0.017	-0.004	-0.016	0.007	-0.008
	(0.021)	(0.007)	(0.010)	(0.008)	(0.007)
SEAM X math >p(50) X female	0.040	0.001	0.010	-0.012	0.019*
	(0.031)	(0.011)	(0.012)	(0.011)	(0.010)
Sum (<p50)	-0.002	0.008	0.011	0.009	0.000
P-value (<p50)	0.885	0.170	0.125	0.032	0.905
Sum (>p50)	0.021	0.005	0.006	0.005	0.011
P-value (>p50)	0.201	0.156	0.324	0.373	0.021
B. By high school type (academic vs vocational)					
SEAM	-0.040	-0.015*	-0.026**	-0.016	0.003
	(0.031)	(0.008)	(0.011)	(0.010)	(0.004)
SEAM X female	0.060*	0.025	0.049***	0.018*	0.008
	(0.034)	(0.016)	(0.017)	(0.010)	(0.005)
SEAM X academic	0.019	0.016*	0.025**	0.015	-0.006
	(0.034)	(0.009)	(0.013)	(0.011)	(0.005)
SEAM X university X female	-0.038	-0.019	-0.043**	-0.009	-0.001
	(0.039)	(0.016)	(0.018)	(0.011)	(0.008)
Sum (vocational)	0.020	0.009	0.023	0.002	0.011
P-value (vocational)	0.395	0.526	0.114	0.716	0.003
Sum (academic)	0.000	0.006	0.006	0.008	0.004
P-value (academic)	0.973	0.070	0.323	0.054	0.183

Notes: All models are estimated with the full set of background controls, cohort, high school and program fixed effects. Background controls include indicators for parental income decile, and indicators for parental educational level (separate for each parent). Standard errors clustered on the high school program level in parenthesis. Asterisks indicate significance at the following levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Focusing on the results shown in column (5), we observe that for girls of parents with an income below the median, SEAM does not increase the probability of graduating with a Math demanding degree (the sum of SEAM and SEAM X female is 0.001, p=0.804), however, for girls of parents with an income above the median, the combined effect is 0.010 (p=0.012). Similary, but less clear, for girls where at least one parent has a university degree, the combined effect is 0.013 (p=0.079),

compared to 0.005 (p=0.084), for girls of parents without a university degree. Focusing again on column (5), the results in panel C. suggest that effects are stronger for students with above median written Math exam marks with a combined effect of 0.011 (p=0.021) compared to 0.000 (p=0.905) for girls below the median. Interestingly, the pattern in column (4) is different, effects are strongest for students below the median for graduating from a Math requiring degree. Panel D shows a similar pattern to panel C: effects are strongest for girls in academic high schools when focusing on column (4), graduating with a Math required degree, but strongest for girls in vocational high schools when looking at column (5), graduating with a Math demanding degree.

Finally, it is worth noting that the triple interaction between SEAM, female, and the dimension of heterogeneity, is negative and significant in panel B., column (2), suggesting that the gender difference of SEAM is less pronounced for children of highly educated parents, in terms of the impact on graduation. Similarly, the coefficient is negative and significant in panel C., column (3), suggesting that gender differences in the effect of SEAM on graduation are less pronounced in academic high schools. However, it is worth noting that we are testing 20 interaction terms, and while two significant (at the 5 percent level) is more than expected by chance, overall Table 5 suggests little evidence of strong treatment effect heterogeneity.

5 Conclusion

During the COVID-19 pandemic many countries replaced exams with teacher assessments. We use a feature of the Danish education system that provides a random allocation to examinations at the end of high school. The contribution of the paper is to show that these exams are important for human capital development. Concretely, we show that if girls and boys are allocated to an oral exam in advanced math at the end of high school, it reduces the gender gap in graduating from some of the most math demanding higher education degrees ten years later, compared to a situation where they are allocated to an oral exam in another subject.

In practical policy terms, our results emphasize the importance of students having the opportunity to demonstrate what they can actually do, assessed at least in part by an external assessor, as opposed to purely what their teacher thinks they can do. The shift back to relying on teacher assessments that many countries have seen during the COVID-19 pandemic, while understandable, may well have led to widening gender gaps in STEM careers.

References

Burgess, Simon, and Ellen Greaves. 2013. "Test Scores, Subjective Assessment, and Stereotyping of Ethnic Minorities." *Journal of Labor Economics* 31 (3): 535–76.

Carlana, Michela. 2019. "Implicit Stereotypes: Evidence from Teachers' Gender Bias." The Quarterly Journal of Economics 134(3): 1163–1224.

Dee, Thomas S. 2015. "Social Identity and Achievement Gaps: Evidence from an Affirmation Intervention." *Journal of Research on Educational Effectiveness* 8 (2): 149–68.

Dee, Thomas S, Will Dobbie, Brian A Jacob, and Jonah Rockoff. 2019. "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations." *American Economic Journal: Applied Economics* 11 (3): 382–423.

Diamond, Rebecca, and Petra Persson. 2016. "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests." National Bureau of Economic Research.

Falch, Torberg, and Linn Renée Naper. 2013. "Educational Evaluation Schemes and Gender Gaps in Student Achievement." *Economics of Education Review* 36: 12–25.

Gershenson, Seth, Stephen B Holt, and Nicholas W Papageorge. 2016. "Who Believes in Me? The Effect of Student–Teacher Demographic Match on Teacher Expectations." *Economics of Education Review* 52: 209–24.

Joensen, Juanna Schrøter, and Helena Skyt Nielsen. 2016. "Mathematics and Gender: Heterogeneity in Causes and Consequences." *The Economic Journal* 126 (593): 1129–63.

Ministry of Higher Education and Science (2016): The Danish Education System. The Ministry of Higher Education and Science. Copenhagen.

OECD. 2017. "The under-representation of women in STEM fields", in *The Pursuit of Gender Equality: An Uphill Battle*, OECD Publishing, Paris, https://doi.org/10.1787/9789264281318-10-en.

Rangvid, Beatrice Schindler. 2015. "Systematic Differences Across Evaluation Schemes and Educational Choice." *Economics of Education Review* 48: 41–55.

Terrier, Camille. 2020. “Boys Lag Behind: How Teachers’ Gender Biases Affect Student Achievement.” *Economics of Education Review* 77: 101981.

Online Appendix

Table A1: List of subjects and levels

	Subject	Level	Percentage of students attending
1	Danish	A	98.1
2	History	A	69.4
3	Classics	C	68.7
4	Religion	C	68.7
5	Geography	C	67.9
6	Art	C	63.2
7	Biology	C	61.1
8	Music	C	58.1
9	PE	C	57.1
10	German	B	50.9
11	English	A	50.8
12	English	B	47.6
13	Math	A	40.2
14	Physics	B	38.8
15	Chemistry	C	32.3
16	Math	B	29.9
17	Psychology	C	29.7
18	Natural sciences	C	28.8
19	Latin	C	28.6
20	Spanish	C	25.4
21	International economics	B	21.6
22	Business law	C	21.6
23	Contemporary history	B	21.4
24	IT	B	18.4
25	Social sciences	A	18.3
26	Business case	C	16.3
27	PE	B	12.5
28	Sales	A	12.0
29	Social sciences	B	11.8
30	Business economics	B	11.6

Notes: This table lists the 30 most common subject times level combinations for the students in our sample. For example, 98 percent of the students attended A-level math and 40.2 percent attended A-level Math. To create fixed effects for the high school "program" attended we compute all combinations of the 20 most common level times combinations.

Table A2: Covariate balance. Dependent variable: A-level Math exam

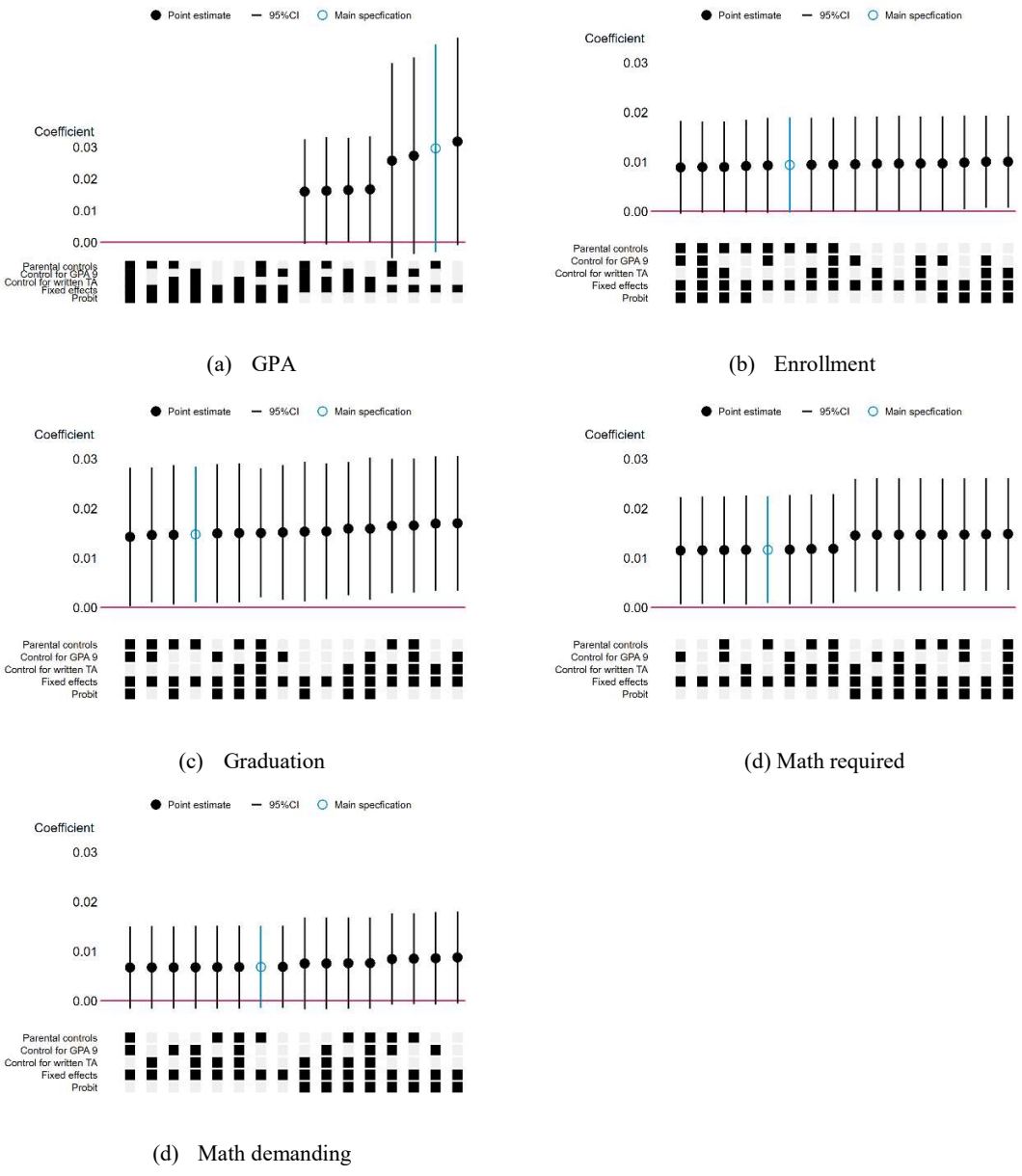
Parental income decile: 2	-0.015	-0.013
	(0.010)	(0.010)
Parental income decile: 3	-0.009	-0.007
	(0.010)	(0.010)
Parental income decile: 4	-0.004	-0.002
	(0.010)	(0.010)
Parental income decile: 5	-0.021**	-0.020**
	(0.010)	(0.010)
Parental income decile: 6	0.004	0.006
	(0.010)	(0.010)
Parental income decile: 7	-0.016	-0.014
	(0.010)	(0.010)
Parental income decile: 8	-0.007	-0.005
	(0.010)	(0.010)
Parental income decile: 9	-0.002	-0.000
	(0.010)	(0.010)
Parental income decile: 10	-0.007	-0.005
	(0.011)	(0.011)
Parental income observed: one	0.004	0.004
	(0.030)	(0.030)
Parental income observed: both	0.007	0.006
	(0.028)	(0.028)
Mother education: compulsory	0.015	0.016
	(0.010)	(0.010)
Mother education: prepatory	-0.077	-0.079
	(0.054)	(0.054)
Mother education: academic high school	0.008	0.013
	(0.015)	(0.015)
Mother education: vocational high school	0.005	0.009
	(0.020)	(0.020)
Mother education: vocational training	0.001	0.003
	(0.011)	(0.011)
Mother education: adv. prepatory	0.010	0.012*
	(0.007)	(0.007)
Mother education: short professional	0.009	0.012
	(0.014)	(0.014)
Mother education: vocational college	0.007	0.010
	(0.008)	(0.008)
Mother education: university (Bachelor)	-0.005	-0.006
	(0.029)	(0.030)

Continued the next page

Table A2 (*continued)*

Mother education: university (Master)	0.007	0.010
	(0.010)	(0.011)
Mother education: university (PhD)	0.003	0.012
	(0.027)	(0.027)
Father education: compulsory	0.010	0.012
	(0.012)	(0.013)
Father education: prepatory	-0.141	-0.141
	(0.093)	(0.093)
Father education: academic high school	0.009	0.012
	(0.013)	(0.014)
Father education: vocational high school	-0.005	-0.004
	(0.018)	(0.019)
Father education: vocational training	-0.018	-0.017
	(0.017)	(0.017)
Father education: adv. prepatory	-0.007	-0.007
	(0.009)	(0.009)
Father education: short professional	-0.004	-0.002
	(0.011)	(0.011)
Father education: vocational college	0.002	0.003
	(0.009)	(0.010)
Father education: university (Bachelor)	0.011	0.009
	(0.021)	(0.022)
Father education: university (Master)	-0.006	-0.004
	(0.012)	(0.012)
Father education: university (PhD)	-0.013	-0.009
	(0.021)	(0.021)
Parental education observed: one	-0.004	-0.002
	(0.007)	(0.007)
Female	0.001	-0.002
	(0.005)	(0.005)
Teacher assessment: written		-0.003
		(0.002)
Constant	0.604***	0.633***
	(0.027)	(0.028)
Observations	48165	47639
Clusters	598	557
F-stat	1.00	1.09
P-val	0.47	0.33

Notes: P-val shows the p-value for the test of the null hypothesis that all coefficients are zero. All models are estimated with the full set of high school, cohort, and program fixed effects. Asterisks indicate significance at the following levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.



(a) GPA

(b) Enrollment

(c) Graduation

(d) Math required

(d) Math demanding

Figure A1: Specification curves

Notes: These five charts show the coefficients and 95% confidence intervals on the interaction term from estimating 16 different specifications of equation 1 using the dependent variables listed in Table 2. The specifications are indicated by the markers below the plot. In the first row a black square indicates that the specification includes parental controls. In the second row a black square indicates that the specification includes control for 9th grade GPA (i.e. before enrollment in high school). In the third row a black square indicates that we also control for the teacher assessment for written high-school Math. In the fifth row a black square indicates that the specification is estimated as a probit instead of a linear probability model.

Table A3: Regression results: the effect of SEAM Math GPA and written Math exam grade.

	Math GPA (1)	Written Math Exam (2)
Female	0.436***	0.425***
	(0.033)	(0.040)
SEAM	-0.105***	0.014
	(0.021)	(0.029)
SEAM X female	0.085***	-0.017
	(0.031)	(0.041)
SEAM+SEAM X female	-0.020	-0.004
P-val	0.334	0.893
MDV Female	8.067	8.106
MDV Male	7.613	7.794
Observations	47560	48143

Notes: SEAM+SEAM X female shows the sum of the coefficient on the variable SEAM and the coefficient on the interaction between SEAM and female. P-val shows the p-value for the test of the null hypothesis that the sum of the coefficient on SEAM and the coefficient on SEAM interacted with female is zero. MDV is the mean of the dependent variable. All models are estimated with the full set of background controls, cohort, high school and program fixed effects. Background controls include indicators for parental income decile, and indicators for parental educational level (separate for each parent). Standard errors clustered on the high school program level in parenthesis. Asterisks indicate significance at the following levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.