CESifo
AREA CONFERENCES 2019

Economics of Education

Munich, 30–31 August 2019

Gender Gap Variation across Assessment Forms: Explanations and Implications

Georg Graetz and Arizo Karimi

20 YEARS CESifo

Gender Gap Variation across Assessment Forms: Explanations and Implications*

Georg Graetz† Arizo Karimi‡

August 23, 2019

Abstract

Females consistently outperform males in school grades, while standardized achievement tests rarely exhibit pronounced gender gaps. We document that in Sweden, females outperform males on compulsory and high school GPAs by a third of a standard deviation, while males outperform females on the Swedish version of the SAT by the same magnitude. We establish that GPAs capture different attributes and skills compared to SAT scores. Differences in motivation and effort explain up to 60 percent of the female advantage in GPAs, while differences in cognitive skills explain 40 percent of the male advantage in SAT scores among the self-selected sample of SAT takers. Because motivation predicts SAT participation but is not rewarded in terms of scores, women are more likely to participate in the SAT, but are negatively selected on both observed and unobserved ability. Quantifying this mechanism using a selection model, we are able to explain 70 percent of the SAT gender gap. Our results show that understanding the origins of gender gaps in non-cognitive traits is key to informing the debate about why boys under-perform, relative to girls, in school. Our findings also imply large effects of the choice of university admission criterion on students' characteristics, which has policy relevance not least because relative returns to non-cognitive skills in the labor market appear to be increasing. Finally, we demonstrate the implications of self-selection for the interpretation of gender gaps.

Keywords: gender gaps, student assessment, cognitive skills, non-cognitive skills, university admissions, selection

JEL Classification: I21, I24, J16

*We thank Peter Arcidiacono, Susan Dynarski, Nicole Fortin, Helena Holmlund, Erica Lindahl, Mattias Nordin, Tuomas Pekkarinen, Oskar Nordström Skans and seminar participants at the Uppsala Center for Labor Studies for helpful comments and suggestions.

†Corresponding author. Department of Economics, Uppsala University; Centre for Economic Performance, CESifo, IZA, and Uppsala Center for Labor Studies. Email: georg.graetz@nek.uu.se.

‡Department of Economics, Uppsala University; Institute for Evaluation of Labor Market and Education Policy (IFAU), and Uppsala Center for Labor Studies. Email: arizo.karimi@nek.uu.se.

1 Introduction

Women outnumber men in university attendance rates in most OECD countries (Goldin, Katz, and Kuziemko, 2006), giving rise to both public debate and research into why females systematically outperform males in compulsory and high school grade point averages (GPAs), and are less likely to drop-out of high school (Murnane, 2013; OECD, 2017; SCB, 2017). However, a challenge for understanding gender gaps in educational outcomes is that differences in measured performance consistently vary with assessment form. In particular, despite the female advantage in overall GPAs from school, males tend to perform at least as well as females on standardized aptitude and achievment tests (see for instance Duckworth and Seligman, 2006), and typically outperform females on tests measuring quantitative skills (see for instance Fryer Jr and Levitt, 2010).[1] As both GPAs and standardized tests are widely used in university admission procedures, and as academics and policy makers debate their relative merits (Borghans, Golsteyn, Heckman, and Humphries, 2016; Rothstein, 2004; Diamond and Persson, 2016), it is important to understand what attributes each type of assessment captures, not least in order to shed light on the variation in gender gaps across these assessment formats.[2]

Using administrative data on the Swedish population, we document that females, on average, outperform males on both compulsory school and high school GPAs by about a third of a standard deviation, but that the reverse is true for the Swedish SAT, where females *under*-perform relative to males by a third of a standard deviation. These gaps are stable across the cohorts covered by our data, born between 1977 and 1996. The pattern of a flipped gender gap across school GPAs and SAT scores is present also within subject areas (verbal and quantitative), and there is a sizeable gender gap in within-individual score differences: a female student's position in the score distribution deteriorates by half a standard deviation, relative to males, when moving from school grades to the SAT.

We test for two potential explanations for the flipping gender gap across GPAs and SAT scores. First, we investigate whether school grades capture different individual attributes or skills than do SAT scores, and if so, to what extent gender differences in the endowments of these attributes can account for the gaps. While standardized achievement tests and grades both measure students' acquired knowledge and skills, previous evidence suggests that the two assessment formats differ in important ways. In particular, relative to standardized achievement tests, course grades seem more strongly associated with personality traits like conscientiousness, which is generally higher in girls than in boys.[3]

Second, we ask whether self-selection into taking the SAT differs across the genders. This may happen if, for example, there are gender differences in the endowment of attributes that predict SAT

[1] The latter finding is particularly puzzling as there appears to be no systematic gender difference in early-age mean numerical ability (Kersey, Braham, Csumitta, Libertus, and Cantlon, 2018), nor in regions or countries with a more gender-equal culture (Guiso, Monte, Sapienza, and Zingales, 2008; Pope and Sydnor, 2010; Gevrek, Neumeier, and Gevrek, 2018). Some evidence, however, points to greater variance in test scores in the male population (see, for instance Machin and Pekkarinen, 2008).

[2] In particular, one of the main rationales for standardized university admissions tests is to encourage a diversified student body (Dynarski, 2017). See Edwards, Coates, and Friedman (2012) for an overview of countries using central admissions tests.

[3] See Borghans, Golsteyn, Heckman, and Humphries (2016); Heckman and Kautz (2012); Almlund, Duckworth, Heckman, and Kautz (2011). Duckworth and Seligman (2006) find that female advantage in self-discipline among eighth graders accounts for a larger portion of the female dominance in report card grades than in achievement test scores. Similarly, Cornwell, Mustard, and Van Parys (2013) find that non-cognitive factors account for the female advantage in teacher assessed grades among primary school students in the US. Fortin, Oreopoulos, and Phipps (2015) focus on the shift in the mode of girls' high school GPA from B to A that occurred between the 1980s and 2000s in the US, leaving boys behind, and conclude that gender differences in expectations for attending higher education are the most important factors accounting for this trend.

participation. Self-selection potentially interacts with the first explanation, as it may create or reinforce gender imbalances in attributes that are differentially rewarded in the SAT. More generally, in any setting where there is choice involved and potential for systematically different decisions made by males and females, gender gaps cannot be taken at face value (see for instance Ahn, Arcidiacono, Hopson, and Thomas, 2015 and Card and Payne, 2017 for evidence on the sources of the gender gap in STEM enrolment).

We test these explanations using data on cognitive skills, motivation, and study effort measured at ages 13 and 16 for a representative 10-percent sample of Swedish students born in 1992, which we link to population-wide individual-level data on compulsory school and high school GPAs (measured at ages 16 and 19, respectively) as well as SAT scores (typically measured at ages 19-20). Our paper thus informs the literature on which types of traits and skills different assessment formats reflect, which focuses mostly on personality correlates with grades, or on the relationship between IQ and test scores (see e.g. Borghans, Golsteyn, Heckman, and Humphries, 2016; Heckman and Kautz, 2012; Noftle and Robins, 2007). Our data, with detailed information on skills, attributes, grades, and scores for a representative longitudinal sample of students, combined with population-wide data allow us to make several contributions to this literature. In particular, we are able to characterize the predictive power of cognitive and non-cognitive skills, both along multiple dimensions, for grades as well as standardized test scores; and we are able to explore the sources of gender gap variation across assessment formats, including a rich characterization of differential patterns of self-selection.

We find that cognitive skills, motivation, and effort are strongly positively related to compulsory school GPA (CSGPA) and high school GPA (HSGPA). While SAT scores are also highly informative about cognitive skills, they show no correlation with motivation or effort. This suggests that school-level assessments capture different attributes than the SAT.[4] With respect to gender gaps, we find a pronounced female advantage in motivation and effort in the representative sample, which accounts for over 60 percent of the female advantage in CSGPA, and for 30 percent in the case of HSGPA.

These results extend to alternative measures of cognitive and non-cognitive skills which are available for a much larger sample (65 percent or more of birth cohorts 1990-1996): The skills of students' fathers measured at military enlistment.[5] We find that the unconditional relationships between father's cognitive skills and a student's grades and SAT scores are both strongly positive. The unconditional relationship between father's *non*-cognitive skills and own grades is similarly strong, but that with own SAT scores is markedly weaker, and disappears once controlling for father's cognitive skills. In the case of grades, both cognitive and non-cognitive skills retain their predictive power when analyzed jointly. The cognitive tests from the two data sets (military enlistment and the 1992-cohort study) are highly similar, while the non-cognitive tests or questionnaires are designed to capture different traits across the two assessments. Nevertheless, the relative importance of cognitive vis-a-vis non-cognitive skills for grades and achievement tests is maintained across the two data sources, strengthening our conclusion that school level assessments capture partly different traits than standardized achievement tests. Since father's skills are balanced across genders in the population, they do however not contribute to explaining the gender

[4]This is consistent with, e.g., Borghans, Golsteyn, Heckman, and Humphries (2016), who report that personality is more important in predicting grades than achievement tests, and IQ more important in predicting scores on achievement tests.

[5]Grönqvist, Öckert, and Vlachos (2017) show that the intergenerational transmission of cognitive and non-cognitive skills between fathers and sons is 0.32–0.35 and 0.2, respectively, using Swedish military enlistment data.

gap in grades.

Turning to the gender gap in SAT scores, we find that among SAT takers, males have higher (own) cognitive skills than females, especially along the dimensions rewarded by the test. These differences account for more than 40 percent of the female disadvantage in SAT scores. (In the representative sample as a whole there are no gender differences in cognitive skills, so cognitive skills do not help explain the female advantage in GPAs.) Moreover, we find that females are 8 percentage points more likely to take the SAT than males, which appears largely driven by their higher motivation. This suggests an important role for differential self-selection into taking the test in explaining the SAT gender gap.

To explore the self-selection channel, we formally model the choice of taking the SAT. In our model, motivation decreases the costs of taking the test while being unrelated to test scores, and cognitive ability increases the expected test score while also lowering the test-taking costs. Under these assumptions, students with higher cognitive ability or higher motivation are more likely to take the test, but cognitive skills among test-takers are lower for students with higher motivation. Moreover, the expected test score's unobserved component is also lower for test takers with higher motivation. Applying these results to differences across genders, the model predicts that females are more likely to take the SAT; that there are no gender differences in expected SAT scores in the population; and that among test takers, females have lower (expected and actual) SAT scores than males due to negative selection on both observed and unobserved test-taking skills.

We confront these predictions with the data by jointly estimating a selection equation and an equation linking SAT scores to attributes—an application of the Heckman (1979) correction method; and by subsequently performing an Oaxaca-Blinder decomposition taking into account differential selection on both observed and unobserved test-taking ability. Our previous conclusions regarding the importance of cognitive skills for SAT scores hold after correcting for self-selection.

Accounting for selection on unobserved test-taking ability helps us explain 70 percent of the SAT gender gap, significantly improving upon the 40 percent from the uncorrected decomposition. Females' negative selection on unobserved ability appears to be twice as important as that on observed ability in explaining the gender gap among test-takers. Observed attributes become relatively less important since a subset of motivational measures—which favor females—are more strongly associated with SAT scores after correcting for selection. This also explains why, in the population, females have higher predicted SAT scores than males. A pure emphasis on cognitive skills would have implied the absence of a gender gap in the population.

Taken together, our results show that understanding the origins of gender gaps in non-cognitive traits will be key to informing the debate about why boys under-perform relative to girls in school. Our findings also have broader policy implications. They suggest that abolishing the SAT in favor of the HSGPA as the sole admission criterion to university programs would cause the fraction of females among college students to increase, and that admitted students would have higher motivation but lower cognitive skills. In light of such effects, an obvious question is which of the two evaluation instruments is more suited to identify college and career preparedness (Mattern, Burrus, Camara, O'Connor, Hansen, Gambrell, Casillas, and Bobek, 2014). We document that the HSGPA is a stronger predictor of graduation—as well as of earnings—than SAT scores, which implies a student pool with greater preparedness using the former instrument.

These findings have clear parallels to the case of the GED testing program in the US.[6] The GED captures cognitive skills, but fails to capture important non-cognitive skills that are valued in the labor market and predict college graduation. Controlling for cognitive ability, GED recipients appear to earn less, have lower hourly wages, and obtain lower levels of schooling than other high-school dropouts (Heckman, Humphries, and Kautz, 2014; Heckman and Rubinstein, 2001; Cameron and Heckman, 1993). Thus, standardized achievement tests as sorting instruments, in particular if subject to self-selection, appear to favor individuals with higher cognitive skills but lower non-cognitive skills. A growing literature highlights the increasing importance of *non*-cognitive skills in the labor market (Deming, 2017; Edin, Fredriksson, Nybom, and Öckert, 2018). If human capital investments made in college are complementary to non-cognitive skills, this might suggest that standardized achievement tests are becoming less appropriate as a tool for college admissions.

Finally, we note that even after controlling for individual differences in attributes and skills—and also after correcting for selection in the case of the SAT—there are sizeable gender gaps in performance: of 0.14 and 0.24 standard deviations (in favor of females) in CSGPA and HSGPA, respectively, and of 0.11 standard deviations (in favor of males) in SAT scores. Possible explanations for these (in the context of our study) 'unexplained' portions of the gender gaps may be found in related strands of literature. First, there is evidence of women performing worse than men, on average, in multiple-choice formats compared to in free-response exams.[7] In general, evidence from the lab and the field suggests a gender gradient in performance in competitive environments, which has potential implications for gender gaps in SAT scores or other high-stakes achievement tests.[8] In a second strand of literature, teacher-student interactions—such as teacher gender effects due to role models or teacher discrimination—are studied as explanations for the gender gap in grades and GPAs. The evidence from this literature is, however, inconclusive on the empirical relevance of teacher (gender) effects for gender gaps in school performance.[9]

This paper is organized as follows. Sections 2 and 3 describe our institutional setting and our data, respectively. Section 4 documents the variation in gender gaps across the various assessment forms. Section 5 investigates whether different assessment forms capture different attributes, and whether this can account for the variation in gender gaps. Section 6 presents a model of self-selection into taking the SAT, tests the model's predictions in the data, and uses the model to quantify the importance of self-selection for the SAT gender gap. Section 7 concludes the paper with a discussion of policy implications.

[6] The GED is a second-chance program that administers cognitive tests to self-selected high school dropouts.

[7] See for instance Bolger and Kellaghan (1990). However, the male advantage in multiple-choice tests seems more prevalent when wrong answers are penalized with negative points as women exhibit a lower likelihood of guessing relative to men in such tests, which may be attributed to differences in risk preferences (Pekkarinen, 2015; Akyol, Key, and Krishna, 2016; Baldiga, 2013). Wrong answers are not penalized on the Swedish SAT.

[8] Gneezy, Niederle, and Rustichini (2003) find that men's performance increases as the competitiveness of the test increases, while that of females does not. Similarly, Niederle and Vesterlund (2007) find that males show a stronger preference for competitive tasks than females. In the context of education, results presented by Jurajda and Münich (2011) suggest that men perform better than women in entrance exams for more prestigious schools, but not in the exams for less competitive schools. Similarly, Ors, Palomino, and Peyrache (2013) find that females tend to perform worse in more competitive examinations with high future payoffs than do men.

[9] For instance, Holmlund and Sund (2008), Puhani (2018), and Lindahl (2016) find no evidence in support of the hypothesis that a same-sex teacher improves student outcomes, while Dee (2005, 2007) and Falch and Naper (2013) suggest that students benefit from having a same-sex teacher. Hinnerich, Höglin, and Johannesson (2011) find no evidence of discrimination using blind- and non-blind grading of the same exam, while Lavy (2008), Terrier (2016), and Berg, Palmgren, and Tyrefors (2019) find that boys face discrimination in teacher grading.

2 Setting

The Swedish education system consists of nine years of compulsory schooling, followed by three years of (voluntary) high school, the completion of which is required for university eligibility.[10] For oversubscribed high school programs, slots are allocated among applicants based on compulsory school GPA (CSGPA). Similarly, slots to oversubscribed university programs are allocated based on high school GPA (HSGPA) and SAT scores through a centralized process. In the case of oversubscribed programs, Swedish universities are legally required to fill at least one third of slots based on a GPA ranking, and at least one third of slots based on a SAT ranking.

The SAT is voluntary, but nevertheless a high stakes test given the admission process. For instance, all medical programs in Sweden typically require the top score on the SAT or a HSGPA exceeding the mean by more than two standard deviations. Aside from medical programs, the top score on the SAT is a sufficient condition for admission to nearly all university programs in Sweden. In general, a higher SAT expands the set of programs that a student has access to, especially for students with a relatively low HSGPA (Graetz, Öckert, and Skans, 2018).

The assessment formats producing CSGPA, HSGPA, and SAT scores differ along several dimensions, as summarized in Table 1. Apart from natural differences in purpose, participation requirements, and timing, it is noteworthy that both CSGPA and HSGPA are based on more than a dozen separate written and oral assessments occurring during periods of several years. The SAT score, in contrast, is determined in a single one-day exam involving 120-150 multiple-choice questions.[11] In addition to written exams, school grades can be based on several other test formats. For the majority of cohorts studied in this paper, the grading system is a criterion-referenced grade scale. According to the Swedish National Agency for Education (Skolverket), the grades should reflect the students' acquired skills and knowledge based on a holistic assessment of written examinations, lab reports, in-class discussions, oral presentations etc. Thus, teachers have considerable discretion in setting questions for tests underlying the GPAs as well as in setting the course grades that go towards the final GPA. The SAT is instead a centralized test administered by the Swedish Council for Higher Education, with the same questions faced by all students on a given test date. Unlike in the case of GPAs, grading of the SAT is done blindly and graders do not have any discretion, given the multiple-choice format. There is no negative marking in the SAT: each correct answer is rewarded with one point, and wrong answers yield zero points. All assessment formats that we consider have in common that they test for knowledge in various areas. GPAs are based on subjects tests including math, Swedish, and English. The SAT has two parts, one testing for language, and one testing for numerical skills.[12]

[10] All students follow the same curriculum in compulsory school, while there is a range of high school programs, both vocational and academic. The vocational tracks include academic subjects (such as mathematics, English, and Swedish) granting access to some university programs.

[11] Students may repeat the SAT, however, and only the best score counts in admission (the test takes place twice each year).

[12] The Swedish SAT is designed based on the American SAT (SOU, 2004), but differs from the latter in that it does not contain an essay component.

3 Data

Our data come from population-wide individual-level administrative registers. The data include year and country of birth, gender, parents' country of birth and educational attainment, grade point averages (GPAs) from compulsory school and high school, as well as SAT scores. For the purpose of documenting gender gaps in test scores, we focus on the compulsory school graduation cohorts of 1993-2012, corresponding to birth cohorts 1977-1996 as students typically graduate compulsory school at age 16. For high school and SAT, we focus on the years 1996-2015, corresponding to the same birth cohorts given the typical high school graduation age of 19, and given that most students take the SAT around the time of high school graduation.[13] Among the students who have ever taken the SAT, about half have taken the test more than once (Graetz, Öckert, and Skans, 2018). For repeaters, we use the results from the first test throughout. We have checked that our results are robust to using the highest life-time score instead.

Our main sample, however, consists of individuals born in 1992, for whom we have data on cognitive skills in grade 6 from the Evaluation Through Follow-up (UGU) study [14] conducted by the Department of Education at Göteborg University. The cognitive tests measure inductive (number sequences), spatial (plate folding), and verbal (synonyms and opposites) skills. We use these detailed measures as well as a composite index obtained by principal component analysis. In addition to cognitive tests, the UGU administers a comprehensive questionnaire in grade 9 to the same students who took the cognitive tests at age 13, to elicit their motivation and time spent on homework. Using principal component analysis we create three measures of motivation: a general one that captures students' motivation to work towards getting admitted to a high-quality university program, achieving higher pay, becoming a productive member in society, etc.; a school-specific one capturing students' interest and motivation to learn in school; and a composite index combining the two. Our measure of effort exerted is the time spent on homework that the students report.[15]

The UGU data cover a 10-percent stratified random sample of students born in 1992, which corresponds to some 10,000 individuals.[16] Due to non-response in the survey, our final sample consists of roughly 4,300 individuals. The UGU data further include sampling weights to allow nationally representative statistics. We adjust these weights to make the final sample representative in terms of gender, immigrant status, and compulsory school GPA decile. The cognitive and non-cognitive measures from the UGU survey are standardized to have zero mean and unit variance within the final estimation sample. GPAs and SAT scores are standardized within each cohort.

[13] In our final sample, the maximum time between graduating from compulsory school and taking the SAT is seven years, and most individuals take the SAT within four years from compulsory school graduation.

[14] In Swedish: Utvärdering Genom Uppföljning (UGU).

[15] Similar measures of cognitive skills, as well as non-cognitive (psychological) ability, are available for a large fraction of Swedish males born between 1955-1985, as these cohorts were subject to military conscription and underwent extensive enlistment examinations (Lindqvist and Vestman, 2011). These data include only a small number of female volunteers, and thus are less useful in our context as we seek to explain gender gaps in test scores. However, in results available on request, we document that GPAs correlate strongly and positively with a student's father's cognitive as well as non-cognitive skills as measured at the enlistment exams. In contrast, and consistent with our findings from the UGU data, SAT scores only correlate with cognitive skills, not non-cognitive skills.

[16] The UGU study performs a two-stage stratified cluster sampling, where municipalities are drawn at random in the first stage, and catchment areas within municipalities in the second stage. All students in the relevant cohort of the included catchment areas are covered in the sample. The UGU data contain 10 percent random samples of nine cohorts born between 1948 and 1998. While the cognitive tests are identical across the samples, the survey questions do not overlap. Therefore, we focus here on one cohort for which we have relevant data on traits, and for which we also have data on GPAs and SAT scores.

4 Documenting gender gaps across assessment forms

Figure 1 reports standardized test scores and grades over time, by gender.[17] Females typically outperform males on both compulsory school GPA (CSGPA) and high school GPA (HSGPA) by about a third of a standard deviation (average gaps of 0.34 and 0.37 for CSGPA and HSGPA, respectively). But the reverse is true for the SAT, where females *under*-perform by a third of a standard deviation (an average gap of -0.32). These gaps are largely stable over time, with the exception of the narrowing of the HSGPA gap in 2011 and the widening of the SAT gap also in 2011.[18] The gaps are present across the score distributions, in the sense that across deciles of CSGPA and HSGPA the fraction of females (and hence the chance that a female student scores in a given decile) increases nearly monotonically, but decreases monotonically across deciles of the SAT; and the average SAT score of males is higher than that of females at all deciles of CSGPA and HSGPA (see Figures A1 and A2). It is also worth noting that despite the flipping gender gaps, the individual-level correlations between GPAs and SAT scores are quite similar across the genders (CSGPA, HSGPA, and SAT scores are all strongly positively correlated, although the correlation between HSGPA and SAT scores is somewhat weaker than that between CSGPA and SAT scores—see Table A2).

Gender gaps flip also within subject areas between the CSGPA and SAT score. In CS mathematics, girls outperform boys by 0.09, while in the quantitative part of the SAT, females under-perform by 0.53.[19] On a standardized score averaging CS grades in Swedish and English, girls outperform boys by 0.42, while in the verbal part of the SAT, females under-perform by 0.14. Between CS grades and SAT scores, the decreases in the gender gaps are very similar across overall (-0.66), mathematical content (-0.62), and language content (-0.56). In other words, the genders' relative advantages are largely constant across tests: Subtracting females' overall score from their average subject scores, we obtain -0.26 for CS math and -0.21 for the quantitative part of the SAT; and we obtain 0.08 for CS Swedish and English, and 0.19 for the verbal part of the SAT. These results are shown in Figure A3. There is some convergence in the genders' relative advantages over time. Gender gaps also flip, both overall and within subject areas, between CSGPA and SAT score when using a matched sample of individuals: with observations on CSGPA; non-missing HSGPA; and who took the SAT 1–6 years after graduating compulsory school. These results are shown in Figures A4–A5.

Finally, there is also a sizeable gender gap in within-individual score differences. For each student in the UGU sample who took the SAT at least once, we calculate the difference between their first SAT score and their CSGPA, and between the SAT score and the HSGPA (see Table A3).[20] This difference is 0.5 less on average for females than males, in both the SAT-CSGPA and SAT-HSGPA comparisons. Put differently, a female student's position in the score distribution deteriorates by half a standard deviation, relative to males, when moving from school grades to the SAT.

In the following sections, we test for two potential explanations for the flipping gender gap across GPAs and SAT scores. First, we investigate whether GPAs reflect different individual attributes than SAT

[17] The means plotted in the figure, along with their standard errors, are listed in Table A1.

[18] Although these shifts, which both favor males, coincide in timing, they were likely caused by two separate changes: a grading reform affecting the HSGPA gap, and an expansion of the quantitative section of the SAT. It is beyond the scope of this paper to investigate this issue.

[19] All figures mentioned in the text are in units of standard deviations, unless noted otherwise.

[20] We take the SAT score from the first test, but the results are largely unchanged when taking the best SAT score instead. The results are also largely unchanged when using all cohorts.

scores, and if so, whether gender differences in the endowments of these attributes are large enough to account for meaningful portions of the gaps. Second, we ask whether gender differences in attributes among SAT takers arise due to non-random selection into taking the test, and calculate what the gender gap in SAT scores would be in the population, if all individuals took the test.

5 What attributes do GPAs and SAT scores capture?

Panel A of Figure 2 plots standardized indices of cognitive skills and motivation measured at age 13 and 16, respectively, against deciles of various scholastic assessments at older ages for our UGU sample, which is representative of the 1992 birth cohort. Both sets of attributes are strongly positively related to CSGPA and HSGPA: There is a two standard deviation difference in cognitive skills, and a one standard deviation difference in motivation, between the bottom and top deciles of the CSGPA. For the HSGPA, the differences between bottom and top are 1.5 for cognitive skills, and 0.5 for motivation. SAT scores are even more strongly informative about cognitive skills, with a greater than two standard deviation gap between bottom and top deciles. However, motivation shows no correlation with SAT scores whatsoever. (The fact that motivation was measured three years prior to taking the SAT is not a likely explanation for this zero-correlation, since motivation does have predictive power for the HSGPA, and most students take the SAT around the time of high school graduation.) It is also worth noting that CSGPA and HSGPA reflect the sub-components of cognitive skills (as well as motivation and effort) in a relatively uniform way. In contrast, SAT scores reflect verbal skills to a greater extent than inductive and spatial skills (see Figure A7).

In Panel B of Figure 2, we show that the relationships between attributes and scores generalize when using a different data source for measuring skills, namely cognitive and non-cognitive tests administered at the military enlistment. Since the enlistment mainly covers the male population, we use the father's test scores as measures of own skills, for the cohorts of students born 1990–1996 (for whom there is high coverage of fathers' scores).[21] The cognitive test consists of four parts: synonyms, inductions, plate folding, and technical comprehension. The non-cognitive measure is based on behavioral questions capturing psychological energy (e.g. focus), intensity, and emotional stability (stress tolerance). Thus, while the cognitive tests are highly similar across the UGU and the military enlistment data, the non-cognitive traits are not likely to capture the same traits. Importantly, however, both types of measures from the Swedish enlistment data are predictive of labor earnings and employment (see e.g. Lindqvist and Vestman, 2011; Fredriksson, Hensvik, and Skans, 2018).

There is a one standard deviation difference in (fathers') cognitive *and* non-cognitive skills between the bottom and top decile of (own) CSGPA, and roughly half a standard deviation difference in both types of skills between the bottom and top deciles of HSPGA. There is a larger than one standard deviation difference in cognitive skills between the bottom and top deciles of SAT scores, while the corresponding difference in non-cognitive skills is less than half a standard deviation.

The overall patterns in Figure 2 thus suggest that school-level assessments test for different attributes

[21] Using Swedish enlistment data, Grönqvist, Öckert, and Vlachos (2017) show that the intergenerational transmission between fathers and sons is 0.32–0.35 for cognitive skills and 0.2 for non-cognitive sills. The authors also use cognitive scores from the UGU data in an IV strategy to correct for measurement error in fathers' abilities, by which the intergenerational correlation increases to 0.42 for cognitive and non-cognitive abilities.

than the SAT. In particular, GPAs are highly informative of students' motivation and non-cognitive skills, while the SAT scores are not.[22] Table 2 shows that females exhibit higher motivation than males (and they also report spending more time on homework). The gender difference in cognitive skills are less systematic in the representative sample, although they seem to favor males in the sub-sample of SAT takers. Taken together, this suggests that differences in the attributes that are tested for, together with gender differences in endowments, may go some way towards explaining the flipping gender gap. We investigate this issue next.

5.1 Gender differences in the endowments of attributes

We first regress GPAs and SAT scores on only a female dummy, which shows that the gender gaps are very similar in this sample of 1992-born students to the gaps we see in the population and over time in Figure 1 (see columns (1), (5), and (9) of Table 3). Next, we add four components of cognitive skill on the right-hand side. In terms of standardized coefficients, inductive skills are most predictive of both CSGPA and HSGPA, with the other three components spatial, synonyms, and verbal opposites skills being less than half as important, although they still have predictive power (columns (2) and (6) of Table 3). Given the lack of systematic gender differences in these attributes, adding cognitive skills does not affect the gender gaps in GPAs.

Motivation and time spent on homework are both highly predictive of CSGPA and HSGPA (columns (3) and (7) of Table 3. This, together with the fact that females exhibit higher motivation and effort (time on homework), implies that these measures account for a large portion of the gender gaps: over 60 percent in the case of CSGPA, and one third in the case of HSGPA.

A summary index of cognitive skills suggests no substantial differences in their importance across GPAs and SAT scores (Figure 2). However, there are such differences when it comes to the sub-components. Similar to the unconditional relationships discussed above, the most predictive for SAT scores among them are synonyms skills and verbal opposites skills, followed by inductive skills; spatial skills are relatively unimportant. But among SAT takers in this sample, spatial skills is the only component of cognitive abilities where females outperform males (Table 2). Taken together, the four components of cognitive skills account for more than 40 percent of the gender gap in SAT scores (column (10) of Table 3). However, general motivation and effort are slightly negatively correlated with SAT scores, while school-specific motivation is slightly positively correlated. Overall, adding these variables has no effect on the gender gap in SAT scores (column (11) of Table 3).

The conclusions that motivation and effort help explain the gender gaps in GPAs, and cognitive skills help explain the gender gap in SAT scores do not change when we enter these sets of variables jointly, as shown in columns (4), (8), and (12) of Table 3.

Table 4 reports the conditional relationships between father's skill measures from the military enlistment and own grades, and SAT scores. Columns (1), (6), and (11) show that the raw gender gaps in grades and test scores are quantitatively very similar in the sample of students born 1990–1996 with non-missing observations on their father's skill to those in the full population of students born 1977–1996. In columns (2) and (7) we show that cognitive and non-cognitive scores are highly predictive of CSGPA and HSGPA, respectively, also when entered jointly, although the non-cognitive skill measure

[22] The relationships between attributes and scores shown in panel A of Figure 2 do not vary by gender, see Figure A6.

is less predictive than the motivational measures from the UGU data. Columns (3) and (9) show that differences in attributes explain substantial portions of the gender gaps in grades also in the sample of UGU students with non-missing information on the father's test scores. However, since fathers' skills are balanced across men and women in the population, they do not contribute to explaining the gender gaps. With respect to the SAT, father's non-cognitive skills are completely uninformative about test scores once controlling for father's cognitive skills (columns (12) and (13)). Taken together, the results in Table 4 strengthens our conclusion that school level assessments capture both cognitive- and non-cognitive skills, while standardized achievement tests are mainly informative about cognitive skills.

We also explore to what extent subject-level gender gaps can be explained by gender differences in (own) cognitive skills, motivation, and effort. Differences in motivation and effort account for all of the (modest) female advantage in CS math, and half of the female advantage in languages. Gender differences in cognitive skills among SAT takers account for 20 percent of the male advantage in the quantitative part, and for nearly all the male advantage in the verbal part. As with overall scores, cognitive skills do not help explain gender differences in CS subject grades, while motivation and effort do not help explain gender gaps in the two parts of the SAT (see Table A4).

Finally, we find that the gender gap in within-individual score differences is accounted for in part by cognitive skills, motivation, and effort. These variables are capable of explaining one third of the gender gap in the CSGPA-SAT comparison, and one fifth of the gap in the HSGPA-SAT comparison (Table A3).

Table 2 shows that females are 8 percentage points more likely to take the SAT. Moreover, the distribution of cognitive skills among test takers differs from that in the full sample, with male test takers exhibiting relatively higher skills; in contrast, the gender gap in measures of motivation is nearly identical among test takers to that in the full sample (Table 2). Thus, non-random selection into taking the SAT may potentially account for parts of the SAT gender gap. In the next section, we set up a model of self-selection into taking the SAT and explore its implications for the gender gap in test scores.

6 The implications of self-selection into taking the SAT

The previous section has shown that differential self-selection on observable cognitive skills partly accounts for the female disadvantage in SAT scores. Here we explore the selection channel in more depth. Section 6.1 formally models the choice of taking the SAT, derives testable predictions, and describes how to take the model to the data. Section 6.2 then takes the model's predictions to the data, and uses the model to quantify the importance of observed and unobserved selection for the SAT gender gap.

Before introducing the model, we present additional motivating evidence about SAT participation. Recall that females are 8 percentage points more likely than males to take the SAT (Table 2). What might account for this difference? To answer this question, we regress an indicator for ever having taken the test on cognitive skills and our measures of motivation. Column (2) in Table 5 shows that cognitive skills are indeed positive predictors of taking the test. But as these traits are balanced across genders in the population, they do not explain the gender gap in test-taking. Column (3) shows that measures of motivation are also positive predictors of participation. As discussed above, and as shown in Table 2, females score higher on these motivational measures. In fact, the gender gap in participation drops to 2 percentage points when controlling for motivation, though it increases slightly to 3pp when also

controlling for cognitive skills (column (4)).[23] In sum, gender differences in motivation account for the majority of the gender gap in participation.

6.1 Modeling the choice of taking the SAT

Suppose that student i's SAT score is given by $z_i = f^z(\mathbf{x}_i) + \varepsilon_i^z$ and her cost of taking the test by $c_i = f^c(\mathbf{x}_i) + \varepsilon_i^c$, where $\mathbf{x}_i$ is a vector of characteristics observable to the econometrician; ε_i^z and ε_i^c are aspects of test-taking and studying ability that are unobserved to the econometrician, assumed to be independent of the observable characteristics (though they may be correlated with each other); and the functions f^z and f^c are yet to be specified. The student observes ε_i^c. In contrast, ε_i^z is only realized if and when she takes the test. However, the student's expectation of ε_i^z will be informed by ε_i^c if the two noise terms are correlated. We assume that $\varepsilon_i^z = -\kappa\varepsilon_i^c + u_i$, where $\kappa > 0$ and u_i is i.i.d. noise, so that $\mathbb{E}[\varepsilon_i^z|\varepsilon_i^c] = -\kappa\varepsilon_i^c$. Realistically, we thus assume that a student with greater unobserved study ability (a lower ε_i^c) has on average a higher unobserved component in the test score.

Assuming risk neutrality, the student's expected utility from taking the test can be written as

$$U_i = \alpha f^z(\mathbf{x}_i) - f^c(\mathbf{x}_i) - (1+\alpha\kappa)\varepsilon_i^c,$$

where α is some positive constant. Normalizing the outside option to zero, the student takes the test if and only if expected utility from taking the test is non-negative. Denoting test-taking by $D_i \in \{0,1\}$ and the CDF of ε_i^c by $G(\cdot)$, we have

$$\mathbb{P}(D_i = 1) = G\left(\frac{\alpha f^z(\mathbf{x}_i) - f^c(\mathbf{x}_i)}{1+\alpha\kappa}\right).$$

Now, for concreteness, assume that $\mathbf{x}_i = (s_i, m_i)$, where s_i stands for cognitive ability and m_i is a measure of motivation. Here, we assume for simplicity that each of these are scalars. Further assume that $\partial f^z/\partial s > 0$, $\partial f^z/\partial m = 0$, $\partial f^c/\partial s < 0$, $\partial f^c/\partial m < 0$: cognitive ability increases the expected score and decreases test-taking costs, and motivation decreases costs while having no effect on the expected score. Under these assumptions—which are of course inspired by our above findings—we have that students with higher cognitive ability, or higher motivation, are more likely to take the test,

$$\frac{\partial\mathbb{P}(D_i=1)}{\partial s_i} > 0, \ \frac{\partial\mathbb{P}(D_i=1)}{\partial m_i} > 0.$$

Moreover, expected observed cognitive skills among test-takers, conditional on motivation, are lower for students with higher observed motivation,[24]

$$\frac{\partial\mathbb{E}[s_i|D_i=1,m_i]}{\partial m_i} < 0.$$

[23] Figure A8 indicates that the relationships between participation and composite indices of motivation and cognitive skills do not vary much by gender. We have also re-estimated the regressions reported in Table 5 interacting allowing the slopes of all right-hand side variables to differ by gender. The interaction terms were never statistically significant.

[24] Observe that $\mathbb{E}[s_i|D_i=1,m_i] = \int(\mathbb{E}[s_i|(1+\alpha\kappa)\varepsilon_i^c \le \alpha f^z(s_i,m_i) - f^c(s_i,m_i), m_i, \varepsilon_i^c])dG(\varepsilon_i^c)$. Since $f^z(s_i,m_i) - f^c(s_i,m_i)$ monotonically increases in s_i and m_i, we have that for a given ε_i^c, a higher m_i allows for s_i to be lower without violating the inequality. Thus, s_i is lower in expectation.

And finally, expected unobserved test-taking ability among test-takers, conditional on motivation, is lower for students with higher observed motivation, both conditionally and unconditionally on cognitive skills,[25]

$$\frac{\partial \mathbb{E}[\varepsilon_i^z | D_i = 1, s_i, m_i]}{\partial m_i} < 0, \ \frac{\partial \mathbb{E}[\varepsilon_i^z | D_i = 1, m_i]}{\partial m_i} < 0.$$

Applying these results to differences across the genders, under additional assumptions motivated by our empirical setting, we obtain the following:

Theoretical predictions. Suppose that in the population, there are no gender differences in cognitive ability, but that females have higher motivation than males on average. It follows that

1. Females are more likely to take the SAT.
2. In the population, there are no gender differences in the expected SAT score.
3. Among test takers, females are expected to have lower SAT scores than males. This female disadvantage is due to negative selection both on observed skills and on unobserved factors.

To take the choice model to the data, we assume linearity, $f^z(\mathbf{x}_i) = \mathbf{x}_i'\beta^z$ and $f^c(\mathbf{x}_i, \varepsilon_i^c) = \mathbf{x}_i'\beta^c$. The vector of observable characteristics is

$$\mathbf{x}_i' = (\ [\text{constant}],\ [\text{inductive skills}]_i,\ [\text{spatial skills}]_i,\ [\text{synonyms skills}]_i,\ [\text{verbal opposite skills}]_i,$$
$$[\text{motivation (general)}]_i,\ [\text{motivation (school)}]_i,\ [\text{time spent on homework}]_i\ \ ).$$

As we discuss below, some elements of β^z and β^c may be zero; and the vector of observables may in addition include a female dummy, the GPAs, and father's skills.

We further assume normality of the unobserved components. Let σ_c denote the standard deviation of ε_i^c.[26] Using a well-known property of the truncated normal distribution, the conditional expectation function of the SAT score—that is, conditioning on observed characteristics, and on taking the test—can now be written as

$$\begin{aligned}
\mathbb{E}[z_i | \mathbf{x}_i, D_i = 1] &= \mathbf{x}_i'\beta^z + \mathbb{E}[\varepsilon_i^z | \mathbf{x}_i, D_i = 1] \\
&= \mathbf{x}_i'\beta^z - \kappa \mathbb{E}\left[\varepsilon_i^c | \mathbf{x}_i, \varepsilon_i^c \leq \tfrac{\mathbf{x}_i'(\alpha\beta^z - \beta^c)}{1+\alpha\kappa}\right] \\
&= \mathbf{x}_i'\beta^z - \kappa\sigma_c \lambda\left(\tfrac{\mathbf{x}_i'(\alpha\beta^z - \beta^c)}{\sigma_c(1+\alpha\kappa)}\right),
\end{aligned}$$

where $\lambda(t) = \varphi(t)/(1-\Phi(t))$ is the hazard function, or inverse Mills ratio, of the standard normal distribution. Defining $\lambda_i \equiv \lambda\left(\frac{\mathbf{x}_i'(\alpha\beta^z - \beta^c)}{1+\alpha\kappa}\right)$ and $\delta \equiv -\kappa\sigma_c$, we can thus write

$$\mathbb{E}[z_i | \mathbf{x}_i, D_i = 1] = \mathbf{x}_i'\beta^z + \delta\lambda_i. \tag{1}$$

[25] Observe that $\mathbb{E}[\varepsilon_i^z | D_i = 1, s_i, m_i] = -\kappa\mathbb{E}[\varepsilon_i^c | (1+\alpha\kappa)\varepsilon_i^c \leq \alpha f^z(s_i, m_i) - f^c(s_i, m_i), s_i, m_i]$. Since $f^z(s_i, m_i) - f^c(s_i, m_i)$ monotonically increases in s_i and m_i, a higher m_i allows for ε_i^c to be larger without the inequality being violated, and hence ε_i^z is lower in expectation. To obtain the result when not conditioning on cognitive skills, one only needs to take expectations over s_i.

[26] For completeness, one may write σ_u for the standard deviation of u_i. The variance of ε_i^z is then $\kappa^2\sigma_c^2 + \sigma_u^2$.

Equation (1) is an application of the Heckman (1979) correction method, demonstrating that selection into taking the SAT must be controlled for by including λ_i in the regression, or else β^z cannot be consistently estimated. One can obtain λ_i by estimating a probit model for taking the test, noting that we now have

$$\mathbb{P}(D_i = 1|\mathbf{x}_i) = \Phi\left(\frac{\mathbf{x}_i'(\alpha\beta^z - \beta^c)}{\sigma_c(1+\alpha\kappa)}\right). \tag{2}$$

The results from estimating this probit model are reported in column (5) of Table 5. The estimated coefficients (marginal effects at the mean) are nearly identical to the ones obtained from the linear probability model reported in column (4).

Moreover, equation (1) lets us take our theoretical prediction regarding unobserved test-taking ability to the data. Analogously to the above discussion, suppose that $\beta^z_{\text{motivation (gen.)}} = 0$, that $\beta^c_{\text{motivation (gen.)}} < 0$, and that females have higher general motivation. As the hazard function is increasing, a higher λ_i corresponds to a greater probability of taking the test. Thus, females would have a higher λ_i even if all other observable characteristics besides general motivation were balanced across the genders.[27] Therefore, they will be negatively selected on unobserved ability. This negative impact is captured by the coefficient $\delta = -\kappa\sigma_c < 0$, and it can readily be quantified as part of an otherwise standard Oaxaca-Blinder (OB) decomposition. From equation (1) we obtain

$$\begin{aligned}
&\underbrace{\mathbb{E}_F[z_i|D_i=1] - \mathbb{E}_M[z_i|D_i=1]}_{\text{SAT gender gap}} \quad = \\
&\underbrace{\left(\mathbb{E}_F[\mathbf{x}_i|D_i=1] - \mathbb{E}_M[\mathbf{x}_i|D_i=1]\right)'\beta^z_M + \left(\mathbb{E}_F[\lambda_i|D_i=1] - \mathbb{E}_M[\lambda_i|D_i=1]\right)'\delta_M}_{\text{explained component}} \\
&+\underbrace{\mathbb{E}_F[\mathbf{x}_i|D_i=1]'(\beta^z_F - \beta^z_M) + \mathbb{E}_F[\lambda_i|D_i=1](\delta_F - \delta_M)}_{\text{unexplained component}},
\end{aligned} \tag{3}$$

with F and M indicating female and male, respectively. In this selection-corrected OB decomposition (Fortin, Lemieux, and Firpo, 2011), the explained component not only contains the usual difference in observed characteristics, but also a term due to differential selection on unobserved ability.

In practice, we estimate equations (1) and (2) jointly via maximum likelihood, using Stata's `heckman` command, both on the full sample and separately by gender. In our preferred specification, we identify the model by excluding general motivation from (1), imposing $\beta^z_{\text{motivation (gen.)}} = 0$. Alternatively, we relax this assumption and identify the model off our functional form assumptions. Both approaches yield very similar results, and the latter approach suggests that $\beta^z_{\text{motivation (gen.)}} = 0$ is indeed a reasonable assumption. Estimation of the model yields estimates of λ_i, which we then use for the decomposition (3). We also obtain estimates of $\mathbb{P}(D_i = 1)$, and predicted values of z_i for both test-takers and non-takers.

To sum up, the Heckman correction method fulfils the following purposes in our context. First, it lets us check whether our finding that GPAs and SAT scores reflect different attributes is not an artefact of

[27] As we discuss below, cognitive skills are balanced across genders in the population, though females have higher school-specific motivation and report spending more time on homework.

self-selection into taking the test. Second, it helps us assess to what extent self-selection on observed and unobserved characteristics can account for the SAT gender gap. Third, it yields predicted SAT scores for the full sample, allowing us to estimate the gender gap in the population. Finally, we are able to confront our theoretical predictions to the data, namely that as a result of higher motivation (which is not reflected in SAT scores) females are more likely to take the test, are negatively selected, and should nevertheless not have a disadvantage in terms of predicted SAT scores in the population.

6.1.1 Alternative modeling assumptions

Before presenting the results from implementing our selection model, we further discuss some of our assumptions, and the consequences of relaxing them.

First, one could argue that both the compulsory school GPA and the high school GPA should enter the model. The GPAs likely contain additional information about test-taking propensity and ability. And the HSGPA arguably affects the utility that students derive from the SAT score. For instance, being at the top of the HSGPA distribution means having access to all university programs regardless of one's SAT score. One may thus hypothesize that females are less likely to take the SAT because they have higher HSGPAs. Nevertheless, we omit the GPAs from our preferred specification for three reasons. First, our goal is to explain the variation in the *unconditional* gender gap across GPAs and SAT scores. The gender gap in SAT scores conditional on GPAs is a function of the gender gap in GPAs, and will be harder to explain than the unconditional SAT gender gap as long as we do not succeed in fully explaining the gender gap in GPAs. Second, the GPAs are not predetermined with respect to the cognitive skills and motivational variables, which are measured prior to graduation from compulsory school. In a regression of SAT scores, the coefficient on say spatial skills conditional on the CSGPA and HSGPA is thus hard to interpret. Third, the HSGPA is a positive predictor of SAT participation in our data, even though a rather weak one when the CSGPA is also controlled for—likely because the HSGPA captures both positive and negative influences on test-taking propensity. Females are in fact more likely to take the SAT despite having higher HSGPAs, contradicting the above-mentioned hypothesis. We will nonetheless discuss results from controlling for the GPAs, and report them in the appendix.

Another concern is that our model does not allow for study effort to be chosen optimally. A way of addressing this is to include a proxy of study effort among the vector of observables. Indeed, our measures of school-specific motivation and time spent on homework arguably serve this function. Furthermore, a model allowing for optimal effort yields qualitatively similar theoretical predictions.[28]

[28]Denoting effort by e_i, suppose that $z_i = f^z(e_i, \mathbf{x}_i) + \varepsilon_i^z$ and $c_i = f^c(e_i, \mathbf{x}_i) + \varepsilon_i^c$. Further assume that $\partial f^z/\partial e > 0$, $\partial^2 f^z/(\partial e)^2 < 0$, $\partial f^c/\partial e > 0$, and $\partial^2 f^c/(\partial e)^2 > 0$, so the optimization problem is well defined. Via backward induction, we solve for the optimal effort level conditional on taking the test, which is pinned down by the firs-order condition

$$\alpha \frac{\partial}{\partial e} f^z(e_i, \mathbf{x}_i) = \frac{\partial}{\partial e} f^c(e_i, \mathbf{x}_i).$$

This yields optimal effort as a function of observable characteristics, $e_i^* = e(\mathbf{x}_i)$. The student takes the test if and only if expected utility given optimal effort is strictly positive, $\alpha f^z(e(\mathbf{x}_i), \mathbf{x}_i) + \alpha \mathbb{E}[\varepsilon_i^z | \varepsilon_i^c] > f^c(e(\mathbf{x}_i), \mathbf{x}_i) + \varepsilon_i^c$.

Our results regarding negative selection due to higher motivation are still valid under endogenous effort: The conditional expectations discussed above are affected by a marginal change in motivation only directly, but not through effort, since they involve the objective function and hence the envelope theorem applies. However, when considering the expected SAT score, indirect effects do matter, as the offsetting indirect effect from the cost function is absent. Thus, the main difference relative to the case without effort choice is that a variable that affects the cost function but not the expected score directly, will now affect the expected score indirectly through its effect on effort. This means for instance that in the population, females now have a

6.2 Self-selection: Results

In Table 6, we report results from the Heckman correction described in Section 6.1. For convenience, we show OLS results, partly replicating Table 3, in columns (1)-(3). Column (4) presents results from the Heckman correction when general motivation is excluded from the second-stage equation. The coefficients on the various cognitive traits are up to 50 percent larger when attempting to correct for selection. But the general pattern that spatial skills appear less important relative to inductive and verbal skills, remains unchanged. Both school-specific motivation and self-reported study time now gain in importance relative to the OLS results. Finally, the female disadvantage has shrunk to 13 percent of a standard deviation. Before investigating the sources of this improved explanatory power, we note that the results are virtually unchanged when we allow general motivation to enter the second stage, and thus identify the model from the functional form. This specification also delivers a selection-corrected estimate of the importance of general motivation. The coefficient is close to zero—as in the raw data and in the OLS regressions, there appears to be no association between SAT scores and general motivation.

To better understand why the Heckman method allows us to explain more of the gender gap in SAT scores than OLS, we carry out the selection-corrected OB decomposition discussed in Section 6.1. For reference, column (1) in Table 7 shows results from the standard decomposition based on OLS, which can explain 0.11 of the original 0.29-gap, as seen in panel A. Panel B shows that almost all of this is due to gender differences in cognitive skills among test takers. The selection-corrected decomposition, based on estimation of the choice model separately by gender, yields an explained component of 0.21, shown in column (2) of panel A. Two thirds of this are due selection on unobserved test-taking ability—the inverse Mills ratio contributes 0.14 (panel B). That implies a contribution of observable traits of 0.07, which is lower than in the uncorrected decomposition. The reason is that correcting for selection increases the coefficients on school-specific motivation and study time relatively more than those on the cognitive skills, shifting the explained component somewhat in favor of females.

We can use the various regression models to predict the SAT score for the entire sample, not just the test-takers. The simple OLS regressions from Table 3, together with the descriptive statistics in Table 2, suggest that the gender gap should be zero in the population, as cognitive traits are balanced and motivational measures are unimportant. However, as just noted, the selection-corrected coefficients imply a greater importance of school-specific motivation and study time. In fact, the population gender gap in the predicted SAT score, based on the Heckman model, is 0.12 percent of a standard deviation in favor of females, as shown in column (3) of Table 7. Two-thirds of this gap are explained by females' greater school-specific motivation and study time.[29] As an additional illustration, we regress the predicted SAT score on a female dummy, but re-weighting the regression by the test-taking probabilities estimated as part of the Heckman correction. The coefficient in the re-weighted regression is -0.08, the same as the unexplained component from the selection-corrected OB decomposition. And when conditioning the sample on having taken the test, the gender gap in the predicted score is -0.09 (Table A6).

higher expected SAT score due to having greater general motivation; and among test takers, females may still have lower scores but the difference will be smaller than in the case without effort choice.

Estimation of the model would of course be more challenging under optimal effort, since linearity could not be justified except as an approximation.

[29] The predicted score is calculated from applying the Heckman correction separately by gender, so it contains an unexplained component that is due to differences in the intercept and slope coefficients.

Thus, we have illustrated and quantified the forces highlighted by the theoretical model in Section 6.1. General motivation predicts test-taking but does appear to be rewarded by better SAT scores, a conclusion that survives our attempts at selection-correction. Given females' greater general motivation, we expect them to be negatively selected on both observed and unobserved test-taking ability. Indeed, we find this to be the case, and differences in unobserved ability appear to be twice as important as those in observed ability in explaining the gender gap among test-takers. Observed traits become relatively less important since a subset of motivational measures are more strongly associated with SAT scores after correcting for selection. This also explains why in the population, females actually have higher predicted SAT scores than males, whereas a pure emphasis of cognitive traits would have implied the absence of a gender gap in the population.

7 Discussion

In this paper we document a flipped gender gap between school-level assessments and standardized achievement test scores using Swedish longitudinal administrative data. Females outperform males on cumulative compulsory school GPA and on high school GPA, by about a third of a standard deviation in both cases. At the same time females under-perform by about a third of a standard deviation in the Swedish SAT. Our results suggest that differences in the endowments of non-cognitive traits—in particular, motivation and effort—account for a sizeable portion of the female advantage in school performance. In contrast, motivation has no predictive power for SAT scores. Turning to cognitive skills, we account for 40 percent of the male advantage in SAT scores by observing gender gaps in the endowments of inductive, spatial, and verbal skills among SAT takers. Moreover, women are more likely to take the SAT, which is largely driven by their higher motivation, implying that they are negatively selected on both observed and unobserved test-taking ability. Quantifying this mechanism using a selection model, we are able to explain 70 percent of the SAT gender gap. Taken together, our findings show that school-level assessments and standardized achievement tests capture different skills, and that differences in endowments of skills and selection effects go a long way towards explaining gender gaps in school- and test-performance.

Our findings have implications for the measurement of university preparedness and the design of university admission systems. Countries or institutions can choose between admissions being based on standardized test results, on school grades, or on a combination of the two. And in the latter case, they can choose between combining different scores into a composite measure or creating separate quotas for each measure. To inform the choice of which measures to emphasize, it is important to know which skills and traits they reflect, and hence what the potential distributional impacts of choosing different measures are.

How, then, would a change in admission criteria at Swedish universities affect the composition of students enrolled? This is a difficult question, because students may change how they allocate their effort across different assessments in response. If admissions were solely based on SAT scores, for instance, the distribution of gender, cognitive skills, and motivation conditional on SAT scores might look very different from what we find in this paper, because a much broader population would take the SAT. However, in the opposite scenario, where admissions are solely based on HSGPA, selection concerns are less severe, and our findings should be quite informative. This is because under the current system,

all students must graduate high school to be eligible for university.[30] We thus conclude, based on our results, that abolishing SAT scores as an admission criterion would lead to an increase in the fraction of females, an increase in motivation, and a decrease in cognitive skills (because HSGPA does not reflect cognitive skills as strongly as SAT scores), among students enrolled in university.

To get a sense of the magnitudes involved, consider a stylized scenario where competitive programs are filled by drawing randomly from the top two deciles of the score distributions. Start with the case where HSGPA and SAT are weighted equally, as is the spirit of the current system. Among admitted students, the fraction female would be 0.55, and the indices of motivation and cognitive skills would be 0.34 and 1.18 on average, respectively. If HSGPA were the sole criterion, the fraction female would increase to 0.68, motivation would increase to 0.49, and cognitive skills would decrease to 0.8. These figures are suggestive of large effects of the choice of admission criterion on the characteristics of admitted students, at least in competitive, oversubscribed programs.

How would college preparedness be affected if either HSGPA or SAT scores were more strongly emphasized in admissions? When we regress the probability of having graduated from university by age 30 on GPAs and SAT scores, we find that the HSGPA is a much stronger predictor of graduation than SAT scores, with the ratio of standardized coefficients exceeding five for both females and males. We also find that GPAs are much stronger predictors of earnings than SAT scores.[31] Based on this evidence, and on our finding that SAT scores are not informative about motivation and other non-cognitive skills, we tentatively conclude that putting more emphasis on HSGPA at the expense of SAT scores might lead to greater college preparedness among admitted students.

[30] One may still worry that some students allocate less effort towards achieving a good HSGPA if there is also the option of using SAT scores when applying to university. However, we have documented that the HSGPA reflects measures of cognitive skills, motivation, and effort in a very similar way to the CSGPA, where such strategic concerns could not play any role.

[31] Educational attainment and earnings are measured in 2014, and the sample includes the population of Swedish residents born between 1977-1984. These results are shown in Table A8.

References

AHN, T., P. ARCIDIACONO, A. HOPSON, AND J. THOMAS (2015): "Equilibrium Grade Inflation with Implications for Female Interest in STEM Majors," .

AKYOL, Ş. P., J. KEY, AND K. KRISHNA (2016): "Hit or Miss? Test Taking Behavior in Multiple Choice Exams," Discussion Paper 22401, National Bureau of Economic Research.

ALMLUND, M., A. L. DUCKWORTH, J. HECKMAN, AND T. KAUTZ (2011): "Personality psychology and economics," in *Handbook of the Economics of Education*, vol. 4, pp. 1–181. Elsevier.

BALDIGA, K. (2013): "Gender differences in willingness to guess," *Management Science*, 60(2), 434–448.

BERG, P., O. PALMGREN, AND B. TYREFORS (2019): "Gender Grading Bias in Junior High School Mathematics," Working Paper Series 1263, Research Institute of Industrial Economics.

BOLGER, N., AND T. KELLAGHAN (1990): "Method of measurement and gender differences in scholastic achievement," *Journal of Educational Measurement*, 27(2), 165–174.

BORGHANS, L., B. H. GOLSTEYN, J. J. HECKMAN, AND J. E. HUMPHRIES (2016): "What grades and achievement tests measure," *Proceedings of the National Academy of Sciences*, 113(47), 13354–13359.

CAMERON, S. V., AND J. J. HECKMAN (1993): "The nonequivalence of high school equivalents," *Journal of labor economics*, 11(1, Part 1), 1–47.

CARD, D., AND A. A. PAYNE (2017): "High school choices and the gender gap in STEM," Discussion paper, National Bureau of Economic Research.

CORNWELL, C., D. B. MUSTARD, AND J. VAN PARYS (2013): "Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school," *Journal of Human Resources*, 48(1), 236–264.

DEE, T. S. (2005): "A teacher like me: Does race, ethnicity, or gender matter?," *American Economic Review*, 95(2), 158–165.

——— (2007): "Teachers and the gender gaps in student achievement," *Journal of Human Resources*, 42(3), 528–554.

DEMING, D. J. (2017): "The Growing Importance of Social Skills in the Labor Market*," *The Quarterly Journal of Economics*, 132(4), 1593–1640.

DIAMOND, R., AND P. PERSSON (2016): "The long-term consequences of teacher discretion in grading of high-stakes tests," Discussion paper, National Bureau of Economic Research.

DUCKWORTH, A. L., AND M. E. SELIGMAN (2006): "Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores.," *Journal of educational psychology*, 98(1), 198.

DYNARSKI, S. (2017): "Make Everyone Take the SAT or ACT. And Make It Free," *The New York Times*.

EDIN, P.-A., P. FREDRIKSSON, M. NYBOM, AND B. ÖCKERT (2018): "The rising return to non-cognitive skill," .

EDWARDS, D., H. COATES, AND T. FRIEDMAN (2012): "A survey of international practice in university admissions testing," *Higher Education Management and Policy*, 24(1), 1–18.

FALCH, T., AND L. R. NAPER (2013): "Educational evaluation schemes and gender gaps in student achievement," *Economics of Education Review*, 36, 12–25.

FORTIN, N., T. LEMIEUX, AND S. FIRPO (2011): "Chapter 1 - Decomposition Methods in Economics," vol. 4 of *Handbook of Labor Economics*, pp. 1 – 102. Elsevier.

FORTIN, N. M., P. OREOPOULOS, AND S. PHIPPS (2015): "Leaving boys behind: Gender disparities in high academic achievement," *Journal of Human Resources*, 50(3), 549–579.

FREDRIKSSON, P., L. HENSVIK, AND O. N. SKANS (2018): "Mismatch of talent: Evidence on match quality, entry wages, and job mobility," *American Economic Review*, 108(11), 3303–38.

FRYER JR, R. G., AND S. D. LEVITT (2010): "An empirical analysis of the gender gap in mathematics," *American Economic Journal: Applied Economics*, 2(2), 210–40.

GEVREK, Z. E., C. NEUMEIER, AND D. GEVREK (2018): "Explaining the Gender Test Score Gap in Mathematics: The Role of Gender Inequality," Discussion Paper 11260, IZA Discussion Papers.

GNEEZY, U., M. NIEDERLE, AND A. RUSTICHINI (2003): "Performance in competitive environments: Gender differences," *The Quarterly Journal of Economics*, 118(3), 1049–1074.

GOLDIN, C., L. F. KATZ, AND I. KUZIEMKO (2006): "The homecoming of American college women: The reversal of the college gender gap," *Journal of Economic perspectives*, 20(4), 133–156.

GRAETZ, G., B. ÖCKERT, AND O. N. SKANS (2018): "College admission opportunities and educational outcomes," working paper.

GRÖNQVIST, E., B. ÖCKERT, AND J. VLACHOS (2017): "The intergenerational transmission of cognitive and noncognitive abilities," *Journal of Human Resources*, 52(4), 887–918.

GUISO, L., F. MONTE, P. SAPIENZA, AND L. ZINGALES (2008): "Culture, gender, and math," *Science*, 320(5880), 1164.

HECKMAN, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

HECKMAN, J. J., J. E. HUMPHRIES, AND T. KAUTZ (2014): *The myth of achievement tests: The GED and the role of character in American life*. University of Chicago Press.

HECKMAN, J. J., AND T. KAUTZ (2012): "Hard evidence on soft skills," *Labour economics*, 19(4), 451–464.

HECKMAN, J. J., AND Y. RUBINSTEIN (2001): "The importance of noncognitive skills: Lessons from the GED testing program," *American Economic Review*, 91(2), 145–149.

HINNERICH, B. T., E. HÖGLIN, AND M. JOHANNESSON (2011): "Are boys discriminated in Swedish high schools?," *Economics of Education review*, 30(4), 682–690.

HOLMLUND, H., AND K. SUND (2008): "Is the gender gap in school performance affected by the sex of the teacher?," *Labour Economics*, 15(1), 37–53.

JURAJDA, Š., AND D. MÜNICH (2011): "Gender gap in performance under competitive pressure: Admissions to Czech universities," *American Economic Review*, 101(3), 514–18.

KERSEY, A. J., E. J. BRAHAM, K. D. CSUMITTA, M. E. LIBERTUS, AND J. F. CANTLON (2018): "No intrinsic gender differences in childrens earliest numerical abilities," *npj Science of Learning*, 3(1), 12.

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.

LINDAHL, E. (2016): "Are teacher assessments biased?–Evidence from Sweden," *Education Economics*, 24(2), 224–238.

LINDQVIST, E., AND R. VESTMAN (2011): "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment," *American Economic Journal: Applied Economics*, 3(1), 101–128.

MACHIN, S., AND T. PEKKARINEN (2008): "Global sex differences in test score variability.," *Science*.

MATTERN, K., J. BURRUS, W. CAMARA, R. O'CONNOR, M. A. HANSEN, J. GAMBRELL, A. CASILLAS, AND B. BOBEK (2014): "Broadening the Definition of College and Career Readiness: A Holistic Approach," Research Report Series, ACT.

MURNANE, R. J. (2013): "US high school graduation rates: Patterns and explanations," *Journal of Economic Literature*, 51(2), 370–422.

NIEDERLE, M., AND L. VESTERLUND (2007): "Do women shy away from competition? Do men compete too much?," *The Quarterly Journal of Economics*, 122(3), 1067–1101.

NOFTLE, E. E., AND R. W. ROBINS (2007): "Personality predictors of academic outcomes: big five correlates of GPA and SAT scores.," *Journal of personality and social psychology*, 93(1), 116.

OECD (2017): *Education at a Glance 2017: OECD Indicators*. OECD Publishing, Paris.

ORS, E., F. PALOMINO, AND E. PEYRACHE (2013): "Performance gender gap: does competition matter?," *Journal of Labor Economics*, 31(3), 443–499.

PEKKARINEN, T. (2015): "Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations," *Journal of Economic Behavior & Organization*, 115, 94–110.

POPE, D. G., AND J. R. SYDNOR (2010): "Geographic variation in the gender differences in test scores," *Journal of Economic Perspectives*, 24(2), 95–108.

PUHANI, P. A. (2018): "Do boys benefit from male teachers in elementary school? Evidence from administrative panel data," *Labour Economics*, 51(C), 340–354.

ROTHSTEIN, J. M. (2004): "College performance predictions and the SAT," *Journal of Econometrics*, 121(1-2), 297–317.

SCB (2017): "Young people left behind? The situation on the labour market for those born in the 1990s without completed upper secondary education," Discussion paper, Statistics Sweden.

SOU (2004): *Tre vägar till den öppna högskolan*, no. SOU: 2004:29 in Statens Offentliga Utredningar. Government of Sweden.

TERRIER, C. (2016): "Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement," IZA Discussion Papers 10343, Institute for the Study of Labor (IZA).

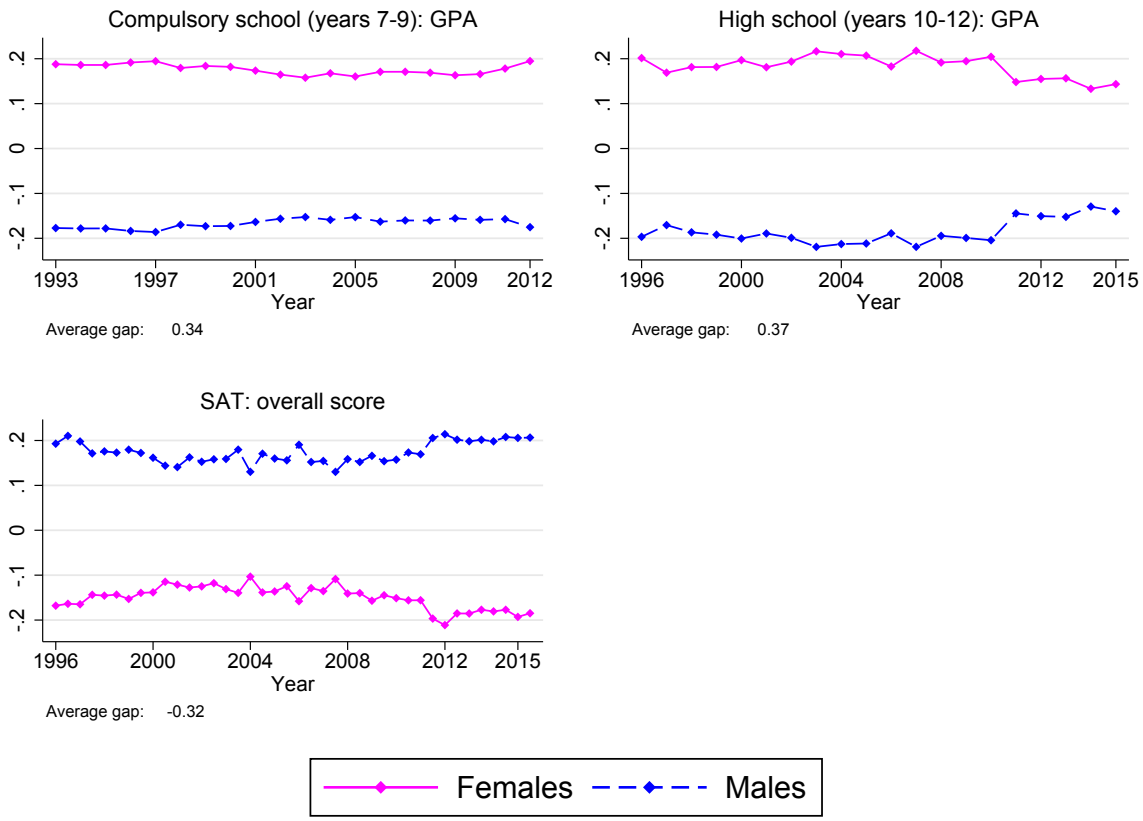
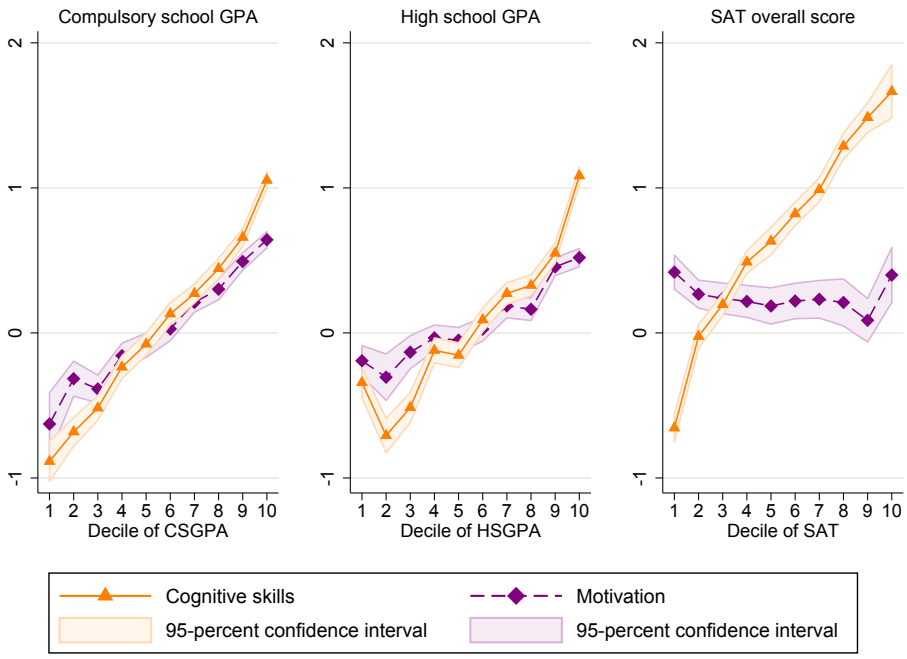
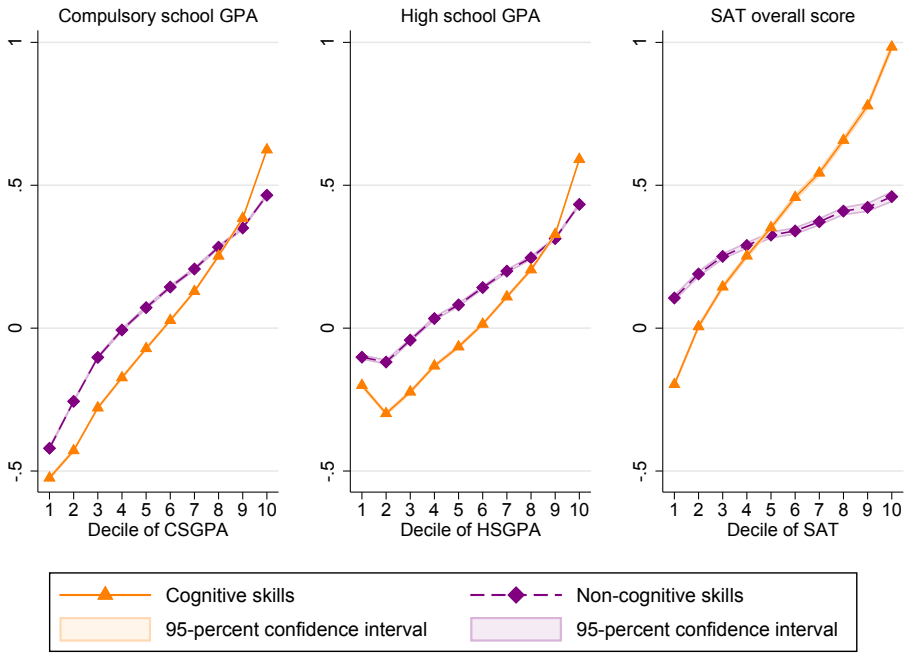


Figure 1: Standardized scores over time, by assessment form and gender



(a) Own cognitive skills and motivation



(b) Father's cognitive and non-cognitive skills

Figure 2: Test scores, cognitive skills, motivation, and father's skills

Table 1: Overview of assessments

	CSGPA	HSGPA	SAT
Purpose	Progression to HS	Progression to university	
Participation	Compulsory	Voluntary	Voluntary
Timing	Years 7-9	Years 10-12	Usually year 12+
Number of tests	15+	15+	1, may repeat
Format	Written & oral	Written & oral	Multiple choice
Content	Mixed	Mixed	Mixed
Teacher discretion in choosing content	✓	✓	✗
Blind grading	✗	✗	✓
Teacher discretion in grading	✓	✓	✗

Table 2: Descriptive statistics

	All				SAT-takers			
	Female	Male	Diff.	*t*-stat	Female	Male	Diff.	*t*-stat
Cognitive skills, comp. index	0.024	-0.023	0.047	1.54	0.35	0.46	-0.11	-2.63
Inductive skills	-0.046	0.044	-0.089	-2.95	0.23	0.51	-0.28	-6.67
Spatial skills	0.13	-0.12	0.25	8.47	0.34	0.14	0.19	4.50
Synonyms skills	-0.068	0.065	-0.13	-4.40	0.21	0.44	-0.23	-5.38
Verbal opposites skills	0.058	-0.055	0.11	3.72	0.31	0.34	-0.030	-0.67
Motivation, comp. index	0.17	-0.16	0.34	11.2	0.41	0.076	0.34	8.42
Motivation (general)	0.15	-0.14	0.30	9.85	0.38	0.072	0.31	7.62
Motivation (school)	0.23	-0.22	0.45	15.4	0.44	0.064	0.38	9.80
Time spent on homework	0.21	-0.20	0.41	13.8	0.47	0.011	0.46	9.94
Mother graduated university	0.22	0.23	-0.0069	-0.55	0.29	0.34	-0.046	-2.17
Father graduated university	0.16	0.15	0.0024	0.22	0.22	0.25	-0.025	-1.31
Took SAT	0.41	0.33	0.082	5.64	1	1	0	
Observations	4,351				1,940			

Table 3: Gender gaps, cognitive skills, and motivation

	Compulsory school GPA				High school GPA				SAT			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Female	0.32	0.32	0.12	0.14	0.33	0.33	0.22	0.24	-0.29	-0.17	-0.29	-0.19
	(0.045)	(0.038)	(0.042)	(0.035)	(0.044)	(0.039)	(0.044)	(0.040)	(0.051)	(0.033)	(0.055)	(0.035)
Inductive skills		0.29		0.25		0.20		0.18		0.21		0.20
		(0.023)		(0.020)		(0.024)		(0.023)		(0.021)		(0.021)
Spatial skills		0.11		0.12		0.064		0.074		0.072		0.071
		(0.022)		(0.019)		(0.023)		(0.022)		(0.019)		(0.019)
Synonyms skills		0.13		0.12		0.088		0.083		0.29		0.29
		(0.033)		(0.028)		(0.032)		(0.031)		(0.027)		(0.027)
Verbal opposites skills		0.14		0.15		0.12		0.13		0.26		0.25
		(0.030)		(0.027)		(0.030)		(0.029)		(0.023)		(0.024)
Motivation (general)			0.21	0.24			0.12	0.14			-0.098	-0.010
			(0.029)	(0.021)			(0.029)	(0.025)			(0.038)	(0.020)
Motivation (school)			0.10	0.027			0.058	0.0050			0.17	0.065
			(0.028)	(0.023)			(0.029)	(0.025)			(0.033)	(0.023)
Time spent on homework			0.22	0.21			0.12	0.12			-0.072	-0.018
			(0.020)	(0.016)			(0.021)	(0.019)			(0.022)	(0.015)
R-squared	0.03	0.33	0.22	0.49	0.03	0.17	0.08	0.21	0.03	0.53	0.06	0.53
Observations			4,351				4,114				1,940	

Notes: The dependent variables are standardized test scores as indicated in the column headings. All right-hand side variables, except the female dummy, are standardized. Regressions are weighted using sampling weights. Robust standard errors in parentheses.

Table 4: Gender gaps, cognitive skills, motivation, and father's skills

	Compulsory school GPA					High school GPA					SAT				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Female	0.34	0.34	0.32	0.17	0.18	0.32	0.33	0.28	0.21	0.22	-0.34	-0.30	-0.33	-0.20	-0.19
	(0.0025)	(0.0023)	(0.043)	(0.037)	(0.037)	(0.0027)	(0.0026)	(0.041)	(0.042)	(0.042)	(0.0041)	(0.0039)	(0.050)	(0.045)	(0.044)
Father's cognitive skills		0.26	0.28		0.075		0.18	0.19		0.074		0.31	0.29		0.092
		(0.0013)	(0.023)		(0.020)		(0.0015)	(0.023)		(0.024)		(0.0022)	(0.028)		(0.022)
Fathers's non-cognitive skills		0.14	0.13		0.079		0.086	0.089		0.057		-0.014	-0.023		-0.0043
		(0.0013)	(0.025)		(0.019)		(0.0015)	(0.022)		(0.020)		(0.0022)	(0.027)		(0.019)
Inductive skills				0.22	0.20				0.19	0.17				0.18	0.17
				(0.018)	(0.018)				(0.024)	(0.025)				(0.024)	(0.024)
Spatial skills				0.10	0.096				0.034	0.029				0.091	0.085
				(0.020)	(0.020)				(0.023)	(0.023)				(0.023)	(0.023)
Synonyms skills				0.13	0.12				0.075	0.065				0.30	0.29
				(0.034)	(0.033)				(0.029)	(0.029)				(0.035)	(0.035)
Verbal opposites skills				0.18	0.17				0.11	0.096				0.27	0.26
				(0.027)	(0.028)				(0.029)	(0.029)				(0.028)	(0.028)
Motivation (general)				0.23	0.24				0.13	0.13				-0.033	-0.032
				(0.022)	(0.021)				(0.026)	(0.026)				(0.024)	(0.023)
Motivation (school)				0.020	0.014				0.025	0.018				0.058	0.053
				(0.023)	(0.023)				(0.024)	(0.024)				(0.026)	(0.026)
Time spent on homework				0.19	0.18				0.11	0.10				-0.015	-0.022
				(0.017)	(0.016)				(0.021)	(0.021)				(0.018)	(0.018)
R-squared	0.033	0.16	0.16	0.50	0.52	0.029	0.082	0.077	0.19	0.20	0.037	0.14	0.13	0.51	0.52
Observations	548510	548510	3199	3199	3199	483543	483543	3027	3027	3027	185812	185812	1410	1410	1410

Notes: The dependent variables are standardized test scores as indicated in the column headings. Results in columns (1)–(2), (6)–(7), and (11)–(12) are based on the full sample of students born 1990–1996 with non-missing observations on father's skills; columns (2)–(5), (8)–(10), and (13)–(15) use the representative sample of 1992-born students from the UGU database with non-missing observations on fathers' skills. All right-hand side variables, except the female dummy, are standardized. Regressions are weighted using sampling weights. Robust standard errors in parentheses.

Table 5: Gender gaps in SAT participation

	(1)	(2)	(3)	(4)	(5)
Female	0.082	0.086	0.019	0.030	0.032
	(0.017)	(0.017)	(0.017)	(0.017)	(0.018)
Inductive skills		0.088		0.076	0.077
		(0.011)		(0.010)	(0.010)
Spatial skills		0.018		0.022	0.021
		(0.0090)		(0.0087)	(0.0088)
Synonyms skills		0.040		0.038	0.039
		(0.012)		(0.011)	(0.012)
Verbal opposites skills		0.042		0.042	0.039
		(0.011)		(0.011)	(0.011)
Motivation (general)			0.025	0.032	0.035
			(0.011)	(0.010)	(0.011)
Motivation (school)			0.064	0.043	0.045
			(0.0099)	(0.0093)	(0.0099)
Time spent on homework			0.065	0.060	0.058
			(0.0089)	(0.0084)	(0.0078)
R-squared	0.01	0.11	0.07	0.15	

Notes: The dependent variable is an indicator for ever having taken the SAT. Coefficients shown in columns (1)-(4) are estimated by OLS, and those shown in column (5) are marginal effects, evaluated at sample means, estimated by probit. The number of observations is 4,351. All right-hand side variables, except the female dummy, are standardized. Regressions are weighted using sampling weights. Robust standard errors in parentheses.

Table 6: Selection-corrected regressions of SAT scores

	(1)	(2)	(3)	(4)	(5)
Female	-0.29	-0.19	-0.19	-0.13	-0.13
	(0.051)	(0.035)	(0.035)	(0.040)	(0.040)
Inductive skills		0.20	0.20	0.31	0.31
		(0.021)	(0.021)	(0.027)	(0.027)
Spatial skills		0.071	0.071	0.097	0.097
		(0.019)	(0.019)	(0.023)	(0.023)
Verbal opposites skills		0.25	0.25	0.30	0.30
		(0.024)	(0.024)	(0.028)	(0.028)
Synonyms skills		0.29	0.29	0.34	0.34
		(0.027)	(0.027)	(0.030)	(0.030)
Motivation (general)		-0.010			0.039
		(0.020)			(0.026)
Motivation (school)		0.065	0.060	0.14	0.12
		(0.023)	(0.020)	(0.026)	(0.027)
Time spent on homework		-0.018	-0.021	0.073	0.064
		(0.015)	(0.015)	(0.021)	(0.021)
Heckman correction				✓	✓
R-squared	0.03	0.53	0.53		
Observations: test-takers		1,940		1,940	
Observations: full sample				4,351	

Notes: The dependent variable is the standardized SAT score. Results from OLS regresions are shown, adjusted for selection as indicated. Columns (1) and (2) replicate columns (9) and (12) from Table 3. Columns (4) and (5) report results from the Heckman correction as described in Section 6.1. In column (4), identification is achieved through excluding general motivation from the 2nd stage equation, whereas in column (5), the model is identified off the functional form. All right-hand side variables, except the female dummy, are standardized. Regressions are weighted using sampling weights. Robust standard errors in parentheses.

Table 7: Oaxaca-Blinder decompositions of the SAT gender gap

	Actual score		Predicted score
	(1)	(2)	(3)
A: Overall decomposition			
Gender gap	-0.29	-0.29	0.12
	(0.051)	(0.051)	(0.037)
Explained	-0.11	-0.21	0.079
	(0.045)	(0.057)	(0.039)
Unexplained	-0.17	-0.081	0.041
	(0.037)	(0.049)	(0.0044)
B: Detailed decomposition of explained component			
Inductive skills	-0.065	-0.13	-0.034
	(0.017)	(0.039)	(0.015)
Spatial skills	0.0083	0.017	0.015
	(0.0053)	(0.0075)	(0.0024)
Verbal opposites skills	-0.0084	-0.013	0.042
	(0.015)	(0.023)	(0.014)
Synonyms skills	-0.062	-0.082	-0.043
	(0.019)	(0.024)	(0.014)
Motivation (school)	0.025	0.092	0.074
	(0.011)	(0.030)	(0.0064)
Time spent on homework	-0.012	0.053	0.026
	(0.011)	(0.026)	(0.0023)
Inverse Mills ratio		-0.14	
		(0.059)	
Observations	1,940		4,351

Notes: Results are reported from the Oaxaca-Blinder decomposition of the SAT gender gap given in equation (3). The inverse Mills ratio and the predicted SAT score are obtained from estimating the Heckman selection model separately for males and females, where general motivation is excluded from the 2nd-stage equation. Robust standard errors in parentheses.

Appendix figures and tables

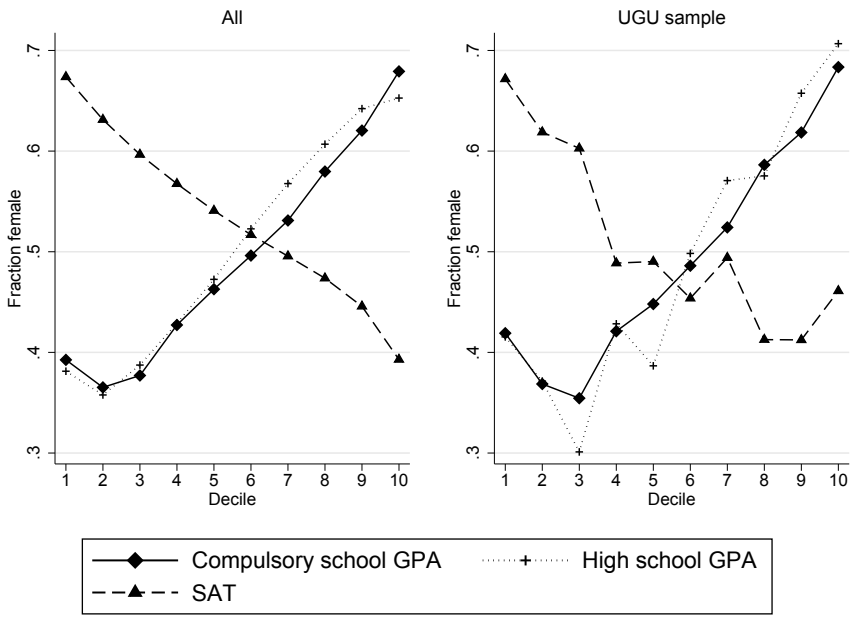


Figure A1: Fraction female by score decile, across assessment forms

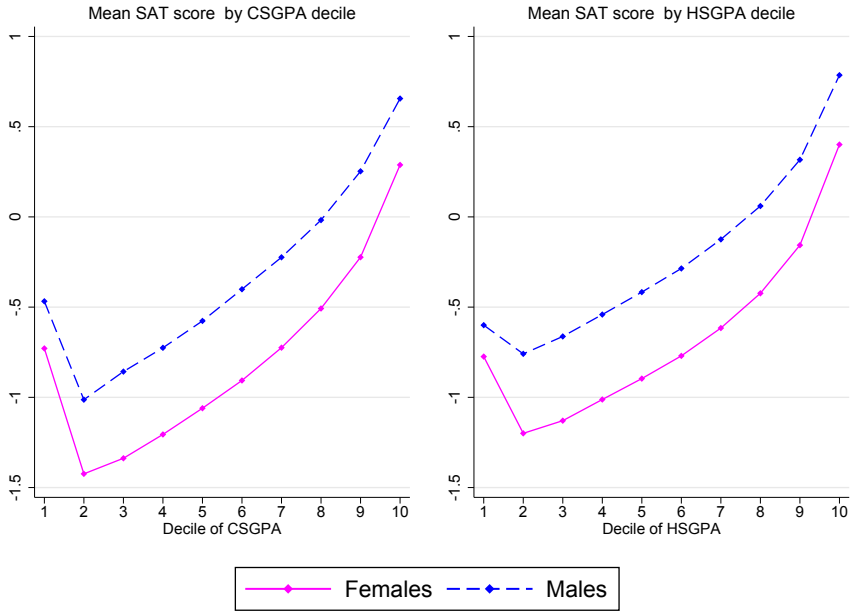
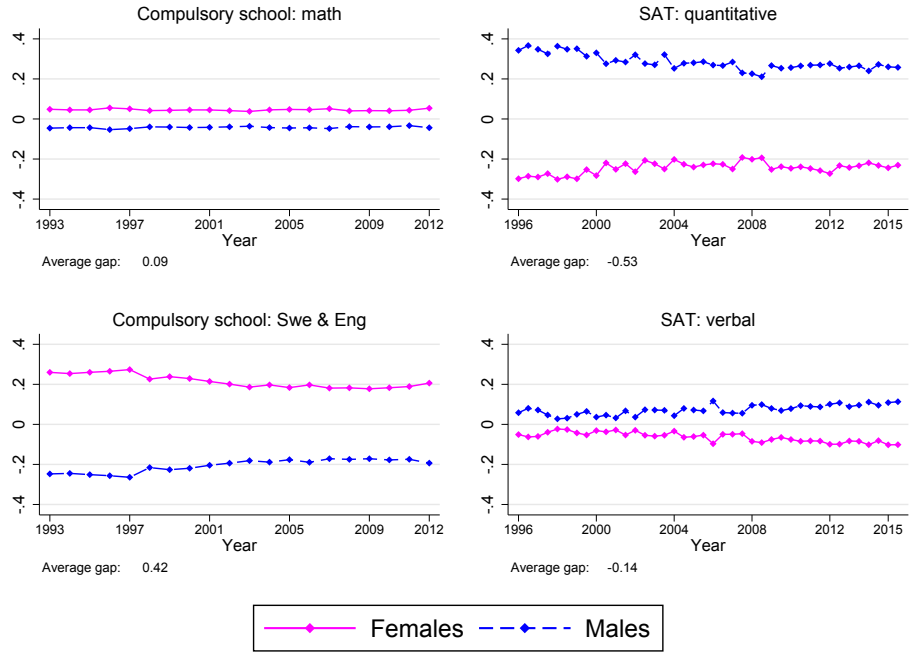
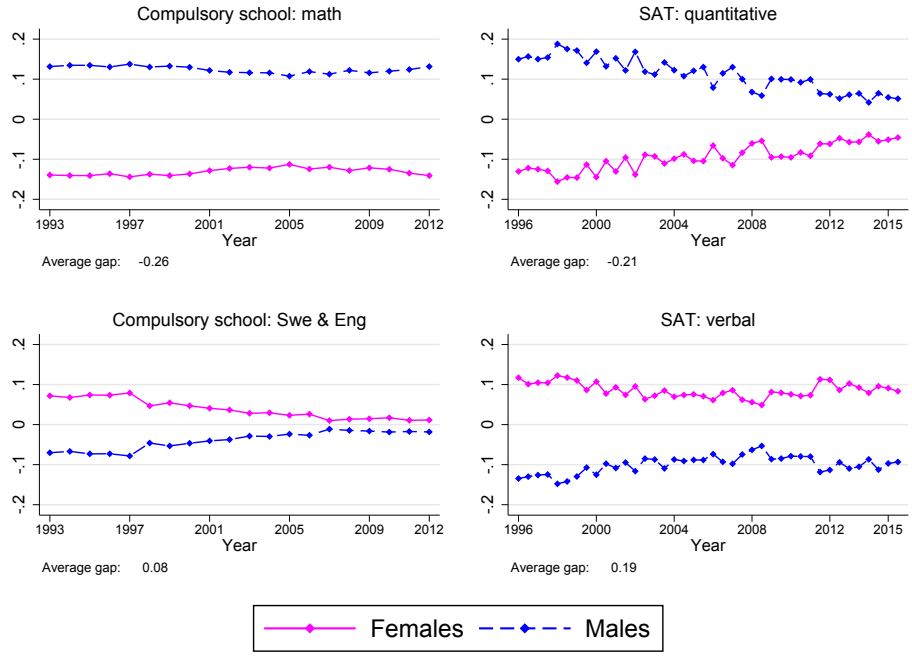


Figure A2: Average SAT score by gender and decile of GPA



(a) Standardized subject scores



(b) Relative standardized subject scores

Notes: Relative standardized subject scores are defined as the difference between the within-group mean subject score and the within-group mean overall score.

Figure A3: Standardized subject scores over time, by assessment form and gender

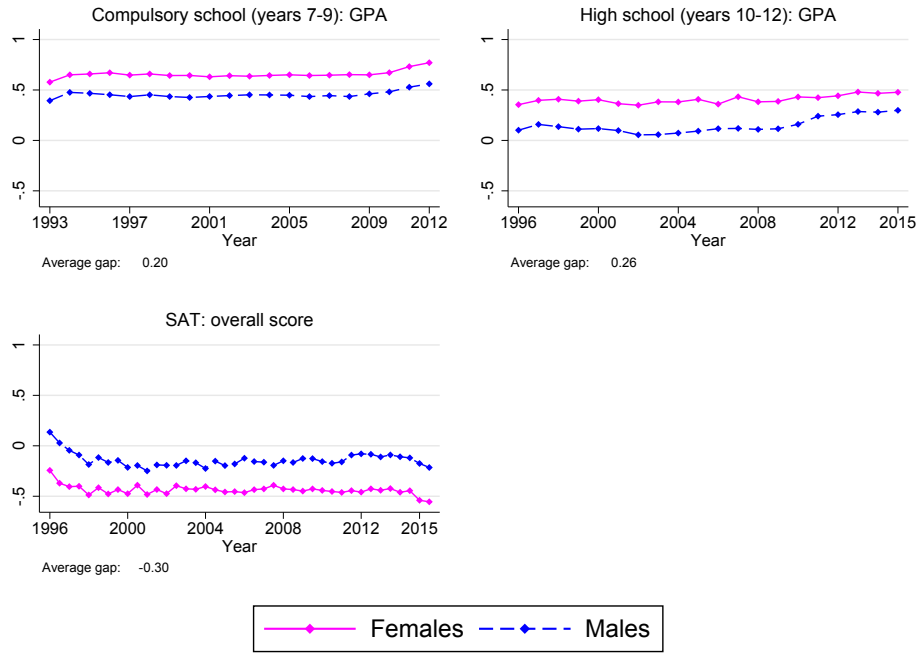
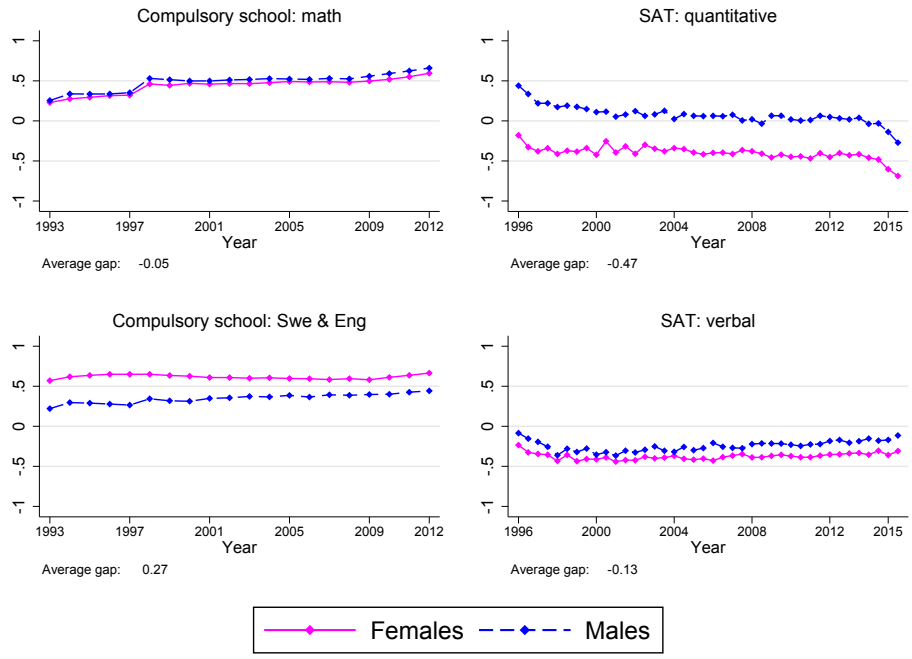
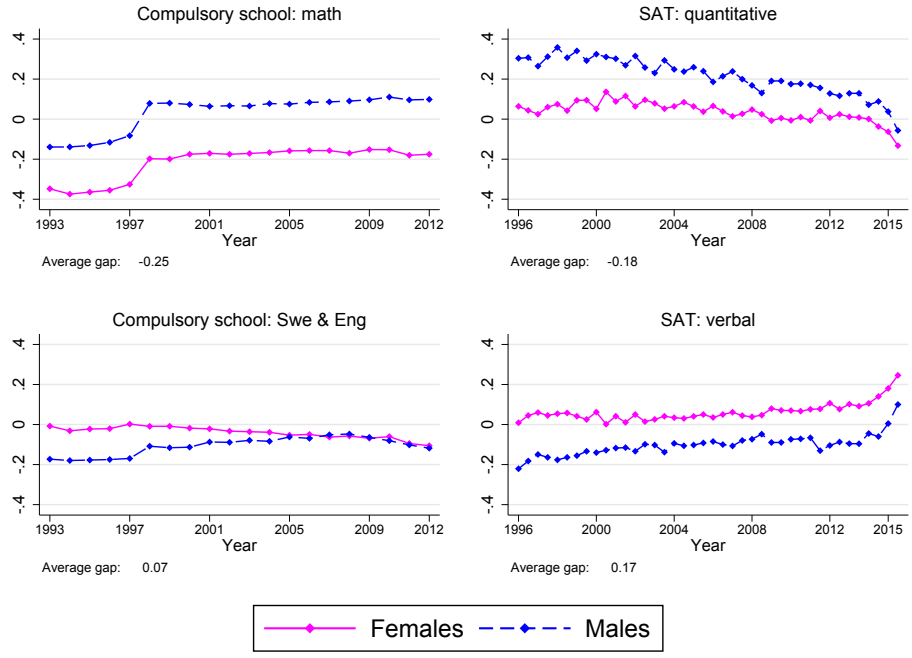


Figure A4: Standardized scores over time, by assessment form and gender—matched GPA-SAT sample



(a) Standardized subject scores



(b) Relative standardized subject scores

Notes: Relative standardized subject scores are defined as the difference between the within-group mean subject score and the within-group mean overall score.

Figure A5: Standardized subject scores over time, by assessment form and gender—matched GPA-SAT sample

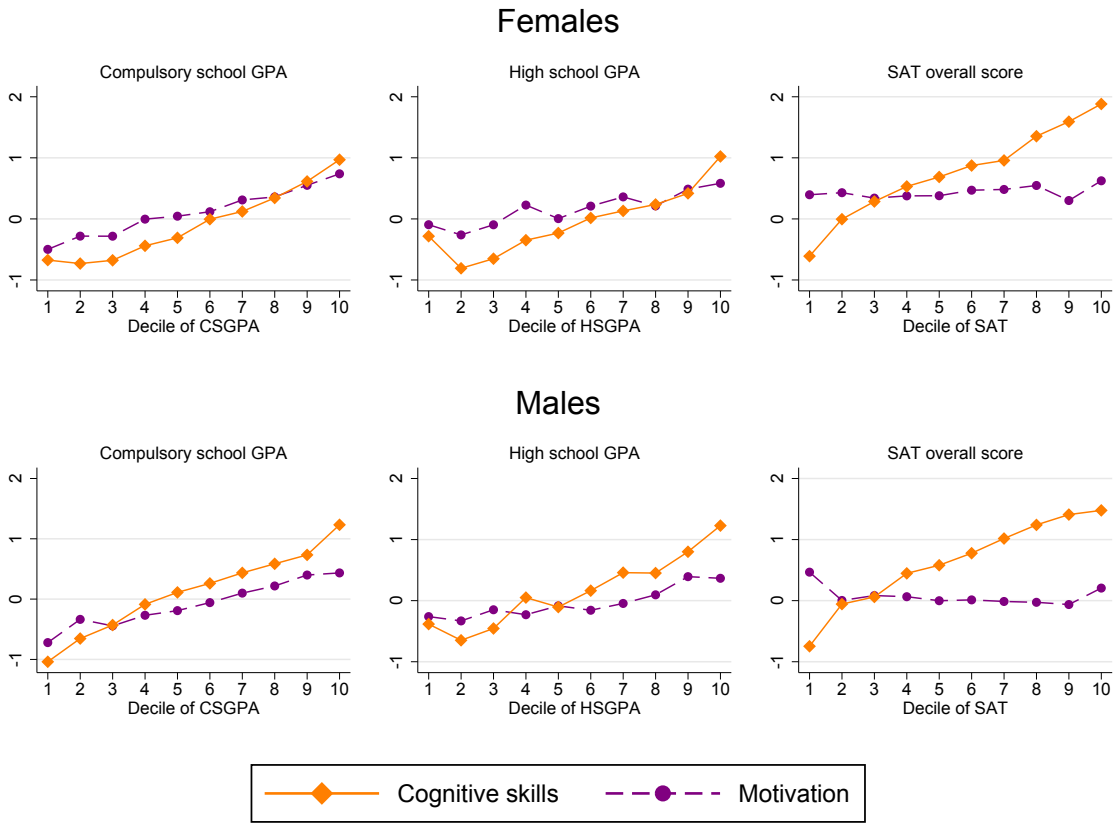


Figure A6: Test scores, cognitive skills, and motivation by gender

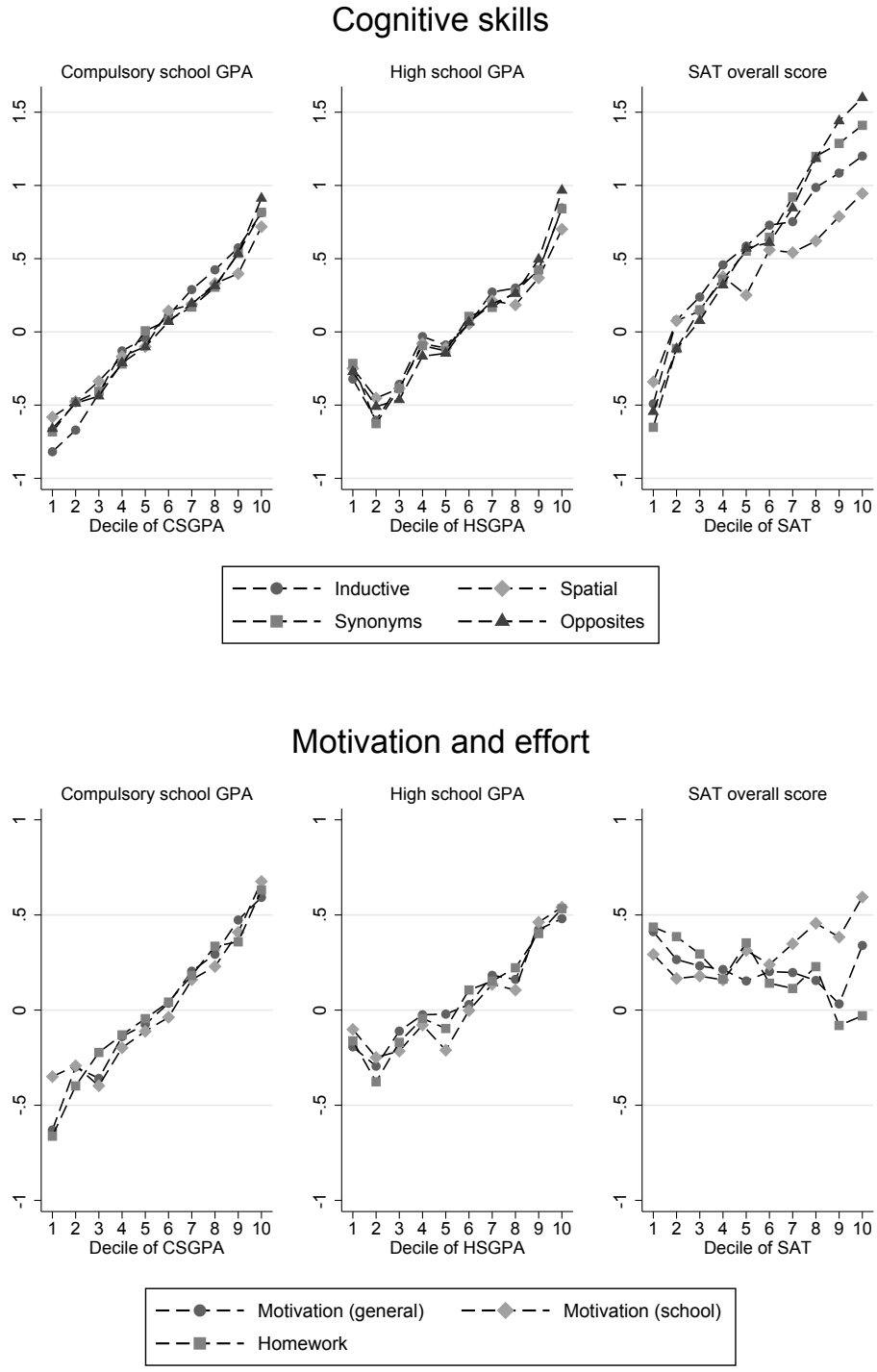


Figure A7: Test scores, detailed cognitive skills, and detailed measures of motivation and effort

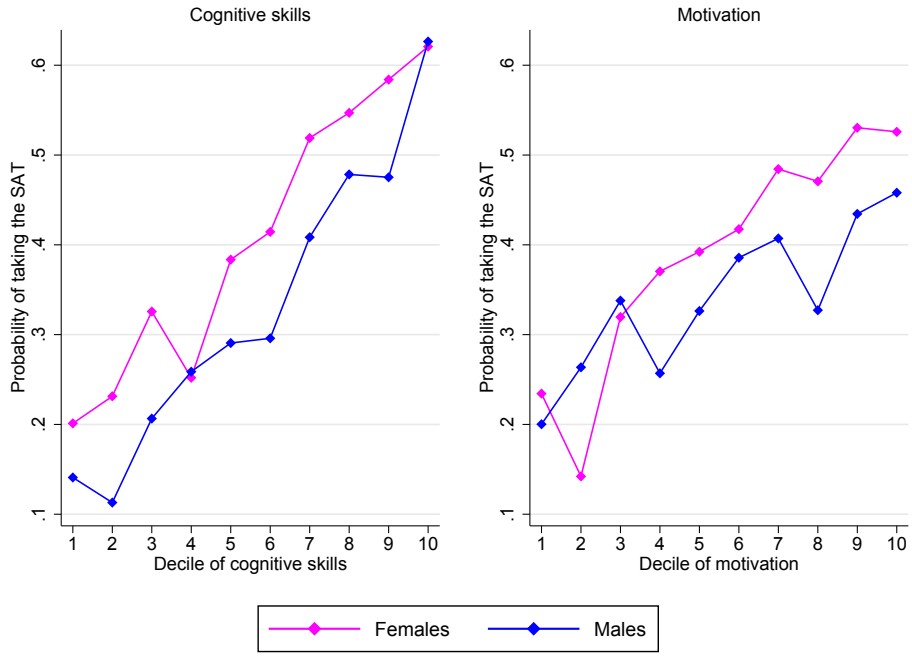


Figure A8: SAT participation, cognitive skills, and motivation

Table A1: Standardized scores over time

	CSGPA			HSGPA			SAT	
Year	Females	Males	Year	Females	Males	Year	Females	Males
1993	0.19 (0.004)	-0.18 (0.004)	1996	0.20 (0.005)	-0.20 (0.005)	1996	-0.17 (0.005)	0.19 (0.005)
1994	0.19 (0.005)	-0.18 (0.005)	1997	0.17 (0.005)	-0.17 (0.005)	1997	-0.16 (0.005)	0.20 (0.005)
1995	0.19 (0.004)	-0.18 (0.004)	1998	0.18 (0.005)	-0.19 (0.005)	1998	-0.15 (0.005)	0.18 (0.005)
1996	0.19 (0.004)	-0.18 (0.004)	1999	0.18 (0.005)	-0.19 (0.005)	1999	-0.15 (0.005)	0.18 (0.006)
1997	0.19 (0.004)	-0.19 (0.004)	2000	0.20 (0.005)	-0.20 (0.005)	2000	-0.14 (0.005)	0.16 (0.006)
1998	0.18 (0.005)	-0.17 (0.004)	2001	0.18 (0.005)	-0.19 (0.005)	2001	-0.12 (0.006)	0.14 (0.007)
1999	0.18 (0.005)	-0.17 (0.004)	2002	0.19 (0.005)	-0.20 (0.005)	2002	-0.12 (0.007)	0.15 (0.007)
2000	0.18 (0.004)	-0.17 (0.004)	2003	0.22 (0.005)	-0.22 (0.005)	2003	-0.13 (0.007)	0.16 (0.007)
2001	0.17 (0.004)	-0.16 (0.004)	2004	0.21 (0.005)	-0.21 (0.005)	2004	-0.10 (0.006)	0.13 (0.007)
2002	0.16 (0.004)	-0.16 (0.004)	2005	0.21 (0.005)	-0.21 (0.005)	2005	-0.14 (0.006)	0.16 (0.007)
2003	0.16 (0.004)	-0.15 (0.004)	2006	0.18 (0.005)	-0.19 (0.005)	2006	-0.16 (0.007)	0.19 (0.007)
2004	0.17 (0.004)	-0.16 (0.004)	2007	0.22 (0.005)	-0.22 (0.005)	2007	-0.14 (0.007)	0.15 (0.007)
2005	0.16 (0.004)	-0.15 (0.004)	2008	0.19 (0.005)	-0.19 (0.004)	2008	-0.14 (0.007)	0.16 (0.007)
2006	0.17 (0.004)	-0.16 (0.004)	2009	0.19 (0.004)	-0.20 (0.004)	2009	-0.16 (0.007)	0.17 (0.007)
2007	0.17 (0.004)	-0.16 (0.004)	2010	0.20 (0.004)	-0.20 (0.004)	2010	-0.15 (0.006)	0.16 (0.006)
2008	0.17 (0.004)	-0.16 (0.004)	2011	0.15 (0.004)	-0.14 (0.004)	2011	-0.16 (0.005)	0.17 (0.006)
2009	0.16 (0.004)	-0.16 (0.004)	2012	0.15 (0.004)	-0.15 (0.004)	2012	-0.21 (0.005)	0.21 (0.006)
2010	0.17 (0.004)	-0.16 (0.004)	2013	0.16 (0.004)	-0.15 (0.004)	2013	-0.19 (0.005)	0.20 (0.006)
2011	0.18 (0.004)	-0.16 (0.004)	2014	0.13 (0.005)	-0.13 (0.005)	2014	-0.18 (0.005)	0.20 (0.005)
2012	0.19 (0.004)	-0.18 (0.004)	2015	0.14 (0.005)	-0.14 (0.005)	2015	-0.19 (0.005)	0.21 (0.005)

Notes: For each year and gender, the mean of the indicated score is reported, with its standard error in parentheses.

Table A2: Correlations between scores

	All			Females			Males		
A1: Population of high-school graduates, graduation years 1996-2015 (1,726,166 observations)									
	CSGPA	HSGPA		CSGPA	HSGPA		CSGPA	HSGPA	
CSGPA	1			1			1		
HSGPA	0.64	1		0.63	1		0.63	1	
A2: Population of high-school graduates who took the SAT, graduation years 1996-2015 (633,126 observations)									
	CSGPA	HSGPA	SAT	CSGPA	HSGPA	SAT	CSGPA	HSGPA	SAT
CSGPA	1			1			1		
HSGPA	0.57	1		0.55	1		0.58	1	
SAT	0.45	0.41	1	0.51	0.46	1	0.45	0.44	1
B1: Population of high-school graduates, 1992 birth cohort (110,223 observations)									
	CSGPA	HSGPA		CSGPA	HSGPA		CSGPA	HSGPA	
CSGPA	1			1			1		
HSGPA	0.56	1		0.54	1		0.56	1	
B2: Population of high-school graduates who took the SAT, 1992 birth cohort (40,021 observations)									
	CSGPA	HSGPA	SAT	CSGPA	HSGPA	SAT	CSGPA	HSGPA	SAT
CSGPA	1			1			1		
HSGPA	0.45	1		0.42	1		0.48	1	
SAT	0.44	0.28	1	0.51	0.30	1	0.45	0.32	1
C1: UGU sample of high-school graduates, 1992 birth cohort (4,351 observations)									
	CSGPA	HSGPA		CSGPA	HSGPA		CSGPA	HSGPA	
CSGPA	1			1			1		
HSGPA	0.57	1		0.55	1		0.58	1	
C2: UGU sample of high-school graduates who took the SAT, 1992 birth cohort (1,940 observations)									
	CSGPA	HSGPA	SAT	CSGPA	HSGPA	SAT	CSGPA	HSGPA	SAT
CSGPA	1			1			1		
HSGPA	0.48	1		0.45	1		0.50	1	
SAT	0.45	0.24	1	0.51	0.25	1	0.48	0.28	1

Notes: High school graduates whose compulsory school GPA is unknown were dropped from each sample.

Table A3: Gender gaps in within-individual score differences

	SAT minus CSGPA				SAT minus HSGPA			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-0.50	-0.45	-0.38	-0.33	-0.50	-0.43	-0.45	-0.39
	(0.043)	(0.045)	(0.043)	(0.049)	(0.052)	(0.052)	(0.054)	(0.054)
Inductive skills		0.022		0.019		0.071		0.068
		(0.028)		(0.031)		(0.032)		(0.032)
Spatial skills		-0.018		-0.034		0.036		0.026
		(0.023)		(0.022)		(0.028)		(0.028)
Synonyms skills		0.19		0.18		0.21		0.20
		(0.039)		(0.038)		(0.038)		(0.038)
Verbal opposites skills		0.14		0.13		0.18		0.18
		(0.033)		(0.033)		(0.033)		(0.033)
Motivation (general)			-0.21	-0.17			-0.14	-0.084
			(0.029)	(0.032)			(0.034)	(0.030)
Motivation (school)			0.038	-0.0051			0.10	0.042
			(0.029)	(0.027)			(0.033)	(0.032)
Time spent on homework			-0.15	-0.14			-0.10	-0.070
			(0.019)	(0.019)			(0.026)	(0.024)
R-squared	0.09	0.22	0.20	0.31	0.06	0.21	0.09	0.22
Observations (raw—weighted)	1,940—45,478							

Notes: The dependent variables are within-individual differences in standardized test scores as indicated in the column headings. All right-hand side variables, except the female dummy, are standardized. Regressions are weighted using sampling weights. Robust standard errors in parentheses.

Table A4: Gender gaps across assessments and subject areas

	Compulsory school: math				Compulsory school: Swe & Eng				SAT: quant				SAT: verbal			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Female	0.14	0.14	-0.018	0.0064	0.39	0.41	0.22	0.26	-0.45	-0.35	-0.46	-0.38	-0.13	-0.025	-0.13	-0.028
	(0.047)	(0.038)	(0.046)	(0.038)	(0.045)	(0.037)	(0.043)	(0.036)	(0.053)	(0.037)	(0.056)	(0.039)	(0.049)	(0.036)	(0.053)	(0.038)
Inductive skills		0.38		0.35		0.25		0.21		0.36		0.36		0.055		0.052
		(0.023)		(0.022)		(0.024)		(0.023)		(0.023)		(0.023)		(0.023)		(0.023)
Spatial skills		0.15		0.16		0.076		0.083		0.16		0.17		-0.0038		-0.0056
		(0.021)		(0.020)		(0.022)		(0.020)		(0.021)		(0.021)		(0.020)		(0.020)
Synonyms skills		0.088		0.082		0.23		0.22		0.091		0.088		0.37		0.37
		(0.034)		(0.031)		(0.034)		(0.030)		(0.027)		(0.027)		(0.031)		(0.032)
Verbal opposites skills		0.13		0.14		0.19		0.20		0.17		0.17		0.27		0.27
		(0.030)		(0.028)		(0.032)		(0.030)		(0.025)		(0.025)		(0.026)		(0.026)
Motivation (general)			0.12	0.15			0.17	0.18			-0.084	-0.019			-0.093	-0.0053
			(0.033)	(0.024)			(0.029)	(0.022)			(0.040)	(0.023)			(0.034)	(0.022)
Motivation (school)			0.13	0.043			0.15	0.054			0.16	0.077			0.14	0.042
			(0.032)	(0.026)			(0.029)	(0.024)			(0.034)	(0.025)			(0.032)	(0.024)
Time spent on homework			0.17	0.14			0.14	0.13			-0.053	0.0038			-0.067	-0.026
			(0.023)	(0.018)			(0.023)	(0.018)			(0.023)	(0.018)			(0.023)	(0.017)
R-squared	0.01	0.36	0.10	0.43	0.04	0.39	0.16	0.48	0.06	0.45	0.08	0.45	0.01	0.46	0.03	0.46
Observations		4,351—123,668				4,270—118,395				1,940—45,478				1,940—45,478		

Notes: The dependent variables are standardized subject test scores as indicated in the column headings. All right-hand side variables, except the female dummy, are standardized. Regressions are weighted using sampling weights. Robust standard errors in parentheses.

Table A5: Selection-corrected regressions of SAT scores—father's skills and GPAs

	Father's skills			GPAs						Both	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Female	-0.33	-0.19	-0.12	-0.43	-0.44	-0.23	-0.24	-0.22	-0.23	-0.23	-0.20
	(0.050)	(0.044)	(0.046)	(0.041)	(0.040)	(0.034)	(0.034)	(0.042)	(0.042)	(0.042)	(0.046)
Inductive skills		0.17	0.27			0.15	0.14	0.20	0.18	0.12	0.18
		(0.024)	(0.030)			(0.022)	(0.022)	(0.027)	(0.027)	(0.025)	(0.029)
Spatial skills		0.085	0.10			0.039	0.038	0.041	0.045	0.049	0.061
		(0.023)	(0.027)			(0.018)	(0.019)	(0.022)	(0.022)	(0.023)	(0.027)
Verbal opposites skills		0.26	0.30			0.22	0.22	0.23	0.23	0.21	0.22
		(0.028)	(0.032)			(0.022)	(0.022)	(0.027)	(0.026)	(0.026)	(0.030)
Synonyms skills		0.29	0.35			0.26	0.25	0.28	0.28	0.27	0.31
		(0.035)	(0.037)			(0.024)	(0.024)	(0.029)	(0.029)	(0.030)	(0.033)
Motivation (school)		0.040	0.12			0.018	0.019	0.066	0.073	0.014	0.074
		(0.024)	(0.032)			(0.020)	(0.019)	(0.025)	(0.025)	(0.023)	(0.030)
Time spent on homework		-0.030	0.031			-0.068	-0.067	-0.027	-0.030	-0.077	-0.055
		(0.017)	(0.022)			(0.015)	(0.015)	(0.019)	(0.019)	(0.018)	(0.021)
Father's cognitive skills	0.29	0.092	0.13							0.086	0.11
	(0.028)	(0.022)	(0.027)							(0.022)	(0.026)
Father's non-cognitive skills	-0.023	-0.0036	0.032							-0.014	0.0063
	(0.027)	(0.020)	(0.023)							(0.020)	(0.023)
CSGPA				0.64	0.64	0.30	0.32	0.51	0.48	0.32	0.43
				(0.043)	(0.049)	(0.035)	(0.036)	(0.047)	(0.049)	(0.042)	(0.055)
HSGPA					0.075		0.022		0.063	0.00059	0.043
					(0.039)		(0.023)		(0.028)	(0.027)	(0.033)
Sample: father's skills	✓	✓	✓							✓	✓
Sample: high school					✓		✓		✓	✓	✓
Heckman correction			✓					✓	✓		✓
R-squared	0.13	0.52		0.29	0.30	0.57	0.58			0.56	
Observations	1,410	1,410	3,199	1,940	1,912	1,940	1,912	4,351	4,114	1,393	3,027

Notes: The dependent variable is the standardized SAT score. Results from OLS regresions are shown, adjusted for selection as indicated. All right-hand side variables, except the female dummy, are standardized. Regressions are weighted using sampling weights. Robust standard errors in parentheses.

Table A6: Gender gaps in predicted SAT scores

	(1)	(2)	(3)
Female	0.12	-0.080	-0.094
	(0.037)	(0.029)	(0.052)
Re-weighted with test-taking propensity		✓	
SAT-takers only			✓
Observations	4,351		1,940

Notes: The dependent variable is the predicted SAT score obtained from running the Heckman model separately for males and females, where general motivation is excluded from the 2nd-stage equation. The propensity to take the test is also estimated from these models.

Table A7: The explanatory power of skills and motivation within the sample of SAT takers

	Compulsory school GPA				High school GPA			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	0.22	0.28	0.095	0.14	0.21	0.26	0.16	0.20
	(0.046)	(0.042)	(0.053)	(0.046)	(0.046)	(0.048)	(0.052)	(0.051)
Inductive skills		0.18		0.18		0.13		0.13
		(0.024)		(0.027)		(0.031)		(0.031)
Spatial skills		0.090		0.10		0.036		0.045
		(0.019)		(0.017)		(0.023)		(0.023)
Synonyms skills		0.097		0.11		0.081		0.085
		(0.044)		(0.044)		(0.037)		(0.037)
Verbal opposites skills		0.12		0.12		0.075		0.080
		(0.034)		(0.033)		(0.031)		(0.030)
Motivation (general)			0.11	0.16			0.044	0.078
			(0.045)	(0.029)			(0.030)	(0.026)
Motivation (school)			0.13	0.070			0.065	0.025
			(0.028)	(0.023)			(0.030)	(0.029)
Time spent on homework			0.082	0.12			0.029	0.054
			(0.018)	(0.015)			(0.022)	(0.021)
R-squared	0.02	0.28	0.12	0.40	0.02	0.11	0.03	0.13
Observations (raw—weighted)	1,940—45,478							

Notes: This table replicates columns (1)-(8) of Table 3 for the sample of SAT takers. All right-hand side variables, except the female dummy, are standardized. Regressions are weighted using sampling weights. Robust standard errors in parentheses.

Table A8: School GPAs and SAT scores as predictors for adult earnings and university graduation

	Females						Males				Males			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
A. Graduated university														
CSGPA	0.25	0.29	0.20				0.09	0.23	0.30	0.24				0.09
	(0.00)	(0.00)	(0.00)				(0.00)	(0.00)	(0.00)	(0.00)				(0.00)
HSGPA				0.24	0.18		0.13				0.24	0.22		0.16
				(0.00)	(0.00)		(0.00)				(0.00)	(0.00)		(0.00)
SAT						0.11	0.02						0.13	0.03
						(0.00)	(0.00)						(0.00)	(0.00)
R-squared	0.24	0.21	0.10	0.22	0.13	0.05	0.14	0.24	0.25	0.12	0.25	0.17	0.06	0.18
B. 2014 annual earnings, '000SEK														
CSGPA	45.25	45.10	45.86				24.49	59.47	55.27	60.60				31.92
	(0.29)	(0.41)	(0.71)				(0.95)	(0.40)	(0.60)	(1.21)				(1.50)
HSGPA				37.41	40.34		28.21				48.27	54.04		48.36
				(0.32)	(0.52)		(0.73)				(0.50)	(0.91)		(1.21)
SAT						20.21	-1.95						8.38	-21.93
						(0.49)	(0.56)						(0.87)	(1.12)
R-squared	0.09	0.06	0.05	0.06	0.06	0.03	0.06	0.09	0.05	0.05	0.06	0.06	0.02	0.07
C. Percentile rank in 2014 annual earnings														
CSGPA	8.00	7.06	6.58				3.87	8.66	7.18	7.38				4.45
	(0.05)	(0.07)	(0.11)				(0.15)	(0.05)	(0.07)	(0.12)				(0.16)
HSGPA				5.80	5.65		3.96				6.22	6.35		5.73
				(0.05)	(0.08)		(0.12)				(0.05)	(0.09)		(0.12)
SAT						2.54	-0.73						0.35	-3.45
						(0.08)	(0.09)						(0.09)	(0.10)
R-squared	0.07	0.04	0.03	0.04	0.03	0.01	0.04	0.09	0.04	0.04	0.05	0.04	0.00	0.06
HSGPA non-missing		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
SAT score non-missing			✓			✓	✓			✓			✓	✓
Observations	361,525	288,539	145,959	288,539	145,959	145,959	145,959	380,558	284,545	124,383	284,545	124,383	124,383	124,383

Notes: The dependent variables are the indicated measures of annual labor earnings and university graduation. The baseline sample includes all individuals born 1977-1984 with non-missing compulsory school GPA. All regressions include cohort dummies. Robust standard errors in parentheses.