Sex differences in academic achievement are modulated by evaluation type[☆]Ava Guez^{c,*}, Hugo Peyre^{a,b,c}, Franck Ramus^c^a Neurodiderot, INSERM UMR 1141, Paris Diderot University, Paris, France^b Department of Child and Adolescent Psychiatry, Robert Debré Hospital, APHP, France^c LSCP, Département d'études cognitives, ENS, EHESS, PSL University, CNRS, Paris, France

ARTICLE INFO

Keywords:

Sex differences
Evaluation characteristics
Standardized tests
National examinations
Teacher grades

ABSTRACT

Studies on sex differences in academic skills have often reported diverging results depending on the type of evaluation used, with girls typically obtaining better school grades and results at national examinations, and boys scoring higher at standardized tests. In this paper, we provide a framework for better understanding and interpreting these differences, integrating previously established factors that affect variations in the gender gap across evaluation types: writing skills, stakes, self-discipline and grading bias. We apply this framework to a dataset containing the results of 23,451 French students in three evaluations characterized by different combinations of these factors: teacher evaluations, national examinations, and standardized tests. We find that, overall, girls show lower performance than boys in mathematics and higher in French. However, this main effect is modulated by evaluation type: relative to boys, girls over-perform in teacher evaluations and under-perform in standardized achievement tests, compared to national examinations. These effects are larger in mathematics than in French. These results offer new insights regarding the extent to which writing skills, stakes, self-discipline and grading bias may influence the observed gap.

1. Introduction

The question of sex differences in academic skills has garnered much attention and concern from researchers and policy-makers in the last decades. However, studies on the topic often report diverging results depending on how these skills are measured, which muddles interpretation.

On the one hand, several meta-analyses analyzing differences in achievement test scores have shown that girls obtain better results at language tests (Hedges & Nowell, 1995; Hyde & Linn, 1988), while boys perform better in mathematics (Else-Quest et al., 2010; Hyde et al., 1990; Reilly et al., 2015), albeit not consistently (Lindberg et al., 2010). The much publicized PISA (Programme for International Student Assessment) studies have confirmed these findings: at age 15, in most participating countries, girls outperformed boys in reading assessments, while the gap was reversed in mathematics (OECD, 2015). On the other hand, studies focusing on school marks and examinations consistently reported an advantage of girls in all subjects. Thus, the meta-analysis by Voyer and Voyer (2014) revealed that girls outperformed boys across all course materials, the largest difference being in language and the smallest in mathematics. Similarly, in a large sample of UK students,

Deary et al. (2007) found that girls performed better than boys in all subjects of the GCSE (with the exception of Physics where there was no difference). Several factors can help explain this apparent contradiction between results stemming from achievement tests and school evaluations.

First, some evaluations are more likely to assess certain aspects of students' personality and behavior, and in particular, *self-discipline*. In this context, self-discipline is defined as the ability to make a conscious effort to resist impulses in order to reach a higher goal (Duckworth & Seligman, 2006). As such, it is likely to affect teacher evaluation grades, since they reflect students' behavior in class and homework assignment completion. Previous research has shown that girls tend to display higher levels of self-discipline than boys (e.g., meta-analysis of gratification delay tasks by Silverman, 2003), and this difference partly explains why girls obtain better report card grades than predicted by their achievement test scores (Duckworth et al., 2015; Duckworth & Seligman, 2006; Kling et al., 2013). Similarly, Steinmayr and Spinath (2008) found that girls' lesser tendency to avoid work (as well as greater agreeableness) partly explained girls' advantage in German school marks (but not in mathematics).

Second, some evaluations tend to assess *writing skills* more than

[☆] We acknowledge funding from Agence Nationale de la Recherche (ANR-17-EURE-0017 and ANR-10-IDEX-0001-02 PSL). We thank the Direction de l'Evaluation, de la Prospective et de la Performance for providing us access to the Panel 2007.

* Corresponding author at: LSCP, Département d'Etudes Cognitives, Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France.

E-mail address: ava.guez@ens.psl.eu (A. Guez).

others, depending on the type of response used (typically, open answers versus multiple choice questions). Since girls display higher writing abilities than boys (Feingold, 1988; Hedges & Nowell, 1995; Reilly et al., 2018; Scheiber et al., 2015), this in turn may alter the observed gender gap. Indeed, when presented with open-ended questions, girls perform relatively better than boys, while boys obtain relatively better results when multiple choice questions are used; these results were found both in language and in mathematics (Bolger & Kellaghan, 1990; Lafontaine and Monseur, 2009b; Lindberg et al., 2010; Reardon et al., 2018 – however, it is possible that this may not hold when item difficulty is high: see Beller & Gafni, 2000; Routitsky & Turner, 2003; Willingham & Cole, 1997).

Third, some evaluations create more *stress* in students due to their high-stakes, which may affect boys' and girls' academic performance differently. Exploiting a pure change in stakes in school examinations (high-, medium- and low-stakes), Azmat et al. (2016) found that girls performed worse in higher- compared to lower-stakes settings (in Catalan, but not in mathematics). Similarly, in settings combining high-stakes and competition, girls obtain lower results than expected across all subjects (Cai et al., 2018; Jurajda & Münich, 2011; Ors et al., 2013). One should note that the effect of stress may be confounded with the motivation to perform well at low-stakes tests. Indeed, there is some indication that girls are more motivated than boys to do their best at low-stakes tests compared to higher-stakes ones, which may affect the sex differences in performance across tests in the same way as stress (DeMars et al., 2013; Eklöf, 2007; OECD, 2015; O'Neil et al., 2005 - note however that it is not clear from these papers whether the motivation difference had any effect on performance at the test).

Fourth, some evaluations, due to their non-blind nature, may generate a *grading bias* based on students' gender. Past studies have shown in a variety of countries that girls obtain higher results in non-anonymous evaluations compared to anonymous ones – thus suggesting that there is a grading bias in favor of girls (Breda & Ly, 2015; Falch & Naper, 2013; Lavy, 2008; Protivinský & Münich, 2018; Terrier, 2015; but see Lafontaine & Monseur, 2009a). It may be that this effect is moderated by ability level, as Lafontaine and Monseur (2009a) found that teachers tend to underestimate the performance of high-achieving girls' in mathematics, and over-estimate the performance of low-achieving girls'.

1.1. Goals of the study

To what extent does each of these evaluation characteristics affect differences between boys and girls in academic performance? How much do sex differences in performance depend on these differences in evaluation characteristics? In the present paper, we aimed to explore these questions by studying differences between boys and girls in measured academic performance. We compared boys' and girls' results in two subjects (French and mathematics) at three different types of evaluations for the same set of students (teacher evaluations, national examinations, and standardized achievement tests) from a large, representative sample of middle school students in France.

2. Method

2.1. Sample

We used data from the DEPP Panel 2007, a large cohort study led by the Direction de l'Évaluation, de la Prospective et de la Performance (DEPP; French Ministry of Education) containing rich data on 34,986 French students from their first year of middle school in 2007 (grade 6, 11 years old) to their last year of middle school (grade 9, 14 years old) (Trosseille et al., 2013). The study was compulsory and approved by the National Council for Statistical Information (CNIS) (visa n°2008A061ED and 2011A082ED), ensuring public interest and conformity with ethical, statistical and confidentiality standards. The sample was

randomly selected from an exhaustive sampling frame, ensuring representativeness by balancing available characteristics (region, public/private status of the school, urban unit, school establishment, age of entry in grade 6). The sample was constituted in such a way as to be representative of the population of French middle school students, with a slight over-representation of students in schools belonging to the *Réseau Ambition Réussite* (Success Ambition Network – schools in disadvantaged areas). The present study focusses on the grade 9 wave, when three different measures of achievement were reported: National examination grades, teacher grades, and standardized test scores. Our working sample includes students for whom results at the three tests in French and mathematics were available ($N = 23,451$). Thus, the sample size was not determined based on expected effect sizes but on available data. With this sample size, we are able to detect a sex difference of size $d = 0.037$ with 80% power. 52% of the participants in our working sample were girls. Students' average age in grade 9 was equal to 14.09 years ($SD = 0.42$).

2.2. Measures

2.2.1. Academic achievement

Three different measures of academic achievement, in French and mathematics, were collected in grade 9:

National examination grades: At the end of grade 9 (in June), all French students have to take the written tests of a national examination, the *Diplôme national du Brevet* (DNB). The tests, lasting 2 h each, are graded anonymously by teachers, and assess school knowledge acquired throughout the school year. They are composed of open-ended questions. In French, the exam is divided in three parts: open-ended questions on reading comprehension and grammar; text dictation; and essay. In mathematics, the exam is divided in three parts as well, including: numerical activities (open-ended questions in arithmetic, algebra and statistics); geometrical activities (open-ended questions in geometry); and problem solving (open-ended questions on a real-world problem). The DNB written examinations are the first official, nationwide examination that students take, and the grade they obtain constitutes about 40% of their final grade at the DNB (the remaining 60% coming from teacher grades in grade 9 in all subjects). Therefore, they are relatively high stakes for students.¹

Teacher grades: Teacher average grades include grades at in-class tests as well as homework grades throughout the year – thereby more influenced by students' self-discipline (both in terms of diligence with respect to school work and behavior in class). The grades count in the final DNB grade and are of great importance for selection into high school. Therefore, they also are relatively high stakes for students.

Standardized tests: For the purpose of the Panel 2007 cohort study, the DEPP administered standardized tests to students. These tests were carried out solely for national statistics and research, and had no impact on students', teachers' or schools' prospects, and were not even sent back to teachers. They are therefore low stakes for students. The mathematics test included short open-ended and multiple choice questions testing students in logic, mental arithmetic, problem solving, units and time calculations, and geometry (45 items) (Aubret & Blanchard, 1992; Blanchard & Berger, 1994; OECD, 2011). In French, two tests were administered: a cloze test (blank-filling task) composed of three short texts with missing logical

¹ We only had access to the final grade of the national examination, not to the different items composing the exam. Therefore, we could not compute reliability and consistency indices. We nevertheless provided the correlations between grades at the national examination, teacher grades and standardized examination as a proxy for reliability and validity (in Table A2).

Table 1
Characteristics of the three types of tests available in grade 9.

Characteristics	Teacher grades	National examinations	Standardized tests
Anonymous scoring	No	Yes	Yes
Stakes	High	High	Low
Assessing writing skills	Yes	Yes	No
Role of self-discipline	Yes	No	No

connectors, determiners or pronouns (20 items) (Aubret et al., 2006); and a reading comprehension test in limited time (12 min) composed of three short texts, each followed by five short open-ended questions (Aubret et al., 2006). The tests were administered from April to May. Internal consistency was good (see Table A1). The same tests were administered in grade 6 (except for some items that changed to match students' level), with high correlations between both mathematics and cloze tests in grade 6 and grade 9, and a moderate correlation for the reading comprehension test, suggesting an acceptable test-retest reliability (see Table A1).

Note that in the standardized tests, both in mathematics and French, open-ended questions did not require students to write full sentences, but merely to write down a short answer. In contrast, both teacher evaluations and national examinations expected student to write full sentences as answers. Table A2 shows correlation between the three types of evaluation, in Mathematics and in French. Table 1 summarizes the main characteristics of the three types of evaluation available. All scores were standardized with a mean 0 and a standard deviation of 1.

2.2.2. Covariates

We used the following variables in our analysis in order to control for students' differences in socio-demographic and cognitive characteristics:

Socio-demographic variables: As part of the cohort study, parents had to fill in questionnaires in grade 9. We used information extracted from this questionnaire in our analysis in order to control for socio-demographic factors: parental education (average of the years of education of the mother and the father), household monthly income (in logarithmic scale), number of books and CDs at home, age of entry in grade 6, extracurricular activities, and being schooled in a priority school.

Cognitive variables: Along with standardized tests assessing academic performance, students completed a nonverbal intelligence test in grade 9, Chartier's Reasoning Test on Playing Cards (*Raisonnement sur Cartes de Chartier*, RCC) (Cronbach's $\alpha = 0.87$) (Terriot, 2014). They also filled in a questionnaire measuring their perceived self-efficacy in three different aspects (autoregulation, social, and academic), the Children's Perceived Self-Efficacy scales (Bandura, 1990), from which three factor scores were constructed using confirmatory factor analysis (Cronbach's α above 0.80 for each of the three scores). Lastly, they answered a questionnaire measuring school related motivation (intrinsic motivation, extrinsic motivation, and amotivation) derived from the Academic Self-Regulation Questionnaire (Ryan & Connell, 1989), from which three factor scores were constructed using confirmatory factor analysis (Cronbach's $\alpha = 0.85$ for intrinsic motivation and 0.68 for extrinsic motivation and amotivation).

2.3. Analyses

Statistical analyses were carried out with the software SAS.²

2.3.1. Descriptive statistics

We first used descriptive statistics to understand how the gender gap in performance varies with the subject and evaluation type. We assessed the effect size of the difference for each evaluation score by computing Cohen's d , i.e. the standardized difference between male and female students' means.

2.3.2. Omnibus tests

In a second step, we performed omnibus tests with a repeated measured ANOVA in order to test whether the main effects of sex, subject and evaluation type, as well as their interactions with sex, are statistically significant. Note that, since scores were standardized for each test, there should be no main effect of subject and evaluation type.

2.3.3. Difference-in-differences

In a third step, we used difference-in-differences regressions to estimate the sign and significance of the effects of evaluation type on the gender gap in achievement in French and mathematics. This method has already been used in sex differences studies, in particular to assess the existence of a grading bias (Breda & Ly, 2015; Lavy, 2008; Terrier, 2015). As is summarized in Table 1, we assume that the three evaluation types (national examinations, teacher evaluations, and standardized tests) have different characteristics. Independently from socio-demographic and individual factors which affect performance similarly across evaluations, we suppose that results at standardized tests solely reflect students' ability in the subject; that results at the national examinations are, in addition, dependent on students' writing skills and their ability to cope with stress (induced by the high stakes involved)³; and that results at teacher evaluations are, in addition to all of the above, influenced by students' self-discipline and teachers' biases (since they are not anonymous). It is likely that these characteristics affect male and female students differently. Based on the literature, we hypothesize for example that female students would have an advantage when the evaluation relies more heavily on writing skills, when it is influenced by self-discipline, and when it is not anonymous. Similarly, we hypothesize that they would be at a disadvantage when the stakes are higher. In order to assess the relative influence of these different factors on the achievement gap between male and female students, we analyzed how the gender gap changes depending on the evaluation used. The presence of the three scores (standardized score, teacher grade, and national examinations grade) in each subject (French and mathematics) allowed us to use a difference-in differences estimation strategy with three different conditions. The score obtained by a student i on evaluation j ($Score_{ij}$ – standardized with a mean 0 and a standard deviation of 1) depends on her sex (dummy variable Fem equal to 1 if the student is a female) and on the type of the evaluation (dummy variables $Teach$ and Std equal to 1 for teacher evaluation or standardized test, respectively – the national examinations constitute the reference condition here). We may thus write the equation of the score for student i at evaluation j as follow:

$$Score_{ij} = \alpha + \gamma Fem_i + \beta_1 Teach_{ij} + \beta_2 Std_{ij} + \delta_1 (Fem_i \times Teach_{ij}) + \delta_2 (Fem_i \times Std_{ij}) + u_{ij} \quad (1)$$

Our coefficients of interest are those of the interaction terms, δ_1 and δ_2 . Estimating δ_1 and δ_2 with Eq. (1) is equivalent to estimating them

² The script is available on the Open Science Framework with the link: https://osf.io/c37hb/?view_only=d227c87f5ffd439fb5d8cc0ec84c5a7. Data are available on request on the Qu  telet PROGEDO French data archives for human and social sciences: <http://www.progedo-adisp.fr/enquetes/XML/lil-0955.xml>.

³ As noted in the introduction, the effect of stress on high-stakes tests versus low-stakes tests may be confounded with motivation to perform at low-stakes tests, but it is not possible to disentangle them.

with a difference equation where the difference in scores is the dependent variable (Lavy, 2008). In order to obtain δ_1 and δ_2 , we estimated Eq. (1) using generalized estimating equations (Liang & Zeger, 1986) with standard errors clustered at the individual level in order to correct for dependence among our repeated observations (PROC GENMOD).⁴ In addition, we estimated Eq. (1) controlling for socio-demographic variables. We subsequently included individual cognitive variables as covariates in order to see whether they would explain the observed differences. Missing data in covariates (see Table 2) were dealt with using multiple imputation (Rubin, 1987) (PROC MI and PROC MIANALYZE with MCMC method and 10 imputed datasets).

δ_1 measures to what extent the discrepancy between teacher evaluation scores and national examinations (DNB) scores differs between male and female students. Similarly, δ_2 reflects to what extent the discrepancy between standardized test scores and national examinations (DNB) scores differs between male and female students. We can indeed derive from the above equation that:

$$(\text{Teach_Score} - \text{DNB_Score})^{\text{Fem}} - (\text{Teach_Score} - \text{DNB_Score})^{\text{Male}} = \delta_1 \quad (2a)$$

$$(\text{Std_Score} - \text{DNB_Score})^{\text{Fem}} - (\text{Std_Score} - \text{DNB_Score})^{\text{Male}} = \delta_2 \quad (2b)$$

We can further express the score at each type of evaluation – hence δ_1 and δ_2 – in terms of the evaluation's hypothesized characteristics. As stated above, standardized tests should solely measure ability in the subject; while both national examinations and teacher evaluation scores should also reflect their writing skills, and their ability to cope with stress. In addition, teacher evaluations may reflect potential gender biases in grading, and students' discipline. We can formalize this as follow:

$$\text{Std_Score}_i = \alpha_1 \text{ability}_i + \epsilon_i \quad (3a)$$

$$\text{DNB_Score}_i = \alpha_1 \text{ability}_i + \alpha_2 \text{writing}_i + \alpha_3 \text{stress}_i + \mu_i \quad (3b)$$

$$\begin{aligned} \text{Teach_Score}_i \\ = \alpha_1 \text{ability}_i + \alpha_2 \text{writing}_i + \alpha_3 \text{stress}_i + \alpha_4 \text{discipline}_i + \alpha_5 \text{Fem}_i + \eta_i \end{aligned} \quad (3c)$$

The variables *ability*, *writing*, *stress* and *discipline* represent, respectively, a student's general ability in the subject, her writing skills, her ability to cope with stress, and her self-discipline – characteristics that we do not measure directly. We assume that each of these unobserved factors affects results at the different evaluations the same way, for males and females (i.e. the coefficients α_1 , α_2 , α_3 and α_4 are assumed to be the same across evaluations and sexes for a given subject). The last coefficient α_5 represents the grading bias (such that if it is positive, there is a bias in favor of girls, and if it is negative, there is a bias in favor of boys). Injecting Eqs. (3a), (3b) and (3c) into Eqs. (2a) and (2b), we obtain:

$$\delta_1 = \alpha_4 (\text{discipline}^{\text{Fem}} - \text{discipline}^{\text{Male}}) + \alpha_5 \quad (4a)$$

$$\delta_2 = -\alpha_2 (\text{writing}^{\text{Fem}} - \text{writing}^{\text{Male}}) - \alpha_3 (\text{stress}^{\text{Fem}} - \text{stress}^{\text{Male}}) \quad (4b)$$

Obviously, it will not be possible to determine the values of all the unknowns, since the observed variables will be fewer. It will nevertheless be possible to make some valuable inferences. Indeed, δ_1 reflects the difference between male and female students that can be attributed to grading bias and discipline, while δ_2 reflects the difference between male and female students that can be attributed to writing skills and the ability to cope with stress. The signs of δ_1 and δ_2 , estimated with Eq. (1), can thus inform us on the relative effect of the different unobserved evaluation factors on the gender gap in performance. Since higher

general ability, writing skills, ability to cope with stress and discipline are associated with higher test scores, α_1 , α_2 , α_3 and α_4 are assumed to be positive. Based on the extant literature, we expect female students to exhibit on average higher levels of self-discipline than their male counterparts. We also expect the grading bias to be in favor of female students (positive α_5). Therefore, we expect to find a positive δ_1 . Besides, we expect female students to have better writing skills, but to cope less well with stress. Thus, the two terms in Eq. (4b) are expected to be of opposite sign. A positive δ_2 would thus mean that the ability to cope with stress plays a more important role than writing skills in the gender gap in performance (whether it stems from the respective sizes of the α_2 and α_3 coefficients or the magnitude of the gender gaps in writing ability and stress). If δ_2 is null, then this means that both factors cancel out. Lastly, a negative δ_2 would suggest that writing skills play a more important role than the ability to cope with stress in the gender gap in performance.

3. Results

3.1. Sex differences between subjects and evaluation types

Table 2 shows descriptive statistics for all academic achievement variables as well as socio-demographic variables and cognitive variables by sex, and the effect size of the difference between male and female students.

Fig. 1 displays the distribution of evaluation scores for male and female students by subject and evaluation type, and Fig. 2 represents the effect sizes of the sex difference in performance across evaluations. Globally, female students performed better than their male peers in French, while the reverse was true in mathematics. However, their respective performance appears to depend on the evaluation used. There is indeed a clear trend as illustrated in Fig. 1, whereby both in French and mathematics female students scored the highest in teacher evaluations, followed by national examinations and lastly by standardized tests. Conversely, male students obtained their highest grades with standardized tests, followed by national examinations and teacher evaluations. Consequently, the gender gap in scores varies considerably across evaluation types, as illustrated in Fig. 2. This is most striking if we look at sex differences in mathematics scores across evaluation types: while the gap is clearly in favor of male students in the standardized test ($d = -0.41$, $p < 0.001$), it shrinks but is still in favor of male students in the national examination ($d = -0.11$, $p < 0.001$), and is almost null and reversed in the teacher examination ($d = 0.05$, $p < 0.001$). In French as well, the gap appears to vary across evaluation types, although it is in favor of female students in all three evaluations: the effect size is medium in the teacher evaluation ($d = 0.43$, $p < 0.001$) as well as in the national examination ($d = 0.40$, $p < 0.001$), but is small in the standardized test ($d = 0.18$, $p < 0.001$).

3.2. Omnibus interaction effects between subject, evaluation type, and sex

In order to assess the significance of these overall moderating effects of subject and evaluation type on the gender gap, we turn to results from the repeated measured ANOVA. Results are reported in Table 3. First, we observe a main effect of sex ($p < 0.0001$), reflecting the overall higher mean performance of girls over boys. Second, the main effects of evaluation type and subject are, as expected, non-significant (given that the scores were standardized). Third, the interaction between sex and evaluation type is significant ($p < 0.0001$), reflecting that teacher evaluations and the national exam yield a mean advantage for girls, while standardized tests yield a mean advantage for boys. Fourth, the interaction between sex and subject is significant ($p < 0.0001$), reflecting that girls perform better than boys in French, and less well in Mathematics. Finally, the three-way interaction between sex, evaluation type, and subject was also significant

⁴ We could not cluster at the school or class level because such information was not available.

Table 2
Descriptive statistics for achievement, socio-demographic, and cognitive variables, by sex.

Variables	N	Girls (N = 12,180)		Boys (N = 11,271)		Difference		
		M or %	SD	M or %	SD	d or O.R. [C.I.]	p	
Academic achievement variables								
French DNB grade (z-score)	23,451	0.18	0.96	−0.21	1.00	0.40	[0.37; 0.42]	< 0.001
French teacher grade (z-score)	23,451	0.20	0.97	−0.22	0.99	0.42	[0.40; 0.45]	< 0.001
French standardized test (z-score)	23,451	0.09	0.99	−0.09	1.00	0.18	[0.15; 0.21]	< 0.001
Mathematics DNB grade (z-score)	23,451	−0.05	0.97	0.05	1.03	−0.11	[−0.13; −0.08]	< 0.001
Mathematics teacher grade (z-score)	23,451	0.03	1.00	−0.03	1.00	0.05	[0.03; 0.08]	< 0.001
Mathematics standardized test (z-score)	23,451	−0.19	0.99	0.21	0.97	−0.40	[−0.43; −0.38]	< 0.001
Socio-demographic variables								
Parental education (years)	22,104	12.13	3.39	12.38	3.43	−0.07	[−0.10; −0.08]	< 0.001
Household monthly income (EUR)	13,412	3309	3783	3448	4070	−0.04	[−0.07; 0.00]	0.0408
Number of books and CDs in the household (score)	22,200	0.08	0.89	0.10	0.89	−0.02	[−0.05; 0.00]	0.0683
Age of entry in grade 6 (years)	23,451	11.08	0.40	11.09	0.00	−0.03	[−0.05; 0.00]	0.0491
Extracurricular activities (score)	21,775	0.27	0.27	0.28	0.27	−0.04	[−0.06; −0.01]	0.0049
Schooled in a disadvantaged area (%; OR)	21,661	17.50	–	16.48	–	1.08	[1.00; 1.15]	0.0379
Cognitive variables								
Non-verbal intelligence (out of 30)	23,306	18.85	5.59	18.82	6.02	0.00	[−0.02; 0.03]	0.7635
Perceived self-efficacy in autoregulation (z-score)	23,236	0.02	1.00	−0.02	1.00	0.04	[0.01; 0.07]	0.0023
Perceived academic self-efficacy (z-score)	22,517	0.10	0.97	−0.11	1.01	0.22	[0.19; 0.24]	< 0.001
Perceived social self-efficacy (z-score)	22,692	−0.19	1.03	0.21	0.92	−0.41	[−0.44; −0.39]	< 0.001
Intrinsic motivation (z-score)	23,089	0.04	0.99	−0.04	1.01	0.08	[0.057; 0.11]	< 0.001
Extrinsic motivation (z-score)	23,089	−0.02	0.98	0.02	1.02	−0.04	[−0.06; −0.01]	0.0038
Amotivation (z-score)	23,089	−0.17	0.85	0.18	1.11	−0.36	[−0.38; −0.33]	< 0.001

Source: MENESR DEPP.

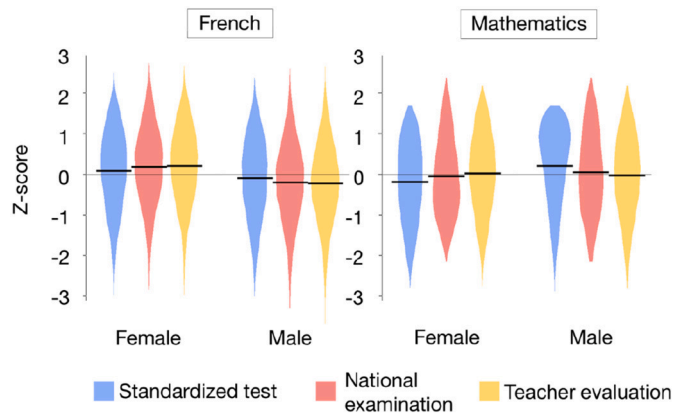


Fig. 1. Distribution of test scores, by subject, sex and test type. The solid black lines represent group means and the beans the smoothed density curves.
Source: MENESR DEPP.

($p < 0.0001$), reflecting that the magnitude of the sex by evaluation interaction is larger in Mathematics than in French (all interactions are clearly visible in Fig. 2).

3.3. Relative effects of evaluation characteristics on sex differences

In order to properly measure this apparent effect of evaluation type on the gender gap in achievement across subjects, we turn to the difference-in-differences estimation.

Table 4 displays results from the difference-in-differences regression without covariates. The coefficients of interest are the two interaction terms.

The coefficient δ_1 of the interaction term between *Female* and *Teacher evaluation* is significantly different from zero and positive, both in French and mathematics. Thus, female students increased their performance in teacher evaluations relative to the national examinations more than their male counterparts. According to our analysis of Eq. (4a), this may be due to greater self-discipline for homework and regular study in females, or teacher bias in favor of females, both

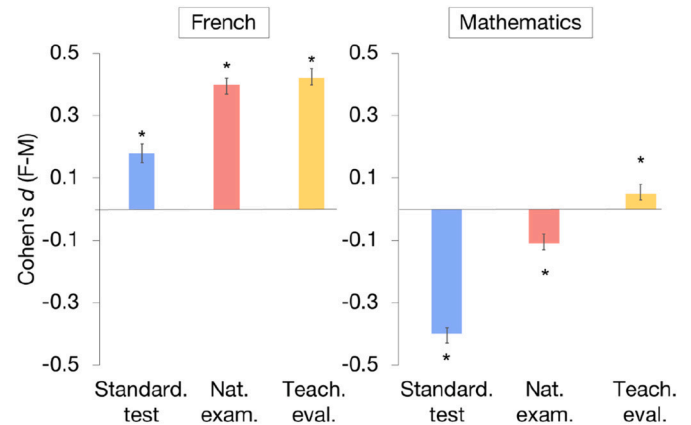


Fig. 2. Difference in test scores between male and female students, by subject and test type. Error bars indicate 95% confidence intervals; and stars (*) indicate $p < 0.001$. “Standard. test” stands for standardized tests; “Nat. exam.” stands for national examinations; and “Teach. eval.” stands for teacher evaluations.

Source: MENESR DEPP.

Table 3

Results from a repeated measures ANOVA with evaluation type and subject as within-subject variables.

Variables	Df	F value	p-Value
Sex	1	295.60	< 0.0001
Evaluation type	2	0.38	0.685
Subject	1	0.33	0.564
Sex * Evaluation type	2	397.15	< 0.0001
Sex * Subject	1	2111.46	< 0.0001
Evaluation type * Subject	2	0.13	0.876
Sex * Evaluation type * Subject	2	35.43	< 0.0001

Source: MENESR-DEPP.

Table 4
Difference-in-differences regression estimates.

Parameters	French			Mathematics		
	β [C.I.]		p	β [C.I.]		p
Female	0.40	[0.37; 0.42]	< 0.0001	−0.11	[−0.13; −0.08]	< 0.0001
Teacher evaluation	−0.01	[−0.03; 0.00]	0.0492	−0.08	[−0.10; −0.07]	< 0.0001
Standardized test	0.11	[0.10; 0.13]	< 0.0001	0.16	[0.14; 0.17]	< 0.0001
Female \times Teacher evaluation (δ_1)	0.03	[0.01; 0.05]	0.0056	0.16	[0.14; 0.18]	< 0.0001
Female \times Standardized test (δ_2)	−0.22	[−0.24; −0.20]	< 0.0001	−0.30	[−0.32; −0.28]	< 0.0001

Source: MENESR-DEPP.

factors likely to increase scores in teacher evaluations relative to national examinations. The effect is very small in French ($\delta_1 = 0.03$, $p = 0.0056$), but larger in mathematics ($\delta_1 = 0.16$, $p < 0.001$).

The second coefficient of interest, δ_2 , is that of the interaction term between *Female* and *Standardized test*. It is significantly different from zero, negative and of moderate size both in mathematics ($\delta_2 = -0.30$, $p < 0.001$) and in French ($\delta_2 = -0.22$, $p < 0.001$). Indeed, female students' performance was lower in standardized tests than at the national examinations, whereas that of their male peers was higher in standardized tests than at the national examinations. As indicated in Eq. (4b), the negative sign of δ_2 suggests that writing skills play a more important role in the observed gender gap in performance than the ability to cope with stress. Indeed, female students scored lower in the standardized tests despite the fact that these tests were lower stakes and therefore less stressful than the national examinations.

As can be seen from Tables A3 and A4, additional adjustment on students' socio-demographic and cognitive factors had little effect on the estimates of δ_1 and δ_2 . Thus, these characteristics do not explain the gender gap in performance across evaluation types. As a last step, in order to ensure that the estimated effects do not reflect the effects of interactions between evaluation type and other variables that would correlate with sex, we added interaction terms between our evaluation type dummies and all other covariates. The results are reported in Table A5. We can see that our coefficients of interest (the interactions between sex and evaluation type) are slightly reduced. In French, δ_1 is still non-significant and reduces from 0.03 in the regression without interactions, to -0.02 when adding in the control interaction terms; while δ_2 reduces from -0.22 to -0.18 ($p < 0.001$). In mathematics, δ_1 reduces from 0.16 to 0.12 ($p < 0.001$) and δ_2 from -0.30 to -0.25 ($p < 0.001$). This decrease is due to the fact that there were significant interaction effects between evaluation type and other covariates, most notably being schooled in a disadvantaged area, and academic self-efficacy. However, these other interactions did not reduce by much the sex by evaluation type interaction terms, which thus strengthens our results.

4. Discussion

This paper aimed at better understanding the relative influences of evaluation characteristics on sex differences in academic achievement. We compared boys' and girls' results at three different types of evaluations (teacher evaluations, national examinations, and standardized achievement tests) both in French and mathematics, from the same set of 23,451 middle school French students. We found that, across subjects and evaluation types, girls performed better than boys. However, this main effect was modulated by subject: relative to boys, girls obtained better results in French, and worse in mathematics. Crucially, the effect of sex was also modulated by the evaluation type, although this moderation was smaller than the sex difference across subjects. This effect was the focus of our paper, which we discuss more thoroughly now.

4.1. Evaluation conditions affect sex differences in achievement

In line with previous findings, we found that girls performed better than boys in French, and worse in mathematics, when assessed with standardized achievement tests (Else-Quest et al., 2010; Hedges & Nowell, 1995; Hyde et al., 1990; Hyde & Linn, 1988; OECD, 2015; but see Lindberg et al., 2010), and that they performed better in both when assessed with teacher evaluations (Voyer & Voyer, 2014). Contrary to previous findings on national examinations by Deary et al. (2007), showing that British female students scored significantly higher than males in both English ($d = 0.41$) and mathematics ($d = 0.03$), we found that French female students performed higher in French ($d = 0.40$), but lower in mathematics ($d = -0.11$). This result may be due to socio-economic and cultural differences. For instance, the UK's Global Gender Gap Index (GGI), which measures gender parity in the areas of health, education, economy and politics, was higher than France's in the 2000s (Hausmann et al., 2010). Moreover, boys reported that their parents find mathematics more important to study and for their future career than girls do, and this difference is lower in the UK (Stoet et al., 2016). Both the GGI and the gender gap in parental mathematics valuation are correlated with the gender gap in mathematics achievement in PISA (Stoet et al., 2016). Although these results contrast with the larger advantage of boys over girls in PISA mathematics test in the UK compared to France, it is possible that such socio-economic and cultural variations are more strongly associated with results at national examinations than at lower-stakes standardized achievement tests such as PISA.

4.2. Interpretations of the effect of evaluation conditions

In both French and mathematics, girls over-performed in teacher evaluations, and under-performed in standardized achievement tests, compared to national examinations. The former effect was expected, given that higher stakes and grading bias are supposed to be detrimental to girls' performance in national evaluations compared to teacher evaluations. The latter is more interesting. Indeed, female students should have an advantage in national examinations compared to standardized tests due to the higher requirements on writing skills in the first; however, the higher stakes in national examinations should play in favor of boys compared to standardized tests. Therefore, the finding that girls under-performed in standardized achievement tests compared to national examinations indicates that writing skills play a more important role than stakes in the gender gap in performance. The fact that these effects were in the same direction in French and mathematics suggests that the mechanisms at play are the same in both subjects.

Interestingly, both girls' over-performance in teacher evaluations and under-performance in standardized tests were larger in mathematics than in French. Females' greater relative performance in mathematics than in French during teacher evaluations (compared to national examinations) may result from superior self-discipline in mathematics, a greater effect of self-discipline on mathematics, or from teachers having a greater grading bias in favor of females in mathematics. While our data does not allow us to distinguish these

possibilities, the third option seems likely, as evidence for a grading bias in favor of girls has previously been reported in France in mathematics but not in French (Breda & Ly, 2015; Terrier, 2015).

Females' lower relative performance in mathematics than in French on standardized tests (compared to national evaluations) might be due to several factors: a larger sex difference in writing skills in mathematics; a larger effect of writing skills on performance in mathematics; a smaller sex difference in stress in mathematics; or a lower effect of stress on performance in mathematics (or, undistinguishably, a lower motivation to perform well at low-stakes standardized tests in mathematics). None of these hypotheses appear very intuitive or plausible, and we know of little data that might adjudicate between them. However, Azmat et al.'s (2016) findings that girls performed worse in higher-stakes settings in Catalan but not in mathematics are in line with the hypothesis that there may be a smaller sex difference in stress in mathematics compared to French.

4.3. Limitations

In light of the following limitations, conclusions must be interpreted with caution. First, our inferences regarding the extent to which each characteristic affects sex differences are purely based on assumptions from the literature, since we did not have the data to measure these characteristics. Further research with more comprehensive datasets is needed to address this issue, as it is also possible that we omitted other evaluation characteristics that may affect the gap across evaluations. For example, in mathematics, the presence of problems with spatially based solution strategies and multiple solution paths may influence the sex difference in favor of boys (Gallagher et al., 2002).

Second, our model is somewhat simplistic: it assumes that there are no interactions between the different evaluation characteristics, and that students' ability and evaluation characteristics affect results in the different evaluations in the same way across sexes. Although these assumptions were necessary in our model to draw simple interpretations regarding the effects of these factors on the gender gap, it is possible that they might not hold.

Another potential limitation is our assumption that male and female students' handwriting are indistinguishable in anonymous evaluations, which is why we only included a potential grading bias in teacher evaluations. However, Baird (1998) showed that grades are not affected by the gender style of handwriting and Breda and Ly (2015) found that the percentage of correct guesses of students' gender based on handwritten anonymous exam is only 68.6%. Furthermore, even if gender could be partly detected and induced a grading bias in anonymous examinations, this would only underestimate the grading bias in our study (and in previous studies on the subject).

Despite these limitations, our data clearly showed that there are differences in the gender gap in achievement across evaluations, which must depend on differences in evaluation characteristics.

Appendix A

Table A1
Internal consistency for standardized tests scores.

Test	Cronbach's alpha	Correlation with the same score in grade 6
Mathematics	0.935	0.83
Cloze test	0.817	0.73
Reading comprehension	0.816	0.56

Source: MENESR-DEPP.

4.4. Conclusions and practical implications

Sex differences in academic achievement is a hot topic for which multiple interpretations have been proposed. One possible interpretation is that such differences do not really exist in terms of sheer academic competence, but that they emerge only as a result of unfair evaluation conditions. Our results confirm that sex differences are modulated by multiple evaluation conditions. This multiplicity makes it difficult to attribute sex differences in a given evaluation or a given subject to a specific factor. For instance, in our study, it seems that stress due to perceived high-stakes only played a limited role in explaining the female disadvantage in mathematics, since girls actually succeeded better in mathematics at the high stakes national examination than at the low stakes standardized tests – and this difference was smaller in French. Modifying these factors can alter sex differences substantially, sometimes even reversing them. However, this does not mean that sex differences are solely the product of evaluation conditions, and that they might entirely disappear under “ideal” testing conditions. In this regard, it is noteworthy that changing some factors may reduce the gap in one subject, while it may increase it in another subject. For example, we observed that using teacher evaluations reduced the gap in mathematics, but increased it in French. Therefore, finding testing conditions that globally reduce sex differences is complex. One may also question whether this is a worthy goal. Indeed, it is important to distinguish adverse impact from fairness (Halpern, 2002): finding significant sex differences in a given evaluation does not imply that the evaluation is unfair. A fair evaluation is one which is comparably valid for both sexes, meaning that it evaluates accurately the skills targeted for each group. To have a correct picture of students' skills and progress, we should thus strive to eliminate factors that are not intended to be measured and that differentially affect female and male students. Hence, fair evaluations would in theory be low-stakes and corrected anonymously. In the context of our study, the evaluation that filled both of these conditions (the standardized test) was the one which reduced the gap the most in French, but increased it the most in mathematics. However, this test was based on multiple choice questions and short answers, in which boys tend to perform better. Ultimately, the response format does not affect fairness per se: the choice should depend on which response mode most accurately taps the target abilities. If anything, it might seem appropriate to balance different response formats as much as allowed by the test setting. For a test designed for research purposes, such as the standardized tests reported here, it may seem difficult to introduce open-ended questions, as the time students can spend on the test is limited and correction time is very costly. In the case of a national examination like the DNB which already consists of open-ended questions, introducing multiple choice items would be much more feasible.

Table A2

Correlations between national examinations, standardized tests and teacher grades in French and Mathematics.

	National examination	Standardized test	Teacher grades
<i>French</i>			
National examination	1.00		
Standardized test	0.67	1.00	
Teacher grades	0.72	0.63	1.00
<i>Mathematics</i>			
National examinations	1.00		
Standardized tests	0.73	1.00	
Teacher grades	0.75	0.65	1.00

Source: MENESR-DEPP.

Table A3

Difference-in-differences regression estimates, controlling for students' socio-demographic characteristics.

Parameters	French			Mathematics		
	β [C.I.]		p	β [C.I.]		p
Female	0.42	[0.40; 0.45]	< 0.0001	−0.08	[−0.10; −0.05]	< 0.0001
Teacher evaluation	−0.01	[−0.04; 0.01]	0.2328	−0.08	[−0.11; −0.06]	< 0.0001
Standardized test	0.11	[0.09; 0.14]	< 0.0001	0.16	[0.13; 0.18]	< 0.0001
Female \times Teacher evaluation (δ_1)	0.03	[0.00; 0.06]	0.0942	0.16	[0.13; 0.19]	< 0.0001
Female \times Standardized test (δ_2)	−0.22	[−0.25; −0.18]	< 0.0001	−0.30	[−0.33; −0.27]	< 0.0001
Parental education	0.07	[0.07; 0.08]	< 0.0001	0.07	[0.07; 0.07]	< 0.0001
Household income	0.11	[0.09; 0.12]	< 0.0001	0.15	[0.13; 0.16]	< 0.0001
Books and CDs in the household	0.16	[0.15; 0.17]	< 0.0001	0.13	[0.12; 0.14]	< 0.0001
Extracurricular activities	0.22	[0.19; 0.25]	< 0.0001	0.16	[0.14; 0.19]	< 0.0001
Schooled in a disadvantaged area	−0.15	[−0.17; −0.13]	< 0.0001	−0.20	[−0.22; −0.18]	< 0.0001

Source: MENESR-DEPP.

Table A4

Difference-in-differences regression estimates, controlling for students' socio-demographic and cognitive factors.

Parameters	French			Mathematics		
	β [C.I.]		p	β [C.I.]		p
Female	0.31	[0.28; 0.33]	−0.19	[−0.21; −0.17]		< 0.001
Teacher evaluation	−0.01	[−0.03; 0.01]	−0.08	[−0.10; −0.06]		< 0.001
Standardized test	0.11	[0.09; 0.13]	0.16	[0.14; 0.17]		< 0.001
Female \times Teacher evaluation (δ_1)	0.03	[0.00; 0.06]	0.16	[0.13; 0.19]		< 0.001
Female \times Standardized test (δ_2)	−0.22	[−0.24; −0.19]	−0.30	[−0.33; −0.27]		< 0.001
Non-verbal intelligence	0.05	[0.05; 0.05]	0.07	[0.07; 0.07]		< 0.001
Parental education	0.05	[0.05; 0.05]	0.05	[0.05; 0.05]		< 0.001
Household income	0.06	[0.04; 0.07]	0.09	[0.07; 0.1]		< 0.001
Books and CDs in the household	0.10	[0.09; 0.11]	0.06	[0.05; 0.06]		< 0.001
Extracurricular activities	0.16	[0.14; 0.19]	0.11	[0.09; 0.13]		< 0.001
Schooled in a disadvantaged area	−0.12	[−0.14; −0.1]	−0.14	[−0.16; −0.13]		< 0.001
Self-efficacy in autoregulation	0.09	[0.08; 0.10]	0.03	[0.02; 0.04]		< 0.001
Social self-efficacy	−0.13	[−0.14; −0.13]	−0.14	[−0.15; −0.14]		< 0.001
Academic self-efficacy	0.29	[0.28; 0.30]	0.27	[0.26; 0.28]		< 0.001
Intrinsic motivation	−0.05	[−0.06; −0.03]	0.02	[0.01; 0.03]		< 0.001
Extrinsic motivation	0.01	[0.00; 0.02]	0.00	[−0.01; 0.01]		0.5113
Amotivation	0.00	[−0.01; 0.01]	0.02	[0.01; 0.02]		< 0.001

Source: MENESR-DEPP.

Table A5

Difference-in-differences regression estimates, controlling for interactions between students' characteristics and evaluation type.

Parameters	French			Mathematics		
	β [C.I.]		p	β [C.I.]		p
Female	0.31	[0.29; 0.33]	< 0.001	−0.20	[−0.22; −0.18]	< 0.001
Teacher evaluation	0.31	[0.06; 0.55]	0.0145	0.21	[−0.03; 0.45]	0.0876
Standardized test	−0.17	[−0.42; 0.07]	0.1553	−0.17	[−0.39; 0.04]	0.1165
Female \times Teacher evaluation (δ_1)	−0.02	[−0.05; 0.01]	0.2378	0.12	[0.09; 0.15]	< 0.001
Female \times Standardized test (δ_2)	−0.18	[−0.21; −0.15]	< 0.001	−0.25	[−0.28; −0.22]	< 0.001

(continued on next page)

Table A5 (continued)

Parameters	French			Mathematics		
	β [C.I.]		<i>p</i>	β [C.I.]		<i>p</i>
Parental education \times Teacher evaluation	−0.01	[−0.01; 0.00]	0.0012	−0.01	[−0.02; −0.01]	< 0.001
Parental education \times Standardized test	0.00	[−0.01; 0]	0.2016	−0.01	[−0.02; −0.01]	< 0.001
Household income \times Teacher evaluation	−0.02	[−0.05; 0.02]	0.3085	−0.01	[−0.04; 0.02]	0.5715
Household income \times Standardized test	−0.02	[−0.05; 0.02]	0.3232	−0.01	[−0.04; 0.02]	0.3979
Books and CDs \times Teacher evaluation	−0.02	[−0.04; 0]	0.0516	−0.03	[−0.05; −0.01]	0.0013
Books and CDs \times Standardized test	0.04	[0.02; 0.06]	< 0.001	0.02	[0; 0.04]	0.0242
Extracurricular activities \times Teacher evaluation	−0.01	[−0.06; 0.05]	0.8535	0.01	[−0.05; 0.06]	0.8210
Extracurricular activities \times Standardized test	−0.01	[−0.06; 0.05]	0.7723	−0.02	[−0.07; 0.04]	0.5127
Schooled in a disadvantaged area \times Teacher evaluation	0.19	[0.15; 0.23]	< 0.001	0.28	[0.24; 0.32]	< 0.001
Schooled in a disadvantaged area \times Standardized test	−0.08	[−0.12; −0.04]	< 0.001	0.01	[−0.02; 0.05]	0.4849
Non-verbal intelligence \times Teacher evaluation	0.00	[0; 0]	0.8188	0.00	[0; 0]	0.1595
Non-verbal intelligence \times Standardized test	0.02	[0.02; 0.02]	< 0.001	0.03	[0.02; 0.03]	< 0.001
Self-efficacy in autoregulation \times Teacher evaluation	−0.03	[−0.06; −0.01]	0.0032	−0.02	[−0.04; 0]	0.0536
Self-efficacy in autoregulation \times Standardized test	0.07	[0.04; 0.09]	< 0.001	0.03	[0.01; 0.05]	0.0039
Social self-efficacy \times Teacher evaluation	−0.05	[−0.07; −0.03]	< 0.001	−0.03	[−0.05; −0.02]	0.0002
Social self-efficacy \times Standardized test	0.04	[0.02; 0.06]	< 0.001	0.07	[0.06; 0.09]	< 0.001
Academic self-efficacy \times Teacher evaluation	0.13	[0.1; 0.15]	< 0.001	0.09	[0.07; 0.11]	< 0.001
Academic self-efficacy \times Standardized test	−0.08	[−0.1; −0.06]	< 0.001	−0.11	[−0.13; −0.09]	< 0.001
Intrinsic motivation \times Teacher evaluation	0.03	[0; 0.06]	0.0211	0.02	[0; 0.05]	0.0777
Intrinsic motivation \times Standardized test	−0.02	[−0.05; 0.01]	0.1142	−0.05	[−0.08; −0.03]	< 0.001
Extrinsic motivation \times Teacher evaluation	0.02	[−0.01; 0.04]	0.1783	0.02	[0; 0.04]	0.1221
Extrinsic motivation \times Standardized test	−0.03	[−0.06; −0.01]	0.0083	0.00	[−0.02; 0.02]	0.8370
Amotivation \times Teacher evaluation	0.02	[−0.01; 0.04]	0.1450	−0.01	[−0.03; 0.01]	0.5485
Amotivation \times Standardized test	0.02	[0; 0.04]	0.0729	0.00	[−0.03; 0.02]	0.6708
Parental education	0.06	[0.05; 0.06]	< 0.001	0.06	[0.05; 0.06]	< 0.001
Household income	0.07	[0.05; 0.09]	< 0.001	0.09	[0.07; 0.12]	< 0.001
Books and CDs in the household	0.09	[0.08; 0.11]	< 0.001	0.06	[0.05; 0.07]	< 0.001
Extracurricular activities	0.17	[0.13; 0.21]	< 0.001	0.11	[0.07; 0.15]	< 0.001
Schooled in a disadvantaged area	−0.15	[−0.18; −0.13]	< 0.001	−0.24	[−0.27; −0.21]	< 0.001
Non-verbal intelligence	0.04	[0.04; 0.05]	< 0.001	0.06	[0.06; 0.06]	< 0.001
Self-efficacy in autoregulation	0.08	[0.06; 0.09]	< 0.001	0.03	[0.01; 0.04]	0.0010
Social self-efficacy	−0.13	[−0.14; −0.12]	< 0.001	−0.16	[−0.17; −0.15]	< 0.001
Academic self-efficacy	0.28	[0.26; 0.29]	< 0.001	0.28	[0.26; 0.29]	< 0.001
Intrinsic motivation	−0.05	[−0.07; −0.03]	< 0.001	0.03	[0.01; 0.05]	0.0010
Extrinsic motivation	0.02	[0; 0.03]	0.0707	−0.01	[−0.02; 0.01]	0.3072
Amotivation	−0.02	[−0.03; 0]	0.0564	0.02	[0; 0.03]	0.0135

Source: MENESR-DEPP.

References

- Aubret, F., & Blanchard, S. (1992). Valeur prédictive du test analytique de mathématique. *L'orientation scolaire et professionnelle*, 21(4), 449–454.
- Aubret, J., Blanchard, S., & Sontag, J.-C. (2006). Évaluer les compétences des collégiens en 6e/5e. *L'orientation scolaire et professionnelle*, 35(3), 446–473.
- Azmat, G., Calsamiglia, C., & Iriberrí, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association*, 14(6), 1372–1400. <https://doi.org/10.1111/jeea.12180>.
- Baird, J. (1998). What's in a name? Experiments with blind marking in A-level examinations. *Educational Research*, 40(2), 191–202. <https://doi.org/10.1080/0013188980400207>.
- Bandura, A. (1990). *Multidimensional scales of perceived self-efficacy*. Stanford University.
- Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, 42(1–2), 1–21. <https://doi.org/10.1023/A:1007051109754>.
- Blanchard, S., & Berger, S. (1994). Valeurs prédictives d'épreuves psychopédagogiques en français et en mathématiques utilisées en classe de 6e. In M. Huteau (Ed.), *Les techniques psychologiques d'évaluation des personnes* (pp. 595–598). E.A.P.
- Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, 27(2), 165–174. <https://doi.org/10.1111/j.1745-3984.1990.tb00740.x>.
- Breda, T., & Ly, S. T. (2015). Professors in Core science fields are not always biased against women: Evidence from France. *American Economic Journal: Applied Economics*, 7(4), 53–75. <https://doi.org/10.1257/app.20140022>.
- Cai, X., Lu, Y., Pan, J., & Zhong, S. (2018). Gender gap under pressure: Evidence from China's National College Entrance Examination. *The Review of Economics and Statistics*, 101(2), 249–263. https://doi.org/10.1162/rest_a_00749.
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35(1), 13–21. <https://doi.org/10.1016/j.intell.2006.02.001>.
- DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research and Practice in Assessment*, 8, 69–82.
- Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, 98(1), 198–208. <https://doi.org/10.1037/0022-0663.98.1.198>.
- Duckworth, A. L., Shulman, E. P., Mastronarde, A. J., Patrick, S. D., Zhang, J., & Druckman, J. (2015). Will not want: Self-control rather than motivation explains the female advantage in report card grades. *Learning and Individual Differences*, 39, 13–23. <https://doi.org/10.1016/j.lindif.2015.02.006>.
- Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, 7(3), 311–326. <https://doi.org/10.1080/15305050701438074>.
- Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, 136(1), 103–127. <https://doi.org/10.1037/a0018053>.
- Falch, T., & Naper, L. R. (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review*, 36, 12–25. <https://doi.org/10.1016/j.econedurev.2013.05.002>.
- Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, 43(2), 95–103. <https://doi.org/10.1037/0003-066X.43.2.95>.
- Gallagher, A., Levin, J., & Cahalan, C. (2002). Cognitive patterns of gender differences on mathematics admissions tests. *ETS Research Report Series*, 2002(2), i–30. <https://doi.org/10.1002/j.2333-8504.2002.tb01886.x>.
- Halpern, D. F. (2002). Sex differences in achievement scores: Can we design assessments that are fair, meaningful, and valid for girls and boys? *Issues in Education*, 8(1), 2.
- Hausmann, R., Tyson, L. D., Zahidi, S., & World Economic Forum (2010). The global gender gap report 2010. http://www3.weforum.org/docs/WEF_GenderGap_Report_2011.pdf.
- Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science (New York, N.Y.)*, 269(5220), 41–45.
- Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, 107(2), 139–155. <https://doi.org/10.1037/0033-2909.107.2.139>.
- Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, 104(1), 53–69. <https://doi.org/10.1037/0033-2909.104.1.53>.
- Jurajda, Š., & Münich, D. (2011). Gender gap in performance under competitive pressure: Admissions to Czech universities. *American Economic Review*, 101(3), 514–518. <https://doi.org/10.1257/aer.101.3.514>.
- Kling, K. C., Nofle, E. E., & Robins, R. W. (2013). Why do standardized tests underpredict women's academic performance? The role of conscientiousness. *Social Psychological*

- and *Personality Science*, 4(5), 600–606. <https://doi.org/10.1177/1948550612469038>.
- Lafontaine, D., & Monseur, C. (2009a). Les évaluations des performances en mathématiques sont-elles influencées par le sexe de l'élève? *Mesure et évaluation en éducation*, 32(2), 71–98.
- Lafontaine, D., & Monseur, C. (2009b). Gender gap in comparative studies of reading comprehension: To what extent do the test characteristics make a difference? *European Educational Research Journal*, 8(1), 69–79. <https://doi.org/10.2304/eeerj.2009.8.1.69>.
- 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. <https://doi.org/10.1016/j.jpubeco.2008.02.009>.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13. <https://doi.org/10.2307/2336267>.
- Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, 136(6), 1123–1135. <https://doi.org/10.1037/a0021276>.
- OECD (2011). *Résultats du PISA 2009. Savoirs et savoir-faire des élèves : Performances des élèves en compréhension de l'écrit, en mathématiques et en sciences* (OECD Editions). Vol. I.
- OECD (2015). *The ABC of gender equality in education*. OECD Publishing <https://doi.org/10.1787/9789264229945-en>.
- O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment*, 10(3), 185–208. https://doi.org/10.1207/s15326977ea1003_3.
- Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: Does competition matter? *Journal of Labor Economics*, 31(3), 443–499. <https://doi.org/10.1086/669331>.
- Protivínský, T., & Münich, D. (2018). Gender Bias in teachers' grading: What is in the grade. *Studies in Educational Evaluation*, 59, 141–149. <https://doi.org/10.1016/j.stueduc.2018.07.006>.
- Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educational Researcher*, 47(5), 284–294. <https://doi.org/10.3102/0013189X18762105>.
- Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology*, 107(3), 645–662. <https://doi.org/10.1037/edu0000012>.
- Reilly, D., Neumann, D. L., & Andrews, G. (2018). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist*. <https://doi.org/10.1037/amp0000356>.
- Routitsky, A., & Turner, R. (2003). *Item format types and their influence on cross-national comparisons of student performance*. Chicago: Paper presented at the Annual Meeting of the American Educational Research Association.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.
- Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, 57(5), 749–761. <https://doi.org/10.1037/0022-3514.57.5.749>.
- Scheiber, C., Reynolds, M. R., Hajovsky, D. B., & Kaufman, A. S. (2015). Gender differences in achievement in a large, nationally representative sample of children and adolescents: Gender and achievement. *Psychology in the Schools*, 52(4), 335–348. <https://doi.org/10.1002/pits.21827>.
- Silverman, I. (2003). Gender differences in delay of gratification: A meta-analysis. *Sex Roles*, 49(9), 451–463. <https://doi.org/10.1023/A:1025872421115>.
- Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality*, 22(3), 185–209. <https://doi.org/10.1002/per.676>.
- Stoet, G., Bailey, D. H., Moore, A. M., & Geary, D. C. (2016). Countries with higher levels of gender equality show larger national sex differences in mathematics anxiety and relatively lower parental mathematics valuation for girls. *PLoS One*, 11(4), Article e0153857. <https://doi.org/10.1371/journal.pone.0153857>.
- Terrier, C. (2015). Giving a little help to girls? Evidence on grade discrimination and its effect on students' achievement. *CEP discussion paper no 1341*. Centre for Economic Performance London School of Economics and Political Science.
- Terriot, K. (2014). Testons les tests: Le RCC (Raisonnement sur Cartes de Chartier). *ANAE-Approche Neuropsychologique des Apprentissages chez l'Enfant*, 26(129), 179–183.
- Trosseille, B., Champault, F., & Lieury, A. (2013). Évaluation de 30 000 élèves de 6^e du collège français. Présentation et introduction. *Bulletin de psychologie*, Numéro 523(1), 3. <https://doi.org/10.3917/bupsy.523.0003>.
- Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, 140(4), 1174–1204. <https://doi.org/10.1037/a0036620>.
- Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Lawrence Erlbaum Associates Publishers.