Journal of the Institute for Educational Research
Volume 57 • Number 1 • June 2025 • 5–28
UDC: 37.091.3::51
37.091.26-057.874(497.11)"2013/2019"
Received 16.12.2024; accepted 26.05.2025.

ISSN 0579-6431
ISSN 1820-9270 (Online)
DOI: 10.2298/ZIPI2501005V
Original research paper

SIMPSON'S PARADOX IN MATHEMATICS GRADING: A CASE STUDY OF SERBIAN PRIMARY SCHOOLS*

Srđan Verbić** • ORCID: 0000-0003-1908-1421
FEFA, Metropolitan University, Belgrade, Serbia

Marija Kaplar • ORCID: 0000-0002-0920-8276
Faculty of Technical Sciences, University of Novi Sad, Serbia

Zorana Lužanin • ORCID: 0000-0002-7215-2252
Faculty of Sciences, University of Novi Sad, Serbia, Novi Sad

Dragana Trnavac • ORCID: 0000-0002-0919-5049
Faculty of Sciences, University of Novi Sad, Serbia, Novi Sad

Branislav Randelović • ORCID: 0000-0002-0643-0955
Faculty of Electrical Engineering, University of Niš, Niš, Serbia

ABSTRACT

The paper presents an empirical study that examines the academic performance in mathematics of fifteen-year-old students at the end of elementary school in Serbia. In conducting the analysis, the student results of the national test and their mathematics grades were utilized. The study covers a seven-year period 2013-19 and over 440 thousand students. Empirical findings affirm that girls exhibit superior math grades and higher achievements in national testing. However, the occurrence of Simpson's paradox indicates a grading bias favoring girls. The results further highlight that girls exhibit a notably higher interest in attempting to solve open-ended tasks, even when unsuccessful, in contrast to boys, who leave those tasks unanswered during tests more frequently. The qualitative component of the research involved a focus group comprising teachers. It was conducted to gain insights into the teachers' perspectives and experiences regarding the gender grading gap.

Key words:
grading, national exam, mathematics, gender bias, Simpson's paradox.

* **To cite this article:** Verbić, S., Kaplar, M., Lužanin, Z., Trnavac, D. & Randelović, B. (2025). Simpson's paradox in mathematics grading: A case study of Serbian primary schools. *Zbornik Instituta za pedagoška istraživanja, 57*(1), 5–28. DOI: 10.2298/ZIPI2501005V.

** E-mail: sverbic@fefa.edu.rs

INTRODUCTION

Enrollment in secondary schools, the granting of student scholarships, and decisions regarding further education are typically contingent on the grades obtained in elementary school and/or the results of external testing. Therefore, it is not surprising that there is significant interest in students' grades and their achievements in external examinations. Consequently, a substantial volume of research has focused on grades and academic achievements, with numerous scientific studies conducting comprehensive investigations into gender. These studies significantly emphasize examining disparities in students' academic achievements and grades, particularly focusing on the 'gender achievement gap' and the 'gender grading gap'. The term 'gender gap', also known as the 'gender achievement gap', typically denotes the disparity in performance between girls and boys. This phenomenon is commonly quantified by measuring the differences in their test scores (Protivínský & Münich, 2018). While the 'gender grading gap' particularly highlights the disparity in grades assigned to boys and girls of similar skill levels, as determined through anonymous testing, in non-anonymous evaluations it is often carried out by teachers.

This study is dedicated to exploring the presence and extent of the gender gap and the gender grading gap in mathematics within the Serbian educational framework. The advantage of this research lies in its extensive dataset, encompassing data spanning seven years from 2013 to 2019, and includes over 60,000 participants for each year. Additionally, in this study, unanswered tasks have been analyzed concerning both gender and task type. The purpose of this analysis is to gain deeper insights into the gender gap. It involves examining data on the instances where students skipped certain test items, which could indicate either a reduced motivation or a reluctance to put forth greater effort in solving mathematical problems. Numerous prior studies that encompass various educational systems have consistently indicated that a gender grading gap in mathematics frequently arises, typically favoring female students (Protivínský & Münich, 2018). Existing studies suggest that many factors can contribute to the gender grading gap where one of them is teacher-student interactions (Protivínský & Münich, 2018). Taking this into consideration, this study integrates a focus group of primary school teachers, an approach distinct from previous methodologies, to enrich the understanding of teachers' opinions about grading. This qualitative component of the research was

conducted to gain insights into the teachers' perspectives and experiences regarding the gender grading gap.

The structure of this paper is organized in the following manner: Section 2 provides a literature review upon which this study is founded. Section 3 presents an overview of the institutional setting and data used in our analysis, as well as the methodologies applied. Section 4 details the results of our quantitative research, which investigates gender disparities in different assessment styles. Section 5 focuses on the insights gathered from a qualitative study involving a focus group. Finally, Section 6 encompasses a thorough discussion of our findings, followed by reflections on the implications of modifying educational practices.

LITERATURE REVIEW

Gender gap and gender grading gap

Research into gender disparities in academic abilities consistently reveals similar trends across various assessment methods. Generally, it is observed that girls outperform boys in school grades, whereas boys tend to achieve higher scores on standardized tests. The study by Terrier (2020) found that middle school teachers in France are inclined to award higher grades to girls. Further, in the study conducted by Guez et al. (2020), the performance of boys and girls in French and mathematics was examined, employing three distinct assessment types: teacher evaluations, national exams, and standardized achievement tests. Analyzing a representative sample of middle school students in France, the findings indicated that girls generally performed better in French, whereas their performance in mathematics was not as strong as that of boys. Giofrè et al. (2020) reported that also in Italian schools, male students demonstrated superior performance compared to female students in mathematics, with the reverse trend observed in reading. Notably, this discrepancy in performance increased progressively from the 2nd to the 8th grade. Di Liberto et al. (2022) compared teacher-assigned grades and standardized test scores of students in Italy. They found that boys were graded less favorably than girls in both mathematics and language. Falch and Naper (2013) observed similar patterns in Norway, with their study revealing that girls get significantly higher

grades than boys when the same skills are assessed by their teacher. This gender grading gap in favor of the girls is found in both languages and mathematics. The research by Protivínský and Münich (2018) investigated gender grading biases in the Czech Republic by analyzing teacher evaluations against entrance exam scores for 15-year-olds in mathematics and their native language. The study's findings pointed to a prevailing bias in favor of female students. Ferman and Fontes (2022) discovered evidence of gender grading bias against boys in mathematics assessments within their study, which focused on a sample of students from middle schools and the initial two years of high school in Brazil. When it comes to Serbia, the results of the Trends in International Mathematics and Science Study (TIMSS) conducted in 2003 indicated that fifteen-year-old girls outperformed boys on average in mathematics. Similar results were obtained in 2019, although that cycle included only fourth-grade students (Đerić et al., 2021; Mullis et al., 2020). In contrast, results from the Programme for International Student Assessment (PISA) showed that boys in Serbia achieved better results in mathematics than girls, although the difference was not always statistically significant (OECD, 2023; Videnović & Čaprić, 2020).

Zanga and De Gioannis (2023) conducted a systematic review of 37 published studies to investigate the student characteristics that lead to biases or discrimination in grading. The majority of these studies focused on gender bias, with just over half (52%) affirming the presence of gender-discriminatory practices in grading. The research indicates that, on average, teachers tend to award higher evaluations to female students than those reflected in objectively measured blind test scores. In contrast, male students often receive lower grades than females in academic settings, despite similar performances in standardized tests. It is important to note, however, that a small subset of studies have reported instances of bias favoring male students (Guttmann & Boudo, 1988; Lavy & Sand, 2015). In contrast, Rakshit and Sahoo's (2023) study in India reported an absence of bias in teacher evaluations within mathematics. Correspondingly, investigations in Sweden by Hinnerich et al. (2011) and another study in India by Hanna and Linden (2012) also indicated a lack of a gender-based grading gap. The studies mentioned above testify to the frequent presence of gender grading biases favoring girls, particularly in the context of mathematics. Furthermore, it has become apparent that this phenomenon is widespread, manifesting in various educational systems across different countries.

Teachers' Math-Gender Beliefs

The assessment practices of mathematics teachers are crucial as they often impact students' motivation, attitudes, and beliefs. Numerous studies investigating teachers' grading practices have shown that school grades are not exclusively based on students' academic achievements; they also reveal teachers' biases toward certain social categories. Lavy (2008) examines the root cause of the gender grading gap, investigating whether it arises from the behaviors of teachers or students. His findings indicate that the gap is closely associated with the characteristics of the teachers, including their gender, age, experience, and family size, thereby implying that teacher behavior, rather than student conduct, is likely the primary cause of the gap. Given the empirical evidence of bias in grading students in mathematics, the question of whether there are gender-related beliefs among teachers is raised (Dersch et al., 2022; Li, 1999; Lindner et al., 2022; OECD, 2015; Sumpter, 2016). Li (1999) asserts that the existing body of literature collectively suggests that, even in the absence of conclusive evidence, teachers often have distinct beliefs about male and female students. They tend to stereotype mathematics as a male domain, which is evident in their inclination to overestimate the mathematical abilities of male students, hold higher expectations for them, and maintain more favorable attitudes toward male students. Dersch et al. (2022) identify three distinct misconceptions among teachers. 1) The empathizing-systemizing theory posits that boys, being more systematic thinkers, and girls, being more empathic thinkers, contribute to boys being naturally more adept at math. 2) The notion of girls' compensation suggests that boys need to exert less effort to achieve the same level of success in mathematics due to their innate talent, which is supposedly greater than that of girls. 3) The concept of girls' non-compensability implies that, despite generally putting in more effort, girls are typically less skilled in math compared to boys. Additionally, Tiedemann (2002) notes that teachers often attribute girls' failures in math to a lack of talent, whereas boys' failures are more likely attributed to a lack of effort. This reflects a gender bias in evaluation practices. In experimental settings, when teachers assume that female students achieved certain performance outcomes, they tend to underrate these outcomes compared to those of male students, as demonstrated in the studies by Avitzour et al. (2020) and also observed by Holder and Kessels (2017). These findings contrast with those from numerous studies where a grading gap is present but favors girls. On one side, there are teacher beliefs favoring boys, and on

the other, teacher grades that favor girls. In light of this, the current study seeks to illuminate this phenomenon through focus groups, examining teachers' perspectives on gender-based grading.

Format of tasks and omitting answers

Over several decades, the task formats used in assessments have been the subject of extensive scrutiny, as evidenced by numerous empirical studies and their subsequent meta-analyses. Multiple empirical studies have documented that students' performance varies depending on the format of the tasks. These studies have consistently shown that boys generally perform better in multiple-choice formats, whereas girls excel in tasks requiring more open-ended responses (Beller & Gafni, 2000; Birenbaum & Feldman, 1998; Garner & Engelhard, 1999; Riener & Wagner, 2018; Wester & Henriksson, 2000). However (Beller & Gafni, 2000), through two separate studies, presented conflicting findings. In one study, gender effects were more noticeable in multiple-choice (MC) items compared to open-ended (OE) items, while the other study showed a reverse trend, with gender effects favoring OE items over MC items. These contradictory results question the widely held belief that girls invariably excel over boys in OE items, thereby challenging the idea that item format solely accounts for gender differences in mathematical performance. It is significant to note that Beller and Gafni's research emphasized the influence of item difficulty over its format in determining gender-related performance variations. Their findings propose that the complexity of an item, rather than its structural format, plays a more critical role in these differences. In particular, the research observed an enhanced performance by boys compared to girls as the item difficulty escalated (Beller & Gafni, 2000).

The inclination to omit tasks, specifically choosing not to attempt solving them, varies between genders (Riener & Wagner, 2017). Von Schrader and Ansley (2006) conducted a study to analyze gender differences in the rate of omitting answers on multiple-choice tests. They discovered that girls more frequently skipped tasks in mathematics, while boys were more inclined to omit tasks in vocabulary and reading. This pattern was consistent with academic performance trends, where girls typically outperformed boys in reading, but boys had a slight edge in mathematics, as also noted by Becker and Forsyth (1990) and Han and Hoover (1994). Consequently, in many cases, the gender group more prone to omitting tasks corresponded with the

group achieving lower in those specific subjects (Von Schrader & Ansley, 2006). Riener and Wagner (2017) found that girls are more likely to skip answering tasks, particularly when they are challenging. However, this gender disparity vanishes when external incentives are offered (Riener & Wagner, 2017).

DATA AND METHODOLOGY

The research design comprises two consecutive parts. The first part involved quantitative analysis focused on the results of the final exam and student grades, while the second part entailed qualitative research that engaged mathematics teachers.

Quantitative research

In this study, we have made a secondary analysis of the results of the mathematics tests at the Final exam and students' final grades at the end of compulsory education for seven consecutive years, the period before the COVID-19 pandemic (2013–2019). This timeframe was selected to avoid potential inconsistencies that may have arisen from changes in teaching or grading practices due to the altered mode of instruction during the COVID-19 pandemic. In the period examined in this study (2013–2019), the Final exam in Serbia consisted of three tests: Serbian language (or other mother tongue), Mathematics, and the combined test (natural and social sciences). All students who finish compulsory education after the 8th grade of primary school, typically at the ages of between 14.5 to 15.5 years old, are required to take the Final exam. The main purpose of these tests is to enable the selection of students when enrolling in secondary school. The number of students who participated in the Final exam for these seven years was 446814 and varied between 62618 and 65223 students per year. The ratio of girls among examinees was between 48.2% and 48.9%.

The mathematics test at the Final Exam in Serbia consistently consists of 20 tasks to be completed within two hours. The tasks cover five areas: numbers and operations, algebra and functions, geometry, measurement, and data analysis. These tasks are classified into three levels of difficulty: basic, intermediate, and advanced, and are presented in various formats. Over the seven years of administering the Final Exam, five types of tasks have been used in the mathematics tests: multiple

choice (23 tasks), fill-in-the-blank (3 tasks), short answer (36 tasks), open-ended (74 tasks), and matching (4 tasks). The total test score is calculated as the sum of points for each task, with scores ranging from 0 to a maximum of 20 points.

School grades in Serbia range from 1 (insufficient) to 5 (excellent). At the end of a school year, students receive the final grade for all school subjects. These grades are quoted in the certificate of completed education. Students who have a final grade of 1 in any of the subjects are required to take the correctional exam, and they were not included in this research. A linear combination of test results and the final grades for a few selected school subjects represents the total number of points used for students' ranking.

Grading gap research is typically conducted by comparing grades with some form of external testing. Although both types of assessments are intended to evaluate students' academic progress, they differ significantly in many aspects (Marcenaro-Gutierrez et al., 2023). External testing primarily employs standardized tests, which are graded anonymously by external evaluators and are designed to produce comparable results across all students, ensuring consistency across schools (Marcenaro-Gutierrez et al., 2023; Willingham et al., 2002). Standardized tests in external assessments generally focus on testing students' competencies, aiming to measure their skills and abilities (Hanushek & Woessmann, 2012), whereas teacher assessments often use a content-based approach, typically associated with the evaluation of knowledge (Borghans et al., 2016; Lauermann et al., 2020; Marcenaro-Gutierrez et al., 2023). The type of exam for these two types of assessments also differs significantly: external tests are considered high-stakes exams, while school grading is categorized as low-stakes. Another key difference lies in the time of testing where in schools, knowledge is tested immediately after instruction, without the time gap present in external testing. Additionally, teacher assessments are based on several months of testing, while external testing consists of a single test conducted on one day, with only one opportunity for completion. Despite these differences, existing literature suggests that comparing these two forms of assessment is possible to some extent and, furthermore, provides a satisfactory indication of the presence of a grading gap (Falch & Naper, 2013; Lavy, 2008; Matějů & Smith, 2014). Moreover, in many educational systems, both teacher grades and external tests are included in the evaluation of students' achievements, which are used to rank students and determine their admission to higher levels of education. Given that this is also the

case in Serbia, this study compares these two methods of evaluation to identify potential differences.

Data on the results of the final examination in mathematics were sourced from the Institute for Education Quality and Evaluation, the designated entity responsible for the preparation and implementation of the final examination. Each student who participated in the final examination had their performance on each task of the test recorded, along with grades in mathematics, physics, chemistry, biology, history, and geography, as well as their gender and the schools they attended. The data obtained from the Institute for Education Quality and Evaluation and used in this study were provided in a fully anonymized form, ensuring the privacy and confidentiality of individuals' information.

To test the hypotheses, in addition to descriptive statistics, we employed appropriate statistical tests including the t-test and test of equality of proportions; Cohen's d and Cohen's h, which accounts for effect size, were considered due to the large sample size. Statistical analysis and data visualization were conducted using R programming language.

Qualitative research

The focus group was established to gain deeper insights into gender disparities in student performance, particularly in terms of grades and achievements (Adler et al., 2019; Lindqvist et al., 2020). The research was guided by two specific questions: i) What observations do teachers have regarding gender differences in students' mathematics grades and achievements? ii) How do these teachers interpret or explain their observations?

The participants in this study are seven teachers with varying levels of professional experience, employed in primary schools during the spring semester of the 2022-2023 academic year from four distinct school districts. All participants had been teaching for at least five years during the period covered by the quantitative study. The focus group interview was conducted with one moderator and two assistant moderators.

Initially, the teachers were prompted to share their observations on gender disparities in student achievements and grades, drawing from their personal experiences. The first question posed to the teachers by the researcher asked them to share their impressions regarding whether they observe a gender gap and a gender

grading gap in their own teaching practice, as well as to express their opinions and reasoning on these issues. Subsequently, they were presented with data illustrating that female students generally received higher grades than their male counterparts. Specifically, the teachers were shown a graphical representation of results derived from the quantitative data, clearly displaying the presence of a grading gap, after which they were asked to provide their explanations of such findings. The researcher then guided the discussion through various subtopics, which emerged organically from the conversation rather than being predefined. These subtopics encompassed a range of issues, including differences in grade distribution, reasons for grading gap, and the influence of grades on future career paths.

The discussion was conducted in a synchronous online discussion board environment. This choice was motivated by its potential to facilitate reflective engagement with each other's comments, as highlighted by Newell et al. (2002), and to encourage participation from individuals who might be less vocal in face-to-face settings, as noted by Groth (2006). The conversation spanned two hours. Three researchers independently reviewed and transcribed all participant responses during the focus group. After the discussion, the researchers jointly reconciled their notes to create a comprehensive and unified transcript. Thematic coding of the data was conducted using an inductive approach, with codes emerging directly from the content rather than being predefined. To ensure the reliability of the coding process, both researchers independently developed initial codes and then met to compare, discuss, and harmonize their interpretations. Discrepancies were resolved through discussion until full agreement was reached. Although formal inter-coder agreement metrics such as Cohen's Kappa were not calculated, the collaborative nature of the coding process served as a measure of internal consistency. The matrix developed in accordance with Miles and Huberman's (1994) framework was used to structure the data, allowing for the identification of recurring themes and relationships between teachers' observations and their explanatory frameworks. This matrix included the teachers' observations as well as their explanations for these observations. Data within the matrix were systematically categorized according to the subtopics discuswsed.

RESULTS

Grades and achievements in the final exam

Table 1 and Table 2 show the mean values of the final grades in Mathematics and achievement on Mathematics test for boys and girls. The relative difference between mean values is given as Cohen's *d*. In all of the selected years, girls outperformed boys in both final grades and test results. It is visible that the mean value of the final grade in Mathematics is quite stable, and it does not vary significantly across years (Fig. 1). It is also visible that the relative difference in final grades between boys and girls is much greater than the difference between achievements' mean values. Since the number of degrees of freedom is more than 60,000 in all these cases, all these differences are statistically highly significant.

TABLE 1. Number of examinees, percent of final grades by gender in Mathematics, mean values of the final grades for boys and girls, standard deviation, t-value of the difference, and Cohen's d

year	gender	N	grade				mean grade	SD	t-value	Cohen's d
			2	3	4	5				
2013	boys	32981	42.3%	21.1%	16.1%	20.5%	3.15	1.21	48.3*	0.37
	girls	31433	26.5%	21.1%	18.0%	34.4%	3.60			
2014	boys	32452	43.1%	21.0%	15.3%	20.7%	3.14	1.22	46.5*	0.37
	girls	30166	27.4%	21.0%	17.6%	33.9%	3.58			
2015	boys	33445	42.1%	21.0%	15.6%	21.3%	3.16	1.22	49.1*	0.38
	girls	31778	26.1%	20.4%	18.6%	34.9%	3.62			
2016	boys	33134	42.2%	21.2%	15.6%	21.0%	3.16	1.22	50.3*	0.39
	girls	31609	25.9%	20.5%	18.5%	35.1%	3.63			
2017	boys	32019	43.2%	20.9%	15.2%	20.7%	3.13	1.22	48.1*	0.38
	girls	30599	27.2%	20.3%	18.3%	34.2%	3.59			
2018	boys	32948	43.8%	20.7%	15.1%	20.4%	3.12	1.22	49.3*	0.38
	girls	31472	27.4%	20.3%	18.5%	33.8%	3.59			
2019	boys	32157	43.8%	20.8%	15.2%	20.2%	3.12	1.21	46.3*	0.36
	girls	30621	27.9%	21.0%	18.7%	32.5%	3.56			

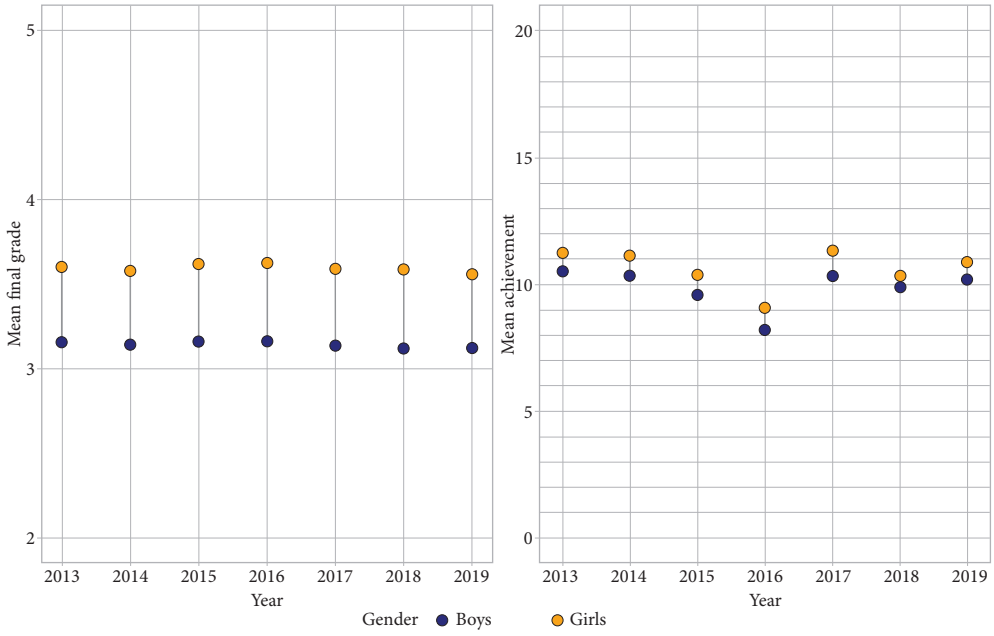
*Note.**p<0.01

TABLE 2. Number of examinees and mean values of the achievement on Mathematics exam for boys and girls, standard deviation of achievement, t-value of the difference, and Cohen's d

Year	Gender	N	Mean	SD	T-value	Cohen's d
2013	Boys	32981	10.49	5.03	19.1*	0.15
	Girls	31433	11.25			
2014	Boys	32452	10.33	5.26	19.0*	0.15
	Girls	30166	11.13			
2015	Boys	33445	9.58	5.03	19.3*	0.15
	Girls	31778	10.34			
2016	Boys	33134	8.18	5.27	21.9*	0.17
	Girls	31609	9.08			
2017	Boys	32019	10.32	4.98	25.0*	0.20
	Girls	30599	11.31			
2018	Boys	32948	9.9	4.28	11.7*	0.09
	Girls	31472	10.3			
2019	Boys	32157	10.19	4.28	20.0*	0.16
	Girls	30621	10.87			

Note. *p<0.01

FIG. 1. Comparison of the final grades in Mathematics and achievement in Mathematics test for boys and girls



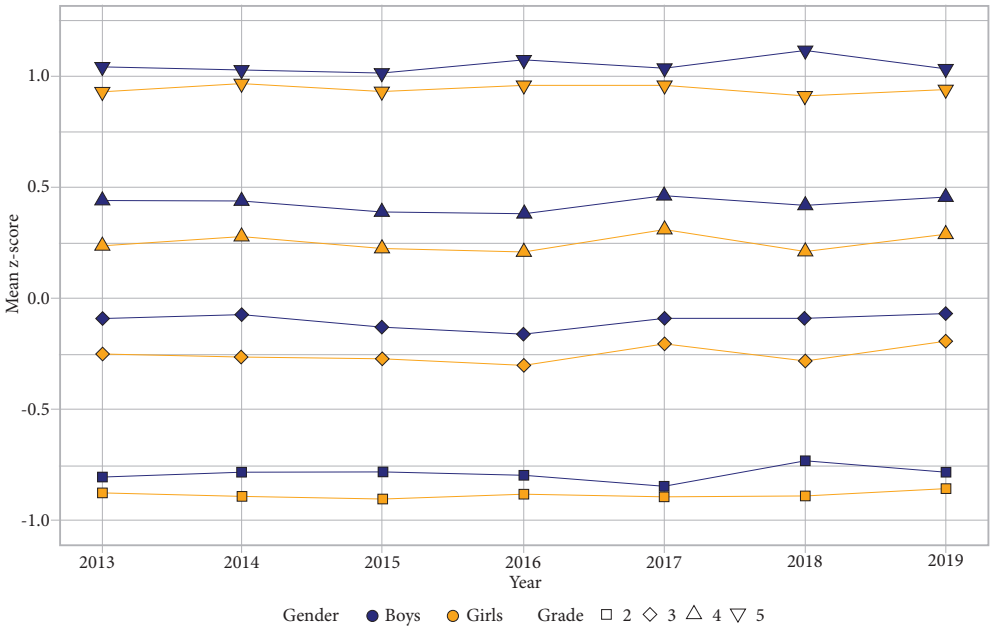
From Fig.1, we can assume that test difficulty is not always the same. For instance, in year 2016, the test was more difficult so the mean results for both boys and girls were lower. For the sake of comparison between tests in different years, as well as comparison with similar studies in the literature, we have transformed achievement to the *z*-score. In Table 3, we can see the mean z-score on the mathematics tests cross-tabulated across examinees' final grades in Mathematics and gender. The standard errors of all these scores are 0.01 or less. Curiously, *z*-scores of boys' achievement were greater than girls' for all final grades and all years. Despite the greater overall achievement of girls, boys had a higher mean achievement for each grade group. This manifestation of Simpson's paradox is a consequence of the great disbalance in the final grade received by boys and girls. Simpson's paradox occurs when a trend appears in several different groups of data but disappears or reverses when the groups are combined. This happens because a hidden or confounding variable influences the results, leading to misleading conclusions if not properly considered.

TABLE 3. Mean z-score on Mathematics exam and the number of students cross-tabulated across to their final grades in Mathematics and gender. Standard errors for all given z-scores are 0.01 or less which makes all displayed gender based differences statistically highly significant

		Mean z-score by final grade			
Year	**Gender**	**2**	**3**	**4**	**5**
2013	Boys	-0.80	-0.09	0.45	1.04
	Girls	-0.87	-0.25	0.24	0.93
2014	Boys	-0.78	-0.07	0.44	1.02
	Girls	-0.89	-0.26	0.28	0.97
2015	Boys	-0.77	-0.12	0.39	1.01
	Girls	-0.90	-0.27	0.23	0.93
2016	Boys	-0.80	-0.16	0.38	1.07
	Girls	-0.88	-0.30	0.22	0.96
2017	Boys	-0.84	-0.08	0.47	1.03
	Girls	-0.89	-0.20	0.31	0.96
2018	Boys	-0.73	-0.08	0.42	1.11
	Girls	-0.89	-0.28	0.21	0.91
2019	Boys	-0.78	-0.07	0.46	1.03
	Girls	-0.85	-0.19	0.29	0.94

Achievement differences between boys and girls (represented as *z*-score) and final grade follow the same pattern in all observed years. These differences are visually presented for all years, in Fig. 2.

FIG. 2. Comparison of examinees' z-scores for all final grades and all testing years.



Unanswered tasks

We hypothesized that boys receiving lower grades might be due to teachers inadvertently assessing and grading noncognitive attributes of students, such as effort, attitude toward school, behavior, and responsibility, which they ideally should not consider.

Among these characteristics, the only one we can directly relate to variables in our testing datasets is effort. Instances of omitting various tasks, i.e., any response to the task, could be interpreted as a lack of motivation or unwillingness to put greater effort into solving mathematical problems. We have found that boys fail to answer test tasks more often than girls. In total, boys omit 24.1% of answers to tasks, while girls omit 18.4%.

As different types of tasks require different levels of effort and ways of response, we expected that open-ended and more complex tasks are more likely to remain unanswered. We have statistically compared the frequency of omitted responses for three types of tasks multiple choice, short answer, and open answer.

Table 4 shows percentages of unanswered and incorrect tasks for three task types for both boys and girls. As one could expect, the frequency of omitting answers for multiple-choice tasks is much lower than for the constructed response task types, especially open-ended tasks. It is also noticeable that girls have fewer unanswered tasks than boys but they also provide more incorrect answers in open-ended tasks.

TABLE 4. Percentages of unanswered tasks per task type (MC – multiple-choice, SA – short answer, OA – open answer) by student gender for all Mathematics tests from 2013 to 2019

	Skipped			Incorrect		
	MC	**SA**	**OA**	**MC**	**SA**	**OA**
Boys	1.20%	6.39%	41.3%	20.7%	20.3%	26.7%
Girls	0.78%	4.62%	31.7%	20.1%	19.1%	30.8%
Cohen's h	0.04	0.08	0.20	0.01	0.03	-0.09

Results show that boys display a significantly higher tendency to disengage compared to girls, and this difference is statistically highly significant. The size effect is most pronounced for open-ended answers, suggesting that the most significant discrepancy in the reluctance to provide answers between boys and girls occurs particularly in tasks requiring open-ended responses. This observation implies that girls are more inclined to attempt answering tasks. This characteristic of boys' behavior could be one of the causes why boys generally exhibit lower achievement levels in Mathematics.

Teachers' opinions

In various educational settings, encompassing both individual classrooms and broader school contexts, the teachers, drawing from their own perspectives, observed no significant disparities in the academic grades between male and female students. This pattern held true in both urban and rural school settings, as illustrated in Table 5. In response to the presented data, which showed higher grades for girls compared to boys, the teachers offered explanations grounded in perceived gender characteristics. They noted that girls generally exhibit greater diligence, responsibility, and meticulousness in their academic work. Girls were observed to practice more for tests and meticulously check their solutions, but tended to take fewer risks and showed a greater fear of making mistakes. Conversely, the teachers

observed that boys are more inclined to engage in competition, perform better with technology, and are more likely to take risks. However, they also noted that boys generally prepare less for tests and are more prone to give up quickly. The teachers also discussed the impact of disciplinary measures on boys' grades, noting occasions where the establishment of discipline through assessment resulted in lower grades for boys (Table 5).

The teachers collectively agreed that the observed lower academic performance of boys compared to girls does not significantly influence the students' self-perception or their opportunities for enrolling in secondary education, as detailed in Table 6. Additionally, they noted that these variations in grades do not seem to have a lasting impact on the student's future academic success.

TABLE 5. Teachers' observation – behavior of boys and girls

Teachers' observation	Teachers' explanation
No difference	I don't have the impression that girls have higher grades. It changes from year to year. My school is small, rural. Maybe the sample is small. In our school (city school) there is no such difference. They have the same grades. With boys, it's probably forced, it's an external motivation for them.
Characteristics of girls	Girls are more persistent, more diligent, work to the end, and check solutions. Girls are more obedient, more meticulous, and more responsible. The attitude towards work is different. The girls will practice for the test. Girls are afraid of making mistakes.
Characteristics of boys	Boys are more competitive. They prefer to compete. Boys don't prepare so they make trivial mistakes. Boys are dominant in relating to technology. When doing some application of mathematics, they are better. We often get top results from boys, but they often give up. More boys show up to get on the board, take more risks, and are more relaxed. In our school, every year a boy is a student of the generation who won something in a math competition.
Punishment	Enforcement of discipline brings bad grades. When someone is restless, they bring him to the board, and then he gets a bad grade. These are mostly boys.

TABLE 6. Teachers' observations –lower grades of boys and consequences

Teachers' observations	Teachers' explanation
A lower grade does not affect students	I think high school students have no idea what grade they got in eighth grade and don't really care. I believe that lower grades have no consequences. They don't care.
A lower grade does not affect school enrollment	Everyone enrolls a school, and everyone goes on regardless of the grade. It doesn't affect them much. In Vrbas, everyone enrolls in the mathematics major. And those with twos. There are always places in high schools. Except in Belgrade.
A lower grade does not affect future achievements	It is not decisive what they knew and had as a grade at the age of 15. I have a colleague who got his doctorate in mathematics in Ghent, and he had a three in high school.

DISCUSSION

Grading gap

The findings of this study indicate that girls outperform boys in mathematics, as evidenced by their superior average grades in mathematics and higher average achievements on the mathematics test in the final exam (see Table 1, Table 2, and Fig. 1). However, the disparity in average grades between boys and girls is greater than the difference observed in the final exam scores (Table 1, Table 2, and Fig. 1). This phenomenon suggests the potential existence of a grading bias in favor of girls. The observation that girls, on average, achieve higher scores in the final exam contrasts with findings from other standardized tests, such as the PISA, where boys from Serbia outperform girls in mathematics, but that difference is not always statistically significant (OECD, 2023; Videnović & Čaprić, 2020). Additionally, in the present study, the effect size for average achievements on the mathematics test ranged from 0.15 to 0.20 standard deviations. This finding diverges from the results presented by Protivínský and Münich (2018), who identified a gender difference of approximately 0.29 standard deviations, but in favor of boys. Furthermore, our findings contrast with those reported by Machin and Pekkarinen (2008), who, upon analyzing global data from the Programme for International Student Assessment (PISA), observed that 15-year-old boys outperformed girls in mathematics by a margin of 0.10 standard deviation.

Further analysis by grades corroborated the presence of a grading bias, revealing that boys outperform girls in the mathematics test of the final exam among students with equivalent grades (Table 3). This evidence, illustrating Simpson's paradox, further underscores the existence of a grading bias favoring girls. The obtained results align with the findings of other studies (Di Liberto et al., 2022; Graetz & Karimi, 2022; Protivínský & Münich, 2018; Terrier, 2020) but diverge from those of Lavy and Sand (2015), who identified a grading gap favoring boys, albeit through a slightly different methodology.

The findings of this study reveal that the grading gap transcends any single cohort of students and manifests consistently year after year. Analysis of data spanning seven years uncovers a recurring pattern: on average, girls receive higher grades and overall, on average, outperform boys on the final mathematics exam. Yet, among students with equivalent grades, boys achieve higher scores on the final exam than girls. This discussion highlights the grading gap's enduring nature, a phenomenon that persists over consecutive years. Prior research has investigated the factors contributing to this gap, identifying variables such as the intensity of competition during evaluations, teacher-student interaction dynamics, and the impact of non-cognitive skills on grading (Falch & Naper, 2013; Lavy, 2008). Other influences include the degree of anonymity in assessments and curriculum emphasis, which may highlight different competencies and variations in the interaction or timing of exams (Lavy, 2008). Diverging from previous studies, this research focuses on teachers' perspectives to understand their views on the grading gap and its underlying causes.

How do teachers perceive the grading gap?

While the results of this study reveal an evident grading gap, teachers often do not perceive this discrepancy in their teaching practices or within their schools. They tend to attribute the grading gap to student characteristics rather than teacher bias. Teachers recognize different student characteristics, which they believe contribute to a distribution of grades where girls tend to receive higher grades (Li, 1999). The observations of teachers align somewhat with those of studies indicating that boys outperform girls in competitive environments and under increased stress (Gneezy et al., 2003; Ors et al., 2013). Furthermore, these observations are consistent with the findings of Jurajda and Münich (2011), which showed that girls often exhibit

poorer performance in high-risk tests. However, a study by Falch and Naper (2013) in Norway investigated the presence of a grading gap across various types of testing, extending beyond the comparison between teacher-given grades and high-risk tests. That study, along with the findings of Lavy (2008) and Matějů and Smith (2014), confirmed that the grading gap is equally prevalent in both high-risk and low-risk tests, as well as in anonymous testing. Consequently, based on these studies, the impact of differences in the competitiveness of the assessment environment on the grading gap remains questionable (Protivínský & Münich, 2018).

Teachers have noted that qualities such as diligence, persistence, dedication, meticulousness, and the thorough verification of solutions are more pronounced in girls, which could potentially influence their academic achievements. This study analyzed the number of tasks left unanswered by students, categorized by task type. Findings indicate that girls consistently attempt answers across various task types, with a statistically significant difference evident in each category. The most noticeable disparity in task non-response is observed in open-ended tasks, where girls demonstrate a significantly higher propensity to engage. This finding is to some extent aligned with studies (Beller & Gafni, 2000; Graetz & Karimi, 2022; Riener & Wagner, 2018; Wester & Henriksson, 2000) indicating that girls tend to be more successful in open-ended tasks. This propensity aligns with teacher observations that girls exhibit greater persistence and diligence, making more effort to solve tasks even when unsure of the answers. These observations contrast with findings from Von Schrader and Ansley (2006) and Riener and Wagner (2017), who reported lower rates of omitted answers among boys in mathematics tasks. The key insight is the persistence of girls in attempting to answer tasks, regardless of their certainty, that stands out. This trait among girls is identified as a factor that may contribute to the grading gap.

The teachers involved in this research expressed the belief that lower grades for boys will not negatively affect them, their high school enrollment, or their future accomplishments. They attribute this perspective to the fact that there are occasionally vacancies in some high schools. This opinion of teachers contrasts with the findings of studies Lavy and Sand (2015), and Federičová (2016), which demonstrate that grades play a significant role in shaping students' perceptions of their own achievements. Therefore, bias in teachers' grading can impact students' future lives in various ways. Students and parents typically view grades as accurate feedback on a student's academic performance and progress, and they rely on this

information to make decisions about the student's educational trajectory and career. If such feedback is biased, these decisions may be suboptimal, potentially leading to labor market inefficiencies (Protivínský & Münich, 2018). In Serbia, there is relatively low interest in enrollment at grammar schools, often resulting in vacancies within these institutions. However, in many schools high admission scores are mandatory where mathematics grades play an important role. Furthermore, overall academic performance, including mathematics grades, significantly influences access to student dormitories and scholarships.

CONCLUSIONS

This study analyzed student grades and their performance on the final mathematics exam. Initially, the observation that girls, on average, received higher grades and performed better on the final exam did not appear unusual. However, a more detailed analysis uncovered a significantly larger discrepancy in grades between boys and girls than in their average exam scores, indicating a grading gap that favors girls. The subsequent grade-level analysis confirmed a grading bias, showing that among students with the same grades, boys surpass girls in mathematics performance in the final exam. This phenomenon, consistent with previous years, suggests a persistent grading gap within the Serbian education system. Numerous studies have explored the potential causes of this gap. As a novel contribution, this research engaged a focus group of teachers to delve into their perceptions of the gender gap. Despite the identification of a grading gap, teachers largely did not recognize its existence, attributing differences to intrinsic characteristics of boys and girls, which indeed could contribute to the observed gap. One notable finding from this study is that girls are more inclined to attempt an answer to a problem, even without knowing the solution. The teachers articulated a somewhat unconventional viewpoint, suggesting that marginally lower grades for boys might not have any significant impact on them. This perspective overlooks the well-documented influence of grades on opportunities such as high school and college admissions, scholarship eligibility, residence hall placements, and the overall perception of student capabilities. This finding highlights the importance of closely examining teachers' attitudes and increasing awareness of the role and implications of grading practices. Although the qualitative part of this study provided new insights, it is subject to certain limitations.

These include the limited number of teachers who participated in the focus group and the time gap between the qualitative and quantitative phases of the research. Such limitations suggest the need for future studies to further explore, confirm, or challenge the findings presented in this study.

The current study raises the critical issue of the reliability of teacher-assigned grades versus scores from final examinations. Currently, both components are integral to evaluating students' eligibility for further education. The study, along with the broader educational framework, encounters limitations in assessing whether the content tested by teachers aligns with that of the final exam. Furthermore, it questions the comparability of continuous assessment, which encompasses several months of student work and effort, with a final exam that primarily assesses mathematical knowledge. It becomes clear that a grade in mathematics encompasses more than mere knowledge; it reflects a broad spectrum of student characteristics and behaviors. Despite the observable differences between teacher grades and final exam scores, the debate persists on whether these evaluation metrics should be harmonized, considering they assess distinct dimensions of student learning and achievement.

Competing interests. The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. No funding.

Note. This research was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Grants No. 451-03-65/2024-03/200125, 451-03-65/2024-03/200156, 451-03-65/2024-03/200102, 451-03-137/2025-03/200251, 451-03-137/2025-03/200102, 451-03-65/2025-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad" (No. 01-3394/1).

REFERENCES

- Adler, K., Salanterä, S., & Zumstein-Shaha, M. (2019). Focus group interviews in child, youth, and parent research: An integrative literature review. *International Journal of Qualitative Methods, 18*, 1–15. DOI: 10.1177/1609406919887274
- Avitzour, E., Choen, A., Joel, D., & Lavy, V. (2020). On the origins of gender-biased behavior: The role of explicit and implicit stereotypes (NBER Working Paper No. 27818). *National Bureau of Economic Research*. DOI:10.3386/w27818
- Becker, D. F., & Forsyth, R. A. (1990, April). Gender differences in academic achievement in grades 3 through 12: A longitudinal analysis. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.
- Beller, M., & Gafni, N. (2000). Can item format (Multiple Choice vs. Open-Ended) Account for gender differences in mathematics achievement? *Sex Roles*, *42*, 1–21. DOI:10.1023/A:1007051109754
- Birenbaum, M., & Feldman, R. A. (1998). Relationships between learning patterns and attitudes towards two assessment formats. *Educational Research*, *40*(1), 90–98. DOI:10.1080/0013188980400109
- Borghans L. Golsteyn B. H. Heckman J. J., & Humphries J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences*, *113*(47) 13354–13359. DOI:10.1073/pnas.1601135113
- Đerić, I. Gutvajn, N., Jošić, S. & Ševa, N. (Eds.). *TIMSS 2019 in Serbia*. Belgrade: Institute for Educational Research.
- Dersch, A. S., Heyder, A., & Eitel, A. (2022). Exploring the nature of teachers' math-gender stereotypes: The math-gender misconception questionnaire. *Frontiers in Psychology, 13*, Article 820254. DOI:10.3389/fpsyg.2022.820254
- Di Liberto, A., Casula, L., & Pau, S. (2022). Grading practices, gender bias and educational outcomes: evidence from Italy. *Education Economics*, *30*(5), 481–508. DOI:10.1080/09645292.2021.2004999
- Falch, T., & Naper, L. R. (2013). Educational evaluation schemes and gender gaps in student achievement. *Economics of Education Review*, *36*, 12–25.
- Federičová, M. (2016). Gender gap in application to selective schools: Are grades a good signal? (CERGE-EI Working Paper No. 550).
- Ferman, B., & Fontes, L. F. (2022). Assessing knowledge or classroom behavior? Evidence of teachers' grading bias. *Journal of Public Economics, 216*, Article 104773. DOI:10.1016/j.jpubeco.2022.104773
- Garner, M., & Engelhard, G., Jr. (1999). Gender differences in performance on multiple-choice and constructed response mathematics. *Applied Measurement in Education, 12*(1), 29–51.
- Giofrè, D., Cornoldi, C., Martini, A., & Toffalini, E. (2020). A population level analysis of the gender gap in mathematics: Results on over 13 million children using the INVALSI dataset. *Intelligence, 81*, 101467. DOI:10.1016/j.intell.2020.101467
- Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics, 118*(3), 1049–1074.
- Graetz, G., & Karimi, A. (2022). Gender gap variation across assessment types: Explanations and implications. *Economics of Education Review*, *91*, 102313. DOI:10.1016/j.econedurev.2022.102313
- Groth, R. E. (2006). Analysis of an online case discussion about teaching stochastics. *Mathematics Teacher Education and Development, 7*, 53–71.
- Guez, A., Peyre, H., & Ramus, F. (2020). Sex differences in academic achievement are modulated by evaluation type. *Learning and Individual Differences, 83–84*, 101935. DOI:10.1016/j.lindif.2020.101935
- Guttmann, J., & Boudo, M. (1988). Teachers' evaluations of pupils' performance as a function of pupils' sex, family type and past school performance. *Educational Review, 40*(1), 105–113. DOI:10.1080/0013191880400108

- Han, L., & Hoover, H. D. (1994, April). Gender differences in achievement test scores. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy, 4*(4), 146–168.
- Hanushek, E. A. & Woessmann L. (2012). Do better schools lead to more growth? Cognitive skills economic outcomes and causation. *Journal of Economic Growth*, *17*(4) 267–321. DOI:10.1007/s10887-012-9081-x
- Hinnerich, B. T., Höglin, E., & Johannesson, M. (2011). Are boys discriminated in Swedish high schools? *Economics of Education Review, 30*(4), 682–690.
- Holder, K., & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: A new look from a shifting standards perspective. *Social Psychology of Education, 20*, 471–490.
- Jurajda, Š., & Münich, D. (2011). Gender gap in performance under competitive pressure: Admissions to Czech universities. *The American Economic Review, 101*(3), 514–518.
- Lauermann, F., Meißner A., & Steinmayr R. (2020). Relative importance of intelligence and ability self-concept in predicting test performance and school grades in the math and language arts domains. *Journal of Educational Psychology*, *112*(2), 364. DOI:10.1037/edu0000377
- Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics, 92*(10), 2083–2105.
- Lavy, V., & Sand, E. (2015). On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases (NBER Working Paper No. 20909). National Bureau of Economic Research. https://www.nber.org/papers/w20909
- Li, Q. (1999). Teachers' beliefs and gender differences in mathematics: A review. *Educational Research*, *41*(1), 63–76. DOI:10.1080/0013188990410106
- Lindner, J., Makarova, E., Bernhard, D., & Brovelli, D. (2022). Toward gender equality in education – Teachers' beliefs about gender and math. *Education Sciences, 12*(6), Article 373. DOI:10.3390/educsci12060373
- Lindqvist, A.-K., Weurlander, M., Wernerson, A., & Thornberg, R. (2020). A focus group interview study of the experience of stress amongst Swedish schoolchildren. *International Journal of Environmental Research and Public Health, 17*(11), 4021. DOI:10.3390/ijerph17114021
- Machin, S., & Pekkarinen, T. (2008). Global sex differences in test score variability. *Science, 332*(5906), 1331–1332.
- Marcenaro-Gutierrez, O. D., Prieto-Latorre, C., & Sánchez Rodriguez, M. I. (2023). Gender differences between teachers' assessments and test-based assessments: Evidence from Spain. *Assessment in Education: Principles, Policy & Practice*. DOI:10.1080/0969594X.2023.2251715
- Matějů, P., & Smith, M. (2014). *Are boys that bad? Gender gaps in measured skills, grades and aspirations in Czech elementary schools*. British Journal of Sociology of Education, *36*(6), 871–895.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *Average achievement by gender – TIMSS 2019 International results in mathematics and science*. TIMSS & PIRLS International Study Center, Boston College.
- Newell, G., Wilsman, M., Langenfeld, M., & McIntosh, A. (2002). Online professional development: Sustained learning with friends. *Teaching Children Mathematics, 8*, 505–508.
- OECD. (2015). *The ABC of Gender Equality in Education*. OECD. DOI: 10.1787/9789264229945-en
- OECD. (2023). Serbia student performance (PISA 2022). Education. https://gpseducation.oecd.org
- Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: Does competition matter? *Journal of Labor Economics, 31*(3), 443–499.

- Protivínský, T., & Münich, D. (2018). Gender bias in teachers' grading: What is in the grade. *Studies in Educational Evaluation, 59*, 141–149. DOI:10.1016/j.stueduc.2018.07.006
- Rakshit, S., & Sahoo, S. (2023). Biased teachers and gender gap in learning outcomes: Evidence from India. *Journal of Development Economics, 161*, Article 103041. DOI:10.1016/j.jdeveco.2022.103041
- Riener, G., & Wagner, V. (2017). Shying away from demanding tasks? Experimental evidence on gender differences in answering multiple-choice questions. *Economics of Education Review*, *59*, 43–62. DOI:10.1016/j.econedurev.2017.06.005
- Riener, G., & Wagner, V. (2018). Gender differences in willingness to compete and answering multiple-choice questions – The role of age. *Economics Letters*, *164*, 86–89. DOI:10.1016/j.econlet.2018.01.012
- Sumpter, L. (2016). Investigating upper secondary school teachers' conceptions: is mathematical reasoning considered gendered? *International Journal of Science and Mathematics Education*, *14*, 347–362. DOI: 10.1007/s10763-015-9634-5
- Terrier, C. (2020). Boys lag behind: How teachers' gender biases affect student achievement. *Economics of Education Review,* 77, 101981. DOI:10.1016/j.econedurev.2020.101981
- Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. *Educational Studies in Mathematics, 50*, 49–62. DOI:10.1023/A:1020518104346
- Videnović, M., & Čaprić, G. (2020). *PISA report for 2018, for the Republic of Serbia*. Ministry of Education, Science and Technological Development. http://www.obrazovanje.org/rs/uploaded/dokumenta/PISA-2018_Izvestaj-za-Republiku-Srbiju_ceo.pdf
- Von Schrader, S., & Ansley, T. (2006). Sex differences in the tendency to omit items on multiple-choice tests: 1980–2000. *Applied Measurement in Education*, *19*(1), 41–65.
- Wester, A., & Henriksson, W. (2000). The interaction between item format and gender differences in mathematics performance based on TIMSS data. *Studies in Educational Evaluation* , *26*(1), 79–90. DOI:10.1016/s0191-491x(00)00007-9
- Willingham, W. W., Pollack, J. M. & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, *39*(1) 1–37. DOI:10.1111/j.1745-3984.2002.tb01133.x
- Zanga, G., & De Gioannis, E. (2023). Discrimination in grading: A scoping review of studies on teachers' discrimination in school. *Studies in Educational Evaluation*, *78*, 101284.