DE GRUYTER
OPEN

STUDIES IN LOGIC, GRAMMAR
AND RHETORIC 50 (63) 2017
DOI: 10.1515/slgr-2017-0025

Alicja Zawistowska
University of Bialystok

GENDER DIFFERENCES IN HIGH-STAKES MATHS TESTING. FINDINGS FROM POLAND

Abstract. The present research investigates gender gaps in the results of secondary school exit exams (Matura) in mathematics in Poland in 2015. The analysis shows that, in the basic level exam, males are highly overrepresented at the upper end of the score distribution. The same pattern did not exist in the extended-level Matura. Two explanations are offered here. The differences are driven by gender self-selection in high school programs. Students who decide on maths-related tracks have more maths lessons than other students. Secondly, a student who takes the extended Matura also has to take the basic Matura exam. As a result, the population of students taking the basic Matura is highly differentiated in terms of maths competence and motivation. Additionally, the analysis of differential item functioning (DIF) shows that only a few items were flagged as having DIF.

Keywords: maths, gender, Poland, Matura, DIF, Mantel-Haenszel.

School tests are expected to properly reflect the levels of skills of the students taking them. Students who have the same abilities should have the same chances to give the correct answers to test questions, regardless of sex, skin colour, or social background. However, in reality there are many factors which may potentially interfere with test results. Some of them concern the student, his/her mood, motivation, and pre-examination stress levels, as well as the construction of the test itself. This last includes the choice of questions, their format, or the the effect of teachers through different teachers making differing assessments of the same answers (Pokropek, Jakubowski 2009). The necessity to preserve the high quality of the didactic measurement is especially important in relation to high-stakes tests – that is, those which have a crucial impact on the next stage of education. In the Polish system of education, Matura is such an examination. Its results determine the chances of continuing an education in tertiary institutions, and indirectly influence the future professional career. It is obvious that this will be the case, based on the assumption that Matura results are a good

ISBN 978-83-7431-527-2 ISSN 0860-150X

predictor of future achievements. However, irrespective of how much the scores of that examination influence the future of the individual, they are a source of information on the general abilities and preferences of a student.

One of the factors which crucially differentiates the school test score is sex. This matter raises particular controversy with respect to mathematics, which is considered to be a discipline with great influence on the social position an individual will achieve in adult life. It is these disparities in that discipline – their scale and causes – that have long been the subject of research among sociologists, pedagogues, and psychologists. The attention devoted to this issue is significantly greater than that devoted to, for example, language competence, which also presents great difference between the sexes.

Since the moment when, in 2010, the Matura exam became compulsory for all students in Poland, it has been possible to follow the results of the examination with regard to sex differences. In this article I shall analyze the results of the mathematics scores of the Matura examination in 2015, at both the basic and extended level. I shall limit my analysis to students who took the "new" Matura examination, prepared on the basis of the core curriculum introduced in 2012. These were mostly students at general secondary schools. In comparison to the "old" curriculum, these changes introduced or removed some elements of the curriculum and described the detailed abilities the student who graduates from secondary school should possess.

At first, I shall compare the percentage of females and males at particular levels of results, and then I shall conduct an analysis of the differential item functioning of test questions (DIF) using the Mantel-Haenszel method.

Differentiation of scores of mathematical tests

The research on female and male scores in mathematical tests may be divided into two main categories. The first includes studies in which there are basic results distribution measurements achieved by men and women. Usually, these are limited to average and variance (Halpern 2012, Lindberg et al. 2010). Such research does not provide unambiguous results. Some suggest that there has been a gradual blurring of distinctions between the sexes within recent decades, while others state that there are slight differences in favour of one sex. On the other hand, some argue that – at least in some countries – there are no crucial variations between the sexes (Lindberg et al. 2010, Kenney-Benson 2006). However, cross-case analysis conducted at the international scale shows that average scores in most countries are still systematically higher for boys (PISA 2012). The situation among Polish students is not yet well known. Gruniewska and Kondratka's research (2012) on average scores in the mathematical-science part of junior-high school tests show that the differences between the sexes were statistically crucial in favour of girls in 2009–2011, while in 2002–2008 they were lower.

Aside from this, one may wonder whether the reversion of that trend might have been connected to the introduction in 2008 of changes to the core curriculum in preparation for the introduction in 2010 of the compulsory Matura in mathematics.

The second group includes studies analyzing the proportion or ratio of females to males on the edges of the result distribution, and especially among students with the highest scores (Stoet, Geary 2013; Ellison & Swanson 2009). These show the systematic predominance of boys among the best 1% or 5% of students and laureates of maths-subject contests (Zawistowska 2013). It is interesting that the size of the gap on the upper edge of the distribution is not the same over a student's entire educational career; instead, it widens with age and the progressing complexity of mathematical problems. As a result, there are greater disproportions among the participants of more advanced courses than in primary education. (Halpern 2013). This may be the result of many, varying, factors. Hyde and Mertz (2009), who analyzed data from the International Mathematical School Contests stress the role of institutional and cultural aspects. Paradoxically, in these competitions a relatively high number of female participants came from countries where there is a rather traditional model of sex roles, e.g. China, S. Korea, or Bulgaria. Another example – of the impact of institutional regulations on the decrease of disproportions – is the USA. There, for the first 23 years of the Mathematical Competition, the country did not have a single female among its representatives. It changed only in the beginning of the 70s when STEM, the special governmental system of development in secondary schools, was introduced (Hyde and Mertz 2009).

Another indicator in gender gap analysis in maths is school grades. Here, the results of research are quite unequivocal and show the predominance of girls in different subjects, including mathematics (King et al.; Duckworth and Seligman 2006). The case is similar among Polish junior-high school students (Skórska, Świst 2014). In mathematics this difference in favour of girls was lower than for Polish language and other subjects, yet it was statistically significant.

Generally speaking, girls have higher grades at school, yet they have weaker scores in standarised tests. This effect is called the "female underprediction effect" (King et al. 2012). This term was initially used for the American SAT test's predictive power with regard to grade point average. It was noticed that girls' school grades were higher than their SAT scores had predicted. King et al. (2012) suggest that this inflation results from gender differences in conscientiousness. That is to say, it was argued that women were more conscientious, and at the same time, that students who were more conscientious earn higher grades than their SAT score would predict. Carefulness, conscientiousness or diligence turn out to be closely correlated with high school grades, yet they do not have the same effect in tests, and even indirectly lower test scores (Duckworth, Seligman 2006). This discrepancy may result from the different atmospheres when assessing a student in class and in an exam. It was stated several times that not only abilities but also teacher preferences have an influence on school grades. In one of the studies on that subject, teachers assessed the mathematical abilities of girls more highly, but this was changed when the analysis monitored assessments of students' behaviour and previous achievements. The abilities of boys who behaved and had the same scores as girls were assessed higher (Robinson-Cimpian et al. 2013). It is interesting that this correlation existed only for the relation between sex and mathematics, not for any other subjects or individual features of students, such as skin colour. These results may be explained using the famous "Pygmalion effect", according to which, the higher performance of students is a response to the higher expectations of the teacher. The existence of a conviction that boys have better natural abilities will be reflected in them being presented more challenging tasks, which will result in higher abilities. On the other hand, girls' "good behaviour" may have a greater impact on their assessment. One may wonder whether this effect might be restricted, at least partially, if the ability to make direct comparisons between students of different sexes disappeared. The ideal environment to check this is single-sex schools. However, the example of Ireland, where approximately 1/4 attend such schools, shows that differences between the sexes in mathematics, and especially at the upper edge of the distribution, are even larger. Boys in single-sex schools had higher scores than their peers in co-educated schools, yet the girls did not have the same benefits (O'Neill, Sweetman 2013).

Assessing in the classroom significantly differs from that in exams. The element of subjective assessment disappears, while the stress and awareness of test consequence appears (pressure). The research concerning the intensity of examination anxiety shows that women suffer from stress more. Kosmala-Anderson (2006) proves that among Matura students she studied taking the final Matura exam and taking the mock exam, females experienced diverse psycho-somatic reactions to stress in greater numbers and more intensely. In her opinion, the explanation for these differences was a greater physiological reactivity in women, and different ways of processing information related to stress situations (Kosmala-Anderson 2006). Perhaps women assigned greater importance to the examination, excessively focused on the consequences, or feared failure more. Nevertheless, one may presume that the level of anxiety shall have a different impact on the scores of persons with higher abilities (to a lesser extent) and lower abilities (to a greater extent).

The exam is also connected with competition and risk-taking. Women prefer less competitive situations, even if that means receiving lesser benefits. Among others, Niederle and Vesterlund (2007) came to such conclusions within a series of research conducted on the effectiveness of both sexes in performing different tasks in competitive contexts.

In one research, a group of women and men solved a simple mathematical task in two experimental conditions: under piece-rate compensation and in a tournament. After experiencing both schemas and receiving feedback on whether their answer was correct or incorrect, participants chose which schema they wanted to apply for in the next round. If they selected the piece-rate condition they received 50 cents per correct answer. In the tournament condition only the person with the largest number of correct answers received 2 dollars, while the other participants received no payment. This study showed that twice as many men as women selected the tournament, even after controlling for previous performances. Niederle and Vesterlund (2007, 2010) explained that the reason was a higher overconfidence about relative performance among men compared to women.

On a more general level, these findings signal that individual features such as self-confidence, inclination to risk or risk aversion, and score anxiety along with cognitive abilities, influence test scores. Some research even suggests that risk aversion may explain approx. 40% of the difference in the mathematical SAT test (Tannenbaum 2012). Individual differences between the sexes may manifest in the smaller tendency of women to undertake more difficult tasks, lesser perseverance, or resignation from participation in more selective mathematical courses (see Weinberger 2005).

Aversion to competitive situations might be an important intermediate factor in exam performances, especially those with high stakes. A very illustrative example of that phenomenon is provided by Ors, Palomino and

Peyrache (2013). To determine the influence of high pressure on performance, they utilised real-world data from the entrance examination to a highly selective French business school (ranked as the first in Europe in its category). As a point of reference they compare the results of entrance exams with the performance from two less competitive exams for the same group of individuals. One of the exams was a *baccalauréat*, which is taken before the entrance exam, and the other was taken in the first year of core courses. The study shows that women performed worse than men on the selective exam, but they performed as well or even better on two other less competitive exams (Ors, Palomino and Peyrache 2013).

To sum up all the quoted results, one may presume that to a certain extent the cause of females' lower test scores is a lower ability to cope with the examination itself. Lesser confidence in ability or their own self-efficacy, are not allies of effective exam performance.

Do women react differently to test questions?

It is also suggested that the form itself of the standarised test favours boys, through contrary reactions to test items; there are hypotheses that boys better cope with multiple-choice items, while girls are better at open-ended questions, fill-in items and essays. The first occur more frequently in standarised tests, which may be the source of the additional points scored by boys. Girls' dominance in open-ended questions is usually explained by higher verbal competence. It is also pointed out that women have a lesser tendency to guess answers in a test and a greater tendency to omit items. These differences explain only a small part of the variances in differences between the sexes in tests (Beller, Gafni 1996). What is more, these findings are not universal. From the analysis conducted for Poland, it transpires that boys are more prone to omit multiple-choice questions than girls (Świst et. al 2015). This pattern exists even after controlling for ability, and is true for both humanities and science exit exams. The difference between genders is small, but statistically significant. Świst et. al (2015) claim that the omissions pattern might to some degree be caused by the test administration procedure in Polish exit exams. After solving the test, students have to mark answers on a separate sheet of paper. If boys are less conscientious, they may unintentionally omit some boxes during rewriting or skipping from one question to another. However, only 2–3% of students omitted at least one question, and further investigation is needed to explain the nature of this phenomenon.

Other studies show that differences in test scores are influenced by the vocabulary used in questions. Tasks whose content is related to situations that are characteristic for boys, for example competitive sport games, (Loewen, Rosser, and Katzman 1988) favour this sex. Other regularities concern the particular subject of mathematics (Romains). For example, it is proven that girls do better at items connected with calculations, symbols, or questions referring to social issues, while boys have better scores at questions including tables, figures, or proportions (Bennet 1993). The influence of these factors on the size of the gender gap depends on the characteristics of a given test, so it is hard to make generalisations here.

Apart from the aforementioned hypothesis, there are many others explaining the gender gap in mathematics from the cultural perspective. Nevertheless, the ability to reliably verify and prove a cause and effect dependency remains the weak point of many of them. For example, Frey and Levitt (2009) tested hypotheses of different socialisation of boys and girls in several ways, but they did not show any support for them. While it turned out to be true that parents have lower mathematical expectations towards their daughters, using this variable as a factor explaining their lower scores did not have any impact. Furthermore, it is still a challenge to create a research plan which would properly detect the influence of nuanced cultural factors in a strictly statistical model.

There is no unanimous opinion as to the influence of stereotype threat on test scores. Although within the last years it is a popular hypothesis, some scientists remain sceptical about treating it as a direct cause of lower scores or the auto-selection of women to more advanced mathematical courses. For example, Stoet and Geary (2012), on the basis of a meta-analysis, argue that less than half the studies on that subject do not unanimously confirm the existence of that effect, due to, among other things, deficiencies in methodology. However, it does not mean that stereotypes do not influence the lowering of scores in a more indirect manner, for example through the building of maths self-esteem.

Data used for the analysis

The analysis in the further part of the article covers students of secondary school who took the "New Matura" in 2015. The exam covered students of post-gymnasium schools and artistic ones, who graduated from school in the school year 2015/2016, and graduates of the general secondary schools, who graduated from school in the school year 2014/2015. The population did not include graduates who in 2015 took the Matura exam according to the phased out curriculum (almost all students of secondary technical schools)

Data used for the analysis was made available in the public domain by the Educational Research Institute [Instytut Badań Edukacyjnych] (Szaleniec et al. 2015). They constitute the combined sets of national institutions, that is, the Central Examination Board [Centralna Komisja Egzaminacyjna] and Regional Examination Boards [Regionalna Komisja Egzaminacyjna], concerned with the organisation of the exam. To download them I used a dedicated package "zpd" supported by R (Szaleniec et al. 2015).

The size of the population of students taking Matura exam as presented by CKE and that available in the zpd database slightly differ from one another, which is a result of the complex process of gaining and combining databases. According to the official results, the population of students taking basic Matura exam in Mathematics counted in 2015 more than 177 000 people, while only 50 000 students took the extended one. After excluding the lack of data in test scores, the set on which I did my calculations was slightly smaller. The gender ratio in the zpd database was also concurrent with the data given by CKE, and in the case of basic Matura it was 63% women and 37% men.

Matura exam

The Matura exam is taken at the end of a three-year period of education in secondary school. The Matura exam in Mathematics can be taken at two levels – basic and extended. Since 2010, the basic level exam has been a compulsory subject for all students who want to pass Matura and to start tertiary education. The extended Matura is optional, and students can decide for themselves if they want to take it or not. This examination is necessary to start education on most engineering majors and some nature-oriented ones. Since 2012, after graduating from the first grade of secondary school finishing with the Matura exam – students have been obliged to choose at least 2 subjects, which they are going to take at the extended level in the Matura. The choice of profiles may differ between schools, yet in each school one may choose among the more mathematical (or polytechnical), nature-oriented and humanities specialisations. The choice of profile is connected with an increased number of lessons in a given subject or teaching module, to the detriment of other subjects. For example, students who decide on exact science or polytechnical profiles have more classes in physics, and mathematics, and less in history or civic studies. In the case of Mathematics, 300 hours are designed for teaching this subject on the basic level and an additional 180 hours for the extended level. Nonetheless, even the students who are taking the extended Matura exam must also take the basic (as a compulsory exam for all students). In both tests, the maximum score was 50 points, and the pass threshold was 30%. The questions were in the form of multiple-choice and open-ended questions. Both tests are anonymised and checked by external examiners. The examination sheets from previous years are available to the public.

The changes in the Matura exam concordant with the new core curriculum introduced in 2012 mainly concerned the extended level. In accordance with the expectations of CKE, this test was, to a greater extent, to check the ability of students to understand mathematical notions and to create their own strategies to solve atypical tasks.

Results of New Matura 2015 by gender

Due to the large disproportion of sexes in the examined population, the analysis shows the percentage of women and men (separately) who achieved a given score. Figure 1 shows then the percentage of women out of all women and percentage of men out of all men who achieved a given score. The critical difference between the sexes is clearly seen in the final part of the distribution – the percentage of females decreases while the percentage of males rapidly increases. Approximately 1.5% of all female secondary-school graduates achieved the maximum score, in comparison to 3.8% of males. The number of males was relatively smaller in the medium part of the distribution. The size of the gap, as well as the fact that it concerns basic Matura, seems surprisingly significant. According to the studies previously referred to, in less advanced tests this gap should be small, or not present at all. The organisation of education in secondary schools, precisely the existence of profiles and the rules for passing Matura exam, are responsible to a great extent for the existence of this gap. As mentioned above, basic Matura in mathematics is compulsory for all students, so students who have decided to take the extended level take it as well. One may suppose that that the size of the "fork" on the right side of the distribution is strengthened by the results of students for whom the basic Matura is just a trial before the more important exam at the extended level. Thus, using average or variance as a measure of gender gap would be misleading here.

Figure 1. Percentage of males and females on given score level on basic Matura exam 2015

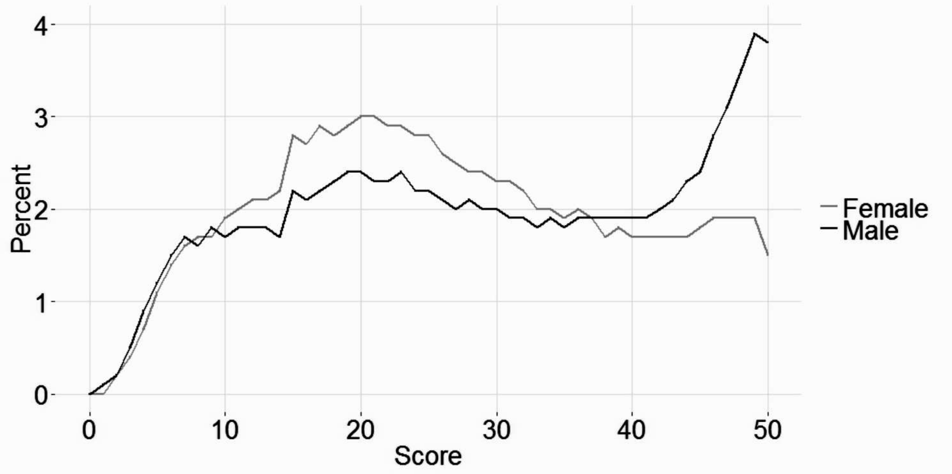
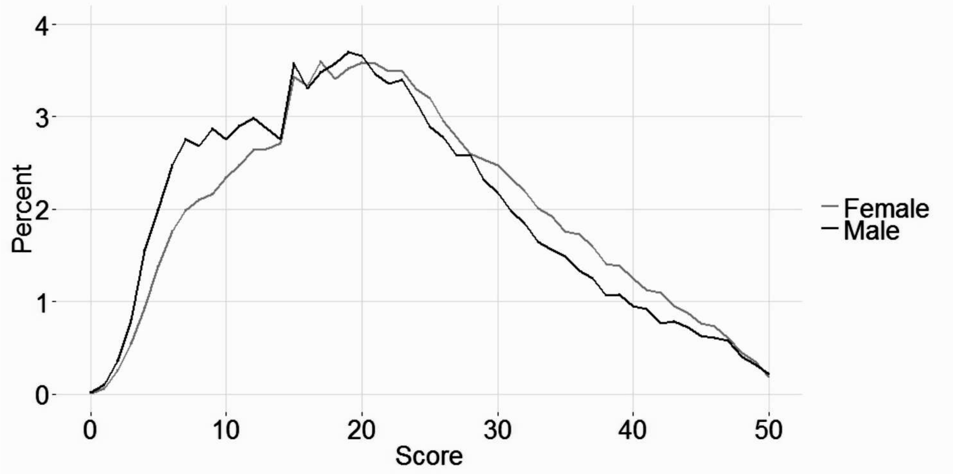


Figure 1 then presents the scores of persons who completed education in different profiles – both those with an extended humanities program and those with a mathematical one. I assume that a strong proxy for taking the extended Matura is participation in a program with an increased number of hours. So, Matura at this level is taken by a population differentiated not by abilities but also by level of motivation. For students who decided on a humanities major in tertiary education, where during the recruitment process scores of non-mathematical examinations are taken into account, the motivation may be to pass the exam at the minimum acceptable level. Future graduates of engineering or technical universities will be more motivated. In some sense, Figure 1 compares "apples and oranges", women with lower and higher abilities and men with different abilities. One may presume that "excluding" people who took the extended Matura from the score distribution, will crucially decrease this differentiation. The differentiation of scores after such a procedure can be seen in Figure 2.

As Figure 2 shows, excluding students who passed the extended exam leveled the difference at the upper end of the score distribution. Furthermore, there are relatively more males than females among students with the lowest scores. Since this population is more homogenous in terms of mathematical skills, maths background in high school, and further aspiration towards a major at the tertiary level, it may represent more accurately gender differentiation in maths exam scores.

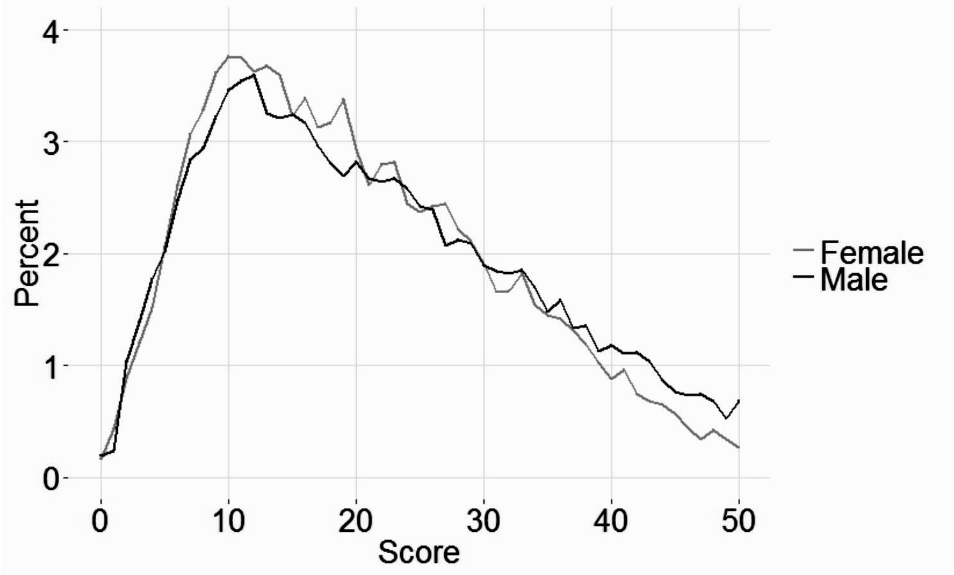
Figure 2. Percent of female and male on basic Matura after excluding population taking extended Matura



Relying on such a subpopulation of secondary school graduates, in many points of score distribution the relative differences are more favourable to girls. However, these charts only represent relative results expressed as a percentage of males or females with a given score. Taking into account that the share of male and female students who took the basic Matura was highly unequal (63% females and 37% males), those small differences in percentage are significant in terms of absolute values. Far more females than males have been “cut out” from Figure 1, having taken the extended Matura. Out of the group of females who took the basic Matura, 20% also took the extended, in comparison to 40% of males. This is a considerable difference. The effect of this is the smaller spread of scores between the two populations of students taking the extended exam seen in Figure 3. Here, the differences between the sexes are significantly smaller than in the case of the basic Matura, yet still in the upper end of distribution there are more boys. Taking absolute numbers into account, the number of females is notably smaller than of males, and they represent a more select group than the male group.

To sum up, among the students taking the basic Matura, boys achieved a higher average of points (29) than girls (26). This population constitutes a mixture of people with higher and lower abilities among women and men, where the differentiation of boys’ scores was higher – although this population was smaller, there was a greater number of more and less talented representatives of this sex than among women. After excluding

Figure 3. Percentage of male and female on a given score level in advanced Matura exam 2015



people who took extended Matura (females and males), the gap existing in the original situation disappears. Those people will still have influence on the size of the gap as seen in Figure 1 because – it may be assumed – they achieved higher scores on the basic Matura than people who did not take the extended Matura. On average, four out of five women do not decide to take the extended level exam. It must be emphasised that the problem of aversion to mathematics does not concern only females – boys also share this unwillingness, yet not to the same degree.

In the present system of "early choice" of Matura subjects, students' decisions are influenced by their mathematical experience from junior-high school, and perhaps from even earlier stages. These initial stages seem strategic in terms of maths-avoiding and maths-seeking attitudes and learning strategies because, as different studies show, early mathematical failures are rarely possible to catch up and vice versa – early successes are an announcement of future ones.

The size of the gap expressed in absolute values at the extended Matura exam in Mathematics may have long-term consequences, the most obvious symptom of which is the gender disproportion in many fields of STEM. Not having passed this exam prevents women from studying and in the future excludes them from STEM-related jobs.

Differential item functioning of test items

One of the hypotheses explaining the differentiation in test scores argues that men and women may react differently to the same test questions. It may happen, for example, in the situation where the content of the question will be cognitively closer to one sex than another by the vocabulary used or its content.

In the tradition of research on text properties, this problem is described as Differential Item Functioning. DIF occurs when examinees from different groups with the same level of ability manifest different probabilities of success in answering a given test item. The presence of DIF means that giving the correct answer to a test task depends on factors correlated with belonging to a group, and not only on the level of ability that the test measures (Kondratek, Sikorska, Świst 2015). For example, the fact of belonging to a group of women, or an ethnic minority shall have an impact on the probability of giving the correct answer in the test.

The notion of DIF, due to its semantic field, may bring with it associations with other terms used for describing systematic differentiations in test scores such as bias, injustice, or partiality (Hornowska 2000). It is worth explaining that different methods to detect test biases were developed in the USA in the early 60s and were originally dedicated to a better understanding of systematic differences in test scores between White, Black, and Hispanic minorities. The word "bias", which was initially used to depict this problem, became too confusing, because in colloquial language it referred to the event of prejudice, and in the statistical sense to non-random error (Angoff 1993). In turn, Kondratek, Sikorska, Świst (2015) point out the necessity to distinguish the notion of DIF from the other two possible situations. The source of differentiation in test scores may be differences in the psychometric properties of tasks (e.g. if the task is easy or difficult), as well as differences in the level of differences between the groups. In the case of DIF, the easiness of a task – what percentage of individuals from those compared groups shall give the correct answer – is conditioned by the level of ability measured by the test. DIF describes the differences in item functioning after groups have been matched with respect to ability (the proxy for ability is usually total test score) (Kondratek, Sikorska, Świst 2015; Holland, Thayer 1986). In other words, individuals are equated at the level of test score before they are compared. "Sorting out" students according to their ability measured in the test distinguishes the DIF approach from the unconditional analysis of the frequency of existence of the correct answer among groups. Only the control of the difference in ability level distribution among groups enables the statement that members of the examined groups react differently to the same task.

Summing up, DIF appears if giving the correct answer in the test question is influenced by belonging to a group, ability level aside. Such a task shall be systematically more difficult for one group. In its simplest form, DIF refers to the situation of a single-point item (for which only one answer is correct), where the population answering is divided into two groups. In test theory jargon these are usually known as the reference group and focal group.

In Polish educational studies, DIH analysis has been used several times. Koniewski et al. (2014) used it to analyse the functioning of two versions of the same test, to find out if the location of distracters has an impact on the probability of giving the correct answer. Moreover, Gruniewska and Kondratek (2012) showed the number of questions with DIF for both sexes in the junior-high school tests for 2002–12. However, DIF is still rarely used to evaluate fairness of test items, and no systematic studies using this method have been undertaken to evaluate Polish exit exams.

Procedure of the Mantel-Haenszel test

The most popular method for assessing DIF is the Mantel-Haenszel procedure (Holland and Thayer 1988). The test is based on the classic form of a 2 × 2 contingency table. It crosses two variables: correct and incorrect answers for a single test item and membership of either a focal or reference group. The latter is the one of interest and the former serves as a comparison. The test items are analyzed using two groups of examiners by two levels of responses split into each of the score levels. The number of 2 × 2 tables constructed corresponds with the number of unique scores for the test, which gives $2 \times 2 \times m$, where m is the total score level (Narayanan, Swarminathan 1994, Kondratek, Gruniewska 2013). This is illustrated in the following table:

	Response		
Group	1	0	
Focal Group (f)	a_i	b_i	N_{f_i}
Reference Group (r)	c_i	d_i	N_{r_i}
Total (t)	N_{1_i}	N_{0_i}	T_i

In the table, a, b, c and d are the frequencies for the corresponding cells, N_{f_j} and N_{r_j} are the numbers of reference and focal group for a given score, and N_{1_j} and N_{0_j} are the numbers of correct and incorrect responses.

MH uses the classic procedure of odds ratio calculated as:

$$\alpha_i = \frac{a_i d_i}{b_i c_i} \tag{1}$$

The null hypothesis of no-DIF claims that odds for a correct answer for a given item and on a given m are equal for the reference group and focal group (Narayanan, Swaminathan 1994: 316). This means that at each level of total score m, examinees have the same probability of correct answers regardless of group membership, taking into account the score of the entire test (Kondratek, Sikorska, Świst 2015). An alternative hypothesis claims that all odds ratios α_i will equal the common odds ratio α (Kondratek, Gruniewska 2014).

MH statistics are evaluated against standard chi-square values with one degree of freedom. It has the following form:

$$MH_{\chi^2} = \frac{\left(\sum_j^{\square}(a_i - E(a_j)) - 0.5\right)^2}{\sum_j^{\square} D^2(a_i)}, \tag{2}$$

where $E(a_i)$ is an expected value and $D^2(a_i)$ is a variance of a_i (Gruniewska, Kondratek 2012). The statistics show whether the odds of a correct answer for the focal group differ significantly from the odds of success for the reference group (Ayala 2013). It has been proven that MH_{χ^2} is uniformly the most powerful unbiased test of H_0, compared with H_1 (Gruniewska, Kondratek 2012, Holland and Thayer 2013).

To obtain the strength of the relation in $2 \times 2 \times m$ tables, an estimate of the common odds ratio is used. It is given by:

$$\hat{\alpha}_{MH} = \frac{\sum\limits_i \frac{a_i d_i}{T_i}}{\sum\limits_i \frac{b_i c_i}{T_i}}, \tag{3}$$

In this expression, the individual odds ratios are summed to obtain a common estimate on a given score level. When $\hat{\alpha}_{MH}$ equals one, it implies that on average there is no difference in item functioning between groups in term of odds. When the value of $\hat{\alpha}_{MH}$ is greater than one it indicates that on average members of the reference group perform better than members of the focal group on a given item, and when the value is less than 1 it means that, on average, the reference group perform worse (de Ayala 2013). Due to the lack of symmetry in the scale of $\hat{\alpha}_{MH}$ (values range from 0 to $+\infty$) it is transformed to a natural logarithm to obtain a log odds relation $\beta = \ln(\alpha_{HM})$ or difficulty delta scale. The latter is expressed by: $\Delta_{MH} = -2.35 \ln(\hat{\alpha}_{MH})$. Negative values indicate that an item favours the reference group, and positive that an item favours the focal group. The magnitude of both indicators expresses the degree of DIF. Based on this value, each test item can be classified into one of the three categories with regard to the strength of DIF. Classification usually uses information on the absolute value of DIF and the statistical significance of the MH test. Meyer (2014), based on Rebecca Zwick and Kadriye Ercikan (1989), describes the rules of classification for test items as follows:

- "A", when the common odds ratio is between 0.65 and 1.53 or MH test is non-significantly different from 0. These items are free of DIF.
- "B", when the MH test is significantly different from 0 and either has absolute value between 1 and 1.5 or absolute value
- "C", when the common odds ratio less than 0.53 and the upper bound of the 95% confidence interval for the common odds ratio is less than 0.65, or the common odds ratio is greater than 1.89 and the lower bound of the 95% confidence interval for the common odds ratio is greater than 1.53.

The HM test is relatively straightforward in terms of computation procedure, but it has one significant constraint. It only detects one type of DIF, namely uniform DIF. Using a framework of Item Response Theory, a uniform DIF takes place when the probability of a correct answer is higher throughout the continuum of latent variable (e.g. ability) for one group than for the other group. The second type of DIF, called non-uniform DIF, occurs when the lines representing the probability of a correct response for the focus and reference group cross at some point of latent variable. The probability lines cross because part of the continuum members of the reference group perform better, but in a different part it is reversed and members of the reference group perform worse (see Kondratek et.al 2015). Other methods are recommended to detect both types of DIF, for example logistic regression.

Secondly, the MH test in the form discussed above is dedicated only to dichotomous test items. The procedure intended for polytomous items is called the generalised MH test (GMH). Similarly to the case of dichotomous items, one contingency table is arranged for each item at each score level. To test the null hypothesis on lack of DIH, Wlad statistics is used. As with dichotomous items, this statistic has chi-square values with one degree of freedom (Henderson 2001, Golia 2012, Kondratek et.al 2015).

Analysis of DIF items

For analysis of DIF, I used jMetric open source software. One of the advantages of jMetric compared to other open source applications is that it allows the computation of DIF for polynomial items (for example difR supported by R allow to compute DIF only for dichotomous items). For the purpose of analysis, females were set as the focal group and males were the reference group.

Taking all 34 questions included in the basic level of Matura, only three manifested an intermediate level of DIF (B). Among them, two favour women and one favours men. All the other items manifested negligible magnitude (classified as A) (Table 1).

Table 1

"New" Matura exam items flagged as DIF

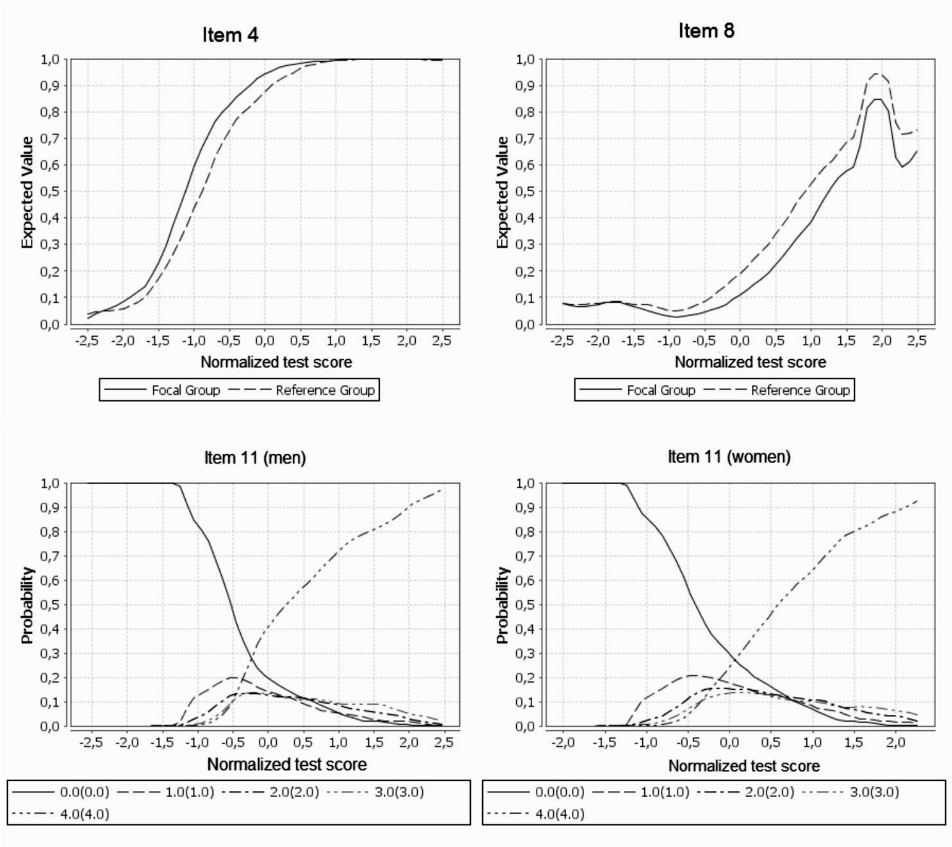
Item	Δ_{MH}	ETS Classification	Favored gender
Basic			
item 4	1.41	B+	females
item 8	1.23	B–	males
item 26*	0.03	BB+	females
Extended			
item 11*	–0.06	BB–	males

* These are polytomous items. Effect size estimated by standardised mean difference (SEM) and has been divided by the item score range (see Golia 2012, Meyer 2012).

Items flagged as DIF concerned different fields of mathematics. In item 4, students had to solve an equation with one unknown. In item 8 students had to find a range of function the shape of which had been given on a figure. Item 26 concerned solving inequality with one unknown, but compared with item 4 it was more advanced with a constructed response. Students could earn from 0 to 2 points for this task. In extended Matura only one item was flagged as DIF. To solve this problem students had to be familiar with probability laws and combinatorics. The score ranged from 0 to 4 points in this item. It wasn't the most difficult item in the test – in the most highly valued item a student could score 7 points.

To visualise the results, items with DIF are represented graphically in Figures 4.1–4.4. The plots show non-parametric item characteristic curves for the focal and reference group (Meyer 2014). It represents the relation

Figure 4.1–4.4. Non-parametric item characteristic curve for DIF items



between the probability of a correct answer and examinee ability defined as total score and transformed into a normalised score. Theoretically this relation should be monotonic – non-decreasing – because it is expected that the higher abilities an examinee will have, the higher the probability that he or she will give the correct answer to a given item.

While that is true in the case of item 4, where women were more likely to give the correct answer, item 8 presents a totally different form. Here, the probability of giving the correct response was not the same for both genders across all levels of scores. More importantly, the upper part of the distribution has an inverted U-shape. In this item – opposite to item 4 – men had a higher probability of giving the correct answer across the whole distribution.

Plots on the lower panel show non-parametric curves for all range of scores in item 11, separate for males and females. For both genders, the probability of receiving the maximal number of points (4) increases with ability. However, small inter-gender differences are present in all score levels. For example, among the students who received the maximal number of points in the test (50 points), 0.75 men and 0.25 women scored the maximal number of points for this individual task.

The existence of DIF in the indicated tasks on the average level does not necessarily indicate their bias or injustice. Nevertheless, it may signal that both sexes "read" the content of the task differently, or that a given task measures a different ability than it should (Gruniewska and Kondratek 2012). The number of questions with DIF and its size seem marginal, taking into account the number of questions in the entire test. One may exclude the hypothesis that the different functioning of test questions influences the differentiation of Matura scores. However, this conclusion requires an additional analysis of Matura tests from previous years to be conducted using methods other than MH.

Conclusion

The results of the analysis presented in this article are concordant with other findings concerning differentiation in mathematics. The average differences in the mathematical test scores are marginal, yet in the upper edge of the scores distribution they are more pronounced. However, it turns out that in Poland the system of education in secondary schools is responsible for that to a great extent, specifically, through the existence of a profiled curriculum and the obligation for individuals with greater mathematical skills to pass the exit exams at both levels. In more homogenous groups, that is, among students taking only the extended level or after elimination of that group from the population of individuals taking basic Matura, the relative differences become almost invisible, or in favour of women. Such a result was not obvious at all in respect to studies on self-assessment of the mathematical abilities of both sexes, or considering aspirations for further education.

The analysis of functional differentiation of test questions enabled the exclusion, with a high degree of probability, of the hypothesis that the questions themselves – their content and form – are the source of possible variations. Apart from 3 questions in the basic examination and 1 question in the extended one, sex did not crucially differentiate the probability of giving the correct answer. However, it is worth complementing these analyses with more robust methods for controlling the interaction between abilities and the probability of giving the correct answer than the Meantel-Heanshel test.

Nonetheless, besides the relative differences, more concerning are gender differences expressed in the form of absolute numbers. Although females are over-represented compared to male high school students, only a small minority of them tend to take extended Matura in maths. One may propose the hypothesis that a portion of female graduates of junior-high schools, having the possibility of free choice, decides to avoid mathematics and minimizes time spent on learning that subject, investing it into others. As a result, a significantly larger population of men has the formal ability to apply for mathematical majors at universities.

Following the process of selection of females and males who take the extended Matura also requires further study. Is the hypothesis that men have better self-assessment of their mathematical abilities also true for Polish students? Do women and men with identical scores at basic Matura have the same chances to take the extended level? Answers to these questions shall help better understand the mechanism of selection and what factors cause them. This knowledge is the more necessary, as the results of studies devoted to gender gap in maths are being used as an argument in the ideological dispute on the "superiority" of one sex over the other, or they become the convenient justification of the inequality between the sexes, and very rarely do they serve to level the existing differences.

REFERENCES

Angoff, W. H. (1993) Perspectives on Differential Item Functioning in Differential Item Functioning. In P. W. Holland & Howard Wainer (Eds.), *Differential Item Functioning* (pp. 3–25). New York: Routledge.

Ayala de, R. J. (2013). The theory and practice of item response theory. New York: Guilford Publications.

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.

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

Doris, A., O'Neill, D., & Sweetman, O. (2013). Gender, single-sex schooling and maths achievement. *Economics of Education Review*, 35, 104–119.

Ellison, G., & Swanson, A. (2009). *The gender gap in secondary school mathematics at high achievement levels: Evidence from the American Mathematics Competitions* (No. w15238). National Bureau of Economic Research.

Fryer Jr, R. G., & Levitt, S. D. (2009). *An empirical analysis of the gender gap in mathematics* (No. w15430). National Bureau of Economic Research.

Good, C., Aronson, J., & Harder, J. A. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology*, 29(1), 17–28.

Golia, S. (2012). Differential Item Functioning classification for polytomously scored items. *Electronic Journal of Applied Statistical Analysis*, 5(3), 367–373.

Halpern, D. F. (2013). *Sex differences in cognitive abilities.* New York: Psychology press.

Henderson, D. L. (2001). Prevalence of Gender DIF in Mixed Format High School Exit Examinations. Paper presented at the Annual Meeting of the American Educational Research Association (Seattle).

Holland, P. W., & Thayer, D. T. (1986). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco).

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. Proceedings of the National Academy of Sciences, 106(22), 8801–8807.

Jurajda, S., & Munich, D. (2008). Gender Gap in Admission Performance under Competitive Pressure. *CERGE-EI Working Paper Series*, (371).

Kling, K. C., Noftle, E. E., & Robins, R. W. (2012). Why do standardized tests underpredict women's academic performance? The role of conscientiousness. *Social Psychological and Personality Science*, 4(5), 600–606.

Kosmala-Anderson, J. (2006). Płeć a natężenie i rodzaj psychosomatycznych reakcji na stres egzaminacyjny. Przegląd terapeutyczny 1/2006.

Kondratek, B., & Grudniewska, M. (2014). Comparison of Mantel-Haenszel with IRT procedures for DIF detection and effect size estimation for dichotomous items. *Edukacja Quarterly*, 128(3).

Kenney-Benson, G. A., Pomerantz, E. M., Ryan, A. M., & Patrick, H. (2006). Sex differences in math performance: The role of children's approach to schoolwork. *Developmental psychology*, 42(1), 11.

Meyer, J. P. (2014). Applied measurement with jMetrik. New York: Routledge.

Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: does competition matter? *Journal of Labor Economics*, 31(3), 443–499.

Stoet, G., & Geary D.C. (2013) Sex Differences in Mathematics and Reading Achievement Are Inversely Related: Within- and Across-Nation Assessment of 10 Years of PISA Data. PLoS ONE 8(3): e57988.

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, 16(1), 93.

Skórska, P., & Świst, K. (2014). Wielkość efektu płci w wewnątrzszkolnych i zewnątrzszkolnych wskaźnikach osiągnięć ucznia. Konferencja PTDE.

Świst, K., Skórska, P., Koniewski, M., & Jasińska-Maciążek, A. (2015). Sex differences in guessing and item omission. Edukacja, 3, 48–62.

Szaleniec, H., Kondratek, B., Kulon, F., Pokropek, A., Skórska, P., Świst, K., & Żółtak, M. (2015). Porównywalne wyniki egzaminacyjne. Warszawa: Instytut Badań Edukacyjnych.

Jakubowski, M., & Pokropek, A. (2009). Badając egzaminy: Podejście ilościowe w badaniach edukacyjnych. Centralna Komisja Egzaminacyjna.

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.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 1067–1101.

Niederle, M., & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. The Journal of Economic Perspectives, 24(2), 129–144.

Penner, A. M. (2008). Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. American Journal of Sociology 114: S138–S170.

PISA, O. (2012). Results in Focus: What 15-year-olds know and what they can do with what they know. [2014–12–03].

Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental psychology*, 50(4), 1262.

Tannenbaum, D. I. (2012). Do gender differences in risk aversion explain the gender gap in SAT scores? Uncovering risk attitudes and the test score gap. Unpublished paper, University of Chicago, Chicago.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, 26(1), 55–66.