Gender Differences in Mathematics and Science Achievement of Eighth Grade Students in Ukraine on TIMSS 2011

TETIANA V. LISOVA, YURIY O. KOVALCHUK
Department of Applied Mathematics and Educational Measurement
Nizhyn Mykola Gogol State University
Krapivianskogo 2, Nizhyn, 16600
UKRAINE

Abstract: The purpose of the paper is in-depth analysis of gender differences in Ukrainian student achievement in mathematics and science on TIMSS 2011 results because the average scores similarity hides some significant differences between the achievements of boys and girls. Boys significantly outperform girls in some cognitive and content domains, while the benefit of girls never was significant. The procedure of Differential Item Functioning (DIF) on the basis of the Mantel-Haenszel method was used to test the fairness of assessment. The results show that the set of TIMSS 2011 items does not have any gender bias on Ukrainian students since there are insignificant number of items with large DIF, among which some favouring boys, and some favouring girls. Each of gender groups is equally represented among both stronger and weaker participants. Any significant relationship between the DIF size and the difficulty or type of the item is not detected. But the items in number, physics and earth science, and math problems on applying function in favour of boys. This is likely caused by the fact that in Ukrainian society still there is attitude to such fields of study as masculine.

Key-Words: Gender differences, TIMSS, bias, uniform DIF, non-uniform DIF, Mantel-Haenszel method

1 Introduction

Given the decisive influence of education on the realization of human life, the policy of minimizing the gap between educational opportunities for men and women was implemented in most countries. In the second half of the 20th century there have been significant improvements in reducing gender inequalities regarding access to all levels of education as well as educational achievements of boys and girls. In recent years in Ukraine, women demonstrate higher success rate of university admissions than men. For example, in 2010 the proportion of women enrolled in universities was 55.3% of all female applicants, while there were only 48.8% among men [1]. As in many developed countries, there are more women than men among university students in Ukraine. However, such success of women is some paradoxical. There is a gender differentiation in the choice of specialty, in which applicants want to get higher education. In humanities specialties number of women exceeds 75%, whereas in specialties related to the exact sciences, their number does not exceed 25%. As a result, women often cannot find work according to their specialty because of low labour market demand for their diplomas.

We know that the majority of gender differences is formed and fixed in the period of study in secondary school. The results of various international projects such as TIMSS, PIRLS or PISA, significantly deepened the understanding of the formation of gender differences, needs, capabilities and effectiveness of introducing special measures to improve alignment and academic success of boys and girls, because of traditions, cultural and religious peculiarities of countries. For some countries, many researchers have noted a significant gender gap in reading in favour of girls, and the gender gap in mathematics and physics in favour of boys [2, 3, 4, 5]. Girls tend to show much higher interest in reading, devote more time to reading. However, they show less interest in math, lower confidence in their mathematical abilities, higher levels of anxiety, uncertainty and stress in mathematics classes. In addition, the magnitude of gender inequality may depend on the characteristics of the test (topics, cognitive level, item format), characteristics of the target population (sampling, age, level of education) and the type of statistical analysis [5].

As shown from TIMSS 2011 results in science [6], on average across the eighth grade countries, girls had a 12-point advantage in biology and a 10-point advantage in chemistry, while boys had a 2-point advantage in earth science. There was no significant difference between the achievement of girls and boys in physics. Also girls outperformed boys in all three of the cognitive domains. In mathematics [7], on average across the eighth grade countries, boys had higher achievement than girls in number (468 vs. 464), but girls had higher achievement in algebra (476 vs. 464), geometry (464 vs. 461), and data and chance (459 vs. 456). Girls outperformed boys on average in mathematics in both the knowing and reasoning domains.

However, for Ukrainian participants there were no significant gender differences in general indicators. Average scores of girls and boys are very similar: 478 and 481 respectively in mathematics, and 499 and 503 in Science. However, such similarities of average values may conceal some significant differences. For example, scores were significantly higher for boys in numbers, physics (13 points) and earth science (15 points). Also, the average score of boys was 13 points higher in mathematics applying cognitive domain [6, 7]. Girls did not demonstrate significant advantages neither in content nor in cognitive domains. The reasons for this can be explained better by means of in-depth analysis of the test results at item level.

2 Problem Formulation

Sometimes a test score can have a bias, the causes of which may be associated with external to the basic construct factors (such as belonging of person to a certain group - cultural, ethnic, social, gender, etc.). The term bias is usually associated with unfair, biased evaluation of the group. In most cases, Differential Item Functioning (DIF) procedure is used to identify bias. In the past, DIF and bias terms were interchangeable, but since 1988 Holland P. and Thayer D. [8] distinguished these two concepts. The introduction of the more palatable term DIF allowed one to distinguish item impact from item bias.

It is now accepted that DIF appears where respondents from different groups, say boys and girls, have the same ability but different probability to solve an item correctly. Item impact described the situation in which DIF exists, because there were true differences between the groups in the underlying ability. Item bias described the situations in which there is DIF because of some characteristic of the test item that is not relevant to the underlying ability. Therefore, DIF is a necessary but not sufficient condition for bias. Depending on interaction between group membership and the ability levels, two classes of DIF are distinguished: uniform and non-uniform. In the case when no interaction is found it is uniform DIF, otherwise non-uniform DIF is present.

For each TIMSS assessment, examining item statistics to detect any gender bias is an important stage of item selection. It is therefore reasonable to assume that where significant differences do occur, they result from differences in performance rather than problem situations favoring one gender or the other. The interaction between items or cognitive levels by gender may differ significantly for different countries because of differences in cultures and educational systems. So, before analyzing gender differences in mathematics and science for Ukrainian participants it is valuable to check if the instrument of the measurement is free of gender biases. Only after proving that differences in abilities (and not tests) are the ones which cause the gender gap in achievement it is possible to start a proper analysis of factors behind these abilities. Therefore, the main question in this paper is: is there are differential item functioning in the TIMSS 2011 study comparing boys and girls and does it give an advantage to any of genders in Ukraine?

Various statistical methods for detecting DIF were developed within the framework of three main approaches: modeling item responses via contingency tables and/or regression models, Item Response Theory (IRT), and multidimensional models [9]. For example, staff of Educational Testing Service (ETS), which has been a leader in fairness assessment, published around 100 research bulletins, memoranda, or reports on the topics of item fairness, DIF, or item bias [10]. ETS has found original implementation of the Mantel-Haenszel procedure based on analysis of contingency tables for DIF assessing. A major disadvantage of this method is that they have low power in detecting non-uniform DIF.

2.1 Mantel-Haenszel procedure for DIF detection

Nonparametric Mantel-Haenszel method (MH) is based on the assumption of the equality of chances for overall success in each group (reference and focal). If there are no differences between the groups, they have the same chances for success, so the odds ratio should be close to 1 [8]. To calculate the required statistics, such as dichotomous items, at each score level *j*, a 2-by-2 contingency table is created for each item *i*, as shown in Table 1.

Then, the odds ratio is calculated for all strata *j* using the formula:

$$\alpha_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j}. \qquad (\qquad 1)$$

Score level j	Score on Studied Item		Total
	1	0	
Reference Group	A_j	B_j	N_{rj}
Focal Group	C_j	D_j	N_{fj}
Total	T_{1j}	T_{0j}	T_j

Table 1. Contingency table for dichotomous item.

If α_{MH} is greater than 1, this means that the members of the reference group performed the item better than the members of the focus group. On the contrary, if the value is less than 1, this means that the reference group performs the item worse than the focal group.

To test the null hypothesis of no deviation in the chances of success in both groups (no DIF) the statistics χ^2_{MH} is used, which have Chi-Square distribution with one degree of freedom:

$$\chi^2_{MH} = \frac{\left\{\left|\sum_j (A_j - E(A_j)\right| - 0.5\right\}^2}{\sum_j \operatorname{var}(A_j)}, \quad (2)$$

where $E(A_j) = \dfrac{N_{rj}T_{1j}}{T_j}$, $\operatorname{var}(A_j) = \dfrac{N_{rj}N_{fj}T_{1j}T_{0j}}{T_j^2(T_j - 1)}$.

α_{MH} is often transformed to Δ_{MH} to enhance the interpretability of the result using the formula $\Delta_{MH} = -2.35\ln(\alpha_{MH})$. Research at the ETS has resulted in proposed Δ_{MH} values for classifying DIF as negligible, moderate, or large.

In our study, we use the classification equivalent to one used by ETS but in a slightly altered form as the computer program Winsteps (version 3.80.1) is employed. The latter provides means for tests results modeling based on the Rasch-family models. The students' ability measures are sliced into strata in Winsteps instead of raw scores. The *DIF contrast* is defined as the difference in item difficulty for two participants groups. Items flagged as h aving DIF after Chi-Square tests can be classified as exhibiting negligible, moderate or large DIF based on t he following criteria for the DIF size [11]: C (large) – if $|DIF\ contrast| \geq 0.64$; B (moderate) – if $0.43 \leq |DIF\ contrast| < 0.64$ and A (negligible) – if $|DIF\ contrast| < 0.43$. Typically, items with significant DIF of C level should cause concern [12]. All decisions in this work were taken at the 0.05 significance level. Winsteps has several advantages in detecting DIF. It allows exploring non-uniform DIF as opposed to classical Mantel-Haenszel procedure. Dichotomous and polytomous items can be studied simultaneously.

3 Problem Solution

In this study the data for 3378 Ukrainian participants (1723 girls and 1655 boys) from TIMSS 2011 study have been used. The performed analysis of Differential Item Functioning for 215 mathematics items and for 216 science items is based on the Partial Credit model, which in the case of dichotomous items coincide with the Rasch model. The missing values and non-reached items in the students' responses were considered as incorrect.

In addition to DIF analysis Winsteps allows carry out the analysis of Differential Test Functioning (DTF) and Differential Group Functioning (DGF) at the test level. DTF investigates whether the test functions in the same way for different gender groups, through a comparison of how the test items function.

Figures 1 and 2 displays a scatterplot for the item difficulties in the boys' group compared to those in the girls' group. 53 (24.7%) mathematics items remains outside 95% two-sided confidence bands, when compared the difficulties of items obtained for Ukrainian boys and girls. The dotted identity line goes through the origin of the two axes. The maximum value of the Student's t-statistics for the item 7 (algebra, reasoning) is $t_7 = 4.47$. This item is significantly more difficult for the boys. The item 30 (number, knowing), conversely, is significantly easier for boys ($t_{30} = -5.87$).

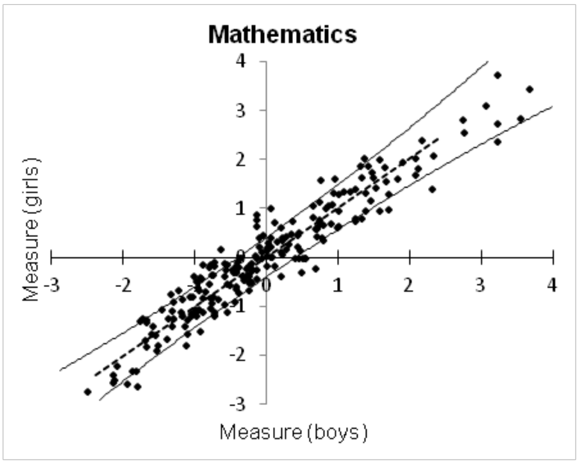

Fig.1. Scattering of items measures for mathematics.

Much more scattering is observed when comparing the difficulties of science items. There are 56 (25.9%) items outside the confidence bands. The item 218 (biology, knowing, $t_{218} = 11.6$) is much harder for boys than for girls, whereas the item 274 (physics, knowing, $t_{274} = -4.34$) is much easier. DTF procedure cannot be a substitute for bias research, because the data fit to the Rasch model not ideally (some outfit statistics for persons exceeds 3.5, indicating a possible guessing), but at least it indicates the direction of search.

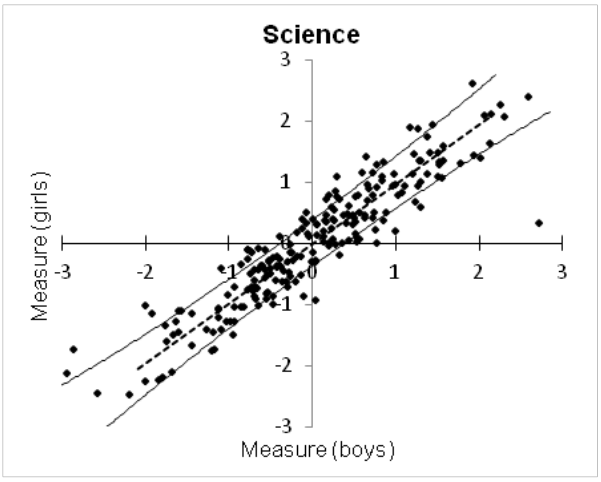

Fig.2. Scattering of items measures for science.

DGF allows revealing interactions between classification-groups of person and classification-groups of items. The difference in difficulty of the item between two groups (*DGF Contrast* in Winsteps) should be at least 0.5 logits for DGF in order to be noticeable [11]. There is no statistically significant difference between the averages in mathematics and science for Ukrainian girls and boys. Nor is there significant difference between ability of boys and girls in content domains of chemistry, geometry, and data and chance (Fig.3).

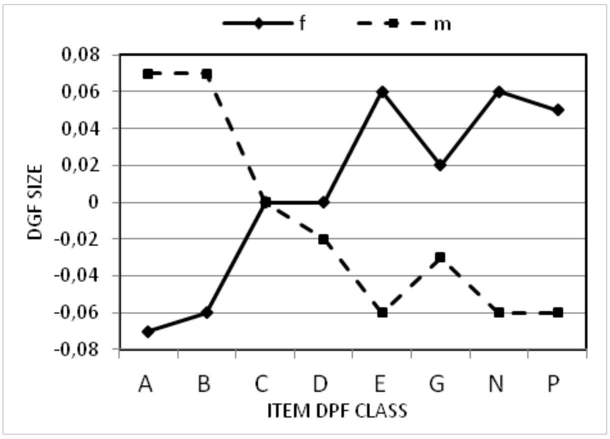

Fig.3. Differential Group Functioning by content domains.

The advantage of girls is statistically significant ($p = 0.00$) only in algebra and biology, but the effect sizes (*DGF Contrast* = −0.14) does not give reason to believe that this is an essential advantage. Boys traditionally show no significant advantage in physics, earth science, and number content domains. As to cognitive domains, there is no significant gender gap, though the girls demonstrate a little advantage in knowing whereas the boys in applying.

For DIF analysis has been chosen not too thin strata in 1 logit (MHSLISE = 1.0), because data matrix has many gaps which are not administered [11]. A number of items that demonstrate large DIF (C-level) in comparison of their difficulty for different groups according to content and cognitive domains are shown in Table 2. There are more such items for science than for mathematics: 16 (7,4%) and 9 (4,2%) accordingly. There are twice more items in favor of boys. This is mainly the items in number, physics and earth science. This is consistent with previous results. All math items with large DIF in applying cognitive domain also function in favor of boys. The constructed-responses items (4 of 6) prevail among those math problems that function in favor of boys, whereas in science those are the multiple-choice items (8 of 11). However, none of items in algebra, geometry and biology show large DIF in favor of boys. Instead, two algebraic items show large DIF in favor of girls. Many studies argue that women tend to be better at algebra, arithmetic and algebraic operations [13, 14]. But in our case on the multiple-choice item 92 (algebra, knowing), girls outperform boys also because it requires having of certain verbal abilities:

What does xy +1 mean?
A. Add 1 to y, then multiply by x.
B. Multiply x and y by 1.
C. Add x to y, then add 1.
D. Multiply x by y, then add 1.

C-DIF items	All	In favour of girls	In favour of boys
Mathematics	**9**	**3**	**6**
number	5	1	4
geometry	0	0	0
algebra	2	2	0
data and chance	2	0	2
Science	**16**	**5**	**11**
physics	6	1	5
chemistry	5	3	2
biology	1	1	0
earth science	4	0	4
Cognitive domains	**25**	**8**	**17**
knowing	10	5	5
applying	13	2	11
reasoning	2	1	1

Table 2. Number of items with large DIF by content and cognitive domains.

Some researchers point out that gender differences in the tails of the distributions can occur in different ways. For example, in mathematics larger gender gap in favor of boys is most prominent at the very high levels of achievement [5]. For Ukrainian participants such dependence was not found. Each gender group is almost equally represented among both stronger and weaker participants. There were 54.4% and 52.7% of boys among the 10% of participants with highest scores and the 10% of participants with largest scores respectively.

Also correlation between the DIF size and items difficulty was studied for each group. All correlation coefficients were negative and statistically non-significant due to the small number of items with DIF. But it is interesting that the size of DIF is the greater, the easier is the item. This trend is more pronounced for boys in science and for girls in mathematics.

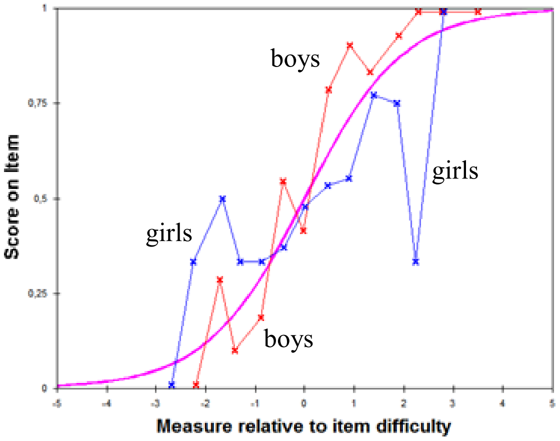

Fig.4. Non-uniform DIF for item 39 (geometry, knowing).

Winsteps also allows analyzing non-uniform DIF by visualization of empirical curves for different groups. Classical MH statistics can detect no DIF when the deviations between groups have opposite signs in different segments of the ability axis and so the result is compensated. In our case, 11 items with a noticeable non-uniform DIF was found. For most of them, the location of the empirical curves is typical as in Fig.4 for the item 39 (geometry, knowing). Weak girls showed a higher chance of success than the weak boys, while for above-average levels of ability it was vice versa. This would be possible due to more guessing the correct answer by girls. But the hypothesis about predisposition of girls to guessing requires additional proof. Such a study can be conducted within three-parameter models of Item Response Theory, which contain additional guessing parameter.

4 Conclusion

Analysis of TIMSS 2011 test items shows that they are absolutely appropriate for assessments of Ukrainian students' achievements in mathematics and science, both – boys and girls. Only 25 (5.8%) of 431 items observed by Mantel-Haenszel method demonstrate large DIF, more in science than in mathematics. Among them there are the items favouring girls and the items favouring boys, but more in favour of boys. At the same time, there are no items favouring boys in algebra, geometry and biology. No differences between boys and girls are found among the items on knowing and reasoning cognitive domain. But among the DIF items on applying the vast majority is in favour of boys.

There is no difference in differential item functioning between multiple choice and constructed-responses items. It is found that the higher the item difficulty the lower the differential item functioning which means that if abilities of both genders are in the same advanced level, items to measure them work more similar for boys and girls than in the case of low abilities. Non-uniform DIF analysis show that often weak girls outperform weak boys and strong boys outperform strong girls.

Content analysis of released items which demonstrate DIF does not give reason to state item bias in favor of any of sex. The achievement gap between boys and girls in some content or cognitive domains is caused by the differences in their abilities and not by a differential item functioning. This difference is likely due to the fact that attitude towards subjects like mathematics and physics as masculine still exists in Ukrainian schools. This occurs both through the content of the curriculum and through gendered interactions between teachers and students, and between students themselves. Despite girls' increased educational attainment, the gender gap in such fields of study as well as in focus on the future professional activity will be kept until women do not feel more confidence in their abilities and do not change their self-esteem.

References:

[1] Kovtunets V.V., Rakov S.A. (Ed.) *Research of quality of competitive selection of university students on the results of an independent external evaluation: analytical materials*, Kiev: Nora-Druk, 2015.

[2] Johnson S., The contribution of large-scale assessment programmes to research on gender differences. *Educational Research and Evaluation*, Vol.2, No.1, 1996, pp. 25-49.

[3] Lafontaine D., Monseur C. Gender gap in comparative studies of reading comprehension: to what extent do the test characteristics make a difference? *European Educational Research Journal*, Vol.8, No.1, 2009, pp. 69-79.

[4] Le L.T., Investigating gender Differential Item Functioning across countries and test languages for PISA Science items, *International Journal of Testing*, Vol.9, No.2, 2009, pp. 122-133.

[5] Baye A., Monseur C., Gender differences in variability and extreme scores in an international context, *Large-scale Assessments in Education*, Vol.4, No.1, 2016, DOI 10.1186/s40536-015-0015-x.

[6] Martin M.O., Mullis I.V.S., Foy P., Stanco G.M., *TIMSS 2011 International Results in Science*, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College, 2012.

[7] Mullis I.V.S., Martin M.O., Foy P., Arora A. *TIMSS 2011 International Results in Mathematics*, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College, 2012.

[8] Holland P.W., Thayer D.T., Differential item performance and the Mantel-Haenszel procedure. In Wainer H. and Braun H. (ed) *Test validity*, Hillsdale, NJ: Erlbaum, 1988, pp. 129-145.

[9] Zumbo B.D., Three generations of DIF analyses: considering Where it has been, Where it is now, and Where it is going, *Language Assessment Quarterly*, Vol.4, No.2, 2007, pp. 223-233.

[10] Dorans N.J., *ETS Contributions to the Quantitative Assessment of Item, Test, and Score Fairness*, Educational Testing Service: Princeton, New Jersey, 2013.

[11] Linacre J.M., *A user's guide to Winsteps*, 2010. http://www.winsteps.com/winman/index.htm?guide.htm.

[12] Gierl M.J., Rogers W.T., Klinger D.A. Using statistical and judgmental reviews to identify and interpret DIF, *The Alberta Journal of Educational Research*, Vol.45, 1999, pp. 353-376.

[13] Abedlaziz N., Ismail W., Hussin Z., Detecting a Gender-Related DIF Using Logistic Regression and Transformed Item Difficulty, *US-China Education Review*, Vol.5, 2011, pp. 734-744.

[14] Sullivan A., Academic self-concept, gender and single-sex schooling, *British Educational Research Journal*, Vol.35, No.2, 2009, pp. 259-288.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)