COMPARABILITY OF SLOVENIAN GENERAL MATURA MATHEMATICS EXAMS AT BASIC AND HIGHER LEVEL

Aleš Drolc[1], Janez Žerovnik*[2,3]

[1]RIC, Državni izpitni center, Kajuhova ulica 32 U, SI-1000 Ljubljana, Slovenia

[2]University of Ljubljana, FME, Aškerčeva 6, SI-1000 Ljubljana, Slovenia

[3]Rudolfovo - Science and Technology Center, Podbreznik 15, SI-8000 Novo mesto, Slovenia

Abstract

At the General Matura exam in Slovenia, mathematics is obligatory subject, the candidates can only choose among two levels. Until 2020, the exams at two levels were treated as two practically independent exams. The recent reform brought a solution that allows well-grounded comparability between the achievements. We outline some details of the approach and discuss preliminary results after first three years.

Keywords: *external exam, comparability of achievements, item response theory.*

1. INTRODUCTION

The General Matura exam in Slovenia (GM) is a school-leaving exam required for the completion of secondary education and for admission to undergraduate university courses. It is not only a final exam but also regulates the transition from secondary to tertiary education, similar to education systems in many European countries.

Here we focus on the mathematics exam that has been recently reformed. The motivation for the changes is explained first in section Motivation. Section 3 recalls some basic notions from testing theory that partly inspired the changes. The new structure, which consists of two exams sharing a common part, along with a procedure for synchronous definition of grade boundaries is explained in sections 4 and 5. Some preliminary statistics are provided in Section 6. The last section concludes the report and mentions an open question that may be of considerable interest.

2. MOTIVATION

First, we recall some basic facts regarding the GM in Slovenia (Ric, 2024). We mention some issues that need to be addressed, some of which are considered in this text. The first goal of GM is to define national standards for secondary schools. By passing the GM, candidates prove that they have achieved the standards of knowledge as determined by general upper secondary education programs and, by doing so, have successfully completed secondary education.

The importance and public interest in GM are sizeable because it is also a formal test of candidates' abilities for academic studies. By passing the GM, candidates prove that they can be admitted to undergraduate university courses. Legally, in Slovenia, the pass in the GM is a general admission requirement for any undergraduate university course and a minimal admission requirement for those undergraduate university courses with restricted admission. This fact implies major public interest in the GM regulations.

Regarding the selection of courses by candidates, in accordance with the Higher Education Act, achievement in the GM and achievement in the last two years of schooling are taken into consideration where there is restricted admission. In addition, there are other requirements for admission to certain programs, defined by the universities. For example, special talents or psycho-physical aptitude can be assessed by entrance examinations set by universities.

GM consists of five exams, three among them are compulsory (mother tongue, mathematics and foreign language). The mother tongue exam is offered only at one difficulty level and graded with grades 1-8, while mathematics and foreign language can be taken at either of two difficulty levels, foundation or basic level (BL, grades 1-5) or higher level (HL, grades 1-8). Optional subjects can be taken only at the BL. The maximal achievement at GM is thus 34 points.

Due to importance of the GM, significant effort is provided by the The National Examinations Centre (Ric), the National Committee (NC) for the GM, and national exam committees for the GM (NECs) to ensure equal conditions for all candidates, who sit the exam simultaneously following the same rules and procedures and are assessed in accordance with the same assessment criteria. There are several issues that are beyond reach of Ric and the committees, and which may relativize the uniform exam conditions. For example, all the exams have, in addition to the external part, an internal part that is left to responsibility of the schools. The internal part is worth 20% of the total possible score, for all exams taken. Statistics show very weak correlation between the external and internal parts in some cases (Zmazek, 2012). There are several other things that do not contribute to equal conditions to all candidates. Namely, the possibility of choice of the second foreign language, and in particular the choice of two optional subjects can possibly make a significant difference. Second, each year there is an Autumn exam series for candidates that for various reasons were not able to attend the Spring exam series. Below we emphasize two further issues that motivated the recent reform of the mathematics exam~~s~~.

2.1. Basic and higher level of some exams

As already mentioned, the mother tongue exam can only be taken at one level while the other two compulsory exams come at two levels. Each candidate can choose between two levels, but the information on which they can base their decision is rather unreliable because it depends on several factors that may even not be known in advance. In this paper we will discuss the currently used solution to the problem of comparability of achievements at the BL and HL of mathematics. The approach that gives some less uncertainty for the candidates has been first adopted in GM 2021, when the reform of the mathematics exams was first adopted following the new math exam catalogue (Banič et al, 2019).

Comparability of achievements at BL and HL of mathematics at GM has been recognized as an important question early. The matura catalogue for mathematics (Legiša et al, 1993) that was valid for the first matura in 1995 writes "Če kandidat, ki se je prijavil za izpit na višji ravni, ne opravi izpita, bi pa ga po številu točk in ustreznem vrednotenju opravil na osnovni ravni, se mu avtomatično prizna izpit iz matematike na osnovni ravni." Google translates it to "If a candidate who applied for a higher-level exam does not pass the exam, but would have passed it at the basic level based on the number of points and the appropriate evaluation, he or she will automatically be recognized for the maths exam at the basic level." The statement repeats in the catalogue for matura 1996, but disappears in the next catalogue for matura 1997. Thus, the rule was abandoned very soon, for reasons unknown to the present authors. The question however remained to be alive although not as a primary problem. There were only some discussions of the issue in the academic community. As the problem seems to be of a local nature, specific for GM in Slovenia, the publications are mainly in Slovene language, see (Cankar et al, 2020), and the references there. A study available in english (Hauptman, 2010) has unfortunately some serious drawbacks as argued in (Žerovnik, 2015).

In practice until 2020, exams in mathematics at BL and HL were treated as two different exams. On formal level, the only thing they had in common was that both exams were determined by the same subject specifications and that the exam-papers were prepared by the same committee. As exam papers for both levels were always written simultaneously, it was possible to include in the first ~~exam-paper~~ for HL the same tasks as used in the ~~exam-paper~~ for BL. (There was only one ~~exam paper~~ at BL.) Indeed, the first test on the BL and the first test on the HL were always identical, only the time allowed for solving was slightly different.

The level of the exam in mathematics is chosen by the candidates themselves, so it is very important for them to achieve a score suitable for their knowledge and skills. In many cases the decision is conditioned by the ambitions of candidates to apply for popular university programmes. The decision on the choice is for the candidates by no means easy. Even their teachers are faced with dilemmas on how to advise pupils. From the point of view of knowledge of mathematics, it is definitely good that more high school graduates decide on a HL as this means that they will study more material, as a result of which they will have more knowledge and will be better prepared for demanding studies. In order to gain more knowledge and invest more time, these candidates generally have to get at least as many points in the result as they would get if they chose the BL. Unfortunately, in the past, after choosing the HL, candidates (too) often got the feeling that they did not make the right decision, and although objective data usually does not confirm this, such opinions influence the decisions of the following generations. In reality, there were indeed a few candidates whose abilities were not sufficient for the HL exam. Consequently, they failed to pass and would possibly pass the exam at the BL, too. Therefore, it makes sense to synchronously assess both levels in a way that will ensure with greater reliability that the candidate will receive the same assessment for the same knowledge, regardless of their choice of the difficulty level.

The aim to assist the candidates to be better informed before choosing the difficulty level was important motivation for the NEC for mathematics to plan the reform many years ago. Another reason is that it is believed that it makes much sense to compare different generations. The committee always tries to achieve as much as possible by carefully preparing exam papers and by the annual process of setting grade boundaries. The important goals include (Erker et al, 2020):

• comparability of grades, regardless of the choice of level (BL or HL), the candidate should receive the same grade for the same knowledge,

• comparability of grades between the spring and autumn exam series,

• comparability between generations.

The first and the third goal are hopefully well solved by the solution presented here.

To the best of authors' knowledge, mathematics is offered at different levels at external state exams in some other countries, and the levels are often treated as independent exams. The present authors hope that this presentation may trigger some future exchange of ideas based on practical and theoretical solutions of similar cases that may exist and we are not aware of them.

2.2. Use of technology at math exam

The changes in 2021 also addressed the use of technology (calculator) that was very hard to handle due to fast improvements of the digital technology on one, and the rigidity of the legal regulations on the other side. Namely, in order to protect the candidates against sudden changes that may affect their successful preparations, any change in the rules has to be known two, in some cases even four years in advance. »Known« in many cases means published after being approved according to a strict formal procedure. For example, the reform of math exam in 2021 was published in 2019, has been formulated in detail at least a year earlier, after being in preparation for several years before. It is clear that it is impossible to predict advances in the digital technology for five or more years in advance. The issue is mentioned here because it is the reason to divide the external exam in two parts, called first and second exam-paper. In this way, by not allowing use of technology in one of the exam-papers, at least »half of the problem« regarding the use of technology has been solved.

We just mention in passing that the definition of »matura calculator« is used to restrict the use of electronic calculators that may solve too many mathematical problems automatically and in particular that may put the candidates in an uneven situation. Most importantly, the definition forbids use of any communication. Furthermore, some advanced features such as symbolic calculations are not allowed. However, it is obvious that it is very difficult to impose a complete set of rules that would answer any dilemma that may appear five years later.

Here we do not go deeper into the discussion about the role of technology in teaching mathematics. We just wish to say that the issue is worth to be discussed comprehensively. There is no doubt that use of information and communication technology in class may be very welcome, as well as in mathematics lessons. With its help, students can quickly get to understand a complex concept. On the other hand, it is particularly important that the use of such technology does not prevail in mathematics lessons. It can serve as an aid in the study of certain properties; with its help, we can get ideas for the formulation of theorems, but in mathematics we cannot only use technology to establish its theory. The very essence of mathematics is the idea that formulas and theorems must be proved using axioms and assertions that have been previously derived and proved. Logical reasoning and formal language should not be ignored, but on the other hand, they should not prevent the majority of students from successfully finish their secondary education.

It is important for students to get familiar with modern tools as they will also use information and communication technology in their professions, whether it is a profession in the social sciences, natural sciences, engineering, or some other field. In many professions, students will use such technology to help themselves solve mathematical problems. Since it is very important that, in addition to the use of such information and communication technology, the user can also critically interpret the obtained results, it is important that in high school the student learns at least some of the basics of mathematics, on which the tools that are or will be offered by technology are based. Only understanding the basic principles allows educated use of advanced tools. Therefore, with the matura exam in mathematics, we primarily wish to test the basic knowledge of mathematics, and to a lesser extent the skill of using information and communication technology for solving mathematical problems.

The matura exam with two exam papers enables greater adaptability to rapid changes in technology in the future. It should be emphasized that the matura exam is the end of the high school program, it tests the level of comprehension the contents and goals that are defined in the curriculum and that the future graduates studied over their four years of classes at the high school. Any significant change, for example the introduction of tasks that would test the non-trivial use of technology at the final exam, must therefore be introduced gradually and very thoughtfully. The current structure of the math exam however allows significant changes in the field of applied mathematics including advanced use of technology for solving mathematical problems.

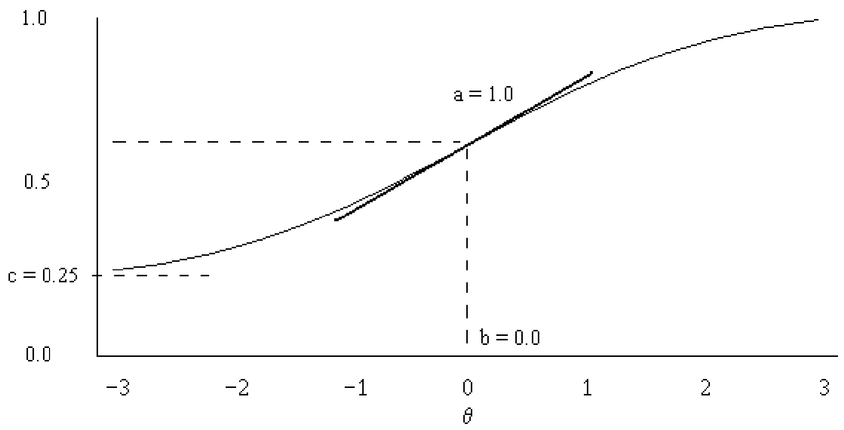
3. MORE PRELIMINARIES

The procedure for assigning grades from scores that is outlined below is based on the ideas from item response theory. Item response theory allows valid ranking of candidates that do not answer the same set of questions/items. The situation we have here is similar, although the items here are long »questions« corresponding to the parts A, B, and C of the exam.

3.1. Classical test theory and item response theory

Classical test theory (CTT) originates from Charles Spearman (Nunnally and Bernstein, 1994), who at the beginning of the twentieth century, while searching for a better way to calculate the correlation coefficient, found that every measured measurement X can be divided into a 'true measurement' T (true score) and the associated measurement error E (error). The true value of T is hypothetical (unknown) and can be defined for one individual as the average value of the measured achievements (X) of the same individual in an infinite number of repetitions of the same test. CTT suffers from three major theoretical shortcomings: (1) it assumes a linear relationship between the latent variable and observed scores, which rarely represents the empirical reality of behavioural constructs; (2) the true score can either not be estimated directly or only by making assumptions that are difficult to be met; and (3) parameters such as reliability, discrimination, location, or factor loadings depend on the sample being used (Rusch et al, 2017).

Three of the pioneers of Item Response Theory (IRT) were the Educational Testing Service psychometrician Frederic M. Lord, the Danish mathematician Georg Rasch, and Austrian sociologist Paul Lazarsfeld, who pursued parallel research independently (Wikipedia, 2024). Item response theory attempts to model the relationship between an individual's ability (or knowledge) and item difficulty. It shows the difficulty of the item and the knowledge of the individual on the same measurement scale, which enables comparison. The probability of an individual's correct answer to a given item depends on the difference between the item's difficulty and the individual's knowledge. The item parameters change the shape of the standard logistic function from which either the difficulty of the question or the ability of the tested person can be observed (see Figure).



By Iulus_Ascanius (talk) (Uploads) - Public Domain,
https://commons.wikimedia.org/w/index.php?curid=33919815

For more details we refer to the literature, for example (Andrich, 1988) or (Wikipedia, 2024) and the references there. Let us only summarize some important features of the Rasch model (Andrich, 1988):

• We can compare item difficulties independently of knowledge individuals and the knowledge of individuals can be compared independently of the difficulty of the items.

• The total sum of the results by items is a sufficient statistic (contains all the necessary information) to compare the items.

• We do not learn any additional information from the sample responses to the items (if we know the total achievement).

4. FORMAL FEATURES OF THE EXAM

In Slovenia, mathematics is an obligatory subject of GM. It can be passed on two levels, BL-and HL. The HL exam brings up to 8 points to the GM score while at the BL, the maximum is 5. Until GM 2022, mathematical exams on the two levels were treated as two independent exams, although there was an overlap of the tasks. More precisely, the grades were given to the candidates based on the slightly corrected normal distributions for each exam separately. An important question, both for candidates and for the analysts, was how the grades from both levels can be compared. Below we provide the structure of the external exams since 2021, and later explain the procedure for assigning grades to candidates.

4.1. Structure

The external part of the exam at both levels consists of two exam papers. In the first exam paper the use of the calculator is not allowed, while in the second one the calculator is a permitted accessory. The time to solve each exam paper is 90 minutes, which means that the total duration of the written part of the exam is 180 minutes, with a 30-minute break between the two question-papers. Three types of assignments are used. Candidates at the BL solve short tasks and short structured tasks, while candidates at the HL solve short structured tasks and structured tasks. Here we do not give the precise definitions of these types of tasks, let us just say that the difference is both in length, taxonomy, and content-related objectives.

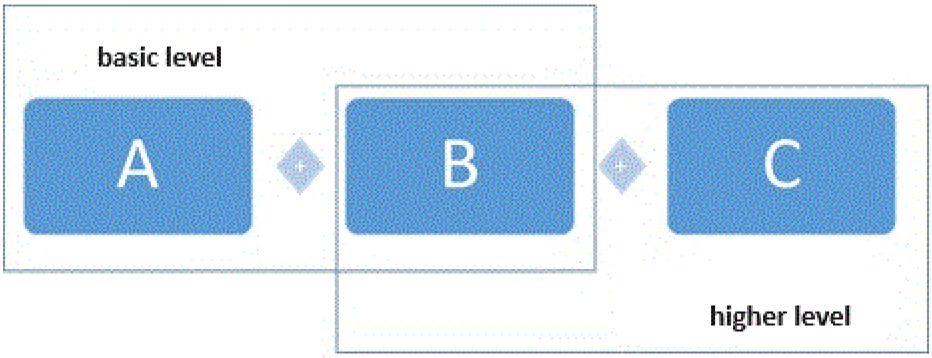


Fig. 1. Structure of the mathematical exam

Fig. 1 shows the structure of the exam at the basic (BL) and higher level (HL). The structure repeats as two exam-papers at each level. At both levels, the use of technology is not allowed for the first exam-papers while for the second a computer satisfying the definition of »matura calculator« is allowed. More precisely, part A brings short tasks, part B shorter structured tasks, and part C structured tasks. Part B is a common part for both levels, supplemented by part A at the BL and part C at the HL. Thus at the BL, the first exam-paper consists of parts A1 and B1 (to be solved without the use of calculator), and the second exam-paper consists of parts A2 and B2 (calculator is a permitted aid). At the HL, the first exam-paper consists of parts B1 and C1 (to be solved without the use of a calculator), and the second exam-paper consists of parts B2 and C2 (calculator is a permitted aid).

Some important features of the parts of the exams are:

• parts B1 and B2 are evaluated with 40 marks, and parts A1, A2, C1 and C2 with 20 marks each, which means that the candidate can in total score 60 points in each exam paper;

• 90 minutes are available for each exam paper;

• each exam paper is 40% of the exam in terms of time and number of marks (there is an oral exam that adds 20% to the final score);

• the expected difficulty of the tasks of part B corresponds to the standards expected to be reachable by the majority of candidates at all schools;

• the expected difficulty of part C corresponds to traditional understanding of HL mathematics, and in addition it also covers some material that is not obligatory for all candidates;

• part A consists of short tasks, most of them of lower taxonomic levels.

5. FROM SCORES TO GRADES

After the exam is completed, it takes about two weeks for coordinated e-marking performed by a large group of examiners, experts who are mainly secondary school teachers. Currently Ric uses e-marking platform RM Assessor[3]. The next step is to use the marks of the external part and internal part to assign grades. Recall that at BL, the grades from 1 to 5 (5 means excellent) are given, while at the HL, the »grades« are from 1 to 8.

When determining the boundaries between grades, the two criteria are used: the absolute or content criterion and relative or statistical criterion. As the populations are large enough, the statistical criterion prevails. The content criterion is heavily used in preparations of the exam papers, which need to be well balanced among the years and each of them must be a blend of tasks of various difficulties that cover most of the material to be tested. The statistical criterion allows good balance among generations. To the contrary of the praxis until 2020, ideas based on item response theory are adopted for ensuring a level of comparability among the two levels. The three parts of the exams are considered as three items, A, B and C. After considering the results of all population on the item B, the abilities of each of the subpopulations are obtained. These are in turn used to evaluate the difficulty of items A and C, allowing to compare the final results of each candidate. In reality, the internal (oral) exam is also taken into account together with items A and C. The procedure is however intuitively quite natural and easy to understand. Below, it will be explained in an informal language.

P*: Procedure FROM SCORES TO GRADES

INPUT: the vector of size (in %) of each class. For example: (0.030, 0.155, 0.245, 0.279, 0.165, 0.047, 0.044, 0,045).

1. Based on the histogram of scores of the total population on part B, determine the boundaries between classes (see Fig. 2).
2. From the boundaries between classes, compute the sizes of classes (in %) for two subpopulations, BL candidates and HL candidates (see Fig. 3).
3. a) The sizes of classes of BL population give boundaries between classes (in points) (see Fig. 4, left). Merge the classes 5, 6, 7, and 8 as all will be awarded grade 5.
b) The sizes of classes of HL population give boundaries between classes (in points) (see Fig. 4)

OUTPUT: The result is two sequences of boundaries. For example: (46, 68, 83, 99, -, -, -) for BL and (44, 49, 61,78, 91, 97, 107) for HL.

The input vector can be (and indeed is) based on history (statistics previous 3 years). For example, the first class can be defined as the average of negative grades over the last five years. This ensures that the distributions of awarded grades do not change significantly among generations.

The output are two sequences of bounds that contain the numbers of points that are needed to get a grade. In the example above, the numbers are rounded to integers for clarity. In reality, the result of procedure is a rough estimate, proposal for the committee to finally determine the boundaries. The least that has to be done is to round the boundaries to integers. Some further modifications are in principle possible based on the content criterion. Good news is that in the first three years the procedure provided estimates that hardly needed any modification beyond rounding. However, there is one exception, namely the number of HL candidates that are given grades 1, 2 and 3 is very small. Therefore, the boundary between grade 1 and 2 has been always determined after detailed inspection of the candidates involved. See Fig. 3 (right) and Fig. 4 (right).

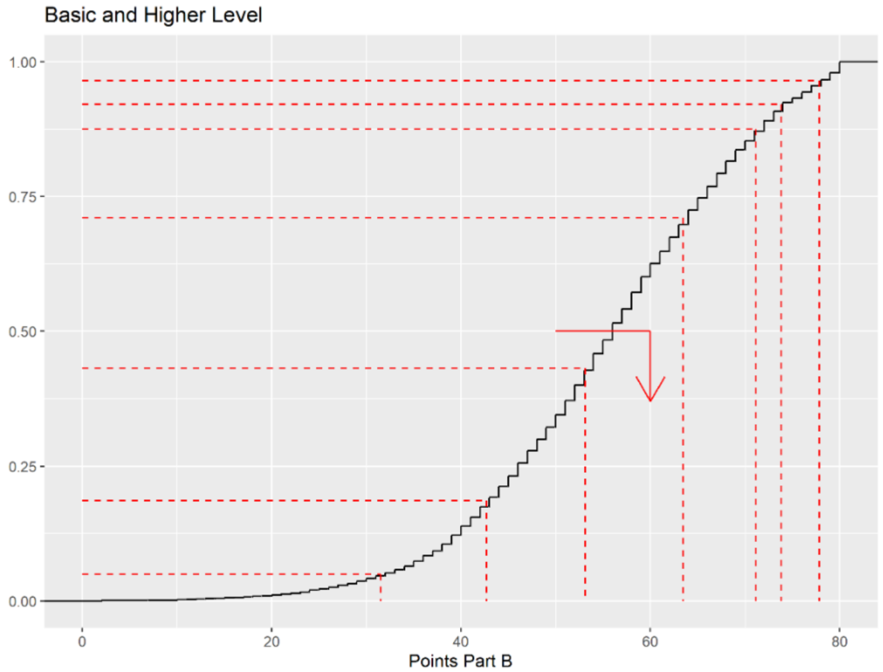


Fig. 2. Empirical cumulative distribution function of achievements on part B, all candidates. Horizontal lines are obtained from input vector that gives the desired/expected shares of grades among the whole population (includes BL and HL candidates). The vertical lines define borders. For example, 18.6 % of candidates are expected to get grade 1 or 2, the line hits the horizontal axis at 42.696 (decimal digit due to interpolation). This implies that candidates that will get grade 3 or more, must on average get more than 42.696 points on part B.

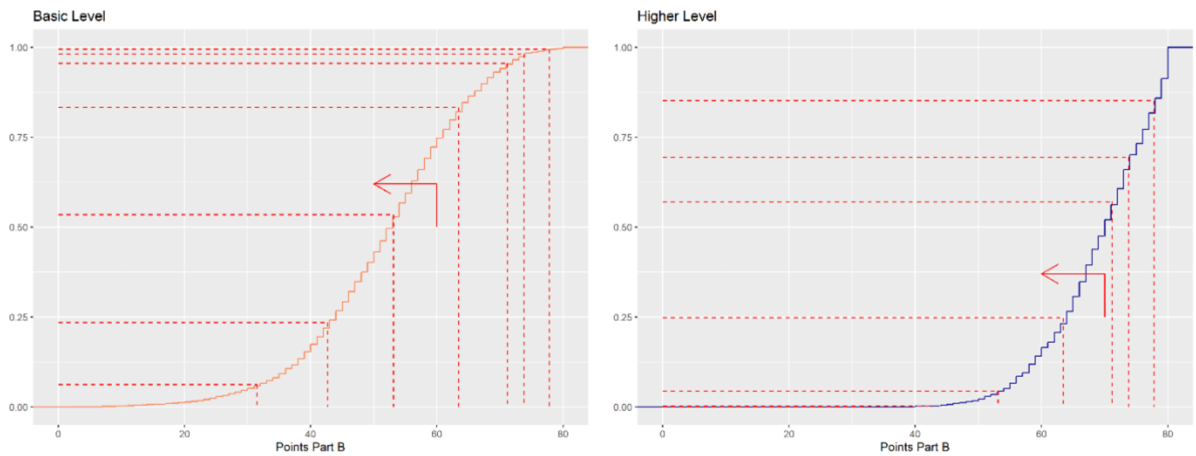


Fig. 3. Empirical cumulative distribution functions of achievements of the two subpopulations on part B, BL candidates (left) and HL candidates (right). Vertical lines are defined by the borders obtained in step 2 (and are drawn in Fig. 2). Horizontal lines define the shares of the population that will be given a particular grade.

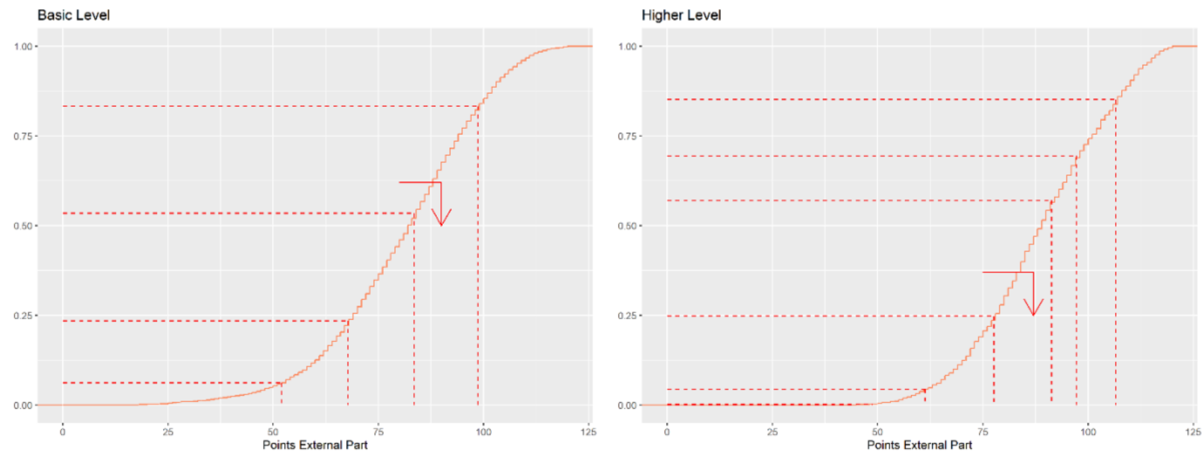


Fig. 4. Empirical cumulative distribution functions of achievements of the two subpopulations on both parts, parts A and B for BL candidates (left) and parts B and C for HL candidates (right). Horizontal lines are obtained from Fig. 3 and as on Fig. 2, the vertical lines define borders. For example, for BL (left), 23.5 % of candidates are expected to get grade 1 or 2, the line hits the horizontal axis at 67.796 (decimal digit due to interpolation). This implies that candidates that will get grade 3 or more, must on average get more than 67.796 points on parts A and B.

6. SOME PRELIMINARY STATISTICS

GM is the final exam for secondary schools excluding the vocational secondary schools. However, there are some other groups that are admitted to GM exams, under some conditions regulated by the law. For example, the students of vocational secondary schools who want to enter university are allowed to sit the GM exams. For statistical analysis, Ric uses a subpopulation that is called the reference group which consists of the candidates who regularly finish a secondary school, and sit all GM exams for the first time in the year studied. In the sequel, we show that the populations at the two levels are significantly different in abilities. Then we give some rough comparisons of grade distributions among generations.

6.1. The two subpopulations are different

First, we show that the two subpopulations of candidates are indeed different. The scores on part B (maximum 80 points) prove that the abilities of two randomly selected candidates from the two populations are expected to be different. Indeed, the two histograms on Fig. 5 indicate that the probability distributions obviously have different means.

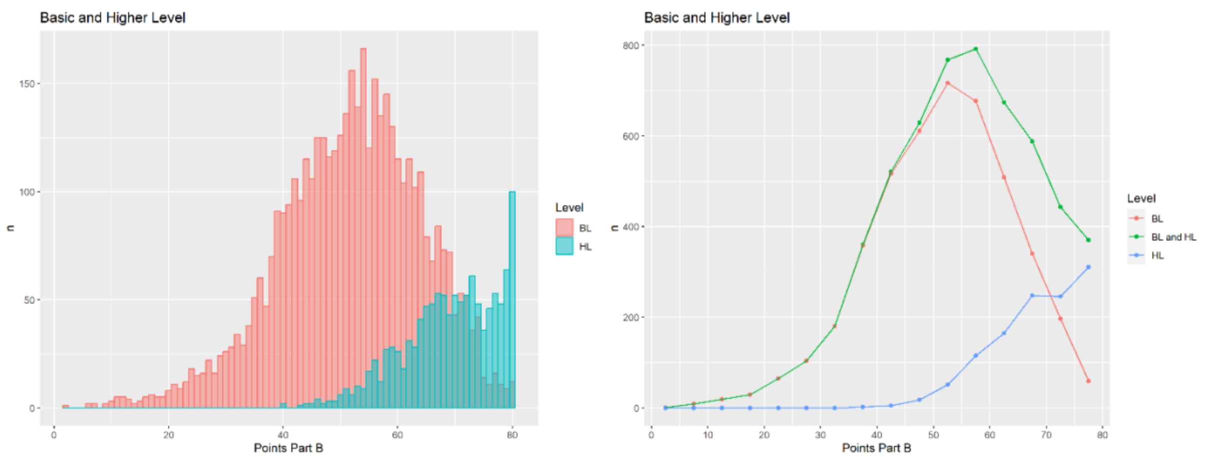


Fig. 5. Histograms of achievements of each of the population on part B. Achievement of the BL population in red, and HL population in blue. Data from exam 2023.

The procedure P* explained above gives a new ranking based on scores including items A and C. The frequencies of grades that would be awarded are on Fig. 6.

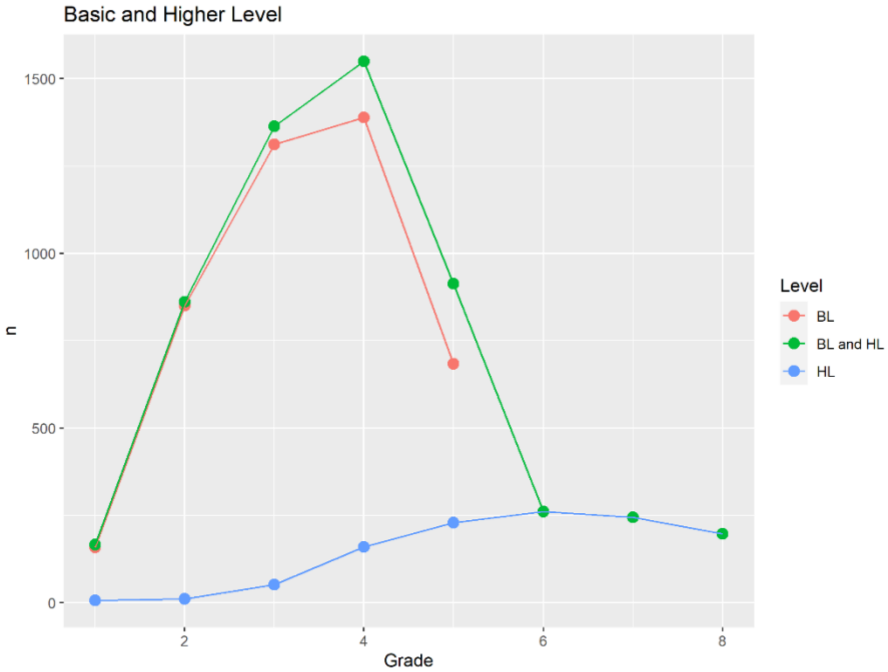


Fig. 6. The grades awarded according to our procedure, if only the external exam score is considered. (Parts A, B, and C only, the internal part of the exam is ignored here.) In reality, the scores of internal part are included, which only slightly changes the graph.

As the populations are obviously different in terms of ability tested on part B, the results given on Fig. 6 are expected and well justified. It should be noted here that by legal regulation that is still valid, the internal part has to be included in the final score. Practically, this means that instead of cumulative score on parts B and (A or C), the total score of the two parts plus the points awarded at the internal exam is taken into account. This implies that the horizontal axis extends to value 150 instead of 120. Otherwise, the procedure is identical to the one explained here.

6.2. Comparison among generations

The next table gives a comparison among the last three generations that have passed the GM exam according to the new rules, the grades awarded using proposition procedure **P***. As the number of candidates in each generation varies, the graph shows the percentages of population that achieved a grade.

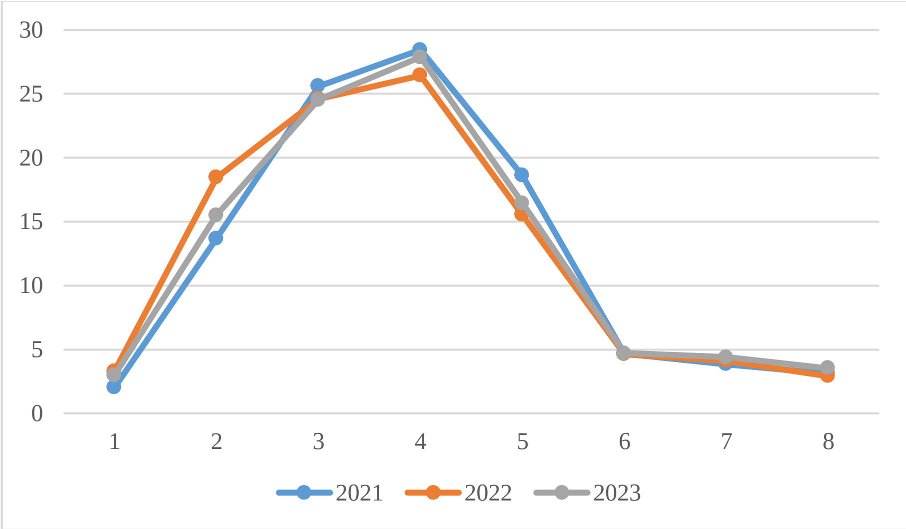


Fig. 7. The shares of grades awarded, the three years when the procedure P* was used. Total population, including BL and HL candidates.

We wish to note that it is important to have some fairness among generations because it is not uncommon that a candidate who passed GM in one year competes for entrance to university a year later with another generation. As the populations count several thousands, it is reasonable to assume that the abilities of the whole generations are equal, hence the overlap of the graphs through generations is expected and fair.

Fig. 7 shows very good overlap over the last three generations. Here it may be worth to say that the generations 2021 and 2022 have, due to pandemic, enjoyed several adaptations of the rigid rules. It is therefore questionable how to interpret the results correctly. Perhaps one can say that at least the grades at math exam at GM were reasonably fairly assigned to candidates in these generations. Maybe it also proves the robustness of the approach explained here, very important being the choice of input shares that has been based on statistical data from previous years.

We wish to emphasize that for a deeper analysis, data of some more years are needed for better grounded conclusions. Therefore, we just present the data here without any advanced statistical analysis. Another reason is that the last years were untypical because of the pandemic.

To get a more complete picture, let us have a look at comparability among some previous generations. Graphs on Fig. 8 and Fig. 9 are based on data from years 2018-2020 and 2015-2017.

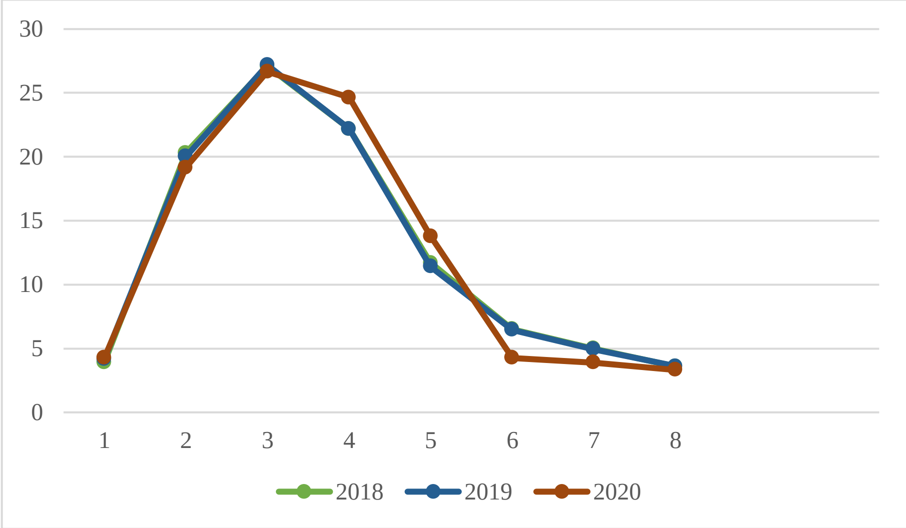


Fig. 8. Grades awarded, years 2018-2020. Total population.

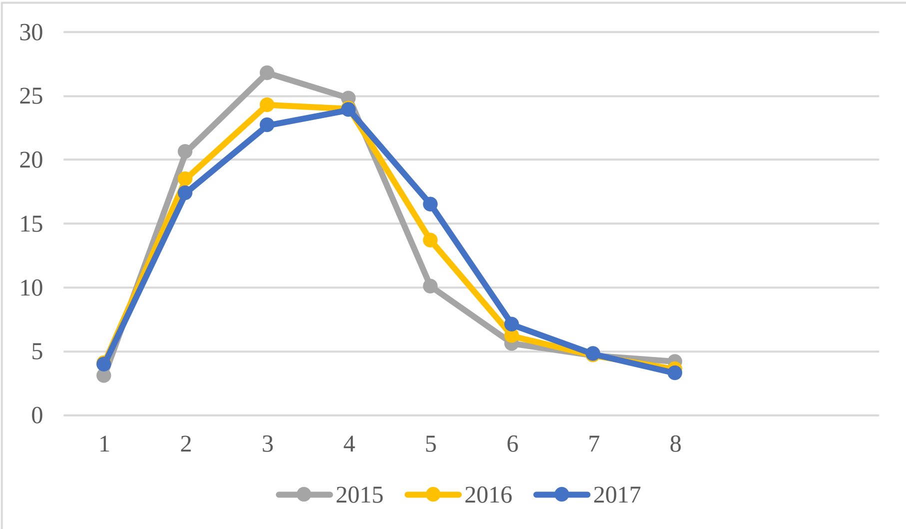


Fig. 9. Grades awarded, years 2015-2017. Total population.

Both graphs show reasonably small variations among three consecutive generations. It may mean that the committees in the past did a good job. Let us compare the past distributions (Figs. 8 and 9) with Fig. 7. Having in mind the limitations mentioned above, it seems obvious that the variation in the last three years is smaller, hence indication that the distributions of grades is become more stable. Finally, Fig. 10 shows that the overlap of distributions for the last nine years is quite good, which still indicates stability.

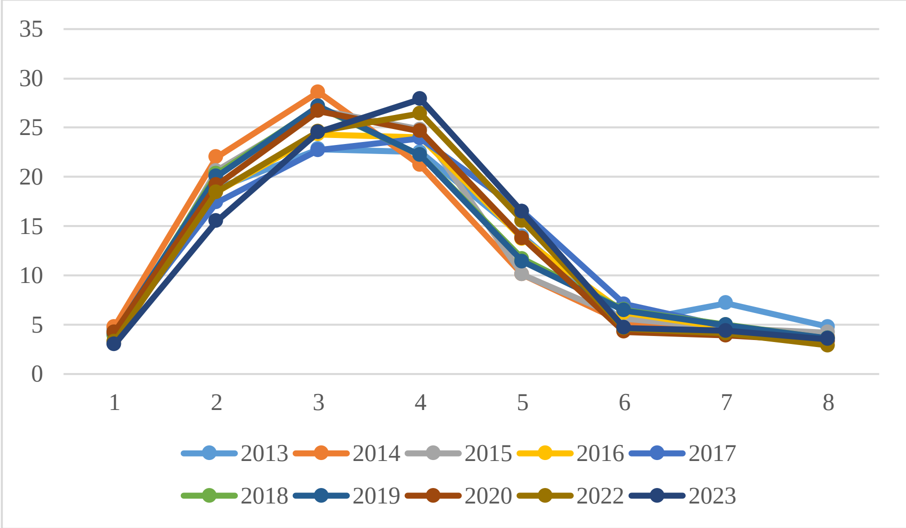


Fig. 10. Grades awarded, year 2013 until 2023. Total population.

6.3. Differences among the two models for candidates

It is worth to recall that before 2021, when the two levels were treated as two independent exams, it was possible that a candidate would get a better grade if he or she chose BL exam instead of HL exam. This indeed did happen in some cases, in particular for candidates with very poor achievement at the HL exam. Such candidates were never many. In addition, there was always some uncertainty when defining the border values for higher grades as there was no well-grounded hint how to assign comparable boundaries at the two levels for grades 3, 4 and 5. This of course may have had affected many more candidates.

In the new model, of course we still have no guarantee that each candidate would achieve the same grade irrespectively which level exam he or she takes. Such an expectation is of course too strong as the final score must also depend on the performance which the candidate has shown on the other parts of the exam. However, the new procedure assures that on average, the candidates that fall within, say, achievement for grade 3 at the common part, will on average indeed be awarded grade 3, the average taken over the whole population.

7. DISCUSSION AND CONCLUSIONS

The new structure of math exams on BL and HL of GM in Slovenia enables well-grounded comparison among the two exams. In particular it is possible to adopt a procedure (P*) that allows to synchronize the grade boundaries in a way that gives certain guarantee that a candidate can expect receive the same grade whichever level of exam he or she takes. In other words, a set of candidates will receive the same average grade on both levels.

When comparing the generations, it is observed that the distribution of grades is quite stable, thus being fair also in this regard. This holds also for the period before the last reform.

The question that remains unsolved, and that seemingly is not easy to answer, is the fairness, or comparability of results between the Spring and Autumn exam series. The latter is known to be passed by a different population. Namely, the autumn population includes a large number of students that did not pass the Spring exam series, but there is also a significant number of students from so-called reference group that had to postpone matriculation for various reasons. The blend of candidates may vary considerably from year to year. Item response theory in combination with a set of pretested items may be a possible avenue for approaching this challenge in the future.

REFERENCES

1. Andrich, D 1988, "*Rasch models for measurement*", Beverly Hills: Sage Publications.
2. Banič, I, Erker, J, Fošnarič, M, Grahor, A, Levstek, T, Škrlec, M, and Žerovnik, J 2019, *Matematika: predmetni izpitni katalog za splošno maturo*. Ljubljana: Državni izpitni center, 2019. https://www.ric.si/mma/M-MAT-2021/2019082714564660/.
3. Cankar, G, and Žerovnik, J 2020, "Primerljivost dosežkov na osnovni in višji ravni izpita iz matematike na splošni maturi" [Comparability of achievement at basic and higher level of mathematics at general matura]. *Psihološka obzorja*, 29, 158–169. https://doi.org/10.20419/2020.29.523
4. Erker, J, Fošnarič, M, Grahor, A, Levstek, T, Škrlec, M, and Žerovnik, J 2020, "Primerljivost izpitov na osnovni in višji ravni pri predmetu matematika na splošni maturi", *Obzornik za matematiko in fiziko*, 67, 1-11.
5. Hauptman, A 2010, "Equating of grades at basic and higher level of mathematics achievement", *Metodološki zvezki*, 7, 167–181.
6. Legiša, P, Pagon, D, Alt, Z, Hvastija, D, Jevnikar, M, Kranjc, B, Perat, Z, and Reisman, D 1993, *Predmetni izpitni katalog za maturo. Matematika*, Ljubljana: Republiški izpitni center.
7. Nunnally, JC and Bernstein, IH 1994, "*Psychometric theory*", New York: McGraw-Hill.
8. Ric 2024, The National Examinations Centre, *General information (on general matura)*, https://www.ric.si/en/general-matura/general-information/ viewed 22 May 2024.
9. Rusch, T, Lowry, PB, Mair, P, and Treiblmaier, H 2017, "Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory", *Information & Management*, 54, 189-203.
10. Wikipedia 2024, "*Item response theory*", https://en.wikipedia.org/wiki/Item_response_theory viewed 22 May 2024.
11. Zmazek, B 2012, "Primerjava rezultatov internih in eksternih delov izpitov splošne mature", in Žakelj A et al (eds), *Razvijanje in vrednotenje znanja : zbornik prispevkov s posveta, 7. in 8. december 2012, Maribor*. 1. izd. Ljubljana: Zavod RS za šolstvo, 13-14. http://www.zrss.si/pdf/Razvijanje, Digitalna knjižnica Slovenije - dLib.si.
12. Žerovnik, J 2015, "About use and misuse of statistics in education: On mathematics exams at general matura in Slovenia", *Education Practice and Innovation*, 2, 1–7.