Michał Bernardelli
Warsaw School of Economics
Warszawa

CORRECTION OF DEVIATIONS FROM THE PROBABILITY DISTRIBUTION OF EXAM RESULTS

KOREKTA ODCHYLEŃ OD ROZKŁADU PRAWDOPODOBIEŃSTWA WYNIKÓW EGZAMINÓW

Key words: grading regulations, matura exam, subjectivism of evaluation
Słowa kluczowe: kryteria oceniania, egzamin maturalny, subiektywizm oceny

1. Introduction

From an early age people are accustomed to evaluating and comparing with others. The newborn baby is evaluated based on five simple criteria in so called Apgar scale (well known as backronym: Appearance, Pulse, Grimace, Activity, Respiration[1]). Mothers are comparing their child with other children in terms of physical and mental growth. On the basis of how quickly a child learned to crawl, walk, speak, count, draw or write, often unauthorized conclusions are drawn on the future development and success in adulthood. On each phase of the education children are tested and compared to the group and other classes. Also in adult life there are methods to

[1] V. Apgar, *A proposal for a new method of evaluation of the newborn infant*, "Current Researches in Anesthesia and Analgesia" 1953, 32 (4), pp. 260-267.

measure efficiency at work, determine the health condition or simply the comparison to the friends' wealth and wellbeing. Comparisons are made between people from the same country, as well as between countries. Over the years measuring systems were developed in almost every area of life. On the one hand clearly defined standards are useful, especially for those unfamiliar with the area under consideration, on the other hand they are often too limited with a tendency to oversimplify the measured issues. Let us give few examples of standards and their limitations.

As a first consider the resting heart rate. For an adult human the norm ranges from 60 to 100 beats per minute[2]. Those are the standard values that could be easily identified by the non-expert in medical field. However the results could vary depending on many different factors like air temperature, body position, used medication, body size, age or fitness level. Accepted standards should be therefore treated as guidelines not strict rules. Values beyond the scale (like bradycardia) not necessarily mean, that there is something wrong with the body. Athlete's heart gets usually bigger, stronger and more efficient with exercise and as a consequence less beats per minute are required. There are many other standards that in fact are only a rough approximation of statistically recognized value in society and are not valid for exceptional, “outside the scheme” people. It is enough to mention such things as body mass index (BMI), intelligence quotient (IQ) or Cooper test.

Standards are usually connected with quantity values. It is difficult to put into standard qualitative data, like the eyes color. What seems even harder, is to measure intangible assets. Expensive, cheap, good, wise or hot – those are the terms that depends most of all of our subjective opinion. Without the possibility of comparison to other objects, memories or feelings from the past, it is impossible to give them the quantitative importance. As an example consider the feeling of happiness. Thinking globally about the national happiness as a development indicator, the gross domestic product may be used for the measurement purpose. It would appear that the richer the country, the happier people. This approach turns to be oversimplification. Of course money are important for many people, but there are other aspects, which should be taken into consideration. The famous happiness and wellbeing paradigm stated by Jigme Thinley[3], the Prime Minister of Bhutan, em-

[2] According to American Heart Association, see http://www.heart.org/HEARTORG.

[3] More in guide to Gross National Happiness Index on http://www. grossnationalhappiness.com.

phasize spiritual values instead of the material development. Based on various criteria (e.g. physical, mental and spiritual health, social, community, cultural and ecological vitality, education, living standards, good governance, and life expectancy) a lot of indicators of national happiness are developed, like Gross National Happiness (GNH[4]), OECD[5] Better Life Index (BLI[6]), The Canadian Index of Wellbeing (CIW[7]), Social Progress Index (SPI[8]) or Happy Planet Index (HPI[9]). The methodology in each case is different and results could be incomparable. According to the HPI ranking (year 2012), the leading country are Costa Rica, Vietnam and Colombia. Switzerland takes 34th place, Poland 71st and the U.S. is ranked 105. In contrast to the HPI, the SPI (year 2015) the countries ranked as best three are Norway, Sweden and Switzerland. Poland is 27th, the U.S. 16th and Costa Rica 28th. As has been pointed out many important to people values are difficult to quantify and different approaches lead to various conclusions. What is more, the indicators are usually biased by the subjectivism and varies depending on the evaluator. Therefore opinions and indicators are hard to compare with each other. Often complex conversion methods are being proposed to allow such comparisons. That leads to the last example: the grading systems in education.

Grading is inextricably linked to the education process. It should not be considered as a tool of repression or punishment, but rather as a form of information and motivation. It could be analyzed in the context of administrative or pedagogical. According to Dorota Klus-Stańska[10] from an administrative point of view the grade largely determines the opportunities and further progress in school and professional life. From a psychological perspective it is possible to determine what mental mechanisms have impact of an assessment process on the man. Regardless, the key aspect from the student point of view is the fairness of the grading system. There is a wide literature on the grading aspects in education in both Polish (Kosińska[11], Niemierko[12]) and

[4] http://www.gnhbhutan.org

[5] Organisation for Economic Co-operation and Development.

[6] http://www.oecdbetterlifeindex.org

[7] https://uwaterloo.ca/canadian-index-wellbeing

[8] http://www.socialprogressimperative.org/data/spi

[9] N. Marks, S. Abdallah, A. Simms, S. Thompson, et al., *The Happy Planet Index 1.0. New Economics Foundation*, London 2006.

[10] D. Klus-Stańska, *Komu potrzebne jest ocenianie w szkole?*, „Edukacja i Dialog" 2006, No. 5 (178) pp. 11-14.

[11] E. Kosińska, Ocenianie w szkole, Kraków 2000.

[12] B. Niemierko, *Pomiar wyników kształcenia*, Warszawa 2000

foreign languages (Black et al.[13]; Gardner ed.[14]). Many articles are devoted to fairness of the evaluation process (Kusiak and Wodnicka[15], Wiczkowski[16]). Undeniably methods of verifying knowledge and skills and ways of measuring it are important both from the point of view of the evaluating person, as well as people assessed. Worldwide there exist number of standardized measurements of student achievements. Grades can be assigned as letters (e.g. A through F in Mongolia, Hong Kong, Japan), as a range (e.g. 1 to 5 in Austria or Czech Republic, 1 to 6 in Poland or Norway, 1 to 10 in Vietnam or Moldova, 0 to 100 in Israel or India), as a ranking list (students' relative position among other students, like in South Korea), etc. "Credential evaluation, credit transfer and grade translation are among the most widely debated and highly sensitive issues in international education, and numerous approaches, solutions, models and formulas have been proposed over the years both in the United States and in Europe"[17]. The multitude of existing conversion methods[18] between different grading systems demonstrates that both comparison to others as well as the assessment itself, is contained in the modern human consciousness. In many cases the education process together with the positive evaluations, affect the whole future life of a young man. That is a main reason why the statistical analysis (from the global point of view) and the fairness of the evaluation (from the individual and global point of view) should be carried with preciseness and attention.

Matura as the high-school exit exam in Poland influences on further career and social position. Based on the results from matura, the selection of candidates to the colleges and universities is conducted. Hundred thousands of people is participating in matura each year and many analysis are made before and after the exam. This is the typical case, where the key aspects of measurement process, which were already presented in this paper, are applied. In this paper we focus only on one aspect of that process, namely the statistical analysis of the results of matura from the basic level in Polish lan-

[13] P. Black, C. Harrison, C. Lee, B. Marshall, D. Wiliam, *Working inside the black box: Assessment for learning in the classroom*, "Phi Delta Kappan" 2004, 86 (1), pp. 8-21.

[14] *Assessment and Learning*, ed. J. Gardner, London 2012.

[15] L. Kusiak, W. Wodnicka, *O ocenianiu słów kilka...*, Kwartalnik Metodyczny »Grono«" 2001, pp. 16-23.

[16] K. Wiczkowski, *Zza i sprzed katedry, czyli jak oceniać sprawiedliwie*, Ostrołęka 1994.

[17] G. Haug, *Capturing the Message Conveyed by Grades. Interpreting Foreign Grades*, "World Education News & Reviews" 1997, Vol. 10, No. 2.

[18] See international grade conversion guide for higher education at http://www.wes.org/gradeconversionguide/index.asp (retrieved on 20 May 2015).

guage. More precisely, the deviations from the probability distribution of exam results are described and propositions of introducing the mathematical correction are presented. The mentioned deviations may be observed near the passing threshold value. There are apparently much more participants, which were allowed to pass the compulsory exam from the Polish language by giving them an extra points, despite not reaching in fact the necessary score of at least 30%. This causes the results and conclusions unreliable. Of course on the one hand those statistics make a better picture of the situation of students, which passed the exam, but on the other hand it is not fair to other students, that get the same amount of points, without any unethical (consciously or unconsciously) help from the person evaluating the exam. These deviation could be also seen from another angle: weak students apply for higher education courses potentially taking places of people who are really entitled to them. What is more it is very likely that these people will have problems with graduation, losing unnecessary time and exposing the country for costs.

That is the reason why not only creating new standards and measuring systems, not only defining rules and guidelines to use them, but also the post-validation process is important for the correctness of reasoning. It allows to identify problems and implement changes afterwards. The aim of this research was to indicate a problem with grading regulation on Polish matura exam and present the method of dealing with that problem by artificial change in results of the exam.

This paper is composed of four sections. The short description of matura exam in Poland with exemplary results of the exam in Polish language are in section 2, whereas section 3 presents in detail proposed methods of correction of the results of the exams, making them more reliable. There have also been added descriptions of numerical experiments on the real-life data. This paper ends with the brief summary in section 4.

2. Matura exam in Poland

In the Polish education system at the end of the high school students take the exam, called matura. Passing this exam is necessary in order to apply to institutions of higher education (university, academy, etc.). The exams are conducted by the Central Examination Board (CKE[19]). They are usually held

[19] http://www.cke.edu.pl/

in May. Since 2005 the written part of the exam is assessed by independent examiners, who are supposed to be more objective than teachers from the same school. As of the school year 2013/2014, the matura consists of three compulsory exams at "basic level" in:

- Polish language (written and oral exam),
- mathematics,
- a selected modern language from the following list: English, French, German, Italian, Spanish or Russian (written and oral exam).

Student have to also take at least one subject at "extended level". This list is expanded with additional, compared to the basic level, subjects, for example physics, biology, chemistry, geography, social studies, history, information technology and many modern languages.

To pass the matura, student need to get at least 30% of all points in each of the three compulsory, basic level exams. The results of the additional, extended level exams are meaningless from the high school graduation point of view, but are usually taken into account while applying for higher education places.

Reports with statistical analysis of the results of the matura each year are published on the websites of the CKE. In this paper we confine ourselves to the results from the exam in Polish language. Basic statistics for the years 2012-2014 are gathered in the Table 1.

Table 1

Statistics concerning the matura in Polish language in 2012-2014

	Year 2012		Year 2013		Year 2014	
	Basic level	Extended level	Basic level	Extended level	Basic level	Extended level
Number of participants	345 159	35 183	329 043	34 770	301 781	29 814
Average	53.54	63.49	55.00	63.00	51.00	65.00
Standard deviation	15.45	15.77	16.00	17.00	17.00	16.00
Median	54	63	56	63	51	65
Success percentage	97	–	96	–	94	–

Source: own elaboration based on the reports of the CKE

In Figures 1 and 2 the histograms of the percentage of participants and corresponding results from the matura exam in Polish language in 2012 are presented. Figure 1 gives the information about basic level, whereas Figure 2 contains data about results of the extended level exam.

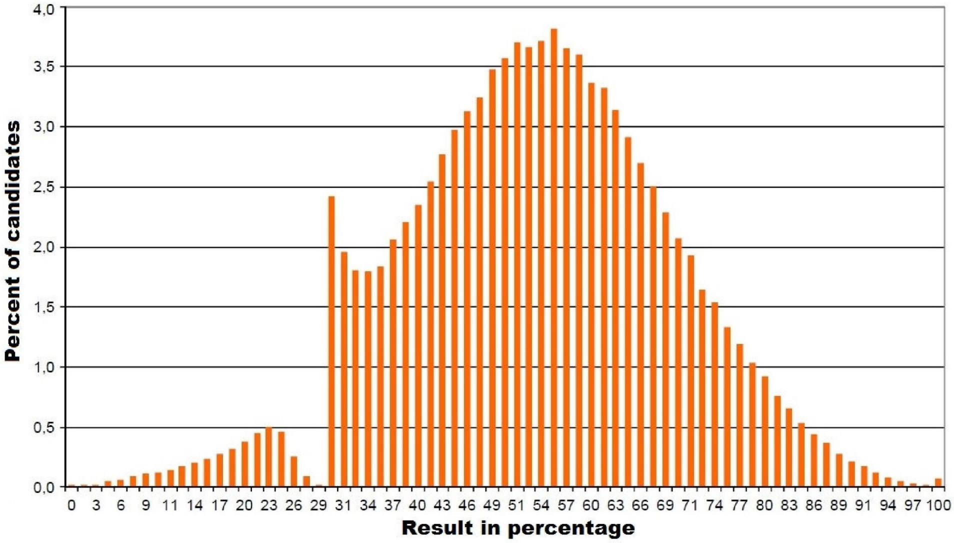

Figure 1. Results of the matura in Polish language in 2012 (basic level)
Source: own elaboration based on the reports of the CKE

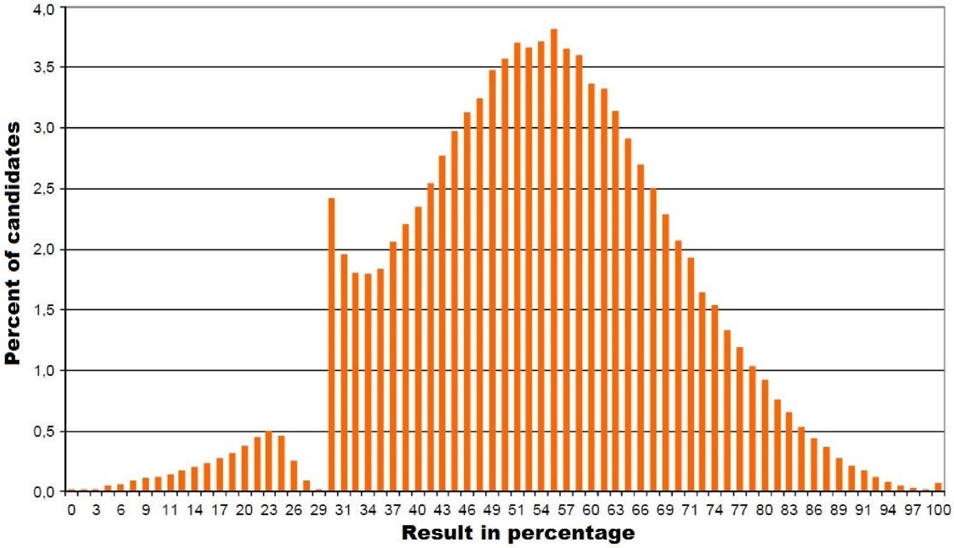

Figure 2. Results of the matura in Polish language in 2012 (extended level)
Source: own elaboration based on the reports of the CKE

There is visible inconsistency around the label "29" on the X axis at the histogram from the Figure 1. Not surprisingly it coincide with the threshold (30%) set by CKE. It seems like the assessments of independent examiners are not as objective as they should be. Likely many results, which in fact are below the threshold, are deliberately changed by adding some extra points to reach the limit of 30 points. It is worth to notice that there are no such discrepancies in case of the extended version of the exam. In this case however the limit of passing has not been determined and additional points are not important in terms of high school graduation. The similar situation may be observed the year later. Figures 3 and 4 present histograms with results on the respectively basic and extended level of the exam in Polish language.

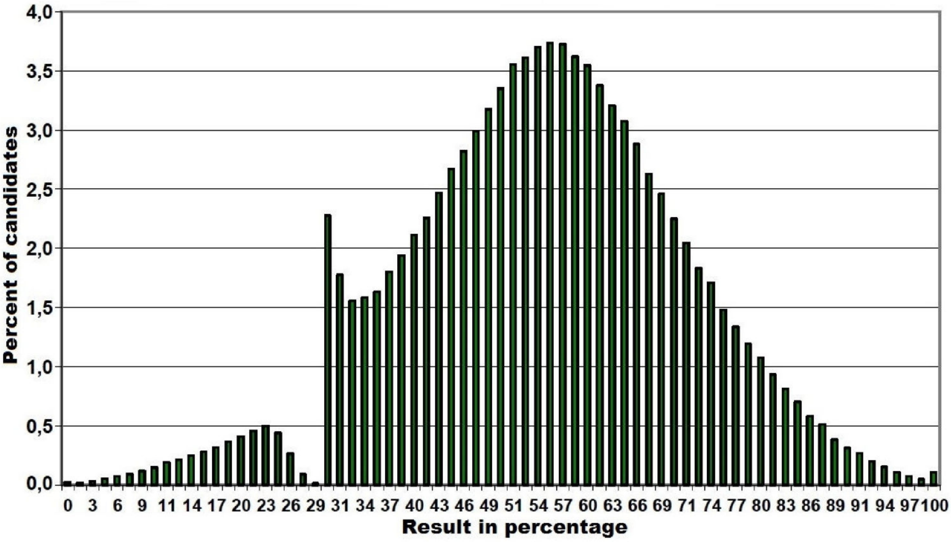

Figure 3. Results of the matura in Polish language in 2013 (basic level)
Source: own elaboration based on the reports of the CKE

The same kind of inconsistency near the 30% threshold can be observed. It provides the empirical proof for the non-accidental relationship and questions the objectivity of the examiners. It also highlights the differences between the results of the matura exams and the actual knowledge of graduates. The aim of this study is to present how the results may be transformed to reflect the actual knowledge state in a better manner. Those methods are described in the next section.

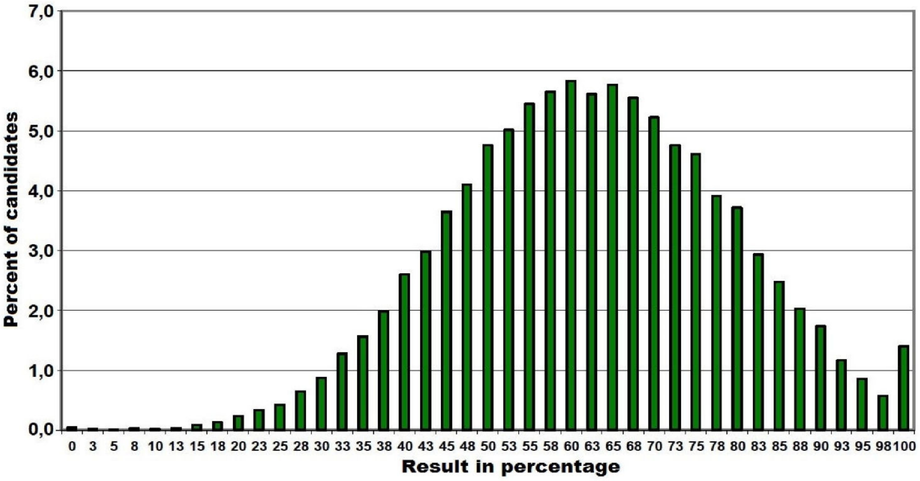

Figure 4. Results of the matura in Polish language in 2013 (extended level)
Source: own elaboration based on the reports of the CKE

The newest available data for the year 2014 are pictured in Figures 5 and 6. The analogous situation exists near the 30% score. It is even more visible than these from the previous years (compare Figure 1 and 3). The percentage of participants, who obtained exactly 30 points, is the highest of all among the possible 0 to 100 score and exceeds 3.5%. In contrast to the graph from Figure 5, the histogram for the extended level of the matura exam in Figure 6 looks quite smooth.

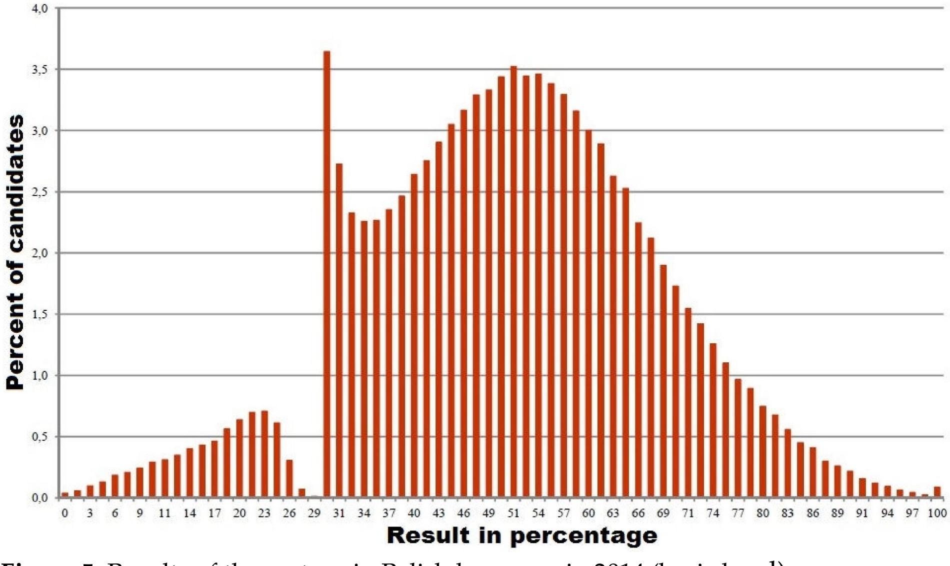

Figure 5. Results of the matura in Polish language in 2014 (basic level)
Source: own elaboration based on the reports of the CKE

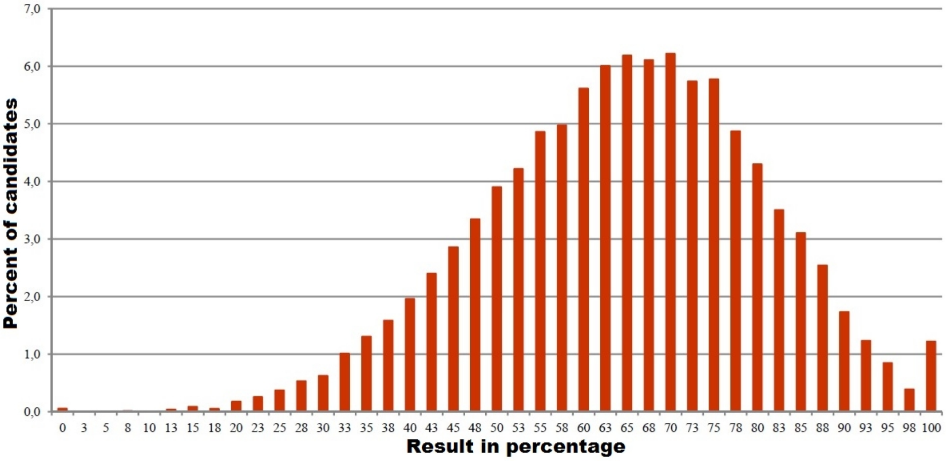

Figure 6. Results of the matura in Polish language in 2014 (extended level)
Source: own elaboration based on the reports of the CKE

There is also noticeable on each of Figures 1-6 value for the maximum number of points possible to gain, which seems to be higher than expected. It is an evidence for the low possibility of differentiation the group of best students. In other words exams are rather too easy than too hard, especially for the extended level.

In modern education the important issue is the phenomenon of grade inflation (see the forcing to reflection article by Stuart Rojstaczer *Where All Grades Are Above Average*[20]). It is the common term for "an apparently continual increase in numbers of students attaining high examination grades, or the practice of awarding grades in this way"[21]. It is the scientifically proven fact that speaks for the difficulties in establishing norms and their observance, associated with a reference point and subjectivity of assessments. Nevertheless the problems in comparison between groups (for example year to year) or individual assessment against the group are not a goal of this paper. It is the introduction of corrections to the probability distribution of the matura exam results at a basic level.

3. Correction methods of post-exam results

To overcome inconsistency in the histograms of the results of the matura in Polish language two methods were developed. The first method is based

[20] "The Washington Post", January 28, 2003, Retrived from http://www.highbeam.com/doc/1P2-234704.html on 20 May 2015.

[21] By CollinsDictionary.com (retrieved on 20 May 2015).

on linear interpolation, whereas the second one uses quadratic interpolation. The idea behind both methods is to maintain the increasing direction of this part of the histogram. Therefore only the mismatched part will be changed. In the given data from years 2012-2014 the score equal 23 was determined as the starting point of the interval taken under consideration. The end of this interval was set as 34 for the year 2013 and 36 for years 2012, 2014. For the first method, linear interpolation, it is assumed that values at ending points of the interval are being constant, while the rest lies on the line connecting these points. The second method, quadratic interpolation, fits the optimal second degree polynomial based on assumption that values at ending points of the given interval are the same as those from the primary data, and additionally in the middle of the interval the value equals the average from the values of scores inside this interval. The schema of the idea of polynomial interpolation for polynomials of degree one and two is shown in the Figure 7.

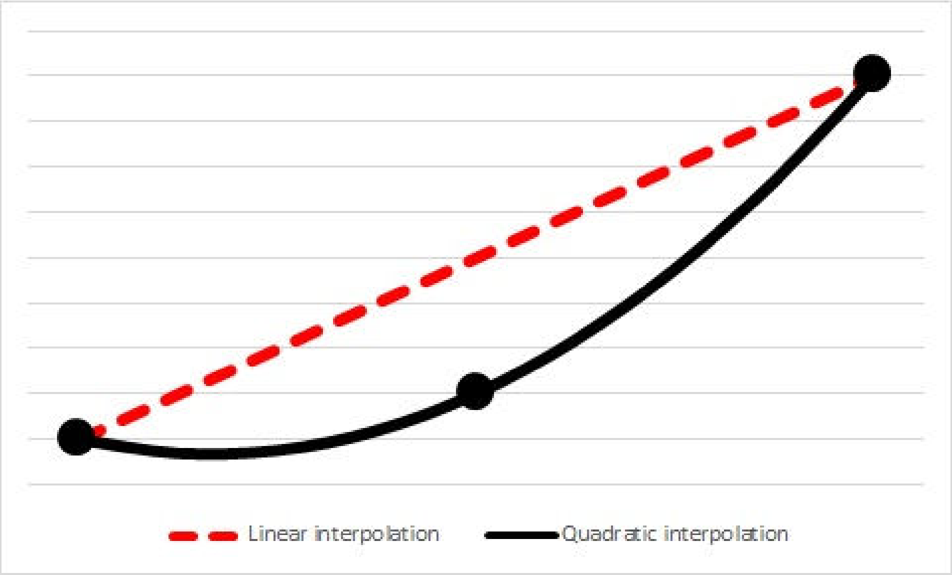

Figure 7. Idea of the linear and quadratic interpolation
Source: own elaboration

Changing the values for the scores in the considered interval may cause that the graph will not represent any more the probability distribution, due to a lack of aggregation of values to 100%. To maintain the sum of changed values fixed with respect to the initial values, we add the difference distributed proportionally to the number of nodes in the interval.

Results of the interpolation methods have been visualized in the form of that part of the histogram, which includes the considered interval with corrected values. In Figure 8 data from the year 2012 were presented, in Figure 9 data from 2013 and in Figure 10 the results of the matura exam from 2014.

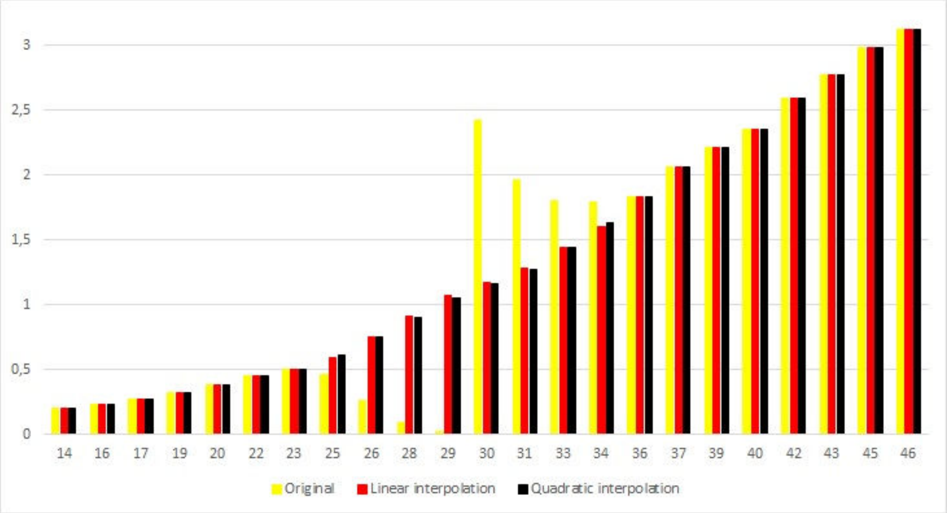

Figure 8. Local linear (red) and quadratic (black) interpolation vs. original (yellow) results of the matura in Polish language in 2012 (basic level)
Source: own elaboration

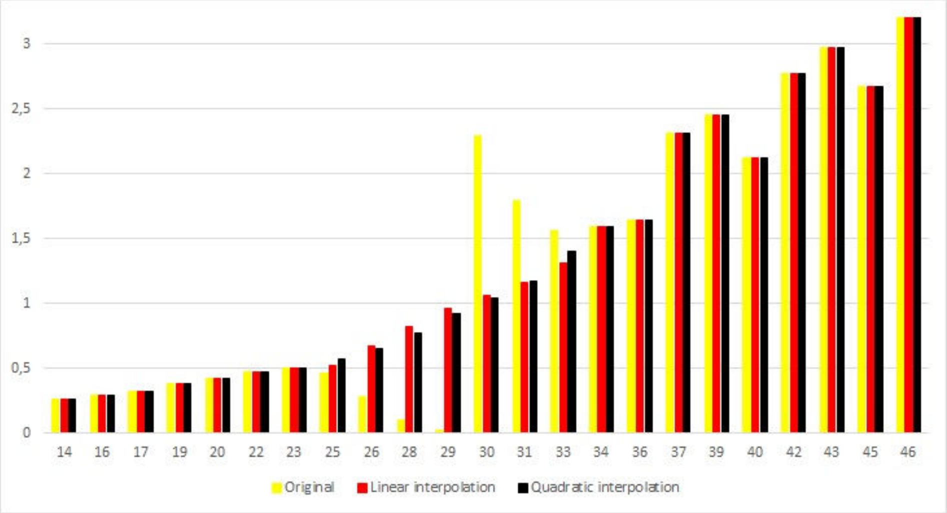

Figure 9. Local linear (red) and quadratic (black) interpolation vs. original (yellow) results of the matura in Polish language in 2013 (basic level)
Source: own elaboration

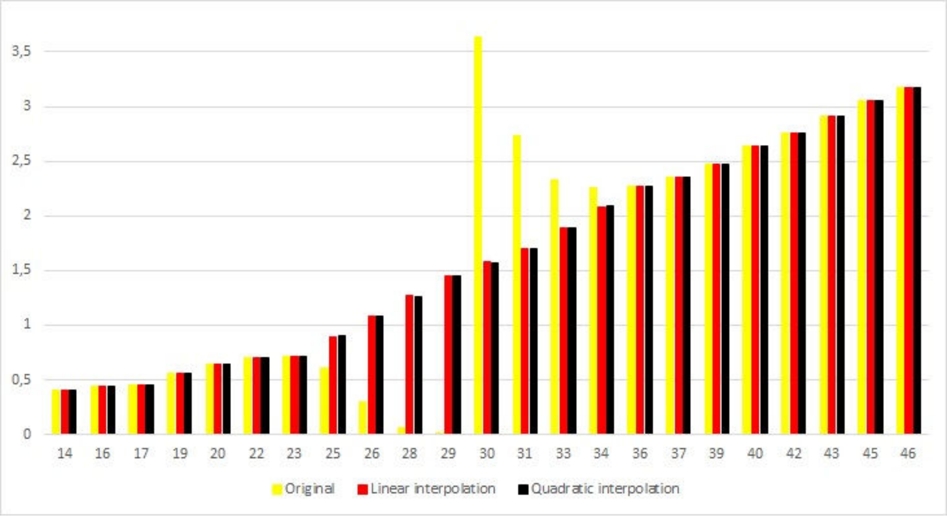

Figure 10. Local linear (red) and quadratic (black) interpolation vs. original (yellow) results of the matura in Polish language in 2014 (basic level)
Source: own elaboration

All histograms (Figures 8-10) are much smoother and free of inconsistencies. Values from both linear and quadratic interpolation are similar, although the use of the higher degree polynomial gives rather lower values comparing to the linear interpolation. Change of values influences the success percentage that is the number of students, who should pass the exam. Estimated values based on the results of the interpolation method are given in Table 2.

Table 2

Comparison of success percentage (rounded up to integers) for different methods of correction of result of the matura exam in Polish language in 2012-2014

Method	Year 2012	Year 2013	Year 2014
CKE original data	97	96	94
Linear interpolation	94	94	90
Quadratic interpolation	94	94	90
Estimate number of people that pass the matura unfairly [in thousands]	10,4	6,6	12,0

Source: own elaboration

Exact values for linear and quadratic interpolation method differ on further decimal places, but after rounding up to integers give the same estimates. However these corrected values are much lower than those given by the CKE. It means that approximately from 2 (year 2013) to 4 (year 2014) percentage of participants of the matura exam in Polish language passed that exam only because of the wrong checking the written work by an examiner. Knowledge of those thousands of students was insufficient to pass the exam fair and square.

4. Summary

This paper describes the issue of standards and measuring systems creation. It focuses on the grading regulation on Polish matura exam and objectivity of assessment the exams. Based on the historical results of matura from the basic level in Polish language published by the Central Examination Board the deviations from the probability distribution in the neighborhood of the 30% threshold can be recognized. This means that there are students, who have been given extra points to reach at least necessary to pass 30 points. Statistical reports of the CKE and based on them conclusions could be unreliable and does not reflect the actual knowledge of graduates.

The aim of this study was to present how the results of the matura exams may be transformed to correct the examiners subjectivity, while assessment the students' exams. There were proposed two methods of dealing with that problem by artificial change in results of the exam. Both methods uses local polynomial interpolation with first degree (method 1) and second degree (method 2) polynomial.

Empirical analysis performed on the data from years 2012-2014 show that the number of students, who should not pass the exam, is much (6-12 thousand) higher than calculated based on the basic data. The post-validation process in this case seems to be an important part of the correctness of reasoning. Conscious or not, the lack of objectivity during checking the exams must be taken into consideration each time the data are used to draw conclusions.

The method based on quadratic interpolation seems to have better accuracy of correction of deviations from the probability distribution of exam results. It may be useful also in other cases, not only the matura exams. As an example, the exam results improved in each of the thresholds for the corresponding grades can be considered. This however requires further study.

Bibliography

American Heart Association, http://www.heart.org/HEARTORG

Apgar V., *A proposal for a new method of evaluation of the newborn infant*, "Current Researches in Anesthesia and Analgesia" 1953, No. 32 (4), pp. 260-267

Assessment and Learning, ed. J. Gardner, London 2012

Black P., Harrison C., Lee C., Marshall B., Wiliam D., *Working inside the black box: Assessment for learning in the classroom*, "Phi Delta Kappan" 2004, No. 86(1), pp. 8-21

Canadian Index of Wellbeing, https://uwaterloo.ca/canadian-index-wellbeing/

CollinsDictionary.com

Gross National Happiness center Bhutan, http://www.gnhbhutan.org

Guide to Gross National Happiness Index, http://www.grossnationalhappiness.com

Haug G., *Capturing the Message Conveyed by Grades. Interpreting Foreign Grades*, "World Education News & Reviews" 1997, Vol. 10, No. 2

International grade conversion guide for higher education, http://www.wes.org/gradeconversionguide/index.asp

Klus-Stańska D., *Komu potrzebne jest ocenianie w szkole?*, „Edukacja i Dialog" 2006, No. 5 (178) pp. 11-14

Kosińska E., *Ocenianie w szkole*, Kraków 2000

Kusiak L., Wodnicka W., *O ocenianiu słów kilka...*, „Kwartalnik Metodyczny »Grono«" 2001, pp. 16-23

Marks N., Abdallah S., Simms A., Thompson S., et al., *The Happy Planet Index 1.0.* New Economics Foundation, London 2006

Niemierko B., *Pomiar wyników kształcenia*, Warszawa 2000

OECD Better Life Index, http://www.oecdbetterlifeindex.org/

Rojstaczer S., *Where All Grades Are Above Average*, "The Washington Post", January 28, 2003 retrived from http://www.highbeam.com/doc/1P2-234704.html on 20.05.2015

Social Progress Index 2015, http://www.socialprogressimperative.org/data/spi

Wiczkowski K., *Zza i sprzed katedry, czyli jak oceniać sprawiedliwie*, Ostrołęka 1994

Summary

This paper describes the reflection on the reliability of standards and measuring systems, with particular emphasis on the grading regulation on Polish matura exam. Based on the historical results of matura from the basic level in Polish language published by the Central Examination Board the deviations

from the probability distribution in the neighborhood of the 30% threshold can be recognized. This means that there are students, who have been given extra points to reach at least necessary to pass 30 points and therefore the objectivity of examiners assessments can be questioned. Statistical reports of the CKE and based on them conclusions could be unreliable and does not reflect the actual knowledge of graduates.

The aim of this study was to present how the results of the matura exams may be transformed to correct the examiners subjectivity, while assessment the students' exams. There were proposed two methods of dealing with that problem by artificial change in results of the exam. Both methods uses local polynomial interpolation.

Empirical analysis performed on the data from years 2012-2014 show that the number of students, who should not pass the exam, is much (6-12 thousand) higher than calculated based on the basic data. The post-validation process in this case seems to be an important part of the correctness of reasoning. Conscious or not, the lack of objectivity during checking the exams must be taken into consideration each time the data are used to draw conclusions.