Edukacja
An interdisciplinary approach 2013, 1
ISSN 0239-6858
pp. 5–24

Results of the 2002–2010 lower secondary school leaving exams on a common scale

Henryk Szaleniec, Magdalena Grudniewska,
Bartosz Kondratek, Filip Kulon, Artur Pokropek
Student Performance Analysis Unit, Educational Research Institute*

The article presents the methodology and results of a survey on equating the lower secondary school examinations from 2002–2010. The survey was carried out by the Student Performance Analysis Unit at the Educational Research Institute. More than 10 000 students were selected for the equating study and information about more than 500 items was used. IRT models were used for equating exams, the results were presented on a latent variable scale and the observed score scale. Using this procedure, it was possible to isolate random difficulty variation between exam papers from specific years and present changes in ability level of students taking the exam. Based on the results, the level of humanities abilities of lower secondary school leavers was stable, whilst maths and science demonstrated a downward trend. Equating was validated by comparison with the results of the international PISA survey. Results for the arts and humanities were consistent with the PISA results for reading literacy. Maths and science, as compared with the PISA survey maths section demonstrated greater divergence.

Keywords: lower secondary school exam, test equating, IRT, PISA.

To compare the school performance of students taking lower secondary school leaving exams in various exam sessions, it is necessary to introduce mechanisms to permit equating. The equating procedures allow for control of random variations in difficulty between papers set for the exam in subsequent years. This is important for the lower secondary school leaving exam, which is used to evaluate school performance and is important for enrolment into upper secondary school. This is a high-stakes exam[1]. Without the application of equating procedures, results of an exam cannot be compared between different years. For this reason, raw scores cannot be used to evaluate changes of the quality of teaching or the level of implementation of educational goals. This makes evaluation of the performance of teachers, schools and the whole education system difficult to assess. Exams which can be equated can provide information to upper secondary schools

Article based on research carried out within the systemic project "Quality and effectiveness of education – strengthening of institutional research capabilities" executed by the Educational Research Institute and co-financed from the European Social Fund (Human Capital Operational Programme 2007–2013, Priority III High quality of the education system). This article was published primarily in Polish language in *Edukacja*, *119*(3) 2012.

* Mail address: Zespół Analiz Osiągnięć Uczniów, Instytut Badań Edukacyjnych, ul. Górczewska 8, 01-180 Warszawa, Poland. Email: h.szaleniec@ibe.edu.pl

[1] A high-stakes exam is an exam, in which information about the result is more important than the teacher's comment.

important for estimation of the ability of new intake. The information may be used for better and more effective planning for teaching a given class.

In many education systems an equating procedure is directly embedded in the structure of examinations[2] and is applied on an ongoing basis with each exam edition. It usually involves concealing a large pool of items the use of which is repeated across sets of exam papers or by organising additional equating sessions. During the development of the Polish examination system the issue of equating exam results was not taken into consideration, equating has not yet been integrated with exam practice and all items are disclosed. This means that in the case of exams in Poland, one cannot answer the simplest question: Do students perform better or worse at an exam than a few years ago? It is not known if the trends observed in the raw scores reflect a change in the difficulty of the exam or a change in the level of abilities.

The article presents the results of a special survey. As the current structure of the lower secondary school leaving exam does not include equating, equating of exams was only possible using an additional survey. In this survey a random sample of students solved items that were chosen from all pre-2011 exam papers. The data gathered, using appropriate statistical techniques allowed equating of exams. This made it possible to show the dynamics of change in terms of student ability and exam difficulty.

Equating survey

Four hundred and forty schools were sampled for the equating survey using stratification in terms of school location and the average exam results in 2010. One class was selected from each school and all from the chosen classes participated. Special-needs, hospital, prison schools, schools for adults and schools with fewer than 11 students were not eligible (the restriction excluded 3.8% of schools and 0.4% of students from the sampling frame). A total of 10 398 students took part in the survey performed on 7–18 March 2011. In maths and science, results were obtained from 9551, and from arts and humanities, 9593 students.

Such a large sample was necessitated by the requirement for a high number of items to equate the results from 9 variants of the 2002–2010 examination in one survey. Twenty two exam booklets were used (11 for the arts and humanities and 11 for the maths and science). Each booklet was printed in two versions, A and B, identical except for the sequence of multiple choice items. Each student solved one booklet from each part of the exam. This survey implied sample design. Each booklet was attempted by at least 800 students.

Table 1 presents the survey design applied for collecting data used for equating the lower secondary school leaving exam in 2002 to 2010. The task of equating covers 9 different populations of students: P_{02}, P_{03}, …, P_{10}, each sitting their respective version of the leaving exam. The variants were identified as: T_{02}, T_{03}, …, T_{10}, where the integer signifies the year.

Students from populations P_{02}, P_{03}, …, P_{10} only took the exam corresponding to the respecting year (T_{02}, T_{03}, …, T_{10}). In the equating survey, students taking the lower secondary school leaving exam in 2011 also participated. The sample was randomly divided into 11 equivalent sub-samples of students: S^1_{11}, S^2_{11}, …, S^{11}_{11}, who participated in the equating session. Students from each of the equating samples solved a test composed of two sub-samples of anchoring items, selected from previous exams ($T_{\cdot}^{A}$). For instance, students from sample S^5_{11} attempted the item-set with sub-samples of anchoring items

[2] The ACT (American College Testing) and the SAT (Scholastic Assessment Test), the Israeli PET (Psychometric Entrance Test) or the Swedish SweSAT (Swedish Scholastic Aptitude Test) are just some examples.

Table 1
Data collection design of equating lower secondary school leaving exams, 2002–2010

Sample symbol of P_{11}	No. of obs. (est.)	Test or set of items (in brackets, symbol of the population taking the test in exam conditions)																	
		T_{02}		T_{03}		T_{04}		T_{05}		T_{06}		T_{07}		T_{08}		T_{09}		T_{10}	
		T^R_{02}	T^A_{02}	T^R_{03}	T^A_{03}	T^R_{04}	T^A_{04}	T^R_{05}	T^A_{05}	T^R_{06}	T^A_{06}	T^R_{07}	T^A_{07}	T^R_{08}	T^A_{08}	T^R_{09}	T^A_{09}	T^R_{10}	T^A_{10}
		(P_{02})		(P_{03})		(P_{04})		(P_{05})		(P_{06})		(P_{07})		(P_{08})		(P_{09})		(P_{10})	
S^1_{11}	800		✓		✓														
S^2_{11}	800				✓		✓												
S^3_{11}	800						✓		✓										
S^4_{11}	800								✓		✓								
S^5_{11}	800										✓		✓						
S^6_{11}	800												✓		✓				
S^7_{11}	800														✓		✓		
S^8_{11}	800																✓		✓
S^9_{11}	800				✓						✓								
S^{10}_{11}	800										✓								✓
S^{11}_{11}	800				✓														✓

T^A_{06} and T^A_{07} (sub-samples of items from the 2006 and 2007 papers). Table 1 does not fully reflect the complexity of the design. Firstly, it only shows the distribution of items for one part of the lower secondary school leaving exam – the arts and humanities or maths and science part. The exam results were equated for both parts, so the design covered twice as many columns – half of them concerned the arts and humanities and the other half the maths and science. Secondly, some previously unpublicised items were added to the equating survey, to be used for equating the exams in future years. Their details are not revealed in Table 1.

The design outlined above reveals two potentially confounding variables which should be taken into account during the analyses:

- student motivation – the students participating in the equating session are not solving items in the context of a high-stakes exam;
- item familiarity – the students participating in the equating session might have had contact with the items from earlier years from exam or revision practice using published previous papers.

These two potentially confounding variables would be expected to have an opposing influence on results. Lower motivation might reduce scores compared with the real high-stakes exam but familiarity with items previously encountered might offer an advantage over peers without prior exposure. If motivation and item familiarity were evenly weighted between the booklets they would not influence the results in a systematic manner, otherwise the estimation of the equated results might be systematically biased.

The need to equate 9 sessions of the lower secondary school leaving exam and to use an appropriately large number of items demanded a complex system for data collection in order to implement the survey.

As a consequence, many classic equating methods that did not apply parametric IRT modelling could not be used.

Equating methods

The concurrent calibration method was used to equate the examinations. In this method, the IRT model was estimated in one step for the equating sample (design presented in Table 1) and 9 data sets submitted by the Central Examination Board (*Centralna Komisja Egzaminacyjna* – CKE), containing scores of all students[3] taking the lower secondary school leaving exam in the years 2002–2010. The advantage of the concurrent calibration method is that the model explicitly estimates the differences of ability distribution between populations. However, there are some limitations. The whole data set is a matrix with an order of 5 million students by 500 items, with a high proportion of data missing. For example, a student who took the exam in 2002 in the full combined data set would only have data in the fields assigned to the items from the lower secondary school leaving exam of 2002. Remaining fields assigned to items from other leaving exams would contain missing data. Estimating parameters of the IRT model for such a vast data set exceeded the computational capacity of the available hardware and software.

To circumvent this problem a sub-sample of 2000 students taking the exam in the years 2002–2010 was used. Therefore, the number of response vectors sampled from actual examinations was similar to the number of responses for anchoring items from the survey samples S^1_{11}, S^2_{11}, …, S^{11}_{11}. In order to: (a) use a larger proportion of the exam data set than only the 2000 response vectors randomly selected at single equating and (b) be able to estimate the equating error resulting from sampling, the procedure was iterated $R = 500$ times. The following algorithm was iterated five hundred times:

1. Sampling of sub-samples of 2000 students from each population P_{02}, P_{03}, …, P_{10}.
2. Sampling with replacement of the same number of students from the survey sample that it contained (the so-called bootstrap sample).
3. Fitting an IRT model to such data and obtaining estimates of the mean and standard deviation of the distribution of ability in each population P_{02}, P_{03}, …, P_{10}.

The IRT model in step 3 was estimated using the MIRT software (Glas, 2010). After $R = 500$ iterations, the mean and standard deviation of student ability levels from a specific population were estimated by averaging the estimates from single replications:

$$\widehat{\mu_{\theta|\mathcal{P}_.}} = \frac{\sum_{r=1}^{R} {}^{r}\widehat{\mu_{\theta|\mathcal{P}_.}}}{R}$$

$$\widehat{\sigma_{\theta|\mathcal{P}_.}} = \frac{\sum_{r=1}^{R} {}^{r}\widehat{\sigma_{\theta|\mathcal{P}_.}}}{R}$$

Equated lower secondary school leaving exam results were thus presented on a latent variable (θ) scale, resulting from the estimated item response model. The results of equating were anchored in 2003. During the process of equating the mean student exam ability was set at 0, and standard deviation at 1. These were the default software anchoring values needed to estimate all parameters. The year 2003 was an arbitrary choice and as one of the first lower secondary school leaving exams was seen as a convenient starting point. The first school leaving exam was in 2002 but since the psychometric qualities of that first year were relatively poor and also since the exam procedures followed were not the same as those used in subsequent years it was not considered a good choice for base year. To improve readability of the results ability was rescaled to have a mean of

[3] Students writing the exam paper for students without dysfunctions and with developmental dyslexia.

100 with a standard deviation of 15 for 2003. This type of scale is more convenient since it does not yield negative results. It is one of the best known standard scales and is used to present the results of Polish surveys, e.g. the survey concerning the development of methodology for estimation of the educational value added (*Edukacyjna Wartość Dodana* – EWD) and the Nationwide testing of skills of third graders (*Ogólnopolskie Badanie Umiejętności Trzecioklasistów* – OBUT).

The benefits of using a scale based on θ include the fact that the results have approximately a normal distribution for each year. The application of a commonly used standard scale (with a mean of 100 and a standard deviation of 15) additionally facilitates interpretation of results. It is common practice in the Polish examination system to report exam results on a scale of observed total score obtained in the test or a linear transformation of that scale, constant from year to year (the percentage of the maximum test score achieved by one student). To reflect this practice the equated results are also presented on the same scale as the 2003 observed score scale. This is calculated using the "observed score equating method".

Observed score equating – theoretical description

In classical test theory it is assumed that the result of a single test assessment of a student sampled from a certain population is a random variable which is called the observed score. The observed score X is broken down into the true score τ and the random measurement error *e*:

$$X = \tau + e$$

For a single student j, the true score is a constant value that characterises their level of ability and is equal to the expected value from the student's observed score: $\tau_j = E(X_j)$. The true score τ at the whole population level is, therefore, a latent variable analogous to the θ variable in the IRT model. Indeed, the relationship between τ and θ is a 1–1 function. The scale used to report lower secondary school leaving exam is an observed score scale. It is necessary to equate observed scores when converting the raw scores from one exam to another. The following descriptions show how the observed scores of two tests, *X* and *Y*, taken by nonequivalent populations *P* and *Q*, are equated based on the θ variable scale for the IRT model.

For the two populations *P* and *Q*, taking tests *X* and *Y* respectively, equating the observed scores in the most general form takes the form of equi-percentile equating. The idea of equi-percentile equating is based on the fact that for continuous and strictly increasing cumulative distribution functions (CDF) F_X and F_Y, there occurs:

$$Y = F_Y^{-1}\big(F_X(X)\big) \tag{1}$$

that is, the compound $F_Y^{-1} \circ F_X$ maps the random variable *X* into random *Y* variable.

Unfortunately, due to the discreteness of observed scores, the CDFs F_X and F_Y for observed scores in tests *X* and *Y* are step functions, so the formula provided cannot be applied directly. So, in all equi-percentile observed score equating methods it is necessary to incorporate some appropriate form of continuization of the CDFs in order to obtain their reversible versions ${}^{(cont)}F_X$ and ${}^{(cont)}F_Y$. The function that equates *X* with *Y* takes the following form:

$${}^{(Equip)}eq_Y(x) = {}^{(cont)}F_Y^{-1}\left({}^{(cont)}F_X\ (x)\right) \tag{2}$$

The above equi-percentile equating function is a combination of the CDF of the scores in test X transformed to the continuous form with the inverse of the CDF of the scores in test Y transformed to the continuous form. The two most popular methods for continuization of distribution functions for discrete variables are: (a) local linear interpolation and (b) kernel smoothing. An in-depth review of the first approach can be found in Kolen and Brenan (2004), and the second in von Davier et al. (2004). The last step in the equating procedure is rounding the equated observed scores.

Equating observed scores with the use of the IRT (IRT Observed Score Equating) requires estimation of the CDFs of observed scores $F_{X|Q}$ or $F_{Y|\mathcal{P}}$ by referring to the parameters of the IRT model expressed on a scale common for population P and Q. Taking into account $F_{X|Q}$ infers the need to integrate after obtaining the distribution $\psi_0(\theta)$ of the conditional probability of obtaining each of the scores:

$$p_{x|Q} = \int_{\theta} \mathbb{P}(X = x|\theta)\psi_Q(\theta)d\theta \qquad (3)$$

The conditional probabilities $\mathbb{P}(X = x \,|\, \theta)$ are combinations of the conditional probabilities of observed vectors that sum up to x. Estimation $F_{X|Q}$ of is, therefore, a complex combinatorics problem combined with numerical integration. The recursive algorithm that calculates the sought-after probabilities is provided by Kolen and Brenan (2004). Glas and Béguin (1996) also identify the possibility of estimating the sought $F_{X|Q}$ by carrying out an appropriate Monte Carlo experiment based on an estimated and equated IRT model.

For the purposes of the survey, a simulation strategy was adapted, generating observed scores expressed on a common scale of the base year (2003) in the maths and science test and the arts and humanities test of the lower secondary school leaving exam. For each examined year 5 million observed scores on the scale of 2003 were generated, in compliance with the parameters of items for 2003 and estimated mean and standard deviation of the distribution of ability θ for that year.

As a result of equating, the MIRT software provides only the first two moments of distribution of ability. To increase the precision of modelling the shape of the distribution θ when generating observed scores, observations from distribution θ were generated using plausible values (PV). These constitute realisations from an a posteriori distribution of the ability of student with the response vector u (Wu, 2005):

$$\mathbb{P}(\theta|U = u) = \frac{\mathbb{P}(U = u|\theta,\beta)\psi_0(\theta)}{\int \mathbb{P}(U = u|\theta,\beta)\psi_0(\theta)\,d\theta} \qquad (4)$$

Table 2
Means of equated results of the arts and humanities part of the exam, 2002–2010[*]

Year	*Mean*	SE_r (bootstrap)	95% CI (bootstrap)	
2002	101.86	0.72	100.71	103.05
2003	100.00	0.51	99.10	100.78
2004	99.96	0.59	99.00	100.92
2005	100.30	0.58	99.36	101.35
2006	102.42	0.50	101.57	103.32
2007	100.40	0.62	99.40	101.42
2008	101.07	0.61	99.99	102.08
2009	100.29	0.57	99.40	101.24
2010	102.16	0.52	101.29	102.98

[*] *Means of equated scores of the arts and humanities part of the exam, 2002–2010.*

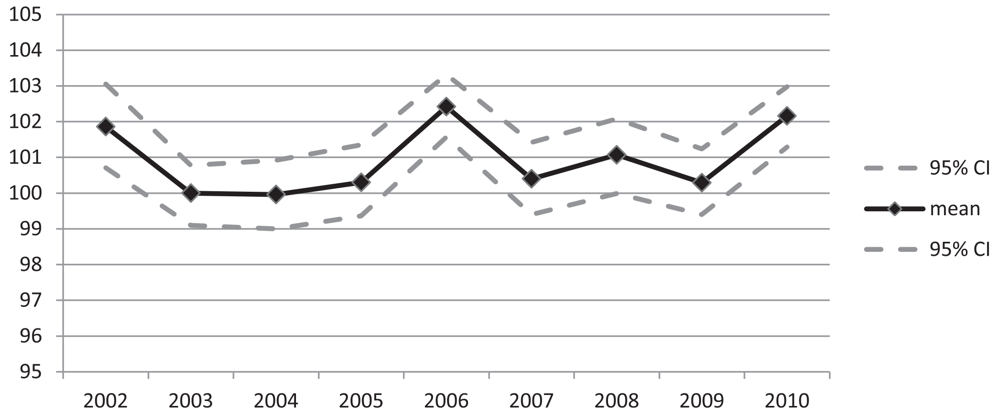

Figure 1. Means of equated scores of the arts and humanities part of the exam, 2002–2010.

where $\psi_0(\theta)$ is an a priori distribution of ability, and $\mathbb{P}(U = u|\theta, \beta)$ is a classical likelihood function dependent on the ability and item parameters. Obtaining the PVs in accordance with the above formula also requires the application of advanced numerical solutions based on the MCMC (Markov Chain Monte Carlo) methodology. In the survey, Markov chains used to generate PVs were created following the Metropolis-Hastings approach with a symmetrical function that generates "candidates" for subsequent points in the chain (c.f. Patz and Junker, 1999; Torre, 2009).

Results of equating

Results of equating on a latent variable scale (θ)

This section presents the equating results on the latent variable scale anchored in 2003 so that the mean for that year is 100 and the standard deviation is 15. Table 2 presents the average ability level of students taking the arts and humanities part of the lower secondary school exam in the years 2002–2010. The first column contains the year, the second the average ability level (the mean of rescaled θ), the next presents the equating

Table 3
Standard deviation of equated scores of the arts and humanities part of the exam, 2000–2010

Year	*SD*	SE$_r$ (*bootstrap*)	95% CI (*bootstrap*)	
2002	15,13	0,68	14.10	16.27
2003	15.00	0.51	14.18	15.86
2004	16.49	0.56	15.56	17.39
2005	15.54	0.54	14.67	16.43
2006	14.01	0.48	13.23	14.81
2007	17.25	0.75	16.11	18.54
2008	15.92	0.61	14.92	16.97
2009	14.77	0.56	13.89	15.71
2010	15.81	0.46	15.06	16.53

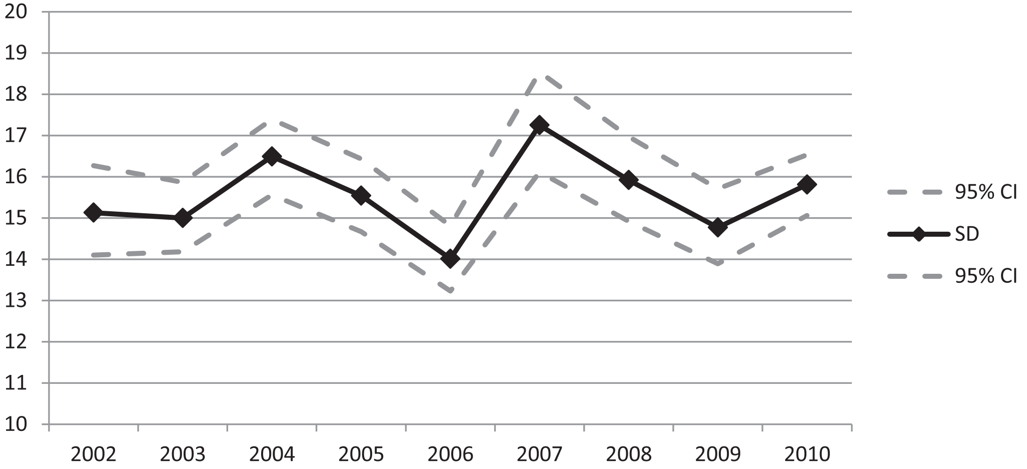

Figure 2. Standard deviation of equated scores of the arts and humanities part of the exam, 2002–2010.

error resulting from the sampling error. The error was estimated by means of the bootstrap procedure. Table 2 also provides the 95% confidence intervals (95% CI bootstrap). Confidence intervals were estimated not on the basis of the standard error but on the empirical distribution of replications from the bootstrap procedure. The 5th and 95th centile of the results of equating over different sub-samples of students are shown. This structure of confidence intervals is more precise and robust to errors arising from deviations of the distributions of interest from the normal distribution.

Figure 1 shows the equating results from Table 2. The solid line combines the average levels of ability in subsequent years (when, as mentioned above, the scale is anchored in 2003). The dashed lines show the confidence intervals constructed by the bootstrap procedure. Student ability appeared rather stable over the years. In general average student ability level in arts and humanities remained stable over the 9 years surveyed. Small but clearly visible variation in student ability is observable in 2002, 2010 and, in particular, 2006 (the year that demonstrated the greatest

Table 4
Means of equated scores for the maths and science, 2002–2010

Year	*Mean*	SE_r (bootstrap)	95% CI (bootstrap)	
2002	102.50	0.56	101.60	103.41
2003	100.00	0.52	99.14	100.86
2004	97.60	0.60	96.61	98.63
2005	96.89	0.59	95.90	97.84
2006	98.23	0.51	97.37	99.04
2007	98.30	0.56	97.37	99.18
2008	99.47	0.65	98.36	100.52
2009	97.85	0.67	96.74	99.05
2010	96.65	0.59	95.66	97.63

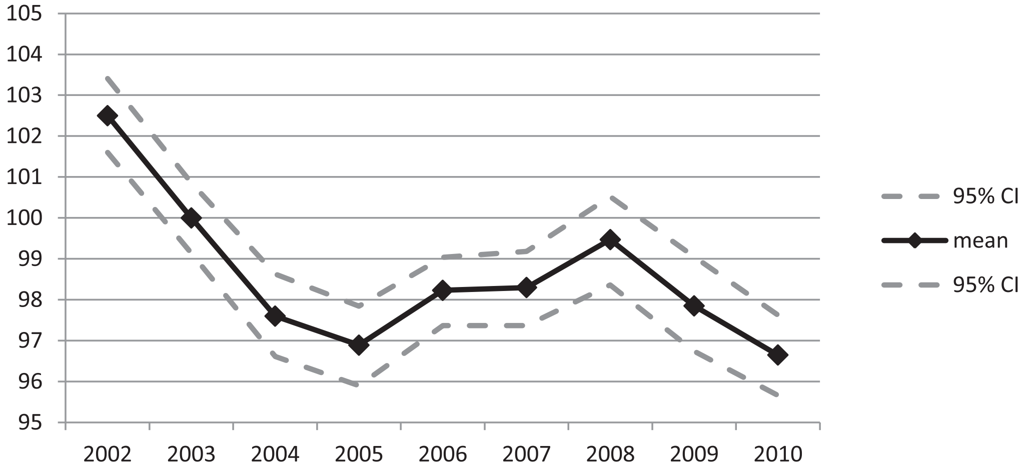

Figure 3. Means of equated scores of the maths and science part of the exam, 2002–2010.

level of ability). It is difficult to determine, though, whether this was the consequence of specific properties of the cohort, the exam paper or of the equating design adopted.

Table 3 and Figure 2 present the estimates of standard deviation for the distribution of student scores on the arts and humanities paper on the scale of the 2003 results. The results are presented in a way analogous to the average values of distributions of ability measured by the lower secondary school leaving exam in the arts and humanities part. For each year the standard deviation of the distribution (*SD*), the bootstrapped standard error and confidence intervals are provided.

As in the case of the average level of ability measured by the leaving exam, no clear trend is perceptible in the dynamics of change in the dispersion of the exam results. The greatest changes to the standard deviation are seen in the years 2006–2007. The difference however, does not influence the general picture of stability of standard deviations between years.

Table 4 and Figure 3 present the average equated scores in maths and science. Similar to the case of arts and humanities, average student

Table 5
Standard deviation of equated results of the maths and science part of the exam, 2002–2010

Year	*SD*	SE_r (*bootstrap*)	95% CI (*bootstrap*)	
2002	14.60	0.46	13.89	15.38
2003	15.00	0.45	14.25	15.72
2004	16.50	0.58	15.55	17.47
2005	16.84	0.54	16.00	17.78
2006	16.09	0.46	15.33	16.84
2007	16.68	0.55	15.79	17.57
2008	16.97	0.57	16.05	17.92
2009	17.81	0.62	16.85	18.87
2010	15.49	0.53	14.64	16.37

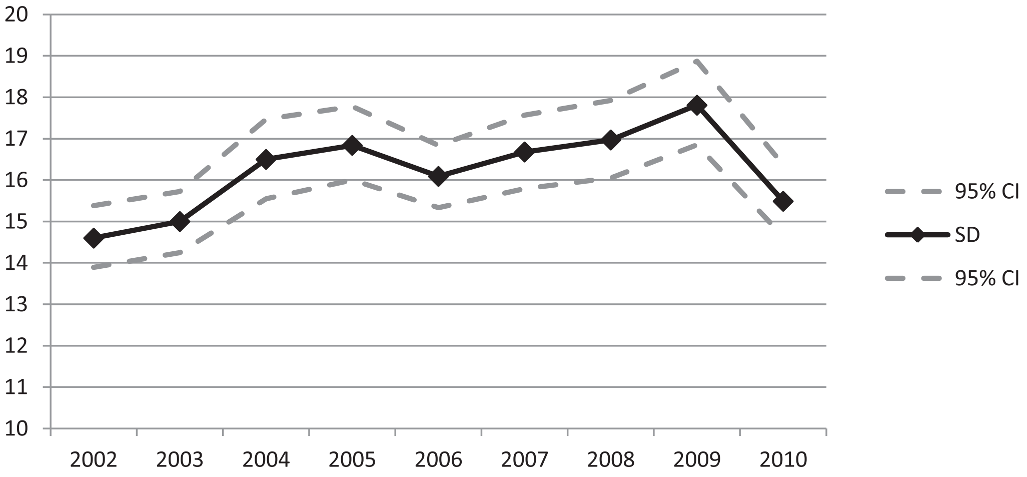

Figure 4. Dispersion (standard deviation) of equated results of the maths and science part of the exam in the years 2002–2010.

ability level anchored in 2003 is provided. The mean is established at 100 and the standard deviation at 15. The table also provides standard error of estimation and confidence intervals estimated on the basis of the bootstrap procedure. The average level of ability with confidence intervals is presented in Figure 3.

The equated lower secondary school leaving exam results in maths and science show a decline in mean ability of Polish lower secondary school students between 2002 and 2005. There is a slight upward trend in the years 2005–2008 followed by another slight downturn in the years 2008–2010. It should be noted that both trends are weak and should be interpreted with caution.

Similarly as in the case of arts and humanities, we present the standard deviations of scores after equating, which indicate dispersion of individual scores. Results are presented in Table 5 and graphically in Figure 4. Presentation of the standard deviation of the exam results on an equated scale for maths and science is close to the presentation of the arts and humanities in the previous section. Standard deviations, bootstrapped standard error as estimates and 95% confidence intervals area shown. As regards standard deviation, a significant and continuous increase in the years 2002–2009 and a sudden fall in 2010, when the level of individual diversity of students was similar to that of 2003, is observed.

Results of equating observed scores

As a result of the applied equating procedure, we obtain the distribution of the ability level θ for each year anchored to a common scale and parameters for each item that describe the probability for a response depending on θ. This allows estimation of what the distribution of the total score from any exam between 2002–2010 would have been if any cohort between 2002–2010 had had to sit it. In particular, it is possible to estimate results that students would have obtained if they had taken the exam in the base year 2003. Histograms that illustrate how the distribution of results from the arts and humanities part of the lower secondary school leaving exam of 2003 would have appeared in other years are presented in Figure 5, Histograms of analogous predictions for maths and science are shown in Figure 6. Histograms for the years

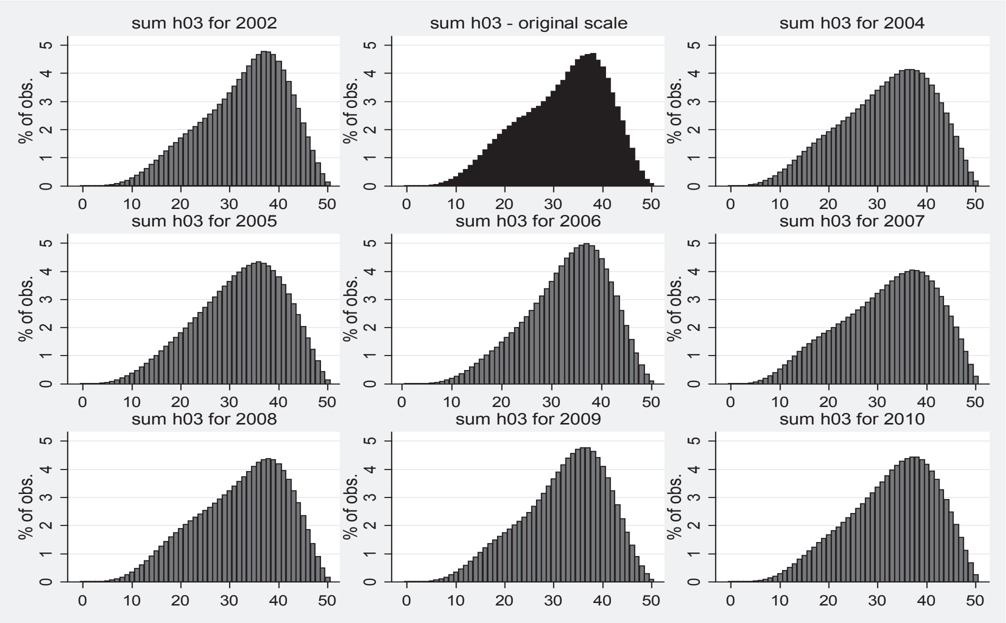

Figure 5. Distribution of observed scores of the arts and humanities part of the lower secondary school leaving exam presented on a scale of observed scores of the exam of 2003 (year 2003 is marked with a darker colour).

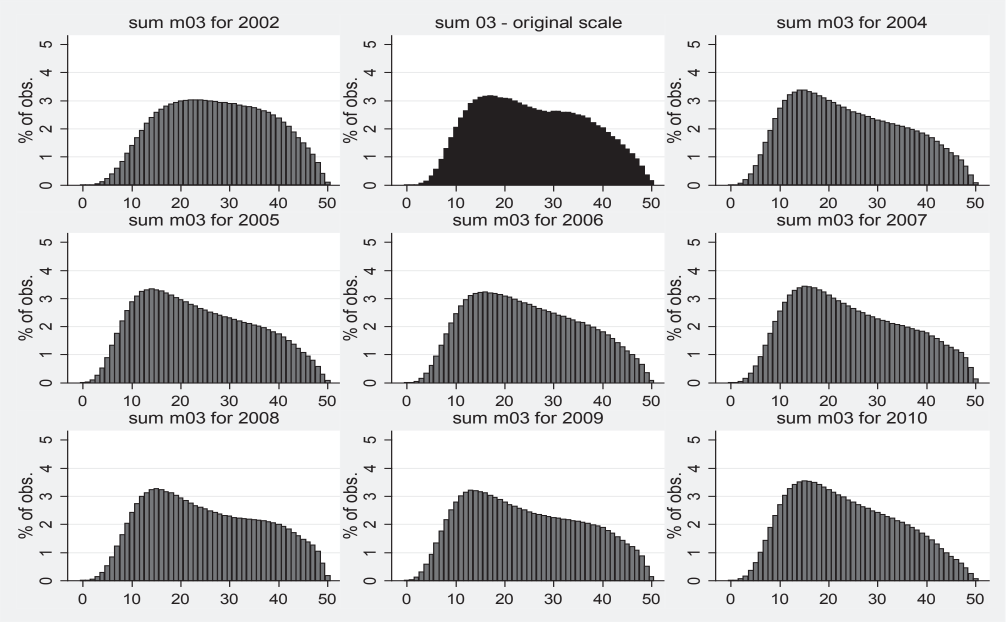

Figure 6. Distribution of observed scores of the maths and science part of the lower secondary school leaving exam presented on a scale of observed scores of the exam of 2003 (year 2003 is marked with a darker colour).

Table 6
Means and standard deviations of observed scores of lower secondary school leaving exams for the original test and on the scale of exam results of 2003

Year	Arts and humanities				Maths and science			
	Mean, origin. scale	Mean, scale of 2003	*SD* on origin. scale	*SD* on scale of 2003	Mean, origin. scale	Mean, scale of 2003	*SD* on origin. scale	*SD* on scale of 2003
2002	30.2	32.7	8.8	8.8	28.2	27.3	8.9	10.6
2003	31.8	31.8	8.9	8.8	25.7	25.7	10.9	10.8
2004	27.0	31.7	9.2	9.4	24.5	24.2	11.0	11.3
2005	33.2	31.9	8.7	9.0	24.3	23.8	10.1	11.3
2006	31.4	33.0	8.4	8.3	23.9	24.6	10.3	11.1
2007	31.5	31.9	9.8	9.7	25.3	24.5	10.2	11.3
2008	30.7	32.3	9.8	9.2	27.1	25.4	10.7	11.6
2009	31.7	32.0	8.7	8.7	26.0	24.4	11.0	11.8
2010	30.3	32.8	8.4	9.0	23.9	23.6	9.6	10.7

2002 and 2004–2010 in Figures 7 and 8 were created on the basis of 5 million simulated vectors of student solutions to the 2003 exams. The highlighted distribution for 2003 is the original distribution of scores obtained by students in that year.

The graphs in Figures 5 and 6 illustrate how student ability levels between 2002–2010 would translate into observed student scores, had each exam replicated equivalent psychometric characteristics of the 2003 exam, by providing the means and standard deviations on the scale of θ (Figures 1–4 and Tables 2–5). Any differences in the shape of the distributions in Figures 5 and 6 are a consequence of estimated differences in the level of ability between cohorts of lower secondary school students.

As an example, in maths and science compared with the scale anchored in 2003 (with a mean of 100 and standard deviation of 15), students from 2002 scored the highest (with a mean of 102.50) and from 2010 the lowest (with a mean of 96.65). The resulting distribution of observed scores in the 2002 exam is almost symmetrical (on the 2003 scale) but the 2010 student observed scores are markedly skewed to the right. The less the distributions of ability differ from a scale with a mean of 100 and a standard deviation of 15, the more subtle are the respective differences in distributions of observed scores.

Distributions shown in Figures 5 and 6 are described in Table 6 which gives mean values and standard deviations. For the purpose of comparison the table also contains distribution parameters for the actual exams administered between 2002 and 2010. This allows some very interesting observations. Figure 7 (the arts and humanities) and Figure 8 (the maths and science) graphically present means juxtaposed to the estimated means for 2003.

In the particular case of the arts and humanities (Figure 7), it can be seen that fluctuation of the average of the real exam scores between years is more pronounced than that of the predicted 2003 exam scores for respective years. That is, if all cohorts had taken an exam with the same psychometric characteristics as that of the 2003 exam there would be far less fluctuation in average score between years. In particular, large differences

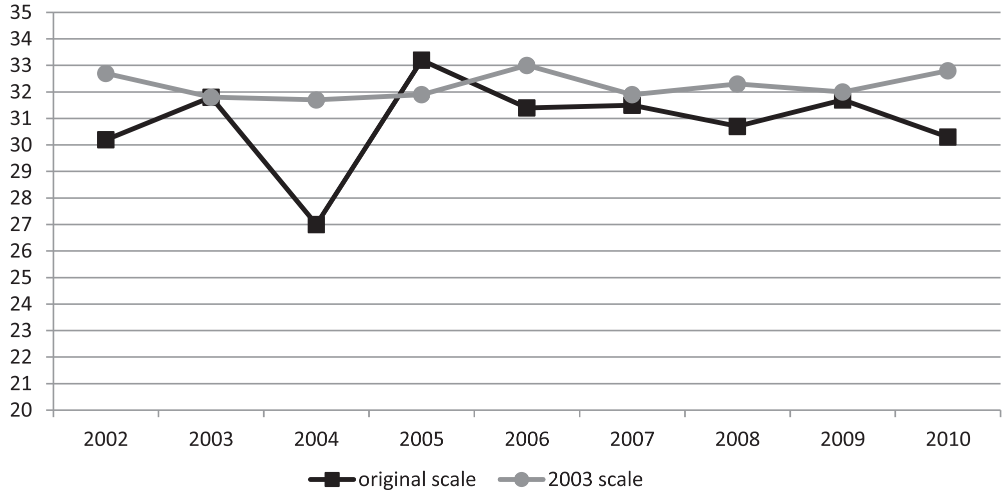

Figure 7. Means of observed scores of the lower secondary school leaving exam (arts and humanities) for the original test on the scale of the results of the 2003 exam.

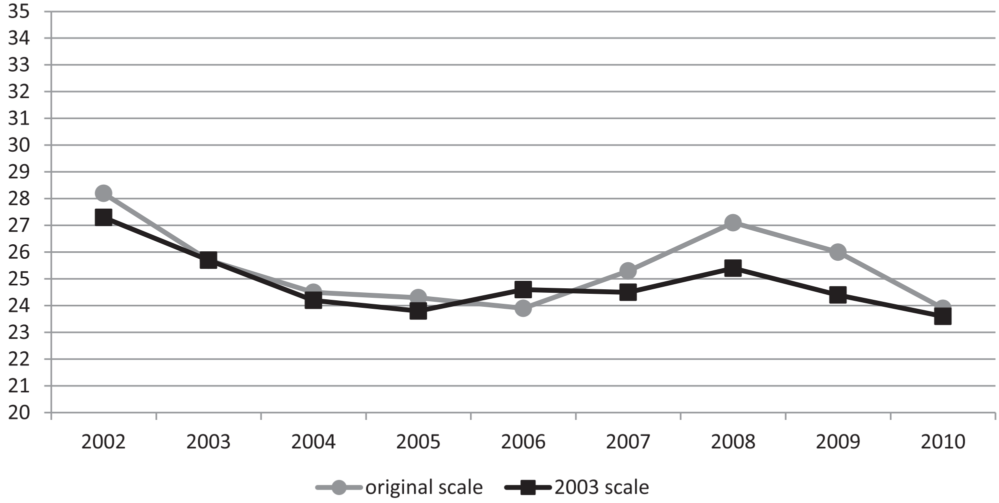

Figure 8. Means of observed scores of the lower secondary school leaving exam (maths and science) for the original test on the scale of the results of the 2003 exam.

between means in the arts and humanities in 2003, 2004 and 2005 (31.8; 27.0 and 33.2, respectively) contrast with minimal differences (to the order of ± 0.1 points) between the means for the same years on the 2003 exam scale. These results demonstrate that the differences are the consequence of major variation of the exam difficulty rather than variation of student ability.

The results presented strongly question the usefulness of scales of total observed scores (or percentage of the maximum score) for comparison of student ability between years. In addition, a question about the effectiveness of control procedures on the level of exam difficulty is raised. Variation in difficulty to the order of 6 points (i.e. 12% of the maximum test result) between consecutive versions of the exam cannot be considered to be good practice.

It is appropriate here to comment on both the original exam parameters and equated to the 2003 exam parameters, for the year 2003 (see Table 6 and Figures 7 and 8). Having the parameters of the IRT model for the distribution of student ability and for the items from 2003, the distribution of observed scores for 2003 may also be estimated in the same way that the procedure is applied to other years. It is unnecessary to estimate this for comparison with other years, however the estimated 2003 observed score distribution for students taking the 2003 exam can be used to evaluate how good is the model in predicting the 2003 observed scores. The IRT-estimated 2003 observed score distribution means for arts and humanities and for maths and science, are identical to one decimal point with the real means of 2003. Standard deviations are underestimated by only a tenth of a point. The IRT model, therefore, provided estimation of observed scores in the exams of 2003 with very high precision, which adds to the reliability of the presented distributions of exam results of 2003 in other years.

Conversion tables were constructed for all populations of students allowing converting of student scores in a given year to the estimated score for 2003. These conversion tables are shown in Table 7 (arts and humanities) and Table 8 (maths and science). These tables also include a column converting the scores of 2003 into the estimated scores of 2003 on the basis of the IRT model. This allows comparison of the observed and estimated data. It is clear that for students who only obtained 0–3 points in arts and humanities and for students who obtained 0 points in maths and science statistical equating model suggests unrealistic conversion. Unreliable parts of the conversion tables are shaded with grey. For remaining score ranges, conversion between scores and the IRT estimate is consistent. In 2003 only 57 students out of 551 150 (0.0103% of observations) were in the range of 0–3 points in arts and humanities and 9 students out of 548 716 (0.0016% of observations) obtained 0 points in maths and science. It is clear that these conversion problems have no practical significance but it does prove how well the IRT model predicts the 2003 observed exam scores and confirms the precision of scores equated on the basis of the IRT model.

The analysis of data shown in the conversion tables reveals several very interesting relationships between the levels of scores from various years. Considering a student who in 2004 scored 27 points in arts and humanities and one who scored 27 points in arts and humanities in 2005, the former would obtain an equated score of 33 points while the second would obtain a score of 25 (Table 7). In this example the difference between equated scores would be as many as 8 points, although non-equated results would suggest the same level of ability.

Differences of a similar order between the years 2004 and 2005 for the arts and humanities test (7–8 points) may be observed in the whole range of scores from 21 to 37. It is an extremely important observation, as the range is located in the centre of the distribution of scores and corresponds to 60.1% and 54.0% of the whole population of students. The consequences of this magnitude of difference in scores might be better to leave unmentioned particularly in the case of the upper secondary school leaving exam, the results of which are used for enrolment purposes.

In the example discussed above, for students in 2004 and 2005 scoring very high (above 45) or very low (under 10 points) scores in arts and humanities, the difference between actual score and the equated score for 2003 is minor (1 to 2 points). However, for

Table 7
Conversion table for observed scores in arts and humanities of the lower secondary school leaving exam into observed scores of 2003

Score	Conversion of scores to the scale of the 2003 exam								
	2002	2003	2004	2005	2006	2007	2008	2009	2010
0	–	2	1	1	2	1	2	1	2
1	2	2	2	2	2	2	2	2	3
2	3	3	3	3	3	3	3	2	4
3	4	4	4	4	4	4	4	3	5
4	5	4	5	5	5	5	5	4	6
5	6	5	6	6	6	6	7	5	7
6	7	6	8	7	7	7	8	6	8
7	8	7	9	8	8	8	9	7	9
8	9	8	10	9	9	9	10	8	10
9	10	9	11	10	10	10	11	9	11
10	12	10	13	10	11	11	12	10	12
11	13	11	14	11	12	12	13	11	13
12	14	12	15	12	13	13	14	12	14
13	15	13	16	13	14	14	15	13	15
14	16	14	18	14	15	15	16	14	16
15	17	15	19	14	16	16	17	15	16
16	18	16	20	15	17	17	18	16	17
17	19	17	21	16	18	18	19	17	18
18	20	18	23	17	20	19	20	18	19
19	22	19	24	18	21	20	21	19	20
20	23	20	25	19	22	20	23	20	21
21	24	21	26	19	23	21	24	21	22
22	25	22	27	20	24	22	25	22	24
23	26	23	28	21	25	23	25	23	25
24	27	24	29	22	26	24	26	24	26
25	28	25	30	23	27	25	27	25	27
26	29	26	32	24	28	26	28	26	28
27	30	27	33	25	29	27	29	27	29
28	31	28	34	26	30	28	30	28	30
29	32	29	34	27	31	29	31	29	31
30	33	30	35	28	32	30	32	30	32
31	34	31	36	29	33	31	33	31	34
32	35	32	37	30	34	32	33	32	35
33	36	33	38	31	35	33	34	33	36
34	37	34	39	32	36	34	35	34	37
35	38	35	40	33	37	35	36	35	38
36	39	36	41	34	38	36	37	36	39
37	40	37	42	36	39	37	38	37	40
38	41	38	42	37	40	38	39	38	41
39	41	39	43	38	41	39	40	39	42
40	42	40	44	39	41	40	41	40	43
41	43	41	45	40	42	41	42	41	44
42	44	42	46	41	43	43	43	42	45
43	45	43	46	43	44	44	44	43	46
44	46	44	47	44	45	45	45	44	47
45	47	45	48	45	46	46	46	45	48
46	48	46	48	46	47	47	47	46	48
47	49	47	49	47	48	48	47	47	49
48	49	48	50	48	48	48	48	48	49
49	50	49	50	49	49	49	49	49	49
50	50	50	50	50	49	50	50	49	50

Table 8
Conversion table for observed scores in maths and science of the lower secondary school leaving exam into observed scores of 2003

Score	Conversion of scores to the scale of the 2003 exam								
	2002	2003	2004	2005	2006	2007	2008	2009	2010
0	0	1	0	0	0	0	0	0	0
1	1	1	1	0	1	0	1	0	1
2	1	2	1	1	2	1	1	1	1
3	1	3	2	2	3	1	2	1	2
4	2	4	3	3	4	2	2	2	3
5	3	5	4	3	5	3	3	3	4
6	4	6	5	4	6	3	4	4	5
7	5	7	6	5	7	4	5	5	6
8	5	8	7	6	8	5	6	6	7
9	6	9	9	7	9	6	6	7	7
10	7	10	10	8	10	7	7	7	8
11	8	11	10	9	11	8	8	8	9
12	9	12	11	10	12	9	9	9	10
13	10	13	12	11	13	10	10	10	11
14	11	14	13	12	14	11	11	11	12
15	12	15	14	13	15	13	12	12	13
16	13	16	15	14	16	14	13	13	14
17	14	17	16	15	17	15	14	14	15
18	15	18	17	16	18	16	15	16	17
19	16	19	18	17	19	17	16	17	18
20	17	20	19	19	20	19	17	18	19
21	18	21	20	20	21	20	18	19	20
22	19	22	21	21	23	21	20	20	21
23	21	23	23	22	24	23	21	21	23
24	22	24	24	23	25	24	22	22	24
25	23	25	25	25	26	25	23	23	25
26	25	26	26	26	27	26	24	24	26
27	26	27	27	27	29	27	25	26	28
28	27	28	28	28	30	28	27	27	29
29	29	29	29	29	31	29	28	28	30
30	30	30	30	31	32	31	29	29	31
31	31	31	31	32	33	32	30	30	32
32	32	32	32	33	34	33	31	31	33
33	34	33	33	34	35	34	32	32	34
34	35	34	34	35	36	35	33	33	35
35	36	35	35	36	37	36	34	34	37
36	37	36	36	37	38	37	35	35	38
37	38	37	37	38	39	38	36	36	38
38	39	38	38	39	40	38	37	37	39
39	40	39	39	40	41	39	38	38	40
40	41	40	40	41	42	40	39	39	41
41	43	41	41	42	43	41	40	40	42
42	44	42	42	43	43	42	42	41	43
43	45	43	43	44	44	43	43	42	44
44	46	44	44	45	45	44	44	43	45
45	47	45	45	46	46	45	45	45	46
46	47	46	46	47	47	46	46	46	47
47	48	47	47	48	48	47	46	47	47
48	49	48	48	49	48	48	47	47	48
49	49	49	49	49	49	48	48	48	48
50	50	50	50	50	49	49	49	49	49

those scoring close to the mean scores there is a large difference between their observed score and the equated score on the 2003 scale. There are large differences for students with average scores and small difference at the extremes. The function that would convert observed results into equated scores is nonlinear and it should be concluded that no linear transformations of scores, would properly convert exam results between different years.

The above example from arts and humanities from 2004 and 2005 of students that obtained 27 points but differed highly in difficulty level contrasts with maths and science from the same years and the same score of 27 points (Table 8). A student scoring 27 points in maths and science in 2004 would also obtain 27 points on the 2003 test, the same for student scoring 27 points in maths and science in 2005. In 2003–2005, in maths and science, the score of 27 points corresponds to the same level of ability. It is evident that in maths and science the exams in 2003–2005 were close in terms of difficulty. At the same time (c.f. Figures 5 or 6) a significant change in ability emerged from those years. For arts and humanities the difference between mean results was the result of differences in the difficulty of the test whereas in the case of maths and science ability between years was the source of discrepancy. This analysis was made possible by equating the scores, without it, explanation of these discrepancies would have been impossible.

Verification of the equating procedure

A convenient approach to verification of the equating procedure is comparison of results with those obtained with a different tool to measure similar skills, the quality of which has been recognised and confirmed. PISA (Programme for International Student Assessment), an international survey carried out since 2000 by the Organisation for Economic Cooperation and Development (OECD) and repeated every three years provided such an opportunity. The method used to construct the tests for PISA is more sophisticated than those used in the Polish education system. Items prepared for the measurements undergo a rigorous series of tests and pilot studies, are reviewed by experts from all participating countries and the applied statistical analyses ensure high quality of scales.

It is also important here that the results of subsequent editions of the PISA survey are linked so that their results are directly comparable. PISA results are linked using an internal anchor design, different from the design adopted here. In each version of the survey, students solve a certain group of items that have already appeared in prior versions (around 20 items from each field). The results are then compared with the implementation of a multi-dimensional Rasch model.

The application of the internal anchor to equate scores is a good solution, since students solve items in subsequent editions in comparative motivational conditions. In this respect, the PISA equating methodology surpasses the post hoc equating design imposed on our survey by the structure of the Polish examination system, which does not provide internal anchor items. Motivation in the reported equating of exam results was additionally controlled for.

It should be stated that the measurement assumptions in the PISA survey are highly consistent with the curriculum of the Polish lower secondary school leaving exam. In PISA, the measurement focuses on assessment of the ability to use and understand concepts, as well as to use a range of generic skills. The measurement is intended to measure knowledge and skills needed by students in their adult life and on the labour market and to enable their full participation in contemporary democratic society (PISA 2003; PISA 2006; PISA 2009).

The similarities between the concepts for the PISA tests and the Polish lower secondary

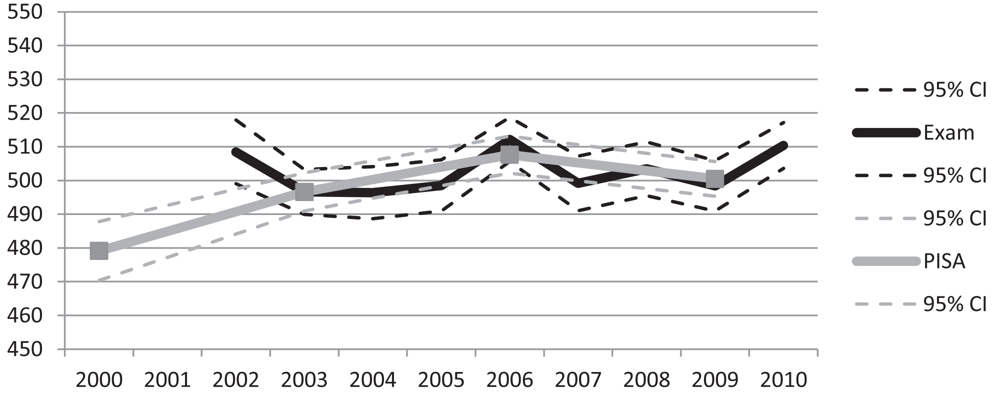

Figure 9. Mean scores of lower secondary school students in the years 2002–2010, equated scores (arts and humanities part), scale anchored in the PISA survey (reading literacy) of 2003 and the results of the PISA survey (reading literacy) years: 2000, 2003, 2006 and 2009.

school leaving exam do not suggest that they are identical. The lower secondary school leaving exam is a high-stakes state exam, mandatory for all students who graduate from the lower secondary school. The PISA survey is performed on a random sample, it is not obligatory and nor is it a high-stakes test for the student. The lower secondary school leaving exam is aimed at measuring the individual ability of the student, whereas PISA is to provide the best possible estimation of the level of performance of the whole population of 15-year-olds in particular countries. This latter difference means that the PISA survey maximises the number of items solved at test, although students solve various subsets of items. As a result, the measurement of a given field of knowledge may be broader; in addition, it prevents failure to cover some areas of knowledge due to insufficient number of items in the test. To estimate the individual score, there is a less advantageous situation, as an additional source of measurement error emerges

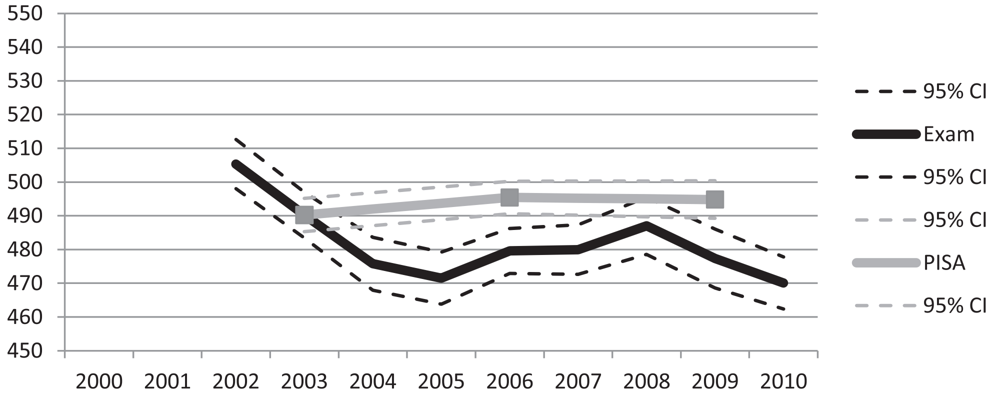

Figure 10. Mean lower secondary school students' scores in the years 2002–2010, equated results (the maths and science part), scale anchored in the PISA survey (mathematics) of 2003 and the results of the PISA surveys (mathematics) for the years: 2003, 2006 and 2009.

– sampling of items within the full set constituting a given test.

Yet the above-mentioned differences are so major that they constitute an obstacle in comparison of the two surveys. We expect high convergence between our approach to equating and the approach represented in PISA. The results of equating may not be identical, but similarities should confirm the validity of the method employed.

Figure 9 presents estimation of the reading literacy skill in PISA and estimation of the ability in humanities measured on the basis of the arts and humanities part of the lower secondary school leaving exam. The data were scaled on the PISA scale, so that the level of ability of students in 2003 would be identical in both surveys. Equality between the two surveys in 2003 is therefore imposed. The differences in other years result from the application of other tests and equating methodology. In addition, Figure 9 presents 95% confidence intervals for the results of the equating survey and PISA.

The outcomes of equating are almost identical for the equating survey and PISA. In the years 2006 and 2009, the results almost overlap. Unfortunately, the equating survey offers no way to confirm the growth of the ability of Polish lower secondary school students between 2000 and 2003. The equating survey only concerns data from 2002, which seem to contradict such conclusions. It should be remembered that the scores of 2002 should be analysed with caution, since, as mentioned above, it was the first year of the lower secondary school leaving exam, when not all procedures had already been perfected and the peculiarity of the first exam must have affected exam results.

Similarly as in the case of the arts and humanities part of the lower secondary school exam and reading literacy in PISA, the maths and science part of the lower secondary school leaving exam and the verification of mathematical skills in PISA were compared. The results of that comparison are presented in Figure 10. The dark-grey line shows the results of the equating survey and the light-grey line the result obtained in PISA. As previously, the scales of the two surveys were anchored in the year 2003 in the PISA result (this time for maths). In the case of the mathematical component, the PISA survey ensures comparability only for the years 2003–2009, so only those data are presented in the figure.

In the case of maths and science, the outcome of equating in the Polish survey differs from PISA. The results of PISA from the years 2003–2009 demonstrate stability, yet the results of the equating study show a downwards trend. On the other hand, if the years 2002 and 2003 were removed from the equating survey, a claim of relative stability could be defended, since variation of student ability in the equating survey from the years 2004 to 2010 is not substantial.

Literature

Davier von, A. A., Holland, P. W. and Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer Verlag.

Glas, C. A. (2010). *Preliminary manual of the software program Multidimensional Item Response Theory (MIRT)*. Enschede: University of Twente.

Glas, C. A. W. and Béguin A. A. (1996). *Appropriateness of IRT observed-score equating.* Research Report 1996–2, Enschede: University of Twente.

Kolen, M. J. and Brennan R. L. (2004). *Test equating, scaling, and linking: Method and practice* (2nd ed.). New York, NY: Springer Verlag.

Patz R. J. and Junker B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for Item Response Models. *Journal of Educational and Behavioural Statistics, 24*(2), 146–178.

OECD [Organization for Economic Co-operation and Development] (2003). Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2003 w Polsce [Programme for International Student Assessment. Results of the study in 2003 in Poland]. Unpublished typescript.

OECD (2006). *Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2006 w Polsce*. [Programme for International Student Assessment. Results of the study in 2006 in Poland] Retrieved from http://www.ifispan.waw.pl/pliki/pisa_raport_2006.pdf

OECD (2009). Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2009 w Polsce [Programme for International Student Assessment. Results of the study in 2009 in Poland]. Retrieved from http://www.ifispan.waw.pl/pliki/pisa_2009.pdf

Wu, M. (2005). The role of plausible values in large -scale surveys. *Studies in Educational Evaluation*, *31*, 114–128.

Torre, J. de la (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement*, *33*(6), doi: 10.1177/0146621608329890