Edukacja 2012, 3 (119), 9-30
ISSN 0239-6858

Wyniki egzaminu gimnazjalnego 2002–2010 na wspólnej skali

HENRYK SZALENIEC*, MAGDALENA GRUDNIEWSKA*,
BARTOSZ KONDRATEK*, FILIP KULON*, ARTUR POKROPEK*

W artykule prezentowane są metodologia oraz rezultaty badania nad zrównaniem wyników egzaminu gimnazjalnego dla lat 2002–2010 przeprowadzonych przez Pracownię Analiz Osiągnięć Uczniów w Instytucie Badań Edukacyjnych. Do badania zrównującego wylosowano ponad 10 tys. uczniów i wykorzystano informacje o ponad 500 zadaniach. Do zrównania wyników egzaminu wykorzystano modele IRT, wyniki przedstawiono na skali zmiennej ukrytej oraz na skali wyników obserwowanych. Dzięki zastosowanej procedurze udało się wyizolować losowe wahania trudności między arkuszami egzaminacyjnymi w poszczególnych latach i przedstawić zmiany w poziomie umiejętności uczniów zdających egzamin gimnazjalny. Na podstawie rezultatów badania można stwierdzić, że poziom umiejętności humanistycznych gimnazjalistów jest stabilny, natomiast poziom umiejętności matematyczno-przyrodniczych wykazał trend spadkowy. W analizie dokonano walidacji zrównywania, porównując przedstawione wyniki z wynikami badania międzynarodowego, porównywalnego w kolejnych cyklach badania PISA. Wyniki dla części humanistycznej wykazują wysoką zbieżność z wynikami PISA dla czytania ze zrozumieniem. W przypadku części matematyczno-przyrodniczej egzaminu, która porównywana była z matematyką w badaniu PISA, zaobserwowano większe różnice pomiędzy rezultatami obydwu badań.

Aby możliwe było porównywanie osiągnięć szkolnych uczniów, którzy zdawali egzaminy gimnazjalne w różnych sesjach egzaminacyjnych, niezbędne jest wprowadzenie mechanizmów, które pozwolą na ich zrównanie. Procedury zrównywania pozwalają na kontrolę losowych wahań poziomu trudności między arkuszami egzaminacyjnymi zastosowanymi do przeprowadzenia tego samego egzaminu w kolejnych latach. Jest to ważne w przypadku egzaminu gimnazjalnego, którego wyniki są stosowane do ewaluacji pracy szkoły oraz stanowią istotną część składową wskaźnika wykorzystywanego w rekrutacji do szkół ponadgimnazjalnych – jest to doniosły egzamin[1]. Bez zastosowania procedur zrównujących nie można porównać wyników danego egzaminu przeprowadzonego w różnych latach. Na pod-

Artykuł powstał w ramach projektu systemowego „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego" realizowanego przez Instytut Badań Edukacyjnych i współfinansowanego ze środków Europejskiego Funduszu Społecznego (Program Operacyjny Kapitał Ludzki 2007–2013, priorytet III: Wysoka jakość systemu oświaty).

* Pracownia Analiz i Osiągnięć Uczniów, Instytut Badań Edukacyjnych. E-mail: h.szaleniec@ibe.edu.pl

[1] Egzamin doniosły (*high-stakes*) – to egzamin, w którym znaczenie informacji o wyniku jest większe niż znaczenie komentarza dydaktycznego.

stawie wyników surowych nie można zatem rozstrzygnąć, czy jakość nauczania na danym poziomie edukacji, a wraz z nią poziom realizacji celów kształcenia, zmieniają się w ciągu kolejnych lat, czy też nie. Utrudniona jest tym samym ewaluacja pracy nauczycieli, szkoły i całego systemu oświatowego. Ponadto zrównane wyniki egzaminu dostarczają istotnej informacji dla szkół ponadgimnazjalnych do oszacowania potencjału intelektualnego kolejnych roczników rozpoczynających edukację w konkretnej szkole. Informacja ta może być wykorzystana do lepszego, skuteczniejszego planowania pracy dydaktycznej z danym rocznikiem.

W wielu systemach oświatowych procedura zrównywania włączona jest bezpośrednio do konstrukcji egzaminów[2] i stosowana jest na bieżąco, z każdą edycją egzaminu. Zazwyczaj wiąże się to z utajnieniem znacznej puli zadań, które stosowane są kilkakrotnie w różnych edycjach egzaminu lub organizowanych dodatkowo sesjach zrównujących. W trakcie tworzenia polskiego systemu egzaminacyjnego problematyki zrównywania wyników egzaminacyjnych nie wzięto pod uwagę, zrównywanie nie stało się częścią praktyki egzaminacyjnej, a wszystkie zadania są jawne. Tym samym w przypadku polskich egzaminów nie można odpowiedzieć nawet na najprostsze pytanie: Czy uczniowie wypadają na egzaminie lepiej, czy gorzej niż kilka lat wcześniej? Nie wiadomo bowiem, czy trendy obserwowane na wynikach surowych odzwierciedlają zmianę trudności egzaminu, czy zmianę poziomu umiejętności.

W artykule zostaną zaprezentowane wyniki specjalnie przygotowanego badania. Przy obecnej konstrukcji egzaminu gimnazjalnego, nieprzewidującej zrównywania w sposób systemowy, zrównanie wyników możliwe było tylko poprzez przeprowadzenie dodatkowego badania. W badaniu tym losowa próba uczniów rozwiązywała zadania ze wszystkich edycji egzaminu przed rokiem 2011, co dzięki odpowiedniej technice analizy statystycznej pozwoliło na zrównanie wyników kolejnych egzaminów i przedstawienie dynamiki zmian zarówno poziomu umiejętności populacji gimnazjalistów, jak i trudności egzaminów.

Badanie zrównujące

Do badania zrównującego wylosowano 440 szkół, wykorzystując warstwowanie ze względu na lokalizację szkoły i średnie wyniki egzaminacyjne z roku 2010. Z każdej szkoły wylosowano po jednym oddziale szkolnym. W badaniu brali udział wszyscy uczniowie z wylosowanego oddziału szkolnego. Z losowania wykluczono szkoły specjalne, przyszpitalne, przywięzienne, szkoły dla dorosłych i szkoły liczące mniej niż 11 uczniów (przyjęte ograniczenie wyłączyło z operatu 3,8% szkół i 0,4% uczniów). W badaniu, przeprowadzonym w dniach 7–18 marca 2011 roku, wzięło udział łącznie 10 398 uczniów. W części matematyczno-przyrodniczej uzyskano wyniki 9551 uczniów, a w części humanistycznej – 9593 uczniów.

Tak liczna próba niezbędna była ze względu na konieczność wykorzystania dużej liczby zadań potrzebnych do zrównania wyników z aż 9 edycji egzaminacyjnych (2002–2010) w jednym badaniu. Wykorzystano 22 zeszyty zrównujące (11 dla części humanistycznej i 11 dla części matematyczno-przyrodniczej). Każdy zeszyt występował w wersji A i B, które różniły się jedynie kolejnością odpowiedzi do wyboru w zadaniach zamkniętych. Każdy uczeń rozwiązywał jeden zeszyt zrównujący z każdej z części egzaminu. Schemat badania implikował sposób doboru próby, któ-

[2] Amerykański ACT (*American College Testing)* i SAT (*Scholastic Assessment Test)*, izraelski PET (*Psychometric Entrance Test*) czy szwedzki SweSAT (*Swedish Scholastic Aptitude Test*) to tylko kilka przykładów.

ra została zaprojektowana tak, aby jeden zeszyt rozwiązywany był przynajmniej przez 800 uczniów.

W Tabeli 1 przedstawiono plan badawczy zastosowany do zebrania danych wykorzystanych do zrównywania wyników egzaminu gimnazjalnego 2002–2010. Zadanie zrównania wyników obejmuje 9 różnych populacji uczniów: $\mathcal{P}_{02}, \mathcal{P}_{03}, ..., \mathcal{P}_{10}$, z których każda pochodziła z odpowiedniej edycji egzaminu gimnazjalnego od 2002 do 2010 roku. Egzaminy 2002–2010 oznaczono odpowiednio: $T_{02}, T_{03}, \ldots, T_{10}$.

Uczniowie z populacji $\mathcal{P}_{02}, \mathcal{P}_{03}, ..., \mathcal{P}_{10}$, pisali wyłącznie egzamin odpowiadający ich rocznikowi ($T_{02}, T_{03}, \ldots, T_{10}$). W badaniu zrównującym wzięli również udział wylosowani uczniowie zdający egzamin gimnazjalny w roku 2011. Próbę tę podzielono losowo na 11 równoważnych podprób uczniów: $S_{11}^1, S_{11}^2, ..., S_{11}^{11}$, którzy wzięli udział w sesji zrównującej. Uczniowie z każdej próby w sesji zrównującej rozwiązywali test składający się z dwóch podzbiorów zadań kotwiczących wybranych z poprzednich egzaminów ($T_\cdot^A$). Przykładowo uczniowie z próby S_{11}^5 rozwiązywali test z podzbiorami zadań kotwiczących T_{06}^A oraz T_{07}^A (podzbiory z edycji egzaminów 2006 oraz 2007).

Należy zaznaczyć, że Tabela 1 nie oddaje w pełni złożoności zastosowanego planu. Po

Tabela 1
Schemat zrównywania wyników egzaminów gimnazjalnych 2002–2010

Symbol próby z $\mathcal{P}_{11}$	N obs. (szac.)	Test lub zbiór zadań (w nawiasie symbol populacji rozwiązującej test w warunkach egzaminacyjnych)																	
		T_{02}		T_{03}		T_{04}		T_{05}		T_{06}		T_{07}		T_{08}		T_{09}		T_{10}	
		T_{02}^R	T_{02}^A	T_{03}^R	T_{03}^A	T_{04}^R	T_{04}^A	T_{05}^R	T_{05}^A	T_{06}^R	T_{06}^A	T_{07}^R	T_{07}^A	T_{08}^R	T_{08}^A	T_{09}^R	T_{09}^A	T_{10}^R	T_{10}^A
		($\mathcal{P}_{02}$)		($\mathcal{P}_{03}$)		($\mathcal{P}_{04}$)		($\mathcal{P}_{05}$)		($\mathcal{P}_{06}$)		($\mathcal{P}_{07}$)		($\mathcal{P}_{08}$)		($\mathcal{P}_{09}$)		($\mathcal{P}_{10}$)	
S_{11}^1	800		✓		✓														
S_{11}^2	800				✓		✓												
S_{11}^3	800						✓		✓										
S_{11}^4	800								✓		✓								
S_{11}^5	800										✓		✓						
S_{11}^6	800												✓		✓				
S_{11}^7	800														✓		✓		
S_{11}^8	800																✓		✓
S_{11}^9	800				✓						✓								
S_{11}^{10}	800										✓								✓
S_{11}^{11}	800				✓														✓

pierwsze, pokazuje ona rozkład zadań tylko dla jednej części egzaminu gimnazjalnego – części humanistycznej lub matematyczno-przyrodniczej. Wyniki egzaminu zrównywane były dla obydwu części, zatem plan obejmował dwa razy więcej kolumn – połowa z nich dotyczyła części humanistycznej, a połowa matematyczno-przyrodniczej. Po drugie, do badania zrównującego dołączono nowe, nieupublicznione zadania, które będą służyć do zrównywania wyników w kolejnych latach, co również nie zostało zaznaczone w tabeli.

Nakreślony plan zrównywania ujawnia dwie potencjalnie zakłócające (*confounding*) zmienne, które warto uwzględnić podczas analiz:

- motywacja uczniów – uczniowie biorący udział w sesji zrównującej nie rozwiązują zadań w warunkach egzaminu doniosłego;
- ujawnienie zadań – uczniowie biorący udział w sesji zrównującej mogli mieć styczność z zadaniami z wcześniejszych edycji egzaminów, ćwicząc swoje umiejętności na upublicznionych arkuszach egzaminacyjnych z lat wcześniejszych.

Wymienione dwie, potencjalnie zakłócające, zmienne są względem siebie w opozycji. Czynnik motywacyjny powinien obniżać wyniki podczas sesji zrównującej w porównaniu z warunkami testu doniosłego. Natomiast czynnik ujawnienia zadań stawia w uprzywilejowanej pozycji uczniów z sesji zrównującej w porównaniu z ich kolegami i koleżankami widzącymi zadania po raz pierwszy podczas egzaminu. Jeżeli czynniki motywacji oraz ujawnienia zadań będą równomiernie działały pomiędzy zeszytami zrównującymi, to nie powinny one wpłynąć na wyniki zrównania w sposób systematyczny w innym przypadku oszacowanie zrównanych wyników może być systematycznie obciążone.

Konieczność zrównania jednocześnie 9 edycji egzaminu gimnazjalnego i użycia w tym celu odpowiednio dużej liczby zadań wymusiła skomplikowany system zbierania danych do przeprowadzenia tego badania. W konsekwencji nie można było wykorzystać wielu klasycznych metod zrównywania, niekorzystających z parametrycznego modelowania IRT.

Metoda zrównywania

Do zrównywania wyników egzaminacyjnych wykorzystano metodę kalibracji łącznej, zgodnie z którą w jednym kroku szacowano model IRT dla próby zrównującej (schemat przedstawiony w Tabeli 1) i 9 zbiorów danych dostarczonych przez Centralną Komisję Egzaminacyjną (CKE), zawierających odpowiedzi wszystkich uczniów[3] zdających egzamin gimnazjalny w latach 2002–2010. Atutem metody kalibracji łącznej jest uwzględnienie w modelu różnic między populacjami i szacowanie ich. Wiążą się z tym jednak pewne ograniczenia. Cały zbiór danych stanowi macierz rzędu 500 odrębnych kategorii punktowych oraz 5 milionów uczniów, a dodatkowo z dużym odsetkiem braku danych (uczeń zdający egzamin w 2002 roku w połączonym zbiorze danych będzie miał przypisane wyłącznie odpowiedzi za zadania z egzaminu gimnazjalnego 2002, w pozostałych komórkach będzie figurowała informacja o braku danych). Dopasowanie modelu IRT do tak rozległego zbioru danych przerasta możliwości obliczeniowe dostępnego sprzętu oraz oprogramowania.

Aby przezwyciężyć problem wielkości pełnego zbioru danych, zamiast korzystać z całego zbioru zdecydowano się na zrównanie egzaminów, wykorzystując podpróby 2000 uczniów zdających egzaminy w latach 2002–2010. Dzięki temu osiągnięto liczebność prac egzaminacyjnych podobną do liczebności prac z prób badawczych $S_{11}^{1}, S_{11}^{2}, \ldots, S_{11}^{11}$.

[3] Uczniów piszących egzamin z wykorzystaniem arkusza egzaminacyjnego przeznaczonego dla uczniów bez dysfunkcji i z dysleksją rozwojową.

Aby: (a) wykorzystać większą część zbioru danych egzaminacyjnych niż 2000 prac losowanych przy pojedynczym zrównywaniu, (b) móc oszacować błąd zrównania wynikający z doboru próby, procedurę powtórzono $R = 500$ razy. Zatem pięćsetkrotnie powtórzono następujący algorytm:

1. Wylosowanie z każdej z populacji $\mathcal{P}_{02}$, $\mathcal{P}_{03}$, ..., $\mathcal{P}_{10}$, podprób liczących 2000 uczniów.
2. Wylosowanie ze zwracaniem z próby badawczej takiej samej liczby uczniów, jaka się w niej znajdowała (tzw. próba *bootstrap*).
3. Dopasowanie do takiej podpróby danych modelu IRT, *explicite* szacującego średnią i odchylenie standardowe rozkładu umiejętności w każdej z populacji $\mathcal{P}_{02}, \mathcal{P}_{03}, ..., \mathcal{P}_{10}$.

Model IRT w kroku 3. był szacowany z wykorzystaniem oprogramowania MIRT (Glas, 2010). Po $R = 500$ replikacjach średnią $\mu_{\theta|\mathcal{P}_\cdot}$ i odchylenie standardowe $\sigma_{\theta|\mathcal{P}_\cdot}$ poziomu umiejętności uczniów z określonej populacji $\mathcal{P}_\cdot$ oszacowano uśredniając oszacowania z pojedynczych replikacji:

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

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

Tak zrównane wyniki egzaminów gimnazjalnych zostały przedstawione na skali zmiennej ukrytej (θ), wynikającej z dopasowanego modelu odpowiedzi na zadania testowe (IRT). Wyniki zrównywania zostały zakotwiczone w roku 2003. W czasie procesu zrównywania średnia umiejętności egzaminacyjnych uczniów została ustawiona na 0, a odchylenie standardowe na 1. Był to standardowy zabieg techniczny, konieczny do oszacowania wszystkich parametrów modelu. Rok 2003 wybrany został arbitralnie, przy czym wzięto pod uwagę, iż był to jeden z pierwszych egzaminów gimnazjalnych i jako taki stanowił dogodny punkt wyjścia (pierwszym był egzamin z roku 2002, lecz właściwości psychometryczne tego egzaminu były relatywnie słabe, a procedury egzaminacyjne różniły się od stosowanych w kolejnych latach – z tego względu nie wybrano 2002 jako roku bazowego). Aby zwiększyć czytelność informacji o wynikach, przeskalowano ich surową postać na skalę o średniej 100 i odchyleniu standardowym 15 dla roku 2003. Taka skala jest łatwiejsza do prezentacji, ponieważ nie daje ujemnych wyników. Jest to jedna z najbardziej znanych skal standardowych, ponadto używa się jej już do przedstawiania wyników polskich badań np. badania dotyczącego rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej (EWD) oraz *Ogólnopolskiego badania umiejętności trzecioklasistów* (OBUT).

Korzyścią z zastosowania skali wyników bazujących na θ jest to, że wyniki mają w przybliżeniu rozkład normalny dla każdego roku, a zastosowanie powszechnie używanej skali standardowej (o średniej 100 i odchyleniu standardowym 15) dodatkowo ułatwia interpretację wyników. Przyjętą w polskim systemie egzaminacyjnym praktyką jest jednak raportowanie wyników egzaminów na skali sumarycznej liczby punktów zdobytych w teście lub na skali będącej stałym, między latami, liniowym przekształceniem tejże skali (procent maksymalnej liczby punktów testu zdobyty przez ucznia). W związku z tym opis zrównanych wyników z egzaminów gimnazjalnych w latach 2002–2010 przedstawiamy także na skali wyników surowych w egzaminie z 2003 roku, co wymaga przeprowadzenia tak zwanego „zrównania wyników obserwowanych".

Zrównywanie wyników obserwowanych – opis teoretyczny

W klasycznej teorii testów przyjmuje się, że wynik z pojedynczego badania testem ucznia wylosowanego z pewnej populacji jest zmienną losową, którą nazywa się wynikiem obserwowanym. Wynik obserwowany X jest rozbijany na wynik prawdziwy τ oraz losowy błąd pomiaru e:

$$X = \tau + e$$

Dla pojedynczego ucznia j wynik prawdziwy jest stałą wartością charakteryzującą jego poziom umiejętności, jednocześnie równą wartości oczekiwanej z wyniku obserwowanego tego ucznia: $\tau_j = E(X_j)$. Wynik prawdziwy τ na poziomie całej populacji jest zatem ciągłą zmienną ukrytą analogiczną do zmiennej θ w IRT. Faktycznie zależność między τ oraz θ jest funkcyjna – jest to ta sama rzecz tylko wyrażona na różnej skali.

Skala wyników surowych egzaminu gimnazjalnego, na której raportuje się jego wyniki, jest skalą wyników obserwowanych. Chcąc przeliczyć wyniki surowe z jednej z edycji egzaminu na inną konieczne jest zatem przeprowadzenie zrównania wyników obserwowanych. W kolejnych akapitach zostanie opisane w jaki sposób w oparciu o zrównanie na skali zmiennej θ modelu IRT dokonuje się zrównania wyników obserwowanych dwóch testów, X oraz Y rozwiązywanych przez nierównoważne populacje $\mathcal{P}$ oraz $\mathcal{Q}$.

Dla dwóch populacji $\mathcal{P}$ i $\mathcal{Q}$ piszących odpowiednio testy X oraz Y zrównywanie wyników obserwowanych w najogólniejszej postaci przyjmuje formę tzw. zrównywania ekwicentylowego (*equipercentile equating*). Idea zrównywania ekwicentylowego opiera się na tym, że dla ciągłych i ściśle rosnących dystrybuant F_X oraz F_Y zachodzi:

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

czyli złożenie $F_Y^{-1} \circ F_X$ przekształca zmienną losową X w zmienną losową Y.

Niestety, dystrybuanty F_X oraz F_Y dla wyników obserwowanych w testach X oraz Y, ze względu na dyskretność tychże wyników, są funkcjami skokowymi, więc podany wzór nie może zostać bezpośrednio zastosowany. W efekcie, we wszystkich ekwicentylowych metodach zrównywania wyników obserwowanych niezbędne jest uwzględnienie jakiejś formy odpowiedniego uciąglania dystrybuant do ich odwracalnych postaci ${}^{(cont)}F_X$ oraz ${}^{(cont)}F_Y$. Funkcja zrównująca X z Y przyjmuje wtedy następujący kształt:

$${}^{(Equip)}eq_Y(x) = {}^{(cont)}F_Y^{-1}\Big({}^{(cont)}F_X(x)\Big).$$

Ekwicentylowa funkcja zrównująca podana powyżej jest złożeniem przekształconej do postaci ciągłej dystrybuanty rozkładu wyników w teście X z odwrotnością przekształconej do postaci ciągłej dystrybuanty rozkładu wyników w teście Y. Dwoma najpopularniejszymi metodami uciąglania dystrybuant zmiennych dyskretnych są: (a) lokalna interpolacja liniowa, (b) wygładzanie za pomocą estymatora jądrowego (*kernel smoothing*). Pogłębiony przegląd pierwszego podejścia można znaleźć u Michaela J. Kolen i Roberta L. Brenana (2004), a drugiego u Aliny von Davier i in. (2004). Ostatnim krokiem w procedurze zrównywania jest zaokrąglenie zrównanych poprzez funkcję (podaną tym wzorem) wyników.

Zrównywanie wyników obserwowanych z wykorzystaniem IRT (*IRT Observed Score Equating*) wymaga estymacji dystrybuant

obserwowanych wyników $F_{X|Q}$ lub $F_{Y|\mathcal{P}}$, lub obu tych dystrybuant poprzez odwołanie się do parametrów modelu IRT wyrażonych na wspólnej dla populacji $\mathcal{P}$ i Q skali. Biorąc pod uwagę $F_{X|Q}$, oznacza to konieczność scałkowania po rozkładzie $\psi_Q(\theta)$ warunkowego prawdopodobieństwa uzyskania każdego z wyników:

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

Warunkowe prawdopodobieństwa $\mathbb{P}(X = x|\,\theta)$ są kombinacją warunkowych prawdopodobieństw zaobserwowania wektorów odpowiedzi sumujących się x. Oszacowanie $F_{X|Q}$ stanowi zatem skomplikowany problem kombinatoryczny połączony z całkowaniem numerycznym. Rekursywny algorytm obliczający szukane prawdopodobieństwa podają Kolen i Brenan (2004). Cees A. Glas i Anton Béguin (1996) wskazują również na możliwość oszacowania szukanego $F_{X|Q}$ poprzez przeprowadzenie stosownego eksperymentu Monte Carlo bazującego na oszacowanym i zrównanym modelu IRT.

W przeprowadzonym badaniu zaadaptowano symulacyjną strategię generowania wyników obserwowanych wyrażonych na wspólnej skali roku bazowego (2003) w teście matematyczno-przyrodniczym oraz humanistycznym egzaminu gimnazjalnego. Dla każdego rocznika wygenerowano 5 milionów wyników obserwowanych na skali z 2003 roku zgodnie z oszacowaną dla tego rocznika średnią i odchyleniem standardowym rozkładu umiejętności θ oraz przy uwzględnieniu parametrów zadań dla 2003 roku.

W wyniku zrównania program MIRT dostarcza jedynie dwóch pierwszych momentów rozkładu umiejętności. Dla zwiększenia precyzji odwzorowania kształtu rozkładu θ przy generowaniu wyników obserwowanych – obserwacje z rozkładu θ generowano z wykorzystaniem tak zwanych *plausible values* (PV). Stanowią one realizacje z rozkładu a posteriori parametru umiejętności ucznia o wektorze odpowiedzi u (Wu, 2005):

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

gdzie $\psi_0(\theta)$ jest rozkładem a priori umiejętności, a $\mathbb{P}(\boldsymbol{U} = \boldsymbol{u}|\theta, \boldsymbol{\beta})$ klasyczną funkcją wiarygodności zależną od parametru umiejętności oraz parametrów zadań.

Uzyskanie PV zgodnie z powyższym wzorem wymaga również zastosowania zaawansowanych numerycznych rozwiązań opartych na metodologii MCMC (*Markov Chain Monte Carlo*). W badaniu łańcuchy Markova służące do wygenerowania PV utworzono zgodnie z podejściem Metropolis Hastings z symetryczną funkcją generującą „kandydatów" na kolejne punkty w łańcuchu (por.: Patz i Junker, 1999; Torre, 2009).

Wyniki zrównywania

Wyniki zrównywania na skali zmiennej ukrytej (θ)

W tej sekcji prezentowane są wyniki zrównania na skali zmiennej ukrytej zakotwiczonej w roku 2003 tak, że średnia dla tego roku wynosi 100, a odchylenie standardowe 15. W Tabeli 2 przedstawiono średni poziom umiejętności uczniów zdających część humanistyczną egzaminu gimnazjalnego w latach 2002–2010. W pierwszej kolumnie podany został rok, w drugiej średni poziom umiejętności (średnia), w kolejnej przedstawiony jest błąd zrównania wynikający z błędu losowania. Jako że do badania wykorzystana została próba badawcza, a nie cała populacja, tak jak we wszystkich parametrach szacowanych na podstawie próby losowej

Tabela 2
Średnie zrównanych wyników części humanistycznej egzaminu w latach 2002–2010[*]

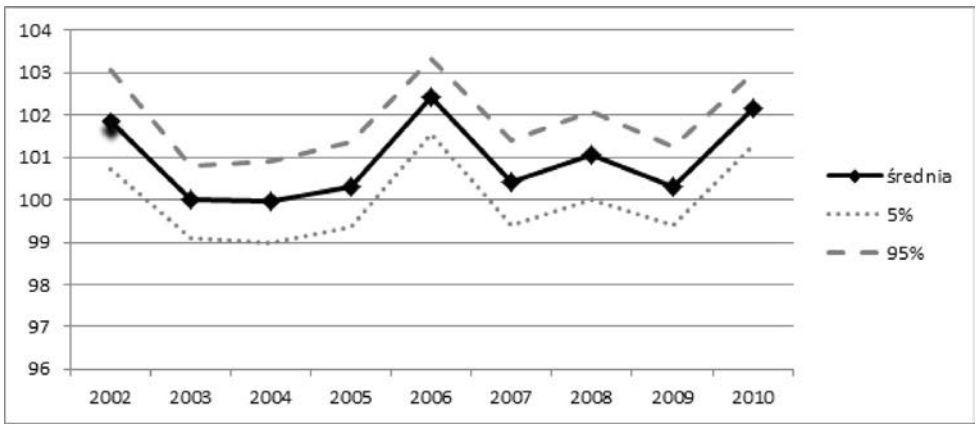
Rok	Średnia	SE_r (*bootstrap*)	95% CI (*bootstrap*)	
2002	101,86	0,72	100,71	103,05
2003	100,00	0,51	99,10	100,78
2004	99,96	0,59	99,00	100,92
2005	100,30	0,58	99,36	101,35
2006	102,42	0,50	101,57	103,32
2007	100,40	0,62	99,40	101,42
2008	101,07	0,61	99,99	102,08
2009	100,29	0,57	99,40	101,24
2010	102,16	0,52	101,29	102,98

[*] Średnie zrównanych wyników części humanistycznej egzaminu w latach 2002–2010 na skali 100;15.

mamy do czynienia z losowym błędem (SE_r (*bootstrap*))[4]. Błąd został oszacowany za pomocą procedury *bootstrap*. Obok błędu standardowego w Tabeli 2 znaleźć można wartości wyznaczające 95-procentowy przedział ufności (95% CI (*bootstrap*)). Przedziały ufności oszacowane zostały nie na podstawie błędu standardowego, ale na podstawie empirycznego rozkładu replikacji z procedury *bootstrap*: pokazują 5 i 95 centyl wyników zrównania na różnych próbach uczniów. Taki sposób konstrukcji przedziałów ufności jest bardziej precyzyjny i bardziej odporny na błędy wynikające z odstępstw badanych rozkładów od rozkładu normalnego.

Na Rysunku 1 w graficzny sposób przedstawiono wyniki zrównywania z Tabeli 2. Ciągła linia oznacza średni poziom umiejętności w danym roku (gdzie, jak przypomi-

[4] Nie jest to jedyne źródło błędu mogące wpływać na precyzję szacowania. Oprócz błędu wynikającego z doboru próby badawczej (błąd próbkowania) w procesie zrównywania – w przyjętym schemacie badawczym uwidacznia się również błąd związany z wyborem zadań (błąd zrównywania).



Rysunek 1. Średnie zrównanych wyników części humanistycznej egzaminu w latach 2002–2010.

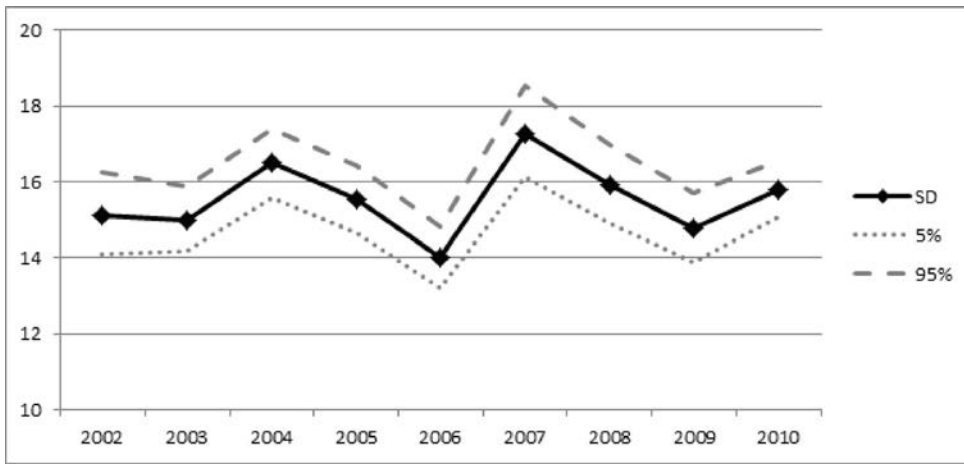
Tabela 3
Zróżnicowanie (odchylenie standardowe) zrównanych wyników części humanistycznej egzaminu w latach 2002–2010

Rok	Odchylenie standardowe (*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

namy, skala zakotwiczona została w roku 2003). Przerywane linie wyznaczają przedziały ufności skonstruowane dzięki procedurze *bootstrap*. Jak widać, poziom umiejętności uczniów w kolejnych latach okazał się bardzo stabilny. Jest to jedyny jednoznaczny wniosek wynikający z przedstawionych danych: umiejętności humanistyczne uczniów nie zmieniły się znacząco w ciągu badanych 9 lat.

Jedyne wyraźne, choć niewielkie, zmiany poziomu umiejętności uczniów można odnotować w roku 2002, 2010, a zwłaszcza w roku 2006 (rocznik ten odznacza się bowiem najwyższym poziomem umiejętności). Trudno jednak stwierdzić, czy jest to wynik jakieś specyficznej cechy tej kohorty, przeprowadzanego egzaminu czy właściwości przyjętego schematu zrównywania.

W Tabeli 3 i na Rysunku 2 przedstawiono oszacowania odchylenia standardowego rozkładu wyników uczniów z egzaminu gimnazjalnego z części humanistycznej przedstawio-



Rysunek 2. Zróżnicowanie (odchylenie standardowe) zrównanych wyników części humanistycznej egzaminu w latach 2002–2010.

Tabela 4
Średnie zrównanych wyników części matematyczno-przyrodniczej egzaminu w latach 2002–2010[*]

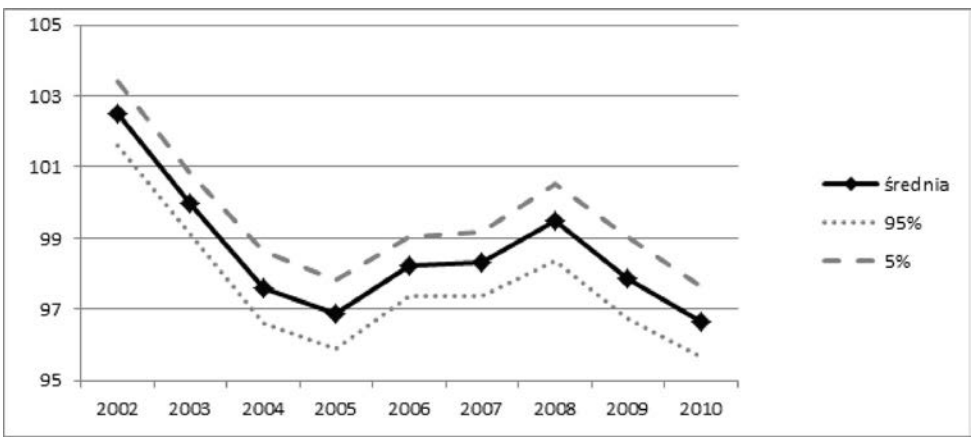
Rok	Średnia	SE_r (*bootstrap*)	95% CI (*bootstrap*)	
2002	102,50	0,56	101,60	103,41
2003	100,00	0,52	99,14	100,86
2004	97,60	0,60	96,61	98,63
2005	96,89	0,59	95,90	97,84
2006	98,23	0,51	97,37	99,04
2007	98,30	0,56	97,37	99,18
2008	99,47	0,65	98,36	100,52
2009	97,85	0,67	96,74	99,05
2010	96,65	0,59	95,66	97,63

[*] Średnie zrównanych wyników części humanistycznej egzaminu w latach 2002–2010 na skali 100;15.

ne na skali wyników z 2003 roku. Wyniki zaprezentowano w analogiczny sposób jak średnie wartości rozkładów umiejętności mierzonych egzaminem gimnazjalnym w części humanistycznej. Dla każdego roku podano odchylenie standardowe rozkładu (*SD*), błąd standardowy *bootstrap* i przedziały ufności.

Tak jak w przypadku średniego poziomu umiejętności mierzonego egzaminem gimnazjalnym, w dynamice zmian zróżnicowania wyników egzaminacyjnych nie widać wyraźnego trendu. Największą różnicę zmiany odchylenia standardowego odnotujemy w latach 2006–2007. Różnica ta nie wpływa jednak znacząco na ogólny obraz stabilności odchylenia standardowego rozkładu wyników między latami.

W Tabeli 4 i na Rysunku 3 przedstawiono średnie wyniki uczniów z egzaminu gimnazjalnego w części matematyczno-przy-



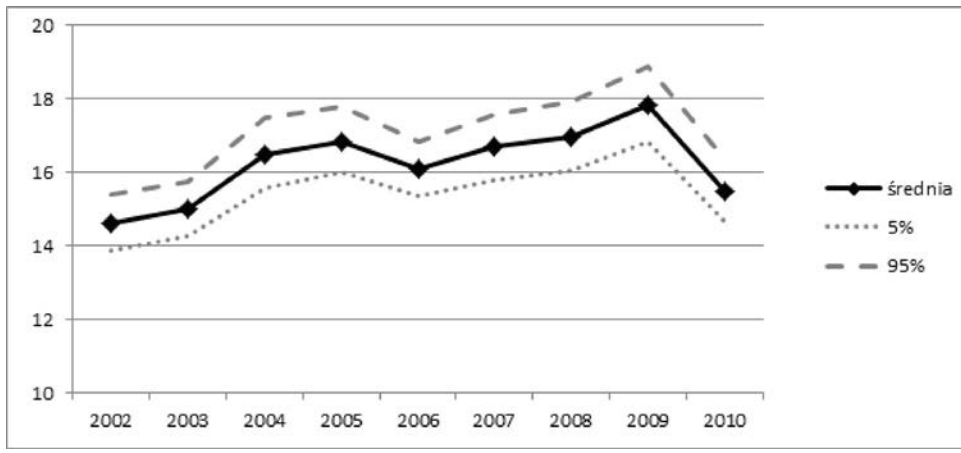
Rysunek 3. Średnie zrównanych wyników części matematyczno-przyrodniczej egzaminu w latach 2002–2010.

Tabela 5
Zróżnicowanie (odchylenie standardowe) zrównanych wyników części matematyczno-przyrodniczej egzaminu w latach 2002–2010

Rok	Odchylenie standardowe (*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

rodniczej po dokonaniu zrównania. Tak jak w przypadku części humanistycznej, podano średni poziom umiejętności uczniów zakotwiczony w roku 2003, gdzie średnią ustalono na 100, a odchylenie standardowe na 15. W tabeli podano również błąd standardowy oszacowania oraz przedziały ufności oszacowane na podstawie procedury *bootstrap*. Średni poziom umiejętności wraz z zarysowanymi przedziałami ufności przedstawiono na Rysunku 3.

Zrównane wyniki egzaminu gimnazjalnego w części matematyczno-przyrodniczej pokazują spadek średniego poziomu umiejętności polskich gimnazjalistów – umiejętności mierzonych testem matematyczno-przyrodniczym od roku 2002 do 2005. Występuje również nieznaczny trend wzrostowy w latach 2005–2008 i kolejny nieznaczny trend spadkowy w latach 2008–2010. Należy przy tym zaznaczyć, iż obydwa trendy są słabe i nale-



Rysunek 4. Zróżnicowanie (odchylenie standardowe) zrównanych wyników części matematyczno-przyrodniczej egzaminu w latach 2002–2010.

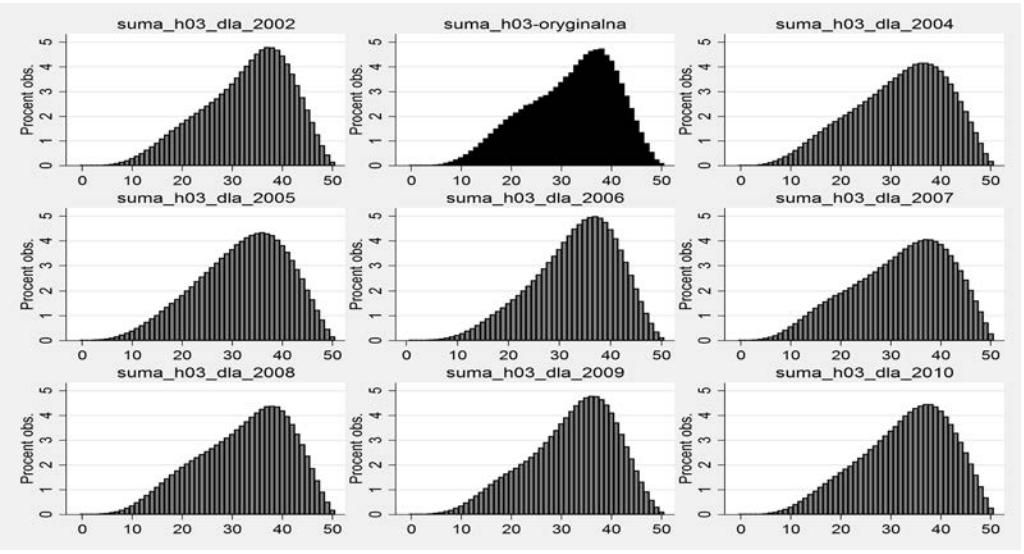
ży być bardzo ostrożnym w interpretacji tych zmian.

Podobnie jak w przypadku części humanistycznej egzaminu gimnazjalnego, prezentujemy wartości odchyleń standardowych rozkładów wyników po zrównaniu, będące wyznacznikiem zróżnicowania indywidualnych wyników. Wyniki te przedstawione zostały w Tabeli 5, a w graficznej formie na Rysunku 4. Prezentacja odchylenia standardowego wyników egzaminacyjnych na zrównanej skali z części matematyczno-przyrodniczej egzaminu gimnazjalnego nie odbiega od prezentacji przedstawionej w poprzednim punkcie dla części humanistycznej. Podano tutaj wartość odchylenia standardowego, błąd standardowy *bootstrap* jego oszacowania oraz 95% przedział ufności.

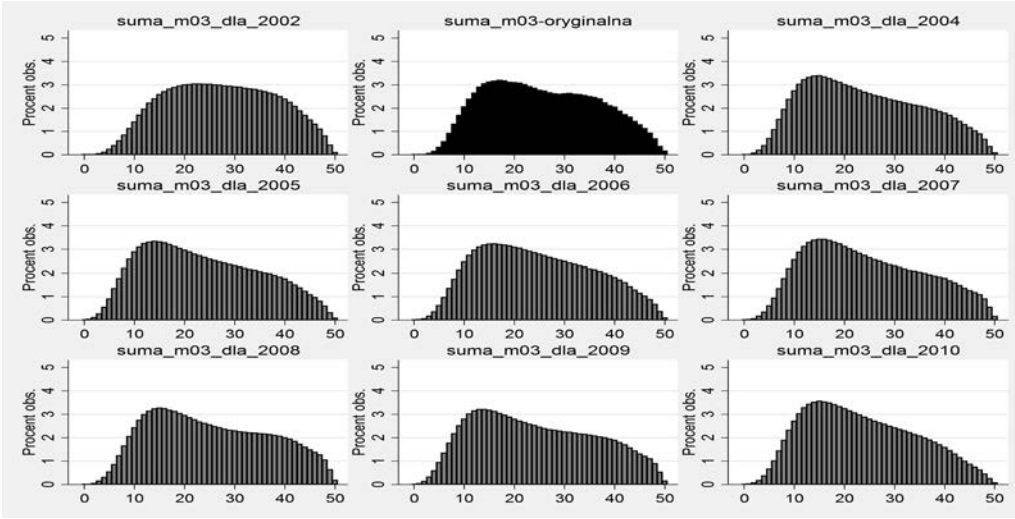
W przypadku odchylenia standardowego obserwujemy istotny i ciągły jego wzrost w latach 2002–2009 i nagły spadek w roku 2010, kiedy poziom indywidualnego zróżnicowania uczniów był podobny jak w roku 2003.

Rezultaty zrównywania wyników obserwowanych

W wyniku zastosowanej procedury zrównywania dysponujemy zakotwiczonym na wspólnej skali rozkładem poziomu umiejętności θ dla każdego roku oraz parametrami zadań opisującymi prawdopodobieństwo udzielenia określonej odpowiedzi na zadania egzaminacyjne w zależności od θ. Informacje te pozwalają na oszacowanie, jak wyglądałby rozkład sumarycznej liczby punktów z dowolnego egzaminu w latach 2002–2010, gdyby rozwiązywany był przez populację uczniów z dowolnej kohorty w latach 2002–2010. W szczególności możliwe jest oszacowanie, jak wyglądałyby wyniki uczniów z lat 2002–2010, gdyby rozwiązywali egzamin z bazowego roku 2003. Histogramy ilustrujące, jak wyglądałby rozkład wyników z części humanistycznej egzaminu gimnazjalnego z 2003 roku w innych latach, przedstawiono na Rysunku 5, analogiczne wykresy dla części matematyczno-przyrodniczej egzaminu znajdują się na Rysunku 6.



Rysunek 5. Rozkład wyników obserwowanych humanistycznej części egzaminu gimnazjalnego przedstawiony na skali wyników obserwowanych egzaminu z 2003 roku (rok 2003 wyróżniony ciemniejszym kolorem).



Rysunek 6. Rozkład wyników obserwowanych matematyczno-przyrodniczej części egzaminu gimnazjalnego przedstawiony na skali wyników obserwowanych egzaminu z 2003 roku (rok 2003 wyróżniony ciemniejszym kolorem).

Histogramy dla lat 2002 oraz 2004–2010 na Rysunkach 7 oraz 8 zostały stworzone na podstawie $5{\cdot}10^6$ zasymulowanych wektorów uczniowskich odpowiedzi na zadania egzaminu z 2003 roku, natomiast wyróżniony rozkład dla 2003 roku jest oryginalnym rozkładem punktów zdobytych przez uczniów z tego roku.

Wykresy na Rysunkach 5 oraz 6 ilustrują, w jaki sposób różnice w poziomie umiejętności uczniów między latami 2002–2010, które przedstawiono wcześniej, podając średnie i odchylenia standardowe na skali θ (Rysunki 1–4 oraz Tabele 2–5), przekładałyby się na wyniki obserwowane uczniów, gdyby w każdym roku egzamin miał w pełni równoważne właściwości psychometryczne do egzaminu z 2003 roku. Wszelkie różnice w kształcie rozkładów na Rysunkach 5 oraz 6 są konsekwencją oszacowanych różnic w poziomie umiejętności między kohortami gimnazjalistów.

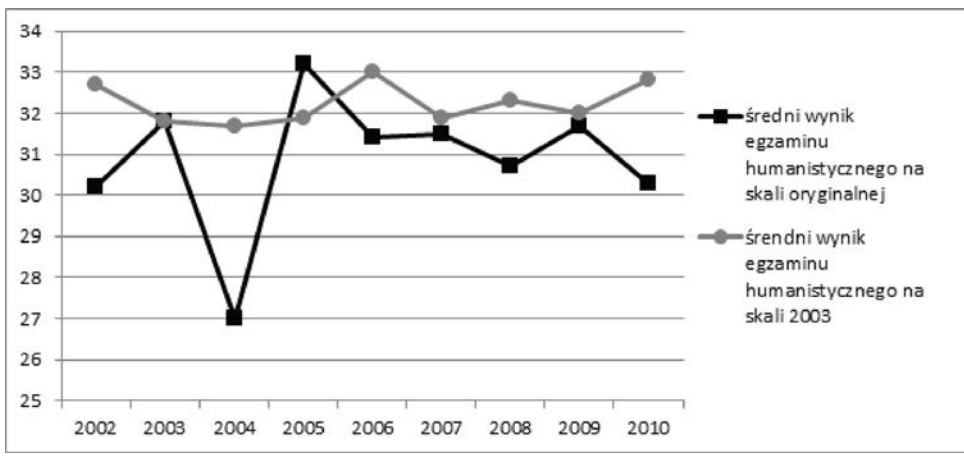
Przykładowo w teście matematyczno-przyrodniczym najwyższą średnią na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2003 uzyskali uczniowie z roku 2002 (102,50), a najniższą z roku 2010 (96,65). Skutkuje to w 2002 roku praktycznie symetrycznym rozkładem wyników obserwowanych egzaminu (w skali z roku 2003) oraz zdecydowanie prawostronnie skośnym rozkładem wyników obserwowanych w tym teście dla uczniów z 2010 roku. Natomiast, im mniej rozkłady umiejętności na skali o średniej 100 i odchyleniu standardowym 15 się różnią, tym subtelniejsze są odpowiednie różnice w rozkładach wyników obserwowanych.

Rozkłady na Rysunkach 5 oraz 6 opisano za pomocą średniej oraz odchylenia standardowego w Tabeli 6. W tabeli zamieszczono dla porównania również parametry rozkładu wyników obserwowanych egzaminów faktycznie przeprowadzanych w latach 2002–2010, co pozwala na poczynienie bardzo interesujących obserwacji. Graficznie średnie z oryginalnych egzaminów w zestawieniu ze średnimi, jakie uzyskaliby uczniowie, gdyby rozwiązywali test z 2003 roku, przedstawiono na Rysunku 7 (część humanistyczna) oraz na Rysunku 8 (część matematyczno-przyrodnicza).

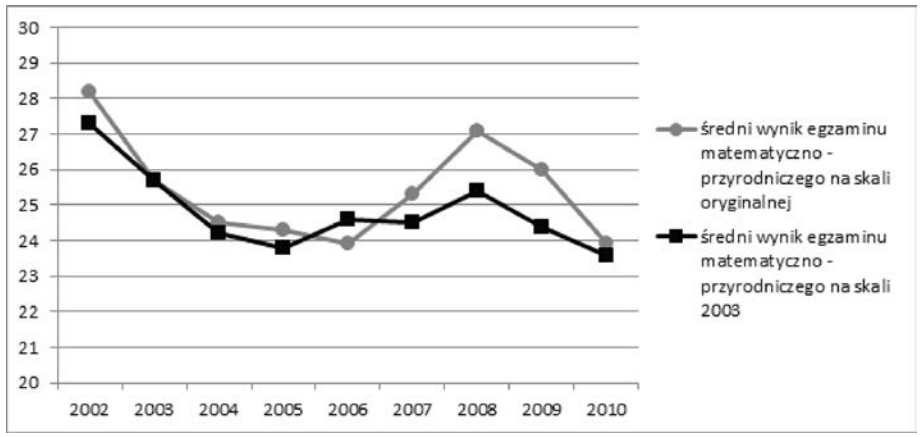
Tabela 6
Średnie oraz odchylenia standardowe wyników obserwowanych egzaminów gimnazjalnych dla oryginalnego testu oraz na skali wyników egzaminu z 2003 roku

Rok	Część humanistyczna				Część matematyczno-przyrodnicza			
	średnia na skali oryg.	średnia na skali 2003	odch. stand. na skali oryg.	odch. stand. na skali 2003	średnia na skali oryg.	średnia na skali 2003	odch. stand. na skali oryg.	odch. stand. na skali 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

Widzimy (zwłaszcza dla części humanistycznej – Rysunek 7), że fluktuacja średnich z prawdziwych egzaminów gimnazjalnych między latami jest o wiele większa niż fluktuacja między średnimi, jakie uzyskiwaliby uczniowie, gdyby rozwiązywali w każdej edycji test o takich samych właściwościach psychometrycznych, jakie miał egzamin w 2003 roku. W szczególności bardzo duże różnice w średnich wynikach testu humanistycznego w trzech kolejnych latach 2003–2005 (średnie odpowiednio: 31,8; 27,0 oraz 33,2) kontrastują z minimalnymi różnicami (rzędu ± 0,1 punktu) między średnimi dla tych samych lat na skali egzaminu z 2003 roku. Wyniki zrównania pokazu-



Rysunek 7. Średnie wyników obserwowanych humanistycznej części egzaminu gimnazjalnego dla oryginalnego testu oraz na skali wyników egzaminu z 2003 roku.



Rysunek 8. Średnie wyników obserwowanych matematyczno-przyrodniczej części egzaminu gimnazjalnego oryginalnego testu oraz na skali wyników egzaminu z 2003 roku.

ją, że te różnice są konsekwencją znacznych zmian w poziomie trudności egzaminów, a nie zmian poziomu umiejętności uczniów.

Przedstawione wyniki przekonują o wątpliwej przydatności skali sumarycznych wyników obserwowanych (lub procentu maksymalnej liczby punktów) do porównywania umiejętności uczniów między latami. Ponadto nasuwa się pytanie o skuteczność procedur kontroli poziomu trudności egzaminów podczas ich konstrukcji – fluktuację poziomu trudności egzaminów rzędu 6 punktów (12% maksymalnego wyniku w teście) między dwoma sąsiednimi edycjami egzaminu nie sposób zaliczyć do dobrych praktyk!

Pewnego komentarza wymaga pojawienie się oryginalnych parametrów egzaminu, jak i parametrów na skali zrównanej do egzaminu z roku 2003 także dla roku 2003 w Tabeli 6 oraz na Rysunkach 7 i 8. Nie ma oczywiście potrzeby zrównywania egzaminu z roku 2003 do siebie samego – wykonano je jednak dla celów diagnostycznych. Dysponując parametrami modelu IRT dla rozkładu umiejętności uczniów oraz dla zadań w 2003 roku, możemy także w analogiczny sposób jak dla innych lat oszacować rozkład wyników obserwowanych z roku 2003. Mimo że nie ma potrzeby szacowania tego rozkładu dla porównania z innymi latami (gdyż jest on zaobserwowany), oszacowanie go na podstawie modelu IRT pozwala ocenić dobroć, z jaką statystyczny model wykorzystany do zrównywania przewiduje wyniki dla 2003 roku. Okazuje się, że dla obu części egzaminu gimnazjalnego (humanistycznej i matematyczno-przyrodniczej) średnie rozkładu wyników obserwowanych oszacowane na podstawie modelu IRT są do pierwszego miejsca po przecinku identyczne z prawdziwymi średnimi, a odchylenia standardowe są niedoszacowane jedynie o dziesiątą część punktu. Oznacza to, że model IRT pozwala na oszacowanie wyników obserwowanych w egzaminach z 2003 roku z bardzo dużą precyzją, co w konsekwencji uwiarygadnia prezentowane rozkłady wyników egzaminów z 2003 roku w innych latach.

Dzięki rozkładom wyników obserwowanych egzaminów humanistycznego oraz matematyczno-przyrodniczego z 2003 roku dla wszystkich populacji uczniów w latach 2002–2010 (Rysunek 5 i 6) można stworzyć tablice przeliczeniowe pozwalające przyporządkować uczniowi piszącemu egza-

min w roku X wynik, jaki uzyskałby na egzaminie z 2003 roku – na podstawie wyniku uzyskanego w roku X. Wystarczy w tym celu dokonać zwykłego ekwicentylowego zrównania wyników sumarycznych testu z roku X (zaobserwowane) z wynikami sumarycznymi w teście 2003 dla tego roku (zasymulowane zgodnie z modelem IRT). Taką tablicę przeliczeniową dla egzaminów humanistycznych przestawiono w Tabeli 7, a dla egzaminów matematyczno-przyrodniczych w Tabeli 8.

Podobnie jak wcześniej, do celów diagnostycznych w Tabelach 7 oraz 8 zamieszczono kolumnę pozwalającą na przeliczenie wyników roku 2003 na wyniki roku 2003 oszacowane na podstawie modelu IRT. Pozwala ona ocenić wiarygodność przeliczania punktów sumarycznych między egzaminami na podstawie wykorzystanego do zrównania modelu statystycznego. Okazuje się, że dla testu humanistycznego model statystyczny sugeruje błędne przeliczenia dla uczniów uzyskujących w egzaminie 0–3 punkty, a dla testu matematyczno-przyrodniczego dla uczniów uzyskujących 0 punktów. Opisany obszar „niepewnych" przeliczeń między egzaminami został w tabelach zaznaczony szarym tłem. W pozostałych zakresach punktowych przeliczenie pomiędzy punktami faktycznie uzyskanymi w 2003 a punktami sugerowanymi przez model IRT jest w pełni zgodne. Biorąc pod uwagę, że w roku 2003 w teście humanistycznym 0–3 punkty uzyskało 57 uczniów z 551 150 (0,0103% obserwacji), a w teście matematyczno-przyrodniczym 0 punktów uzyskało 9 uczniów z 548 716 (0,0016% obserwacji), należy uznać, że ów obszar „niepewności" przeliczeń nie ma żadnego praktycznego znaczenia. Jest to jeszcze jeden dowód na dobroć dopasowania modelu IRT do danych z egzaminu z 2003 roku, a także na poprawność ekwicentylowego zrównania wyników obserwowanych na podstawie modelu IRT.

Analiza danych ukazanych w tabelach przeliczeniowych pozwala na odnotowanie kolejnych kilku bardzo interesujących zależności między poziomami wyników uczniów z różnych lat. Rozważmy (Tabela 7) przykład ucznia, który w 2004 roku z testu humanistycznego zdobył 27 punktów, oraz ucznia, który z testu humanistycznego w 2005 również zdobył 27 punktów. Pierwszy na wspólnej skali testu z 2003 uzyskałby 33 punkty, natomiast drugi na tej samej skali uzyskałby 25 punktów. W tym przykładzie różnica między wynikami dwóch uczniów zdających egzaminy w następujących po sobie latach powinna wynosić faktycznie aż 8 punktów, choć niezrównane wyniki z egzaminów, które zdawali, sugerują taki sam poziom umiejętności!

Różnice podobnego rzędu między rokiem 2004 a 2005 dla testu humanistycznego (7–8 punktów) można zaobserwować w całym przedziale punktów od 21 do 37. Jest to niezwykle ważna obserwacja gdyż ten przedział znajduje się w centrum rozkładu wyników i odpowiada 60,1% oraz 54,0% całej populacji uczniów w tych rocznikach odpowiednio! O tym, jakie konsekwencje tego typu różnice w punktacji miałyby, gdyby wystąpiły np. w przypadku egzaminu maturalnego, którego wyniki są stosowane do celów rekrutacyjnych, nie trzeba wspominać…

W omawianym przykładzie istotne jest to, że dla uczniów, którzy w teście humanistycznym w latach 2004 i 2005 uzyskali bardzo wysokie (powyżej 45 punktów) lub bardzo niskie (poniżej 10 punktów) wyniki, różnica na skali wyników testu z roku 2003 maleje do 1–2 punktów. Zaobserwowana różnica średnich wyników egzaminów na skali z 2003 roku, wynosząca 6,2 punktu (Tabela 6), jest zatem średnią bardzo dużych różnic dla uczniów o przeciętnym poziomie umiejętności (a takich jest z definicji najwięcej) i stosunkowo małych różnic na jego skrajach. Skoro funkcja przeliczająca wyniki mię-

Tabela 7

Tablica przeliczeniowa obserwowanych wyników humanistycznej części egzaminu gimnazjalnego na wyniki obserwowane w roku 2003

Liczba punktów	Przeliczenie wyników na skalę egzaminu z roku 2003								
	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

Tabela 8

Tablica przeliczeniowa obserwowanych wyników matematyczno-przyrodniczej części egzaminu gimnazjalnego na wyniki obserwowane w roku 2003

Liczba punktów	Przeliczenie wyników na skalę egzaminu z roku 2003								
	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

dzy egzaminami ma charakter nieliniowy, to należy wyciągnąć wniosek, że stosowanie wszelkich przekształceń wyników o charakterze liniowym (jak na przykład standaryzacja bez wcześniejszej normalizacji) nie będzie w stanie rozwiązać problemu nierównoważności wyników egzaminu między różnymi edycjami.

Dla skontrastowania do przytoczonego w poprzednich akapitach skrajnego przykładu różniących się znacznie łatwością testów humanistycznych z lat 2004 i 2005, można posłużyć się testami matematyczno-przyrodniczymi (z tych samych lat) uczniów również uzyskujących 27 punktów (Tabela 8). Z przeliczeń wynika, że uczeń, który w 2004 roku uzyskał w teście matematyczno-przyrodniczym 27 punktów, na skali testu z 2003 roku również powinien uzyskać 27 punktów, podobnie jak uczeń zdający egzamin w 2005 roku. Wynik 27 punktów opowiada w teście matematycznym z lat 2003–2005 takiemu samemu poziomowi umiejętności. Okazuje się, że egzaminy matematyczno-przyrodnicze w latach 2003–2005 były bardzo zbliżone pod względem poziomu trudności, a jednocześnie (por. Rysunek 5 lub 6) ze zrównania wynika znaczna zmiana poziomu umiejętności uczniów z tych roczników.

Zrównanie wyników ukazało diametralnie odmienne interpretacje różnic w średnich wynikach egzaminu gimnazjalnego między latami 2004 i 2005 dla części matematyczno-przyrodniczej i humanistycznej. W przypadku części humanistycznej egzaminu różnice średnich wynikały głównie z różnic w trudności testów, natomiast w przypadku części matematyczno-przyrodniczej przyczyn różnicy średnich należy głównie upatrywać w różnicy poziomów umiejętności matematyczno-przyrodniczych uczniów, którzy zdawali egzamin w kolejnych latach. Bez zrównania wyników taka analiza byłaby niemożliwa.

Weryfikacja procedury zrównywania

Jedną z możliwości empirycznej weryfikacji procedury zrównywania egzaminu jest porównanie wyników zrównania opracowanych przez Pracownię Analiz Osiągnięć Uczniów IBE z wynikami uzyskanymi innym narzędziem mierzącym podobne umiejętności, którego jakość została uznana i potwierdzona. Sposobności do takiego porównania dostarcza przeprowadzane w cyklu trzyletnim badanie PISA (*Programme for International Student Assessment*). Jest to międzynarodowe badanie prowadzone przez Organizację Współpracy Gospodarczej i Rozwoju (OECD) od 2000 roku. Sposób konstrukcji testów używanych w PISA jest bardziej wyrafinowany niż metody stosowane w polskim systemie oświaty. Test PISA został zaprojektowany przez czołowych ekspertów z całego świata. Zadania przygotowane do pomiaru przechodzą rygorystyczną serię testów i badań pilotażowych, recenzowane są przez ekspertów ze wszystkich krajów uczestniczących w badaniu, a stosowane analizy statystyczne zapewniają wysoką jakość skal.

Co ważne dla nas, w badaniu PISA wyniki z kolejnych cykli również są wiązane ze sobą tak, by wyniki z kolejnych edycji były bezpośrednio porównywalne. Wiązanie wyników PISA odbywa się za pomocą schematu wewnętrznej kotwicy, odmiennego niż schemat, który przyjęliśmy w naszym zrównaniu. W każdej edycji badania uczniowie rozwiązują pewną pulę zadań pojawiających się w poprzednich edycjach (około 20 zadań z każdej dziedziny). Następnie wyniki są zestawiane przy użyciu wielowymiarowego modelu Rascha.

Zastosowanie wewnętrznej kotwicy do zrównania wyników jest dobrym rozwiązaniem dlatego, że uczniowie rozwiązują zadania w kolejnych edycjach w porówny-

walnych warunkach motywacyjnych. Pod tym względem metodologia zrównywania PISA przewyższa schemat zrównywania *post hoc*, wymuszony w naszych badaniach przez konstrukcję systemu egzaminacyjnego w Polsce, który nie uwzględnia zadań wewnętrznie kotwiczących egzaminy. Czynnik motywacji w raportowanym zrównaniu wyników egzaminacyjnych był dodatkowo kontrolowany.

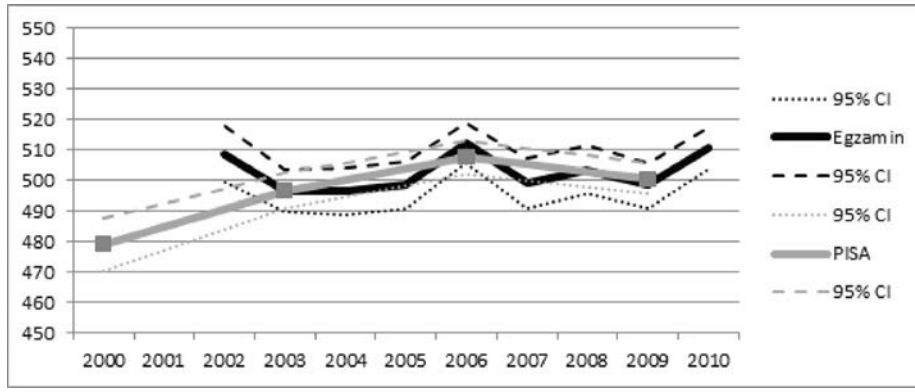
Należy dodać, że założenia pomiarowe w badaniu PISA w dużym stopniu są zgodne z programowymi zapisami wyznaczającymi cele polskiego egzaminu gimnazjalnego. W PISA pomiar skupia się na ocenie umiejętności posługiwania się pojęciami i ich rozumienia oraz posługiwania się wachlarzem ogólnych umiejętności. Z założenia pomiar ma dotyczyć wiadomości i umiejętności niezbędnych uczniom w życiu dorosłym, na rynku pracy i do tego, aby w pełni funkcjonować we współczesnym społeczeństwie demokratycznym (PISA 2003; PISA 2006; PISA 2009).

Podobieństwa koncepcji testów PISA oraz polskiego egzaminu gimnazjalnego nie oznaczają oczywiście tożsamości samych testów. Test gimnazjalny to doniosły egzamin państwowy, obowiązkowy dla wszystkich uczniów kończących gimnazjum. Badanie PISA przeprowadzane jest na próbie losowej, nie jest obowiązkowe i nie jest dla ucznia testem wysokiej stawki. Egzamin gimnazjalny ma na celu pomiar indywidualnych umiejętności ucznia, zadaniem PISA jest jak najlepsze oszacowanie poziomu osiągnięć całej populacji uczniów 15-letnich w poszczególnych krajach. Ta ostatnia różnica oznacza, że w badaniu PISA maksymalizuje się liczbę zadań wykorzystywanych w teście, choć uczniowie rozwiązują ich różne podzbiory. Dzięki temu pomiar danej dziedziny wiedzy może być szerszy; ponadto zapobiega się temu, że ważne obszary wiedzy nie zostają poddane pomiarowi z powodu braku wystarczającej liczby zadań w teście. Dla oszacowania wyniku indywidualnego jest to jednak sytuacja mniej korzystna, ponieważ dochodzi dodatkowe źródło błędu pomiaru – próbkowanie zadań w obrębie pełnego zestawu tworzącego dany test.

Wymienione różnice nie są jednak na tyle duże, by były przeszkodą w porównaniu obydwu badań. Oczekujemy wysokiej zbieżności między naszym podejściem do zrównywania oraz podejściem reprezentowanym w badaniu PISA. Oczywiście wyniki zrównania nie mogą być tożsame, ale wszelkie podobieństwa będą świadczyć na korzyść zastosowanej przez nas metody.

Na Rysunku 9 przedstawiono oszacowanie umiejętności czytania ze zrozumieniem w badaniu PISA oraz oszacowanie umiejętności humanistycznych mierzonych na podstawie części humanistycznej egzaminu gimnazjalnego. Dane zostały przeskalowane do skali PISA, aby poziom umiejętności uczniów w roku 2003 był w obydwu badaniach identyczny. Równość między obydwoma badaniami w 2003 jest zatem narzucona. Różnice w innych latach wynikają z zastosowania innych testów i metodologii zrównywania. Dodatkowo na Rysunku 9 przedstawiono 95% przedział ufności dla wyników badania zrównującego i badania PISA.

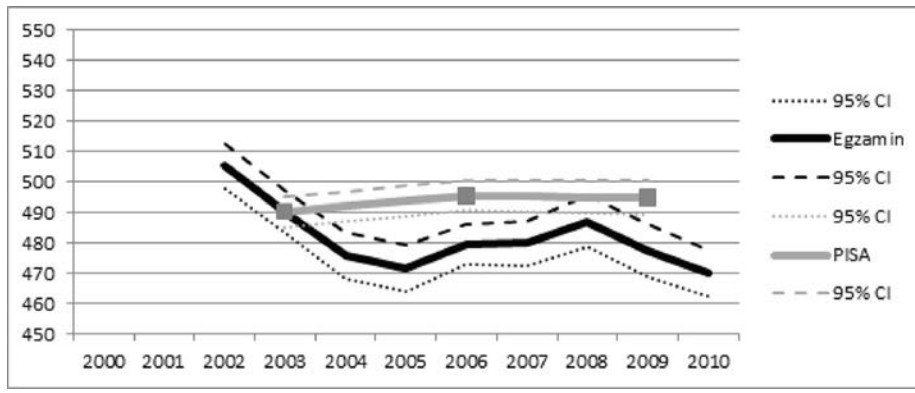
Wyniki zrównania w badaniu zrównującym, jak i w badaniu PISA, są niemal identyczne. W latach 2006 i 2009 wyniki prawie się pokrywają. Niestety, w badaniach zrównujących nie możemy znaleźć żadnego potwierdzenia wzrostu umiejętności polskich gimnazjalistów między rokiem 2000 a 2003. W badaniu zrównującym dysponujemy jedynie danymi sięgającymi roku 2002, które raczej przeczą takim wnioskom. Należy pamiętać, że wyniki z roku 2002 trzeba analizować z dużą ostrożnością, gdyż, jak już wcześniej pisaliśmy, był



Rysunek 9. Średnie wyniki uczniów szkół gimnazjalnych w latach 2002–2010, wyniki zrównane (część humanistyczna), skala zakotwiczona w badaniu PISA (czytanie ze zrozumieniem) z roku 2003 oraz wynik badań PISA (czytanie ze zrozumieniem) lata: 2000, 2003, 2006 i 2009.

to pierwszy rok egzaminu gimnazjalnego, kiedy nie wszystkie procedury były dopracowane, a wyjątkowość pierwszego egzaminu musiała wpływać na wyniki egzaminacyjne.

Podobnie jak w przypadku części humanistycznej egzaminu gimnazjalnego i czytania ze zrozumieniem w badaniu PISA dokonano porównania części matematyczno-przyrodniczej egzaminu gimnazjalnego i badania umiejętności matematycznych w badaniu PISA. Wyniki tego porównania przedstawione zostały na Rysunku 10. Ciemnoszara linia przedstawia wyniki badania zrównującego, jasnoszara wyniku uzyskanego w badaniu PISA. Tak jak poprzednio, skale obydwu badań zostały zakotwiczone w 2003 roku w wyniku PISA (tym razem z matematyki). W przypadku komponentu matematycznego badanie PISA zapewnia porównywalność jedynie dla lat 2003–2009, dlatego tylko te dane prezentujemy na wykresie.



Rysunek 10. Średnie wyniki uczniów szkół gimnazjalnych w latach 2002-2010, wyniki zrównane (część matematyczno-przyrodnicza) skala zakotwiczona w badaniu PISA (matematyka) z roku 2003 oraz wynik badań PISA (matematyka) lata: 2003, 2006 i 2009.

W przypadku umiejętności matematyczno-przyrodniczych wynik zrównywania w polskim badaniu zrównującym różni się od wyników PISA. Wyniki PISA z lat 2003–2009 pozwalają wnioskować o ich stabilności, wyniki badania zrównującego ujawniają natomiast tendencję spadkową. Z drugiej strony gdyby z badania zrównującego usunąć lata 2002 i 2003, można by bronić tezy o względnej stabilności wyników, ponieważ poziomy umiejętności uczniów w badaniu zrównującym z roku 2004 i 2010 nie różnią się od siebie statystycznie.

Literatura

Davier von, A. A., Holland, P. W. i Thayer, DT (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. i Béguin A. A. (1996). *Appropriateness of IRT observed-score equating.* Research Report 1996–2. Enschede: University of Twente.

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

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

OECD (2003). Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2003 w POLSCE [Niepublikowany maszynopis].

OECD (2006). Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2006 w POLSCE. Pobrano z: http://www.ifispan.waw.pl/pliki/pisa_raport_2006.pdf

OECD (2009). Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2009 w POLSCE. Pobrano z: http://www.ifispan.waw.pl/pliki/pisa_2009.pdf

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

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