dr Henryk Szaleniec
Instytut Badań Edukacyjnych

Możliwości wykorzystania zrównanych wyników egzaminu gimnazjalnego w szkolnej diagnozie

Wstęp

W artykule przedstawiono w zarysie główne założenia studium badawczego, którego rezultatem są porównywalne między latami wyniki egzaminu gimnazjalnego w części humanistycznej i matematyczno-przyrodniczej za lata 2002-2010. Zostały także zaprezentowane zrównane wyniki średnie dla kraju w analizowanym dziewięcioletnim okresie. W drugiej części tekstu zaproponowano wykorzystanie do szkolnej diagnozy edukacyjnej zrównanych i wykalibrowanych wyników egzaminów gimnazjalnych odniesionych do roku 2003, który został wybrany jako rok bazowy. Wyniki zostały zaprezentowane w standardowej skali o średniej 100 i odchyleniu standardowym 15, która od szeregu lat stosowana jest przy okazji prezentacji edukacyjnej wartości dodanej (EWD). Ponieważ jednym z celów studium badawczego było szerokie udostępnienie zrównanych wyników do lokalnych diagnoz na różnych szczeblach struktury systemu edukacyjnego, Instytut Badań Edukacyjnych (IBE) przygotował serwis internetowy, który powinien zapewnić dostęp do zrównanych wyników zagregowanych dla szkoły, gminy, powiatu, województwa i kraju. Dlatego też w niniejszym artykule zasygnalizowano zasady funkcjonowania tego serwisu i prezentowania wyników zarówno w przedstawieniach graficznych, jak i tabelarycznych. To pierwsza w Polsce próba komunikowania wyników egzaminacyjnych w taki sposób.

Zrównywanie wyników egzaminacyjnych jako istotny problem systemu zapewniania jakości egzaminowania

W założeniach i w praktyce polskiego systemu egzaminów zewnętrznych główny akcent został położony na porównywalność egzaminów w danej sesji. Jak dotychczas nie jest brana pod uwagę równoważność wyników egzaminacyjnych dla kolejnych populacji przystępujących do tego samego egzaminu. Innymi słowy, z powodu nieuniknionych wahań trudności arkuszy egzaminacyjnych stosownych w kolejnych latach, skale, w których corocznie komunikowane są wyniki, nie są ekwiwalentne. Problem ten, choć był podejmowany pilotażowo przez CKE w latach 2004-2005 dla sprawdzianu (Niemierko B., 2004; Szaleniec H., 2005) oraz w roku 2006 i 2007 dla gimnazjum, nie doczekał się jeszcze rozwiązań systemowych.

Brak porównywalności wyników pomiędzy latami ma kilka istotnych konsekwencji w zakresie komunikowania rezultatów egzaminacyjnych. Po pierwsze, obecny sposób prezentowania wyników nie jest fair w stosunku do maturzystów podczas rekrutacji na wyższe uczelnie. Maturzyści z różnych lat, konkurując o indeks w tym samym roku, przedstawiają na świadectwach wyniki, które obciążone są efektem fluktuacji trudności arkuszy egzaminacyjnych. Dla sprawdzianu i dla

egzaminu gimnazjalnego ten problem nie jest tak wysokiej wagi, gdyż rekrutacja do szkoły wyższego szczebla obejmuje tylko jeden rocznik (tę samą populację). Po drugie, przy obecnym systemie komunikacji wyników egzaminacyjnych utrudnione jest monitorowanie rezultatów egzaminowania w przedziałach czasowych dłuższych niż jeden rok. A także wskaźnik edukacyjnej wartości dodanej obciążony jest efektem fluktuacji trudności arkuszy egzaminacyjnych.

Warto także podkreślić, że wiele ugruntowanych od lat na świecie systemów zajmujących się przeprowadzaniem egzaminów i krajowych oraz międzynarodowych badań testowych stosuje już na poziomie konstrukcji narzędzi egzaminacyjnych mechanizmy pozwalające na zrównywanie wyników w trakcie danej sesji. Między innymi w USA SAT (Scholastic Assessment Test), ACT (American College Testing), w Izraelu - (*Psychometric Entrance Test*), w Szwecji - (*Swedish Scholastic Assessment Test*), w Kanadzie - testy przygotowywane i prowadzone przez EQAO (*Education Equality and Accountability Office*), w Australii - (*National Assessment Program - Literacy and Numeracy*) (Pokropek A., 2011). W Japonii również prowadzone są badania i próby zrównywania post factum wyników egzaminów z języka angielskiego IPET (*Ibaraki Prefecture English Test for High School Students)* (Chisato Saida, Tamaki Hattori, 2008).

W 2010 roku Instytut Badań Edukacyjnych podjął wieloetapowe studium mające na celu zrównanie post factum wyników egzaminacyjnych dla kilku egzaminów. Studium podzielone zostało na cztery etapy obejmujące kolejne egzaminy przeprowadzane przez Centralną Komisję Egzaminacyjną i okręgowe komisje (OKE). Dwa pierwsze etapy dotyczą egzaminów, które wdrożone zostały w 2002 roku, tj. sprawdzian i egzamin gimnazjalny. Dwa kolejne - to egzamin maturalny z matematyki, języka polskiego oraz języka angielskiego. Jako pierwszy egzamin, dla którego przeprowadzono zrównanie i wykalibrowanie wyników do roku bazowego przyjętego jako rok odniesienia, wybrano egzamin gimnazjalny. Wybór nie jest przypadkowy. W 2011 roku, kiedy przeprowadzono badania zrównujące, gimnazjaliści po raz ostatni zdawali egzamin zgodnie z podstawą programową obowiązującą w latach 1999-2011. W 2012 roku do egzaminu przystąpili już uczniowie, których edukacja organizowana była na podstawie nowej podstawy programowej obowiązującej od września 2009 roku[1]. Także począwszy od 2012 roku egzamin gimnazjalny uległ istotnym zmianom. Część humanistyczna została podzielona na egzamin z języka polskiego i egzamin z historii z WOS-em, a część matematyczno-przyrodnicza - na egzamin z matematyki i przedmiotów przyrodniczych. Również egzamin z języka obcego nowożytnego został rozbudowany, obejmując poziom podstawowy dla gimnazjalistów, którzy uczyli się danego języka po raz pierwszy w gimnazjum, i poziom zaawansowany dla uczniów, którzy kontynuowali naukę danego języka.

Studium zrównujące wyniki egzaminu gimnazjalnego

Głównym celem badań, które zostały przeprowadzone w 2011 roku, było doprowadzenie do porównywalności między latami wyników egzaminów gimnazjalnych (część humanistyczna i matematyczno-przyrodnicza) przeprowadzonych

[1] Nowa podstawa programowa została wprowadzona rozporządzeniem Ministra Edukacji Narodowej z dnia 23 grudnia 2008 r. w sprawie podstawy programowej wychowania przedszkolnego oraz kształcenia ogólnego w poszczególnych typach szkół (DzU z 2009 r. nr 4, poz. 17).

w latach 2002-2010 z zastosowaniem arkuszy standardowych (arkusze dla uczniów bez dysfunkcji i uczniów z dysleksją rozwojową) i przedstawienie ich w skali standardowej o średniej 100 i odchyleniu standardowym 15, zakotwiczonej do roku bazowego 2003.

Populację docelową stanowili uczniowie ostatnich klas szkół gimnazjalnych dla młodzieży bez szkół specjalnych i przyszpitalnych. Biorąc pod uwagę efektywność badania (minimalizacja kosztów i minimalizacja błędu pomiaru), rzeczywista populacja losowania została ograniczona do gimnazjalistów w placówkach, w których rocznik przystępujący do egzaminu liczył więcej niż 10 uczniów. Przyjęte ograniczenie spowodowało wyłączenie z operatu losowania około 3,8% szkół i 0,4% uczniów.

Próba losowa uczniów, którzy wzięli udział w dodatkowej sesji zrównującej miesiąc przed egzaminem gimnazjalnym w 2011 roku, wylosowana z powyżej zdefiniowanej populacji składała się z 11 losowo równoważnych podprób uczniów: S_{11}^{1}, S_{11}^{2},..., S_{11}^{11}. Do wyboru próby zastosowano losowanie:

a. warstwowe,
b. proporcjonalne do liczebności uczniów w klasach trzecich gimnazjów,
c. wielostopniowe.

Do przeprowadzenia badań zastosowano plan zgodny ze schematem, który przedstawiony jest w tabeli 1. Zrównywanie przeprowadzone w 2011 roku obejmowało wyniki 10 populacji uczniów od 2002 roku do 2011. Populacje te w schemacie oznaczone są $\mathcal{P}_{02}$, $\mathcal{P}_{03}$, …, $\mathcal{P}_{11}$, z których każda pochodziła z odpowiedniej edycji egzaminu gimnazjalnego i rozwiązywała arkusz egzaminacyjny przeznaczony dla danego roku: T_{02}, T_{03}, …, T_{11}.

Każda z 11 podprób (S_{11}^{1}, S_{11}^{2},..., S_{11}^{11}), w sesji zrównującej rozwiązywała test składający się z dwóch podzbiorów zadań kotwiczących wybranych z poprzednich egzaminów (T^{A}) oraz podzbioru dodatkowych zadań kotwiczących (do zrównywania z przyszłymi egzaminami) - C. Na przykład biorąca udział w badaniu podpróba pierwsza z populacji uczniów 2011 roku S_{11}^{1}, rozwiązywała test zbudowany z podzbiorów kotwiczących zadań T_{02}^{A} (pochodzących z arkusza egzaminacyjnego z 2002 roku) oraz T_{03}^{A} (pochodzących z arkusza egzaminacyjnego z 2003 roku) oraz z zadań dodatkowych C^{1}. Zadania arkuszy egzaminacyjnych z tych dwóch lat (2002 i 2003) nie włączone do testów kotwiczących oznaczone są symbolem T_{02}^{R} oraz T_{03}^{R}. Zadania te rozwiązywane były tylko na egzaminie w odpowiednich latach.

W studium badawczym zastosowano strategię zrównywania opartą na IRT (*Item Response Theory*) i metodę kalibracji łącznej (*concurrent calibration*), która polega na dopasowaniu do wszystkich danych zebranych z podprób S_{11}^{1}, S_{11}^{2},..., S_{11}^{11}), modelu IRT w pojedynczej procedurze estymacji. Istotne jest, że w takim podejściu strukturalny brak odpowiedzi na zadania testu Y w próbie S_{11}^{n} oraz brak odpowiedzi na zadania testu X w próbie S_{11}^{n+1} nie stanowi problemu dla maksymalizujących funkcję wiarygodności metod estymacji wykorzystywanych do szacowania parametrów modelu IRT. W celu uzyskania nieobciążonych

parametrów w planie NEAT[2] (Davier A.A, Holland P.W. i Thayer D.T., 2004), konieczne było bezpośrednie uwzględnienie w modelu IRT odrębnych parametrów rozkładów θ dla populacji $\mathcal{P}_n$ oraz $\mathcal{P}_{n+1}$ i oszacowanie ich z prób $S_{11}{}^n$ oraz $S_{11}{}^{n+1}$.

Tabela 1. Schemat zrównywania wyników egzaminów gimnazjalnych 2002-2011 (źródło: *Analiza porównawcza wyników egzaminów zewnętrznych - gimnazjum. Raport z badań. IBE 2012*)

Dane z sesji zrównującej											Dane egzaminacyjne										Źródło danych	
$\mathcal{P}_{11}$	$\mathcal{P}_{11}$	$\mathcal{P}_{11}$	$\mathcal{P}_{11}$	$\mathcal{P}_{11}$	$\mathcal{P}_{11}$	$\mathcal{P}_{11}$	$\mathcal{P}_{11}$	$\mathcal{P}_{11}$	$\mathcal{P}_{11}$	$\mathcal{P}_{11}$	$\mathcal{P}_{11}$	$\mathcal{P}_{10}$	$\mathcal{P}_{09}$	$\mathcal{P}_{08}$	$\mathcal{P}_{07}$	$\mathcal{P}_{06}$	$\mathcal{P}_{05}$	$\mathcal{P}_{04}$	$\mathcal{P}_{03}$	$\mathcal{P}_{02}$	Populacja	
S_{11}^{11}	S_{11}^{10}	S_{11}^{9}	S_{11}^{8}	S_{11}^{7}	S_{11}^{6}	S_{11}^{5}	S_{11}^{4}	S_{11}^{3}	S_{11}^{2}	S_{11}^{1}											Próba	
800	800	800	800	800	800	800	800	800	800	800	$4 \cdot 10^5$	$4 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	$5 \cdot 10^5$	Liczba obs. (szac.)	
																				✓	T_{02}^R	T_{02} (Test lub zbiór zadań)
										✓										✓	T_{02}^A	T_{02}
																			✓		T_{03}^R	T_{03}
✓		✓							✓	✓									✓		T_{03}^A	T_{03}
																		✓			T_{04}^R	T_{04}
								✓	✓									✓			T_{04}^A	T_{04}
																	✓				T_{05}^R	T_{05}
							✓	✓									✓				T_{05}^A	T_{05}
																✓					T_{06}^R	T_{06}
	✓	✓				✓	✓									✓					T_{06}^A	T_{06}
															✓						T_{07}^R	T_{07}
					✓	✓									✓						T_{07}^A	T_{07}
														✓							T_{08}^R	T_{08}
				✓	✓									✓							T_{08}^A	T_{08}
													✓								T_{09}^R	T_{09}
			✓	✓									✓								T_{09}^A	T_{09}
												✓									T_{10}^R	T_{10}
✓	✓		✓									✓									T_{10}^A	T_{10}
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓										T_{11}	T_{11}
										✓											C^1	Dodatkowe zadania kotwiczące
									✓												C^2	Dodatkowe zadania kotwiczące
								✓													C^3	Dodatkowe zadania kotwiczące
							✓														C^4	Dodatkowe zadania kotwiczące
						✓															C^5	Dodatkowe zadania kotwiczące
					✓																C^6	Dodatkowe zadania kotwiczące
				✓																	C^7	Dodatkowe zadania kotwiczące
			✓																		C^8	Dodatkowe zadania kotwiczące
		✓																			C^9	Dodatkowe zadania kotwiczące
	✓																				C^{10}	Dodatkowe zadania kotwiczące
✓																					C^{11}	Dodatkowe zadania kotwiczące

[2] Schemat zrównywania dla planu nierównoważnych grup z testem kotwiczącym (*Non-Equivalent groups with Anchor Test design).*

Średni wynik egzaminu dla kraju

Wyniki zrównywania dla obydwu części egzaminu gimnazjalnego zakotwiczone zostały względem rezultatów egzaminacyjnych 2003 roku. W procesie zrównywania średnia dla umiejętności egzaminacyjnych uczniów ustawiona została na 0, a odchylenie standardowe na 1. Jako układ odniesienia wybrany został arbitralnie rok 2003. Był to drugi rok egzaminów w Polsce i rok, w którym wprowadzono zewnętrznych obserwatorów podczas sesji egzaminacyjnej. Między innymi brak zewnętrznych obserwatorów był jedną z przyczyn, że nie został wybrany jako punkt wyjściowy rok 2002.

Aby ułatwić komunikowanie wyników, przeskalowano je na skalę o średniej 100 i odchyleniu standardowym 15. Taka skala jest łatwiejsza do prezentacji, ponieważ nie daje ujemnych wyników. Jest także jedną z najbardziej znanych skal standardowych i od szeregu lat używana jest do prezentacji wskaźników edukacyjnej wartości dodanej (EWD). Średnie zrównane wyniki gimnazjalne dla kraju za lata 2002-2010 przedstawione są w tabelach 2. i 3. oraz na rysunkach 1. i 2[3]. Szersze informacje na temat wyników zrównania, metodologii badań i analiz będą dostępne w Kwartalniku „Edukacja" (Szaleniec H., Grudniewska M., Kondratek B., Kulon F., Pokropek A., 2012).

Tabela 2. przedstawia średni dla kraju poziom umiejętności uczniów oszacowany na podstawie wyników egzaminacyjnych w części humanistycznej egzaminu gimnazjalnego w latach 2002-2010. W drugiej kolumnie podany jest średni wynik dla kraju w danym roku. Dla roku 2003 średnia wynosi 100 (rok odniesienia) - w pozostałych latach wynik średni oszacowany jest tak, jak gdyby kolejne populacje pisały egzamin z wykorzystaniem arkusza egzaminacyjnego z 2003 roku. W kolejnej kolumnie 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 mamy do czynienia z losowym błędem (SE_r (bootstrap))[4]. Błąd został oszacowany za pomocą procedury bootstrap. Obok błędu standardowego w tabeli 2. podano wartości wyznaczające 95% 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.

[3] Źródło (tabela 2,3 i rysunek 1,2) - *Analiza porównawcza wyników egzaminów zewnętrznych - gimnazjum. Raport z badań. IBE 2012*

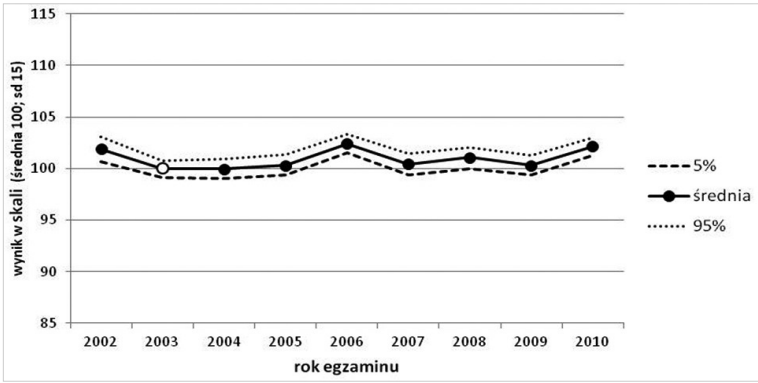
[4] Nie jest to jedyne źródło błędu, które może wpływać na precyzję szacowania. Oprócz błędu wynikającego z doboru próby badawczej uczniów (błąd próbkowania) w procesie zrównywania, w przyjętym schemacie badawczym pojawia się również błąd związany z wyborem próby zadań do testów kotwiczących (błąd zrównywania). Wykorzystanie w procesie badań zrównujących wszystkich zadań z arkuszy standardowych w latach 2002-2010 spowodowałoby znaczny wzrost kosztów badania. Jeżeli chodzi o błąd zrównywania, to jak dotychczas nie istnieją analityczne rozwiązania, a co do procedur replikacyjnych to nie ma zgodności odnośnie ich skuteczności. Dlatego też przedziały ufności i błędy standardowe, konstruowane jedynie na podstawie błędu próbkowania, mogą być nieoszacowane. Proponujemy traktować je jako wskazówkę, a nie jako bazę do formalnych testów statystycznych.

Tabela 2. Średnie zrównane wyniki uczniów szkół gimnazjalnych w latach 2002-2010, skala 100; 15 zakotwiczona w roku 2003, część humanistyczna

Rok egzaminu	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

CI (confidence interval) przedział ufności.

Na rysunku 1. w graficzny sposób przedstawiono wyniki zrównywania dla egzaminu gimnazjalnego w części humanistycznej. Białym kółeczkiem zaznaczony jest rok bazowy 2003, dla którego średnia wynosi 100.



Rysunek 1. Średnie wyniki uczniów szkół gimnazjalnych w latach 2002-2010, skala 100; 15 jest zakotwiczona w roku 2003, część humanistyczna

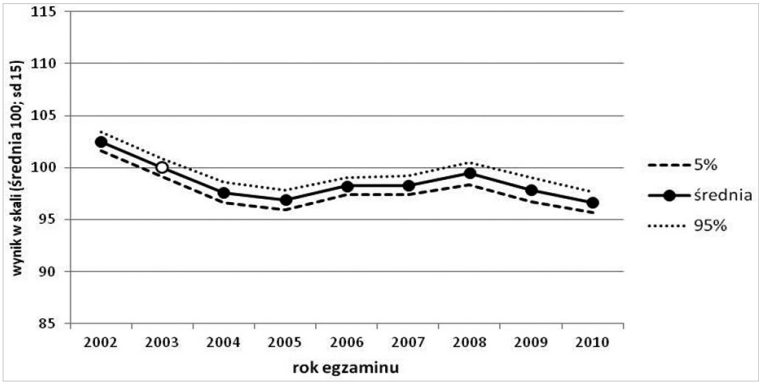
Linią ciągłą oznaczono średnie wyniki egzaminu w poszczególnych latach. Przerywane linie wyznaczają przedziały ufności skonstruowane dzięki procedurze bootstrap. Jak można zauważyć, poziom umiejętności uczniów szacowany na podstawie zrównanych wyników egzaminacyjnych w części humanistycznej w kolejnych latach okazał się być bardzo stabilny i nie wykazuje znaczącego trendu. Innymi słowy możemy powiedzieć, że poziom umiejętności humanistycznych sprawdzanych na egzaminie gimnazjalnym w latach 2002-2010 nie zmienił się znacząco. Niewielkie zmiany poziomu umiejętności uczniów można zaobserwować w roku 2002 oraz w latach 2006, 2010. Szczególnie dla rocznika przystępującego do egzaminu w 2006 roku zaobserwowano stosunkowo największy poziomem umiejętności. Na podstawie dotychczas przeprowadzonych analiz trudno jednoznacznie stwierdzić, czy jest to wynik jakieś specyficznej cechy populacji gimnazjalistów przystępujących do egzaminu w 2006 roku, szczególnych właściwości zastosowanego w danym roku arkusza

egzaminacyjnego lub oceniania zadań otwartych, czy też właściwości przyjętego schematu zrównywania. Zagadnienie to będzie przedmiotem dalszych pogłębionych analiz. Średnie zrównane wyniki dla kraju egzaminu gimnazjalnego w części matematyczno-przyrodniczej przedstawia tabela 3.

Tabela 3. Średnie wyniki uczniów szkół gimnazjalnych w latach 2003-2010, skala 100; 15 zakotwiczona w roku 2003, część matematyczno-przyrodnicza

Rok egzaminu	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

Na rysunku 2. w graficzny sposób przedstawiono wyniki zrównywania dla części matematyczno-przyrodniczej części egzaminu gimnazjalnego.

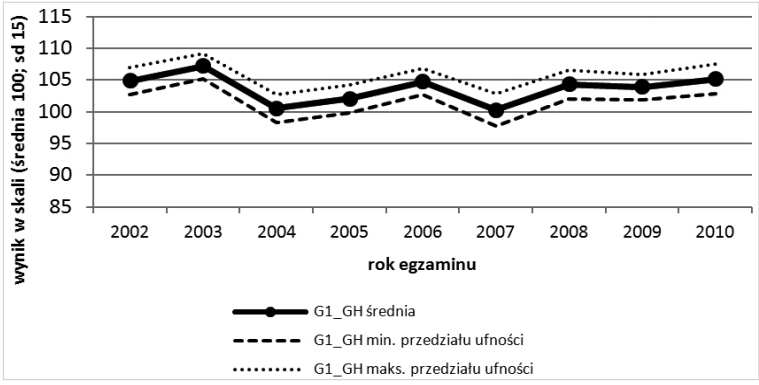


Rysunek 2. Średnie wyniki uczniów szkół gimnazjalnych w latach 2002-2010, skala 100; 15 zakotwiczona w roku 2003, część matematyczno-przyrodnicza

Na podstawie zrównanych wyników egzaminu gimnazjalnego w części matematyczno-przyrodniczej można wnioskować o spadku średniego poziomu umiejętności polskich gimnazjalistów mierzonych testem matematyczno-przyrodniczym od roku 2002 do roku 2005. W latach 2006-2008 obserwowalny jest nieznaczny trend wzrostowy i w latach 2009-2010 kolejny nieznaczny trend spadkowy. Obydwa trendy są niewielkie. Dlatego też przy analizowaniu tych zmian wszelkie interpretacje, które nasuwają się intuicyjnie, powinniśmy traktować ze szczególną ostrożnością. Zagadnienie to jest przedmiotem dalszych badań w Pracowni Analiz Osiągnięć Uczniów Instytutu Badań Edukacyjnych.

Zrównane wyniki egzaminacyjne w szkolnej diagnozie

Wyniki średnie egzaminu gimnazjalnego z poszczególnych lat, wyskalowane w ten sposób, „jak gdyby każdy rocznik rozwiązywał zadania z arkusza egzaminacyjnego z 2003 roku", zostały udostępnione w serwisie internetowym IBE w analogicznym układzie, jak to ma miejsce już od szeregu lat dla komunikowania wskaźników EWD. Dla przejrzystości i łatwości korzystania z serwisu został zachowany identyczny układ wyszukiwania szkół i możliwość bezpośredniego przechodzenia pomiędzy obydwoma systemami komunikowania EWD i wyników zrównanych. Dla przykładu spójrzmy na średnie zrównane wyniki wybranego gimnazjum z gminy Wieliczka.

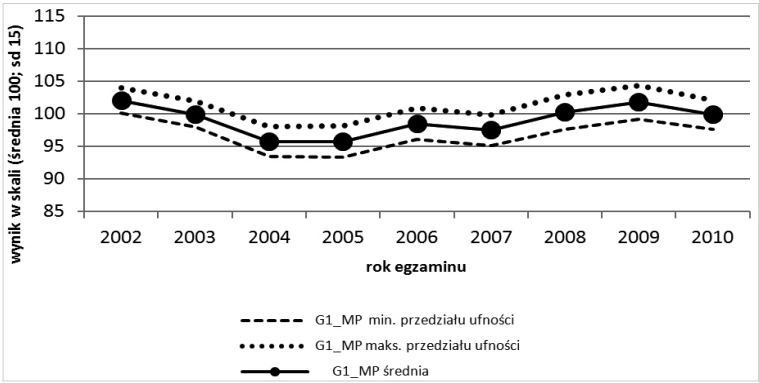


Rysunek 3. Średnie wyniki uczniów przykładowej szkoły G1 w latach 2002-2010 i 95% przedział ufności - część humanistyczna

Dla prezentowanego gimnazjum poziom osiągnięć humanistycznych oszacowany na podstawie wyników egzaminacyjnych poza dwoma latami 2004 i 2007 utrzymywał się powyżej wyniku średniego dla kraju z 2003 roku. Jeżeli dokonamy porównań ze średnią krajową w tych latach (por. rysunek 1.), to także w tej szkole zaobserwowano wynik wyższy niż średni wynik w kraju w tych latach.

We wzmiankowanym powyżej serwisie internetowym nauczyciele, rodzice, uczniowie zainteresowani osiągnięciami tej szkoły w latach 2002-2010 mogą zobaczyć prezentację wyników porównywalnych między latami także w kontekście wyników gminy, powiatu, województwa czy kraju, jak również dokonać porównania z inną wybraną szkołą. We wszystkich porównaniach należy pamiętać (co nie jest powszechną praktyką w naszych szkołach), że prezentowane wyniki obarczone są błędem pomiaru, dlatego też komunikowane są wraz 95% przedziałem ufności.

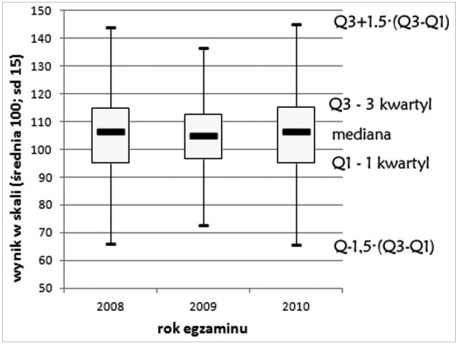
W podobny sposób przedstawiane są wyniki zrównane dla części matematyczno-przyrodniczej gimnazjalnego egzaminu. Dla tej szkoły począwszy od roku 2005 obserwujemy umiarkowany trend wzrostowy osiągnięć matematyczno-przyrodniczych.



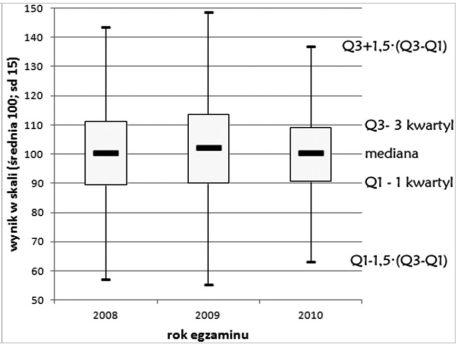
Rysunek 4. Średnie wyniki uczniów przykładowej szkoły G1 w latach 2002-2010 i 95% przedział ufności - część matematyczno-przyrodnicza

Oprócz zaprezentowanych przedstawień zrównanych wyników użytkownicy serwisu mają także sposobność skorzystania z bardziej zaawansowanych przedstawień statystycznych za pomocą wykresów pudełkowych pozwalających nie tylko poznać tendencję centralną wyrażoną poprzez medianę, ale także w przybliżeniu uzmysłowić sobie, jak różniły się rozkłady poziomu umiejętności sprawdzanych danym egzaminem. Dla obydwu typów wykresów można wyświetlić dane tabelaryczne, które dodatkowo zawierają także liczbę uczniów, dla których prezentowane są wyniki.

Ponieważ serwis posiada funkcjonalność wybierania lat, dla których wyniki mają być prezentowane, poniżej przedstawiono dla tej samej szkoły wykres dla trzech wybranych lat 2008-2010.



Rysunek 5. Mediana i porównanie rozkładów zrównanych wyników dla przykładowej szkoły G1 w latach 2008-2010 - część humanistyczna



Rysunek 6. Mediana i porównanie rozkładów wyników dla przykładowej szkoły G1 w latach 2008-2010 - część matematyczno-przyrodnicza

Tabela 4. Fragment tabeli prezentującej opis wyników egzaminacyjnych przykładowej szkoły (część humanistyczna egzaminu)

Rok	2008		2009		2010	
	GH	GMP	GH	GMP	GH	GMP
wynik średni	104,32	100,29	103,89	101,81	105,14	99,92
błąd std.	1,01	1,17	0,87	1,15	1,07	0,98
1. kwartyl	95,29	89,57	96,65	90,26	95,33	90,76
mediana	106,56	100,49	104,86	102,26	106,35	100,44
3. kwartyl	114,75	111,12	112,57	113,60	115,17	109,21
liczba uczniów	221		234		220	

Pozioma kreska w pudełkach odpowiada wartości mediany, która wraz z wartością średnią informuje o tendencji centralnej rozkładu wyników. Dla rozkładów symetrycznych (nieskośnych) mediana równa jest średniej. W kolejnych trzech latach dla wyników w części humanistycznej egzaminu wartość mediany jest wyższa niż wartość średniej. Oznacza to, że rozkłady wyników były ujemnie skośne. Wysokość pudełek na rysunku 5. równa jest różnicy pomiędzy trzecim i pierwszym kwartylem (Q3-Q1) i nazywana jest rozstępem ćwiartkowym lub rozstępem międzykwartylowym oznaczanym często IQR (*Interquartile range*). Informuje on o rozproszeniu środkowej części rozkładu wyników (w naszym przypadku egzaminacyjnych). Z definicji w tym obszarze znajduje się 50% wyników egzaminacyjnych tej szkoły. Dla wyników z lat 2008-2010 prezentowanej szkoły, jak można obliczyć z danych zawartych w tabeli 2., rozstępy międzyćwiartkowe wynosiły odpowiednio 19,46, 15,92 i 19,84, czyli najmniejszy był w 2009 roku. Im większa szerokość rozstępu ćwiartkowego, tym większe zróżnicowanie w danym roku wyników egzaminacyjnych. W tej szkole najmniejsze zróżnicowanie wyników egzaminu (części humanistycznej) w środkowej części rozkładu było w 2009 roku (15,92).

Wąsy łączą pudełko obejmujące 50% wyników z najniższym i najwyższym wynikiem w danej szkole odpowiednio z przedziału (Q1-1,5*(Q3-Q1); Q1) oraz (Q3; Q3+1,5*(Q3-Q1)). Ich długość jest równa półtorej długości przedziału międzyćwiartkowego. W pierwszym przedziale znajduje się 25% wyników egzaminacyjnych o wartościach niższych od dolnego kwartyla Q1, a w drugim przedziale 25% wyników o wartościach wyższych od górnego kwartyla Q3.

To tylko fragment informacji przygotowanych przez IBE do prezentowania w serwisie komunikowania wyników zrównanych. Wybranie dodatkowych podmiotów do prezentacji pozwala on na porównania wyników szkoły z rezultatami dla powiatu, gminy, województwa także pomiędzy szkołami w przedstawieniach graficznych i tabelarycznych.

Podsumowanie

Przedstawione wyniki średnie dla kraju, w których zminimalizowano obciążenie efektem fluktuacji trudności arkuszy egzaminacyjnych, będą przede wszystkim przydatne do analiz długoterminowych osiągnięć gimnazjalistów w całym okresie obowiązywania podstawy programowej, która w 2011 roku przeszła już do historii. Mam nadzieję, że będą one przydatne nie tylko osobom odpowiedzialnym za

kreowanie krajowej polityki oświatowej, ale także w diagnozach, których wyniki pozwolą na lepsze projektowanie działalności dydaktycznej w skali województwa czy powiatu. To może być także, choć w mniejszym stopniu, informacja przydatna dla szkoły. Dlatego też serwis umożliwiający dostęp do zrównanych wyników średnich dla szkoły został przygotowany tak, aby był spójny ze znanym już szkołom systemem komunikowania wskaźników EWD.

To właśnie przede wszystkim szkoła jest środowiskiem, w którym uczniowie nabywają kluczowych kompetencji. Diagnoza wsparta analizą wyników egzaminacyjnych wszystkich roczników, które opuściły szkołę, będzie zapewne razem z rezultatem edukacyjnej wartości dodanej (EWD) cenną informacją dla „małej polityki edukacyjnej" - przydatną do planowania strategicznego w kontekście wszystkich dostępnych informacji (również tych historycznych z egzaminu). Planowania niewątpliwie już dzisiaj także z uwzględnieniem niespotykanego wcześniej (nawet w czasach pierwszego egzaminu gimnazjalnego) tempa zmian w środowisku uczenia się, które już znacznie przekroczyło granice tradycyjnej szkoły, zajmując obszary dostępne przez Sieć między innymi e-learning, b-learning (*blended learning*) m-learning (*mobile learning*), projekty wykonywane w pozaszkolnym środowisku i w Sieci (*WebQuest*).

Opracowana przez IBE i zastosowana podczas prezentowanych badań metodologia zrównywania to zainicjowanie działań, które być może zaowocują przygotowaniem projektu wdrożenia w systemie polskich egzaminów zewnętrznych systematycznego zrównywania wyników na etapie tworzenia narzędzi i przeprowadzenia egzaminu.

Bibliografia:

1. Davier A.A, Holland P.W., Thayer D. T., 2004, *The Kernel Method of Test Equating*, Springer-Verlag, New York.
2. Niemierko B., *Zrównanie wyników sprawdzianu 2004 do wyników sprawdzianu 2003* [w:] Niemierko B., Szaleniec H. (red.), *Diagnostyka edukacyjna. Standardy wymagań i normy testowe w diagnostyce edukacyjnej*, PTDE, Kraków 2004.
3. Pokropek A., *Zrównywanie wyników egzaminów zewnętrznych w kontekście międzynarodowym* [w:] Niemierko B., Szmigel M.K. (red.), *Ewaluacja w edukacji: koncepcje, metody, perspektywy*, PTDE, Kraków 2011.
4. Saida Ch., Hattori T., 2008, *Post-hoc IRT equating of previously administered English tests for comparison of test scores*, Language Testing 25 (2): 187-210, Sage.
5. Szaleniec H., *Wykorzystanie probabilistycznych modeli zadania testowego do zrównywana wyników sprawdzianu 2003-2005 i budowania banku zadań* [w:] Niemierko B., Szyling G. (red.), *Holistyczne i analityczne metody diagnostyki edukacyjnej, perspektywy informatyczne egzaminów szkolnych,* Fundacja Rozwoju Uniwersytetu Gdańskiego, Gdańsk 2005.
6. Szaleniec H., Grudniewska M., Kondratek B., Kulon F., Pokropek A., *Zrównanie egzaminu gimnazjalnego dla lat 2002-2010*, tekst oddany do druku w: Kwartalnik „Edukacja" NR 3 (119) 2012.