KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

IBE entuzjaści edukacji

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

RAPORT TEMATYCZNY Z BADANIA
Raport cząstkowy drugiego etapu czteroetapowego studium badawczego

ANALIZA PORÓWNAWCZA WYNIKÓW EGZAMINÓW ZEWNĘTRZNYCH – SPRAWDZIAN W SZÓSTEJ KLASIE SZKOŁY PODSTAWOWEJ I EGZAMIN GIMNAZJALNY

Autorzy:

Henryk Szaleniec

Magdalena Grudniewska

Bartosz Kondratek

Filip Kulon

Artur Pokropek

Ewa Stożek

Mateusz Żółtak

Konsultacje merytoryczne:
Cees Glas, *Universiteit Twente*

Recenzje:
Dorota Węziak-Białowolska, *European Commission-Joint Research Centre, Institute for the Protection and Security of the Citizen, Econometrics and Applied Statistics Unit*
Joanna Tomkowicz, *CTB/McGraw-Hill Education*

Wydawca:
Instytut Badań Edukacyjnych
ul. Górczewska 8
01-180 Warszawa
tel. (22) 241 71 00; www.ibe.edu.pl

Publikacja została wydrukowana na papierze ekologicznym.

Publikacja opracowana w ramach projektu systemowego*: Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego,* współfinansowanego przez Unię Europejską ze środków Europejskiego Funduszu Społecznego, realizowanego przez Instytut Badań Edukacyjnych.

ISBN wersja .pdf - 978-83-61693-33-8

Egzemplarz bezpłatny

Streszczenie

Badania zrównujące wyniki egzaminacyjne prowadzone przez IBE zaplanowane są na kilka lat i obejmują kolejno egzamin gimnazjalny, sprawdzian na zakończenie szkoły podstawowej i wybrane przedmioty na poziomie maturalnym. Przeprowadzone w 2012 roku badania dotyczyły głównie sprawdzianu i ich podstawowym celem było zrównanie wyników sprawdzianu przeprowadzonego w latach 2002-2011 oraz ich przedstawienie w skali standardowej o średniej 100 i odchyleniu standardowym 15 zakotwiczonej do roku bazowego 2004.

Ponadto kontynuowane były badania zrównujące wyniki egzaminu gimnazjalnego w części humanistycznej i części matematyczno-przyrodniczej mające na celu poszerzenie zakresu porównywalnych między latami wyników zrównanych o rok 2011.

W poszczególnych rozdziałach raportu przedstawiono główne ustalenia metodologiczne, wyniki zrównania sprawdzianu dla lat 2002-2011, rezultaty zrównania wyników egzaminu gimnazjalnego za okres 2002-2011 oraz analizy z wykorzystaniem już zrównanych wyników zarówno dla sprawdzianu, jak i dla egzaminu gimnazjalnego. Szczególną wartość przedstawiają obszerne aneksy zawierające zestawienia ilustrujące rezultaty przeprowadzonych badań oraz informacja na temat dostępu do porównywalnych pomiędzy latami wyników średnich dla szkół, województw, powiatów oraz gmin prezentowana na stronie http://pwe.ibe.edu.pl/.

Streszczenie angielskie

ABSTRACT

Equating studies of examination results carried out by the Student Performance Analysis Section of the Educational Research Institute IBE are planned for several years and will include, in sequence, the lower-secondary school (gimnazjum) examination, primary school examination and selected subjects on the matriculation level (matura). Main purpose of studies conducted in 2012 concerned primary school examination conducted in 2002-2011 and to present changes of students' ability level on a standard scale anchored at the base year 2004 with mean of 100 and standard deviation of 15.

Equating studies of the lower-secondary school examination in the humanities part and the mathematics and science part were also continued, with the aim of expanding the range of compared year-to-year equated results by the year 2011.

The respective chapters of the report present information on both the main methodological determinations, the results of equating the primary school examination for 2002-2011 and of equating the lower-secondary school examination for 2002-2011. Additional contextual analyses utilizing the equated scores for both primary and lower-secondary school examinations are also reported. The report includes extensive appendices with tables illustrating the results of conducted studies and information on access to equated results for schools, voivodeships, districts and gminas presented on http://pwe.ibe.edu.pl/.

Opis zawartości raportu

Niniejszy raport dotyczy drugiego etapu czteroetapowego studium badawczego. Głównym celem tego etapu było zrównanie wyników sprawdzianu w szóstej klasie szkoły podstawowej przeprowadzonego w latach 2002-2011 z zastosowaniem arkuszy standardowych (arkusze dla uczniów bez dysfunkcji i uczniów z dysleksją rozwojową) oraz kontynuacja bieżącego zrównywania wyników egzaminu gimnazjalnego (2011-2012). Pierwszy etap studium obejmował zrównywanie wyników egzaminu gimnazjalnego w części humanistycznej i matematyczno-przyrodniczej z lat 2002 – 2010. Raport z pierwszego etapu badań został przedstawiony w czwartym kwartale 2011 roku[1]. Raport końcowy obejmujący wyniki całego cyklu badawczego zostanie przygotowany do końca pierwszego kwartału 2015 roku.

Pierwszy rozdział będący wstępem do raportu przedstawia charakterystykę sprawdzianu w szóstej klasie szkoły podstawowej przeprowadzonego w latach 2002-2012. Na początku rozdziału przedstawiono ogólne informacje dotyczące struktury arkusza egzaminacyjnego, zasad przygotowywania arkusza i organizacji oceniania. W dalszej kolejności przytoczono rozwiązania stosowane w zakresie komunikowania wyników. Trzecia część to opis właściwości psychometrycznych arkuszy egzaminacyjnych zastosowanych w latach 2002-2012. Szczegółowa charakterystyka poszczególnych zadań egzaminacyjnych załączona jest w Aneksie A – *Psychometryczne właściwości zadań egzaminacyjnych*. Rozdział kończy podsumowanie, w którym podkreślono, że przez wszystkie lata sprawdzian bazował na tym samym konstrukcie – obejmował umiejętności określone w *Standardach wymagań egzaminacyjnych* wyprowadzonych z tej samej *Podstawy programowej kształcenia ogólnego dla pierwszego i drugiego etapu edukacyjnego* a także arkusze egzaminacyjne budowane były według takich samych specyfikacji i procedur. Ponadto biorąc pod uwagę stosunkowo małe fluktuacje wskaźnika rzetelności arkuszy egzaminacyjnych (0,82-0,86) można przyjąć, że w latach 2002-2012 zachowany był konieczny do zrównywania stopień stabilności mierzonego konstruktu oraz precyzji z jaką narzędzie mierzyło badane na sprawdzianie umiejętności.

W rozdziale drugim została przedstawiona koncepcja przeprowadzonego zrównywania wyników sprawdzianu z lat 2002-2011 i bieżącego zrównania egzaminu gimnazjalnego 2011 do roku bazowego (2003). Na wstępie określono cele badania i etapy procesu badawczego. W dalszej kolejności zdefiniowane zostało pojęcie zrównywania wyników, przybliżono zastosowany w badaniu plan zbierania danych, przedstawiono argumentację wyboru 2004 roku jako bazowego dla sprawdzianu stanowiącego punkt odniesienia dla wyników po zrównaniu. Następnie opisano sposób konstrukcji wykorzystanych narzędzi badawczych. W dalszej kolejności przedstawiono opis doboru próby uczniów do sesji zrównującej wyniki sprawdzianów z lat 2002-2011 i próby do zrównania bieżącego wyników egzaminu gimnazjalnego z 2011 roku do roku bazowego 2003.

Rozdział trzeci zatytułowany *Realizacja badania* zawiera podsumowanie najważniejszych etapów przeprowadzenia badania. Rozdział ten obejmuje opis rekrutacji szkół do próby badawczej w podziale na sprawdzian w szóstej klasie szkoły podstawowej i egzamin gimnazjalny, krótkie przedstawienie organizacji szkolenia ankieterów, dostarcza informacji na temat przeprowadzenia badań w szkołach,

[1] Analiza porównawcza wyników egzaminów zewnętrznych – gimnazjum, Warszawa, 2012.

a także przybliża organizację oceniania zadań otwartych występujących w zeszytach kotwiczących. Przedstawione są też założone i osiągnięte progi realizacji badania. Warto podkreślić, że we wszystkich badanych szkołach udało się osiągnąć wymagany próg 85%. W ostatniej części rozdziału przedstawiono także zasady kontroli jakości punktowania zadań otwartych z zastosowaniem podwójnego oceniania minimum 10 procent losowo wybranych prac. Dla podwójnego oceniania zamieszczone zostały tabele oszacowanych wskaźników zgodności *kappa*. Zgodnie z przyjętymi w koncepcji założeniami podczas organizacji badań z należytą starannością zapewniono dla uczniów takie same warunki, jak podczas właściwej sesji egzaminacyjnej. Potraktowanie badań zrównujących, jako egzaminu próbnego i dostarczenie po badaniach wyników dla każdego ucznia jeszcze przed egzaminem właściwym miało szczególne znaczenie dla zwiększenia poziomu motywacji rozwiązywania testów przez poszczególnych uczniów.

Statystyczna koncepcja zrównania stanowi treść czwartego rozdziału prezentowanego raportu. W pierwszej części zarysowana została ogólna typologia metod zrównywania wyników ze szczególnym naciskiem na metody wykorzystujące modelowanie IRT. Następnie omówiono zastosowanie metody łącznej kalibracji do przedstawionego w rozdziale drugim planu zrównywania wyników nierównoważnych grup z testem kotwiczącym (*nonequivalent groups with anchor test design*, NEAT). Omówiono, w jaki sposób do zgromadzonych danych został dopasowany model IRT oraz jak korzystając z wyrażonych na wspólnej skali parametrów modelu IRT przeprowadzono zrównanie wyników obserwowanych. W tej części rozdziału poruszone zostały także zagadnienia związane z szacowaniem błędu zrównywania i szacowaniem rozkładu umiejętności z wykorzystaniem tzw. *plausible values*.

Dobór zadań do zrównywania to kluczowe zadanie w całym procesie. W rozdziale piątym znajduje się opis szeregu działań w tym zakresie zarówno dla sprawdzianu w szóstej klasie szkoły podstawowej, który w 2012 roku stanowił główną część sesji zrównującej, jak i dla egzaminu gimnazjalnego, dla którego w tej sesji kontynuowano zrównanie dla wyników z 2011 roku. Na etapie pracy nad narzędziami do badań i po przeprowadzeniu badań w 2011 roku zadania wszystkich edycji egzaminu gimnazjalnego z lat 2002-2010 poddano kompleksowej analizie psychometrycznej z wykorzystaniem narzędzi klasycznej teorii testów, jak i IRT. Takie samo rozwiązanie przyjęto w 2012 roku dla zadań ze sprawdzianu obejmującego lata 2002-2011 oraz dla egzaminu gimnazjalnego z 2011 roku w części humanistycznej i matematyczno-przyrodniczej. Szczegółowe wyniki tej analizy zebrano w Aneksie A zatytułowanym *Psychometryczne właściwości zadań egzaminacyjnych.* Następny etap wyboru zadań, które finalnie zostały wykorzystane do zrównania wyników egzaminacyjnych, nastąpił po przeprowadzeniu badania zrównującego, które dostarczyło danych pozwalających na zakotwiczenie wyników z różnych lat na wspólnej skali. Szczególne znaczenie dla analizy wrażliwości rezultatu zrównywania na dobór zadań miało zastosowanie opisanych w tym rozdziale metod symulacyjnych.

Rozdział szósty to najważniejsza część raportu. W tym rozdziale zostały przedstawione wyniki zrównywania sprawdzianu z lat 2002-2011 oraz zrównanie egzaminu gimnazjalnego rozszerzone, w stosunku do wyników przedstawianych wcześniej (w raporcie ze zrównywania opracowanym w 2011 r. oraz opublikowanych wynikach), o egzamin gimnazjalny z roku 2011 w podziale na część humanistyczną i matematyczno-przyrodniczą. Włączenie danych z kolejnego roku egzaminu gimnazjalnego dostarczyło nie tylko informacji o tym, jak wyniki z sesji egzaminacyjnej z roku 2011 mają się do wcześniejszych edycji egzaminów, ale również uzyskano nowe dane o zależności pomiędzy egzaminami gimnazjalnymi za cały okres od 2002 do 2011 roku. Ponieważ całą procedurę statystyczną zrównywania egzaminu gimnazjalnego przeprowadzono ponownie dla wszystkich lat, w rozdziale tym przedstawiono zrównane wyniki dla egzaminu gimnazjalnego dla wszystkich edycji. Zarówno dla egzaminu gimnazjalnego, jak i dla sprawdzianu przyjęto taką samą strukturę prezentacji wyników badań. Najpierw zostały omówione różnice pomiędzy rozkładami obserwowanych wyników sumarycznych na oryginalnych skalach oraz na skalach

zrównanych, a następnie przedstawiono tabele pozwalające na przeliczenie wyniku egzaminu z danego roku na wynik (odpowiednio dla egzaminu gimnazjalnego i sprawdzianu) w ustalonym roku bazowym. W szczególności pozwala to na ocenę fluktuacji trudności arkuszy egzaminacyjnych w kolejnych latach. W następnej kolejności zaprezentowane zostały wyniki średnie z lat 2002-2011 na skali zmiennej ukrytej, przekształconej do skali standardowej o średniej 100 oraz odchyleniu standardowym 15.

Rozdział siódmy to analizy wyników egzaminacyjnych na zrównanych wynikach przedstawionych na wspólnej skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej do wyników sprawdzianu z roku 2004, a w przypadku egzaminu gimnazjalnego do roku 2003. Rozdział ten ilustruje zróżnicowanie rezultatów egzaminacyjnych ze względu na płeć, lokalizację szkoły oraz typ szkoły (prywatna a publiczna). Przedstawiono także analizę zróżnicowania międzyszkolnego w zależności od lokalizacji i typu szkoły w trakcie 11 lat. Zróżnicowanie średnich rezultatów z egzaminu gimnazjalnego w poszczególnych latach (2002-2011) przedstawiają mapy zawarte w Aneksie B. W Aneksie B1 przedstawiono rezultaty z podziałem na województwa, w Aneksie B2 w podziale na powiaty, a w Aneksie B3 przedstawiono zróżnicowanie zrównanych wyników sprawdzianu w latach 2002-2011 w podziale na powiaty w obrębie poszczególnych województw. Wszystkie wyniki prezentowane są na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej dla sprawdzianu w roku 2004 przy wykorzystaniu metodologii *plausible values.*

W celu udostępnienia porównywalnych wyników egzaminu gimnazjalnego (PWE) szerokiemu gronu odbiorców, zespół badawczy przygotował specjalny serwis internetowy dostępny pod adresem http://pwe.ibe.edu.pl/. Głównym elementem prezentacji są średnie wyniki (łącznie z przedziałem ufności) wybranej szkoły w zadanym przez użytkownika serwisu okresie z przedziału lat od 2002 do 2011 roku. Można je osadzić w kontekście wyników innych szkół, gminy, powiatu, województwa, jak również wyników ogólnopolskich. Serwis pozwala także na dokonywanie porównań pomiędzy szkołami i jednostkami samorządu terytorialnego. Zawiera narzędzia do wizualizacji wyników, tak by możliwe było przeprowadzenie podstawowych analiz bezpośrednio w serwisie. Możliwe jest również pobranie wyników zrównanych dla interesującej użytkownika grupy szkół i jednostek samorządu terytorialnego w postaci zbioru danych w celu wykorzystania ich do dalszych analiz. W rozdziale dziewiątym zatytułowanym *Komunikowanie porównywalnych wyników egzaminacyjnych – egzamin gimnazjalny* przedstawiono opis serwisu, wyjaśnienia, w jaki sposób prezentowane są wyniki, przykłady dostępnych tabel i wykresów. Przygotowanie serwisu wymagało nie tylko wyczyszczenia bazy danych wyników egzaminacyjnych, ale także bazy szkół, których uczniowie przystępowali do egzaminów w kolejnych latach. Ostatnia część tego rozdziału to opis działań, które przeprowadzono w celu uporządkowania i zweryfikowania bazy danych adresowych szkół (adres szkoły, kod TERYT gminy) oraz działań prowadzących do połączenia bazy porównywalnych między latami wyników egzaminacyjnych z bazą adresową szkół w 11-letnim okresie funkcjonowania egzaminów.

W rozdziale dziewiątym przedstawione zostały rekomendacje wynikające z doświadczeń zespołu badawczego oraz analiz uzyskanych w trakcie zrównywania wyników sprawdzianu i egzaminów gimnazjalnych. Rezultaty pierwszego i drugiego etapu badań zrównujących, dostarczają propozycji rozwiązań, które mogą być wykorzystane do wprowadzenia w systemie egzaminacyjnym, tak aby możliwe było komunikowanie uczniom nie tylko wyników obserwowalnych, ale także wyników porównywalnych pomiędzy latami.

Raport został uzupełniony obszernymi aneksami, które zawierają szereg szczegółowych informacji, które są cennym materiałem do wykorzystania w dodatkowych analizach.

Spis treści

1. Wstęp

Arkusze egzaminacyjne zastosowane na sprawdzianie w szóstej klasie szkoły podstawowej, podobnie jak w przypadku pozostałych egzaminów, różnią się w kolejnych latach pomiędzy sobą poziomem trudności, pomimo że były konstruowane według takiej samej specyfikacji ustalonej przez Centralną Komisję Egzaminacyjną. Nawet najbardziej doskonałe procedury tworzenia arkuszy egzaminacyjnych nie zapewnią w pełni porównywalnych parametrów psychometrycznych. Aby móc oddzielić od siebie faktyczny efekt zmian w poziomie trudności arkuszy egzaminacyjnych od możliwych zmian w poziomie umiejętności uczniów między latami konieczne jest przeprowadzenie zrównania wyników.

Sprawdzian w szóstej klasie szkoły podstawowej pomyślany był jako test diagnostyczny na zakończenie drugiego etapu kształcenia. We wstępie do informatora o sprawdzianie opracowanego przez CKE i okręgowe komisje egzaminacyjne zapisano cele sprawdzianu w następujący sposób.

Jego celem (sprawdzianu) jest sprawdzenie opanowania umiejętności niezbędnych na wyższym etapie kształcenia (w gimnazjum) i przydatnych w życiu. Wyniki sprawdzianu, analizowane z uwzględnieniem ocen szkolnych oraz kontekstów kształcenia, pozwalają na pełniejsze diagnozowanie osiągnięć uczniów i, tym samym, ułatwiają opracowanie indywidualnych zaleceń dotyczących ich dalszej edukacji. Nauczycielom i szkołom takie analizy mogą pomóc w doskonaleniu pracy, a decydentom – w prowadzeniu efektywnej polityki oświatowej[2].

Aby na podstawie wyników sprawdzianu możliwe było wnioskowanie o poziomie osiągnięć szkolnych uczniów kończących szkołę podstawową i rozpoczynających edukację w gimnazjum w kolejnych latach, niezbędne jest wprowadzenie takich mechanizmów, które pozwolą na wyrugowanie z końcowego wyniku losowych wahań trudności między arkuszami zastosowanymi do przeprowadzenia sprawdzianu w kolejnych sesjach egzaminacyjnych.

Do chwili obecnej w naszym kraju nie wprowadzono żadnego systemowego narzędzia umożliwiającego otrzymywanie porównywalnych między latami wyników egzaminacyjnych. Badania realizowane przez Pracownię Analiz Osiągnięć Uczniów IBE mające charakter naukowy dostarczają w tym zakresie istotnych informacji już dla kolejnego egzaminu[3]. Porównywalne między latami wyniki egzaminu gimnazjalnego są dostępne na stronie internetowej Instytutu pod adresem http://pwe.ibe.edu.pl/. Wyniki sprawdzianu zostaną udostępnione w sierpniu 2013 r. Korzystając z tych wyników do dalszych analiz warto pamiętać, że pomimo iż dołożono wszelkich starań, aby warunki, w jakich uczniowie w nich uczestniczyli, były jak najbardziej zbliżone do warunków egzaminacyjnych, to jednak nie był to egzamin, od którego zależą losy ucznia, co mogło wpłynąć na uzyskane wyniki.

---

[2] http://www.cke.edu.pl/files/file/Informatory/informator_od_2009_4.pdf

[2] W 2012 roku opublikowane zostały przez IBE porównywalne między latami wyniki egzaminu gimnazjalnego w części humanistycznej i części matematyczno-przyrodniczej.

Porównywalne między latami wyniki sprawdzianu (niezależne od wahań trudności arkuszy egzaminacyjnych) łącznie z porównywalnymi wynikami gimnazjalnymi to cenne narzędzie do monitorowania poziomu edukacji w skali całego kraju. Nie wypełnią one jednak luki w systemie egzaminacyjnym. Rezultatem mającym o wiele większe znaczenie jest to, że podczas badań przetestowano metodologię i praktyczne rozwiązania, które mogą być przydatne do wprowadzenia w polskim systemie egzaminacyjnym do konstrukcji arkuszy egzaminacyjnych, które pozwoliłyby na zrównywanie wyników w trakcie bieżących sesji egzaminacyjnych.

Przeprowadzone w 2012 roku badania zrównujące wyniki sprawdzianu z lat 2002-2011 i wyniki egzaminu gimnazjalnego w części humanistycznej i matematyczno-przyrodniczej z roku 2011 to druga część czteroetapowego studium obejmującego:

1. Zrównanie wyników egzaminu gimnazjalnego w części humanistycznej i matematyczno-przyrodniczej przeprowadzonego w latach 2002-2010 – badania przeprowadzone w 2011 roku.

2. Zrównanie wyników sprawdzianu z lat 2002-2011 – badania przeprowadzone w 2012 roku.

3. Zrównanie wyników egzaminu maturalnego z matematyki przeprowadzanego w latach 2010-2012 w sesjach majowych i poprawkowych – badania prowadzone w 2013 roku.

4. Zrównanie wyników egzaminu maturalnego z języka polskiego i języka angielskiego przeprowadzanego w latach 2010-2013 w sesjach majowych i poprawkowych – badania planowane na rok 2014.

Począwszy od drugiego etapu (rok 2012) prowadzone jest bieżące zrównywanie wyników egzaminów włączonych do studium w poprzednich latach pozwalające na rozszerzenie zrównanych wyników o kolejne lata. Dlatego też w niniejszym raporcie przedstawiono rezultaty zrównania wyników sprawdzianu z lat 2002-2011 oraz porównywalne wyniki egzaminacyjne dla egzaminu gimnazjalnego uzupełnione o rok 2011.

1.1. Charakterystyka sprawdzianu

1.1.1. Koncepcja egzaminu w szkole podstawowej

Sprawdzian w szóstej klasie szkoły podstawowej jest egzaminem powszechnym i obowiązkowym. Przystąpienie do sprawdzianu warunkuje uzyskanie świadectwa ukończenia szkoły podstawowej, ale dla promocji nie ma znaczenia liczba uzyskanych przez ucznia punktów. Jak już wspomniano wcześniej, celem tego egzaminu zewnętrznego jest sprawdzenie opanowania umiejętności niezbędnych na wyższym etapie kształcenia (gimnazjum), co wyznacza funkcję diagnostyczną sprawdzianu. Zasadniczo wyniki sprawdzianu nie służą do rekrutacji do gimnazjum, które jest szkołą obowiązkową i rejonową. Jednak w przypadku ubiegania się uczniów o przyjęcie do szkoły spoza rejonu wyniki sprawdzianu mogą być i zwykle są kryterium rekrutacyjnym.

Po raz pierwszy sprawdzian przeprowadzono w 2002 roku. Zasady i tryb przeprowadzania sprawdzianu określa rozporządzenie Ministra Edukacji Narodowej[4]. Za organizację i przeprowadzenie sprawdzianu w szkołach podstawowych na terenie całego kraju są odpowiedzialne okręgowe komisje egzaminacyjne (OKE), których działania koordynuje Centralna Komisja Egzaminacyjna (CKE). Sprawdzian przeprowadzany jest w macierzystych szkołach uczniów. W 2003 roku po pierwszych doświadczeniach w organizacji krajowych egzaminów wprowadzono możliwość nadzoru egzaminu przez ekspertów zewnętrznych powoływanych przez dyrektora OKE i obserwatorów, którymi mogli być nauczyciele z innej szkoły.

Na sprawdzianie badany jest poziom opanowania umiejętności określonych w *Standardach wymagań egzaminacyjnych będących podstawą przeprowadzania sprawdzianu w ostatnim roku nauki w szkole podstawowej.* Zadania zawarte w zestawach egzaminacyjnych sprawdzianu nie wykraczają poza treści zawarte w *Podstawie programowej kształcenia ogólnego*.

Standardy wymagań egzaminacyjnych obejmują umiejętności z następujących obszarów:

1. czytanie,
2. pisanie,
3. rozumowanie,
4. korzystanie z informacji,
5. wykorzystanie wiedzy w praktyce.

Sprawdzian ma charakter **ponadprzedmiotowy**, co oznacza, że np. czytanie obejmuje nie tylko umiejętności odczytywania tekstów literackich, ale także wykresów, map itp., a zadania są tak skonstruowane, że sprawdzają umiejętności kształcone w obrębie różnych przedmiotów. Twórcy koncepcji sprawdzianu zakładali, że odejście od ściśle przedmiotowego podziału materiału przeniesie akcent w procesie dydaktycznym z materiału nauczania na kształcenie ogólnych umiejętności uczniów na różnych treściach i w rozmaity sposób. Pierwotnie zakładano, że elementem integrującym zadania sprawdzianu będzie motyw przewodni. Od tej zasady odstąpiono dopiero w 2010 roku. Tabela 1.1 pokazuje, jakie motywy przewodnie zastosowano w arkuszach standardowych[5] w latach 2002-2009.

Tabela 1.1. Motyw przewodni sprawdzianu w latach 2002-2009

Rok sprawdzianu	Tytuł arkusza standardowego
2002	Pory roku
2003	Przed telewizorem
2004	Chleb
2005	W wodzie
2006	Pszczoły i miody

[4] Rozporządzenie Ministra Edukacji Narodowej z dnia 21 marca 2001 r. w sprawie warunków i sposobu oceniania, klasyfikowania i promowania uczniów i słuchaczy oraz przeprowadzania egzaminów i sprawdzianów w szkołach publicznych (DzU Nr 29, poz. 323, zm.: DzU Nr 128, poz. 1419, DzU z 2002 r. Nr 46, poz. 433) z późniejszymi zmianami.

[5] Arkusze egzaminacyjne dla uczniów bez dysfunkcji i uczniów z dysleksją rozwojową.

2007	W szkole
2008	Jasne jak słońce
2009	O zwierzętach

Za rozwiązanie zadań na sprawdzianie uczeń może uzyskać maksymalnie 40 punktów. Zestaw zadań zawiera 20 zadań zamkniętych wielokrotnego wyboru – uczeń wybiera jedną poprawną odpowiedź z czterech możliwości. Zadania otwarte to zadania krótkiej i rozszerzonej odpowiedzi. Umiejętność pisania jest sprawdzana poprzez dłuższą wypowiedź, np. opowiadanie na zadany temat. Egzamin trwa 60 minut.

Przydział punktów za poszczególne grupy umiejętności pokazuje Tabela 1.2. Początkowo 12 punktów w teście (30% możliwych do uzyskania punktów) przeznaczono na pisanie i tylko 2 punkty na sprawdzenie umiejętności korzystania z informacji. W 2005 roku zmieniono te proporcje: na zadania z pisania i czytania przeznaczano po 10 punktów (25% możliwych do uzyskania punktów), a 4 punkty na zadania sprawdzające umiejętność korzystania z informacji. Za zadania sprawdzające umiejętności rozumowania i wykorzystania wiedzy w praktyce uczeń mógł uzyskać 8 punktów (20% możliwych).

Tabela 1.2. Przydział punktów (długość skali) dla poszczególnych standardów wymagań egzaminacyjnych na sprawdzianie

Umiejętności	Lata 2002-2004		Lata 2005-2012	
	maksymalna liczba punktów	udział procentowy	maksymalna liczba punktów	udział procentowy
Czytanie	10	25%	10	25%
Pisanie	12	30%	10	25%
Rozumowanie	8	20%	8	20%
Korzystanie z informacji	2	5%	4	10%
Wykorzystywanie wiedzy w praktyce	8	20%	8	20%

Jak już wcześniej wspomniano, podstawę do przeprowadzenia sprawdzianu stanowią standardy wymagań egzaminacyjnych, które nawiązują do *Podstawy programowej kształcenia ogólnego dla pierwszego i drugiego etapu edukacyjnego*. Są one równocześnie układem odniesienia do zapewnienia trafności sprawdzianu. W latach 2002-2010 przygotowanie arkuszy egzaminacyjnych (w tym propozycji zadań, kluczy do zadań zamkniętych, kryteriów oceny zadań otwartych, modeli odpowiedzi/schematów punktowania, recenzji) odbywało się w okręgowych komisjach egzaminacyjnych według jednolitych zasad i procedur tworzenia arkuszy sprawdzianu ustalonych przez Centralną Komisję Egzaminacyjną. W okręgowych komisjach egzaminacyjnych za przygotowanie arkusza począwszy od planu i kartoteki sprawdzianu, poprzez przygotowanie autorskich zadań, próbne zastosowanie wiązek zadań, recenzje nauczycielskie i akademickie aż do składu gotowego do przekazania do CKE arkusza egzaminacyjnego odpowiedzialny był koordynator sprawdzianu. Koordynatorem najczęściej był kierownik pracowni sprawdzianu w danej okręgowej komisji egzaminacyjnej.

Finalny arkusz egzaminacyjny rekomendowany przez recenzentów i koordynatora sprawdzianu był zatwierdzany przez dyrektora OKE. Zatwierdzenie arkusza oznaczało przyjęcie odpowiedzialności przez dyrektora OKE za jego poprawność merytoryczną, konstrukcyjną i edytorską. Centralna Komisja

Egzaminacyjna podejmowała decyzję o wyborze arkuszy na poszczególne sesje egzaminacyjne i poddawała je kolejnym recenzjom oraz dalszemu doskonaleniu przed skierowaniem do druku. W wyniku tych działań ostateczną odpowiedzialność za jakość arkusza egzaminacyjnego przejmowała CKE. Od 2010 roku zostały zmienione zasady organizacyjne tworzenia arkuszy egzaminacyjnych sprawdzianu. Arkusze przygotowywane są przez ogólnopolski zespół złożony z przedstawicieli poszczególnych okręgowych komisji, którego prace koordynuje dyrektor OKE we Wrocławiu.

Zmieniał się także sposób oceniania odpowiedzi na zadania otwarte, a szczególnie zadania otwarte rozszerzonej odpowiedzi. W pierwszych latach konsekwentnie stosowano kryterialne ocenianie otwartych zadań matematycznych i polonistycznych. Od 2010 roku zaczęto stopniowo odchodzić od oceniania kryterialnego w kierunku oceniania holistycznego. Warto zauważyć, że egzaminatorem sprawdzianu może być każdy nauczyciel, który ukończył kurs na egzaminatora i wpisany został do ewidencji egzaminatorów prowadzonej przez okręgowe komisje egzaminacyjne. W szczególności oznacza to, że polonista może oceniać również zadania matematyczne. Kierując się troską o jakość oceniania, komisje egzaminacyjne stopniowo zaczęły wprowadzać specjalizację: część egzaminatorów ocenia tylko otwarte zadania matematyczne, część – tylko otwarte zadania polonistyczne.

Oprócz arkusza standardowego (S-1) na każdą edycję sprawdzianu przygotowywane są dostosowane arkusze egzaminacyjne dla uczniów o specjalnych potrzebach edukacyjnych: dla uczniów słabo widzących – arkusze S-4 i S-5 ze zwiększoną czcionką, odpowiednio 16 i 24 pkt., arkusze S-6 dla uczniów niewidomych, przygotowywane w piśmie Braille'a, arkusze S-7 dla uczniów słabo słyszących i niesłyszących. Odrębny zestaw zadań jest przygotowywany dla uczniów z upośledzeniem w stopniu lekkim (arkusz S-8). Uczniowie o specjalnych potrzebach edukacyjnych mogą mieć wydłużony czas pracy o 30 minut. Dla uczniów o specjalnych potrzebach edukacyjnych dostosowano także standardy wymagań egzaminacyjnych. Każdego roku sprawdzian z wykorzystaniem arkuszy dostosowanych pisało około 2 procent zdających (średnio 1,94 procent).

Inaczej ma się sprawa z uczniami o specyficznych trudnościach w uczeniu się. Uczniowie z dysleksją rozwojową powinni spełniać określone w standardach, jednolite dla wszystkich uczniów wymagania. Dlatego podczas sprawdzianu otrzymują arkusze o takiej samej treści, jak uczniowie bez specjalnych potrzeb edukacyjnych. Jednak ze względu na trudności w pisaniu tej grupy uczniów zmodyfikowano ogólne kryteria oceniania, w szczególności w zakresie stosowania zasad ortografii i interpunkcji. Z założenia dostosowania kryteriów oceniania powinny mieć charakter kompensacyjny i zrównywać szanse na osiągnięcie porównywalnych wyników. W praktyce jednak, w szczególności w zadaniach krótkiej odpowiedzi, osłabienie kryterium poprawności ortograficznej faworyzowało uczniów z dysleksją. Uczniowie dyslektyczni mogą skorzystać z wydłużonego czasu pracy o 30 minut.

Od 2007 roku na sprawdzianie stosuje się wersje równoległe testu standardowego. Arkusze w wersji A i w wersji B różnią się kolejnością dystraktorów w zadaniach zamkniętych wielokrotnego wyboru.

Sprawdzian odbywa się na początku kwietnia, wyniki ogłaszane są pod koniec maja, a zaświadczenia przekazywane do szkół w czerwcu, na tydzień przed zakończeniem roku szkolnego.

1.2. Organizacja oceniania

Jak już sygnalizowano wcześniej sprawdzian jako egzamin zewnętrzny pełni głównie funkcję diagnostyczną. Dla realizacji tej funkcji ważne jest zapewnienie porównywalności oceniania w całym kraju. Drogą do osiągnięcia porównywalności jest stosowanie jednolitych kryteriów oceniania, jak również odpowiednia organizacja oceniania w całym kraju.

Ocenianie koordynowane jest przez koordynatora krajowego. W każdej okręgowej komisji egzaminacyjnej jedna lub dwie osoby odpowiedzialne są za merytoryczną koordynację oceniania. Na każdą sesję powoływani są przez dyrektora OKE egzaminatorzy, którzy przydzielani są do 18-20 osobowych zespołów egzaminacyjnych. Pracą takiego zespołu kieruje przewodniczący zespołu egzaminacyjnego (PZE). Egzaminatorzy oceniają prace w specjalnie przygotowanych ośrodkach egzaminacyjnych. Jedynie w pierwszych latach funkcjonowania systemu egzaminacyjnego ocenianie pod względem organizacyjnym było mieszane – część oceniania odbywała się w ośrodkach egzaminacyjnych, część w domu egzaminatora. W 2003 r. ocenianie mieszane utrzymano jedynie w OKE Łódź i OKE Poznań, a od 2004 roku uczniowskie prace egzaminacyjne oceniane są w całym kraju wyłącznie w ośrodkach egzaminacyjnych.

Koordynator OKE uczestniczy w spotkaniu koordynacyjnym CKE, organizowanym bezpośrednio po przeprowadzonym sprawdzianie. Celem tego spotkania jest uzyskanie konsensusu co do schematów punktowania poszczególnych zadań otwartych. Koordynatorzy z OKE przywożą[6] na spotkanie dobraną celowo próbę prac uczniowskich, które są oceniane na spotkaniu zgodnie z kryteriami zaproponowanymi przez autorów arkusza egzaminacyjnego. Na tej podstawie analizowane jest funkcjonowanie kryteriów, w szczególności w kontekście nietypowych rozwiązań i interpretacji tematu wypowiedzi pisemnej oraz proponowane są ewentualne korekty. Efektem prac zespołu koordynatorów OKE i CKE są uzgodnione, doprecyzowane kryteria oceniania oraz jednolite materiały pomocnicze. Są one podstawą do szkolenia przewodniczących i egzaminatorów w danej sesji tuż przed ocenianiem.

Koordynatorzy OKE przeprowadzają szkolenie przewodniczących zespołów egzaminatorów (PZE) i koordynują ocenianie prac uczniowskich w obrębie danej OKE. Do koordynacji procesu oceniania wykorzystywane są w komisjach różne rozwiązania z zastosowaniem technologii informacyjno-komunikacyjnych (TIK), w ramach których koordynatorzy mają bezpośredni kontakt synchroniczny lub asynchroniczny z PZE. Bywa to ważne przy rozstrzyganiu powstających podczas oceniania wątpliwości, szczególnie w przypadku nietypowych rozwiązań zadań egzaminacyjnych. Jeśli podczas sprawdzania egzaminator napotka rozwiązanie, które nie mieści się w ramach uzgodnionego schematu punktowania, zgłaszane jest to do koordynatora CKE, ustalane jest poszerzenie klucza odpowiedzi, o czym informowani są wszyscy koordynatorzy oceniania OKE. Należy jednak podkreślić, że w przypadku sprawdzianu niezwykle rzadko dochodzi do takich sytuacji – z reguły jednorazowe doprecyzowanie kryteriów oceniania na spotkaniu koordynatorów po przeprowadzonym sprawdzianie jest wystarczające.

Przewodniczący zespołów egzaminatorów są odpowiedzialni za szkolenie egzaminatorów i zapewnienie jakości i porównywalności oceniania w swoim zespole. Przewodniczący weryfikuje około 3 procent prac ocenionych przez swoich egzaminatorów. Ponadto w zespołach egzaminatorów powołani są weryfikatorzy (egzaminatorzy drugiego oceniania), którzy dodatkowo weryfikują 10 procent wybranych losowo prac.

1.3. Komunikowanie wyników sprawdzianu

Przetwarzanie wyników egzaminacyjnych w celu prezentacji wyników sprawdzianu przebiega kilkuetapowo.

[6] Obecnie przesyłają elektronicznie kopie prac.

1. Po przeprowadzeniu sesji egzaminacyjnej przetworzone dane po anonimizacji przekazywane są do CKE w celu archiwizacji i opracowania sprawozdania. Harmonogram i format przekazywania danych ustala CKE.

2. Wstępna informacja o wynikach sprawdzianu, przygotowywana przez CKE dla całego kraju, a przez OKE dla poszczególnych województw, jest publikowana pod koniec maja; jednocześnie szkoły otrzymują zbiorcze zestawienia wyników swoich uczniów.

3. Na 7 dni przed zakończeniem roku szkolnego okręgowe komisje egzaminacyjne przekazują do szkół imienne zaświadczenia o wynikach sprawdzianu, które wręczane są uczniom razem ze świadectwem ukończenia szkoły podstawowej.

4. Centralna Komisja Egzaminacyjna przygotowuje sprawozdanie[7] o osiągnięciach uczniów kończących szkołę podstawową w danym roku.

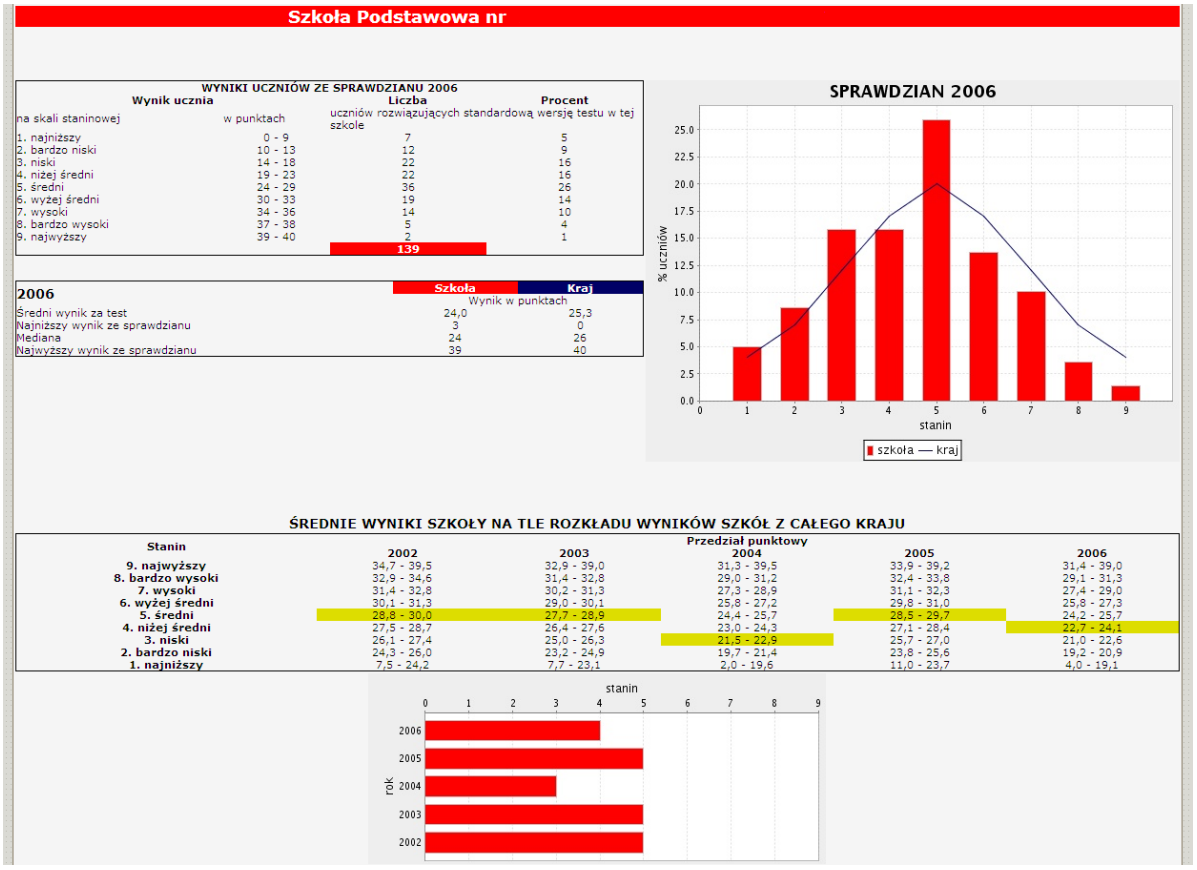
5. Okręgowe komisje egzaminacyjne przygotowują poszerzoną informację o wynikach, którą przekazują różnym odbiorcom – szkołom, kuratoriom oświaty, organom prowadzącym szkoły. Obecnie tego typu informacje udostępniane są najczęściej w dedykowanych serwisach informatycznych na stronach internetowych poszczególnych OKE.

Przez lata funkcjonowania systemu egzaminów zewnętrznych nie wypracowano jednolitego sposobu komunikowania wyników poszczególnych szkół jako publicznie dostępnej informacji. Początkowo uważano, że upublicznianie wyników szkół jest niewłaściwe, ponieważ najczęściej służy tworzeniu rankingów. W latach 2006-2008 podjęto próbę stworzenia jednolitego serwisu internetowego zawierającego wyniki wszystkich szkół podstawowych i gimnazjalnych w Polsce. Przykładową prezentację wyników dla szkoły podstawowej dla roku 2006 pokazano na Rysunku 1.1. Prezentacja zawierała informację o średnich wynikach szkoły na tle kraju, rozkład wyników uczniowskich oraz pozycję szkoły na skali staninowej[8] w perspektywie wieloletniej.

[7] Sprawozdania można znaleźć na stronie internetowej CKE, pod adresem http://www.cke.edu.pl/ (zakładka sprawdzian w klasie VI/informacje o wynikach).

[8] Skala standardowa o średniej 5 i odchyleniu standardowym 2.

Rysunek 1.1. Przykładowa prezentacja wyników szkoły podstawowej z 2006 roku na platformie Scholaris

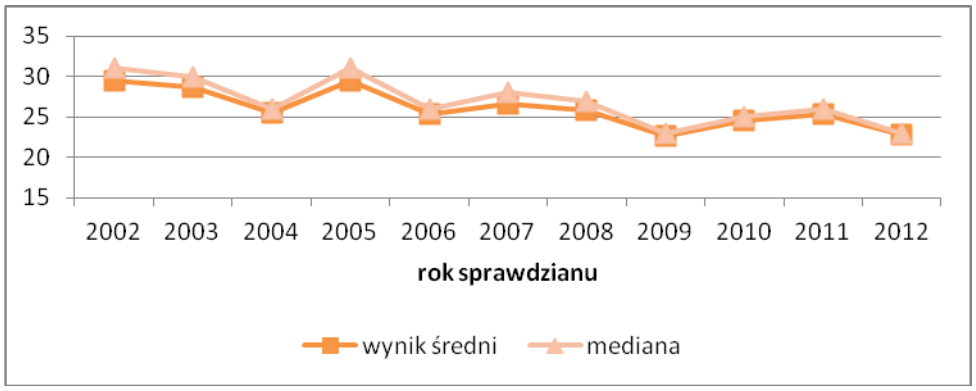


Obecnie nie istnieje analogiczny serwis internetowy. Okręgowe komisje egzaminacyjne publikują na swoich ogólnodostępnych stronach jedynie średnie wyniki szkół wraz z pozycją szkoły na skali staninowej w danym roku.

1.4. Własności psychometryczne sprawdzianów do 2012 roku

Średnie wyniki sprawdzianów, przedstawiane co roku na skali będącej sumą punktów za odpowiedzi na poszczególne zadania (wynik surowy) podlegają dość dużym fluktuacjom (Rysunek 1.2).

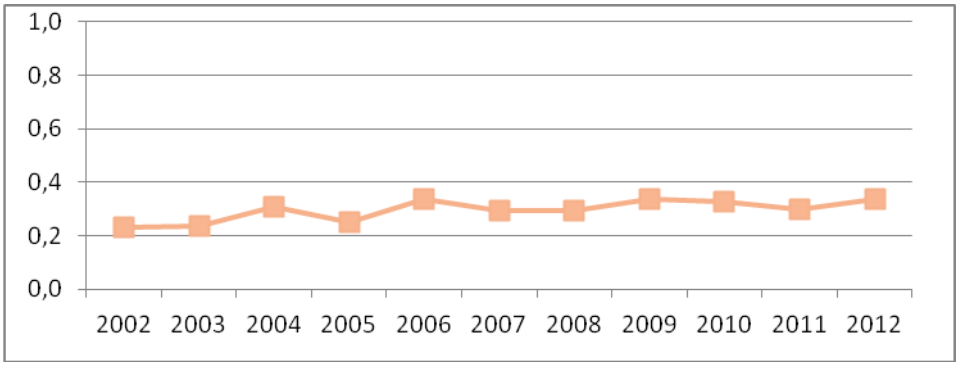
Rysunek 1.2. Średnie krajowe wyniki uczniów i mediana dla sprawdzianu od 2002 roku



Maksymalna różnica pomiędzy wynikami średnimi w 11-letnim okresie wynosi niemal 7 punktów na skali surowego wyniku egzaminacyjnego (wynik najwyższy 29,5; wynik najniższy 22,64), co stanowi 0,9 średniego odchylenia standardowego. Kilkupunktowe różnice z roku na rok np. w latach 2004, 2005, 2006 były niejednokrotnie przyczyną nieuprawnionych wniosków o spadku lub wzroście poziomu osiągnięć szóstoklasistów analizowanych na podstawie rezultatów uzyskiwanych na sprawdzianie. Na wielkość tych fluktuacji niewątpliwie ma wpływ brak odpowiedniej kontroli trudności arkuszy testowych w procesie przygotowywania testów egzaminacyjnych, jak również istotną rolę mogą mieć wahania poziomu umiejętności uczniów między latami. Rozróżnienie efektów tych dwóch czynników na obserwowalny podczas egzaminu wynik to zagadnienie, które jest przedmiotem analiz prezentowanych w kolejnych rozdziałach raportu.

O zróżnicowaniu wyników można wnioskować zarówno na podstawie odchylenia standardowego, jak i względnego współczynnika zmienności (odchylenie standardowe podzielone przez wynik średni). Współczynnik zmienności zmieniał się między latami od 23% w roku 2006 do 34% w latach 2006, 2009, 2012 (Rysunek 1.3).

Rysunek 1.3. Współczynnik zmienności średnich krajowych wyników dla sprawdzianu w latach 2002-2012



Podstawowe charakterystyki psychometryczne dla sprawdzianów z lat 2002-2012 zostały zebrane w Tabeli 1.3.

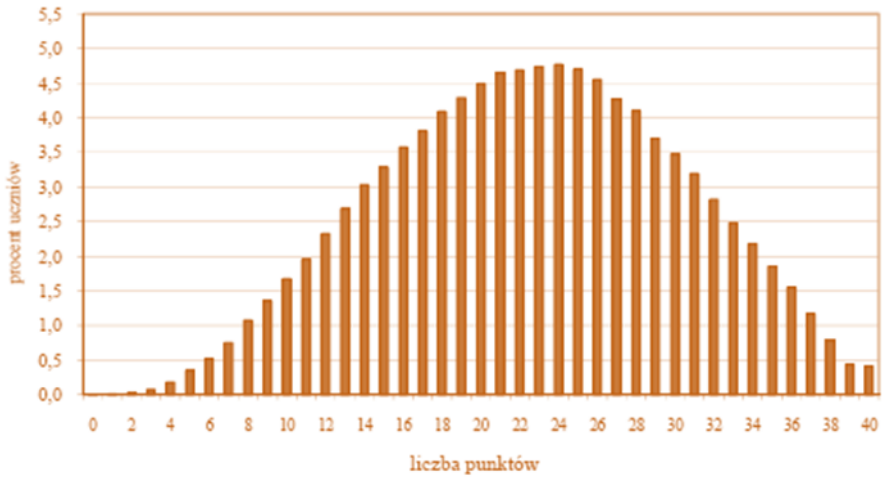
Tabela 1.3. Wybrane charakterystyki psychometryczne opisujące sprawdziany w latach 2002-2012

Rok egzaminu		2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Wynik średni w skali utworzonej z sumy punktów		29,49	28,61	25,55	29,50	25,32	26,57	25,80	22,64	24,56	25,27	22,75
Mediana		31	30	26	31	26	28	27,0	23	25	26	23
Modalna		35	33	25	36	33	33	30,0	24	30	30	23
Odchylenie standardowe		6,83	6,73	7,83	7,43	8,56	7,82	7,53	7,63	8,03	7,51	7,63
Wariancja		46,66	45,33	61,26	55,18	73,34	61,20	56,64	58,18	64,51	56,34	58,24
Skośność rozkładu		-0,765	-0,663	-0,328	-0,822	-0,278	-0,454	-0,446	-0,052	-0,248	-0,318	-0,059
Zmienność sd/wynik$_{śr}$		0,23	0,24	0,31	0,25	0,34	0,29	0,29	0,34	0,33	0,30	0,34
Kwartyle	**Q_1**	25	24	20	25	19	21	21	17	18	20	17
	Q_2	31	30	26	31	26	28	27	23	25	26	23
	Q_3	35	34	32	35	32	33	32	28	31	31	29
Rit (średnia korelacja zadania z testem)		0,44	0,41	0,46	0,45	0,48	0,47	0,50	0,44	0,46	0,42	0,42
Liczba zadań		25	25	25	26	25	26	25	25	25	26	26
Wskaźnik rzetelności alfa Feldt-Raju		0,84	0,82	0,85	0,84	0,86	0,86	0,85	0,83	0,86	0,84	0,84

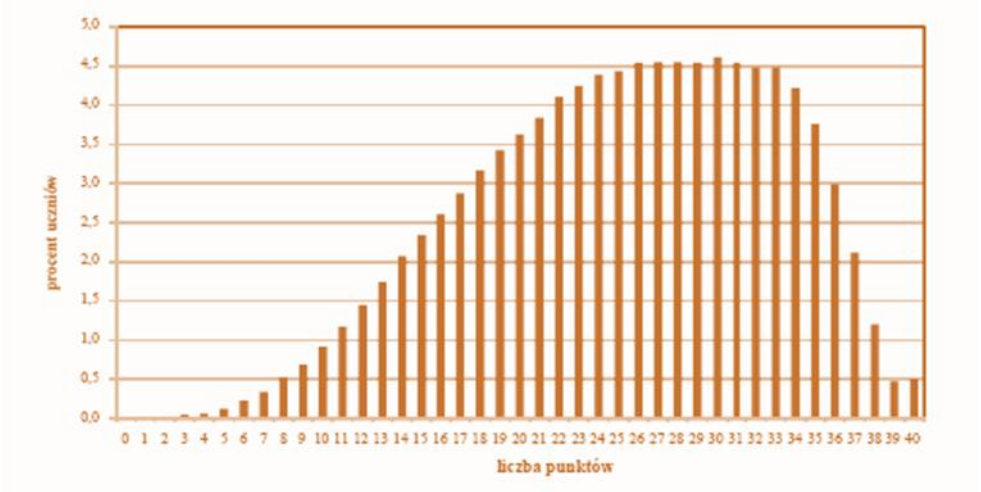
Z założenia przyjętego przez twórców koncepcji sprawdzianu egzamin na zakończenie szkoły podstawowej powinien spełniać przede wszystkim funkcję diagnostyczną. Przyjęto, że sprawdzian powinien być stosunkowo łatwy, co z kolei prowadzi do tego, że pożądany rozkład wyników powinien być lewoskośny, przesunięty w kierunku wyników wysokich. Respektowanie takiego założenia z drugiej strony ujemnie wpływa na jakość pomiaru. Szczególnie uwidacznia się to w latach 2002, 2003 i 2005 (skośność rozkładu wyników w tych latach wynosiła odpowiednio: -0,765, -0,663 i -0,822), w których rozkłady są silnie lewoskośne i różnicowanie uczniów osiągających bardzo dobre wyniki było rzeczywiście słabe, co mogło przypuszczalnie nie najlepiej wpływać także na motywację do dalszej nauki uczniów z wynikami plasującymi się powyżej 3 kwartyla. Jedynie w 2009 i 2012 roku rozkład wyników sprawdzianu zbliżony był do normalnego (por. Tabela 1.3). Na Rysunku 1.4 pokazano przykładowe rozkłady wyników sprawdzianu z lat 2009 i 2011. Rozkłady wyników sprawdzianu dla wszystkich lat przedstawione są na Rysunku 6.7 w rozdziale 6. zatytułowanym *Wyniki zrównania*.

Rysunek 1.4. Przykład rozkładu wyników uczniowskich na sprawdzianie w dwóch wybranych latach

a) rok 2009, skośność rozkładu wyników równa 0,052



b) rok 2011, skośność rozkładu wyników równa -0,318



W 2002 roku arkusz egzaminacyjny zawierał 24 procent zadań (6 zadań), których łatwość była równa lub większa od 0,90. Dla prezentowanych na Rysunku 1.4 rozkładów wyników w 2009 roku sprawdzian zawierał tylko trzy bardzo łatwe zadania (zadanie 3; 12 i 18), a w 2011 aż 7 zadań o łatwości równej lub większej od 0,90 (zadania 1; 2; 4; 5; 6 i 12).

W Tabeli 1.3 zostały przedstawione charakterystyki psychometryczne sprawdzianów z poszczególnych lat z uwzględnieniem wskaźnika rzetelności. Obok trafności, rzetelność wyników testowania jest jednym z kluczowych zagadnień pomiaru dydaktycznego bazującego na klasycznej teorii testu (KTT). Najczęściej stosowaną metodą szacowania wskaźnika rzetelności w jednokrotnym pomiarze danym testem w klasycznej teorii testu jest metoda alfa Cronbacha. Wskaźniki szacowane są zgodnie z formułą podaną przez Cronbacha:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_X^2}\right) \qquad (1.1)$$

gdzie:

n – liczba zadań w teście,

σ_i^2 – wariancja i-tego zadania,

σ_X^2 – wariancja całego testu X.

Alfa Cronbacha, jako wskaźnik rzetelności, daje niedoszacowane wartości rzetelności w przypadku, gdy test zbudowany jest z zadań, które nie są ekwiwalentne pomiarowo, co ma miejsce m.in. wtedy, gdy zadania mają różną formę (zadania wielokrotnego wyboru punktowane 0-1 i zadania punktowane kryterialnie lub holistycznie z wykorzystaniem rozwiniętej skali). Nieekwiwalentne pomiarowo zadania występują we wszystkich testach polskich egzaminów zewnętrznych, również w testach sprawdzianów. Arkusz egzaminacyjny sprawdzianu zawiera od 25 do 26 zadań, z czego 20 zadań stanowią zadania zamknięte wielokrotnego wyboru punktowane 0-1, za rozwiązanie których uczeń może uzyskać połowę możliwych do zdobycia na sprawdzianie punktów. Druga połowa punktów przydzielana jest za 5 lub 6 zadań otwartych krótkiej i rozszerzonej odpowiedzi (np. opisu, opowiadania czy zadania matematycznego).

Próbą przezwyciężenia tej trudności było traktowanie każdego podpunktu rozbudowanego zadania (dokładniej każdego ocenianego kryterium) jako osobnego zadania. W ten sposób np. pięć zadań otwartych w 2003 i 2004 zinterpretowano jako 19 pojedynczych zadań, a w 2011 sześć zadań jako 11 osobnych zadań[9]. W ten sposób jednak nie uzyskano ekwiwalentności elementów testu, co jest wymaganym założeniem do zastosowania wskaźnika, a obliczony wskaźnik alfa Cronbacha prawdopodobnie został przeszacowany.

Innym rozwiązaniem, w przypadku testów, które zawierają zadania o różnej długości skali punktowania, jest oszacowanie rzetelności testu za pomocą wskaźnika alfa Feldt-Raju.

[9] Podział wynika z kryteriów oceniania i skal przypisanych dla poszczególnych kryteriów.

Zdefiniujmy $\lambda_i = \frac{\sigma_{iX}}{\sigma_X^2}$ jako wagę i-tego zadania, gdzie σ_{iX} – kowariancja i-tego zadania z całym testem, σ_X^2 – wariancja testu. Zauważmy, że suma wag w całym teście jest równa 1. Wtedy wskaźnik alfa Feldt-Raju dla n zadań wyraża się wzorem:

$$\alpha_{Feldt-Raju} = \frac{1}{1-\sum_{i=1}^{n}\lambda_i^2}\left(1-\frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_X^2}\right) \quad (1.2)$$

gdzie:

λ_i – waga i-tego zadania,

σ_i^2 – wariancja i-tego zadania,

σ_X^2 – wariancja całego testu X.

Wskaźniki alfa Feldt-Raju, przedstawione w Tabeli 1.3 dla sprawdzianów z lat 2002-2012 są niższe od wskaźników alfa Cronbacha w przypadku potraktowania jako zadania kryteriów oceny zadań rozbudowanych, jednak wartym podkreślenia faktem jest to, że wskaźniki te są stabilne między latami.

Począwszy od sesji wiosennej 2007 roku dla sprawdzianu zostały wprowadzone ściśle równoległe dwie wersje arkuszy sprawdzianu różniące się kolejnością dystraktorów w niektórych zadaniach zamkniętych. Wprowadzenie wersji A i B nie powinno wpływać na wynik przystępującego do sprawdzianu ucznia. Jednak bardziej szczegółowe analizy pokazują, że kolejność dystraktorów i usytuowanie wśród nich poprawnej odpowiedzi ma wpływ na trudność zadania. Zjawisko to zilustrujemy na przykładzie sprawdzianu z 2012 roku.

Tabela 1.4. Porównanie trudności zadań zamkniętych w wersji A i B sprawdzianu 2012

Zadanie	DIF stat	Z (stand)	Różnica istotna (α=0,01) z>=2,58	Trudniejsze		Łatwość zadań dla wersji A i B			Poprawna odpowiedź	
				wersja A	wersja B	pA	pB	pB-pA	wersja A	wersja B
z1	0,94	-3,20	TAK	TAK		0,74	0,75	-0,01	C	B
z2	0,88	-5,95	TAK	TAK		0,83	0,84	-0,01	C	A
z3	1,07	3,85	TAK		TAK	0,46	0,44	0,02	B	C
z4	0,97	-1,22				0,81	0,81	0	D	B
z5	0,92	-4,60	TAK	TAK		0,69	0,70	-0,01	B	A
z6	1,41	18,62	TAK		TAK	0,29	0,23	0,06	A	D
z7	1,16	6,38	TAK		TAK	0,85	0,83	0,02	C	B

z8	1,08	4,45	TAK		TAK	0,56	0,55	0,01	B	B
z9	0,97	-1,63				0,63	0,63	0	C	C
z10	1,00	0,18				0,82	0,81	0,01	B	C
z11	1,24	12,39	TAK		TAK	0,67	0,63	0,04	A	D
z12	0,85	-6,66	TAK	TAK		0,85	0,86	-0,01	D	C
z13	1,34	15,71	TAK		TAK	0,62	0,57	0,05	A	D
z14	0,79	-13,57	TAK	TAK		0,28	0,32	-0,04	C	A
z15	0,99	-0,30				0,40	0,40	0	A	A
z16	1,08	4,34	TAK		TAK	0,51	0,50	0,01	A	D
z17	0,78	-15,06	TAK	TAK		0,43	0,48	-0,05	D	C
z18	0,94	-3,21				0,57	0,58	-0,01	C	B
z19	0,97	-1,95				0,45	0,46	-0,01	C	C
z20	0,99	-0,80				0,52	0,52	0	D	D

We wszystkich przypadkach, dla których różnica łatwości zadań jest istotna i duża, łatwiejsze okazywało się zadanie, w którym prawidłowa odpowiedź była wcześniej w porządku alfabetycznym niż w drugiej wersji. Przykładowo zadanie 6. (odnoszące się do przytoczonego w teście tekstu) w wersji A miało zamieszczoną poprawną odpowiedź na pozycji A (łatwość 0,29), podczas gdy w wersji B poprawną odpowiedzią było D (łatwość 0,23). Podobne efekty obserwujemy w przypadku zadań 11. i 23., których równoległe wersje powstały poprzez zamianę pozycji poprawnej odpowiedzi (werstraktora) z A na D.

Inny schemat rotacji dystraktorów zastosowany został w zadaniu 14. W wersji A poprawną odpowiedzią była odpowiedź C. Przeniesienie poprawnej odpowiedzi na pozycję A spowodowało wzrost łatwości zadania z 0,28 na 0,32.

Podobne zjawisko obserwujemy dla zadania 17., które trudniejsze było w wersji A (odpowiedź D, łatwość – 0,43) niż w wersji B (odpowiedź C, łatwość – 0,48).

Tabela 1.5. Treści zadań 6. i 14. – sprawdzian 2012. Podkreślono poprawne odpowiedzi

Wersja A

6. W pierwszym akapicie autor posługuje się czasownikami w 1. osobie liczby mnogiej, żeby

A. nawiązać bliższy kontakt z odbiorcą.

B. wyrazić swoją fascynację filmami akcji.

C. ocenić żywiołowe reakcje widzów na film.

D. zachęcić widzów do wspólnego oglądania filmów.

Wersja B

6. W pierwszym akapicie autor posługuje się czasownikami w 1. osobie liczby mnogiej, żeby

A. zachęcić widzów do wspólnego oglądania filmów.

B. ocenić żywiołowe reakcje widzów na film.

C. wyrazić swoją fascynację filmami akcji.

D. nawiązać bliższy kontakt z odbiorcą.

Wersja A

14. W jakim znaczeniu użyto w tym wierszu słowa *kocioł*?

A. Duże naczynie do gotowania.

B. Zagmatwana sytuacja, bałagan.

C. Instrument podobny do bębna.

D. Okrążenie wojsk przez nieprzyjaciela.

Wersja B

14. W jakim znaczeniu użyto w tym wierszu słowa *kocioł*?

A. Instrument podobny do bębna.

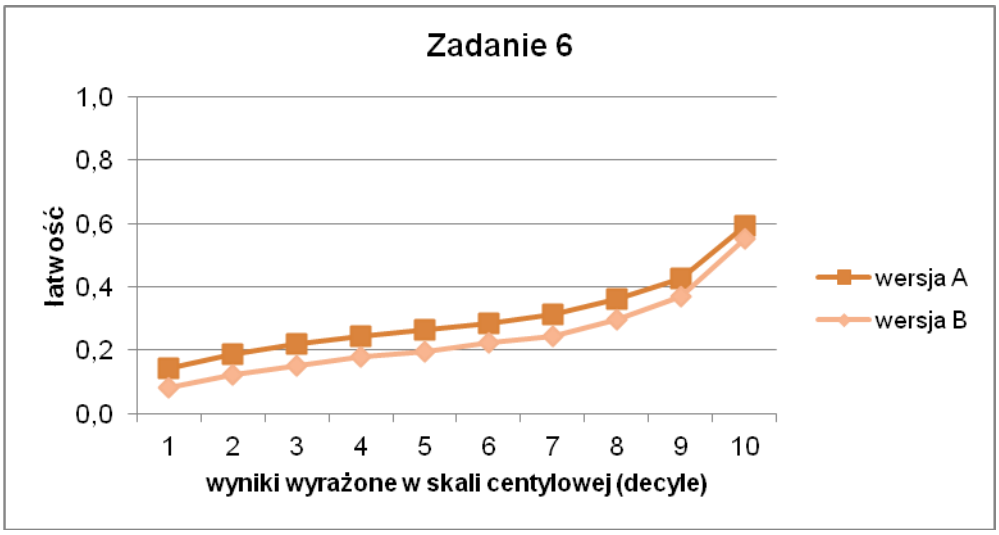
B. Duże naczynie do gotowania.

C. Zagmatwana sytuacja, bałagan.

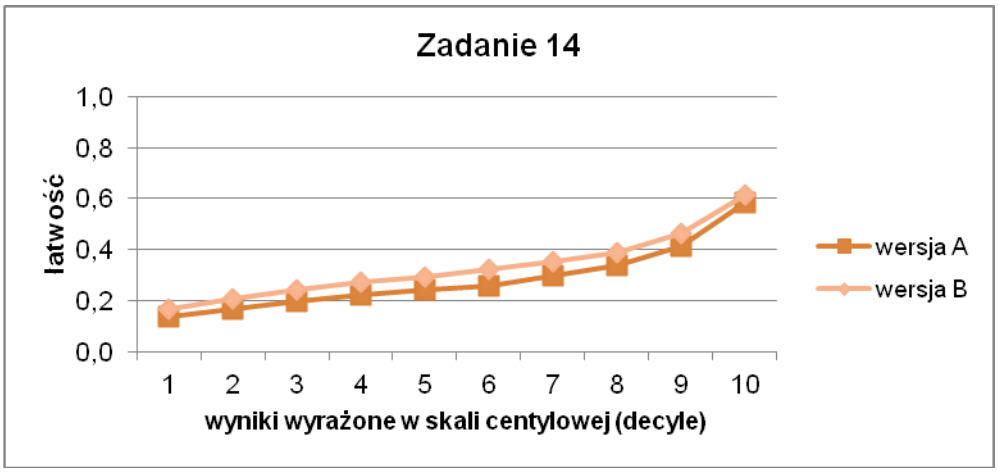
D. Okrążenie wojsk przez nieprzyjaciela.

W przypadku zadań 6. i 14. okazuje się ponadto, że różnica w łatwości zadań dla wersji A i B nie zależy od poziomu ogólnych umiejętności uczniów, sprawdzanych testem. Zamieszczone poniżej rysunki ilustrują różnicę w łatwości zadania dla uczniów, których wyniki sprawdzianu znalazły się odpowiednio w określonym decylu. Można tę sytuację określić jako jednorodne zróżnicowane funkcjonowanie zadań (ang. *uniform differential item functioning*) ze względu na kolejność dystraktorów.

Rysunek 1.5. Charakterystyka zadania 6. ze sprawdzianu 2012 ze względu na wersję arkusza



Rysunek 1.6. Charakterystyka zadań 14. ze sprawdzianu 2012 ze względu na wersję arkusza



Jeżeli weźmiemy pod uwagę wyniki całego testu, to różnica pomiędzy wersjami w danym roku jest praktycznie do pominięcia. Dzieje się tak dlatego, że często wpływy równoległej wersji arkusza egzaminacyjnego w poszczególnych zadaniach znoszą się (jak obserwujemy to na przykładzie zadania 6. i 14. ze sprawdzianu 2012).

W Tabeli 1.6 zestawiono charakterystyki arkuszy egzaminacyjnych z uwzględnieniem wersji testu. W większości edycji różnica średnich wyników dla uczniów piszących wersję A lub B była statystycznie nieistotna lub istotna, ale bardzo mała (praktycznie do zaniedbania, jak w roku 2007 i 2011).

Tabela 1.6. Charakterystyka wyników sprawdzianu w latach 2007-2012 w zależności od wersji arkusza testowego (pominięto wyniki laureatów[10])

Rok sprawdzianu	Wynik średni		Łatwość		Istotność
	wersja A	wersja B	wersja A	wersja B	
2012	22,68	22,66	0,57	0,57	0,481
2011	25,27	25,15	0,63	0,63	0,000
2010	24,55	24,57	0,61	0,61	0,537
2009	22,59	22,61	0,56	0,57	0,286
2008	25,75	25,73	0,64	0,64	0,399
2007	25,50	25,56	0,66	0,66	0,006

1.5. Podsumowanie

Sprawdzian na zakończenie edukacji w szkole podstawowej to pierwszy egzamin zewnętrzny, do którego podchodzą uczniowie w polskiej szkole. Jest to też pierwszy egzamin otwierający każdą sesję egzaminacyjną w danym roku. Odbywa się na początku kwietnia około dwa i pół miesiąca przed zakończeniem roku szkolnego. Ponieważ arkusze egzaminacyjne, klucze odpowiedzi, przykładowe rozwiązania są komunikowane na stronie internetowej CKE zaraz po egzaminie ma on też prawdopodobnie wpływ na proces dydaktyczny w pozostałych miesiącach roku szkolnego przed zakończeniem drugiego etapu edukacji. Sprawdzenie takiej hipotezy wymagałoby jednak dodatkowych badań. Załączona poniżej Tabela 1.7 przedstawia terminy sprawdzianu w poszczególnych latach.

Tabela 1.7. Terminy sprawdzianu 2002-2012

Rok	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Sprawdzian (dzień-miesiąc)	10-04	8-04	1-04	5-04	4-04	12-04	8-04	2-04	8-04	5-04	3-04

Ulokowanie sprawdzianu każdego roku w takim samym okresie (niewielkie fluktuacje spowodowane są terminami Świąt Wielkanocnych) zapewnia porównywalny między latami czas przeznaczony na edukację w poszczególnych latach w miesiącach poprzedzających sprawdzian.

[10] Laureaci konkursów przedmiotowych są zwolnieni z egzaminu i otrzymują maksymalną liczbę punktów bez pisania sprawdzianu.

Przez wszystkie lata sprawdzian obejmował umiejętności określone w *Standardach wymagań egzaminacyjnych* wyprowadzonych z tej samej *Podstawy programowej kształcenia ogólnego dla pierwszego i drugiego etapu edukacyjnego*. Arkusze egzaminacyjne budowane były według takich samych specyfikacji i procedur. Począwszy od 2009 roku uległa zmianie organizacja prac nad przygotowywaniem arkuszy egzaminacyjnych i praktyka oceniania zadań otwartych. Komisje odeszły od ściśle kryterialnego oceniania w kierunku rozwiązań bliższych ocenianiu holistycznemu. Arkusze egzaminacyjne dla uczniów bez dysfunkcji i z dysleksją rozwojową budowane są natomiast przez ogólnopolski zespół.

Na podstawie analizy wyników sprawdzianu w latach 2002-2012 można stwierdzić, że rzetelność egzaminu w poszczególnych latach jest porównywalna. Wskaźnik rzetelności alfa Feldt-Raju waha się od 0,82 do 0,86. Najniższą wartość wskaźnika (0,82) zaobserwowano w 2003 roku. Rzetelność niższa od pożądanej (powyżej 0,9) spowodowana jest niewątpliwie stosunkowo niewielką liczbą zadań (25 lub 26), co jest konsekwencją czasu przeznaczonego na ten egzamin – 60 minut. Także diagnostyczny charakter sprawdzianu i założona łatwość powyżej 0,6 ma wpływ na wartość wskaźnika rzetelności.

Wprowadzone począwszy od 2007 roku dwie wersje sprawdzianu (A i B) różniące się kolejnością dystraktorów nie spowodowały stronniczości wyniku zdających w zależności od przydzielonego arkusza egzaminacyjnego, pomimo że lokalizacja poprawnej odpowiedzi względem dystraktorów nie jest obojętna na proces odpowiadania na zadania testowe.

Ogólnie można przyjąć, że w okresie, dla którego przeprowadzono zrównanie wyników egzaminacyjnych (2002-2011) zarówno zakres sprawdzanych umiejętności, sposób tworzenia narzędzi pomiaru (arkuszy egzaminacyjnych), sposób przeprowadzania egzaminu i oceniania zadań otwartych zachowywały konieczny do zastosowanej metody zrównywania stopień stabilności.

2. Koncepcja zrównywania

2.1. Wstęp

W niniejszym rozdziale przedstawiono koncepcję zrównywania wyników egzaminacyjnych leżącą u podstaw przeprowadzonego badania. Na wstępie określono cele badania i etapy procesu badawczego. W dalszej kolejności zdefiniowane zostało pojęcie zrównywania wyników, przybliżono zastosowane w badaniu plany zbierania danych oraz przedstawiono argumentację wyboru roku bazowego stanowiącego punkt odniesienia dla wyników sprawdzianu na zakończenie szóstej klasy szkoły podstawowej po zrównaniu. Następnie opisano sposób konstrukcji wykorzystanych narzędzi badawczych. Rozdział kończy opis doboru próby uczniów do sesji zrównującej wyniki sprawdzianu i doboru próby uczniów do bieżącego zrównania wyników egzaminu gimnazjalnego (zrównanie wyników egzaminu gimnazjalnego z 2011 roku) oraz opis organizacji sesji. Statystyczna koncepcja zrównania rozwinięta jest w Rozdziale 4.

2.2. Cel badań i etapy procesu badawczego

Głównym celem badań zrównujących w 2012 roku było:

1. Zrównanie wyników sprawdzianu na zakończenie szkoły podstawowej przeprowadzonego w latach 2002-2011 z zastosowaniem arkuszy standardowych (arkusze dla uczniów bez dysfunkcji i uczniów z dysleksją rozwojową).

2. Zrównanie wyników egzaminu gimnazjalnego w części humanistycznej i matematyczno-przyrodniczej przeprowadzonego w 2011 roku do roku bazowego 2003 z zastosowaniem arkuszy standardowych (arkusze dla uczniów bez dysfunkcji i uczniów z dysleksją rozwojową).

3. Przedstawienie wyników sprawdzianu przeprowadzanego w latach 2002-2011 w skali standardowej o średniej 100 i odchyleniu standardowym 15, zakotwiczonej do roku bazowego 2004.

4. Przedstawienie porównywalnych między latami wyników egzaminu gimnazjalnego z włączeniem wyników z 2011 roku w skali standardowej o średniej 100 i odchyleniu standardowym 15, zakotwiczonej do roku bazowego 2003.

5. Przedstawienie wyników sprawdzianu i egzaminu gimnazjalnego przeprowadzanego w latach 2002-2011 w skali standardowej o średniej 100 i odchyleniu standardowym 15, zakotwiczonej do roku bazowego 2003 (dla egzaminu gimnazjalnego) i zakotwiczonej do roku bazowego 2004 (w przypadku sprawdzianu) z podziałem na:

   – chłopców i dziewczęta,
   – wielkość miejscowości, w której zlokalizowana jest szkoła,
   – szkoły publiczne i niepubliczne.

6. Przygotowanie ogólnodostępnego serwisu komunikowania wyników zrównanych egzaminu gimnazjalnego i sprawdzianu.

Studium zrównujące wyniki sprawdzianu w latach 2002-2011 oraz egzaminu gimnazjalnego rozszerzone na lata 2002-2011 zgodnie z koncepcją przygotowaną w IBE w 2010 roku obejmowało przedstawione poniżej 7 głównych obszarów działań. Opisują one proces, którego dwie ostatnie fazy mają istotne znaczenie dla przygotowania następnego etapu badań – sesji zrównującej kolejne egzaminy.

Rysunek 2.1. Główne obszary działań w procesie badawczym dotyczącym jednego etapu badań zrównujących

Podstawową trudnością w przeprowadzeniu zrównania wyników egzaminów zewnętrznych w Polsce jest fakt, że wszystkie zastosowane arkusze egzaminacyjne są publikowane na stronach internetowych Centralnej i okręgowych komisji egzaminacyjnych po przeprowadzeniu egzaminu. Upublicznianie wszystkich zadań arkusza egzaminacyjnego jest w Polsce nie tylko zwyczajowo przyjęte, ale również zagwarantowane prawnie. Arkusze egzaminacyjne są także przedrukowywane w publikacjach o charakterze dydaktycznym oraz w prasie codziennej o zasięgu ogólnokrajowym. Stają się tym samym materiałami dydaktycznymi wykorzystywanymi przez uczniów przygotowujących się do egzaminu. W konsekwencji tego, każdego roku zestaw zadań zastosowanych w arkuszu egzaminacyjnym musi być całkowicie zbudowany z oryginalnych zadań. Nie ma zatem obecnie możliwości, aby w arkuszach egzaminacyjnych stosować utajnioną pulę tych samych zadań w kolejnych egzaminach w celu kotwiczenia arkuszy egzaminacyjnych pomiędzy latami, jak to czyni się w wielu innych systemach egzaminacyjnych (Pokropek, 2011). Ma to konsekwencje dla procedury zrównania. W warunkach polskich przy obecnie stosowanych rozwiązaniach nie jest możliwe przeprowadzenie zrównania w trakcie sesji egzaminacyjnej. Jedyna możliwość, jaka pozostaje dla systemu egzaminacyjnego w obecnym kształcie, to zrównanie po sesji egzaminacyjnej (ang. *post-equating*) poprzez przeprowadzenie specjalnie zaplanowanej sesji zrównującej, w trakcie której losowo dobrana grupa uczniów rozwiązuje zadania z odpowiednio zaprojektowanego zeszytu (lub zeszytów) testowego. Takie rozwiązanie zostało zastosowane dla zrównania wyników egzaminu gimnazjalnego w 2011 roku i ponownie wykorzystane wobec egzaminu gimnazjalnego i sprawdzianu w 2012 roku.

2.3. Definicja zrównywania

Zrównywanie wyników testów ma na celu umożliwienie zamiennego, posługiwania się wynikami z tych testów. W zrównywaniu dąży się do tego, aby wyniki uzyskiwane w różnych narzędziach były jak najbardziej sobie równoważne. Wynika z tego, że jest to procedura dotyczącą testów mierzących ten sam konstrukt i tworzonych zgodnie z tymi samymi specyfikacjami testu (ang. *blueprint*). Konieczność zrównywania wyników testów jest konsekwencją faktu praktycznej niemożliwości stworzenia dwóch testów, które byłyby ściśle równoległe[11] (Holland et al., 2007). Cytując von Davier (2011, s. 1-2):

„Zrównywanie jest konieczne tylko z tego względu, że standaryzowany pomiar edukacyjny korzysta z wielu form testu, które różnią się trudnością, mimo że są tworzone zgodnie z tymi samymi specyfikacjami (…). Zrównywanie można postrzegać jako procedurę statystycznej kontroli zmiennej zakłócającej (ang. *confounding variable*), za którą przyjmuje się właśnie formę testu. Gdyby proces tworzenia testu był idealny, nie byłoby potrzeby zrównywania."[12]

Aby łączenie testów i wyników testowych (ang. *linking*) mogło być uznane za zrównywanie (ang. *equating*), konieczne jest spełnienie szeregu restrykcyjnych założeń. Założenia te (wymogi) sformułowane w zbliżonej formie można znaleźć u wielu autorów (np.: Lord, 1980; Kolen & Brennan, 2004). Poniżej wymieniono je w formie przytoczonej przez Dorans & Holland (2000, s. 282-283):

[11] Dwa testy są ściśle równoległe (ang. *strictly parallel*), jeżeli każda badana osoba z populacji w obu testach będzie miała taką samą wariancję błędu pomiaru oraz taki sam wynik prawdziwy. Mniej formalnie, dwa testy ściśle równoległe są w zupełności sobie równoważnymi (ang. *perfectly equivalent, interchangeable*) narzędziami pomiarowymi. (Gruijter & van der Kamp, 2005).

[12] Tłum. własne.

„(a) Wymóg tożsamego konstruktu (ang. *equal construnt requirement*): testy mierzące różne konstrukty nie powinny być zrównywane;

(b) Wymóg równej rzetelności (ang. *equal reliability requirement*): testy mierzące ten sam konstrukt, ale różniące się rzetelnością, nie powinny być zrównywane;

(c) Wymóg symetrii (ang. *symmetry requirement*): funkcja zrównująca wyniki w teście Y z wynikami w teście X powinna być odwrotnością funkcji zrównującej wyniki w teście X z wynikami w teście Y;

(d) Wymóg równoważności (ang. *equity requirement*): nie powinno mieć żadnego znaczenia dla osoby rozwiązującej test, którą z wersji testu rozwiązuje, gdy testy są zrównane;

(e) Wymóg niezmienniczości względem populacji (ang. *population invariance requirement*): wybór (sub)populacji użytej do obliczenia funkcji zrównującej wyniki w testach X oraz Y nie powinien mieć znaczenia, tj. funkcja zrównująca używana do łączenia wyników w testach X oraz Y powinna być *niezmiennicza względem populacji*."

Dla bliższego wytłumaczenia tych pięciu wymogów można się odwołać do komentarza zawartego w artykule *Equating test scores* (Holland et al., 2007). Wymogi tożsamego konstruktu (a) oraz równej rzetelności (b) oznaczają, że zrównywane testy powinny być skonstruowane tak, aby były zgodne pod względem treści oraz statystycznych właściwości. Wymóg symetrii (c) wyklucza możliwość zastosowania metod regresji do zrównywania testów. Wymóg równoważności (d) poniekąd tłumaczy konieczność wymogu tożsamości konstruktu (a) – jeżeli testy mierzyłyby różne konstrukty, to osoby mające je rozwiązywać preferowałyby podejście do tego testu, w którym, w ich mniemaniu, miałyby szansę uzyskać lepszy wynik (np. preferowałyby test sprawdzający taki zakres umiejętności, który ich zdaniem lepiej opanowały). Wymóg niezmienniczości względem populacji (e) można wykorzystać do uzasadnienia wymogów tożsamego konstruktu (a) oraz równej rzetelności (b). Jeżeli testy byłyby tworzone zgodnie z różnymi wzorcowymi specyfikacjami (ang. *blueprint*), to funkcja zrównująca wyniki mogłaby się różnić w zależności od wyboru subpopulacji. Przykładowo, łącząc wyniki testu badającego umiejętność rozumowania na podstawie „materiału" niewerbalnego z wynikami testu badającego umiejętność rozumowania na podstawie „materiału" werbalnego, zapewne uzyskano by różne funkcje łączące w zależności od płci. Holland et al. (2007) podają również odwołania do krytycznej dyskusji na temat omawianych wymogów dla zrównywania, niemniej zgadzają się, że tworzą one ogólną i intuicyjną „teorię" zrównywania testów.

Pewnego dopowiedzenia wymaga wymóg równoważności (d), gdyż formalnie pojawia się on w dwóch, istotnie różniących się, wersjach (Kolen & Brennan, 2004):

(1) $\forall_\tau \mathbb{P}(eq_Y(X) \leq y|\tau) = \mathbb{P}(Y \leq y|\tau)$,
(2) $\forall_\tau \mathbb{E}(eq_Y(X)|\tau) = \mathbb{E}(Y|\tau)$,

gdzie $eq_Y: X \to Y$ jest funkcją zrównującą test X z Y. Pierwsza wersja równoważności (ang. *equity*) (1) stanowi, że dla każdego wyniku prawdziwego τ warunkowy, względem τ, rozkład wyników otrzymywanych w teście Y jest taki sam jak w zrównanym do niego teście X. Natomiast, druga wersja równoważności (2) osłabia wymóg warunkowej równości dwóch rozkładów do warunkowej równości jedynie względem pierwszego momentu zwykłego (czyli wartości średniej) tych rozkładów. W szczególności wersja (2) nie wymaga równości między warunkowymi wariancjami, czyli nie wymaga równości warunkowego błędu pomiaru. Pierwsza wersja (1), sformułowana przez Lorda (1980), jest bardzo restrykcyjnym wymogiem, który u Kolena i Brennana (2004), spotyka się ze słusznym komentarzem, iż „korzystanie z równoważności Lorda jako kryterium oznacza, że

zrównywanie albo nie jest możliwe, albo nie jest potrzebne" (również: van der Linden, 2011, jak i sam Lord, 1980). Większość metod zrównywania wyników stawia sobie wprost za cel spełnienie słabszej formy równoważności (ang. *weak equity*, *first-order equity*).

Pogłębioną refleksję w kwestii problemu równoważności w „silnym" sformułowaniu Lorda (1) możemy znaleźć u van der Lindena (2011), który zwracając uwagę na lokalny charakter równania proponuje zrównywanie oparte na konstrukcji lokalnych funkcji zrównujących. Pojawia się tu ścisła zależność pomiędzy wymogiem równoważności (d), a wymogiem niezmienniczości względem populacji (e). Mimo iż zależność jest taka, że (e) implikuje (d), van der Linden (2011) sugeruje, że przybliżanie się do spełnienia wymogu niezmienniczości względem populacji również przybliża spełnienie wymogu równoważności. Ujęcie van der Lindena również wskazuje, że istotą problemu jest fakt, że pomiar edukacyjny jest obarczony błędem, co umyka w nielokalnych modelach zrównywania wyników. Zignorowanie tego faktu przy stosowaniu pojedynczej funkcji zrównującej $eq_Y(x)$ prowadzi do lokalnego obciążenia. Niestety, wydaje się, że lokalne funkcje zrównujące wyniki obserwowane nie mogłyby zostać praktycznie wykorzystane do raportowania zależności między zrównywanymi testami – np. przy zastosowaniu IRT, oznaczałoby to różne przekształcenia na test Y dla osób o tym samym wyniku w teście X, jeżeli różniłyby się ich oszacowania poziomu umiejętności θ. Niemniej, koncepcja lokalnych funkcji zrównujących i związek pomiędzy wymogiem równoważności, a niezmienniczością względem populacji dostarczają ważnych narzędzi do empirycznej weryfikacji spełnienia założenia o równoważności.

Liu & Walker (2007) dokonując przeglądu wymogów stawianych dla procedury zrównywania testów przez wymienionych wcześniej autorów, tj. Lorda (1980), Doransa & Hollanda (2000) oraz Kolena & Brennana (2004), zdecydowali się na wyszczególnienie dodatkowych trzech punktów na podstawie pracy tych ostatnich:

„Wymóg takich samych inferencji (ang. *the same inferences*): testy powinny mieć wspólne cele pomiarowe i powinny być zaprojektowane do wyciągania takich samych typów wniosków.

Wymóg takiej samej populacji docelowej (ang. *the same target population*): testy powinny mieć taką samą populację docelową.

Wymóg takich samych charakterystyk/warunków pomiarowych (ang. *the same measurement characteristics/conditions*): testy powinny mieć taką samą specyfikację, być przeprowadzane w takich samych warunkach oraz być równoważne pod względem psychometrycznych właściwości."

Zauważalna jest pewna redundantność zbioru wszystkich, już ośmiu, wymienionych wymogów niezbędnych do przeprowadzenia zrównywania wyników testowych. Jednak wydaje się, że sformułowanie wszystkich *explicite* daje jaśniejszy obraz tego, czym „w teorii" zrównywanie wyników ma być. Natomiast w praktyce, niektóre z wymogów mogą być trudne do weryfikacji (patrz: wymóg (d)) lub mniej istotne. W kwestii wagi poszczególnych wymogów wciąż toczy się dyskusja, którą w skrócie omawiają Holland et al. (2007). Natomiast w kwestii praktycznej weryfikacji wymienionych wymogów, warto odwołać się do Liu & Walker (2007), którzy zastosowali interesujący zestaw kryteriów zrównywalności (ang. *equatability*) testu SAT w wersji funkcjonującej do 2004 roku z nową wersją, która weszła do użycia w 2005 roku. Znamienne jest, że zadanie zrównywania wyników zostało podjęte w obliczu znacznej zmiany w zakresie wzorcowych specyfikacji testu (ang. *test blueprint*), co przy konserwatywnym traktowaniu wszystkich wymogów stawianych przed zrównywaniem, mogłoby zostać uznane za argument dyskwalifikujący możliwość dokonania zrównania. Zaproponowane kryteria zrównywalności były następujące:

1. Podobieństwo konstruktu (ang. *construct similiarity*); weryfikowane przez stopień podobieństwa treści, jak i statystyczne właściwości testu.

2. Empiryczna relacja pomiędzy nowym i starym testem; weryfikowana poprzez współczynnik korelacji między dwoma testami w odniesieniu do współczynnika rzetelności każdego z testów (wyznaczającego górną granicę dla takiej korelacji).

3. Precyzja pomiaru; weryfikowana zarówno poprzez współczynnik rzetelności, jak i poprzez lokalne miary błędu pomiaru umiejętności.

4. Niezmienniczość w podgrupach (ang. *subgroup invariance*); weryfikowana poprzez zbadanie relacji między średnimi wynikami w zależności od istotnych zmiennych grupujących oraz poprzez analizę postaci funkcji łączącej wyniki w zależności od istotnych zmiennych grupujących.

2.4. Zastosowany plan zrównywania wyników i wybór roku bazowego

2.4.1. Plan zrównywania sprawdzianu

Teoretyczne omówienie planu zrównania i implementację metody łącznej kalibracji modelu IRT do zrównania egzaminu gimnazjalnego i sprawdzianu przedstawiono w rozdziale 4 zatytułowanym *Statystyczna koncepcja zrównania*. W tym podrozdziale skupimy się na przedstawieniu, w jaki sposób zostało zaplanowane i zrealizowane zbieranie danych koniecznych do zrównania *post factum* wyników sprawdzianu na zakończenie szkoły podstawowej z lat 2002-2011 oraz wyników egzaminu gimnazjalnego z 2011 roku z wynikami z lat poprzednich.

W Tabeli 2.1 przedstawiono schematycznie plan badawczy zastosowany do zebrania danych wykorzystanych do zrównywania wyników sprawdzianu w szóstej klasie szkoły podstawowej 2002-2011. W tabeli uwzględniono zaplanowaną do 2014 roku kontynuację bieżącego zrównywania. Zadanie zrównania wyników sprawdzianu w badaniach przeprowadzonych w 2012 roku obejmuje 11 różnych populacji uczniów: $\mathcal{P}_{02}$, $\mathcal{P}_{03}$, … , $\mathcal{P}_{12}$, z których każda podchodziła z odpowiedniej edycji sprawdzianu: T_{02}, T_{03}, … , T_{12}.

Uczniowie z populacji $\mathcal{P}_{02}$, $\mathcal{P}_{03}$, … , $\mathcal{P}_{11}$ pisali na zakończenie szóstej klasy szkoły podstawowej wyłącznie sprawdzian odpowiadający ich rocznikowi (T_{02}, T_{03}, … , T_{11}) – są to niepołączone dane, zebrane przez system egzaminacyjny i udostępnione do badań przez Centralną Komisję Egzaminacyjną. Natomiast wszyscy uczniowie z populacji piszącej sprawdzian w 2012 roku – $\mathcal{P}_{12}$ rozwiązywali zadania sprawdzianu T_{12}, ale ponadto, z tej populacji wylosowano 11 losowo równoważnych prób uczniów: S_{12}^{1}, $S_{12}^{2}, \dots, S_{12}^{11}$, którzy wzięli udział w dodatkowej sesji zrównującej. Każda z prób w sesji zrównującej rozwiązywała zadania z zeszytu testowego składającego się z dwóch podzbiorów zadań kotwiczących wybranych z poprzednich egzaminów (T^{A}) oraz podzbioru dodatkowych zadań kotwiczących (do zrównywania pionowego z egzaminem gimnazjalnym i wynikami sprawdzianu w kolejnych latach). Przykładowo, próba S_{12}^{5} rozwiązywała zadania z zeszytu testowego z podzbiorami kotwiczących zadań T_{06}^{A} oraz T_{07}^{A} (podzbiory zadań z arkuszy egzaminacyjnych z lat 2006 oraz 2007) oraz z podzbiorem zadań dodatkowych C^{5}.

W Tabeli 2.1 symbolami T_{02}^{R}, $T_{03}^{\mathrm{R}}, \dots, T_{11}^{\mathrm{R}}$, oznaczono części arkuszy egzaminacyjnych (sprawdzianu), które nie zostały włączone do zeszytów testowych wykorzystywanych w sesji zrównującej. Grubszą

ramką zaznaczono populację uczniów piszących sprawdzian w 2012 roku i 11 równoważnych prób uczniów z tej populacji biorących udział w sesji zrównującej wyniki sprawdzianu.

Tabela 2.1. Schemat planu badawczego – sprawdzian na zakończenie VI klasy szkoły podstawowej – w tabeli uwzględniono zrównywanie bieżące oraz planowane do roku 2014

Źródło danych	Populacja	Próba	N obs. (szac.)	Test lub zbiór zadań																																	
				T_{02}		T_{03}		T_{04}		T_{05}		T_{06}		T_{07}		T_{08}		T_{09}		T_{10}		T_{11}		T_{12}	T_{13}	T_{14}	Dodatkowe zadania kotwiczące										
				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	T_{11}^R	T_{11}^A	T_{12}	T_{13}	T_{14}	C^1	C^2	C^3	C^4	C^5	C^6	C^7	C^8	C^9	C^{10}	C^{11}
Dane egzaminacyjne	$\mathcal{P}_{02}$		$5{,}4 \cdot 10^5$	■	■																																
Dane egzaminacyjne	$\mathcal{P}_{03}$		$5{,}3 \cdot 10^5$			■	■																														
Dane egzaminacyjne	$\mathcal{P}_{04}$		$5{,}2 \cdot 10^5$					■	■																												
Dane egzaminacyjne	$\mathcal{P}_{05}$		$4{,}9 \cdot 10^5$							■	■																										
Dane egzaminacyjne	$\mathcal{P}_{06}$		$4{,}6 \cdot 10^5$									■	■																								
Dane egzaminacyjne	$\mathcal{P}_{07}$		$4{,}4 \cdot 10^5$											■	■																						
Dane egzaminacyjne	$\mathcal{P}_{08}$		$4{,}1 \cdot 10^5$													■	■																				
Dane egzaminacyjne	$\mathcal{P}_{09}$		$4{,}0 \cdot 10^5$															■	■																		
Dane egzaminacyjne	$\mathcal{P}_{10}$		$3{,}8 \cdot 10^5$																	■	■																
Dane egzaminacyjne	$\mathcal{P}_{11}$		$3{,}7 \cdot 10^5$																			■	■														
Dane egzaminacyjne	$\mathcal{P}_{12}$		$3{,}6 \cdot 10^5$		▪		▪		▪		▪		▪		▪		▪		▪		▪		▪	■			▪	▪	▪	▪	▪	▪	▪	▪	▪	▪	▪
Dane egzaminacyjne	$\mathcal{P}_{13}$		$3{,}5 \cdot 10^5$						▫														▫	▫	■							▫					
Dane egzaminacyjne	$\mathcal{P}_{14}$		$3{,}4 \cdot 10^5$						▫															▫	▫	■						▫					
Dane z sesji zrównujących, 2012	$\mathcal{P}_{12}$	S_{12}^1	800		■		■																	■			■										
Dane z sesji zrównujących, 2012	$\mathcal{P}_{12}$	S_{12}^2	800				■		■															■				■									
Dane z sesji zrównujących, 2012	$\mathcal{P}_{12}$	S_{12}^3	800						■		■													■					■								
Dane z sesji zrównujących, 2012	$\mathcal{P}_{12}$	S_{12}^4	800								■		■											■						■							
Dane z sesji zrównujących, 2012	$\mathcal{P}_{12}$	S_{12}^5	800										■		■									■							■						
Dane z sesji zrównujących, 2012	$\mathcal{P}_{12}$	S_{12}^6	800												■		■							■								■					
Dane z sesji zrównujących, 2012	$\mathcal{P}_{12}$	S_{12}^7	800														■		■					■									■				
Dane z sesji zrównujących, 2012	$\mathcal{P}_{12}$	S_{12}^8	800																■		■			■										■			
Dane z sesji zrównujących, 2012	$\mathcal{P}_{12}$	S_{12}^9	800																		■		■	■											■		
Dane z sesji zrównujących, 2012	$\mathcal{P}_{12}$	S_{12}^{10}	800				■																■	■												■	
Dane z sesji zrównujących, 2012	$\mathcal{P}_{12}$	S_{12}^{11}	800												■								■	■													■
Dane z sesji zrównujących, 2013	$\mathcal{P}_{13}$	S_{13}	1600						□														□	□	■							□					
Dane z sesji zrównujących, 2014	$\mathcal{P}_{14}$	S_{14}	1600						□															□	□	■						□					

Legenda:

- (puste pole) uczniowie nie rozwiązują żadnego zadania
- ■ uczniowie rozwiązują wszystkie zadania
- ▪ część uczniów rozwiązuje wszystkie zadania
- ■ uczniowie rozwiązują wszystkie zadania na sesji innego typu
- ▪ część uczniów rozwiązuje wszystkie zadania na sesji innego typu
- □ uczniowie rozwiązują podzbiór zadań
- ▫ część uczniów rozwiązuje podzbiór zadań
- □ uczniowie rozwiązują podzbiór zadań na sesji innego typu
- ▫ część uczniów rozwiązuje podzbiór zadań na sesji innego typu

2.4.2. Plan zrównywania bieżącego egzaminu gimnazjalnego

W 2011 roku zostały zrównane wyniki egzaminu gimnazjalnego z lat 2002-2010. W 2012 roku zebrane zostały dane umożliwiające powtórne przeprowadzenie procedury zrównywania wyników egzaminu gimnazjalnego uwzględniające oprócz egzaminów 2002-2010 także egzamin z roku 2011. Zgodnie z koncepcją badań w kolejnych latach 2013-2014 będzie prowadzone dla egzaminu gimnazjalnego zrównywanie obejmujące wyniki egzaminów z następnych roczników.

2.4.2.1. Zmiana podstawy programowej

W 2012 roku do egzaminu przystąpili gimnazjaliści, którzy uczyli się już według nowej podstawy programowej obowiązującej od 2009 roku. Zmiana podstawy programowej w zróżnicowany sposób dotyczy przedmiotów kształcących umiejętności sprawdzane na egzaminie gimnazjalnym. Nowa postawa w ogólnym zarysie zachowała strukturę starej podstawy programowej. Jednak cele edukacyjne zostały rozbudowane, uszczegółowione i zapisane w formie operacyjnej. Umiejętności zawarte w celach kształcenia w nowej podstawie programowej opisane są znanymi ze standardów wymagań będących podstawą przeprowadzania egzaminu w ostatnim roku nauki w gimnazjum[13] czasownikami operacyjnymi między innymi takimi jak: [uczeń] opisuje, porządkuje, porównuje, rozpoznaje, wyjaśnia, wskazuje, stosuje, dokumentuje, formułuje, weryfikuje, rozróżnia, odczytuje, selekcjonuje i inne. Porównując cele kształcenia nowej podstawy programowej ze standardami wymagań, można stwierdzić, że jest to doprecyzowany i uporządkowany zapis celów, które były treścią wymagań będących podstawą przeprowadzania egzaminu w ostatnim roku nauki w gimnazjum. W wyniku tego nie ma potrzeby definiowania standardów wymagań egzaminacyjnych, jak to miało miejsce w przypadku starej podstawy.

Cele edukacyjne w największym stopniu uległy zmianie w historii, wiedzy o społeczeństwie, w matematyce i biologii. Zadania szkoły i treści nauczania poszerzono w WOS-ie w częściach dotyczących wychowania obywatelskiego i wychowania do aktywnego udziału w życiu gospodarczym. W biologii natomiast zakres treści nauczania został poszerzony o takie działy jak: systematyka organizmów, podstawy botaniki, podstawowa nauka o ewolucji oraz podstawy biochemii i nauki o podstawowych w biologii procesach przetwarzania energii. Wymagania ogólne w starej podstawie programowej analizowanej łącznie ze standardami wymagań egzaminacyjnych i w nowej podstawie są definiowane w podobny sposób. Podstawowa różnica we wszystkich przedmiotach dotyczy umieszczenia w nowej podstawie wymagań szczegółowych zapisanych w sposób czynnościowy. Taki zapis dostarcza wspólnego języka dla nauczycieli organizujących środowisko edukacyjne i dla konstruktorów narzędzi egzaminacyjnych.

2.4.2.2. Zmiana formuły egzaminu

Egzamin gimnazjalny jest egzaminem obowiązkowym zdawanym na zakończenie nauki w gimnazjum. Do 2012 roku egzamin ten składał się z trzech części: (1) część humanistyczna, (2) część matematyczno-przyrodnicza, (3) język obcy nowożytny[14]. Za każdą z części uczeń mógł otrzymać maksymalnie 50 punktów. Każdą z części egzaminu uczniowie zdawali innego dnia, przy czym w pierwszym dniu była to część humanistyczna, w dniu drugim – matematyczno-przyrodnicza i w dniu

[13] Załącznik do rozporządzenia Ministra Edukacji Narodowej z dnia 28 sierpnia 2007 roku (Dz.U. z dnia 31 sierpnia 2007 r. Nr 157 poz. 1102).

[14] Do roku szkolnego 2008/2009 egzamin gimnazjalny obejmował tylko dwie części: (1) część humanistyczną, (2) część matematyczno-przyrodniczą.

trzecim – język obcy nowożytny. Egzamin zwykle odbywa się w kwietniu, zaś jego wyniki ogłaszane w czerwcu są wykorzystywane w procesie rekrutacji do szkoły ponadgimnazjalnej.

Zawartość merytoryczna egzaminu nie zmieniała się w latach 2002-2011. W części matematyczno-przyrodniczej sprawdzane były umiejętności w następujących obszarach:

- Stosowanie terminów, pojęć i procedur z zakresu przedmiotów matematyczno-przyrodniczych niezbędnych w praktyce życiowej i dalszym kształceniu.
- Wyszukiwanie i stosowanie informacji.
- Wskazywanie i opisywanie faktów, związków i zależności, w szczególności przyczynowo-skutkowych, funkcjonalnych, przestrzennych i czasowych.
- Stosowanie zintegrowanej wiedzy i umiejętności do rozwiązywania problemów.

W części humanistycznej sprawdzane były umiejętności w obszarach:

- czytania i odbioru tekstów kultury,
- tworzenia własnego tekstu.

Natomiast w części dotyczącej języka obcego nowożytnego sprawdzane były umiejętności w obszarach:

- odbioru tekstu słuchanego,
- odbioru tekstu pisanego,
- reagowania językowego.

Każda z części zawierała zadania zamknięte (punktowane 0-1) i otwarte. W części humanistycznej standardowe arkusze egzaminacyjne zawsze zawierały 20 zadań zamkniętych wielokrotnego wyboru (ww) z jedną poprawną odpowiedzią do wyboru oraz od 6 (2006) do 13 (2005) zadań otwartych. Wśród zadań otwartych były zadania krótkiej wypowiedzi i zadania szerszej swobodnej wypowiedzi. W arkuszach matematyczno-przyrodniczej części egzaminu gimnazjalnego 25 zadań było zadaniami zamkniętymi (ww), a liczba zadań otwartych zmieniała się w zakresie od 13 do 16.

W związku z wprowadzeniem od roku szkolnego 2009/2010 nowej podstawy programowej, zmieniła się struktura egzaminu gimnazjalnego oraz rozwiązania dotyczące budowy arkuszy egzaminacyjnych.

Nowy egzamin gimnazjalny, który jest zdawany od kwietnia 2012 roku, również jest zdawany przez 3 dni, z tym, że części są podzielone na dwa testy. W części humanistycznej obejmuje:

- Test z języka polskiego złożony z zadań zamkniętych i zadań otwartych – pisany przez 90 minut.
- Test z historii i wiedzy o społeczeństwie złożony tylko z zadań zamkniętych – pisany przez 60 minut.

W części matematyczno-przyrodniczej:

- Test z matematyki złożony z zadań zamkniętych i otwartych – pisany przez 90 minut.

- Test z przedmiotów przyrodniczych złożony tylko z zadań zamkniętych – pisany przez 60 minut.

2.4.2.3. Schemat planu bieżącego zrównywania egzaminu gimnazjalnego

Podczas zrównywania wyników egzaminu gimnazjalnego z 2011, które zostało przeprowadzone na próbie losowej z populacji gimnazjalistów zdających egzamin gimnazjalny w 2012 roku, konieczne było zmierzenie się z problemem wynikającym z dwóch faktów. Po pierwsze, uczniowie przystępujący do egzaminu i biorący udział w badaniach zrównujących uczyli się według nowej podstawy programowej. Po drugie, dla zwiększenia motywacji uczniów do rozwiązywania testów kotwiczących badania musiały być prowadzone w taki sposób, aby dla uczniów mogły być postrzegane jako egzamin próbny (identyczna organizacja, jak na egzaminie właściwym w 2012 roku, dostarczenie uczniom informacji o wynikach w krótkim czasie po badaniach).

Aby rozwiązać pierwszy problem, zeszyty zastosowane do zrównania zostały skonstruowane w ten sposób, aby zawierały zadania, które są zgodne ze starą i nową podstawą programową. W związku z drugim problemem organizacja badań zrównujących była taka, jak organizacja egzaminu w 2012 roku, a nie taka jak w 2011 roku.

Podczas badań zrównujących uczniowie włączeni do próby badawczej rozwiązywali zatem test z języka polskiego, test z historii i wiedzy o społeczeństwie, test z matematyki i test z przedmiotów przyrodniczych. Ponieważ w 2011 roku arkusz egzaminacyjny w części humanistycznej zawierał niewiele zadań z historii, podjęto decyzję, aby test z części humanistycznej badający umiejętności z historii i WOS uzupełnić zadaniami z 2005, 2008 i 2009 roku (por. Tabela 2.2a).

Tabela 2.2. Schemat planu badawczego – gimnazjum 2011-2014

Źródło danych		Populacja	Próba	N obs. (szac.)	Test lub zbiór zadań																														
					T_{02}		T_{03}		T_{04}		T_{05}		T_{06}		T_{07}		T_{08}		T_{09}		T_{10}		T_{11}	T_{12}	T_{13}	T_{14}	Dodatkowe zadania kotwiczące								
					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}	T_{07}	T_{08}^{R}	T_{08}^{A}	T_{09}^{R}	T_{09}^{A}	T_{10}^{R}	T_{10}^{A}	T_{11}	T_{12}	T_{13}	T_{14}	C^1	C^2	C^3	C^4	C^5	C^6	C^7	C^8	C^9
Dane egzaminacyjne		$\mathcal{P}_{02}$		$5{,}6 \cdot 10^5$	■	■																													
		$\mathcal{P}_{03}$		$5{,}5 \cdot 10^5$			■	■																											
		$\mathcal{P}_{04}$		$5{,}4 \cdot 10^5$					■	■																									
		$\mathcal{P}_{05}$		$5{,}3 \cdot 10^5$							■	■																							
		$\mathcal{P}_{06}$		$5{,}2 \cdot 10^5$									■	■																					
		$\mathcal{P}_{07}$		$5{,}1 \cdot 10^5$											■	■																			
		$\mathcal{P}_{08}$		$4{,}7 \cdot 10^5$													■	■																	
		$\mathcal{P}_{09}$		$4{,}5 \cdot 10^5$															■	■															
		$\mathcal{P}_{10}$		$4{,}3 \cdot 10^5$																	■	■													
		$\mathcal{P}_{11}$		$4{,}1 \cdot 10^5$	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■				■	■	■	■	■	■	■	■	■
		$\mathcal{P}_{12}$		$3{,}9 \cdot 10^5$																			□	■							□				
		$\mathcal{P}_{13}$		$3{,}7 \cdot 10^5$																			□	□	■						□				
		$\mathcal{P}_{14}$		$3{,}5 \cdot 10^5$																				□	■	■					□				
Dane z sesji zrównujących	2011	$\mathcal{P}_{11}$	S_{11}^{1}	800		■		■															■				■	■							
		$\mathcal{P}_{11}$	S_{11}^{2}	800				■		■													■				■		■						
		$\mathcal{P}_{11}$	S_{11}^{3}	800						■		■											■						■	■					
		$\mathcal{P}_{11}$	S_{11}^{4}	800								■		■									■							■	■				
		$\mathcal{P}_{11}$	S_{11}^{5}	800										■		■							■					■			■				
		$\mathcal{P}_{11}$	S_{11}^{6}	800												■		■					■					■				■			
		$\mathcal{P}_{11}$	S_{11}^{7}	800														■		■			■										■	■	
		$\mathcal{P}_{11}$	S_{11}^{8}	800																■		■	■								■			■	
		$\mathcal{P}_{11}$	S_{11}^{9}	800				■						■									■							■				■	
		$\mathcal{P}_{11}$	S_{11}^{10}	800										■								■	■											■	■
		$\mathcal{P}_{11}$	S_{11}^{11}	800				■														■	■					■							■
	2012*	$\mathcal{P}_{12}$	S_{12}	1600																		■	□	■							□				
	2013	$\mathcal{P}_{13}$	S_{13}	1600																			□	□	■						□				
	2014	$\mathcal{P}_{14}$	S_{14}	1600																				□	□	■					□				

Legenda

- uczniowie nie rozwiązują żadnego zadania
- uczniowie rozwiązują wszystkie zadania
- część uczniów rozwiązuje wszystkie zadania
- uczniowie rozwiązują wszystkie zadania na sesji innego typu
- część uczniów rozwiązuje wszystkie zadania na sesji innego typu
- uczniowie rozwiązują podzbiór zadań
- część uczniów rozwiązuje podzbiór zadań
- uczniowie rozwiązują podzbiór zadań na sesji innego typu
- część uczniów rozwiązuje podzbiór zadań na sesji innego typu

* badania 2012 rozpisane są szczegółowo w Tabeli 2.2a.

Tabela 2.2a. Schemat planu badawczego – zrównywanie egzaminu gimnazjalnego 2012

Źródło danych		Populacja	Próba	N obs. (szac.)	Test lub zbiór zadań																																	
					T_{02}		T_{03}		T_{04}		T_{05}		T_{06}		T_{07}		T_{08}		T_{09}		T_{10}		T_{11}		T_{12}	T_{13}	T_{14}	Dodatkowe zadania kotwiczące										
					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}	T_{07}	T_{08}^{R}	T_{08}^{A}	T_{09}^{R}	T_{09}^{A}	T_{10}^{R}	T_{10}^{A}	T_{11}^{R}	T_{11}^{A}	T_{12}	T_{13}	T_{14}	C										
Badania zrównujące 2012	GMP – Matematyka	$\mathcal{P}_{12}$	S_{12}																		■		■	■			□											
	GMP – Przyroda	$\mathcal{P}_{12}$	S_{12}																		■		■	■			□											
	GH – Historia i WOS	$\mathcal{P}_{12}$	S_{12}								□						□		□		■		■	■			□											
	GH – Język Polski	$\mathcal{P}_{12}$	S_{12}																		■		■	■			□											

Zarówno w przypadku zrównywania sprawdzianu z lat 2002-2011, jak i bieżącego zrównywania wyników egzaminu gimnazjalnego, mamy do czynienia ze zrównywaniem *post hoc*. Nakreślony plan zrównywania ujawnia dwie potencjalnie zakłócające (*confounding*) zmienne, jakie warto uwzględnić podczas analiz:

1. Motywacja uczniów – uczniowie biorący udział w sesji zrównującej nie rozwiązują zadań w warunkach testu doniosłego.

2. Ujawnienie zadań – uczniowie biorący udział w sesji zrównującej mogli mieć wcześniej styczność z zadaniami z wcześniejszych egzaminów ćwicząc swoje umiejętności na upublicznionych arkuszach egzaminacyjnych z lat wcześniejszych.

Wymienione dwie, potencjalnie zakłócające, zmienne stoją względem siebie w opozycji. Czynnik motywacyjny powinien obniżać wyniki podczas sesji zrównującej w porównaniu z warunkami testu doniosłego. Czynnik ujawnienia zadań natomiast stawia w uprzywilejowanej pozycji uczniów, którzy zetknęli się z zadaniami zawartymi w testach zrównujących podczas przygotowania do właściwego egzaminu. Można jednak sądzić, że jeżeli czynniki motywacji oraz ujawnienia zadań będą równomiernie działały pomiędzy zeszytami testowymi, to nie powinny one wpłynąć na wyniki zrównania w sposób systematyczny – wskazują na to również wyniki przeprowadzonych symulacji.

Należy zaznaczyć, że Tabele 2.1 i 2.2 nie oddają w pełni faktycznego skomplikowania zastosowanego planu. Po pierwsze, w obydwu tabelach nie uwzględniono faktu zbalansowania kolejności testek dla prób $S_{11}^{1}, S_{11}^{2}, \ldots, S_{11}^{11}$ (ang. *counterbalancing*). Po drugie, Tabela 2.2 pokazuje rozkład zadań dla jednej z części egzaminu gimnazjalnego – części humanistycznej lub matematyczno-przyrodniczej do roku 2011. Dlatego też Tabela 2.2 uzupełniona została Tabelą 2.2a, która obejmuje cztery testy zastosowane w badaniach.

Uwzględnienie wszystkich części egzaminu w zrównywaniu jest bardzo istotne także ze statystycznego punktu widzenia, ponieważ stwarza możliwość wykorzystania informacji zawartej w korelacji między wynikami z tych części, co powinno zmniejszyć błędy wnioskowania statystycznego.

Analiza planu zrównywania wyników sprawdzianu i egzaminów gimnazjalnych pozwala na stwierdzenie, że nie ma w tym planie jednej pary testów egzaminacyjnych, która byłaby połączona którymś z klasycznych planów zrównywania (por. Rozdział 4): EG (ang. *equivalent groups design*), SG (ang. *single group design*), CB (ang. *counterbalanced design*), lub NEAT (ang. *nonequivalent groups with anchor test design*). Przykładowo, dla próby S_{11}^{1}, można uznać, że mamy do czynienia z planem CB dla zrównania kotwic T_{02}^{A} oraz T_{03}^{A} względem populacji $\mathcal{P}_{11}$, jednak nie wystarcza to do zrównania egzaminów z lat 2002 oraz 2003, gdyż części T_{02}^{R} oraz T_{03}^{R} nie są rozwiązywane przez żadnych uczniów z populacji $\mathcal{P}_{11}$.

W konsekwencji skomplikowania schematu zbierania danych do przeprowadzenia zrównania nie będzie można wykorzystać wielu klasycznych metod zrównywania nie korzystających z narzędzi parametrycznego modelowania IRT, które specyficznie są związane z wymienionymi klasycznymi planami. Modele IRT mogą zostać bezpośrednio zaadaptowane do sytuacji przedstawionej w omawianym planie, na przykład z wykorzystaniem metody kalibracji łącznej niezależnie estymowanymi rozkładami umiejętności dla poszczególnych populacji. Szczegółowy opis statystycznego instrumentarium wykorzystanego do zrównywania zamieszczono w Rozdziale 4.

2.4.3. Wybór roku bazowego

Dla egzaminu gimnazjalnego rok 2003 został przyjęty jako rok bazowy (rok odniesienia), pomimo że intuicyjnie wydaje się, że racjonalnym rozwiązaniem byłoby przyjęcie za rok bazowy pierwszego roku przeprowadzenia egzaminu gimnazjalnego, czyli 2002. Biorąc jednak pod uwagę fakt, że pierwszy rok (2002) najprawdopodobniej obciążony jest efektem „innowacji", jaką było wprowadzenie w Polsce egzaminu zewnętrznego i wielu działań podjętych przez komisje egzaminacyjne celem przybliżenia uczniom i nauczycielom egzaminu: egzaminy próbne, intensywne szkolenie nauczycieli, analizy wyników z egzaminów próbnych, spotkania okręgowych komisji egzaminacyjnych z dyrektorami i nauczycielami itp. – takie rozwiązanie zostało odrzucone. Ponadto, w 2003 roku zgodnie z zaleceniem Centralnej Komisji Egzaminacyjnej wprowadzono zewnętrzną kontrolę przestrzegania procedur egzaminacyjnych podczas organizacji egzaminu w szkołach, co, jak można przypuszczać, spowodowało większą rzetelność przeprowadzania egzaminów. Ostatecznie ustalono, że rokiem bazowym dla egzaminu gimnazjalnego będzie rok 2003.

Przyjęcie tego samego roku (2003) dla sprawdzianu byłoby wygodne z punktu widzenia komunikowania porównywalnych między latami wyników dla szerokiego spektrum odbiorców. Za odrzuceniem roku 2002 dla sprawdzianu przemawiają te same argumenty, które wykorzystano w przypadku gimnazjum. Niestety podczas analizy wyników egzaminacyjnych sprawdzianu z 2003 roku okazało się, że w trakcie oceniania zadania otwartego 24 (punktowanego kryterialnie od 1 do 8), którego celem było sprawdzenie umiejętności redagowania listu, popełniony został błąd na poziomie koordynacji oceniania w jednej z OKE. Ponadto wskaźnik rzetelności dla sprawdzianu przeprowadzonego w 2003 roku okazał się najniższy (por. Tabela 1.3) w porównaniu z rzetelnością oszacowaną dla pozostałych lat. Biorąc pod uwagę te fakty za rok bazowy przyjęto 2004 rok. Oznacza to, że wyniki sprawdzianu z lat 2002-2011 zostały wyrażone w skali wyników z roku 2004.

2.5. Narzędzia badawcze

W badaniach zastosowane zostały trzy typy narzędzi badawczych:

1. Testy kotwiczące (zeszyty testowe) przeznaczone do rozwiązania przez uczniów włączonych do próby badawczej:

    a. dla sprawdzianu,

    b. dla egzaminu gimnazjalnego.

2. Ankiety ucznia przeznaczone do zastosowania po każdej części badań.

3. Ankiety nauczyciela adresowane do nauczycieli uczących w badanej klasie przedmiotów, których zakres obejmuje treści sprawdzane na sprawdzianie i na egzaminie gimnazjalnym.

2.5.1. Testy kotwiczące (zeszyty testowe)

Test kotwiczący jest to test, który zawiera, oprócz zadań nowych, także tzw. zadania „stare", tzn. te same, które zastosowane zostały w teście referencyjnym. Wszystkie zadania tworzące test kotwiczący określane są mianem zadań „wspólnych".

Zrównanie za pomocą testów kotwiczących stosuje się w celu zrównania wyników uzyskiwanych za pomocą testów zbudowanych według tych samych założeń odnośnie zawartości oraz własności psychometrycznych. Test kotwiczący powinien zostać zbudowany w taki sposób, aby minimalizował błędy zrównania wynikające z różnicy w poziomie umiejętności uczniów rozwiązujących test referencyjny i test kotwiczący wykorzystywany do zrównywania. W omawianym tu przypadku zjawisko to odnosi się do ewentualnych różnic w poziomie umiejętności uczniów rozwiązujących zadania arkusza egzaminacyjnego na egzaminie w danym roku i poziomu umiejętności uczniów rozwiązujących testy kotwiczące zbudowane na podstawie tych arkuszy dla studium zrównującego przeprowadzanego w innym roku (2012) na miesiąc przed sprawdzianem na zakończenie szkoły podstawowej i na miesiąc przed egzaminem gimnazjalnym (2012). Występowanie tego typu błędów jest powszechnie znane (por. Liu et al. 2009).

Jak podkreśla Dorota Węziak (Węziak, 2007), kluczową sprawą w procesie budowania testu kotwiczącego jest ustalenie liczby zadań wspólnych. W literaturze przedmiotu zalecenia w tej dziedzinie nie są jednoznaczne. Większość autorów zaleca, aby była to liczba zadań z przedziału od 5 do 15 (por. Wright i Master, 1982; Wright i Stone, 1979). Dokładniejsze wytyczne podają Afrassa i Keeves (1999). Zalecają oni, aby dla testów liczących 60 pozycji liczba zadań wspólnych w teście kotwiczącym kształtowała się w granicach od 10 do 20. Natomiast Smith (2004), powołując się na Angoffa, twierdzi, że optymalna liczba zadań wspólnych to większa z dwóch liczb: 20 zadań lub 20% zadań całego testu. Ponadto Smith zwraca uwagę, że z przeprowadzonych badań empirycznych wynika, że w przypadku liczby pytań/zadań kotwiczących z przedziału od 15 do 25, dołożenie dodatkowych pytań nie pociąga za sobą znaczącego wzrostu precyzji zrównywania. Należy jednak podkreślić, że cała procedura zrównywania jest wykonalna nawet w przypadku tylko jednego pytania wspólnego. W praktyce liczba pytań/zadań wspólnych, na których ma się opierać zrównywanie, powinna być wyższa od założonej docelowo. Jest to niezbędne, ponieważ często zdarza się, że stopień dopasowania pytań do modelu nie jest wystarczający, aby zapewnić wiarygodność uzyskiwanych wyników. W takim przypadku pytania niedopasowane dostatecznie są usuwane z procesu ostatecznej kalibracji.

Dorota Węziak zwraca także uwagę, że na wynik zrównania ma także wpływ sposób wyboru pytań wspólnych. W przypadku modeli IRT, oprócz już wymienionych zaleceń, znajdują zastosowanie wszelkie wskazówki, jakie w tej kwestii zostały wypracowane w ramach klasycznych[15] metod zrównywania[16]. Wskazane jest, aby zadania kotwiczące charakteryzowały się poziomem trudności zbliżonym do przeciętnego poziomu umiejętności uczniów (w trakcie procesu zrównywania te dwie wielkości są porównywane bezpośrednio – obie są wyrażane w jednostkach zwanych logitem, a właściwość ta wynika z założeń modeli IRT), jako że te mają najmniejsze standardowe błędy oszacowania. Ogólnie nie poleca się wykorzystywania do kotwiczenia zadań, które charakteryzują się ekstremalnymi poziomami trudności (zadania bardzo łatwe i zadania bardzo trudne). Rozstęp oszacowań trudności zadań wspólnych powinien wynosić od 1,5 logita do 2 logitów względem średniego poziomu umiejętności uczniów rozwiązujących dany test, zaś ich rozkład powinien być bardziej zbliżony do rozkładu jednostajnego niż normalnego.

Według badań prowadzonych przez Liu et al. (2009) zastosowanie baterii zadań wspólnych o tej samej zawartości merytorycznej oraz tej samej przeciętnej trudności co cały test, ale mniejszym zróżnicowaniu tej trudności (wyrażonym odchyleniem standardowym) daje wyniki zrównania o tym samym stopniu dokładności (wyrażonym standardowym błędem zrównania oraz średniokwadratowym błędem zrównania), co zastosowanie baterii pytań wspólnych o tym samym zróżnicowaniu trudności co cały test[17].

W literaturze przedmiotu znaleźć można również zalecenia, aby usytuowanie wspólnych pytań w zrównywanym teście było przynamniej zbliżone do usytuowania tychże pytań w teście referencyjnym. Ponadto zwraca się również uwagę na zawartość merytoryczną pytań wspólnych. Według Cook i Paterson (za: Hu, Rogers i Vukmirovic, 2008) zawartość merytoryczna zadań wspólnych ma istotne znaczenie dla dokładności zrównania zwłaszcza wtedy, gdy grupa uczniów rozwiązujących test zrównywany znacząco różni się poziomem umiejętności od grupy referencyjnej. W omawianych tu badaniach sesja zrównująca odbyła się miesiąc przed egzaminem właściwym, zatem można założyć, że poziom umiejętności uczniów rozwiązujących testy zrównujące był porównywalny z poziomem umiejętności uczniów rozwiązujących zadania z arkuszy egzaminacyjnych.

2.5.1.1. Koncepcja budowy zeszytu testowego do zrównywania

Dla sprawdzianu przygotowano jedenaście wersji kotwiczących zeszytów testowych. Każdy zeszyt składał się z 3 części:

1. zadania z egzaminu z roku *n* (część A z odpowiednim numerem),

2. zadania z egzaminu z roku *m* (część B z odpowiednim numerem),

[15] Przez klasyczne rozumie się metody nie wykorzystujące modeli *IRT* (teoria odpowiedzi na zadania testowe).

[16] Więcej na ten temat znaleźć można m.in. u Livingstona (2004).

[17] W swoich badaniach Liu et al. (2009) budowali testy kotwiczące składające się z odpowiednio 35 pytań oraz 20 pytań dla testu składającego się z 78 pytań.

3. zadania nowe, nieznane uczniom kotwiczące sprawdzian z egzaminem gimnazjalnym (część C z odpowiednim numerem).

Zeszyty testowe zawierały średnio 31 zadań (od 29 do 34), za które badani mogli uzyskać 40 punktów, co odpowiada długości skali wyników surowych w sprawdzianie na zakończenie szkoły podstawowej w poszczególnych latach. W założeniach konstrukcyjnych zeszytów kotwiczących przyjęto, że części A i B powinny zawierać po 40 procent zadań, a zadania kotwiczące z egzaminem gimnazjalnym (część C) 20 procent zadań. W rezultacie uzyskano dla zadań w części A i B zeszytów testowych średnio po 39 procent zadań i 22 procent dla części C.

Zeszyty testowe przygotowane zostały przez wybranych specjalistów mających doświadczenie w zakresie budowania zadań i arkuszy egzaminacyjnych. Schemat rozdziału zadań do zeszytów kotwiczących przedstawiono w Tabeli 2.3. Do badań został zastosowany taki sam schemat zrównania, jak w przypadku zrównania dla egzaminu gimnazjalnego przeprowadzonego w 2011 roku, który przed zastosowaniem w 2011 roku został skonsultowany z ekspertem zewnętrznym Antonem Beguinem (CITO, Arnhem).

Grupy zadań kotwiczących sprawdzian z egzaminem gimnazjalnym C1, C2, C3, C4 są podzbiorem zbioru zadań kotwiczących nieznanych uczniom zastosowanym w sesji zrównującej wyniki egzaminu gimnazjalnego w 2011 roku. Z zadań tych zostało utworzonych 11 zestawów C^1, C^2 …….C^{11} w sposób przedstawiony w tabeli 2.3 w wierszu zatytułowanym zadania kotwiczące sprawdzian z egzaminem gimnazjalnym. Zadania kotwiczące z grupy C zostały wybrane do sprawdzianu w ten sposób, aby ich trudność i obszar treściowy możliwie najlepiej przystawały do standardów wymagań dla sprawdzianu.

Z tej samej puli został także wyselekcjonowany podzbiór zadań, który został zastosowany do kotwiczenia egzaminu maturalnego z matematyki z egzaminem gimnazjalnym w części matematyczno-przyrodniczej (badania przeprowadzone w 2013 roku). W podobny sposób przypadku części humanistycznej egzaminu gimnazjalnego, z grupy C zostaną wybrane zadania do kotwiczenia tej części egzaminu z egzaminem maturalnym z języka polskiego (badania planowane na 2014 r.).

Rysunek 2.1. Ilustracja doboru zadań kotwiczących (C) do badań zrównujących wyniki sprawdzianu z puli zadań (C) zastosowanych w badaniach zrównujących wyniki egzaminu gimnazjalnego w części matematyczno-przyrodniczej w 2011 roku

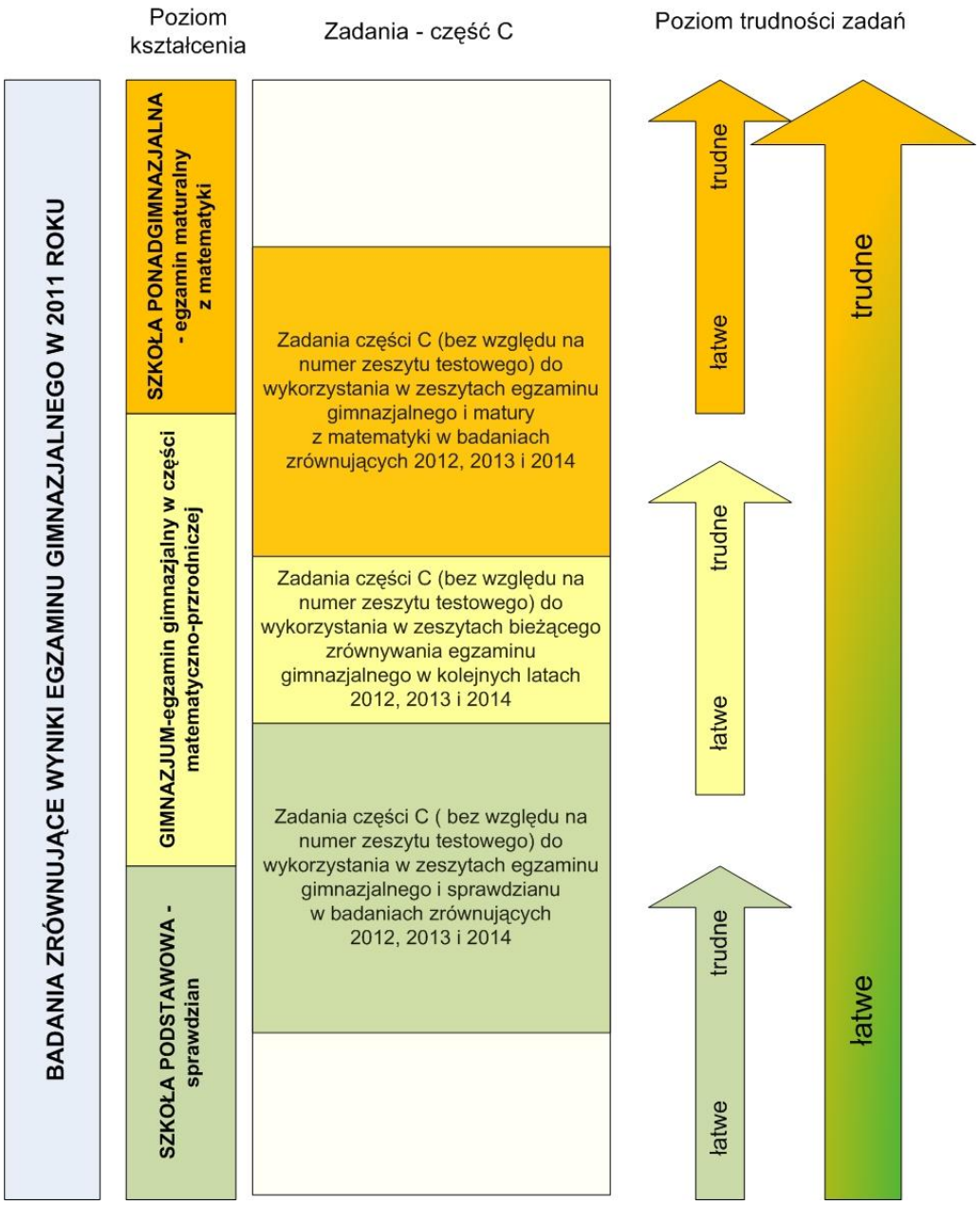


Tabela 2.3. Schemat konstruowania testów kotwiczących zeszytów testowych do sprawdzianu

Numer zeszytu w sesji zrównującej	Zeszyt 1		Zeszyt 2		Zeszyt 3		Zeszyt 4		Zeszyt 5		Zeszyt 6		Zeszyt 7		Zeszyt 8		Zeszyt 9		Zeszyt 10		Zeszyt 11	
Arkusz z roku:																						
2012	Arkusz egzaminacyjny zastosowany w sesji egzaminacyjnej – 3 kwietnia 2012																					
Zadania kotwiczące sprawdzian z egzaminem gimnazjalnym	C^1		C^2		C^3		C^4		C^5		C^6		C^7		C^8		C^9		C^{10}		C^{11}	
	C3	C1	C1	C2	C2	C3	C3	C4	C4	C1	C1	C2	C2	C3	C3	C4	C4	C1	C1	C2	C2	C3
2002	A1																					
2003	B1		B1																B1			
2004			A2		A2																	
2005					B3		B3															
2006							A4		A4													
2007									B5		B5										B5	
2008											A6		A6									
2009													B7		B7							
2010															A8		A8					
2011																	A9		A9		A9	

Przed przystąpieniem do budowy testów zrównujących w części A i B przeprowadzono analizy własności psychometrycznych wszystkich zadań wchodzących w skład arkuszy sprawdzianu na zakończenie szkoły podstawowej w latach 2002-2011 oraz arkuszy egzaminacyjnych egzaminu gimnazjalnego w części humanistycznej i matematyczno-przyrodniczej z 2011 roku. W analizach zastosowano podejście zgodne z klasyczną teorią testów oraz z teorią wyniku zadania testowego (IRT) – model Rascha.

Przy konstrukcji poszczególnych zeszytów testowych założono, że suma punktów możliwych do zdobycia w przypadku każdego testu kotwiczącego (każdego z 11 zeszytów testowych) dla sprawdzianu powinna w miarę możliwości być taka sama (40 punktów). Dopuszczono jednak odchylenie od tej liczby na poziomie maksymalnie ±3 punktów. Zawartość merytoryczna zadań w części A i B w przybliżeniu powinna stanowić reprezentatywną próbkę zawartości merytorycznej całych arkuszy, z których pochodziły.

Z założenia w ramach części A i B zróżnicowanie łatwości/trudności zadań wybranych do zeszytu testowego powinno być zbliżone do zróżnicowania łatwości/trudności zadań całego arkusza, z którego wybierano zadania. Zasada ta nie była ściśle przestrzegana, ponieważ przy wyborze zadań do części A i B priorytetem było unikanie zadań, których parametry psychometryczne były poniżej wartości ustalonych przez zespół badawczy jako krytyczne (m.in. moc dyskryminacyjna bliska 0: Rit<0,2)[18] lub tylko jeden działający dystraktor). Przyjęto ponadto, że kotwiczące zeszyty testowe sprawdzianu i egzaminu gimnazjalnego do części humanistycznej nie będą zawierały zadania rozbudowanej wypowiedzi pisemnej (rozprawka), ze względu na znaczne obciążenie efektem egzaminatora wyników z tej części. Spowodowany tym ubytek punktów w stosunku do sumy punktów możliwych do zdobycia zrekompensowano dołączając do zeszytu testowego inne zadania z arkuszy egzaminacyjnych stanowiących źródło zadań do części A i B oraz z zestawu zadań nieznanych badanym (części C).

Na wypadek konfliktu w jednoczesnym spełnieniu wszystkich wymogów niezbędnych przy konstrukcji zeszytów do sesji zrównującej określono kolejność niezbędnych do wypełnienia kryteriów. Jako priorytet przyjęto kolejno: wykluczenie zadań, które źle funkcjonowały w oryginalnych arkuszach egzaminacyjnych, zapewnienie co najmniej 30 procent dla każdej części A i B oraz co najmniej 20 procent puli C zadań, zapewnienie dla części A i B reprezentacji treści całego źródłowego arkusza egzaminacyjnego z danego roku, zapewnienie w zeszytach podobnego do arkuszy źródłowych rozkładu trudności zadań.

Ponieważ w arkuszach egzaminacyjnych poszczególne zadania występują w wiązkach przyporządkowanych do tekstu źródłowego przyjęto, że źle funkcjonujące zadania (słaba korelacja zadania z testem, ale bez tego zadania, bardzo słabe dopasowanie zadania do modelu zastosowanego w analizie IRT) zostaną usunięte z wiązki. W szczególnym przypadku, jeśli po takim zabiegu zbyt mało zadań pozostałoby przypisanych do danego tekstu źródłowego, można było także usunąć całą wiązkę.

Jak już wspomniano, kluczową składową zeszytów testowych budowanych do zastosowania w sesji zrównującej były zadania części C. Były to zadania odpowiednio przygotowane do wykorzystania w sesji zrównującej, które nie mogły być wcześniej znane badanym. Ponadto te zadania zostały tak dobrane, aby kotwiczyły sprawdzian z egzaminem gimnazjalnym.

[18] Korelacja zadania z całym testem.

Włączenie takich zadań do arkuszy było niezwykle ważne. Po pierwsze, istniało duże prawdopodobieństwo, że zadania z części A i B były wcześniej znane uczniom, a znajomość tych zadań wpływa na wyniki, jakie uczniowie uzyskają w trakcie badań (uzyskanie wyższych wyników niż można było oczekiwać bez wcześniejszego kontaktu z tymi zadaniami). Po drugie, zawarcie nieznanych zadań w arkuszach sprzyja wzrostowi motywacji testowej uczniów. Po trzecie, zadania te stanową podstawę do przeprowadzenia próby zrównywania pionowego wyników egzaminacyjnych w szóstej klasie szkoły podstawowej i ostatniej klasie gimnazjum. Przyjęto wstępne założenie, że taka procedura może umożliwić sprawdzenie, jakie jest zróżnicowanie poziomów umiejętności pomiędzy poszczególnymi poziomami kształcenia (zrównanie pionowe). Zagadnienie to będzie przedmiotem analiz planowanych po zakończeniu czwartego etapu studium zrównującego w 2014 roku.

2.5.2. Badanie ankietowe

Badanie ankietowe nauczycieli i uczniów, to między innymi poszukiwania odpowiedzi na pytanie, w jakim stopniu zadania egzaminacyjne z poprzednich lat są wykorzystywane do ćwiczeń w procesie dydaktycznym w kontekście korzystania z innych komercyjnych testów. To także pytanie, czy efekt niższej motywacji związany z faktem, iż test zrównujący nie był dla uczniów tak zwanym egzaminem doniosłym, może mieć systematyczny wpływ na wyniki badania zrównującego. Obydwa te efekty mają przeciwstawny wpływ na wyniki uczniów uzyskiwane podczas badań zrównujących. Przeprowadzone na ten temat badania ankietowe nie dostarczają jednoznacznej odpowiedzi czy te efekty się kompensują. Ta część badania kontynuowana jest w kolejnych etapach studium zrównującego. Analizy z badań ankietowych zostaną przeprowadzone po zakończeniu czwartego etapu badań zrównujących w 2014 roku i będą przedmiotem odrębnego raportu.

2.6. Uczniowie biorący udział w sesji zrównującej

W rozdziale tym przedstawiony zostanie opis populacji, operat losowania i zastosowany schemat losowania próby szkół i uczniów do sesji zrównującej (informacje szczegółowe zamieszczono w Aneksie C).

Podstawowym sposobem gromadzenia danych koniecznych do zrównania wyników egzaminacyjnych było przeprowadzenie badań z wykorzystaniem zeszytów testowych. Celowi temu podporządkowane są wszystkie inne czynności badawcze. W ramach zrównywania wyników sprawdzianu po szkole podstawowej wykorzystanych zostało jedenaście unikatowych zeszytów testowych występujących w dwóch wariantach (wynikających z rotacji kolejności odpowiedzi w zadaniach zamkniętych). Każdy uczeń rozwiązywał zadania z jednego zeszytu testowego. Taki schemat badania implikował kształt schematu doboru próby. Jako, że każdy zeszyt testowy musiał być wypełniony przez losowo dobranych uczniów pozwala to myśleć o próbie, jako jedenastu oddzielnych losowych próbach (jedna próba to uczniowie rozwiązujący zadania z jednego zeszytu testowego w wersji A i B). Trzeba jednak pamiętać, iż faktycznie w czasie procedury losowania jedenaście „podprób” musiało zostać wylosowanych jednocześnie, co wynika ze złożonego schematu losowania i skończonej liczby placówek w populacji szkół podstawowych. Dla egzaminu gimnazjalnego, dla którego badanie zrównujące było kontynuacją etapu pierwszego przeprowadzonego w 2011 roku, wylosowano jednocześnie dwie „podpróby” szkół.

Dla każdej z „podprób" wylosowanych zostało 40 szkół, co łącznie daje 440 szkół podstawowych i 80 gimnazjów. Przy takiej liczebności można przeprowadzić proporcjonalną alokację próby, która skutkuje (w przybliżeniu) tym, iż wylosowana próba jest próbą „autoważoną", co w wielu wypadkach ułatwia analizę. Z każdej szkoły wylosowana została jedna klasa, co w przybliżeniu pozwoliło uzyskać

dla sprawdzianu próbę liczącą około 8800 uczniów dla gimnazjum odpowiednio 1600 uczniów (przyjęto iż średnia wielkość klasy wynosi 20).

2.6.1. Populacja i operat losowania

Populacja docelowa w badaniu została określona dla sprawdzianu jako: uczniowie klas szóstych szkół podstawowych dla młodzieży, z wyłączeniem szkół specjalnych i przyszpitalnych. Operatem pierwotnym w tym losowaniu była lista szkół udostępniona przez CKE, sporządzona na podstawie zamówień na arkusze egzaminacyjne w 2011 roku. Dla egzaminu gimnazjalnego populację docelową stanowili uczniowie klas trzecich ze szkół gimnazjalnych dla młodzieży, z wyłączeniem szkół specjalnych i przyszpitalnych. Podobnie, jak w przypadku losowania próby dla badań zrównujących wyniki sprawdzianu na zakończenie szkoły podstawowej operatem pierwotnym w tym losowaniu była lista szkół udostępniona przez CKE, sporządzona na podstawie zamówień na arkusze egzaminacyjne w 2011 roku.

Aby zapewnić minimalny błąd pomiaru dla narzuconych ograniczeń kosztowych, postanowiono zarówno w przypadku sprawdzianu, jak i egzaminu gimnazjalnego ograniczyć populację docelową do uczniów uczących się w szkołach nie mniejszych niż jednostki dziesięcioosobowe (na poziomie klasy przystępującej do sprawdzianu/egzaminu). Tę populację nazywamy rzeczywistą populacją losowania, której odpowiada dostosowany do niej operat wtórny. Decyzja ta została podjęta między innymi po uwzględnieniu wyników symulacji niezbędnych do oszacowania błędów próbkowania. Przyjęte ograniczenie populacji wyłącza z operatu około 4% uczniów.

2.6.1.1. Schemat losowania

Losowanie miało charakter losowania warstwowego (a), proporcjonalnego do liczebności szkoły na poziomie klas/klasy (b) oraz wielopoziomowego (c).

(a1) Każda z „podprób" została podzielona na 4 warstwy, według kategorii wielkości miejscowości. Dokonano alokacji zbliżonej do procentowego udziału uczniów w danej warstwie. Szczegóły operatu losowania oraz alokacji próby przedstawia Tabela 2.4. (sprawdzian) i Tabela 2.5. (egzamin gimnazjalny).

(a2) Dodatkowo każda z warstw została podzielona na n/2 warstw drugiego rzędu (gdzie n to liczba wylosowanych szkół w danej warstwie) utworzonych ze względu na wyniki egzaminacyjne szkoły z roku 2011. Z każdej tak powstałej warstwy wylosowane zostały 2 szkoły.

(b) Szkoły losowane były proporcjonalnie do liczby uczniów uczących się w klasach VI szkoły podstawowej i klasach III gimnazjum. Użyta została klasyczna metoda kumulatywna.

(c) W każdej wylosowanej szkole do udziału w badaniu poproszona została jedna klasa (oddział) wylosowana w sposób prosty. Jeżeli w danej szkole była tylko jedna klasa ostatniego rocznika automatycznie ta klasa weszła do badania. Jeżeli klas było więcej, w sposób prosty losowana była jedna klasa. Ponieważ poprawne wylosowanie klasy ma szczególnie znaczenie dla reprezentatywności próby wybieranej w dwustopniowym losowaniu, dla każdej szkoły przypisano unikalne liczby losowe do wyboru klas.

Wraz z próbą główną została wylosowana próba rezerwowa. Każda szkoła miała przypisane dwie szkoły rezerwowe. W razie odmowy tylko te szkoły mogły zastąpić szkołę, która odmówiła udziału w badaniu.

Tabela 2.4. Struktura operatu losowania i alokacji próby – sprawdzian

Warstwa	Operat				Alokacja
	Liczba szkół	Procent szkół	Liczba uczniów	Procent uczniów	Liczba szkół w „podpróbie”
miasto do 20 tys.	1025	10.2%	58773	16.4%	6
miasto od 20 tys. do 100 tys.	1118	11.1%	71213	19.9%	8
miasto powyżej 100 tys.	1586	15.8%	87577	24.4%	10
Wieś	6331	62.9%	140874	39.3%	16
Razem	10,060	100%	358437	100%	40

Tabela 2.5. Struktura operatu losowania i alokacji próby – gimnazjum

Warstwa	Operat				Alokacja
	Liczba szkół	Procent szkół	Liczba uczniów	Procent uczniów	Liczba szkół w „podpróbie”
miasto do 20 tys.	912	14.5%	80338	19.6%	8
miasto od 20 tys. do 100 tys.	895	14.2%	85816	20.9%	8
miasto powyżej 100 tys.	1247	19.8%	98088	23.9%	10
Wieś	3245	51.5%	145999	35.6%	14
Razem	6299	100%	410241	100%	40

W trakcie badań została zapewniona taka sama organizacja sesji zrównującej (zarówno w szkołach podstawowych, jak i w gimnazjach), jak organizacja właściwej sesji egzaminacyjnej. Oznaczało to, że uczniowie wraz z arkuszem egzaminacyjnym otrzymywali kartę odpowiedzi, na którą nanosili odpowiedzi na zadania zamknięte. Natomiast odpowiedzi na zadania otwarte zostały ocenione przez egzaminatorów posiadających certyfikat egzaminatora właściwego egzaminu oraz doświadczenie w sprawdzaniu prac egzaminacyjnych.

Ponieważ w 2012 roku została wprowadzona nowa formuła egzaminu gimnazjalnego, sesja zrównująca gimnazjalne wyniki egzaminacyjne zarówno w części humanistycznej, jak i matematyczno-przyrodniczej została przeprowadzona już według procedur obowiązujących w nowej formule. Oznacza to, że do badań zastosowano odrębny test z języka polskiego i odrębny test z historii z WOS-em. Tę samą zasadę zastosowano w odniesieniu do części matematyczno-przyrodniczej.

3. Realizacja badania

Przeprowadzenie badań terenowych zostało zlecone zewnętrznej firmie wybranej w drodze przetargu. Poniżej opisano poszczególne etapy realizacji badania terenowego – osobno dla szkół podstawowych i gimnazjów.

3.1. Szkoły podstawowe

3.1.1. Rekrutacja szkół

Zgodnie z założeniami projektu rekrutacja szkół była podzielona na trzy etapy:

1) kontakt listowny,
2) kontakt telefoniczny,
3) kontakt osobisty (ankieter).

Wysłanie listów zapraszających do udziału w badaniu nastąpiło w dniach 3-7 października 2011 roku, a więc zgodnie z założonym harmonogramem. Oprócz listów z informacjami o badaniu, kierowanych do dyrektorów, przesyłki zawierały formularz zgody na udział w badaniu. Jednocześnie do szkół, między 5 października a 15 listopada, zostały wysłane emaile zawierające informację o badaniu wraz z elektroniczną wersją pism wysłanych pocztą. Z uwagi na odmowy wzięcia udziału w badaniach, wysyłka listów kontynuowana była do zakończenia rekrutacji.

Na drugim etapie rekrutacji przeprowadzano rozmowy telefoniczne ze szkołami. Ten etap został poprzedzony szkoleniem pracowników odpowiedzialnych za kontakt ze szkołami. Kontakt telefoniczny miał na celu potwierdzenie otrzymania przez szkoły listów z informacjami o badaniu, uzyskanie wstępnej zgody dyrektora na udział w badaniu, zdobycie podstawowych informacji o liczbie klas w szkole, wylosowanie klasy mającej wziąć udział w badaniu, a także ustalenie terminu osobistej wizyty ankietera w szkole.

W dniach od 24 października do 19 listopada 2011 roku ankieterzy odwiedzali dyrektorów, którzy wymagali wizyty osobistej przed udzieleniem zgody na udział w badaniu. W trakcie takich wizyt ankieterzy udzielali szczegółowych informacji o przebiegu badania oraz ustalali termin jego realizacji w szkole.

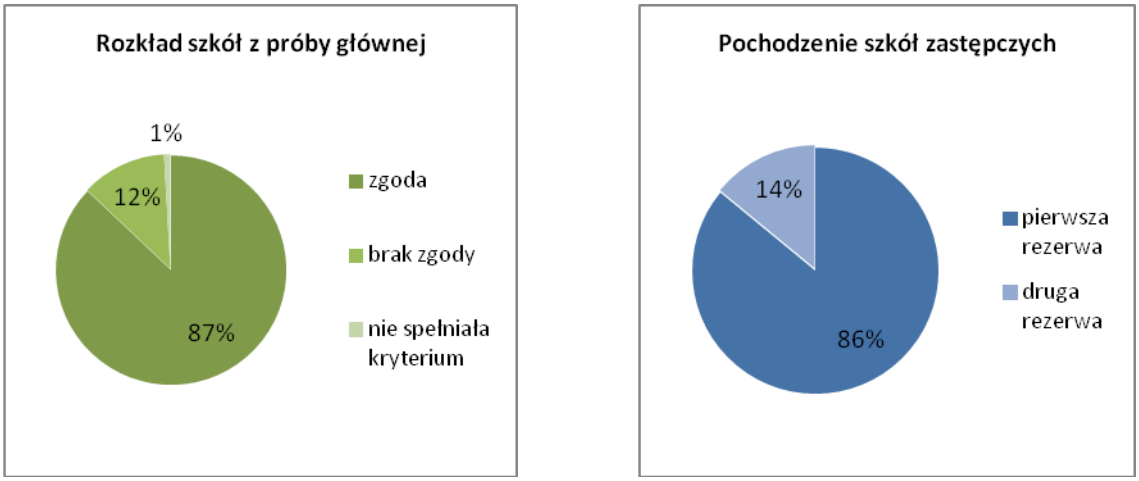
W trakcie rekrutacji dyrektorzy niektórych szkół odmawiali udziału w badaniu. Najczęstszą przyczyną odmów były:

- brak czasu na przeprowadzenie badania – szkoły często brały udział w innych badaniach lub egzaminach próbnych,
- zbyt późna prośba o udział w badaniu – dyrektorzy wskazywali, że decyzje o organizacji roku szkolnego podejmują na przełomie sierpnia i września,

- objęcie badaniem tylko jednego oddziału klasowego, co wymaga organizacji innego testu próbnego dla pozostałych oddziałów.

Pomimo odmów części szkół, na 440 wszystkich szkół podstawowych biorących udział w badaniu, udało się zrealizować badania w 383 szkołach z próby podstawowej, a zatem w 88%. W 53 przypadkach powodem wykluczenia szkoły z próby podstawowej była odmowa dyrektora na udział w badaniu, a w 4 przypadkach szkoła nie spełniała kryteriów (szkoły integracyjne lub tylko z oddziałami I-III). Spośród szkół rezerwowych 49 pochodziło z pierwszej próby rezerwowej, a 8 z drugiej próby rezerwowej. Powyższe dane przedstawia Rysunek 3.1.

Rysunek 3.1. Rozkład szkół z próby głównej i pochodzenie szkół zastępczych



Jako efekt rekrutacji szkół powstał raport z rekrutacji zawierający listę szkół biorących udział w badaniu wraz w datami badania i oznaczeniem wylosowanych klas. W kilku przypadkach zostały zbadane dodatkowe klasy, od czego dyrektorzy uzależnili zgodę na udział w badaniu.

3.1.2. Badanie pilotażowe

W celu przetestowania procedur badania oraz narzędzi w postaci testów umiejętności szkolnych i kwestionariuszy ankiet w dniach 9-12 stycznia 2012 roku odbyły się badania pilotażowe w 12 szkołach w województwach: kujawsko-pomorskim, pomorskim i małopolskim. W pilotażu nie był wymagany losowy dobór szkół, w związku z czym zastosowano dobór dogodnościowy. Pilotaż nie ujawnił poważnych błędów w procedurach i narzędziach, jedynie w jednym przypadku ankieter nie zastosował się do wymogu naprzemiennego rozdawania uczniom wersji A i B testów, co zaowocowało szczególnym podkreśleniem tego elementu procedury podczas szkolenia przed właściwym etapem badania.

3.1.3. Szkolenie koordynatorów i ankieterów

Przed przystąpieniem do realizacji badania w terenie konieczne było zorganizowanie szkolenia koordynatorów oraz ankieterów. Nad jego kształtem czuwali przedstawiciele PAOU. W celu zachowania spójności przekazywanych informacji i zebrania pytań od wszystkich ankieterów zaangażowanych w badanie zdecydowano się na organizację centralnego szkolenia. Ze względu na dużą liczbę uczestników, około 200 osób, szkolenie przeprowadzono w dwóch równoległych grupach. Szkolenie odbyło się w Warszawie, 18 lutego 2012 roku.

Szkolenie było podzielone na dwie części. W pierwszej, omawiana była organizacja badania związana z otrzymaniem narzędzi przez ankieterów, sposób kontaktu ze szkołami oraz z koordynatorami,

a także sposób przekazywania wypełnionych materiałów badawczych. W części drugiej, przedstawiciele PAOU omówili koncepcję badania i szczegółowe procedury dystrybucji narzędzi w szkołach, a także zwrócili uwagę na specyfikę badań edukacyjnych i sposób zachowania się ankieterów w szkołach. Uczestnicy otrzymali komplet materiałów szkoleniowych w postaci procedur realizacji badania oraz próbek narzędzi.

Podczas szkolenia wszelkie niejasności procedur zgłaszane przez ankieterów zostały wyjaśnione na miejscu, a dodatkowo doprecyzowano niektóre sformułowania w dokumentach. Efektem szkolenia było także przystosowanie list uczniów i raportów z realizacji badania w szkołach do potrzeb zgłaszanych przez ankieterów.

3.1.4. Realizacja badań w szkołach i ocenianie

W trakcie rekrutacji szkół do badania dyrektorzy często sygnalizowali zbyt późny termin badania. Według harmonogramu badanie miało nastąpić a dniach 5-15 marca 2012, a przekazanie wyników do szkół w dniach 28 marca-1 kwietnia 2012 roku. Z uwagi na to, iż 3 kwietnia 2012 roku odbywał się sprawdzian szóstoklasisty, zdecydowano się na przesunięcie terminu realizacji badania w szkołach o tydzień wcześniej – na dni 27 lutego-8 marca 2012 roku. Jednocześnie analogicznemu przesunięciu uległa sesja oceniania na weekend 10-11 marca.

W badaniach zebrano zeszyty testowe, jak i kwestionariusze ankiety od 9086 uczniów. Tabela 3.1 zawiera podsumowanie liczby wypełnionych zeszytów testowych w zależności od ich wersji. Oprócz materiałów pochodzących od uczniów, zebrano kwestionariusze ankiety od 1774 nauczycieli.

Tabela 3.1. Podsumowanie wypełnionych zeszytów testowych w zależności od wersji

	Wersja A		**Wersja B**	
	Liczebność	**Procent całości**	**Liczebność**	**Procent całości**
Zeszyt 1	429	5%	415	5%
Zeszyt 2	400	4%	384	4%
Zeszyt 3	396	4%	373	4%
Zeszyt 4	413	5%	396	4%
Zeszyt 5	423	5%	398	4%
Zeszyt 6	426	5%	403	4%
Zeszyt 7	435	5%	418	5%
Zeszyt 8	458	5%	441	5%
Zeszyt 9	412	5%	402	4%
Zeszyt 10	409	5%	388	4%

Zeszyt 11	441	5%	426	5%
Razem	4642	51%	4444	49%

W ocenianiu prac brało udział 157 egzaminatorów podzielonych na pięć zespołów kierowanych przez przewodniczących zespołów egzaminatorów (PZE). Nad całością oceniania czuwało dwóch koordynatorów. Procedury oceniania zostały oparte na procedurach stosowanych w trakcie sesji egzaminacyjnej, a egzaminatorami były osoby posiadające uprawnienia do oceniania prac w sesjach egzaminacyjnych.

W przededniu oceniania odbyło się szkolenie przewodniczących zespołów oceniających (PZE). W szkoleniu uczestniczyli zarówno koordynatorzy oceniania, osoby odpowiedzialne za realizację badania w terenie, jak i przedstawiciele Pracowni Analiz Osiągnięć Uczniów. Pierwszego dnia oceniania przewodniczący zespołów na podstawie wcześniejszego szkolenia przygotowali egzaminatorów do procesu oceniania.

Egzaminatorzy swoje oceny kodowali na kartach oceniania dołączonych do prac uczniów. Po zakończeniu oceniania karty zostały oddzielone od prac i zeskanowane (razem z odpowiedziami na zadania zamknięte zakodowanymi przez uczniów). Po ręcznym uzupełnieniu danych, w przypadku błędów odczytu skanera, wyniki uczniów zostały umieszczone w bazie danych.

Podczas oceniania zastosowano procedurę podwójnego oceniania minimum 10% prac z każdej szkoły. Dzięki takiemu schematowi podwójnie ocenionych zostało łącznie 1106 prac, co stanowi 12% wszystkich prac.

3.1.5. Progi realizacji

Założony próg realizacji badania w szkole wynosił 85% (uczniów niewykluczonych z badania). W przypadku, gdy podczas głównego terminu badania w szkole nie uzyskiwano wymaganego progu, przeprowadzano w szkole badanie uzupełniające. We wszystkich badanych szkołach udało się osiągnąć wymagany próg 85%. W 213 szkołach udało się uzyskać prace od 100% uczniów zakwalifikowanych do badania – stanowi to prawie połowę wszystkich przebadanych szkół. Realizacja badania w szkołach zakończyła się na bardzo wysokim poziomie – średnio w szkołach uzyskano odpowiedzi od 96% uczniów zakwalifikowanych do badania. Badania uzupełniające przeprowadzono w 23 szkołach i uzyskano w nich odpowiedzi od 74 uczniów.

3.1.6. Kontrola badań terenowych

W losowo wybranych szkołach pracownicy terenowi IBE przeprowadzili kontrolę zgodności badań z procedurami opracowanymi przez Pracownię Analiz Osiągnięć Uczniów IBE. W trakcie kontroli nie odnotowano rażącego odstępstwa od przyjętych procedur. Zaobserwowano jedynie niewielkie niedopatrzenia nie mające wpływu na poprawność zebranego materiału badawczego (głównie związane z nieutrzymywaniem dyscypliny wśród uczniów). W niektórych przypadkach uczniowie pisali testy w salach nie odpowiadających wytycznym IBE, lecz było to spowodowane skromnymi warunkami lokalowymi poszczególnych szkół. Warto podkreślić, że w tych szkołach właściwy egzamin odbywa się w takich samych warunkach.

3.2. Gimnazja

3.2.1. Rekrutacja

Rekrutacja szkół do badań w gimnazjach przeprowadzana była w taki sam sposób, jak rekrutacja w szkołach podstawowych i składała się z takich samych etapów. Analogicznie do szkół podstawowych, dyrektorzy gimnazjów otrzymali imienny list z informacją o badaniu i prośbą o wzięcie w nim udziału.

W szkołach gimnazjalnych również zdarzały się odmowy, a ich powody nie były odmienne od powodów odmów w szkołach podstawowych. Ostatecznie w badaniu wzięło udział 81 gimnazjów, przy czym zgodę na realizację badania spośród szkół z próby podstawowej wyraziło 55 placówek, co stanowi 69% tej próby. Z pierwszej próby rezerwowej pochodziło 21 szkół na zastępstwo, a do drugiej próby rezerwowej sięgnięto po zastępstwa dla 4 szkół. Ponadto przebadano jedną, dodatkową szkołę, z pierwszej próby rezerwowej z uwagi na nadmiarową rekrutację. Stosunkowo dużo szkół, z którymi się kontaktowano nie spełniało kryteriów wzięcia udziału w badaniu – łącznie było 6 takich placówek. Ponadto w trakcie rekrutacji okazało się, że jedna ze szkół jest w stanie likwidacji.

3.2.2. Badanie pilotażowe

Pilotaż badania w gimnazjum przebiegał w tym samym czasie, co pilotaż w szkołach podstawowych, a zatem w dniach 9-12 stycznia 2012 roku. Podobnie jak w przypadku szkół podstawowych badania przeprowadzono w województwach: kujawsko-pomorskim, pomorskim i małopolskim w celowo do tego wybranych szkołach. Pilotaż objął 8 gimnazjów i trwał dwa dni – pierwszego dnia uczniowie brali udział w testach z części humanistycznej (j. polski i historia z WOS-em), a drugiego w testach z części matematyczno-przyrodniczej (matematyka i przyroda).

Pilotaż w gimnazjach nie ujawnił błędów w procedurach przewidzianych w badaniach, które miały zapewnić warunki, jak najbardziej zbliżone do warunków na egzaminie właściwym. Organizacja pracy przebiegała sprawnie, a oszacowana liczba zadań okazała się właściwa w stosunku do czasu zaplanowanego na badania. Warto podkreślić, że podobnie jak w przypadku szkół podstawowych, badania zostały zaplanowane w taki sposób, aby uczniowie mogli je traktować, jako egzamin próbny, co miało szczególne znaczenie dla zwiększenia motywacji w trakcie rozwiązywania poszczególnych zadań.

3.2.3. Szkolenie koordynatorów i ankieterów

Szkolenie wszystkich ankieterów odbywało się wspólnie, według schematu opisanego w części dotyczącej szkół podstawowych. Procedury badania w obydwu typach szkół były bardzo podobne i zgodne z procedurami obowiązującymi podczas właściwej sesji egzaminacyjnej. Różnice wynikały z organizacji sesji badawczej. W gimnazjach badania odbywały się podczas kolejnych dwóch dni, podobnie, jak to ma miejsce podczas gimnazjalnej sesji egzaminacyjnej. Ankieterów szkolono do udziału w badaniach w obydwu typach szkół. Wątpliwości zgłaszane przez ankieterów podczas szkolenia dotyczyły elementów procedur stosowanych do badania niezależnie od typu szkoły. Jak już wspomniano w podrozdziale dotyczącym szkół podstawowych, w szkoleniu uczestniczyli przedstawiciele PAOU udzielając odpowiedzi na ewentualne pytania. Podsumowanie szkolenia zorganizowano wspólnie, zarówno dla szkoły podstawowej, jak i gimnazjum. W trakcie podsumowania szczególną uwagę zwrócono na zapewnienie podczas badań warunków zgodnych z tymi, jakie są stosowane podczas właściwej sesji egzaminacyjnej.

3.2.4. Realizacja badań w szkołach i ocenianie

Badanie w gimnazjach przeprowadzono zgodnie z harmonogramem, w dwóch następujących po sobie dniach pomiędzy 5 i 15 marca 2012, czyli ponad miesiąc przed egzaminem gimnazjalnym. Łącznie zebrano: 1688 prac uczniów z historii i WOS wraz z ankietami z tego przedmiotu, 1689 prac uczniów z języka polskiego wraz z ankietami z tego przedmiotu, 1676 prac uczniów z przyrody wraz z ankietami z tego przedmiotu, 1679 prac uczniów z matematyki wraz z ankietami z tego przedmiotu oraz 631 ankiet nauczycieli. W Tabeli 3.2 zawarto zestawienie wypełnionych zeszytów testowych w zależności od ich wersji dla każdego z testów.

Tabela 3.2. Podsumowanie wypełnionych zeszytów testowych w zależności od wersji

	Wersja A		**Wersja B**	
	Liczebność	Procent całości	Liczebność	Procent całości
Historia i WOS	860	13%	828	12%
Język polski	860	13%	829	12%
Przyroda	856	13%	820	12%
Matematyka	863	13%	816	12%
Razem	3439	51%	3293	49%

W ocenianiu prac, które miało miejsce w dniach 17-18 marca 2012, brało udział 44 egzaminatorów: 21 polonistów i 23 matematyków. Oceniano jedynie prace z języka polskiego i matematyki, gdyż arkusze z historii i wiedzy o społeczeństwie oraz z przedmiotów przyrodniczych zawierały jedynie zadania zamknięte samodzielnie kodowane przez uczniów. Egzaminatorzy podzieleni byli na cztery zespoły (po dwa dla każdego rodzaju arkusza) kierowane przez PZE i wspomagane przez asystentów wypełniających zadania organizacyjno-techniczne. Koordynatorzy oceniania i przedstawiciele PAOU monitorowali ocenianie szczególnie pod względem zgodności z zasadami stosowanymi przez okręgowe komisje egzaminacyjne w trakcie sesji oceniania.

Egzaminatorzy swoje oceny kodowali na kartach oceniania dołączonych do prac uczniów. Po zakończeniu oceniania karty zostały oddzielone od prac i zeskanowane (razem z odpowiedziami na zadania zamknięte zakodowanymi przez uczniów). Po ręcznym uzupełnieniu danych w przypadku błędów odczytu skanera, wyniki uczniów zostały umieszczone w bazie danych.

Podczas oceniania zastosowano procedurę podwójnego oceniania minimum 10% prac z każdej szkoły. Dzięki takiemu schematowi podwójnie ocenionych zostało łącznie 260 prac z języka polskiego i 202 prace z matematyki, co stanowi odpowiednio 12% wszystkich prac danego typu. Wyniki podwójnego oceniania zostały zapisane w osobnych plikach.

3.2.5. Progi realizacji

Założony próg realizacji badania w gimnazjach wynosił 85% (uczniów niewykluczonych z badania) dla każdej części testu osobno. Nie było jednak wymagane, aby dany uczeń wypełnił wszystkie rodzaje testu dla niego przewidziane – stopy realizacji obliczane były niezależnie dla wszystkich części testu.

W przypadku, gdy podczas głównego terminu badania w szkole nie uzyskiwano wymaganego progu, przeprowadzano w szkole badanie uzupełniające.

We wszystkich badanych szkołach udało się osiągnąć wymagany próg 85% z każdego testu. W 26 szkołach udało się uzyskać prace od 100% uczniów zakwalifikowanych do badania – stanowi to blisko jedną trzecią wszystkich przebadanych szkół. Realizacja badania w szkołach zakończyła się na bardzo wysokim poziomie – średnio w szkołach uzyskano odpowiedzi od 95-96% uczniów zakwalifikowanych do badania w zależności od części testu. Badania uzupełniające przeprowadzono w 6 szkołach, przy czym w jednej ze szkół konieczna była sesja uzupełniająca zarówno z części humanistycznej, jak i z części matematyczno-przyrodniczej testu. Badania uzupełniające objęły 2-5 uczniów w każdej ze szkół.

3.2.6. Kontrola badań terenowych

Podobnie jak w przypadku badań w szkołach podstawowych także i w gimnazjach podczas kontroli badań terenowych nie odnotowano odstępstw od procedur, które mogłyby mieć wpływ na poprawność zebranego materiału badawczego. Pojawiające się problemy to zapewnienie dyscypliny wśród uczniów, czy też warunki lokalowe (zbyt małe sale) poszczególnych szkół, co było poniżej standardów określonych przez IBE, ale zgodne z warunkami, w których w tych szkołach przeprowadzany jest egzamin właściwy.

3.3. Podwójne ocenianie losowej próby prac

W trakcie ocenia prac z badań zrównujących zachowano organizację zgodną ze stosowaną przez okręgowe komisje egzaminacyjne podczas sesji egzaminacyjnych. Ocenianie składało się z trzech etapów.

1. Spotkanie koordynatora oceniania i przewodniczących zespołów egzaminatorów w dniu poprzedzającym ocenianie.

2. Szkolenie egzaminatorów do oceniania zadań z poszczególnych zeszytów testowych zastosowanych podczas badań.

3. Ocenianie prac.

Podobnie jak na egzaminie właściwym w studium zrównującym 10 procent prac oceniono podwójnie. Do oceny przez drugiego egzaminatora została wylosowana próba prac, która została skopiowana przed rozpoczęciem oceniania. W przypadku różnych ocen któregokolwiek z zadań otwartych ostateczną decyzję podejmował przewodniczący zespołu egzaminatorów. Przyznana przez niego punktacja była ostateczna.

Dla każdego zadania otwartego oszacowany został wskaźnik zgodności kappa. Współczynnik kappa pozwala ocenić stopień zgodności dwukrotnych pomiarów tej samej zmiennej. Współczynnik obliczany jest za pomocą wzoru (Cohen, 1960) :

$$kappa = \frac{P_o - P_e}{1 - P_e}$$

gdzie:

P_o – obserwowana zgodność kodowania,

P_e – oczekiwana zgodność kodowania, czyli zgodność, która powstałaby w przypadku losowego rozłożenia wartości skali punktowania danego zadania.

Kappa może przyjmować wartości od -1 do +1. Wartość 0 oznacza zgodność na takim poziomie jaki powstałby dla losowego rozkładu wartości na całej skali punktowania zadania. Szeroko przyjętą interpretację wskaźnika zgodności kappa przedstawiono w tabeli 3.3.

Tabela 3.3. Interpretacja wskaźnika zgodności (Bland & Altman, 1999)

Poziom zgodności	Wartość kappa	Interpretacja poziomu zgodności
1	poniżej 0	brak zgodności
2	0,00-0,02	niewielki
3	0,21-0,40	dostateczny
4	0,41-0,60	średni
5	0,61-0,81	znaczny
6	0,81-1,00	idealny

3.3.1. Zgodność kodowania w badaniu zrównującym wyniki sprawdzianu

W przypadku badań zrównujących wyniki sprawdzianu na zakończenie szkoły podstawowej procedurze podwójnego oceniania poddano 1106 prac. Dla 280 prac wystąpiły niezgodności w ocenie pierwszego i drugiego egzaminatora w przynajmniej jednym zadaniu.

W Tabeli 3.4 przedstawione zostały wyniki analizy zgodności oceniania zadań otwartych w badaniu zrównującym wyniki sprawdzianu. W tabeli została podana obserwowana i oczekiwana zgodność kodowania oraz obliczona na tej podstawie wartość współczynnika zgodności kappa. W tabeli podano błąd standardowy współczynnika i wartość statystyki testowej **z** pozwalającej stwierdzić czy współczynnik kappa jest statystycznie różny od zera. W tabeli podano również, w ilu pracach podwójnie ocenianych znalazło się dane zadanie oraz informację na temat maksymalnej liczby punktów w tym zadaniu.

Wśród wszystkich zadań wykorzystanych w badaniu zrównującym znalazło się 29 zadań otwartych. W przypadku większości zadań osiągnięto wysoką zgodność kodowania: w 6 zadaniach współczynnik zgodności kappa jest wyższy niż 0,80, co oznacza idealną zgodność, tylko w przypadku 5 zadań współczynnik kappa jest niższy niż 0,80, jednak wyższy niż 0,60, co można interpretować jako znaczną zgodność.

W ostatniej kolumnie została podana maksymalna liczba punktów, jaką można było uzyskać w danym zadaniu. Należy zwrócić uwagę na zgodność kodowania w zadaniach wielopunktowych. Także w przypadku tych zadań osiągnięto wysoką zgodność oceniania pierwszego i drugiego egzaminatora.

Tabela 3.4. Oszacowany wskaźnik zgodności kappa dla zadań otwartych zastosowanych podczas badań zrównujących wyniki sprawdzianu[19]

Nr zadania	Etykieta zadania	Zgodność		kappa	Błąd standardowy	z	Liczba prac z danym zadaniem	Maksymalna liczba punktów w danym zadaniu
		obserwowana	oczekiwana					
1	sp02z22	97,98%	59,43%	0,950	0,101	9,45	99	1
2	sp02z23_s	90,72%	24,68%	0,877	0,053	16,64	97	5
3	sp03z22_s	91,46%	27,26%	0,883	0,031	28,49	281	5
4	mc2z40	94,86%	35,57%	0,920	0,032	28,92	507	2
5	ht2z37	82,78%	51,27%	0,647	0,060	10,70	273	1
6	ht1z29	97,45%	58,32%	0,939	0,053	17,64	354	1
7	mc1z36	100,00%	47,01%	1,000	0,054	18,51	245	2
8	sp06z22	98,79%	46,26%	0,977	0,059	16,65	165	2
9	sp06z23	100,00%	53,75%	1,000	0,075	13,38	179	1
10	sp07z22_s	94,53%	25,30%	0,927	0,032	28,70	275	4
11	sp07z23	99,66%	37,03%	0,995	0,042	23,62	296	2

[19] Wszystkie wyniki istotne statystycznie na poziomie 0,01.

12	c1z28	98,59%	52,53%	0,970	0,119	8,18	71	1
13	sp08z23_s	92,86%	22,12%	0,908	0,037	24,34	197	4
14	sp08z24_s	76,02%	36,48%	0,623	0,048	13,10	197	3
15	c3z29	92,52%	63,96%	0,792	0,058	13,71	294	1
16	sp09z22_s	98,56%	35,12%	0,978	0,050	19,70	208	2
17	sp09z23_s	96,15%	32,80%	0,943	0,045	20,82	182	3
18	c4z39	97,83%	45,32%	0,960	0,083	11,56	92	2
19	mc2z40	94,86%	35,57%	0,920	0,032	28,92	507	2
20	sp10z21_s	97,74%	42,56%	0,961	0,059	16,29	177	2
21	ht1z29	97,45%	58,32%	0,939	0,053	17,64	354	1
22	sp11z22	96,34%	36,39%	0,942	0,044	21,69	273	2
23	sp11z23	99,25%	55,35%	0,983	0,048	20,71	268	2
24	sp11z26_s	89,16%	32,95%	0,838	0,037	22,49	286	3
25	sp03z22_s	91,46%	27,26%	0,883	0,031	28,49	281	5
26	sp03z25_s	87,72%	41,54%	0,790	0,048	16,63	285	2
27	sp04z23_s	96,83%	40,69%	0,947	0,055	17,37	189	2

28	sp04z24_s	94,27%	27,37%	0,921	0,042	21,81	157	5
29	sp05z25_s	81,55%	28,68%	0,741	0,043	17,08	168	6

3.3.2. Zgodność kodowania w badaniu zrównującym wyniki egzaminu gimnazjalnego

W procesie zrównywania wyników egzaminu gimnazjalnego procedurze podwójnego oceniania poddano 293 prace uczniów biorących udział w badaniu. W przypadku 44 prac wystąpiły niezgodności między oceną pierwszego i drugiego egzaminatora w przynajmniej jednym zadaniu.

W Tabeli 3.5 przedstawione zostały wyniki analizy zgodności oceniania zadań otwartych w badaniu zrównującym wyniki egzaminu gimnazjalnego. Każde zadanie ocenione zostało przez 2 różnych egzaminatorów. W tabeli została podana obserwowana i oczekiwana zgodność kodowania oraz obliczona na tej podstawie wartość współczynnika zgodności kappa. W tabeli podano także błąd standardowy współczynnika i wartość statystyki testowej pozwalającej stwierdzić czy współczynnik kappa jest statystycznie różny od zera. W tabeli podano również, w ilu pracach podwójnie ocenianych znalazło się dane zadanie oraz informację na temat maksymalnej liczby punktów w tym zadaniu.

Wśród wszystkich zadań wykorzystanych w badaniu zrównującym wyniki egzaminu gimnazjalnego znalazło się 11 zadań otwartych, w tym 3 zadania z zakresu przedmiotów matematyczno-przyrodniczych i 8 zadań humanistycznych.

We wszystkich zadaniach wartość współczynnika kappa jest wyższa niż 0,80, co oznacza bardzo wysoką zgodność kodowania. W przypadku 5 zadań osiągnięto 100% zgodność kodowania.

Tabela 3.5. Oszacowany wskaźnik zgodności kappa dla zadań otwartych zastosowanych podczas badań zrównujących wyniki egzaminu gimnazjalnego[20]

Zeszyt testowy	Nr zadania	Etykieta zadania	Zgodność		kappa	Błąd standardowy	z	Liczba prac z danym zadaniem	Maksymalna liczba punktów w danym zadaniu
			obserwowana	oczekiwana					
matematyka	1	mc2z41	100,00%	43,65%	1,000	0,060	16,61	174	3
	2	mc1z34_s	91,72%	28,45%	0,884	0,045	19,73	157	4
	3	mc2z40	100,00%	39,48%	1,000	0,061	16,30	164	2
Język polski	1	ht1z14	99,51%	53,63%	0,989	0,070	14,13	204	1
	2	ht1z15	93,65%	52,45%	0,867	0,072	11,99	189	1
	3	ht1z16	96,25%	42,94%	0,934	0,063	14,93	160	2
	4	ht2z04	97,87%	73,16%	0,921	0,073	12,66	188	1
	5	ht1z34	98,99%	50,73%	0,980	0,071	13,79	198	1
	6	ht2z37	96,48%	57,26%	0,918	0,071	12,97	199	1
	7	h11z27	97,96%	42,85%	0,964	0,062	15,47	148	2

[20] Wszystkie wyniki istotne statystycznie na poziomie 0,05.

8	ht1z24	99,39%	50,47%	0,988	0,078	12,69	165	1
9	ht2z14	100,00%	50,04%	1,000	0,075	13,38	179	1
10	ht2z15	100,00%	57,40%	1,000	0,074	13,49	182	1
11	ht2z16	97,42%	51,99%	0,946	0,080	11,78	155	1

4. Statystyczna koncepcja zrównywania

4.1. Wstęp

Rozdział dotyczy metodologicznych oraz statystycznych aspektów badania zrównującego wyniki egzaminów gimnazjalnych z roku 2011 oraz badania zrównującego wyniki sprawdzianu po klasie 6 z lat 2002-2011. Ponieważ raport ze zrównania egzaminu gimnazjalnego 2002-2010 (Szaleniec et al., 2011) zawierał bardziej pogłębiony przegląd różnych (i) planów zrównywania wyników, (ii) metod zrównywania wyników oraz (iii) ogólny opis modeli IRT, w niniejszym rozdziale kwestie te zostały potraktowane jedynie zdawkowo. Nacisk położono przede wszystkim na przedstawienie specyficznych rozwiązań zastosowanych podczas badań zrównujących przeprowadzonych w IBE. Czytelnik zainteresowany bardziej ogólnym omówieniem problemów związanych ze zrównywaniem może sięgnąć do wcześniejszego raportu.

Na wstępie przedstawiono zastosowane plany zbierania danych do badań zrównujących. Mimo iż zaimplementowane plany zrównywania omówiono już na początku raportu, ze względu na ich bezpośredni związek z zastosowanym modelowaniem statystycznym zdecydowano się na ich powtórzenie w tym miejscu. Następnie omówiono, w jaki sposób do zgromadzonych danych dopasowano model IRT oraz jak korzystając z wyrażonych na wspólnej skali parametrów modelu IRT dokonujemy zrównania wyników obserwowanych. Poruszane są też takie tematy jak szacowanie błędu zrównywania, szacowanie rozkładu umiejętności z wykorzystaniem tzw. *plausible values* czy wykorzystane oprogramowanie statystyczne. W szczególności problem szacowania błędu zrównywania został rozwiązany w nieco odmienny sposób niż pierwotnie, tj. niż przy zrównywaniu wyników egzaminu gimnazjalnego z lat 2002-2010.

4.2. Plan nierównoważnych grup z testem kotwiczącym i zrównywanie z wykorzystaniem IRT

Modelowym planem zbierania danych do zrównywania wyników w przypadku różnych populacji uczniów rozwiązujących różne testy jest plan nierównoważnych grup z testem kotwiczącym (ang. *nonequivalent groups with anchor test design*, NEAT). W planie tym mamy dwie różne populacje uczniów $\mathcal{P}$ oraz $\mathcal{Q}$, z których uczniowie rozwiązują dwa różne testy, odpowiednio X oraz Y. Dodatkowo plan też uwzględnia trzeci, dodatkowy podzbiór zadań/test A noszący nazwę kotwicy, który jest rozwiązywany przez uczniów z obu populacji. Schematycznie plan NEAT można przedstawić w następujący sposób:

Populacja	Próba	X	Y	A
$\mathcal{P}$	S_1	✓		✓
$\mathcal{Q}$	S_2		✓	✓

Plan NEAT można formalnie podzielić w zależności od tego czy zbiór zadań wchodzących w skład testu A jest odrębnym testem od X oraz Y (kotwica zewnętrzna – ang. *external anchor*), czy też A stanowi podzbiór zadań testów X oraz Y uwzględniany przy obliczaniu wyników w tych testach (kotwica wewnętrzna – ang. *internal anchor*). Bez zastosowania dodatkowego testu A rozwiązywanego zarówno przez próbę uczniów z populacji $\mathcal{P}$ jak i $\mathcal{Q}$ nie byłoby możliwe porównanie

wyników różnych grup uczniów jakie uzyskują w różnych testach X oraz Y – nie sposób byłoby rozróżnić trudności testów X oraz Y od różnic w poziomie umiejętności między uczniami z $\mathcal{P}$ oraz $\mathcal{Q}$.

Zadanie ustalenia porównywalności wyników z różnych testów mierzących tę samą umiejętność stanowi problem zrównywania wyników testowych (czy ogólniej – linkowania/łączenia wyników testowych). Teoretyczne i metodologiczne niuanse dotyczące problemu zrównywania testów można znaleźć u wielu autorów, na przykład: Lord (1980), Kolen i Brennan (2004), Dorans & Holland (2000) czy Holland et al. (2007), a także w raporcie ze zrównywania wyników egzaminu gimnazjalnego 2002-2010 (Szaleniec et al., 2011). Tu ograniczymy się jedynie do ogólnej typologii metod zrównywania wyników ze szczególnym naciskiem na metody wykorzystujące modelowanie IRT.

Na najogólniejszym poziomie, podziału metod zrównywania wyników testowych można dokonać w zależności od tego czy:

a) zrównywanie odbywa się na skali wyników obserwowanych, czy wyników prawdziwych;
b) zrównywanie odbywa się z bezpośrednim odwołaniem do modelu pomiarowego, czy nie.

Większość technik wykorzystywanych do zrównywania wyników testowych wpada w kategorię zrównywania wyników obserwowanych (ang. *observed score equating*), gdzie przez „wynik obserwowany" rozumie się klasyczny sumaryczny wynik w teście. Nacisk na przeprowadzanie zrównywania na poziomie wyników obserwowanych jest konsekwencją tego, że w przeważającej większości takie właśnie wyniki są wykorzystywane do raportowania rezultatów testowania. Zrównywanie wyników obserwowanych może zostać przeprowadzone bez konieczności odwoływania się w modelu statystycznym do sparametryzowanego mechanizmu leżącego u podstaw obserwowanych wyników, ale także z wykorzystaniem takiego modelu, tj. z wykorzystaniem IRT (ang. *IRT observed score equating*).

W obrębie podejścia opartego na modelach IRT, pojawia się możliwość dokonania zrównywania wyników prawdziwych (ang. *IRT true score equating*). Przez „wynik prawdziwy" danego ucznia rozumie się tu wartość oczekiwaną z wyniku obserwowanego tego ucznia. Aby zrównanie zostało przeprowadzone na skali wyników prawdziwych klasycznej teorii testów, konieczne jest oszacowanie parametrów modelu pomiarowego leżącego u podstaw obserwowanych odpowiedzi. Zrównywania na skali wyników prawdziwych nie można zatem przeprowadzić „ateoretycznie" jak w przypadku wyników obserwowanych. Omawiane zależności między metodami zrównywania wyników schematycznie przedstawiono w Tabeli 4.1.

Tabela 4.1. Schematyczny podział metod zrównywania wyników testowych

	Zrównywanie wyników obserwowanych	**Zrównywanie wyników prawdziwych**
Metody niezależne od modelu pomiarowego	*(non-IRT) observed score equating*	-
Metody oparte na modelu pomiarowym	*IRT observed score equating*	*IRT true score equating*

Korzystanie przy zrównywaniu wyników testowych z modelu pomiarowego IRT nakłada konieczność spełnienia wymagania, że zastosowany model IRT poprawnie opisuje udzielane przez uczniów

odpowiedzi na zadania zrównywanych testów. Jednowymiarowy model IRT stanowi, że dla każdego ucznia j prawdopodobieństwo udzielenia konkretnego wektora odpowiedzi $\boldsymbol{u}_j$ da się sparametryzować w następujący sposób:

$$\mathbb{P}(\boldsymbol{U}_j = \boldsymbol{u}_j|\theta_j) = \prod_i \mathbb{P}(U_{i,j} = u_{i,j}|\theta_j) \tag{4.1a}$$

$$\mathbb{P}(U_{i,j} = u_{i,j}|\theta_j) = f_i(u_{i,j}|\boldsymbol{\beta}_i, \theta_j) \tag{4.1b}$$

gdzie θ_j jest parametrem określającym pozycję ucznia j na zmiennej ukrytej θ (umiejętność, ang. *ability*), a $\boldsymbol{\beta}_i$ jest wektorem parametrów określających kształt f_i – funkcji charakterystycznej zadania i (ang. *item characteristic function*, ICC). IRT zatem wprost postuluje rozdzielenie parametrów określających właściwości ucznia od parametrów określających właściwości zadań/testu w celu opisania mechanizmu udzielania odpowiedzi na zadania. Rozdzielenie parametrów zadań od parametrów uczniów stanowi klucz do zrozumienia popularności metod IRT przy zrównywaniu wyników.

Założenie (4.1a) założeniem o lokalnej homogeniczności pomiarów (ang. *local homogeneity*, Ellis & van der Woldenberg, 1993) i oznacza, że cała informacja o współzmienności zadań zawarta jest w θ – ustaliwszy wartość parametru umiejętności ucznia θ, odpowiedzi na zadania stają się zdarzeniami niezależnymi. Założenie (4.1b) natomiast określa parametryczną postać zależności pomiędzy wartościami θ a prawdopodobieństwem określonych odpowiedzi w zadaniu $U_{i,j}$. Od spełnienia obu tych bardzo silnych założeń zależy jakość wniosków opartych na zastosowaniach IRT, w szczególności jakość wykorzystującego IRT zrównywania wyników testów. Z tego względu (Livingston, 2004; von Davier et al., 2004) nadal często preferowane są, omówione wcześniej, nieczyniące tak daleko idących założeń „ateoretyczne" metody zrównywania wyników. Metody weryfikacji spełnienia założeń IRT w kontekście zrównywania oraz analiza odporności (ang. *robustness*) metod zrównywania IRT na niespełnienie założeń (4.1) można znaleźć u: Béguin (2000) oraz Glas & Béguin (2011)..

Warto wspomnieć, że równania modelu IRT w (4.1) można łatwo uogólnić do postaci w której zmienna umiejętności jest wielowymiarowa (por.: Reckase, 2009). Zastosowanie modelu dwuwymiarowego teoretycznie byłoby z korzyścią dla precyzji zrównywania w przypadku egzaminu gimnazjalnego, w którym mamy dane tych samych uczniów z testów mierzących dwie różne umiejętności. Z rozwiązania dwuwymiarowego zrezygnowano ze względu na obliczeniowe skomplikowanie problemu zrównywania wielowymiarowego, które i tak w wariancie jednowymiarowym wiązało się z wieloma trudnościami natury praktycznej, co zostanie bliżej omówione w dalszej części rozdziału. Potencjalna możliwość zrównania egzaminu gimnazjalnego poprzez wykorzystanie dwuwymiarowego modelu IRT stanowi jednak interesującą alternatywę, która może stać się polem dla dalszych badań. Informacje o zrównywaniu wyników z wykorzystaniem modeli MIRT można znaleźć u: Reckase (2009), Béguin (2000), Li & Lissitz (2000) oraz Yao & Boughton (2009). W dalszej części przyjęto

założenie, że zrównywane w planie NEAT testy X, Y oraz kotwica A są jednowymiarowe[21] i mierzą ten sam pojedynczy konstrukt (Davier & Davier, 2011).

Zmienna umiejętności θ jest w modelu IRT zmienną losową, tzn. umiejętność ucznia θ_j pojawiająca się we wzorze (4.1) nie jest parametrem modelu podlegającym bezpośredniej estymacji, a jedynie losowym efektem zmiennej θ o określonym rozkładzie. To właśnie parametry rozkładu zmiennej umiejętności, oznaczmy go ψ, są estymowane w modelu IRT. Jest to bardzo ważne w kontekście zrównywania – modelowanie rozkładu umiejętności oddzielne dla każdej populacji jest istotą całej procedury, gdyż dąży się do rozdzielenia informacji o poziomie umiejętności różnych populacji od informacji o właściwościach testów. W związku z czym model IRT przedstawiono również w formie bezwarunkowej, tj. scałkowanej po rozkładzie umiejętności w populacji $\mathcal{P}$:

$$P(\boldsymbol{U} = \boldsymbol{u}|\mathcal{P}) = \int f(\boldsymbol{u}, \theta, \boldsymbol{\beta})\psi_{\mathcal{P}}(\theta)\, d\theta.$$

Modele IRT poprzez uwzględnienie wprost parametrów rozkładu populacji oraz bezproblemowe radzenie sobie z niekompletnymi schematami zbierania danych stanowią bardzo dobre narzędzie do rozwiązania problemu zrównywania wyników testów. W obrębie IRT wypracowano wiele metod do umieszczania na wspólnej skali rozkładów θ dla populacji $\mathcal{P}$ oraz $\mathcal{Q}$ oraz parametrów $\boldsymbol{\beta}$ dla testów X, Y oraz A. Wyróżnić można następujące metody (Kolen & Brennan, 2004) :

1) Łączna kalibracja (ang. *concurrent calibration*) wszystkich trzech testów.

2) Oddzielna kalibracja (ang. *separate calibration*), par testów (X, A) oraz (Y, A), po której stosuje się sprowadzające do wspólnej skali przekształcenia oparte na:

   a) liniowej funkcji parametrów kotwicy - metody średnia/średnia lub średnia/sigma (ang. *mean/mean, mean/sigma*);

   b) krzywych charakterystycznych kotwicy - metoda Stockinga-Lorda lub Haebary.

3) Metoda ustalonych parametrów (ang. *fixed parameters method*) dla kotwicy A.

4) Metoda przekształcania umiejętności (ang. *proficiency transformation*).

W zrównywaniu wyników egzaminu gimnazjalnego oraz sprawdzianu zastosowano metodę łącznej kalibracji. Uczyniono tak ze względu na skomplikowanie schematu zbierania danych (Tabele 2.1, 2.2 i 2.2a) oraz ze względu na szacowanie parametrów rozkładu umiejętności w różnych populacjach w tej metodzie w sposób bezpośredni. Łączna kalibracja dla planu NEAT polega na dopasowaniu do wszystkich danych zebranych z prób S_1 oraz S_2 modelu IRT w pojedynczej procedurze estymacji. Strukturalny brak odpowiedzi na zadania testu Y w próbie S_1 oraz brak odpowiedzi na zadania testu X w próbie S_2 nie stanowi problemu dla maksymalizujących funkcję wiarygodności metod estymacji wykorzystywanych do szacowania parametrów modeli IRT. W celu uzyskania nieobciążonych

[21] Dyskusję o ścisłej zależności pomiędzy założeniem (4.1a), a pojęciem „wymiarowości" testu można znaleźć w przełomowej monografii Lorda i Novicka (1968).

parametrów w planie NEAT, konieczne jest bezpośrednie uwzględnienie w modelu IRT odrębnych parametrów dla rozkładów θ dla populacji $\mathcal{P}$ oraz $\mathcal{Q}$ i oszacowanie ich z prób S_1 oraz S_2.

4.3. Implementacja metody łącznej kalibracji modelu IRT do zrównania egzaminów gimnazjalnych i sprawdzianu

W poprzednich akapitach opisano metodę łącznej kalibracji w zastosowaniu do planu NEAT. Logika postępowania w metodzie łącznej kalibracji przy bardziej skomplikowanym planie zrównywania, w skład którego wchodzi więcej populacji oraz więcej różnych testów, jest analogiczna – w pojedynczym kroku estymujemy parametry zadań z wszystkich testów oraz parametry rozkładu umiejętności dla każdej populacji. Niestety dla zastosowanego przy zrównywaniu sprawdzianu oraz gimnazjum planu zrównywania (Tabele 2.1 oraz 2.2) rozmiary macierzy danych przerosły możliwości obliczeniowe dostępnego sprzętu oraz oprogramowania.

W celu przezwyciężenia problemu rozległości danych, zamiast korzystania z całego zbioru danych, zdecydowano się na zrównanie wyników egzaminów wykorzystując podpróby 2000 uczniów piszących egzaminy w latach 2002-2011 (zarówno dla sprawdzianu jak i dla egzaminu gimnazjalnego). Liczba 2000 uczniów wynikała z liczebności przypadających na zadania egzaminacyjne wykorzystywane w poszczególnych próbach badawczych – wynosiły one od około 800 do około 3200, najczęściej około 1600 (Tabele 2.1 oraz 2.2). Aby: (i) wykorzystać większą część zbioru danych egzaminacyjnych niż 2000 prac losowanych przy pojedynczym zrównywaniu (ii) móc oszacować błąd zrównania wynikający z doboru próby, procedurę powtórzono R=500 razy.

W celu obliczenia parametrów rozkładu umiejętności dla populacji $\mathcal{P}_{02}$, $\mathcal{P}_{03}$, … , $\mathcal{P}_{11}$ 500-krotnie powtórzono następujący algorytm:

1) wylosowanie z każdej z populacji uczniów piszących egzamin $\mathcal{P}_{02}$, $\mathcal{P}_{03}$, … , $\mathcal{P}_{11}$ podpróby 2000 uczniów;

2) dołączenie danych z sesji zrównującej (sesji zrównujących);

3) dopasowanie do takiej podpróby danych modelu IRT, *explicite* szacującego średnią i odchylenie standardowe rozkładu umiejętności w każdej z populacji $\mathcal{P}_{02}$, $\mathcal{P}_{03}$, … , $\mathcal{P}_{11}$.

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

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

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

Opisana procedura próbkowania danych jest nieznacznie odmienna od tej jaką zastosowano przy zrównywaniu wyników egzaminu gimnazjalnego na podstawie badań przeprowadzonych w 2011 roku (Szaleniec et al., 2011). We wcześniejszych badaniach w kroku 2) zamiast danych z sesji zrównującej wzięto ich próbę ze zwracaniem (próba *bootstrap*) i błąd estymatorów $\widehat{\mu_{\theta|\mathcal{P}_\cdot}}$ i $\widehat{\sigma_{\theta|\mathcal{P}_\cdot}}$ obliczano na podstawie wariancji ${}^{r}\widehat{\mu_{\theta|\mathcal{P}_\cdot}}$ oraz ${}^{r}\widehat{\sigma_{\theta|\mathcal{P}_\cdot}}$ zaobserwowanej na przestrzeni replikacji. W bieżącym zrównaniu zdecydowano się na zastosowanie dwuetapowego próbkowania danych w celu obliczenia błędów estymatorów. Zastosowaną procedurę oraz uzasadnienie przedstawiono poniżej.

Dla rozróżnienia dwóch kroków zastosowanych w celu oszacowania błędu parametrów $\mu_{\theta|\mathcal{P}_\cdot}$ oraz $\sigma_{\theta|\mathcal{P}_\cdot}$ wprowadzono dodatkową indeksację:

- „A”, na oznaczenie kroku, w którym obliczane są oszacowania $\mu_{\theta|\mathcal{P}_\cdot}$ oraz $\sigma_{\theta|\mathcal{P}_\cdot}$ dokładnie tak jak to opisano powyżej, tj. bez brania próby *bootstrap*;
- „B” na oznaczenie kroku, w którym uwzględniana jest dodatkowa wariancja wynikająca z doboru uczniów do próby w grupie badawczej (grupach badawczych), tj. brana jest próba *bootstrap* danych z sesji zrównującej.

Przy tych oznaczeniach krok „A” jest opisanym wcześniej, powtórzonym 500 razy trzyetapowym algorytmem:

1) wylosowanie z każdej z populacji uczniów piszących egzamin $\mathcal{P}_{02}$, $\mathcal{P}_{03}$, … , $\mathcal{P}_{11}$ podpróby 2000 uczniów;
2) dołączenie danych z sesji zrównującej (sesji zrównujących);
3) dopasowanie do takiej podpróby danych modelu IRT w celu uzyskania oszacowań $\overset{r}{A}\widehat{\mu_{\theta|\mathcal{P}_\cdot}}$ oraz $\overset{r}{A}\widehat{\sigma_{\theta|\mathcal{P}_\cdot}}$

Natomiast krok „B” przebiega 500 razy w następujący sposób:

1) z każdej z populacji uczniów piszących egzamin $\mathcal{P}_{02}$, $\mathcal{P}_{03}$, … , $\mathcal{P}_{11}$ wzięto podpróbę 2000 uczniów wylosowanych w tej samej replikacji w kroku A;
2) dołączenie próby ze zwracaniem (próba *bootstrap*) z danych z sesji zrównującej (sesji zrównujących);
3) dopasowanie do takiej podpróby modelu IRT w celu uzyskania oszacowań $\overset{r}{B}\widehat{\mu_{\theta|\mathcal{P}_\cdot}}$ oraz $\overset{r}{B}\widehat{\sigma_{\theta|\mathcal{P}_\cdot}}$

W dalszej części skupiono uwagę na składnikach błędu występujących przy szacowaniu średniej rozkładu umiejętności (rozumowanie dla odchylenia standardowego/wariancji rozkładu umiejętności jest analogiczne). Kroki „A” oraz „B” różnią się jedynie tym, że w kroku „B” występuje dodatkowe źródło zmienności związane braniem próby *bootstrap* uczniów biorących udział w sesji zrównującej. Jeżeli przyjęto, że oszacowanie średniej dla populacji $\mathcal{P}_\cdot$ w replikacji r nie jest obciążone, to można zdekomponować te oszacowania w następujący sposób:

$$\overset{r}{A}\widehat{\mu_{\theta|\mathcal{P}_\cdot}} = \mu_{\theta|\mathcal{P}_\cdot} + {}^{r}E_{próba\_egz}$$

$$\overset{r}{B}\widehat{\mu_{\theta|\mathcal{P}_\cdot}} = \mu_{\theta|\mathcal{P}_\cdot} + {}^{r}E_{bootstrap} + {}^{r}E_{próba\_egz}$$

gdzie:

$\mu_{\theta|\mathcal{P}_\cdot}$ – prawdziwa wartość szacowanego parametru;

${}^{r}E_{próba\_egz}$ – losowy składnik błędu wynikający z losowania próby 2000 uczniów z egzaminu w danym roku zamiast wykorzystania wszystkich dostępnych danych;

${}^{r}E_{bootstrap}$ – losowy składnik błędu wynikający z wylosowania ograniczonej próby uczniów do sesji zrównującej.

Ponieważ w kroku „A” oraz „B” wykorzystano te same próby uczniów to możemy przyjąć, że składnik ${}^{r}E_{próba\_egz}$ dla danej replikacji jest w obu krokach taki sam, więc jeżeli odejmiemy od siebie oszacowania średnich w obu krokach:

$$= \left(\mu_{\theta|\mathcal{P}} + {}^{r}E_{bootstrap} + {}^{r}E_{próba\_egz}\right) - \left(\mu_{\theta|\mathcal{P}} + {}^{r}E_{próba\_egz}\right) = {}^{r}E_{bootstrap}$$

to uzyskujemy wyizolowany składnik błędu *bootstrap*.

Wariancję ${}^{r}E_{bootstrap}$ wykorzystano do obliczenia błędu standardowego estymatora $\widehat{\mu_{\theta|\mathcal{P}}}$: wynikającego z doboru uczniów do próby zrównującej:

$$SE_{bootstrap} = \left(\frac{\sum_{r=1}^{R}\left({}^{r}_{B}\widehat{\mu_{\theta|\mathcal{P}}} - {}^{r}_{A}\widehat{\mu_{\theta|\mathcal{P}}}\right)^2}{(R-1)}\right)^{\frac{1}{2}}$$

Pozostaje jeszcze składnik błędu ${}^{r}E_{próba\_egz}$, który można obliczyć z wariancji ${}^{r}_{A}\widehat{\mu_{\theta|\mathcal{P}}} = \mu_{\theta|\mathcal{P}} + {}^{r}E_{próba\_egz}$ po replikacjach. Jednakże błąd standardowy $SE_{próba\_egz}$, w odróżnieniu od $SE_{bootstrap}$, zmniejsza się wraz ze zwiększaniem się liczby replikacji – czym więcej replikacji wykonamy tym więcej informacji z prawdziwych egzaminów zostanie wykorzystanych w procedurze zrównywania. Zatem, podobnie jak w klasycznym wyrażeniu $SE = (var/N)^{\frac{1}{2}}$, mamy:

$$SE_{próba\_egz} = \left(\frac{\sum_{r=1}^{R}\left({}_{A}\widehat{\mu_{\theta|\mathcal{P}}} - {}^{r}_{A}\widehat{\mu_{\theta|\mathcal{P}}}\right)^2}{(R-1)R}\right)^{\frac{1}{2}}$$

Ostatecznie oba składniki błędu są dodawane w celu uzyskania całkowitego błędu oszacowania średniej poziomu umiejętności dla danej populacji:

$$SE_{\widehat{\mu_{\theta|\mathcal{P}}}} = \left(SE_{egz\_sampl}{}^{2} + SE_{bootstrap}{}^{2}\right)^{\frac{1}{2}}$$

Analogicznie postąpiono celem oszacowania błędu oszacowania odchylenia standardowego poziomu umiejętności.

4.4. Zrównywanie wyników obserwowanych

Zrównywanie wyników obserwowanych dwóch testów w najogólniejszej postaci przyjmuje formę tzw. zrównywania ekwicentylowego (ang. *equipercentile equating*). Idea zrównywania ekwicentylowego opiera się na fakcie, że dla ciągłych i ściśle rosnących dystrybuant F_X oraz F_Y zachodzi:

$$Y = F_Y^{-1}\left(F_X(X)\right), \tag{4.2}$$

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 i równanie (4.2) nie może zostać bezpośrednio zastosowane. Powoduje to, że wszystkie ekwicentylowe metody zrównywania wyników obserwowanych zmuszone są do uwzględnienia 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:

$$ {}^{(Equip)}eq_Y(x) = {}^{(cont)}F_Y^{-1}\left({}^{(cont)}F_X\ (x)\right). \tag{4.3}$$

Ekwicentylowa funkcja zrównująca podana wzorem (4.3) jest złożeniem uciąglonej dystrybuanty rozkładu wyników w teście X z odwrotnością uciąglonej dystrybuanty rozkładu wyników w teście Y. Dwoma najpopularniejszymi metodami uciąglania skokowych dystrybuant jest (a) lokalna interpolacja liniowa (b) wygładzanie za pomocą estymatora jądrowego (ang. *kernel smoothing*). Pogłębiony przegląd pierwszego podejścia można znaleźć u Kolen & Brenan (2004), a drugiego u von Davier et al. (2004). Ostatnim krokiem w procedurze zrównywania jest zaokrąglenie zrównanych poprzez funkcję (4.3) wyników.

4.5. Zrównywanie wyników obserwowanych z zastosowaniem modelu IRT

Zrównywania wyników obserwowanych z zastosowaniem modelu IRT (por. Tabela 4.1) wymaga estymacji nieobserwowanych dystrybuant obserwowanych wyników $F_{x|Q}$ lub $F_{x|\mathcal{P}}$ lub obu tych dystrybuant na podstawie parametrów modelu IRT wyrażonych na wspólnej dla populacji $\mathcal{P}$ i Q skali. Ustalając uwagę na $F_{x|Q}$, 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 $\mathbb{P}(X = x \mid \theta)$ 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 jest podany w Kolen & Brenan (2004). Glas & Béguin (1996) wskazują również na możliwość oszacowania szukanego $F_{x|Q}$ poprzez przeprowadzenia stosownego eksperymentu Monte Carlo bazującego na oszacowanym i zrównanym modelu IRT.

W przeprowadzonych badaniach zaadaptowano symulacyjną strategię generowania wyrażonych na wspólnej skali wyników obserwowanych roku bazowego $X_{\cdot}^{baz}$. Dla egzaminu gimnazjalnego rok bazowy ustalono na 2003 w teście matematyczno-przyrodniczym oraz humanistycznym, dla sprawdzianu po klasie VI za rok bazowy wybrano 2004. Do wygenerowania wyników obserwowanych dla danego rocznika na skali z roku bazowego (X_r^{baz}) generowano 5 milionów wyników zgodnie z oszacowaną dla tego rocznika średnią i odchyleniem standardowym rozkładu umiejętności θ oraz przy uwzględnieniu parametrów zadań dla roku bazowego.

Aby uzyskać jak najlepsze oszacowanie rozkładu wyników obserwowanych parametry zadań egzaminacyjnych zostały oszacowane w niezależnej procedurze od wielokrotnie replikowanej na podpróbkach danych egzaminacyjnych kalibracji łącznej wielogrupowym modelem IRT. Kalibracja wielogrupowym modelem IRT z wykorzystaniem oprogramowania MIRT, opisana we wcześniejszych akapitach rozdziału, miała na celu oszacowanie pierwszych dwóch momentów rozkładu umiejętności w poszczególnych latach na wspólnej skali wraz z błędami standardowymi tych parametrów. Wykorzystano do tego program MIRT, ze względu na to, że obsługuje on model wielogrupowy. Oszacowania parametrów zadań przeprowadzono natomiast dla każdego z egzaminów osobno, wykorzystując wszystkie wektory uczniowskich odpowiedzi zebrane w danym roku oraz inne niż do kalibracji łącznej oprogramowanie – Parscale 4.1. Zmiana oprogramowania była podyktowana przede wszystkim koniecznością wykorzystania w przypadku części zadań ocenianych 0-1 trójparametrycznego modelu logistycznego (3PLM) zamiast zastosowanego w zrównywaniu modelu dwuparametrycznego (2PLM). Model 3PLM został wykorzystany w przypadku tych zadań, dla których oszacowanie dwuparametrycznej krzywej charakterystycznej wykazywało niedopasowanie dla uczniów o niskim poziomie umiejętności wskazujące na występowanie dolnej asymptoty krzywej charakterystycznej położonej istotnie powyżej zera. Niemodelowanie tego zjawiska poprzez dwuparametryczną krzywą charakterystyczną skutkowałoby obciążonymi oszacowaniami rozkładu wyników obserwowanych.

Najlepszym rozwiązaniem byłoby zastosowanie modelu 3PLM również przy kalibracji łącznej w modelu wielogrupowym, niestety program MIRT nie obsługuje modelu 3PLM, a program Parscale nie obsługuje modelu wielogrupowego. Uzyskane za pomocą Parscale oszacowania parametrów zadań wraz z krzywymi charakterystycznymi oraz empirycznymi proporcjami zdobytych punktów w centylach umiejętności (które pozwalają ocenić dobroć dopasowania) przedstawiono w Aneksie A.

Przeliczenie konkretnego wyniku obserwowanego w roku r (oznaczymy x_r^{obs}), na odpowiedni wynik obserwowany dla bazowego testu (x_r^{przel}) na podstawie wspomnianych 5 milionów zasymulowanych wyników w teście X_r^{baz} zostało wyznaczone jako modalny wynik na skali bazowej uczniów o wyniku x_r^{obs}:

$$x_r^{przel} = \max_{x_r^{baz}} P(X_r^{baz} = x_r^{baz} | X_r^{obs} = x_r^{obs})$$

Przeliczone wyniki x_r^{przel} posłużyły do stworzenia tablic przeliczeniowych zamieszczonych w dalszej części raportu (Tabele 6.2, 6.3 oraz 6.7).

Należy zauważyć, że zastosowana procedura nie spełnia formalnych wymogów nałożonych na zrównywanie. Dokonano przewidywania wyników ze wszystkich lat na skalę z roku bazowego, przez co niezachowany jest chociażby wymóg symetrii. Przełożenie wszystkich lat na jeden rok bazowy jest jednak rozwiązaniem pozwalającym na porównywanie wyników między wieloma latami. Zrównywanie w ścisłym rozumieniu tego terminu, jak opisano wcześniej, na skali wyników obserwowanych doprowadziłoby przy zaokrąglaniu do całkowitych punktów do przekształcenia identycznościowego – tylko takie przekształcenie pozwala przeliczyć wyniki w dwóch testach ocenianych na tą samą liczbę punktów w sposób różnowartościowy.

4.5.1. Generowanie PV z wykorzystaniem MCMC

W wyniku zrównania program MIRT dostarcza jedynie dwa pierwsze momenty rozkładu umiejętności. Dla zwiększenia precyzji odwzorowania kształtu rozkładu θ przy generowaniu wyników obserwowanych, obserwacje z rozkładu θ generowano z wykorzystaniem tak zwanych *plausible*

values – w skrócie *PV*. PV stanowią realizacje z rozkładu a *posteriori* parametru umiejętności ucznia o wektorze odpowiedzi $\boldsymbol{u}$ (Wu, 2005):

$$\mathbb{P}(\theta|\boldsymbol{U}=\boldsymbol{u}) = \frac{\mathbb{P}(\boldsymbol{U}=\boldsymbol{u}|\theta,\boldsymbol{\beta})\psi_0(\theta)}{\int \mathbb{P}(\boldsymbol{U}=\boldsymbol{u}|\theta,\boldsymbol{\beta})\psi_0(\theta)\,d\theta} \tag{4.4}$$

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 (porównaj równanie (4.1)).

Uzyskanie PV zgodnie z równaniem (4.4) wymaga również zastosowania zaawansowanych numerycznych rozwiązań opartych na metodologii MCMC (*Markov Chain Monte Carlo*). W badaniu łańcuchy Markowa służące do wygenerowania PV stworzono zgodnie z podejściem Metropolis Hastings z symetryczną funkcją generującą kandydatów na kolejne punkty w łańcuchu Markowa. Konkretnie algorytm składał się z następujących kroków (por.: Patz & Junker, 1999, oraz de la Torre, 2009):

1) wylosuj punkt kandydujący θ^* zgodnie z generującym rozkładem $N(.|\theta^t, SE^2_{\theta_0})$;

2) oblicz prawdopodobieństwo: $\alpha = min\left\{1, \frac{\mathbb{P}(U=u|\theta^*,\beta)\psi_0(\theta^*)}{\mathbb{P}(U=u|\theta^t,\beta)\psi_0(\theta^t)}\right\}$;

3) wylosuj v z rozkładu jednostajnego na przedziale (0;1);

4) jeżeli $v < \alpha$, to zaakceptuj kandydata ($\theta^{t+1} = \theta^*$), a w przeciwnym razie pozostaw łańcuch w miejscu ($\theta^{t+1} = \theta^t$).

Przy czym spełnione są następujące warunki:

a) wartość startowa łańcucha θ_0 jest punktowym oszacowaniem EAP dostarczonym przez program użyty do dopasowania modelu IRT;

b) stała wartość odchylenia standardowego funkcji generującej kandydatów SE_{θ_0} jest wzięta jako błąd standardowy oszacowania θ_0 również raportowany przez program pierwotnie estymujący parametry IRT; kształt funkcji jest normalny (symetryczność);

c) rozkład a *priori* parametry umiejętności $\psi_0(.)$ jest rozkładem standardowym normalnym.

Ze względu na dobór wartości startowej łańcucha (θ_0) i odchylenia standardowego funkcji generujących kandydatów (SE_{θ_0}) bardzo zbliżony do faktycznego rozkładu a *posteriori* (4.44), łańcuchy MCMC od samego początku znajdowały się w centralnym rejonie swojego docelowego rozkładu stacjonarnego. Rozkład a *posteriori* (4.4) wykorzystany do generowania PV uzyskiwano z 500 replikacji łańcucha po uprzednim odrzuceniu 200 pierwszych replikacji łańcucha (tzw. *burn-in*).

Do generowania PV dla uczniów piszących dany egzamin wykorzystano parametry zadań oszacowane za pomocą programu Parscale, których wartości podano w Aneksie A. Parametry zadań w Parscale są oszacowane w modelu jednogrupowym zakotwiczonym na rozkładzie o średniej 0 i odchyleniu standardowym 1, zatem przed przystąpieniem do dalszych analiz z wykorzystaniem PV przeskalowano je dla każdego roku, tak aby uwzględnić różnice rozkładów umiejętności między latami oszacowane za pomocą programu MIRT.

5. Dobór zadań do zrównywania

5.1. Wstęp

Przed przystąpieniem do zrównywania zarówno wyników egzaminu gimnazjalnego, jak i wyników sprawdzianu po szkole podstawowej dokonano kilkuetapowej analizy właściwości zadań egzaminacyjnych, która miała na celu wybranie odpowiedniego zbioru zadań wykorzystanych do oszacowania funkcji zrównującej wyniki egzaminów ze sobą.

W pierwszym etapie zadania wszystkich egzaminów gimnazjalnych z lat 2002-2010 poddano kompleksowej analizie psychometrycznej z wykorzystaniem narzędzi klasycznej teorii testów, jak i IRT. W drugim etapie badań analizowano również zadania z egzaminu gimnazjalnego z roku 2011 i zadania ze sprawdzianu po szkole podstawowej z lat 2002-2011. W tym raporcie prezentujemy łącznie wyniki wcześniejszych badań oraz nowych analiz.

Wyniki analizy zadań egzaminacyjnych zebrano w Aneksie A – *Psychometryczne właściwości zadań egzaminacyjnych.* Na podstawie tych analiz dokonano wyboru najlepszych zadań egzaminacyjnych, które wykorzystano w arkuszach rozwiązywanych podczas sesji zrównującej. Po przeprowadzeniu badania zrównującego ponownie przeanalizowano właściwości wykorzystanych w badaniu zadań egzaminacyjnych, a także nowych, nieegzaminacyjnych, zadań kotwiczących. Niniejszy rozdział rozpocznie przedstawienie zadań wykluczonych ze zrównywania na podstawie tych analiz.

Następny etap wyboru zadań wykorzystanych do zrównania wyników egzaminacyjnych nastąpił po przeprowadzeniu badania zrównującego, które dostarczyło danych pozwalających na zakotwiczenie wyników z różnych lat na wspólnej skali. Polegał on na wielokrotnym przeprowadzeniu zrównania wyników egzaminów na losowych połówkach zadań (podtesty zawierające połowę zadań z arkusza) z danego roku i analizie stopnia, w jakim średni poziom umiejętności w danym roku był wrażliwy na wykorzystaną do zrównywania kombinację zadań. Procedura ta pozwoliła zidentyfikować grupę zadań egzaminacyjnych, które wpływały na zrównane średnie w sposób odmienny od pozostałych zadań. Obserwacje zostały uzupełnione o analizę zróżnicowanego funkcjonowania zadań (DIF) między prawdziwą sytuacją egzaminacyjną, a sytuacją badania zrównującego, która potwierdziła słuszność wykluczenia z obliczania funkcji zrównującej kolejnej porcji zadań.

5.2. Egzamin gimnazjalny 2002-2010 i 2011

5.2.1. Wykluczenie zadań na podstawie właściwości psychometrycznych

Przyjęto zasadę, iż aby zadanie pozostało w analizach, musi zachowywać minimalnie dobre właściwości psychometryczne zarówno w prawdziwym egzaminie, jak i badaniu zrównującym.

Zadania o bardzo niskiej dyskryminacji (korelacja zadania z resztą testu: poniżej 0,2[22]), bardzo łatwe (łatwość powyżej 95%) oraz niedopasowane do modelu odpowiedzi na zadanie testowe (IRT) zostały usunięte z analiz. Parametry wszystkich zadań egzaminacyjnych można znaleźć w Aneksie A *Psychometryczne właściwości zadań egzaminacyjnych*. Listę zadań usuniętych z analizy dla części humanistycznej oraz matematyczno-przyrodniczej przedstawiają odpowiednio Tabela 5.1 oraz Tabela 5.2. W nawiasach przy numerze zadania podany został powód wykluczenia. Wartość liczbowa wskazuje na moc różnicującą zadania, litery „e": *egzamin*, „b": *badanie zrównujące*. Skrót „nied" oznacza zadanie *niedopasowane do modelu* odpowiedzi na zadanie testowe. Skrót „łatwość" oznacza iż zadanie zostało *usunięte ze względu na zbyt dużą łatwość*.

Tabela 5.1. Zadania humanistyczne usunięte z analiz ze względu na słabe właściwości psychometryczne

Rok/Arkusz	Usunięte zadania humanistyczne (numer zadania)					
2002	1 (0,17e)	9 (0,19e)	12 (-0.03e)	14 (0.07e)	19 (0,14e)	
2003	1 (0,15e)	7 (0,12e)	5 (0,11e)	13 (0,16e)	14 (0,19e)	28 (0,19b)
2004	7 (0,17b)	20 (0,15b)				
2005	3 (0,17b)	22 (nied e)				
2006	6 (0,18e)	7 (0,16e)	12 (0,20e)	14 (0,14e)	19 (0,20e)	20 (0,19b)
2007	17 (0,15e)					
2008	25 (0,11e)					
2009	2 (łatwość e)	6 (łatwość e)	11 (0,14e)	21 (0,15e)	16 (0,19b)	
2010	10 (nied e)	12 (0,16e)	13 (0,07e)	14 (0,15e)	15 (0,17e)	
2011	13 (0,13e)	14 (0,07e)				
ht1*	28 (0,12b)	30 (0,20b)	32 (0,12b)	41 (0,13b)	44 (0,20b)	
ht2*	12 (0,18b)	20 (0,18b)	32 (0,19b)	39 (0,03b)	45 (0,16b)	

* zeszyty z dodatkowymi zadaniami w sesji zrównującej

Tabela 5.2. Zadania matematyczno-przyrodnicze usunięte z analiz ze względu na słabe właściwości psychometryczne

Rok/Arkusz	Usunięte zadania matematyczno-przyrodnicze (numer zadania)					
2002	1 (0,18e)	2 (0,14e)	6 (0,19e)	8 (0,16e)	13 (0,20e)	28 (nied e)
2003	1	16	17	22		

[22] Kryterium to nie było używane, jeżeli zadania charakteryzowało się bardzo dobrym poziomem dopasowania do modelu dwuparametrycznego.

	(0,17e)	(0,07e)	(0,15e)	(0,11e)					
2004	9 (0,16e)	25 (0,17e)							
2005	15 (0,17e)	21 (0,19e)							
2006	13 (0,17e)	25 (0,19e)							
2007	4 (nied e)	18 (nied e)	23 (0,19e)	25 (0,19e)	26 (0,19e)				
2008	8 (0,12e)	18 (0,15e)							
2009	8 (0,13e)								
2010	1 (nied e)	2 (0,08e)	3 (0,12e)	9 (0,18e)	13 (-0,05e)	17 (0,19e)	20 (0,11e)	22 (0,06e)	23 (0,17e)
2011	4 (0,15e)	15 (0,13e)	17 (0,16)	33 (0.06e)					
c1*	7 (0,09b)	16 (0,18b)	21 (0,12b)						
c2*	1 (0,07b)	13 (0,16b)	28 (0,16b)	34 (0,14b)	4 (-0,17b)	87 (0,16b)			

* zeszyty z dodatkowymi zadaniami w sesji zrównującej

Jak widać już na wstępnym etapie, z każdego roku w procedurze zrównywania usuwane są średnio 4 zadania. Przy czym istnieje duża zmienność liczby usuniętych zadań między latami i egzaminami. Dla przykładu dla egzaminów w roku 2009 usunięto 4 zadania w części humanistycznej, ale tylko jedno w części matematyczno-przyrodniczej. Najwięcej zadań z uwagi na niesatysfakcjonujące właściwości psychometryczne usunięto w roku 2010 – w części humanistycznej 4, a w części matematyczno-przyrodniczej aż 9.

5.2.2. Analiza wrażliwości zrównywania na dobór zadań

Aby oszacować wrażliwość zrównywania na dobór zadań, zastosowano metody symulacyjne, które polegają na usuwaniu kolejnych zadań z puli zrównującej i przeprowadzeniu zrównywania. Procedura polegała na wielokrotnym zrównywaniu liczbą n-1 zadań, gdzie n oznacza liczbę zadań zrównujących. Innymi słowy dokonywano wielu zrównań, w każdym ze zrównań pomijając jedno zadanie. Procedura ta jest jedną z przyjętych metod statystycznych i należy do grupy metod replikacyjnych. W niektórych kontekstach nazywana jest metodą *Jacknife*. Zadania wytypowane do usunięcia w punkcie 5.21. również brały udział w tej procedurze.

Metoda ta pozwala na kwantyfikację wrażliwości wybranej metody zrównywania na dobór zadań, a także daje informację o błędzie zrównywania. Jeżeli rozrzut wyników po zrównywaniu dla określonego roku przy różnym doborze zadań będzie stosunkowo wysoki, należy stwierdzić, iż błąd zrównywania jest wysoki, a tym samym precyzja oszacowania wyniku stosunkowo mała. W analizie ważną rolę odgrywają również wyniki odstające, które sugerują iż po wyrzuceniu poszczególnego zadania można mieć do czynienia z istotną zmianą wyników zrównywania. Może być to efektem jednego lub kilku zadań, które z jakiś względów funkcjonowały inaczej podczas egzaminu i badania zrównującego, lub też wielowymiarowością testu i zadań zrównujących. W pierwszym wypadku sprawa wydaje się prosta i polega na usunięciu zadań odmiennie funkcjonujących w obydwu testach tak, aby zwiększyć precyzję zrównywania. Druga sytuacja jest znacznie bardziej skomplikowana, jednak na szczęście, jak się okaże w kolejnych punktach, w przypadku tego zrównywania do niej nie dochodzi.

5.2.2.1. Część humanistyczna

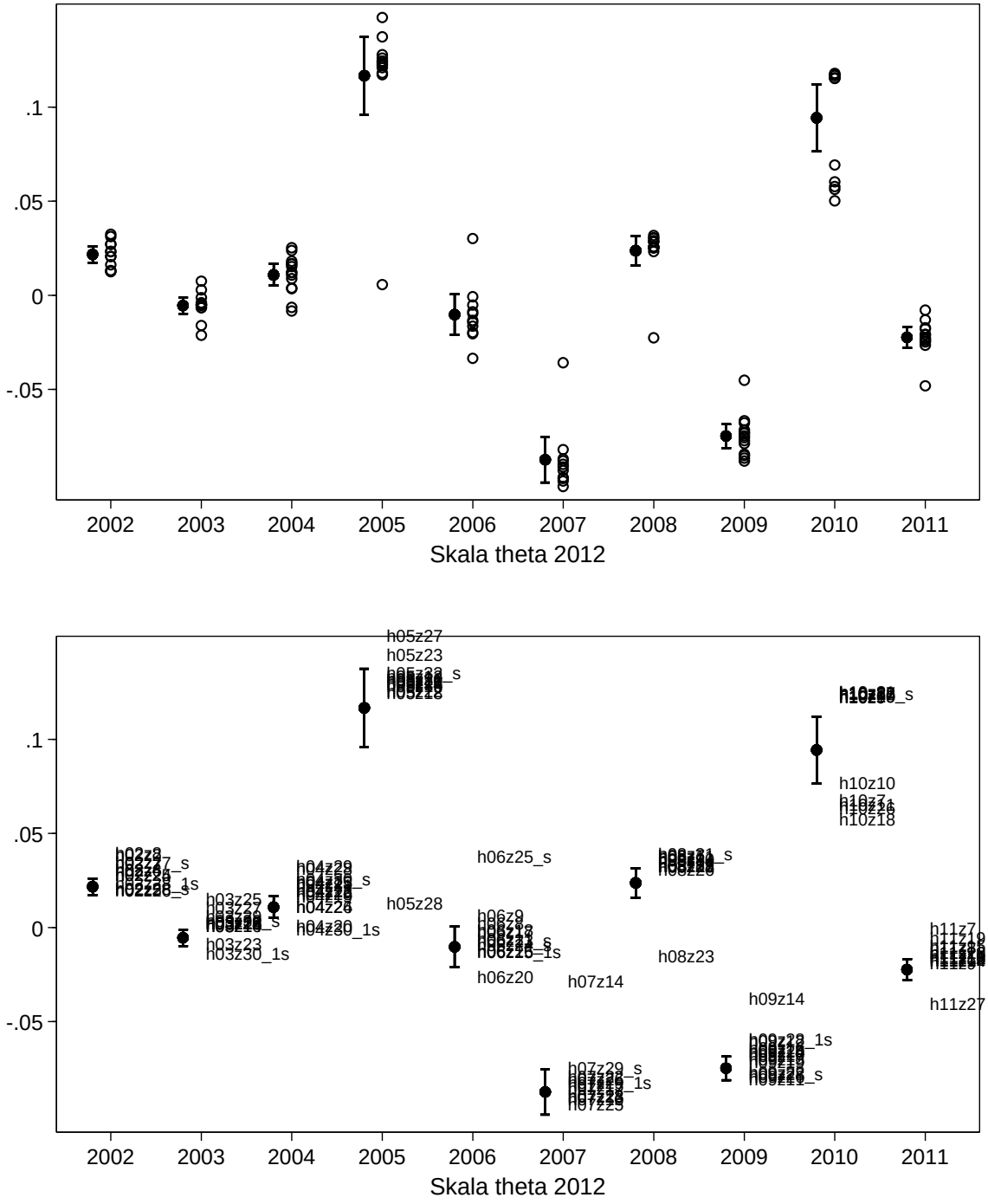
Rozkład wyników zrównania[23] w części humanistycznej dla poszczególnych lat przedstawiony został na Rysunku 5.1. Jak pisano wcześniej zrównanie okazuje się problematyczne w przypadku wyników

[23] Wyniki zrównania w przedstawianych analizach wrażliwości procedury na dobór zadań znajdują się na skali o średniej 0 oraz odchyleniu standardowym 1 dla uczniów biorących udział w sesji zrównującej w 2012 roku. Wyniki ostatecznego zrównania przedstawione są natomiast na skali o średniej 100 oraz 15 dla uczniów piszących egzaminy w 2003 roku. W analizach wrażliwości lokalizacja skali nie ma istotnego znaczenia, istotne jest umiejscowienie wyników na skali pozwalające ocenić kształt rozkładu średniego poziomu umiejętności w danym roku. Ponadto wyrażenie wyników wstępnych analiz zrównywania na innej skali niż ostateczna pozwala uniknąć potencjalnych nieporozumień. Z tego względu wyniki analiz wrażliwości procedury na dobór zadań pozostawiono na skali domyślnie wynikającej z zastosowanej metody zrównywania.

Kolejną techniczną różnicą między zrównywaniem zastosowanym w analizie wrażliwości, a ostatecznie zastosowanym zrównaniem jest wykorzystane oprogramowanie i sposób obliczania średniej w danym roku. W analizie wrażliwości wyniki zrównywano metodą ustalonych parametrów z wykorzystaniem programu Parscale, natomiast ostateczne zrównanie przeprowadzono metodą kalibracji łącznej z wykorzystaniem programu MIRT. Zrównując wyniki programem Parscale średnią w danym roku liczono poprzez uśrednienie oszacowań EAP umiejętności pojedynczych uczniów, natomiast oszacowanie średniej w danym roku w programie MIRT jest dokonywane wprost podczas estymacji parametrów modelu. Mimo iż kalibracja łączna zaimplementowana w MIRT jest metodologicznie słuszniejszym rozwiązaniem niż zrównywanie metodą ustalonych parametrów w programie Parscale, przeprowadzone przez zespół badawczy analizy wskazują na bardzo wysoką korelację wyników zrównywania tymi dwoma metodami. Program Parscale został wykorzystany do analizy wrażliwości ze względu na wydajność obliczeniową przy dużych zbiorach danych.

odstających. W takim przypadku wynik zrównywania zależy szczególnie mocno od doboru zadań do puli zrównującej. Wyraźnie do takiej sytuacji dochodzi w roku 2005, 2007 i 2008, w mniejszym stopniu również w roku 2006.

Rysunek 5.1. Rozkład wyników zrównania dla części humanistycznej, lata 2002-2011 (górny panel bez zaznaczonych nazw zadań, dolny z zaznaczonymi nazwami zadań)



Wyniki odstające sugerują, iż znaczne różnice w wynikach zrównywania mogą być związane z efektem pojedynczego zadania. W sytuacji włączenia takiego problematycznego zadania do

wylosowanej próbki zadań, na której przeprowadzane jest zrównywanie, uzyskuje się rozkład odpowiadający jednej modzie, w przeciwnym przypadku – rozkład odpowiadający drugiej modzie. Tabela 5.3 przedstawia szczegółowe wyniki analizy dla zadań humanistycznych dla „problematycznych" lat. W kolumnie „zadanie" podano numer zadania, które zostało wykluczone ze zrównywania, a w kolumnie „bez zad." przedstawiono średni poziom umiejętności uczniów na skali *theta* bez uwzględnienia danego zadania. Natomiast w kolumnie „różnica" znajduje się wartość bezwzględna różnicy (moduł) pomiędzy średnim poziomem umiejętności uczniów a średnią wartością średnich poziomów umiejętności uczniów ze wszystkich zrównań.

Tabela 5.3. Zrównywanie z pominięciem poszczególnych zadań

2005			2006			2007			2008		
zadanie	bez zad.	różnica	zadanie	bez zad.	różnica	zadanie	bez zad.	różnica	Zadanie	bez zad.	różnica
h05z32_s	0,160	0,017	h06z8	-0,040	0,012	**h07z14**	**-0,035**	**0,050**	h08z7	0,021	0,007
h05z11	0,150	0,007	h06z9	-0,034	0,018	h07z15	-0,091	0,005	h08z8	0,020	0,005
h05z12	0,150	0,006	h06z11	-0,057	0,005	h07z16	-0,100	0,015	h08z9	0,019	0,004
h05z14	0,152	0,008	h06z16	-0,071	0,019	h07z22	-0,088	0,003	h08z11	0,026	0,011
h05z16	0,148	0,005	h06z17	-0,053	0,001	h07z23	-0,096	0,011	h08z13	0,015	0,000
h05z18	0,145	0,002	h06z18	-0,051	0,001	h07z25	-0,095	0,010	h08z14	0,016	0,001
h05z20	0,155	0,012	h06z23_s	-0,065	0,013	h07z26	-0,083	0,002	h08z22	0,016	0,001
h05z23	0,162	0,019	h06z24_s	-0,065	0,013	h07z28	-0,095	0,010	**h08z23**	**-0,030**	**0,045**
h05z26	0,154	0,011	h06z25_1s	-0,071	0,019	h07z29_1s	-0,089	0,004	h08z26	0,008	0,007
h05z27	0,181	0,038	**h06z25_s**	**-0,013**	**0,039**	h07z29_s	-0,080	0,005	h08z27	0,016	0,001
h05z28	**0,013**	**0,130**							h08z28	0,019	0,005
h05z31	0,149	0,006							h08z29	0,019	0,004
									h08z30	0,020	0,005
									h08z31_s	0,023	0,008

Czerwoną czcionką w Tabeli 5.3 zaznaczono zadania, które wytypowano jako „problematyczne". Wyniki zrównywania w przypadku zaznaczonych zadań okazują się znacznie odbiegać od pozostałych. Jak pisano wcześniej, taka sytuacja może być spowodowana odmiennym funkcjonowaniem zadania na egzaminie i w próbie badawczej. Na odmienne funkcjonowanie zadania może wpływać osłabiona motywacja podczas badań zrównujących, kontekst, w którym pojawia się zadanie, miejsce zadania w teście, czy zmiana układu dystraktorów.

W Tabeli 5.4 podano łatwość, moc różnicującą (korelacja wyniku zadania z resztą testu) oraz maksymalną liczbę punktów do zdobycia za zadanie. Już na pierwszy rzut oka widać, iż zadania te znacząco różnią się pod względem funkcjonowania psychometrycznego w obydwu zastosowaniach.

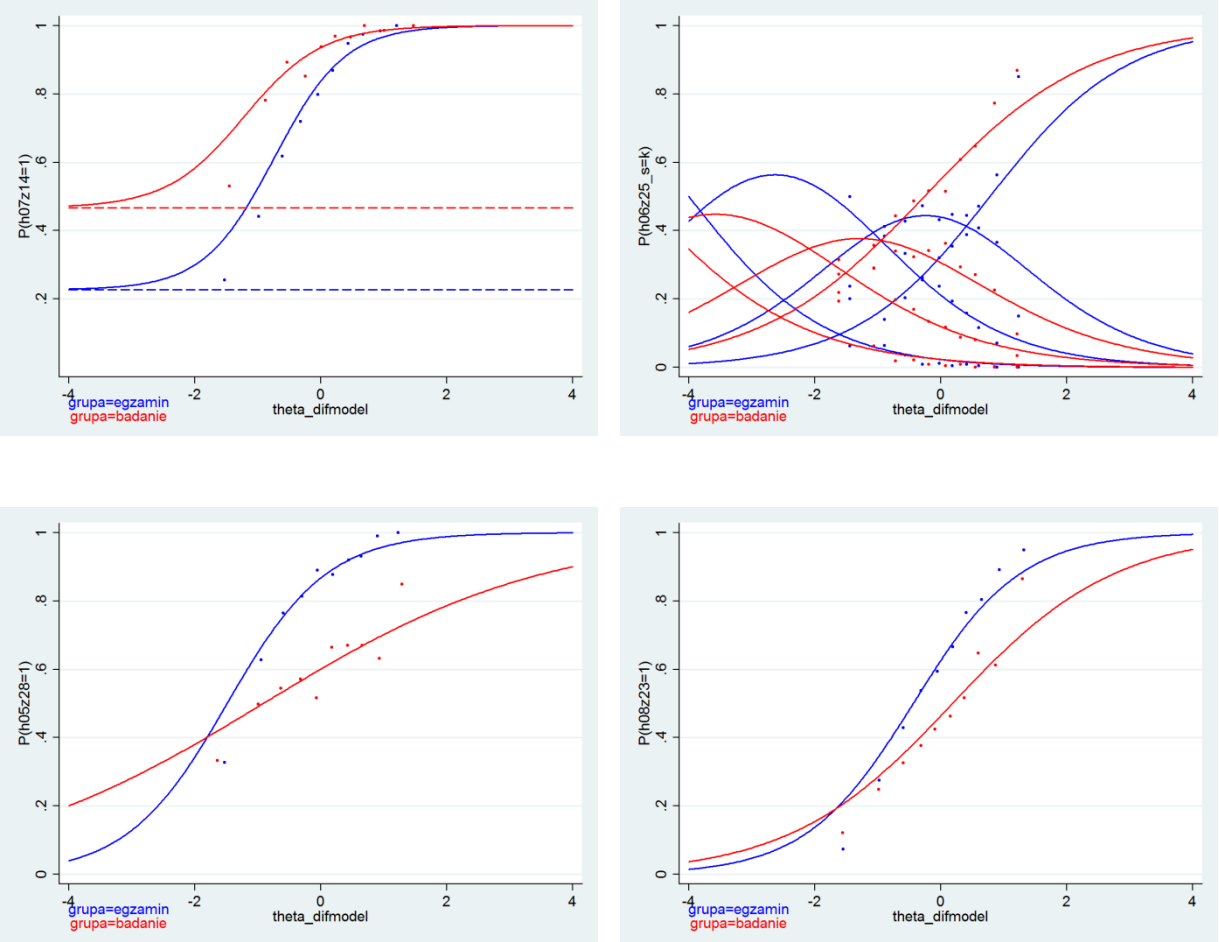
Tabela 5.4. Statystyki zadań „problematycznych" część humanistyczna

	Egzamin			**Próba badawcza**		
	Łatwość	Moc	Max	Łatwość	Moc	Max
h05z28	0,814	0,379	1	0,594	0,207	1
h06z25_s	0,524	0,455	3	0,777	0,321	3
h07z14	0,760	0,454	1	0,891	0,285	1
h08z23	0,601	0,399	1	0,459	0,283	1

Oprócz klasycznej analizy, przedstawionej w Tabeli 5.4, przeprowadzono analizę **zróżnicowanego funkcjonowania zadania**, (ang. *differential item functioning*, DIF). DIF informuje o występowaniu różnic w trudności zadania między grupami uczniów lub zastosowaniami testów, gdy kontrolowany jest ich poziom umiejętności, który operacyjnie jest szacowany poprzez wynik w danym teście. Uczniowie o takim samym wyniku powinni mieć takie same prawdopodobieństwo uzyskania określonej liczby punktów za zadanie bez względu na przynależność do określonej grupy społecznej, etnicznej itp., w przeciwnym wypadku występuje DIF.

Do analizowania zadań z Tabeli 5.4 pod kątem DIF wykorzystano modelowanie IRT (ang. *Item Response Theory*). Modele IRT wyznaczają dla każdego zadania tzw. krzywe charakterystyczne, opisujące prawdopodobieństwa uzyskania określonej liczby punktów w zależności od poziomu umiejętności. Zadanie pod kątem występowania DIF zostało zanalizowane poprzez dopasowanie do niego krzywej charakterystycznej o niezależnych parametrach dla uczniów rozwiązujących je w sytuacji egzaminacyjnej oraz w sytuacji badawczej przy jednoczesnym modelowaniu różnic w poziomie umiejętności miedzy grupami (model dwugrupowy). Jest to bardzo cenne narzędzie do oddzielenia różnic w poziomie umiejętności dwóch grup od różnic we właściwościach zadania między dwoma grupami. Występowanie DIF oceniono poprzez analizę rozbieżności między tak dopasowanymi krzywymi charakterystycznymi. Na Rysunku 5.2 ukazano krzywe charakterystyczne dla rozpatrywanej grupy zadań niezależnie dopasowane dla uczniów piszących egzamin gimnazjalny oraz dla uczniów rozwiązujących arkusze podczas badania zrównującego.

Rysunek 5.2. Analiza DIF zadań „problematycznych" dla zrównywania wyników części humanistycznej egzaminu gimnazjalnego

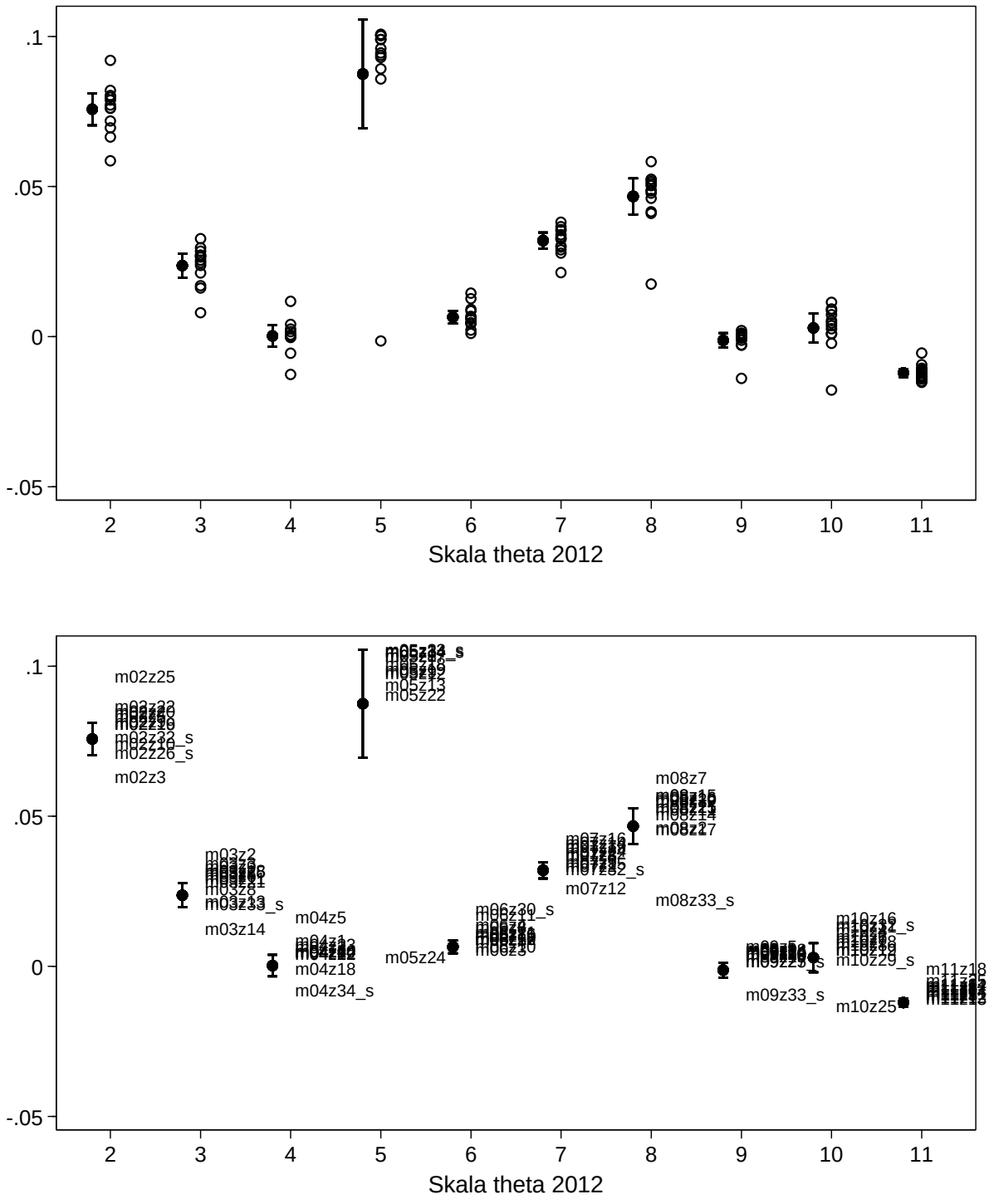


W zadaniu niewykazującym DIF oczekiwano by takiego samego prawdopodobieństwa udzielenia odpowiedzi ocenianej na określoną liczbę punktów w zależności od poziomu umiejętności w obu grupach uczniów. Krzywe charakterystyczne w rozbiciu na zastosowanie zadań (egzamin – badania zrównujące) powinny się ze sobą pokrywać z dokładnością do losowych fluktuacji. Dla rozpatrywanych zadań wspomniane krzywe, wraz z procentem faktycznie zdobytych punktów za te zadania w decylach rozkładu umiejętności (które pozwalają ocenić dobroć oraz precyzję dopasowania), przedstawiono na Rysunku 5.2. Bliskość empirycznego procentu zdobytych za zadania punktów w decylach do oszacowanych krzywych charakterystycznych wskazuje, że różnice między funkcjonowaniem zadania na egzaminie i w badaniu zrównującym są systematyczne, co wskazuje na istotny statystycznie DIF. W związku z powyższym cztery analizowane zadania również wyłączono z puli zadań łączących przy zrównywaniu.

5.2.3. Część matematyczno-przyrodnicza

Analogiczne analizy przeprowadzono dla części matematyczno-przyrodniczej. Tak jak w przypadku części humanistycznej. Rozkład wyników zrównania dla poszczególnych lat przedstawiony został na Rysunku 5.3.

Rysunek 5.3. Rozkład wyników zrównania wyników dla części matematyczno-przyrodniczej, lata 2002-2011 (górny panel bez zaznaczonych nazw zadań, dolny z zaznaczonymi nazwami zadań)



W przypadku zrównania części matematyczno-przyrodniczej dwumodalność rozkładów obserwowana jest w latach 2005 i 2010. Zrównanie roku 2008 zdradza niewielką skłonność do dwumodalności, nie jest ona jednak wyraźna, a odchylenie standardowe rozkładu zrównań w 2008 nie jest duże (nie jest

większe np. od odchylenia standardowego dla roku 2003). Mimo to przeprowadzono analizę efektu wykluczenia pojedynczego zadania ze zrównywania również dla trzech wymienionych lat. Tak jak w przypadku części humanistycznej usuwano po jednym zadaniu z puli zrównującej, po czym dokonywano zrównania. Wyniki tej procedury przedstawiono w Tabeli 5.5.

Tabela 5.5. Zrównywanie z pominięciem poszczególnych zadań (gimnazjum część matematyczno-przyrodnicza)

2005			2008			2010		
Zadanie	bez zad.	różnica	zadanie	bez zad.	różnica	zadanie	bez zad.	różnica
m05z8	0,105	0,027	m08z16	0,050	0,006	**m10z25**	**-0,025**	**0,015**
m05z9	0,099	0,021	m08z17	0,040	0,005	m10z4	-0,007	0,002
m05z12	0,100	0,022	m08z25	0,046	0,001	m10z6	-0,008	0,002
m05z24	**0,008**	**0,070**	m08z12	0,047	0,002	m10z8	-0,008	0,002
m05z18	0,101	0,023	m08z14	0,042	0,002	**m10z11**	**0,004**	**0,013**
m05z19	0,102	0,024	m08z20	0,048	0,003	m10z12	-0,010	0,001
m05z22	0,092	0,014	m08z21	0,047	0,003	m10z7	-0,011	0,001
m05z17	**-0,083**	**0,161**	m08z7	0,056	0,012	m10z16	-0,002	0,008
m05z14	0,105	0,027	m08z11	0,045	0,001	m10z18	-0,012	0,003
m05z13	0,099	0,021	m08z15	0,051	0,007	m10z19	-0,018	0,008
m05z33_s	0,101	0,023	m08z1	0,051	0,007	m10z29_s	-0,008	0,001
m05z34_s	0,106	0,028	m08z2	0,044	0,001			
			m08z33_s	**0,012**	**0,033**			

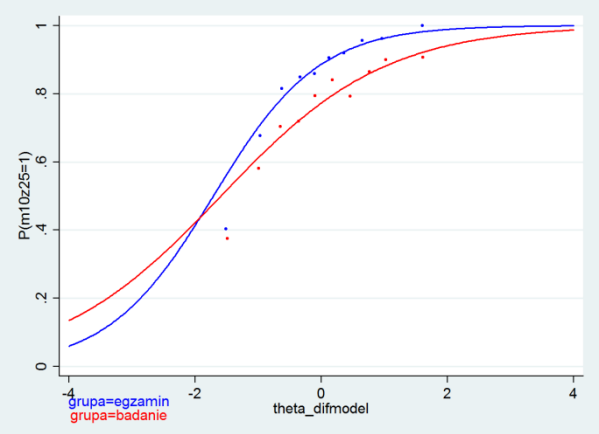
Sytuacja zrównania części matematyczno-przyrodniczej nie jest tak przejrzysta jak w przypadku części humanistycznej. W przypadku roku 2005, jak i 2010, usunięcie przynajmniej dwóch zadań znacznie zmienia wyniki zrównywania (m05z24, m05z17 oraz m10z23, m10z11). Dla roku 2008 tylko usunięcie zadania m08z33_s zmienia wyniki zrównywania, zmiana ta nie jest jednak tak duża jak w przypadku roku 2005 czy 2010 i może być związana z wielopunktowym charakterem tego zadania. W Tabeli 5.6 przedstawiono podstawowe psychometryczne statystyki zadań oszacowane na podstawie egzaminu oraz próby badawczej.

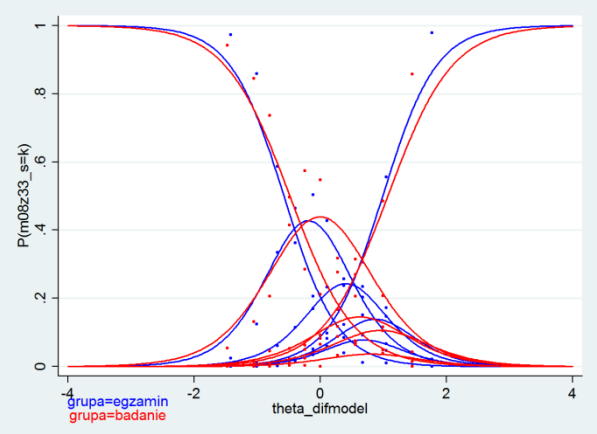
Tabela 5.6. Statystyki zadań „problematycznych" część matematyczno-przyrodnicza

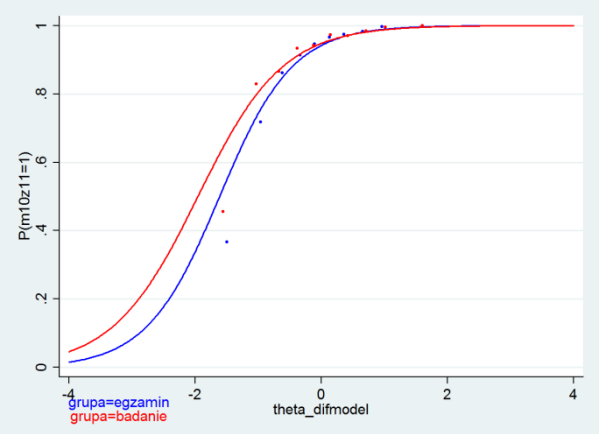
	Egzamin			Próba badawcza		
	Łatwość	Moc	Max	Łatwość	Moc	Max
m05z17	0,483	0,474	1	0,537	0,454	1
m05z24	0,563	0,190	1	0,335	0,190	1
m08z33_s	0,394	0,709	5	0,344	0,585	5
m10z25	0,836	0,321	1	0,751	0,277	1
m10z11	0,879	0,353	1	0,894	0,338	1

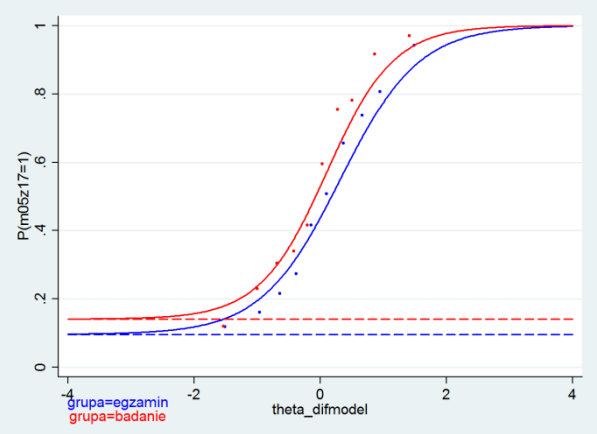
Różnice we właściwościach „problematycznych" zadań w części matematyczno-przyrodniczej są znacznie subtelniejsze niż w przypadku części humanistycznej. Zadania m05z17 i m10z11 charakteryzują się bardzo podobnymi statystykami psychometrycznymi, a zadanie m05z24 znacznie różni się trudnością, podczas gdy dyskryminacja zadania w dwóch zastosowaniach pozostaje niezmienna. Pozostałe trzy zadania różnią się zarówno łatwością, jak i dyskryminacją. Nie są to różnice bardzo duże. Obraz ten potwierdzony jest przez analizę DIF, której graficzną reprezentację przedstawiono na Rysunku 5.4

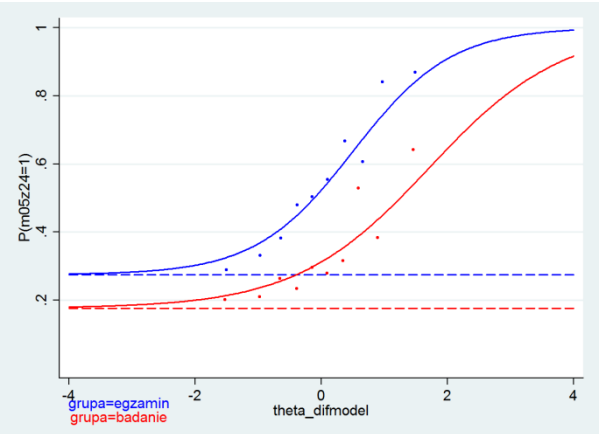
Rysunek 5.4. Analiza zróżnicowanego funkcjonowania zadania (DIF) zadań „problematycznych" dla zrównywania wyników części matematyczno-przyrodniczej egzaminu gimnazjalnego











5.2.4. Ostateczna pula zadań zrównujących wyniki egzaminu gimnazjalnego

W Tabeli 5.7 przedstawiono ostateczną liczbę zadań zastosowanych w zrównywaniu w podziale na każdy rok, co do którego używano procedury zrównywania. W przypadku części matematyczno-

przyrodniczej w zależności od roku użytych zostało od 31% do 42% zadań wykorzystywanych w egzaminie. Jeżeli chodzi o część humanistyczną, procent ten jest mniejszy i zawiera się w przedziale od 24% do 39%. Liczba zadań użytych do zrównania stanowi o precyzji zrównania wyników. Z doświadczeń międzynarodowych wiadomo, że po wykluczeniu w badaniu zadań o słabych właściwościach psychometrycznych około 1/4, 1/3 zadań używanych jest do zrównywania. Taka liczba pozwala na stosunkowo precyzyjne zrównanie wyników i ustrzeżenie się przed poważniejszymi błędami.

Tabela 5.7. Ostateczna liczba zadań używanych do zrównywania wyników egzaminu gimnazjalnego

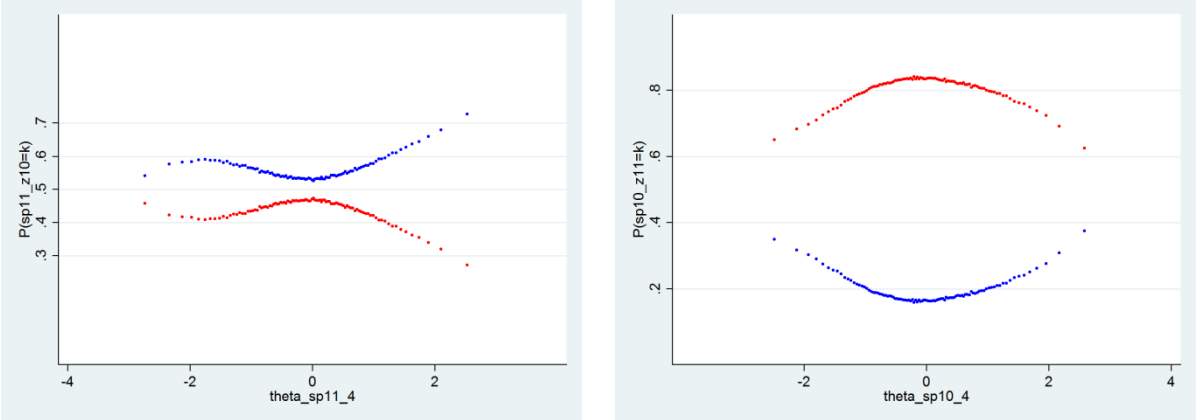
Rok	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Liczba zadań wykorzystanych do zrównania - część matematyczno-przyrodnicza	12	13	12	11	14	13	13	13	12	15
Liczba zadań w arkuszu egzaminacyjnym	36	34	34	35	34	34	33	36	36	36
Procent liczby zadań wykorzystanych do zrównania	33%	38%	35%	31%	41%	38%	39%	36%	33%	42%
Liczba zadań wykorzystanych do zrównania - część humanistyczna	8	11	13	10	10	9	12	11	12	12
Liczba zadań w arkuszu egzaminacyjnym	33	35	35	36	30	34	35	33	32	31
Procent liczby zadań wykorzystanych do zrównania	24%	31%	37%	28%	33%	26%	34%	33%	38%	39%

5.3. Sprawdzian 2002-2011

5.3.1. Wykluczenie zadań na podstawie właściwości psychometrycznych

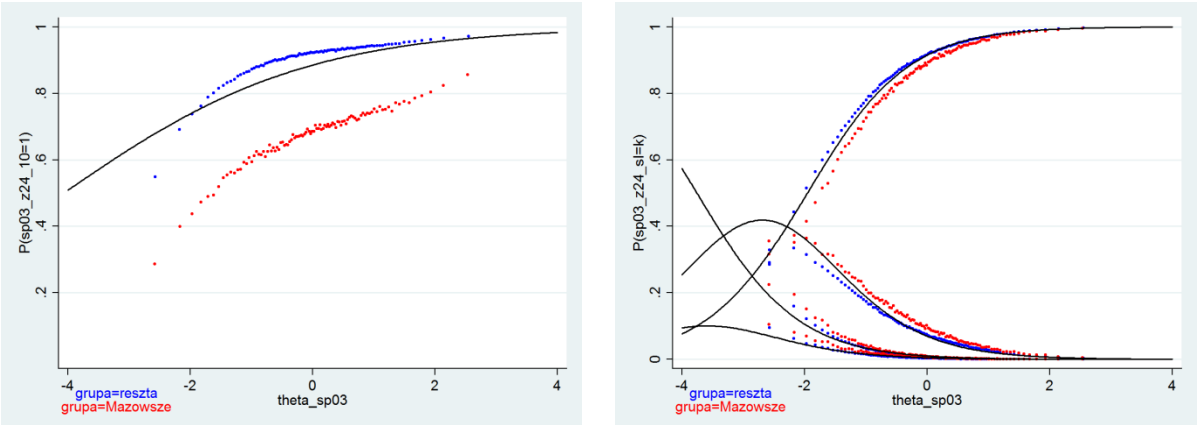
W doborze zadań do zrównywania wyników sprawdzianu po szkole podstawowej przyjęto trochę mniej restrykcyjne kryteria wykluczenia zadań. Odrzucono tylko te zadania, których nie dało się dopasować do modelu IRT: zadanie 11 z roku 2010 i zadanie 10 z roku 2011. Zadania te mają moc różnicującą bliską zeru, a związek prawdopodobieństwa poprawnej odpowiedzi z poziomem umiejętności mierzonym innymi zadaniami egzaminacyjnymi jest nieliniowy i wykluczający te zadania z puli zadań pomiarowo użytecznych (związek został pokazany na Rysunku 5.5).

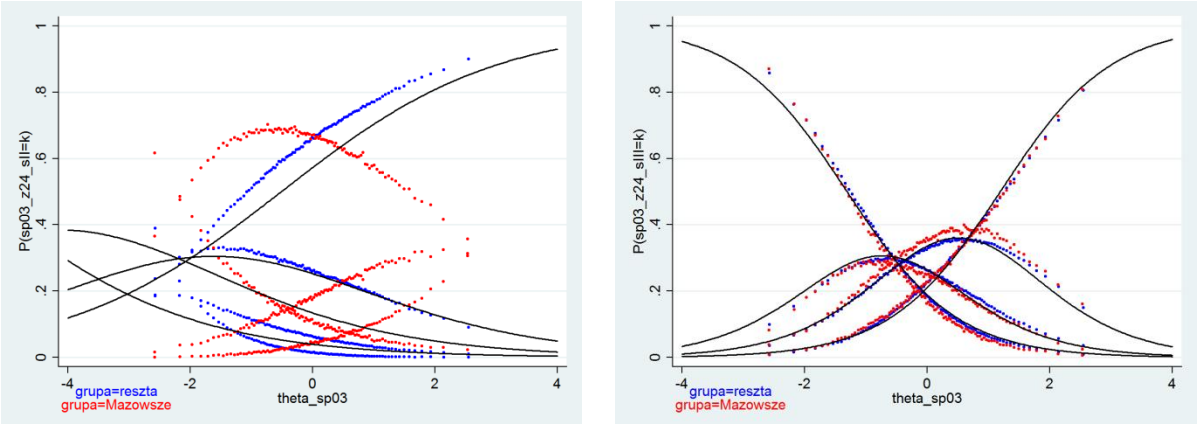
Rysunek 5.5. Odpowiedzi uczniów na zadania niedopasowane do modelu IRT ze względu na poziom umiejętności (dla zrównywania wyników sprawdzianu po szkole podstawowej)



Szczególnej analizie poddane zostało zadanie 24. z roku 2003. W tym roku egzaminatorzy z Mazowsza zostali inaczej (w stosunku do innych egzaminatorów) poinstruowani co do kryteriów oceniania tego otwartego zadania polonistycznego. Skutkowało to tym, iż przyjęli inny sposób oceniania, który miał wpływ na wyniki. Zadanie to zostało podzielone w naszej analizie na 4 części (z24_10, z24_sI, z24_sII, z24_sIII). Przeprowadzono analizę DIF, podobną do opisywanych wcześniej z tym, że grupą referencyjną w tej analizie są egzaminatorzy z całej Polski wyłączając Mazowsze. Grupą odniesienia - egzaminatorzy z Mazowsza. Analizę przedstawiono na Rysunku 5.6. Tylko dwa podpunkty tego zadania charakteryzują się znaczącym zróżnicowaniem funkcjonowania (DIF): z24_10, z24_sII. Podpunkty te zostały usunięte z analiz.

Rysunek 5.6. Analiza DIF zadań „problematycznych" ze względu na schemat oceniania dla zrównywania wyników sprawdzianu po szkole podstawowej

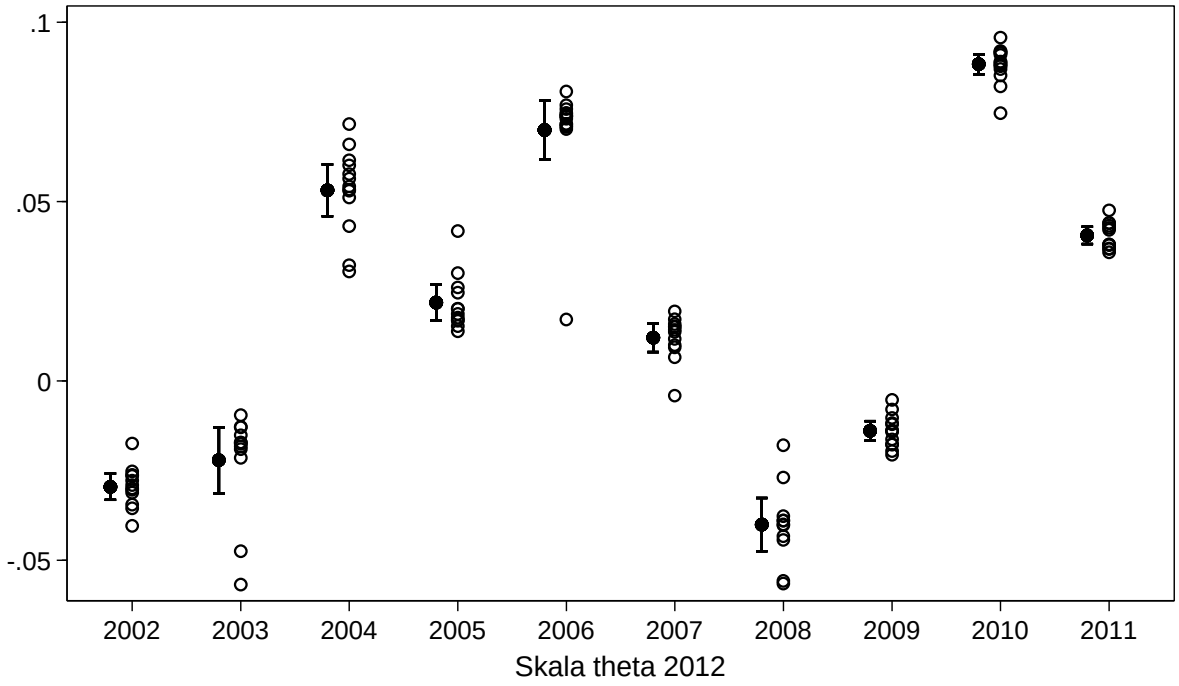


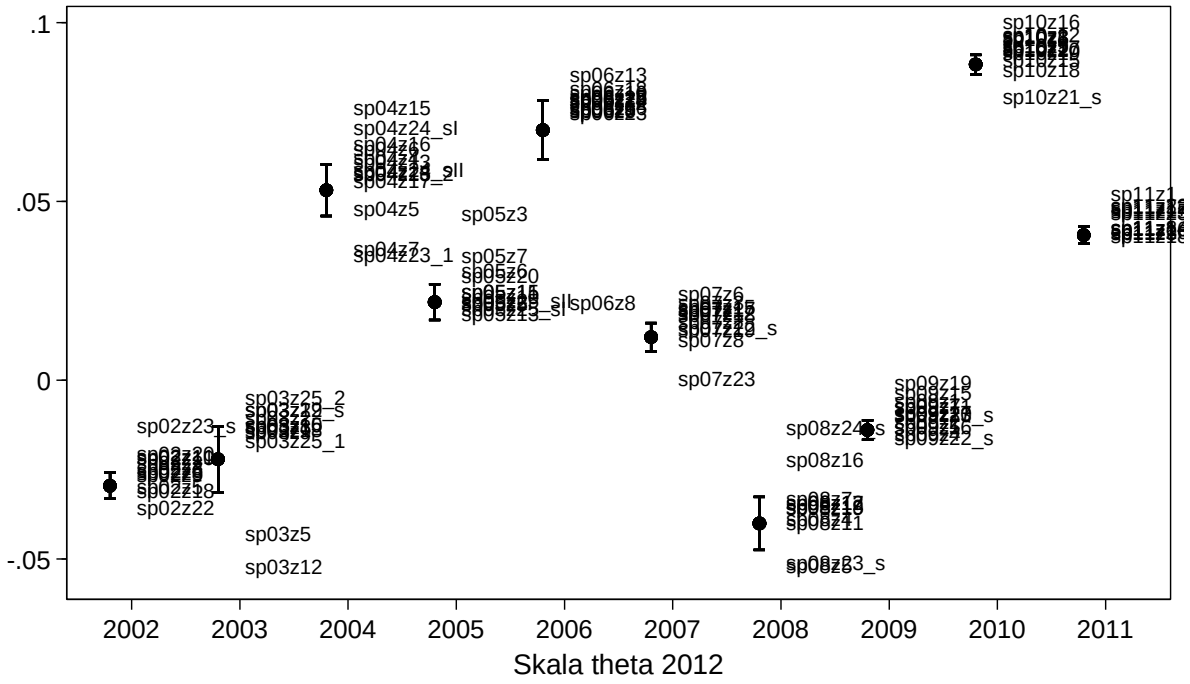


5.3.2. Analiza wrażliwości zrównywania na dobór zadań

Dla sprawdzianu w szóstej klasie szkoły podstawowej przedstawione zostały analogiczne analizy jak dla egzaminu gimnazjalnego opisywanego w punkcie 5.2.2. Zastosowano metody symulacyjne, które polegają na usuwaniu pojedynczych zadań z puli zrównującej i przeprowadzeniu zrównywania. Procedura polegała na wielokrotnym zrównywaniu liczbą n-1 zadań, gdzie n oznacza liczbę zadań zrównujących. Innymi słowy dokonywano wielu zrównań, w każdym ze zrównań pomijano jedno zadanie. Wyniki tej analizy przestawione zostały na Rysunku 5.7.

Rysunek 5.7. Rozkład wyników zrównania dla sprawdzianu po szkole podstawowej, lata 2002-2011 (górny panel bez zaznaczonych nazw zadań, dolny z zaznaczonymi nazwami zadań)





Zadania zrównujące ze sprawdzianu po szkole podstawowej poddane zostały analizom wrażliwości (oraz analizie DIF, której tutaj szczegółowo nie prezentujemy, gdyż daje identyczne rezultaty). Zadania odstające potraktowane zostały inaczej niż w egzaminie gimnazjalnym. W egzaminie gimnazjalnym wszystkie zadania wpływające nienaturalnie na zrównanie zostały usunięte ze zrównywania. W zrównywaniu wyników sprawdzianu zostały one potraktowane jako dodatkowe zadania służące estymacji poziomu umiejętności ucznia. Tym samym zadania te nie wnoszą bezpośrednio informacji do procesu zrównywania, ale pośrednio zwiększają precyzję zrównywania gdyż większa liczba zadań pozwala na precyzyjniejsze oszacowanie poziomu umiejętności w próbie wykorzystanej do zrównywania. Pełna lista takich zadań przedstawiona została w Tabeli 5.8.

Tabela 5.8. Zadania z badania zrównującego wyniki sprawdzianu, które ostatecznie nie zostały wykorzystanie jako zadania zrównujące

Rok					
2002	sp02z2	sp02z23_s	sp02z3		
2003	sp03z12	sp03z22_s	sp03z25_1	sp03z25_2	sp03z5
2004	sp04z15	sp04z23_1	sp04z5	sp04z7	
2005	sp05z10	sp05z13	sp05z25_sII	sp05z3	
2006	sp06z13	sp06z6	sp06z8		
2007	sp07z23	sp07z6			
2008	sp08z16	sp08z23_s	sp08z24_s	sp08z5	
2009	sp09z19	sp09z4			
2010	sp10z21_s	sp10z5			
2011	sp11z22	sp11z23			

5.3.3. Ostateczna pula zadań zrównujących wyniki sprawdzianu

W Tabeli 5.9 przedstawiono ostateczną liczbę zadań zastosowanych w zrównywaniu w podziale na każdy rok, co do którego używano procedury zrównywania. W zależności od roku użytych zostało od 24% do 50% zadań wykorzystywanych w egzaminie. Liczba zadań użytych do zrównania stanowi o precyzji zrównania wyników. Z doświadczeń międzynarodowych wiadomo, że po wykluczeniu w badaniu zadań o słabych właściwościach psychometrycznych około 1/4, 1/3 zadań używanych jest do zrównywania. Taka liczba pozwala na stosunkowo precyzyjne zrównanie wyników i ustrzeżenie się przed poważniejszymi błędami.

Tabela 5.9. Ostateczna liczba zadań używanych do zrównywania wyników sprawdzianu

Rok	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Liczba zadań wykorzystanych do zrównywania	9	7	9	8	12	10	7	11	13	10
Liczba zadań w arkuszu egzaminacyjnym	26	29	28	29	27	27	26	26	26	27
Procent liczby zadań wykorzystanych do zrównania	35%	24%	32%	28%	44%	37%	27%	42%	50%	37%

6. Wyniki zrównania

6.1. Wstęp

W 2012 roku przeprowadzono badanie umożliwiające zrównanie wyników sprawdzianu po klasie szóstej szkoły podstawowej dla lat 2002-2011, a także ponowne zrównanie wyników dla części matematyczno-przyrodniczej oraz humanistycznej egzaminu gimnazjalnego, które poszerzało zakres lat objętych zrównaniem o kolejny rocznik uczniów zdających ten egzamin.

Włączenie nowych danych z egzaminu gimnazjalnego, dostarczyło informacji o tym jak wyniki z roku 2011 mają się do wcześniejszych edycji egzaminów, ale również uzyskaliśmy nowe dane o zależności pomiędzy egzaminami gimnazjalnymi sprzed 2011 roku. W związku z tym procedurę zrównywania egzaminu gimnazjalnego przeprowadzono ponownie dla wszystkich lat, a co za tym idzie, ponownie dla wszystkich lat zrównane wyniki dla egzaminu gimnazjalnego zostaną przedstawione. Warto przy tym zauważyć, że zmiany w średnich zrównanych wynikach jakie uzyskano po badaniach przeprowadzonych w 2011 roku a raportowanych na następnych stronach są minimalne.

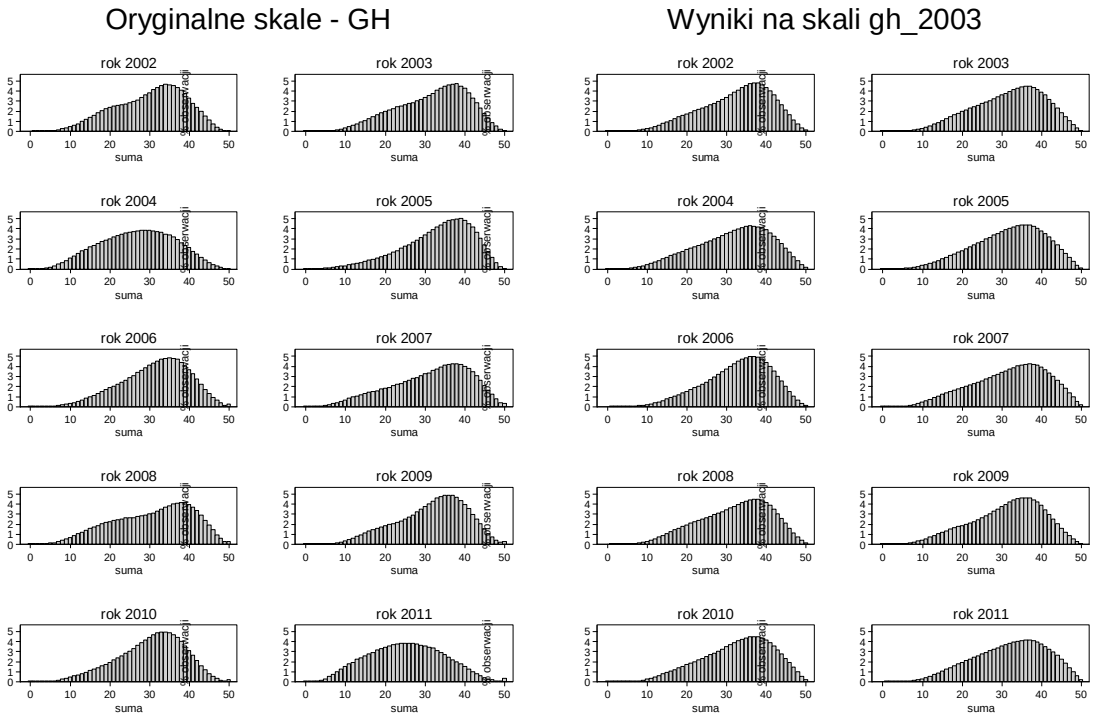
Prezentacja wyników dla egzaminu gimnazjalnego i dla sprawdzianu odbędzie się w analogiczny sposób. Najpierw omówione zostaną różnice pomiędzy rozkładami obserwowanych wyników sumarycznych na oryginalnych skalach oraz na skalach zrównanych. Przedstawienie rezultatów zrównania wyników obserwowanych zakończą tabele pozwalające na przekształcenie wyniku egzaminu z danego roku, na wynik w ustalonym odpowiednio dla egzaminu roku bazowym. W następnej kolejności opisane zostaną wyniki na skali zmiennej ukrytej, przekształconej do średniej 100 oraz odchyleniu standardowym 15.

6.2. Egzamin gimnazjalny

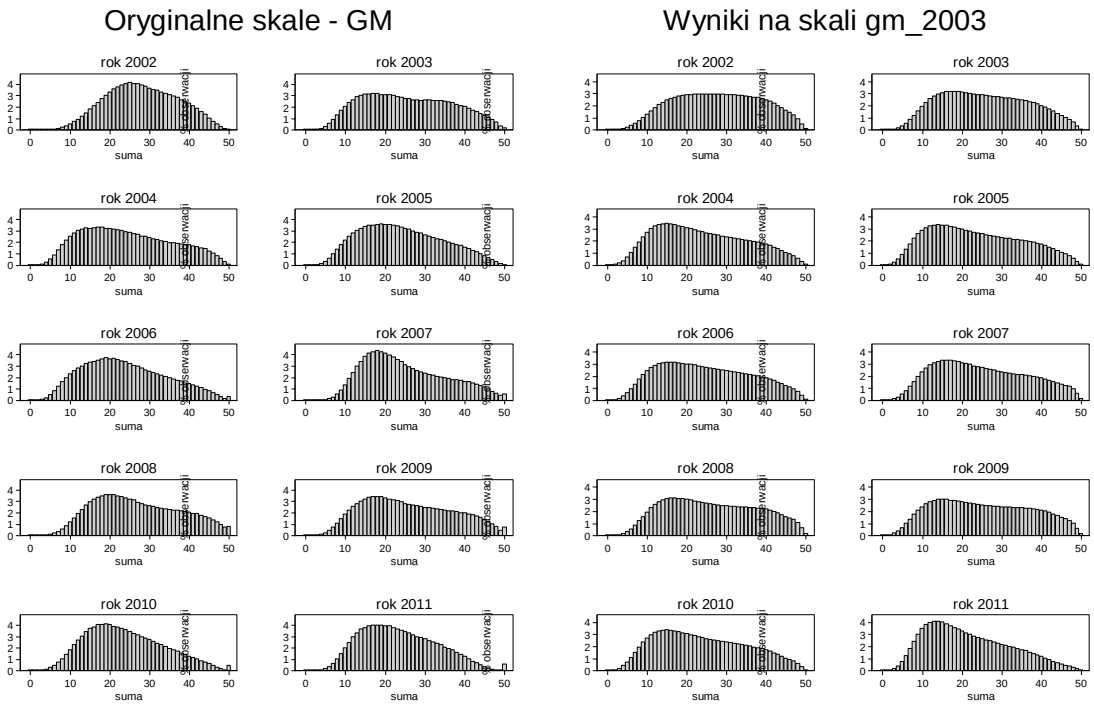
6.2.1. Zmiany trudności egzaminu gimnazjalnego w latach 2002-2011

W wyniku zastosowanej procedury zrównywania dysponujemy zakotwiczonym na wspólnej skali rozkładem poziomu umiejętności dla każdego roku oraz parametrami zadań opisującymi prawdopodobieństwo udzielenia określonej odpowiedzi na zadania egzaminacyjne w zależności od poziomu umiejętności uczniów. Informacje te pozwalają na oszacowanie, jak wyglądałby rozkład „zwykłej" sumarycznej liczby punktów z dowolnego egzaminu w latach 2002-2011, gdyby rozwiązywany był przez populację uczniów z dowolnej kohorty w latach 2002-2011. Za rok bazowy w studium zrównującym przyjęliśmy rok 2003. Rok 2003 wybrany został arbitralnie, przy czym wzięto pod uwagę, iż był to jeden z pierwszych egzaminów gimnazjalnych i jako taki stanowił dogodny punkt wyjściowy (pierwszym był egzamin z roku 2002, lecz właściwości psychometryczne egzaminu były relatywnie słabe, a procedury egzaminacyjne różniły się od stosowanych w kolejnych latach – z tego względu nie wybrano 2002 jako roku bazowego). Rysunki 6.1 i 6.2 przedstawiają porównanie rozkładów uzyskanych na egzaminach w kolejnych latach z rozkładami, które otrzymaliśmy po zrównaniu wyników i przełożeniu ich na skalę roku 2003. Innymi słowy rozkłady te mówią nam jak wyglądałyby wyniki uczniów z lat 2002, 2004-2011, gdyby wszyscy rozwiązywali egzamin z roku 2003.

Rysunek 6.1. Rozkład rzeczywistych wyników egzaminacyjnych (lewy panel) oraz wyników po zrównaniu przedstawionych na skali egzaminu z 2003 roku, część humanistyczna



Rysunek 6.2. Rozkład rzeczywistych wyników egzaminacyjnych (lewy panel) oraz wyników po zrównaniu przedstawionych na skali egzaminu z 2003 roku, część matematyczno-przyrodnicza



Zaobserwowano, że rozkłady przed zrównaniem i po zrównaniu różnią się od siebie znacząco. Po przekształceniu wszystkie rozkłady wyników przypominają bardziej rozkład z 2003 roku. W przypadku części humanistycznej objawia się to tym, iż wszystkie rozkłady wyników są łagodnie lewoskośne, szczególnie zmienia się tutaj rok 2004, który przed zrównaniem zdecydowanie bardziej zbliżał się do rozkładu normalnego. Zrównane rozkłady matematyczno-przyrodnicze są, podobnie jak w przypadku wyników obserwowalnych, prawoskośne, a ich wariancja zwiększa się. Po zrównaniu szczyt rozkładu stał się zdecydowanie mniej ostry (rozkłady bardziej spłaszczone). Takie zmiany rozkładów, przed i po zrównaniu, sugerują, iż kształt rozkładu wyników uwarunkowany był przede wszystkim przez właściwości pomiarowe egzaminów, w mniejszym stopniu przez zmiany umiejętności uczniów.

Warto przyjrzeć się dokładnie wynikom średnim. Trudno jest je bezbłędnie odczytać z rozkładu wyników, przedstawiono je zatem na Rysunkach 6.3 – 6.4. Zamieszczono tam średnie wyniki uzyskane przez uczniów na prawdziwych egzaminach oraz średnie wyniki po przekształceniu na skalę z roku 2003 (tabela wraz z wartościami średnich i odchylenia standardowego dla rozkładów oryginalnych, jak i zrównanych znajduje się w Aneksie). Analiza tych wyników pozwala na poczynienie bardzo interesujących obserwacji.

Zauważyć można (zwłaszcza dla części humanistycznej – Rysunek 6.3), że fluktuacja średnich z prawdziwych egzaminów gimnazjalnych między latami jest o wiele większa, niż fluktuacja między średnimi, jakie uzyskiwaliby uczniowie, gdyby pisali w każdej edycji test o takiej samej trudności, jaki 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 – Tabela 6.1) kontrastują z praktycznie minimalnymi różnicami (rzędu ±0,1 punktu) między średnimi dla tych samych lat na skali egzaminu z 2003 roku. Wyniki zrównania pokazują, że te różnice są konsekwencją znacznych zmian w łatwości egzaminów, a nie zmian w poziomie umiejętności uczniów.

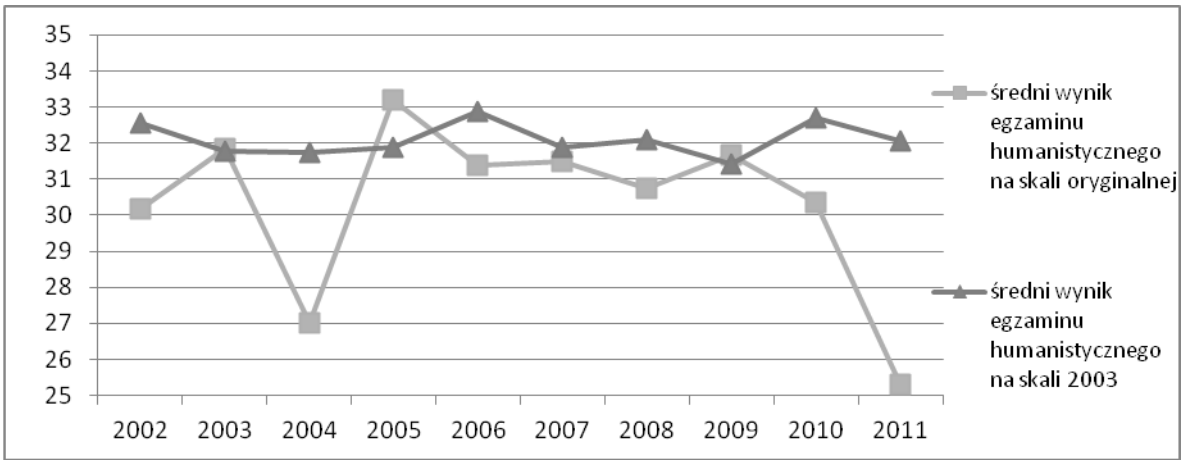
Pierwszy wniosek, jaki należy wyciągnąć z tej obserwacji, to „wątpliwa" przydatność skali sumarycznych wyników obserwowanych (lub procentu maksymalnej liczby punktów) do porównywania umiejętności uczniów między latami. Kolejnym wnioskiem jest podejrzenie braku odpowiednich procedur kontroli łatwości egzaminów podczas ich konstrukcji – wystąpiła fluktuacja łatwości egzaminów rzędu 6 punktów między dwoma sąsiednimi edycjami egzaminu, czyli aż 12% maksymalnego wyniku w teście.

Pewnego komentarza wymaga pojawienie się dla roku 2003 w Tabeli 6.1, a także na Rysunkach 6.3 i 6.4, zarówno oryginalnych parametrów egzaminu, jak i parametrów na skali zrównanej do egzaminu z roku 2003 – nie ma potrzeby zrównywania wyników egzaminu z roku 2003 do siebie samego. Dysponując parametrami modelu IRT dla rozkładu umiejętności uczniów oraz dla zadań w 2003 roku możemy w zupełnie analogiczny sposób jak dla innych lat oszacować rozkład wyników obserwowanych z roku 2003. Mimo iż faktycznie 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 w 2003 roku. Okazuje się, że dla obu egzaminów 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 „hipotetyczne" rozkłady wyników egzaminów z 2003 roku w innych latach.

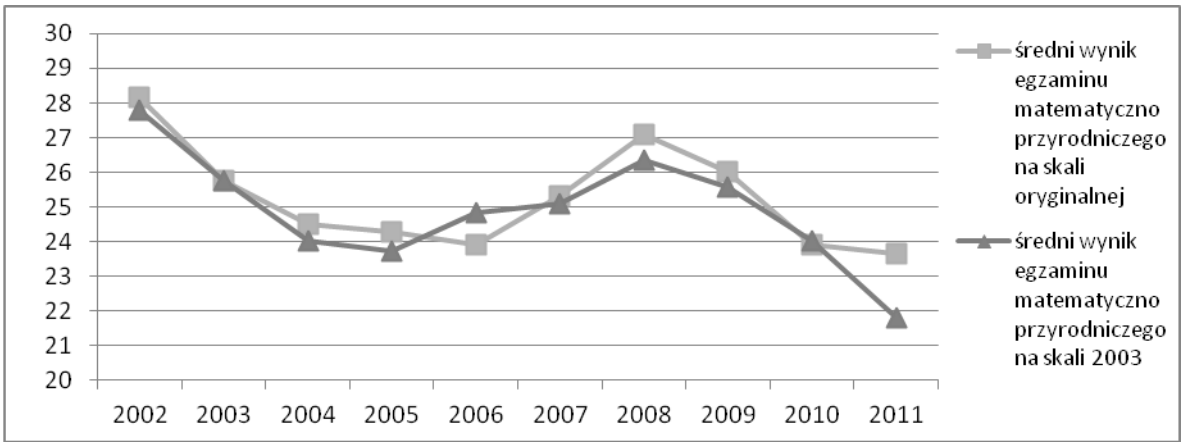
Tabela 6.1. Średnie oraz odchylenia standardowe wyników obserwowanych egzaminów gimnazjalnych dla oryginalnego testu oraz na skali wyników egzaminu z 2003 roku

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	32,6	8,8	8,8	28,2	27,8	8,9	10,7
2003	**31,8**	**31,8**	**8,9**	**8,8**	**25,7**	**25,7**	**10,9**	**10,8**
2004	27,0	31,7	9,2	9,2	24,5	24,0	11,0	11,1
2005	33,2	31,9	8,7	8,9	24,3	23,7	10,1	11,3
2006	31,4	32,9	8,4	8,3	23,9	24,8	10,3	11,2
2007	31,5	31,9	9,8	9,3	25,3	25,1	10,2	11,3
2008	30,7	32,1	9,8	8,9	27,1	26,4	10,7	11,3
2009	31,7	31,4	8,7	8,9	26,0	25,6	11,0	11,7
2010	30,3	32,7	8,4	8,9	23,9	24,0	9,6	11,2
2011	25,3	32,0	9,3	9,1	23,6	21,8	9,4	10,4

Rysunek 6.3. Średnie wyników obserwowanych dla humanistycznej części egzaminu gimnazjalnego dla oryginalnego testu oraz na skali wyników egzaminu z 2003 roku


35
34
33
32
31
30
29
28
27
26
25
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
średni wynik egzaminu humanistycznego na skali oryginalnej
średni wynik egzaminu humanistycznego na skali 2003

Rysunek 6.4. Średnie wyników obserwowanych dla matematyczno-przyrodniczej części egzaminu gimnazjalnego dla oryginalnego testu oraz na skali wyników egzaminu z 2003 roku



Posiadając rozkłady wyników obserwowanych egzaminów humanistycznego oraz matematyczno-przyrodniczego z 2003 roku dla wszystkich populacji uczniów w latach 2002-2011 możliwe jest stworzenie tablic przeliczeniowych pozwalających przyporządkować uczniowi piszącemu egzamin 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 (por. Rozdział 4.4) z roku X (zaobserwowane) z wynikami sumarycznymi w teście 2003 dla tego roku (zasymulowane zgodnie z modelem IRT). Taką tablicę przeliczeniową dla części humanistycznej i matematyczno-przyrodniczej egzaminu gimnazjalnego przedstawiono w tabelach 6.2 i 6.3.

W tabelach 6.2 oraz 6.3 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 w oparciu o wykorzystany do zrównania model statystyczny. Okazuje się, że dla testu humanistycznego model statystyczny sugeruje błędne przeliczenia dla uczniów uzyskujących w egzaminie 0-3 punktów, 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. Mając na względzie, że w roku 2003 w teście humanistycznym 0-3 punkty uzyskało 57 uczniów z 551150 (0,0103% obserwacji), a w teście matematycznym 0 punktów uzyskało 9 uczniów z 548716 (0,0016% obserwacji), należy uznać, że ów rejon „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 z wykorzystaniem modelu IRT.

Analiza danych ukazanych w tabelach przeliczeniowych pozwala na zauważenie kolejnych kilku bardzo interesujących zależności między wynikami uczniów z różnych lat w egzaminach, w których arkusze egzaminacyjne różnią się łatwością. Dla przykładu (Tabela 6.2) przedstawiono ucznia, który w 2004 roku na teście humanistycznym uzyskał 27 punktów oraz ucznia, który na teście humanistycznym w 2005 uzyskał również 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, gdy tymczasem niezrównane wyniki z egzaminów przez nich zdawanych sugerują taki sam poziom umiejętności!

Podobnego rzędu różnice między latami 2004 oraz 2005 dla testu humanistycznego (7-8 punktów) można zaobserwować w całym przedziale punktów od 21 do 37, co jest bardzo doniosłe, gdyż ten przedział znajduje się w centrum rozkładu wyników i odpowiada 60,1% oraz 54,0% całej populacji uczniów! 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 pisać…

Bardzo istotne w omawianym przykładzie jest również to, że dla uczniów uzyskujących w teście humanistycznym w latach 2004 oraz 2005 bardzo wysokie (powyżej 45) lub bardzo niskie (poniżej 10) wyniki, różnica na skali wyników testu z roku 2003 maleje do 1-2 punktów. Zaobserwowana różnica w średniej łatwości całych egzaminów na skali z 2003 roku wynosząca 6,2 punktu (Tabela 6.1) jest zatem średnią z 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 między egzaminami ma nieliniowy charakter, to należy też wyciągnąć wniosek, że stosowanie wszelkich poprawek o liniowym charakterze (jak na przykład standaryzacja) nie będzie w stanie rozwiązać problemu.

Dla kontrastu do skrajnego przykładu różniących się znacznie łatwością testów humanistycznych z lat 2004 oraz 2005 można wziąć testy matematyczno-przyrodnicze z tych samych lat i również uczniów uzyskujących 27 punktów (Tabela 6.3). 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 i tak samo w przypadku ucznia piszącego 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 łatwości, a jednocześnie (por. np. Rysunek 6.2 lub 6.4) ze zrównania wynika znaczna zmiana w poziomie umiejętności między tymi rocznikami.

Bez przeprowadzenia zrównania wyników przedstawione na przykładzie egzaminu gimnazjalnego w 2004 i w 2005 roku diametralnie odmienne interpretacje różnic w zaobserwowanych średnich wynikach byłyby niemożliwe. W przypadku części humanistycznej egzaminu różnica dotyczyła głównie różnic w łatwości testów humanistycznych, natomiast w przypadku części matematyczno-przyrodniczej różnicy średnich należy głównie upatrywać w różnicy poziomów umiejętności matematyczno-przyrodniczych uczniów między latami.

Tabela 6.2. Tablica przeliczeniowa obserwowanych wyników humanistycznej części egzaminu gimnazjalnego na wyniki obserwowane w roku 2003

Wynik z egzaminu	Przeliczenie wyników na skalę egzaminu z roku 2003									
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
0	.	2	1	1	2	2	2	1	2	2
1	2	2	2	2	2	3	3	2	3	3
2	3	3	3	3	3	3	4	2	4	4
3	4	4	4	4	4	5	5	3	5	6
4	5	4	6	5	5	6	6	4	6	7
5	6	5	7	6	6	7	7	4	7	9
6	7	6	8	7	7	8	8	5	8	10
7	8	7	9	8	8	9	9	6	9	12
8	9	8	11	9	9	10	10	7	10	13
9	10	9	12	10	10	11	12	8	11	15
10	12	10	13	11	11	12	13	9	12	16
11	13	11	14	11	12	13	14	10	13	17

12	14	12	16	12	13	14	15	11	14	18
13	15	13	17	13	14	15	16	12	15	19
14	16	14	18	14	15	16	17	13	16	21
15	17	15	19	15	16	17	18	14	17	22
16	18	16	20	15	17	17	19	15	18	23
17	19	17	22	16	18	18	20	16	19	24
18	20	18	23	17	19	19	21	18	19	25
19	22	19	24	18	20	20	22	19	20	26
20	23	20	25	19	22	21	23	20	21	27
21	24	21	26	20	23	22	24	21	23	29
22	25	22	27	20	24	23	25	22	24	30
23	26	23	28	21	25	24	26	23	25	31
24	27	24	30	22	26	25	26	24	26	32
25	28	25	31	23	27	26	27	25	27	33
26	29	26	32	24	28	26	28	26	28	34
27	30	27	33	25	29	27	29	27	29	35
28	31	28	33	26	30	28	30	28	30	36
29	32	29	34	27	31	29	31	29	31	36
30	33	30	35	28	32	30	32	30	32	37
31	34	31	36	29	33	31	32	31	33	38
32	35	32	37	30	34	32	33	32	35	39
33	35	33	38	31	35	33	34	33	36	40
34	36	34	39	32	36	34	35	34	37	41
35	37	35	40	33	37	35	36	35	38	41
36	38	36	41	34	37	36	37	36	39	42
37	39	37	41	35	38	37	38	37	40	43
38	40	38	42	37	39	38	38	38	41	44
39	41	39	43	38	40	39	39	39	42	44
40	42	40	44	39	41	40	40	40	43	45
41	43	41	45	40	42	41	41	41	44	46
42	44	42	45	41	43	42	42	42	45	46
43	45	43	46	43	44	43	43	43	46	47
44	46	44	47	44	45	44	44	44	47	48
45	47	45	48	45	46	45	45	45	47	48
46	48	46	48	46	47	46	46	46	48	48
47	48	47	49	47	48	47	47	47	49	49
48	49	48	49	48	48	48	48	48	49	49
49	50	49	50	49	49	49	49	48	49	49
50	50	50	50	50	49	50	49	49	50	50

Tabela 6.3. Tablica przeliczeniowa obserwowanych wyników matematyczno-przyrodniczej części egzaminu gimnazjalnego na wyniki obserwowane w roku 2003

Wynik z egzaminu	**Przeliczenie wyników na skalę egzaminu z roku 2003**									
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
0	0	1	0	0	0	0	0	0	0	0
1	1	1	1	0	1	1	1	0	0	1
2	1	2	2	1	2	1	1	1	1	1
3	2	3	3	2	3	1	2	2	2	2
4	2	4	4	3	4	2	3	2	3	3
5	3	5	5	3	5	3	4	3	4	3
6	4	6	6	4	6	4	5	4	5	4
7	5	7	7	5	7	4	5	5	5	5
8	6	8	8	6	8	5	6	6	6	6
9	6	9	9	7	9	6	7	7	7	7
10	7	10	10	8	10	7	8	8	8	8

11	8	11	11	9	11	8	9	9	9	9
12	9	12	12	10	12	9	10	10	10	9
13	10	13	13	11	13	11	11	11	11	10
14	11	14	13	12	14	12	12	12	12	11
15	12	15	14	13	15	13	13	13	13	12
16	13	16	15	14	16	14	14	15	14	13
17	14	17	16	15	17	16	15	16	15	14
18	15	18	17	16	18	17	16	17	17	15
19	16	19	18	17	19	18	17	18	18	16
20	17	20	19	18	20	19	19	19	19	17
21	19	21	20	20	22	21	20	20	20	18
22	20	22	21	21	23	22	21	22	22	19
23	21	23	22	22	24	23	22	23	23	20
24	23	24	23	23	25	25	23	24	24	22
25	24	25	25	25	27	26	25	25	26	23
26	25	26	26	26	28	27	26	26	27	24
27	27	27	27	27	29	28	27	27	28	25
28	28	28	28	28	30	29	28	28	30	26
29	29	29	29	29	31	30	29	29	31	28
30	31	30	30	31	32	31	30	30	32	29
31	32	31	31	32	33	32	31	31	33	30
32	33	32	32	33	34	33	32	32	34	31
33	34	33	33	34	36	34	33	33	35	32
34	35	34	34	35	37	35	34	34	37	34
35	37	35	35	36	37	36	35	35	38	35
36	38	36	36	37	38	37	36	36	39	36
37	39	37	37	38	39	38	37	37	40	37
38	40	38	37	39	40	39	38	38	41	38
39	41	39	38	40	41	40	39	39	41	40
40	42	40	39	41	42	41	40	40	42	41
41	43	41	40	42	43	42	41	41	43	42
42	44	42	41	43	44	43	42	42	44	43
43	45	43	42	44	45	44	43	43	45	44
44	46	44	44	45	46	45	44	44	46	45
45	47	45	45	46	46	46	45	45	47	46
46	48	46	46	47	47	46	46	46	47	47
47	48	47	47	48	48	47	47	47	48	47
48	49	48	48	49	48	48	47	48	48	47
49	49	49	49	49	49	48	48	48	48	47
50	50	50	50	50	49	49	49	49	49	48

6.2.2. Wyniki gimnazjalistów w latach 2002-2011 na skali zmiennej ukrytej

W poprzednim punkcie przedstawiono analizę trudności egzaminów z różnych lat. W tej części skupiono się na analizie poziomu umiejętności uczniów. Za pomocą danych egzaminacyjnych postarano się odpowiedzieć na fundamentalne pytanie z perspektywy szeroko rozumianej polityki oświatowej: Czy poziom umiejętności uczniów polskich gimnazjów wzrósł, zmalał, czy może nie zmienił się znacząco w ciągu ostatnich 10 lat?

Wyniki zrównywania zakotwiczone zostały w roku 2003. W czasie procesu zrównywania średnia umiejętności egzaminacyjnych uczniów ustawiona została na 0, a odchylenie standardowe na 1. Był to zabieg formalny, bez którego nie można by dokonać zrównywania wyników. Aby ułatwić prezentację wyników, przeskalowano je na skalę o średniej 100 i odchyleniu standardowym 15 dla roku 2003. Taka skala jest łatwiejsza do prezentacji, ponieważ nie daje ujemnych wyników, jest jedną z najbardziej znanych skal standardowych (IQ), ponadto używana jest już do prezentowania wyników polskich badań np. w EWD oraz OBUT.

6.2.2.1. Wyniki zrównywania w części humanistycznej

W Tabeli 6.4 przedstawiono średni poziom umiejętności uczniów zdających część humanistyczną egzaminu gimnazjalnego w latach 2002-2011. W pierwszej kolumnie podany został rok, w drugiej średni poziom umiejętności (średnia), w kolejnej przedstawiony jest błąd standardowy wokół oszacowania średniej – całkowity i w rozbiciu na dwa składniki. Pierwszy, główny, składnik błędu jest błędem zrównywania, który wynika głównie[24] z doboru do badania zrównującego ograniczonej losowej próby uczniów i został on oszacowany za pomocą procedury *bootstrap* opisanej w Rozdziale 4. Drugi, mniejszy, składnik błędu uwzględnia liczebność kohorty uczniów piszących egzamin w danym roku oraz jego rzetelność – określa on przedział ufności wokół wyniku uczniów w danym roku bez uwzględnienia procedury zrównywania.

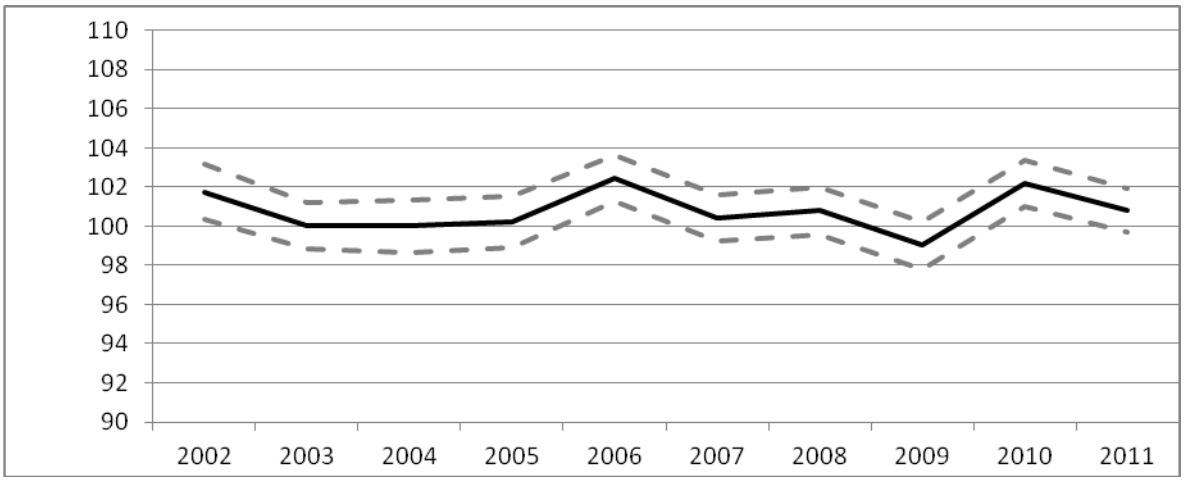
Tabela 6.4. Średnie wyniki uczniów szkół gimnazjalnych w latach 2002-2011, wyniki zrównane skala 100; 15 zakotwiczona w roku 2003, część humanistyczna

Rok	średnia	błąd		
		zrównywania	losowy	razem
2002	101,75	0,717	0,020	0,717
2003	100,00	0,600	0,022	0,600
2004	100,00	0,675	0,023	0,675
2005	100,24	0,667	0,023	0,667
2006	102,48	0,602	0,021	0,602
2007	100,41	0,608	0,024	0,608
2008	100,78	0,604	0,024	0,605
2009	99,02	0,607	0,024	0,607
2010	102,20	0,592	0,025	0,592
2011	100,84	0,574	0,027	0,574

[24] Nie jest to jedyne źródło błędu mogące wpływać na precyzję zrównywania. Oprócz błędu wynikającego z doboru próby badawczej (błąd próbkowania) w procesie zrównywania, w przyjętym schemacie badawczym uwidacznia się również błąd związany z wyborem zadań. W idealnej sytuacji do zrównania można wykorzystać wszystkie zadania pojawiające się w latach 2002-2011. Byłoby to jednak zadanie zbyt kosztowne. Dlatego do zrównywania, co jest powszechną praktyką w metodologii światowej, wykorzystana została tylko część zadań. W takim podejściu istnieje jednak pewien problem: różne konfiguracje zadań prowadzą do różnych wyników zrównania (por. Rozdział 5). Innymi słowy, jeżeli do badania wykorzystujemy jedynie podróbkę zadań, powinno brać się pod uwagę dodatkowy błąd wynikający z tego, iż do zrównania wybrana została specyficzna próbka zadań. Istnieje tutaj analogia między błędem próbkowania a opisywanym typem błędu – tak jak w przypadku uczniów, jeżeli badalibyśmy całą populację, nie trzeba by szacować błędu próbkowania, tak i w przypadku zrównywania dla wszystkich zadań nie trzeba by szacować błędu związanego z doborem zadań. Szacowanie błędu doboru zadań w tak skomplikowanym schemacie zrównywania jak przyjęty w tym badaniu, nie jest rzeczą prostą. Nie istnieją analityczne rozwiązania, a co do procedur replikacyjnych, to nie osiągnięto konsensusu co do ich skuteczności. Uczulamy zatem czytelnika na to, iż przedziały ufności i błędy standardowe, konstruowane w tym raporcie jedynie na podstawie błędu próbkowania, mogą być w pewnym stopniu niedoszacowane. Proponujemy traktować je z należytym rozsądkiem i jako wskazówkę, a nie jako bazę do formalnych testów statystycznych.

Na Rysunku 6.5 w graficzny sposób przedstawiono wyniki zrównywania z Tabeli 6.4. Ciągła linia oznacza średni poziom umiejętności w danym roku (gdzie, 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 być bardzo stabilny i nie wykazuje znaczącego trendu. Jest to jedyny silny wniosek wynikający z przedstawionych danych: umiejętności humanistyczne uczniów nie zmieniły się znacząco w ciągu ostatnich 10 lat.

Rysunek 6.5. Średnie wyniki uczniów szkół gimnazjalnych w latach 2002-2011, wyniki zrównane skala 100; 15 zakotwiczona w roku 2003, część humanistyczna



Jedyne wyraźne choć niewielkie zmiany poziomu umiejętności uczniów można odnotować w roku 2002, 2010, a przede wszystkim 2006 (rocznik ten odznacza się bowiem największym poziomem umiejętności). Wzrostom poziomu umiejętności zaobserwowanym w wymienionych rocznikach towarzyszyło obniżenie poziomu umiejętności w kolejnych, tak, że w perspektywie dziesięcioletniej nie zaobserwowano żadnego trendu, trudno stwierdzić czy mamy tu do czynienia z jakimiś specyficznymi cechami opisywanych kohort, przeprowadzanego egzaminu, czy właściwości przyjętego schematu zrównywania.

6.2.2.2. Wyniki zrównywania w części matematyczno-przyrodniczej

W Tabeli 6.5 i na Rysunku 6.6 przedstawiono średnie wyniki uczniów z egzaminu gimnazjalnego w części matematyczno-przyrodniczej po dokonaniu zrównania. Tak jak w przypadku części humanistycznej dla każdego roku 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. Średni poziom umiejętności wraz z zarysowanymi przedziałami ufności przedstawiono na Rysunku 6.6.

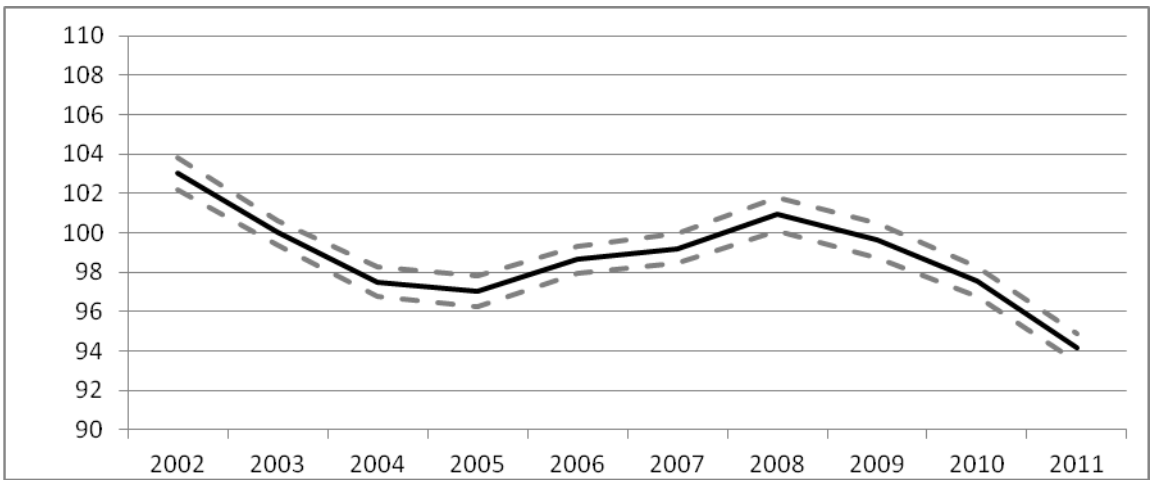
Tabela 6.5. Średnie wyniki uczniów szkół gimnazjalnych w latach 2002-2011, wyniki zrównane skala 100; 15 zakotwiczona w roku 2003, część matematyczno-przyrodnicza

Rok	średnia	błąd		
		zrównywania	losowy	razem
2002	103,01	0,422	0,020	0,423
2003	100,00	0,328	0,021	0,329
2004	97,51	0,380	0,023	0,380
2005	97,04	0,403	0,024	0,404

2006	98,65	0,346	0,024	0,347
2007	99,19	0,381	0,024	0,382
2008	100,93	0,437	0,025	0,438
2009	99,62	0,460	0,027	0,461
2010	97,53	0,374	0,027	0,375
2011	94,15	0,378	0,024	0,379

Zrównane wyniki egzaminu gimnazjalnego w części matematyczno-przyrodniczej pokazują spadek średniego poziomu umiejętności polskich gimnazjalistów 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 2008-2011. Należy przy tym zaznaczyć, iż trendy są bardzo niewielkie i wszelkie interpretacje mogące się nasuwać przy analizowaniu tych zmian trzeba traktować bardzo ostrożnie.

Rysunek 6.6. Średnie wyniki uczniów szkół gimnazjalnych w latach 2002-2011, wyniki zrównane skala 100; 15 zakotwiczona w roku 2003, część matematyczno-przyrodnicza



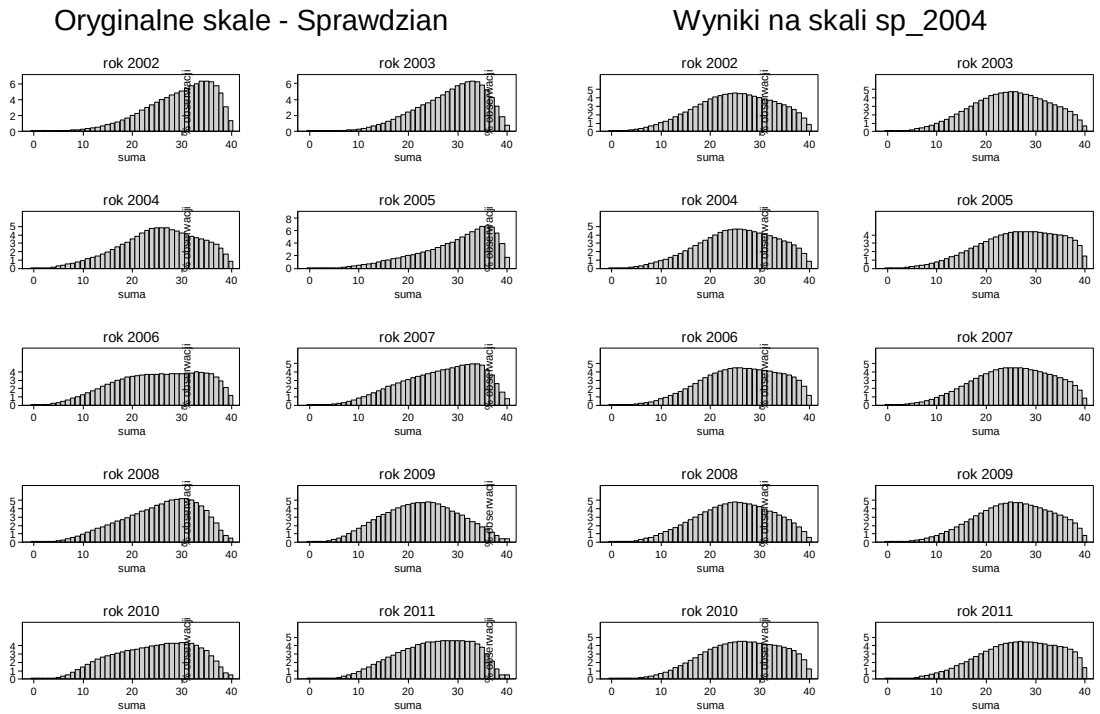
6.3. Sprawdzian po szkole podstawowej

6.3.1.1. Zmiany trudności sprawdzianu po szkole podstawowej 2002-2011

Zrównując wyniki sprawdzianu w szóstej klasie szkoły podstawowej za rok bazowy, do którego dokonano odniesienia wyników pozostałych edycji egzaminów przyjęto rok 2004. Rok 2002 wykluczono jako kandydata na rok bazowy z analogicznych powodów jak to poczyniono w przypadku egzaminu gimnazjalnego – była to pierwsza edycja sprawdzianu. Natomiast rok 2003 nie został przyjęty jako rok bazowy, ponieważ w 2003 roku w jednej z Okręgowych Komisji Egzaminacyjnych odmiennie zinterpretowano klucz oceny zadania 24., co spowodowało istotne zróżnicowane funkcjonowanie tego zadania (Aneks: *Psychometryczne właściwości zadań egzaminacyjnych*).

Rozkłady wyników obserwowanych sprawdzianu na oryginalnej skali oraz po zrównaniu do skali z 2004 roku są przedstawione w postaci histogramów na Rysunku 6.7, średnie w poszczególnych latach naniesiono na wykresie na Rysunku 6.8, zaś średnie i odchylenia standardowe zebrano w Tabeli 6.6.

Rysunek 6.7. Rozkład rzeczywistych wyników egzaminacyjnych (lewy panel) oraz wyników po zrównaniu przedstawionych na skali egzaminu z 2004 roku



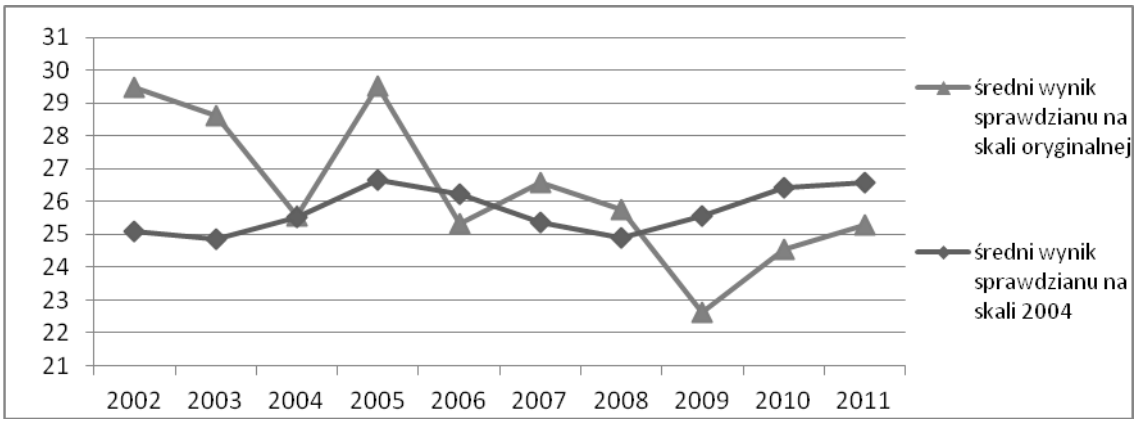
Zestawiając wyniki obserwowane ze zrównanymi do roku bazowego dla sprawdzianu można wyciągnąć podobny ogólny wniosek do tego jaki sformułowano w przypadku części humanistycznej egzaminu gimnazjalnego. Mianowicie rozbieżności pomiędzy obserwowanymi wynikami dla poszczególnych edycji sprawdzianu są o wiele większe niż pomiędzy wynikami zrównanymi. Jest to sygnałem, że występujące między latami fluktuacje rozkładu wyników sprawdzianu są przede wszystkim konsekwencją zmian w łatwości poszczególnych arkuszy egzaminacyjnych, a w mniejszym stopniu zmian w poziomie umiejętności kolejnych populacji uczniów piszących sprawdzian w szóstej klasie szkoły podstawowej. Praktyczny wymiar wielkości tych różnic stanie się bardziej jasny, gdy w dalszych akapitach przyjrzymy się tablicom przeliczeniowym dla wyników obserwowanych.

W szczególności średnie obserwowane wyniki sprawdzianu przed zrównaniem, między latami 2002 a 2011 pokazują trend spadkowy – średnie wyniki w pierwszych dwóch edycjach wynosiły odpowiednio 29,5 oraz 28,5 punktu, gdy w ostatnich trzech badanych latach (2009-2011) obserwujemy odpowiednio średnie 22,6, 24,6 oraz 25,3 punktu, co oznacza spadek rzędu od około 3 do około 7 punktów na 40 możliwych do uzyskania w teście. Po zrównaniu okazuje się natomiast, że wyniki wyrażone na wspólnej skali 2004 dla trzech ostatnich lat są tak naprawdę wyższe niż dla pierwszych edycji sprawdzianu, a jednocześnie różnice między wszystkimi 10 edycjami sprawdzianu nie wykraczają co do absolutnej wartości poza 1,7 punktu.

Wynik ten świadczy, że nie udało się w przypadku sprawdzianu utrzymać stabilnej trudności egzaminu na przestrzeni lat. Trzy z czterech najsilniej odchylających się pod względem trudności sprawdzianów (2002, 2003 oraz 2005) były jednocześnie sprawdzianami na tyle łatwymi, że rozkłady obserwowanych sumarycznych punktów w teście były znacznie lewostronnie skośne (Rysunek 6.7). Taka skośność wyników sumarycznych każe negatywnie ocenić sprawdzian w szóstej klasie szkoły

podstawowej również pod względem pomiarowym – różnicuje on badane osoby w sposób daleki od optymalnego.

Rysunek 6.8. Średnie wyników obserwowanych sprawdzianu dla oryginalnego testu oraz na skali wyników egzaminu z 2004 roku



W Tabeli 6.6 oraz na Rysunku 6.8 zamieszczono, podobnie jak w przypadku gimnazjum, wyniki zrównane dla roku bazowego, które pozwalają na ocenę dobroci przewidywania obserwowanych wyników przez dopasowany model IRT. Z Tabeli 6.6 w szczególności widać, że średnia dla skali oryginalnej w 2004 roku różni się o 0.1 punktu w porównaniu ze skalą zrównaną dla tego roku, a odchylenia standardowe nie różnią się do pierwszego miejsca po przecinku.

Tabela 6.6. Średnie oraz odchylenia standardowe wyników obserwowanych sprawdzianu dla oryginalnego testu oraz na skali wyników egzaminu z 2004 roku

Rok	Sprawdzian			
	średnia na skali oryginalnej	średnia na skali 2004	odch. stand. na skali oryginalnej	odch. stand. na skali 2004
2002	29,5	25,1	6,8	7,9
2003	28,6	24,9	6,7	7,7
2004	**25,6**	**25,5**	**7,8**	**7,8**
2005	29,5	26,6	7,4	7,9
2006	25,3	26,2	8,6	7,8
2007	26,6	25,4	7,8	7,9
2008	25,8	24,9	7,5	7,7
2009	22,6	25,6	7,6	7,6
2010	24,6	26,4	8,0	7,7
2011	25,3	26,6	7,5	7,8

W Tabeli 6.7 przedstawiono tablice przeliczeniowe pozwalające na określenie, jaki wynik uzyskałby uczeń, który w danym roku uzyskał określoną liczbę punktów, na skali zrównanej do roku bazowego 2004. Dla celów diagnostycznych uwzględniono również możliwość przeliczenia roku 2004 do siebie samego – widzimy, że w przypadku każdego wyniku punktowego jest on przeliczany na siebie samego (w przypadku egzaminów gimnazjalnych występowały bardzo wąskie zakresy punktów, przy których w analogicznej sytuacji występowała rozbieżność zaznaczona na szaro w Tabelach 6.2 i 6.3). Pełna zgodność przy przeliczaniu ekwicentylowym wyników zaobserwowanych w 2004 roku na wyniki

przewidywane dla tego roku na podstawie modelu IRT stanowi jeszcze silniejsze pozytywne świadectwo dobroci dopasowania modelu IRT dla celów zrównywania wyników obserwowanych dla tego roku niż zgodność pierwszych dwóch momentów (średniej i odchylenia standardowego), jaką zobaczyć można we wcześniej omówionej Tabeli 6.6.

W sprawdzianie zaobserwować można między latami podobnie znaczące fluktuacje średniego wyniku ze względu na zmiany w trudności arkuszy jak to miało miejsce w przypadku części humanistycznej egzaminu gimnazjalnego (por. Rysunek 6.8 oraz 6.3). W związku z tym spodziewać się można analogicznie jak wcześniej, że dla odstających lat uczniowie uzyskujący takie same niezrównane wyniki będące sumą punktów będą mieli faktycznie bardzo duże różnice wyników po przeliczeniu na skalę zrównaną, niezależnie od faktycznych zmian poziomu umiejętności między latami. Przy czym ta prawidłowość powinna w największym wymiarze przejawiać się w środku skali.

Analizując wartości w Tabeli 6.7 obserwujemy dokładnie taką prawidłowość jakiej się spodziewaliśmy. Dla przykładu, wynik ucznia, który w 2002 lub 2003 roku uzyskał 20 punktów, powinien zostać przeliczony na zrównany wynik niższy odpowiednio o 6 lub 5 punktów na skali zrównanej, co odpowiada różnicom rzędu 15-12,5% wyniku sumarycznego. Jednocześnie różnice między średnimi wynikami po zrównaniu między latami 2002-2004 są poniżej jednego punktu. Dokładnie tak duże rozbieżności dla wybranych lat dotyczą nie tylko uczniów o wynikach surowych 20, ale dość szerokiego przedziału punktowego powyżej tego punktu – aż do wartości 30. Rozbieżności poniżej 20 oraz powyżej 30 nie są o wiele mniejsze, zaczynają raptownie maleć dopiero na krańcach skali. Taka nieliniowość przekształcenia zrównującego, jak już wspomniano przy okazji egzaminu gimnazjalnego, wyklucza stosowanie do zrównywania wyników przekształceń liniowych.

Wybrane zestawienie sąsiadujących lat 2002-2004 bynajmniej nie jest jednak najbardziej skrajnym przypadkiem wahań trudności arkuszy egzaminacyjnych sprawdzianu. Najbardziej skrajne zestawienie powstaje, jeżeli porównamy ze sobą zrównane wyniki między rokiem 2002 (znacznie łatwiejszy arkusz niż w 2004), a rokiem 2009 (znacznie trudniejszy arkusz niż w 2004). Większość uczniów uzyskujących bez zrównania takie same wyniki w sprawdzianach 2002 oraz 2009 (zakres punktów od 15 do 34) po zrównaniu uzyskuje wyniki różniące się aż o 8-9 punktów, co stanowi już 20-22,5% całej rozpiętości skali. Oczywiście różnica średnich zrównanych wyników między populacjami 2002 oraz 2009 jest w tym kontekście zupełnie niezauważalna – pół punktu (Tabela 6.6). Taka sytuacja nie powinna mieć miejsca.

Tabela 6.7. Tablica przeliczeniowa obserwowanych wyników sprawdzianu po szkole podstawowej na wyniki obserwowane w roku 2004

Wynik z egzaminu	Przeliczenie wyników na skalę egzaminu z roku 2004									
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
0	0	0	0	0	1	0	0	0	1	0
1	1	0	1	1	2	1	1	1	1	1
2	1	1	2	1	3	1	1	2	2	2
3	2	1	3	2	4	2	2	3	3	2
4	2	2	4	3	5	3	3	5	4	3
5	3	3	5	3	7	4	4	6	6	5
6	4	4	6	4	8	5	5	7	7	6
7	4	4	7	5	9	6	6	9	8	7
8	5	5	8	6	10	7	7	10	10	8
9	6	6	9	7	11	8	9	11	11	9
10	6	6	10	8	12	9	10	13	12	10
11	7	7	11	9	13	10	11	14	14	12
12	8	8	12	10	14	11	12	15	15	13

13	9	9	13	11	15	12	13	16	16	14
14	9	10	14	12	16	13	14	17	17	15
15	10	11	15	13	17	14	15	18	18	16
16	11	12	16	14	18	15	15	19	19	17
17	12	13	17	14	19	16	16	20	20	18
18	13	13	18	15	20	17	17	21	21	19
19	14	14	19	16	21	18	18	22	21	20
20	14	15	20	17	22	19	19	23	22	21
21	15	16	21	18	23	20	20	24	23	22
22	16	17	22	19	23	21	21	25	24	23
23	17	18	23	20	24	22	22	26	25	24
24	18	19	24	20	25	22	23	27	26	25
25	19	20	25	21	26	23	24	28	27	26
26	20	21	26	22	27	24	24	29	27	27
27	21	22	27	23	28	25	25	30	28	28
28	22	23	28	24	28	26	27	31	29	29
29	23	24	29	25	29	27	28	32	30	30
30	25	25	30	26	30	28	29	33	31	31
31	26	27	31	27	31	29	30	34	32	33
32	27	28	32	28	32	30	31	35	34	34
33	28	29	33	29	33	32	32	36	35	35
34	30	31	34	31	34	33	34	37	36	36
35	31	33	35	32	35	34	35	37	37	37
36	33	34	36	34	36	36	36	38	38	38
37	35	36	37	36	37	37	37	39	38	39
38	37	37	38	37	38	38	38	39	39	39
39	38	39	39	39	39	39	39	40	40	40
40	40	40	40	40	40	40	40	40	40	40

6.3.1.2. Wyniki uczniów szkół podstawowych 2002-2011 na skali zmiennej ukrytej

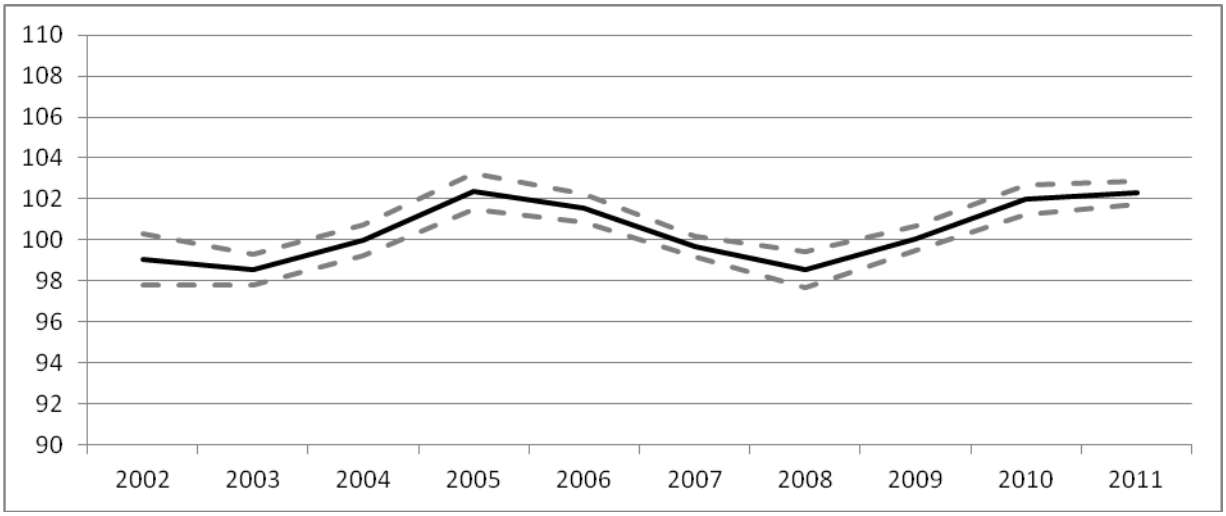
Zrównane wyniki sprawdzianu na skali zmiennej ukrytej określonej przez model IRT przekształcono w taki sposób, aby średnia w roku bazowym 2004 wynosiła 100 oraz odchylenie standardowe było równe 15. W Tabeli 6.8 (analogicznie jak dla egzaminu gimnazjalnego w Tabelach 6.4 oraz 6.5) zestawiono tak przeskalowane średnie dla lat 2002-2011 wraz z błędami wokół średniej.

Tabela 6.8. Średnie wyniki uczniów szkół podstawowych w latach 2002-2011, wyniki zrównane skala 100; 15 zakotwiczona w roku 2004.

Rok	średnia	błąd		
		Zrównywania	losowy	razem
2002	99,06	0,627	0,023	0,627
2003	98,56	0,391	0,022	0,392
2004	100,00	0,384	0,022	0,385
2005	102,36	0,449	0,025	0,449
2006	101,55	0,342	0,023	0,342
2007	99,69	0,247	0,024	0,248
2008	98,55	0,448	0,024	0,449
2009	100,08	0,314	0,025	0,315
2010	101,97	0,359	0,027	0,360
2011	102,29	0,288	0,028	0,290

Wykorzystując wielkości całkowitych błędów wokół średnich skonstruowano 95% przedziały ufności i naniesiono wraz ze średnimi na wykres ukazujący zmiany poziomu umiejętności mierzonych sprawdzianem na wspólnej skali na przestrzeni dekady 2002-2011. Wewnątrz granic wydzielonych przez przedział ufności zauważamy sinusoidalną fluktuację wyników, w cyklu około 6-7-letnim. Najwyższe wyniki zaobserwowano w latach 2005 oraz 2011, a najniższe w latach 2003 oraz 2008. Oczywiście brak jakichkolwiek przesłanek pozwalających na sformułowanie tezy, że zmienność zrównanych wyników w przyszłych latach utrzyma się w ramach opisanego wzoru. Ponieważ przedziały ufności wokół średnich dla wymienionych roczników są w odpowiednich parach rozłączne, można się zastanowić nad poszukiwaniem możliwych przyczyn zaobserwowania takich a nie innych zmian w zrównanych wynikach uczniów dla roczników piszących sprawdzian w latach 2002-2011.

Rysunek 6.9. Średnie wyniki uczniów szkół podstawowych w latach 2002-2011, wyniki zrównane skala 100; 15 zakotwiczona w roku 2004



7. Analizy wyników egzaminacyjnych na zrównanych wynikach

Zrównanie wyników egzaminacyjnych dokonane w ramach projektu pozwala na przeprowadzenie szeregu istotnych z perspektywy polskiej oświaty analiz, które w pełnym kształcie nie mogły zostać przeprowadzone wcześniej. Po raz pierwszy w historii analiz polskiego systemu egzaminacyjnego wyniki badań przekrojowych opisujące zmiany wyników kształcenia na szczeblu szkoły podstawowej przedstawione zostaną na wspólnej skali, co pozwoli na bezpośrednie i poprawne porównania oraz interpretacje zmian jakie zachodzą w poziomie osiągnięć uczniów szacowanym na podstawie wyników egzaminacyjnych. Rozdział ten przedstawia również kontynuację analiz wykorzystujących zrównane wyniki egzaminu gimnazjalnego, które szczegółowo prezentowane były w raporcie z roku 2011. Analizy wyników egzaminu gimnazjalnego przedstawione zostaną tutaj z rozszerzeniem o zrównane wyniki egzaminacyjne z roku 2011, które mogły zostać uzyskane dopiero po przeprowadzeniu dodatkowego badania zrównującego w roku 2012.

W rozdziale tym podjęto trzy najważniejsze tematy poruszane w badaniach edukacyjnych: zróżnicowanie wyników egzaminacyjnych ze względu na lokalizację szkoły, zróżnicowanie wyników ze względu na płeć oraz zróżnicowanie wyników ze względu na typ szkoły (prywatna a publiczna) – po raz pierwszy przy użyciu zrównanych danych egzaminacyjnych. W ostatnich podrozdziałach krótko omówiono różnice w umiejętnościach uczniów z uwzględnieniem podziału terytorialnego kraju.

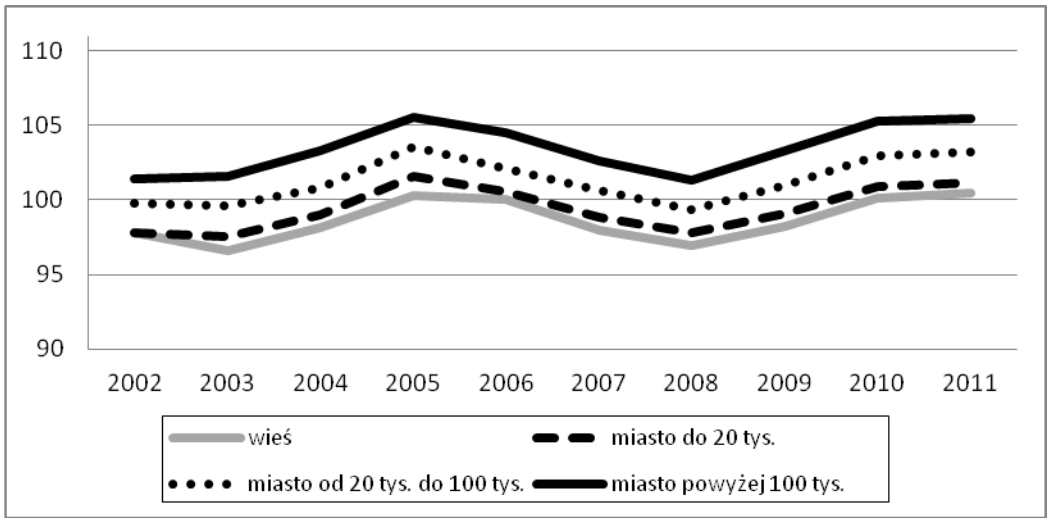
7.1. Sprawdzian po szkole podstawowej

Wszystkie wyniki w tej części rozdziału prezentowane są na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2004 w przypadku sprawdzianu po szkole podstawowej i 2003 dla egzaminów gimnazjalnych. Wyniki szacowano z wykorzystaniem 10 *plausible values* wygenerowanych dla każdego ucznia. Oznacza to, iż każda analiza wykonana została 10 razy, za każdym razem na innym oszacowaniu punktowym umiejętności ucznia (będącego estymacją wyniku prawdziwego z modelu IRT). Następnie wyniki tych analiz zostały uśrednione. Taka procedura jest szczególnie ważna dla oszacowania zróżnicowania wyników uczniów, a co za tym idzie również zróżnicowania międzyszkolnego (por. Wu 2005). Prezentowane tutaj analizy wyników w podziale ze względu na lokalizację szkoły po raz pierwszy prezentowane są za pomocą metodologii, która precyzyjnie szacuje zróżnicowanie uczniów (*plausible values*) oraz przedstawia wyniki na wspólnej skali (po zrównaniu wyników egzaminacyjnych).

7.1.1. Lokalizacja szkoły

Na Rysunku 7.1 przedstawiono średni poziom umiejętności uczniów na zrównanej skali wyników sprawdzianu, w podziale ze względu na wielkość miejscowości, w której zlokalizowana jest szkoła. Wyróżnione klasy wielkości miejscowości: wieś, miasto do 20 tysięcy mieszkańców, miasta od 20 tys. do 100 tys. mieszkańców oraz miasta powyżej 100 tys. mieszkańców.

Rysunek 7.1. Średni poziom umiejętności uczniów – sprawdzian, w podziale ze względu na lokalizację szkoły; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2004



Powyższy wykres pokazuje, że najwyższe wyniki sprawdzianu osiągają uczniowie z dużych miast, natomiast najniższe uczniowie ze szkół położonych na wsi. Różnica między wynikami szkół położonych na wsi i w mniejszych miastach (do 20 tys. mieszkańców) jest nieznaczna. Różnice między poszczególnymi kategoriami lokalizacji są stabilne i w analizowanym okresie nie ulegają znaczącym zmianom.

7.1.2. Zróżnicowanie międzyszkolne

Zróżnicowanie międzyszkolne oszacowane zostało za pomocą wielopoziomowego modelu „pustego", który można zapisać następująco (więcej o modelu pustym i modelowaniu poziomowym można znaleźć w Domański i Pokropek 2011):

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

gdzie:

- Y_{ij} jest wynikiem i-tego ucznia, z j-tej szkoły
- γ_{00} - jest średnią[25] średnich szkolnych wyników w populacji,
- u_{0j} - jest odchyleniem średniej Y dla j-tej szkoły od średniej w populacji, czyli „resztą" dla drugiego poziomu,

[25] W skrócie będziemy posługiwali się terminem średnia mając na myśli średnią ze średnich w modelu dwupoziomowym.

- r_{ij} - jest odchyleniem Y dla i-tego ucznia z j-tej szkoły, od średniej w tej kategorii, czyli „resztą” dla pierwszego poziomu.

Jest to model „pusty”, nie zawierający żadnych zmiennych niezależnych, które mogłyby być związane z wynikami uczniów. Trzema podstawowymi parametrami, których dostarcza ten model są:

- wartość oczekiwana dla zmiennej wyjaśnianej w populacji szkół (γ_{00}),
- wariancja wewnątrzszkolna ($Var(r_{ij}) = \sigma^2$),
- wariancja międzyszkolna ($Var(u_{0j}) = \tau_{00}$).

Podobnie jak cała klasa modeli z efektami losowymi, model ten opiera się na kilku założeniach, od których zależy efektywność estymacji parametrów i statystyk testowych. Założenia odnoszą się do wielkości błędów:

$$r_{ij} \sim N(0, \sigma^2) \qquad u_j \sim N(0, \tau_{00})$$

Założenie o niezależności u_j i r_{ij} jest równoznaczne ze stwierdzeniem o braku korelacji między stałą – średnią wyników uczniów w szkołach – a resztami dla pierwszego poziomu. Kolejne założenie – o braku korelacji między poszczególnymi u – oznacza niezależność między obserwacjami z różnych kategorii drugiego poziomu (szkoły), a założenie o braku korelacji między *r* – niezależność między obserwacjami w ramach pierwszego poziomu (po wyłączeniu wpływu grupowego). Spełnienie tych założeń umożliwia poprawne szacowanie parametrów. Należy równocześnie pamiętać, że nie eliminuje to – fundamentalnego dla danych hierarchicznych – faktu występowania korelacji między resztami w ramach kategorii drugiego poziomu.

Całkowitą wariancję dla modelu wielopoziomowego (u_{ij}), równoznaczną wariancji y_{ij}, można zapisać w postaci równania:

$$Var(y_{ij}) = Var(\gamma_{00} + u_j + r_{ij}) = Var(u_j + r_{ij}) = \tau_{00} + \sigma^2$$

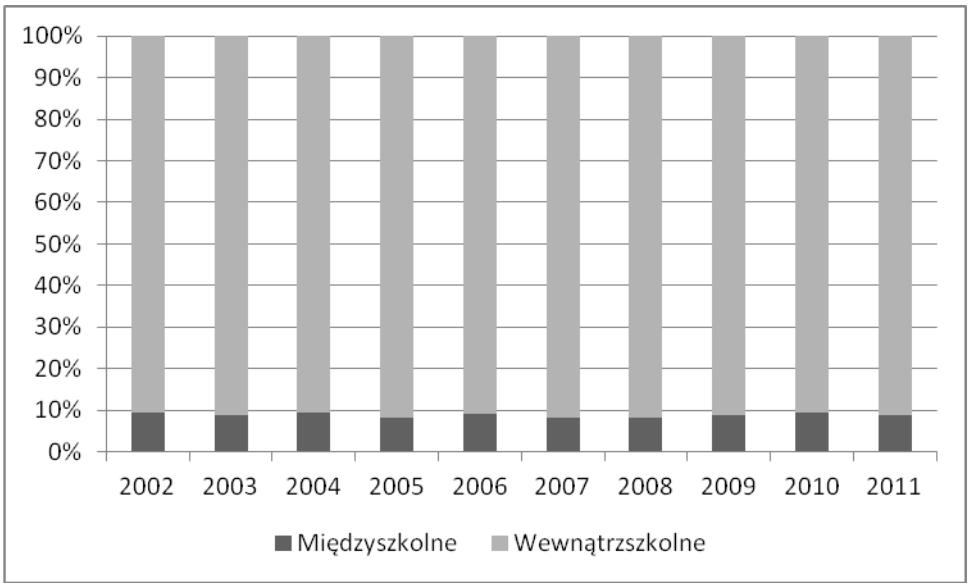
Z założenia o niezależności u_j i r_{ij} wynika, że całkowita wariancja Y_{ij} jest sumą wariancji między- i wewnątrzgrupowej. Obliczając stosunek wariancji międzygrupowej do wariancji całkowitej uzyskujemy współczynnik korelacji wewnątrzgrupowej (ang. *inter-class correlation*), który jest miernikiem homogeniczności kategorii drugiego poziomu (w naszym przypadku szkół).

$$\rho = \frac{Var(u_i)}{Var(y_{ij})} = \frac{\tau_{00}}{\tau_{00} + \sigma^2}$$

Miernik ten (analogiczny do stosunku korelacyjnego) nazywany jest współczynnikiem korelacji wewnątrzgrupowej również z tego powodu, że jego wartość równa jest średniej korelacji między wynikami losowo dobranych uczniów w poszczególnych szkołach. Do naszych analiz stosujemy ten miernik, gdyż daje on nieobciążone estymatory zróżnicowania międzyszkolnego również w przypadku danych o niezbalansowanej strukturze, tj. takich w których liczba jednostek z pierwszego poziomu (uczniów) różni się między grupami (szkołami) (por. Snijders 2002).

W kolejnej części omówione zostanie międzyszkolne zróżnicowanie wyników sprawdzianu z uwzględnieniem lokalizacji szkoły. Wcześniej jednak zobaczmy, jak wyglądało międzyszkolne zróżnicowanie wyników sprawdzianu na poziomie ogólnopolskim.

Rysunek 7.2. Międzyszkolne i wewnątrzszkolne zróżnicowanie wyników sprawdzianu w latach 2002 – 2011, analiza na zrównanych wynikach sprawdzianu

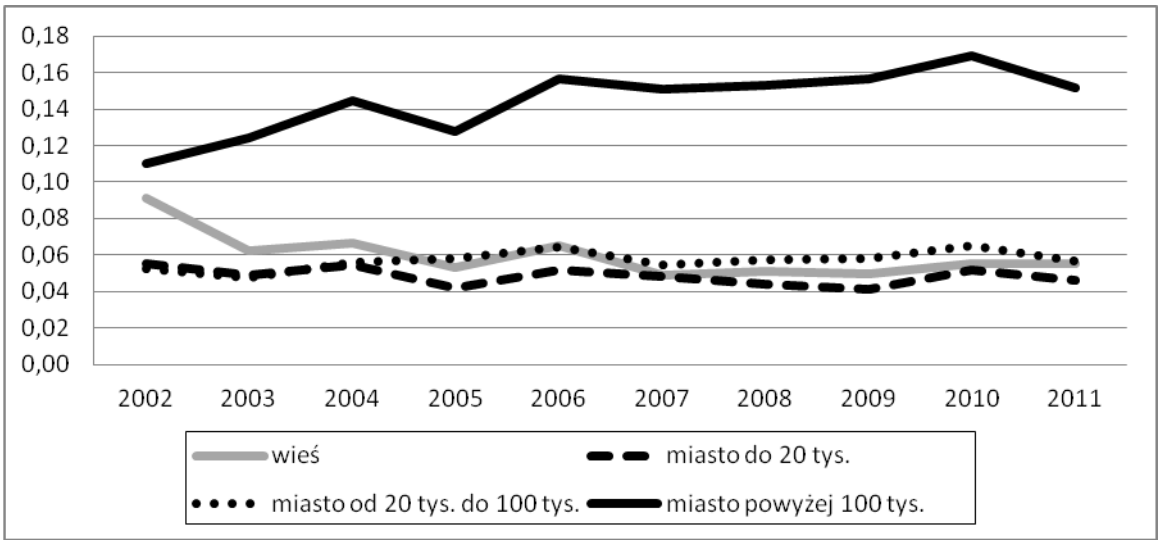


Na Rysunku 7.2 przedstawione zostało międzyszkolne zróżnicowanie wyników sprawdzianu na przestrzeni lat 2002-2011. W analizowanym okresie wskaźnik zróżnicowania międzyszkolnego pozostaje na stałym poziomie i wynosi około 10%. Oznacza to, iż 1/10 całkowitego zróżnicowania wyników sprawdzianu wyjaśnia zróżnicowanie wyników między szkołami. Zróżnicowanie to jest niezwykle stabilne.

7.1.3. Zróżnicowanie międzyszkolne i lokalizacja szkoły

Rysunek 7.3 pokazuje zróżnicowanie międzyszkolne wyników sprawdzianu w zależności od wielkości miejscowości, w której zlokalizowana jest szkoła. Dane przedstawione na rysunku pozwalają ocenić, jaki jest stopień zróżnicowania szkół w zależności od stopnia urbanizacji oraz prześledzić zmiany tego zróżnicowania w ramach poszczególnych kategorii na przestrzeni lat 2002-2011. Międzyszkolne zróżnicowanie wyników w szkołach położonych na wsi, w miastach do 20 tysięcy oraz w miastach od 20 tys. do 100 tysięcy na przestrzeni lat 2002-2011 jest stabilne i wynosi od 0,04-0,06. Większym zróżnicowaniem charakteryzują się natomiast szkoły zlokalizowane w miastach liczących powyżej 100 tysięcy mieszkańców. Co więcej, na przestrzeni lat 2002-2011 obserwujemy nieznaczny, lecz systematyczny wzrost zróżnicowania wyników sprawdzianu w szkołach położonych w miastach powyżej 100 tysięcy mieszkańców. Wyniki te są interesujące w kontekście zróżnicowania międzyszkolnego wyników sprawdzianu na poziomie ogólnopolskim: na ile dla całego kraju zróżnicowanie międzyszkolne w latach 2002-2011 nie wykazuje większych zmian, w przypadku największych miast mamy do czynienia z trendem wzrostu zróżnicowania.

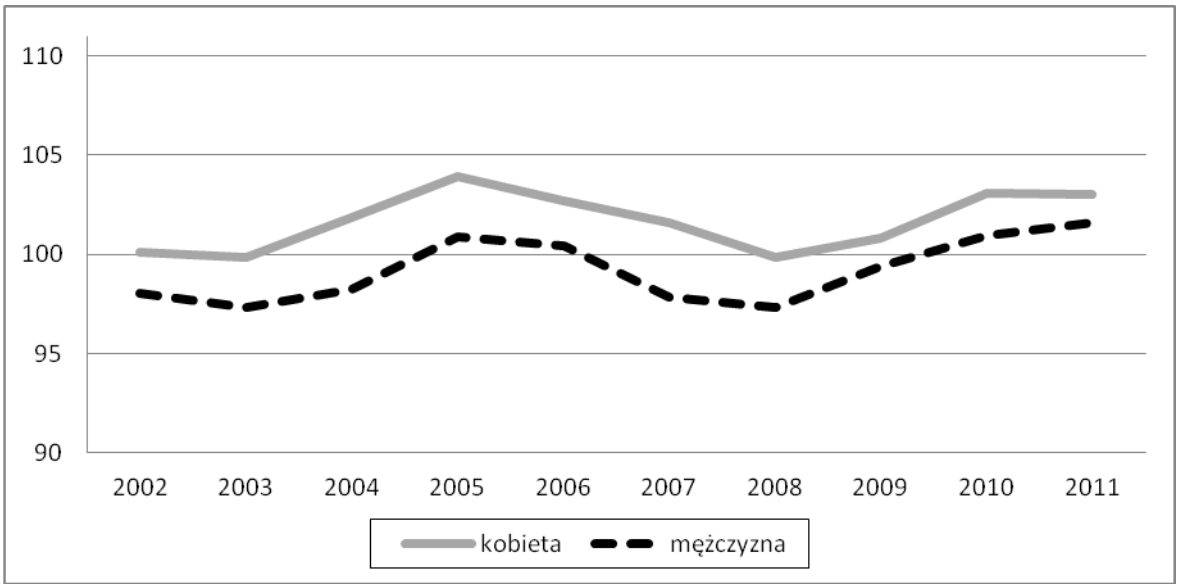
Rysunek 7.3. Zróżnicowanie międzyszkolne (wyrażone wskaźnikiem korelacji wewnątrzszkolnej) wyników sprawdzianu w podziale na lokalizację szkoły; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2004



7.1.4. Płeć uczniów

Na Rysunku 7.4 przedstawione zostało zróżnicowanie wyników sprawdzianu ze względu na płeć. Dziewczęta osiągają średnio nieznacznie wyższe wyniki aniżeli chłopcy.

Rysunek 7.4. Średni poziom umiejętności uczniów – sprawdzian, w podziale ze względu na płeć; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2004

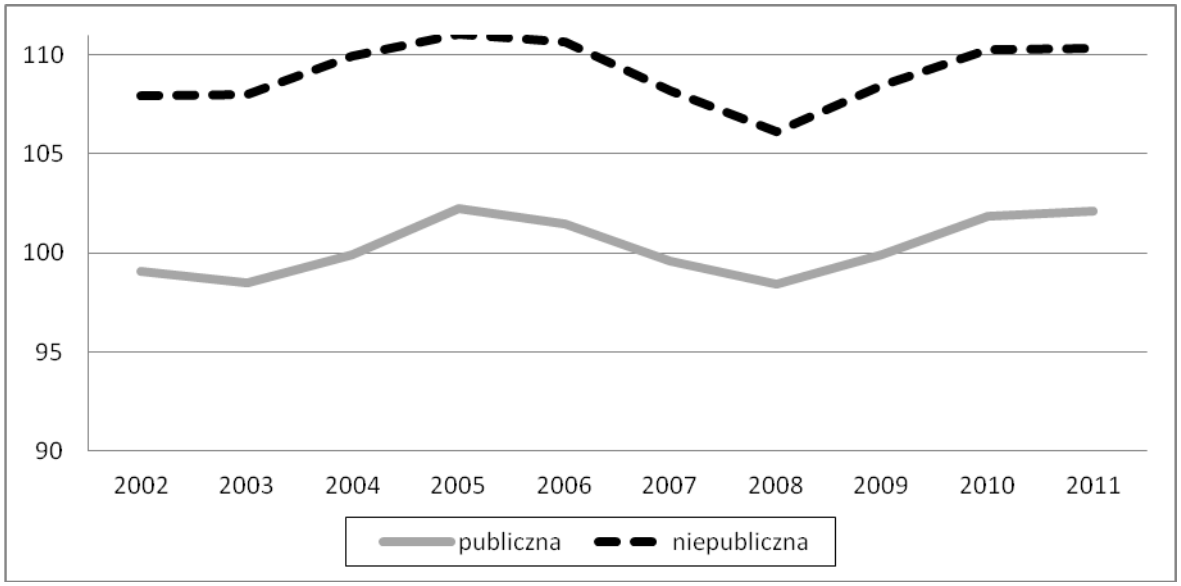


7.1.5. Szkoły publiczne i niepubliczne

W tej części rozdziału przedstawione zostaną analizy na zrównanych wynikach sprawdzianu z uwzględnieniem typu szkoły. Celem tej części raportu jest odpowiedź na pytanie, jak przedstawiały się zrównane wyniki sprawdzianu w okresie 2002-2011 dla szkół publicznych i niepublicznych.

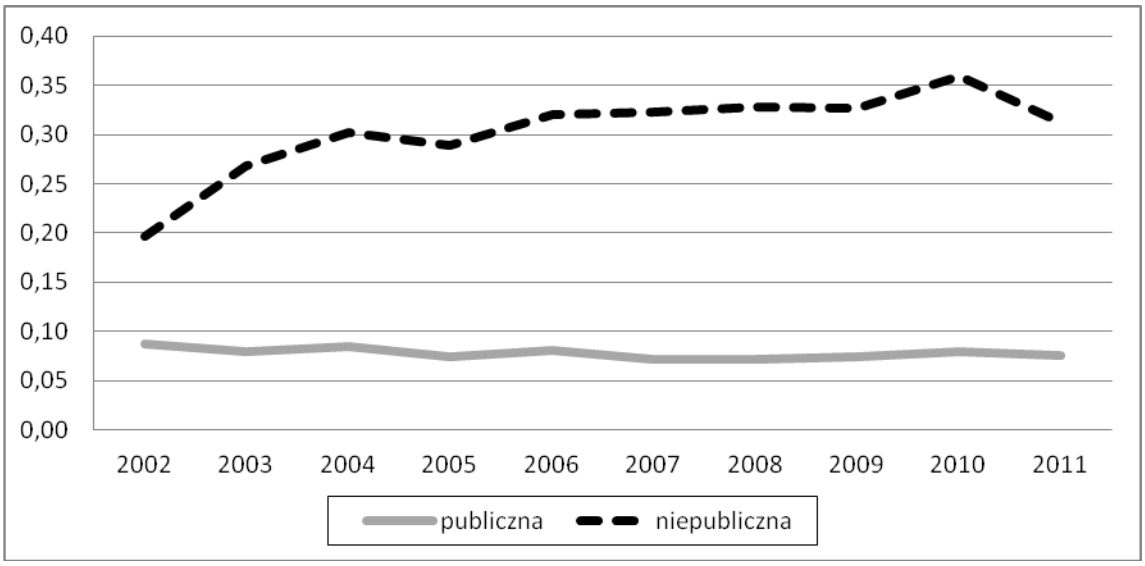
Rysunek 7.5 przedstawia średnie wyniki, jakie na przestrzeni lat 2002-2011 osiągali uczniowie szkół publicznych i niepublicznych. Uczniowie szkół niepublicznych osiągają wyniki wyższe o około 2/3 odchylenia standardowego i różnica ta utrzymuje się w całym analizowanym okresie.

Rysunek 7.5. Średni poziom umiejętności uczniów – sprawdzian, w podziale na szkoły publiczne i niepubliczne; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2004



Rysunek 7.6 przedstawia współczynnik zróżnicowania międzyszkolnego dla szkół publicznych i niepublicznych. Zróżnicowanie międzyszkolne wyników szkół publicznych wynosi około 0,8-07 i utrzymuje się na stałym poziomie. W przypadku szkół niepublicznych w latach 2002-2011 obserwujemy natomiast systematyczny wzrost międzyszkolnego zróżnicowania wyników.

Rysunek 7.6. Zróżnicowanie międzyszkolne (wyrażone wskaźnikiem korelacji wewnątrzszkolnej) wyników sprawdzianu w podziale na szkoły publiczne i niepubliczne; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2004

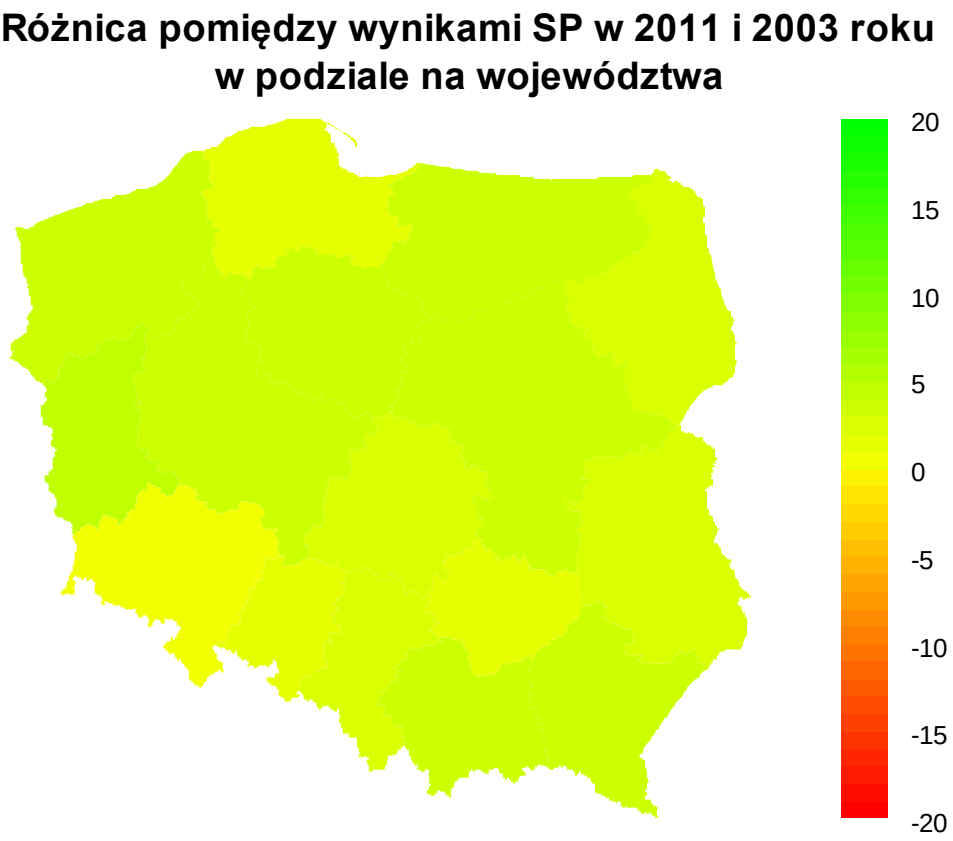


7.1.6. Zróżnicowanie terytorialne

Zrównanie wyników i sprowadzenie ich do wspólnej skali pozwala na obliczenie różnic w poziomie umiejętności uczniów w poszczególnych latach. Dla celów niniejszej analizy wybrano skrajne lata zrównanych wyników, z pominięciem roku 2002. Rok 2002 pominięto z racji tego, że była to pierwsza edycja sprawdzianu.

W niniejszym podrozdziale prezentowane są mapy Polski w podziale na województwa i powiaty. Kolory poszczególnych jednostek podziału terytorialnego odpowiadają różnicom w średnich wynikach pomiędzy rokiem 2011 i 2003. Wyniki zostały przedstawione na zrównanej skali o średniej 100 i odchyleniu standardowym 15 (zatem różnica rzędu 15 punktów oznacza różnicę jednego odchylenia standardowego). Paleta kolorów została ustalona tak, aby brak różnic (0) oznaczony był kolorem żółtym, różnice ujemne (czyli spadek wyników) rozciągał się do koloru czerwonego, a różnice dodatnie (czyli wzrost wyników) – do koloru zielonego.

Rysunek 7.7. Różnica średniego poziomu umiejętności ze sprawdzianu po szóstej klasie szkoły podstawowej pomiędzy rokiem 2011 i 2003 w podziale na województwa na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2004



Na powyższej mapie (Rysunek 7.7) dominują odcienie żółte i zielone, co oznacza, iż pomiędzy rokiem 2003 i 2011 nie zanotowano spadku średniego wyniku żadnego z województw. Najmniejsza różnica wynosi niecałe 2 punkty (dla województwa dolnośląskiego), a największa różnica wynosi nieco ponad 5 punktów (dla województwa lubuskiego). Oznacza to, że po ośmiu latach od pierwszej rozpatrywanej edycji sprawdzianu umiejętności uczniów kończących szkołę podstawową w każdym z województw nieznacznie wzrosły. Warto pamiętać o tym, że zrównane wyniki sprawdzianu po szóstej klasie szkoły podstawowej w skali całego kraju podlegały okresowym wahaniom – od 2003 do 2005 roku następował wzrost wyników, od 2005 do 2008 spadek, a od 2008 do 2011 znów wzrost. Żadne z województw nie odbiega znacząco od tego ogólnopolskiego trendu.

Średnie wyniki poszczególnych województw nie różnią się zbytnio od siebie – różnica pomiędzy średnią najlepszego i najgorszego województwa w każdym roku waha się w granicach od nieco ponad 2 do 4 punktów (wyjątek stanowi tutaj rok 2002, gdzie różnica ta wynosi nieco powyżej 5 punktów). Jak już wspomniano wyżej, jest to pierwszy rok przeprowadzania sprawdzianu w szóstej klasie szkoły podstawowej jako ogólnopolskiego egzaminu zewnętrznego – do jego wyników należy podchodzić z rezerwą. Wyniki poszczególnych województw w kolejnych latach są do siebie bardzo zbliżone – nie można wyróżnić województwa bardzo różniącego się średnimi wynikami od innych (zarówno *in plus*, jak i *in minus*). Taka obserwacja powinna cieszyć, gdyż oznacza, że na poziomie dużych agregatów,

którymi są województwa, nie ma dużych różnic w umiejętnościach uczniów. Różnice regionalne na tak wysokim poziomie praktycznie nie istnieją.

Wyniki sprawdzianu w poszczególnych latach podlegają okresowej fluktuacji i dla większości województw rosną one lub maleją według ogólnego trendu. Wyjątkiem jest województwo kujawsko-pomorskie, które w żadnym roku od początku istnienia sprawdzianu w szóstej klasie szkoły podstawowej nie miało średnich wyników statystycznie istotnie niższych niż w roku poprzedzającym. Nawet w roku 2007, gdzie pozostałe województwa miały statystycznie istotny spadek średnich umiejętności uczniów mierzonych na sprawdzianie, w stosunku do roku 2006 spadek wyników w województwie kujawsko-pomorskim był nieistotny statystycznie.

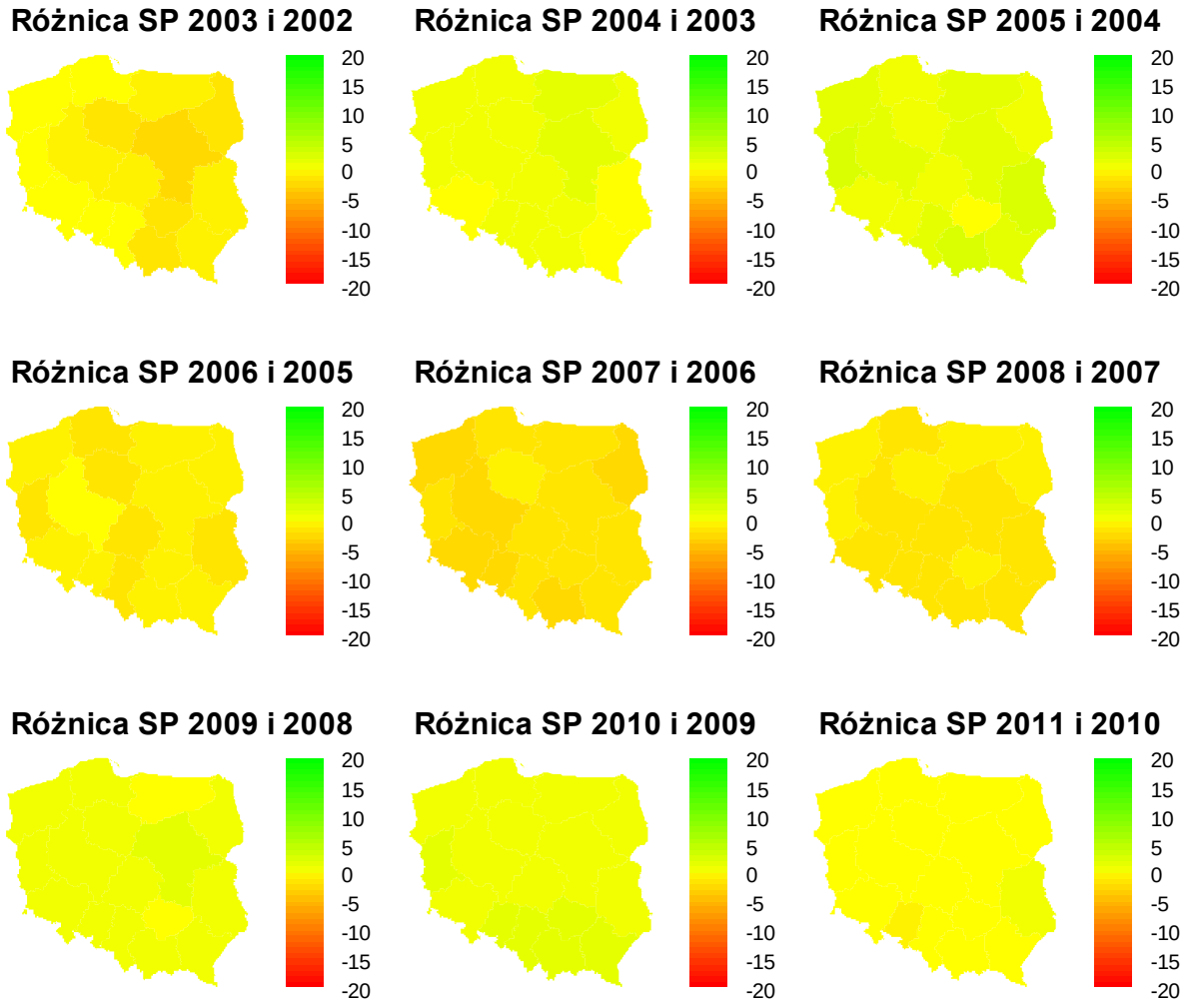
Analiza średnich zrównanych wyników w podziale na województwa nie prowadzi do wniosków znacząco różniących się od analizy surowych średnich. Najwyższe średnie wyniki mają województwa: małopolskie i mazowieckie, lecz po pierwsze, różnice w wynikach tych województw we wszystkich latach są nieistotne statystycznie, a po drugie, jedynie w 2007 roku średnie wyniki tych województw są statystycznie istotnie wyższe od średnich wyników pozostałych województw. Przedziały ufności dla średnich wyników dla wszystkich edycji egzaminu w podziale na województwa przedstawiono w Tabeli 7.1.

Tabela 7.1. Przedziały ufności dla średniego poziomu umiejętności ze sprawdzianu w szóstej klasie szkoły podstawowej dla poszczególnych edycji w podziale na województwa na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2004

Województwo\Lata	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
dolnośląskie	<98,31; 100,79>	<99,22; 100,79>	<100,05; 101,59>	<101,69; 103,49>	<101,23; 102,62>	<99,34; 100,37>	<97,14; 98,93>	<99,07; 100,35>	<100,87; 102,33>	<101,07; 102,28>
kujawsko-pomorskie	<97,18; 99,67>	<96,58; 98,15>	<98,34; 99,90>	<99,87; 101,68>	<98,95; 100,34>	<98,20; 99,25>	<96,84; 98,64>	<98,68; 99,97>	<100,40; 101,87>	<100,87; 102,09>
Lubelskie	<97,88; 100,36>	<98,28; 99,85>	<98,40; 99,94>	<101,97; 103,77>	<100,55; 101,95>	<99,31; 100,35>	<97,54; 99,34>	<99,05; 100,33>	<100,77; 102,24>	<101,97; 103,18>
lubuskie	<94,71; 97,22>	<95,47; 97,09>	<97,09; 98,69>	<100,30; 102,15>	<99,32; 100,76>	<97,53; 98,65>	<97,05; 98,91>	<98,49; 99,84>	<100,61; 102,14>	<100,90; 102,20>
łódzkie	<98,68; 101,16>	<98,47; 100,04>	<99,92; 101,47>	<101,54; 103,33>	<100,72; 102,10>	<99,29; 100,33>	<97,75; 99,54>	<99,57; 100,85>	<101,00; 102,46>	<101,91; 103,11>
małopolskie	<99,64; 102,11>	<98,92; 100,48>	<100,23; 101,77>	<103,20; 104,99>	<102,48; 103,86>	<100,66; 101,67>	<99,24; 101,02>	<100,60; 101,87>	<103,01; 104,45>	<103,27; 104,45>
mazowieckie	<100,04; 102,51>	<98,43; 99,98>	<100,71; 102,24>	<103,01; 104,79>	<102,53; 103,89>	<100,71; 101,72>	<98,76; 100,54>	<101,43; 102,70>	<102,90; 104,33>	<103,10; 104,27>
opolskie	<97,77; 100,28>	<98,59; 100,21>	<100,12; 101,72>	<101,77; 103,61>	<101,29; 102,74>	<99,41; 100,54>	<97,44; 99,29>	<99,05; 100,41>	<101,21; 102,75>	<101,12; 102,45>
podkarpackie	<98,00; 100,48>	<97,53; 99,10>	<98,53; 100,07>	<101,36; 103,16>	<101,22; 102,62>	<99,57; 100,61>	<98,16; 99,96>	<99,92; 101,21>	<102,11; 103,57>	<102,54; 103,74>
podlaskie	<99,38; 101,88>	<98,60; 100,19>	<100,09; 101,67>	<101,51; 103,33>	<101,57; 103,00>	<99,50; 100,61>	<98,17; 100,00>	<99,91; 101,25>	<101,53; 103,05>	<102,29; 103,58>
pomorskie	<97,01; 99,50>	<97,79; 99,36>	<99,00; 100,55>	<100,87; 102,67>	<99,99; 101,38>	<98,61; 99,65>	<97,01; 98,81>	<98,86; 100,16>	<100,57; 102,03>	<100,88; 102,09>
śląskie	<97,20; 99,67>	<97,94; 99,49>	<99,18; 100,71>	<101,29; 103,07>	<100,12; 101,49>	<98,54; 99,55>	<97,02; 98,80>	<98,88; 100,14>	<100,88; 102,31>	<101,15; 102,32>
świętokrzyskie	<98,44; 100,94>	<97,59; 99,19>	<99,37; 100,94>	<100,14; 101,98>	<99,35; 100,77>	<97,91; 99,00>	<97,26; 99,09>	<98,41; 99,75>	<100,35; 101,86>	<100,72; 101,99>
warmińsko-mazurskie	<95,85; 98,34>	<95,95; 97,54>	<98,08; 99,65>	<100,76; 102,58>	<99,99; 101,41>	<98,50; 99,59>	<97,44; 99,26>	<98,40; 99,72>	<100,31; 101,80>	<100,83; 102,08>
wielkopolskie	<96,09; 98,57>	<96,47; 98,02>	<98,17; 99,70>	<100,45; 102,23>	<100,69; 102,06>	<98,38; 99,40>	<96,93; 98,71>	<98,30; 99,57>	<100,14; 101,58>	<100,77; 101,95>
zachodniopomorskie	<94,86; 97,35>	<96,13; 97,71>	<97,64; 99,20>	<99,98; 101,80>	<100,09; 101,51>	<97,93; 98,99>	<96,85; 98,66>	<98,89; 100,22>	<100,00; 101,47>	<100,31; 101,56>

W początkowych latach przeprowadzania sprawdzianu w szóstej klasie szkoły podstawowej poszczególne województwa charakteryzowały się odmiennymi różnicami pomiędzy średnimi wynikami z sąsiadujących edycji sprawdzianu – w niektórych spadki/wzrosty były większe niż w pozostałych. Z czasem wahania poziomu umiejętności uległy stabilizacji i od 2008 roku różnice te są podobne we wszystkich województwach i wpisują się w trend dla całego kraju. Zjawisko to można obserwować na Rysunku 7.8, na którym zestawiono ze sobą mapy obrazujące różnice pomiędzy sąsiadującymi edycjami sprawdzianu ze wszystkich lat w podziale na województwa. Dwa górne rzędy map charakteryzują się większym zróżnicowaniem kolorów pomiędzy województwami (co oznacza różny poziom spadku/wzrostu umiejętności w poszczególnych województwach). Kolorystyka poszczególnych map w dolnym wierszu jest bardziej jednolita, co wskazuje na zbliżone różnice w wynikach dla poszczególnych województw.

Rysunek 7.8. Różnice średniego poziomu umiejętności ze sprawdzianu po szóstej klasie szkoły podstawowej pomiędzy sąsiadującymi edycjami w podziale na województwa na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2004

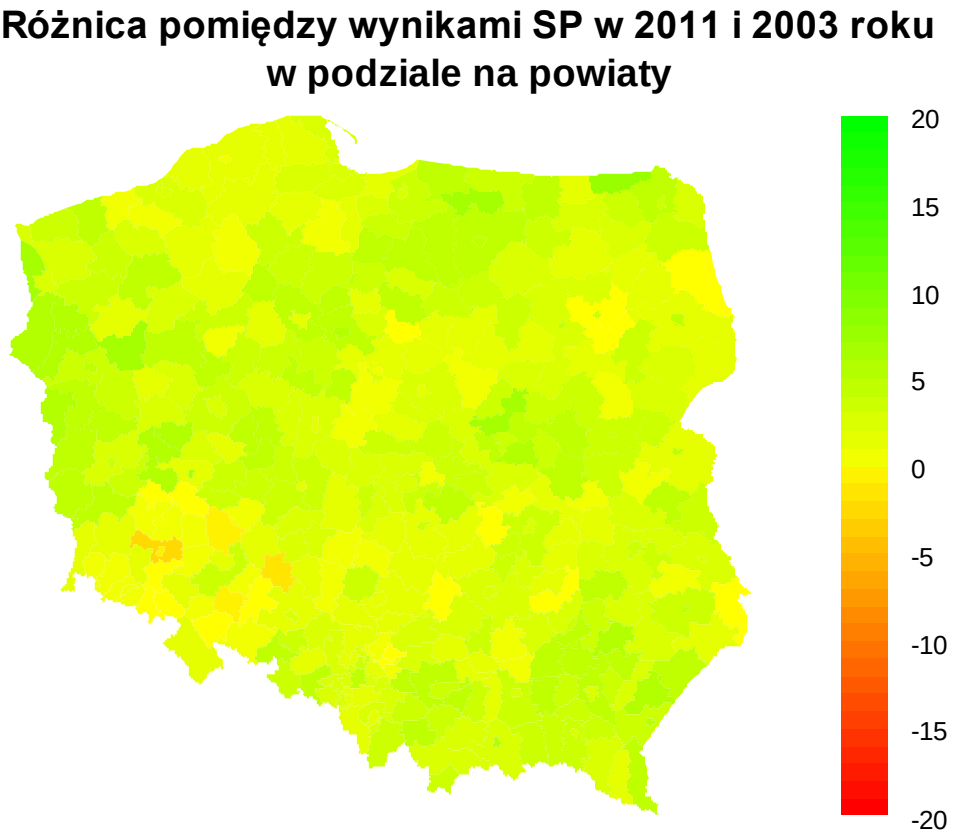


Od 2008 roku nie zaobserwowono spadku umiejętności mierzonych na sprawdzianie w szóstej klasie szkoły podstawowej w żadnym z województw. W roku 2009 siedem województw miało wyniki statystycznie istotnie wyższe niż w roku 2008, a w roku 2010 jedynie województwo

zachodniopomorskie nie miało wyników statystycznie istotnie wyższych niż rok wcześniej. Pomiędzy latami 2011 i 2010 różnice w średnich wynikach nie są istotne statystycznie w żadnym z województw.

Na Rysunku 7.9 przedstawione zostały różnice średnich poziomów umiejętności szóstoklasistów pomiędzy latami 2011 i 2003 w podziale na powiaty. Ogólne wnioski płynące z analizy w podziale na powiaty nie różnią się od tych w podziale na województwa. W zdecydowanej większości (trzy czwarte powiatów) mamy do czynienia ze wzrostem średniego poziomu umiejętności. W żadnym z powiatów nie nastąpił statystycznie istotny spadek umiejętności pomiędzy omawianymi edycjami sprawdzianu w szóstej klasie szkoły podstawowej. Procentowo najmniej powiatów, gdzie nastąpił wzrost wyników pomiędzy 2003 i 2011 rokiem znajduje się w województwie dolnośląskim – tylko jedna szósta powiatów z tego województwa w 2011 roku miała statystycznie istotnie wyższe wyniki niż w roku 2003. W pozostałych województwach była to minimum połowa powiatów, a w województwie lubuskim wszystkie.

Rysunek 7.9. Różnica średniego poziomu umiejętności ze sprawdzianu po szóstej klasie szkoły podstawowej pomiędzy rokiem 2011 i 2003 w podziale na powiaty na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2004



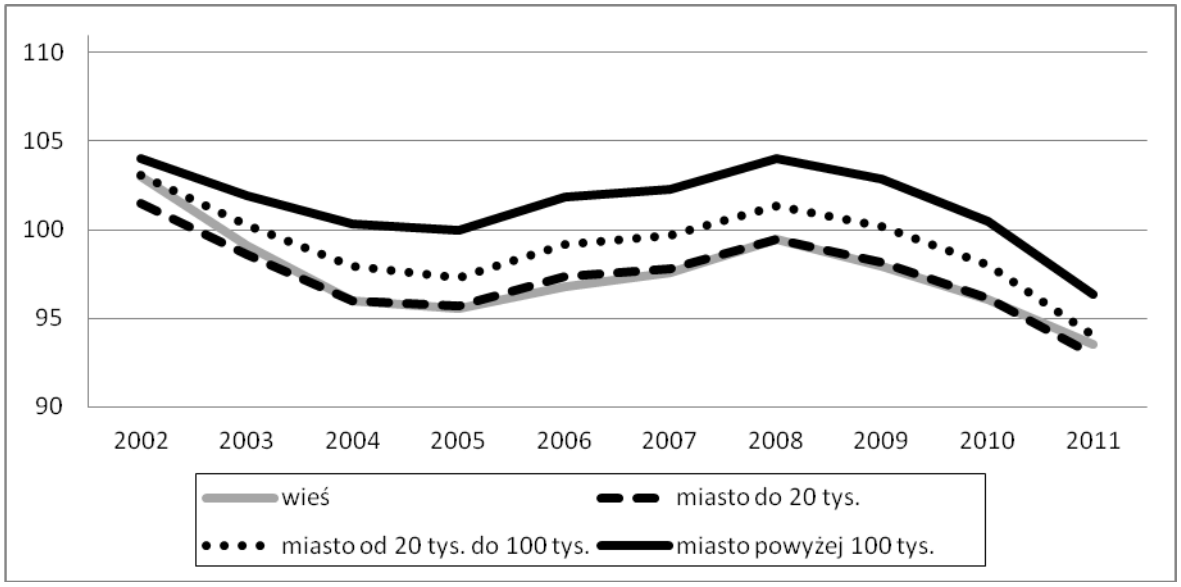
Zróżnicowanie średnich rezultatów ze sprawdzianu w poszczególnych latach przedstawiają mapy zawarte w aneksie. W aneksie B1 przedstawiono podział na województwa, w aneksie B2 podział na powiaty, a w aneksie B3 podział na powiaty w ramach województw (na jednej mapie zawarto obszar jednego województwa).

7.2. Egzamin gimnazjalny

7.2.1. Lokalizacja szkoły

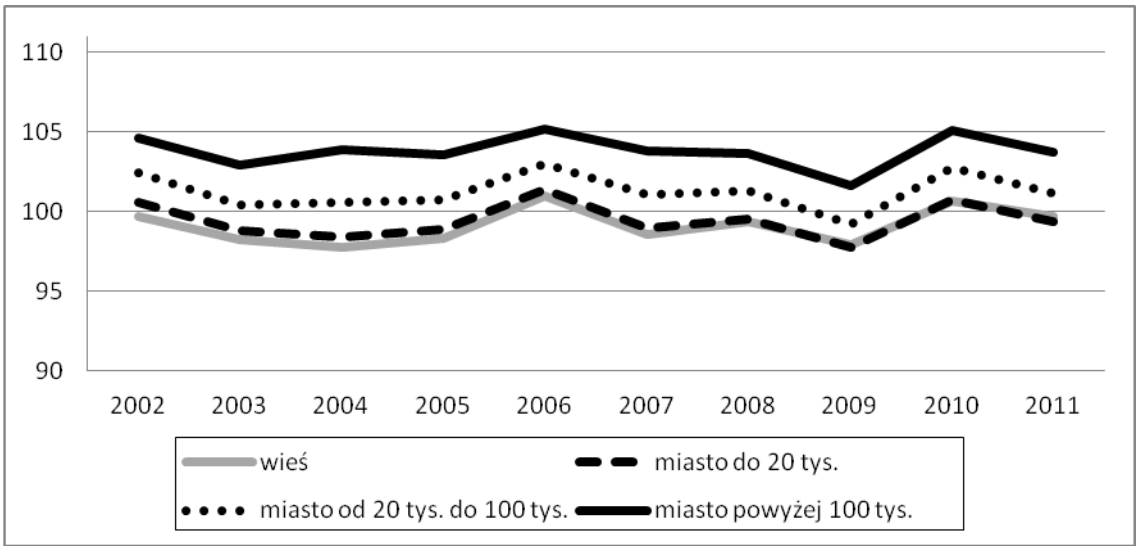
Na Rysunkach 7.10 i 7.11 przedstawione zostało zróżnicowanie umiejętności uczniów ze względu na lokalizację szkoły dla części humanistycznej oraz matematyczno-przyrodniczej egzaminu gimnazjalnego. W analizach uwzględniono podział na cztery klasy wielkości miejscowości, w której zlokalizowana jest szkoła: wieś, miasta do 20 tysięcy mieszkańców, miasta od 20 do 100 tysięcy mieszkańców oraz miasta liczące powyżej 100 tysięcy mieszkańców.

Rysunek 7.10. Średni poziom umiejętności uczniów – egzamin gimnazjalny część matematyczno-przyrodnicza, w podziale ze względu na lokalizację szkoły; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2003



Powyższy rysunek pokazuje, jak różnią się wyniki egzaminacyjne gimnazjalistów z zakresu przedmiotów matematyczno-przyrodniczych, w zależności od wielkości miejscowości, w której zlokalizowana jest szkoła. Dane na wykresie potwierdzają, iż średnio najniższe wyniki osiągają uczniowie uczęszczający do szkół zlokalizowanych na wsiach oraz w mniejszych miastach (do 20 tys. mieszkańców), a najwyższe średnio wyniki uczniowie ze szkół zlokalizowanych w miastach liczących powyżej 100 tys. mieszkańców. Warto zaznaczyć również, iż nie ma różnicy w osiągnięciach uczniów uczęszczających do szkół położonych na wsi oraz w mniejszych miastach. Niemal przez cały badany okres utrzymuje się systematyczna różnica między wynikami uczniów ze szkół położonych na wsi oraz w mniejszych miastach, a wynikami uczniów ze średnich miast i dużych miast. Wyjątkiem jest rok 2002, kiedy wyniki uczniów pochodzących ze szkół położonych na wsi zbliżają się do wyników uczniów ze średnich miast oraz rok 2011, w którym wyniki uczniów ze szkół wiejskich oraz ze szkół zlokalizowanych w małych i średnich miastach są zbliżone.

Rysunek 7.11. Średni poziom umiejętności uczniów – egzamin gimnazjalny część humanistyczna, w podziale ze względu na lokalizację szkoły; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2003

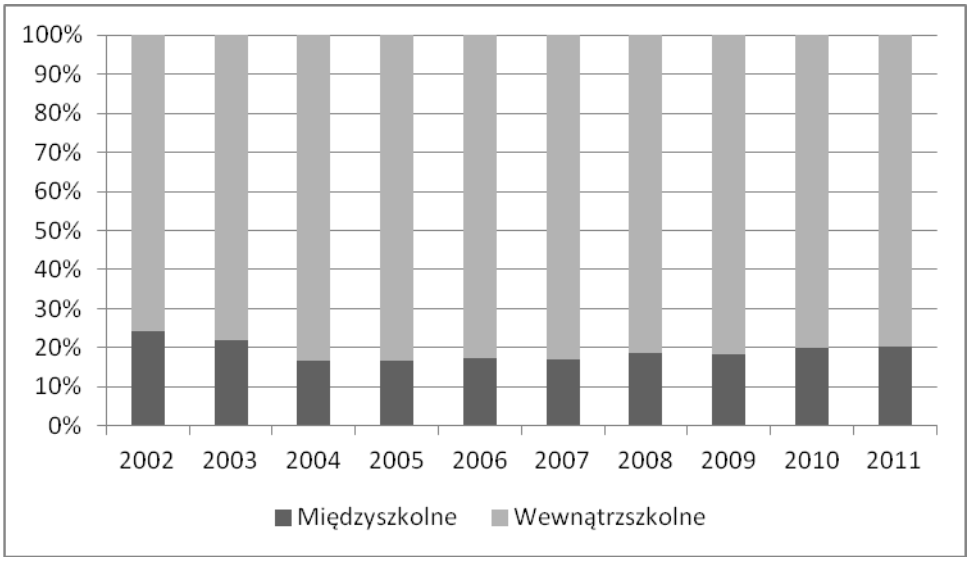


7.2.2. Zróżnicowanie międzyszkolne

Na Rysunkach 7.12 i 7.13 możemy zobaczyć, jak na przestrzeni lat 2002-2011 zmieniało się zróżnicowanie międzyszkolne zrównanych wyników z zakresu przedmiotów matematyczno-przyrodniczych oraz humanistycznych.

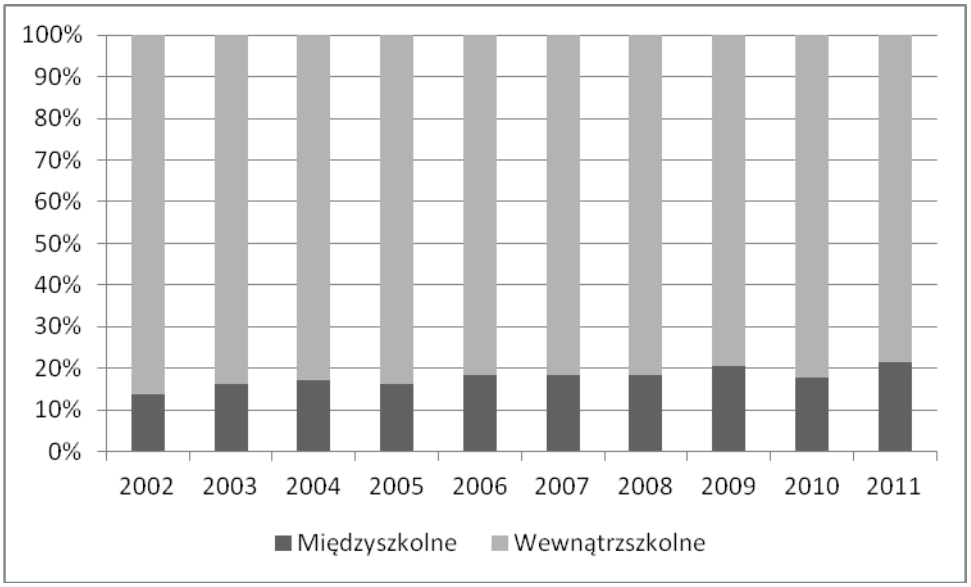
W przypadku wyników z zakresu przedmiotów matematyczno-przyrodniczych, dla roku 2002 i 2003 współczynnik zróżnicowania międzyszkolnego wynosi nieco ponad 20% ogólnej wariancji wyniku, w kolejnych latach ulega niewielkim fluktuacjom spadając niewiele poniżej 20%, a w roku 2010 i 2011 znów wraca do około 20%.

Rysunek 7.12. Zróżnicowanie międzyszkolne i wewnątrzszkolne – egzamin gimnazjalny cześć matematyczno-przyrodnicza



W przypadku wyników z zakresu przedmiotów humanistycznych, na przestrzeni lat 2002-2011 widoczny jest niewielki wzrost zróżnicowania międzyszkolnego. W roku 2002, gdy pierwsi gimnazjaliści zdawali nowy egzamin, zróżnicowanie to wynosiło ok. 14% wariancji wyniku ogółem, w kolejnych latach zróżnicowanie to nieznacznie rosło i w roku 2011 wynosiło już 22%.

Rysunek 7.13. Zróżnicowanie międzyszkolne i wewnątrzszkolne – egzamin gimnazjalny część humanistyczna

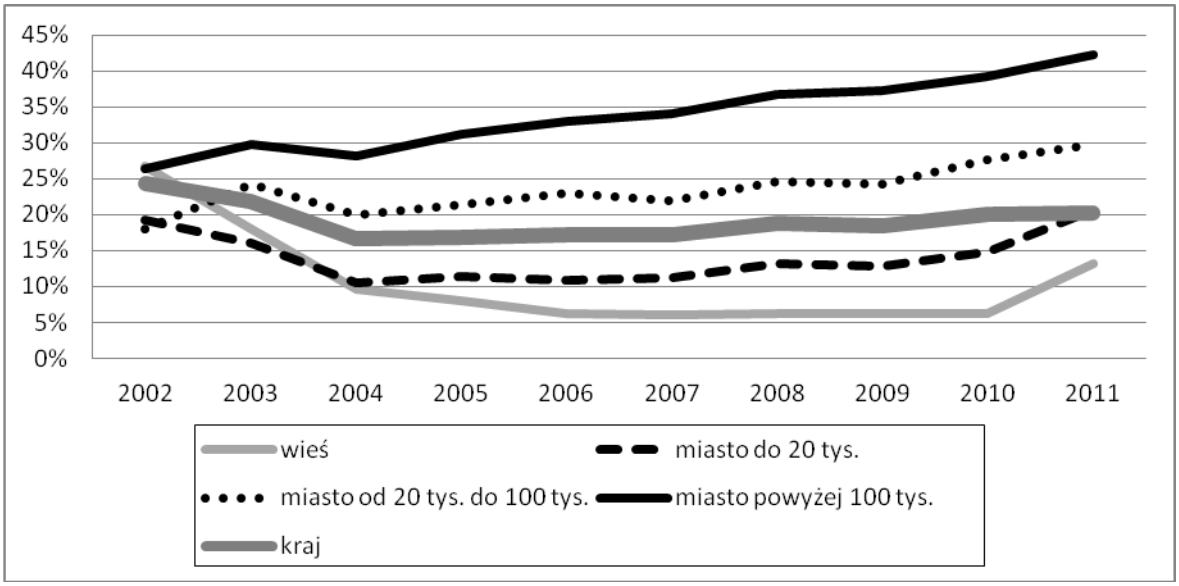


7.2.3. Zróżnicowanie międzyszkolne i lokalizacja szkoły

Proces różnicowania się szkół jest zjawiskiem złożonym, zależnym od wielu czynników. Jednym z mechanizmów, które mogą wpływać na proces różnicowania się szkół gimnazjalnych jest segregacja lokalizacyjna.

Rysunki 7.14 i 7.15 pokazują dynamikę procesu różnicowania się szkół z uwzględnieniem lokalizacji szkoły dla obydwu części egzaminu. Na wykresach został przedstawiony również trend ogółem dla całego kraju.

Rysunek 7.14. Zróżnicowanie międzyszkolne (wyrażone wskaźnikiem korelacji wewnątrzszkolnej) wyników egzaminu gimnazjalnego (cześć matematyczno-przyrodnicza) w podziale na lokalizację szkoły; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2003

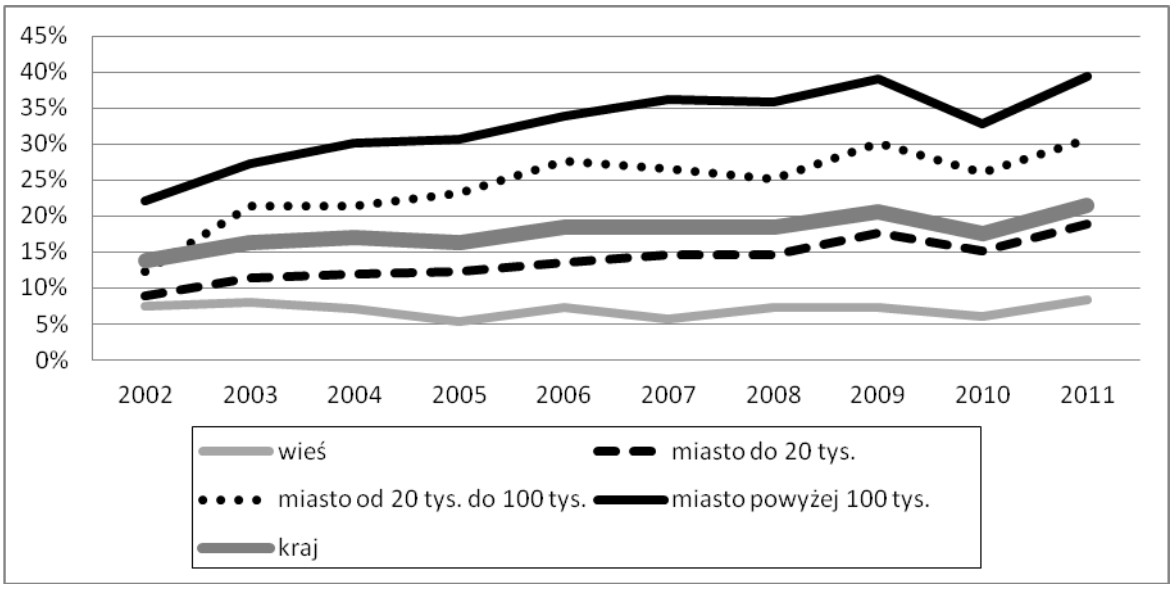


Rysunek 7.14 pokazuje dynamikę różnicowania się szkół gimnazjalnych w zakresie nauczania przedmiotów matematyczno-przyrodniczych. Przedstawiony został zarówno trend dla całego kraju, jak również dla szkół w podziale na 4 typy lokalizacji: wieś, miasto do 20 tysięcy mieszkańców, miasto od 20 do 100 tysięcy mieszkańców oraz miasta powyżej 100 tysięcy mieszkańców. Warto zauważyć, iż biorąc pod uwagę wyniki ogółem dla całego kraju, zróżnicowanie szkół gimnazjalnych spada: z około 25% wariancji wyniku w 2002 roku do 20% wariancji w roku 2011. Biorąc pod uwagę ten nieznaczny, a jednak spadkowy trend na poziomie kraju, interesujące są wyniki w podziale ze względu na wielkość miejscowości, w której zlokalizowana jest szkoła. W roku 2002 gimnazja położone na wsi oraz w dużych miastach cechowały się takim samym poziomem zróżnicowania (ok. 25%), nieco niższe było zróżnicowanie wyników szkół zlokalizowanych w miastach od 20 do 100 tysięcy oraz w miastach do 20 tysięcy. Generalnie rzecz biorąc, w roku 2002 poziom zróżnicowania wyników matematyczno-przyrodniczych szkół ze względu na ich lokalizację był znacznie mniejszy niż w 2012 roku. Na przestrzeni lat 2002-2012 następuje znaczny wzrost zróżnicowania szkół położonych w dużych (powyżej 100 tysięcy mieszkańców) oraz średnich miastach (liczących od 20 do 100 tysięcy mieszkańców). Zauważono, iż podczas gdy następuje znaczne zróżnicowanie i segregacja szkół położonych w dużych i średnich miastach, około dwukrotnie zmniejsza się zróżnicowanie szkół wiejskich, natomiast zróżnicowanie szkół położonych w miastach do 20 tysięcy mieszkańców – po niewielkim spadku zróżnicowania w latach 2003-2010 – w roku 2011 wraca do poziomu z roku 2002.

Można zatem zauważyć, iż na przestrzeni lat 2002-2011 dochodzi do znacznego zróżnicowania i polaryzacji szkół gimnazjalnych położonych w średnich i dużych miastach. W efekcie procesów, które nastąpiły w analizowanym okresie, w 2011 wskaźnik zróżnicowania szkół gimnazjalnych w największych miastach (powyżej 100 tysięcy mieszkańców) wynosi ponad 40%.

Analiza zróżnicowania wyników uczniów szkół gimnazjalnych w zakresie przedmiotów humanistycznych prowadzi do analogicznych wniosków: na przestrzeni lat 2002-2011 nastąpił wzrost zróżnicowania międzyszkolnego wyników. Na poziomie całego kraju zróżnicowanie wyników egzaminu gimnazjalnego w części humanistycznej wzrosło z około 14% do 21% – na poziomie ogólnopolskim obserwujemy zatem niewielki wzrost. Podobnie jak w przypadku części matematyczno-przyrodniczej, zróżnicowanie międzyszkolne wyników jest różne w zależności od lokalizacji szkoły. Generalizując, szkoły położone w miastach są bardziej zróżnicowane niż szkoły wiejskie, a im większe miasto, w którym zlokalizowana jest szkoła, tym większe zróżnicowanie międzyszkolne wyników. Należy również zwrócić uwagę, iż wzrost zróżnicowania wyników w części humanistycznej dotyczy tylko szkół zlokalizowanych w miastach. W szkołach wiejskich w analizowanym okresie zróżnicowanie międzyszkolne jest stabilne i utrzymuje się na poziomie poniżej 10%, podczas gdy w szkołach położonych w średnich i dużych miastach zróżnicowanie wzrosło w tym czasie około dwukrotnie.

Rysunek 7.15. Zróżnicowanie międzyszkolne (wyrażone wskaźnikiem korelacji wewnątrzszkolnej) wyników egzaminu gimnazjalnego (cześć humanistyczna) w podziale na lokalizację szkoły; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2003



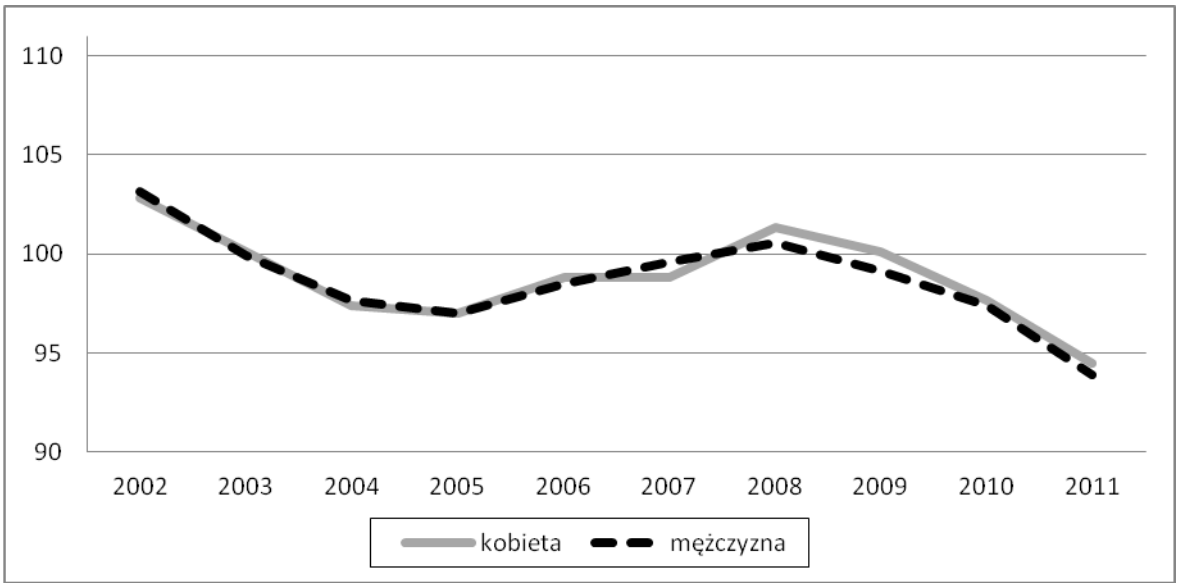
Analiza z uwzględnieniem klasy miejscowości, w której zlokalizowana jest szkoła pozwalają na lepszy wgląd w proces różnicowania się i segregacji szkół gimnazjalnych: możemy zaobserwować wyraźny wzrost zróżnicowania w największych (miasta powyżej 100 tys.) oraz średnich miastach (od 20 do 100 tys. mieszkańców). W przypadku szkół położonych w największych miastach współczynnik zróżnicowania międzyszkolnego w 2011 roku wynosił około 40%, co dotyczy zarówno przedmiotów humanistycznych, jak i matematyczno-przyrodniczych.

7.2.4. Płeć ucznia

Rysunki 7.16 i 7.17 przedstawiają zróżnicowanie wyników chłopców i dziewcząt w części matematyczno-przyrodniczej i humanistycznej egzaminu gimnazjalnego.

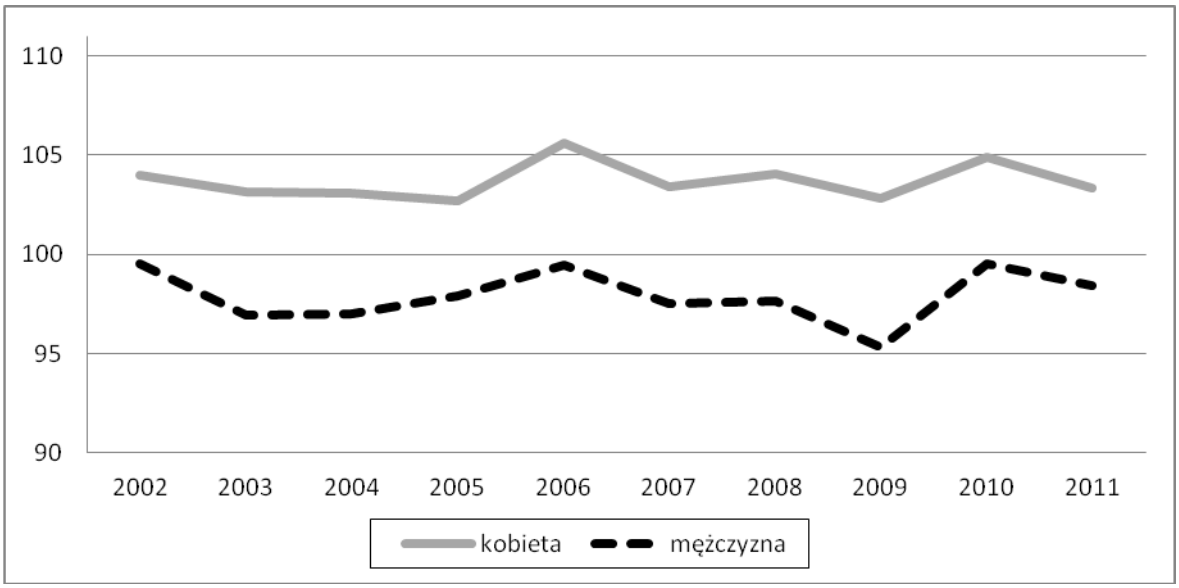
W przypadku części matematyczno-przyrodniczej chłopcy i dziewczęta osiągają średnio takie same wyniki.

Rysunek 7.16. Średni poziom umiejętności uczniów – egzamin gimnazjalny, część matematyczno-przyrodnicza, w podziale ze względu na płeć; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2004



Natomiast w przypadku części humanistycznej, dziewczynki uzyskują lepsze wyniki niż chłopcy. Przewaga dziewcząt nad chłopcami utrzymuje się przez cały badany okres i wynosi od 1/3 do 1/2 odchylenia standardowego.

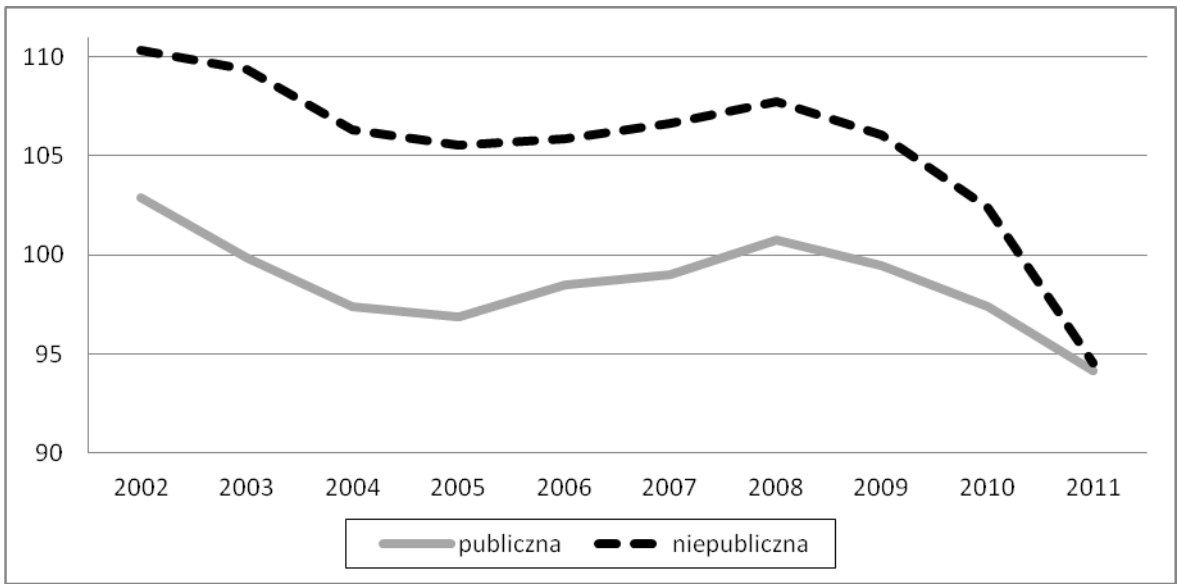
Rysunek 7.17. Średni poziom umiejętności uczniów – egzamin gimnazjalny, część humanistyczna, w podziale ze względu na płeć; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2003



7.2.5. Szkoły publiczne i niepubliczne

Na Rysunkach 7.18 i 7.19 przedstawione jest zróżnicowanie wyników gimnazjalistów z zakresu przedmiotów matematyczno-przyrodniczych i humanistycznych w podziale na szkoły publiczne i niepubliczne. Rysunki pozwalają ocenić, jak na przestrzeni lat 2002-2011 zmieniała się różnica w wynikach egzaminacyjnych uczniów uczęszczających do różnych typów szkół.

Rysunek 7.18. Średni poziom umiejętności uczniów – egzamin gimnazjalny, część matematyczno-przyrodnicza, w podziale na szkoły publiczne i niepubliczne; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2003

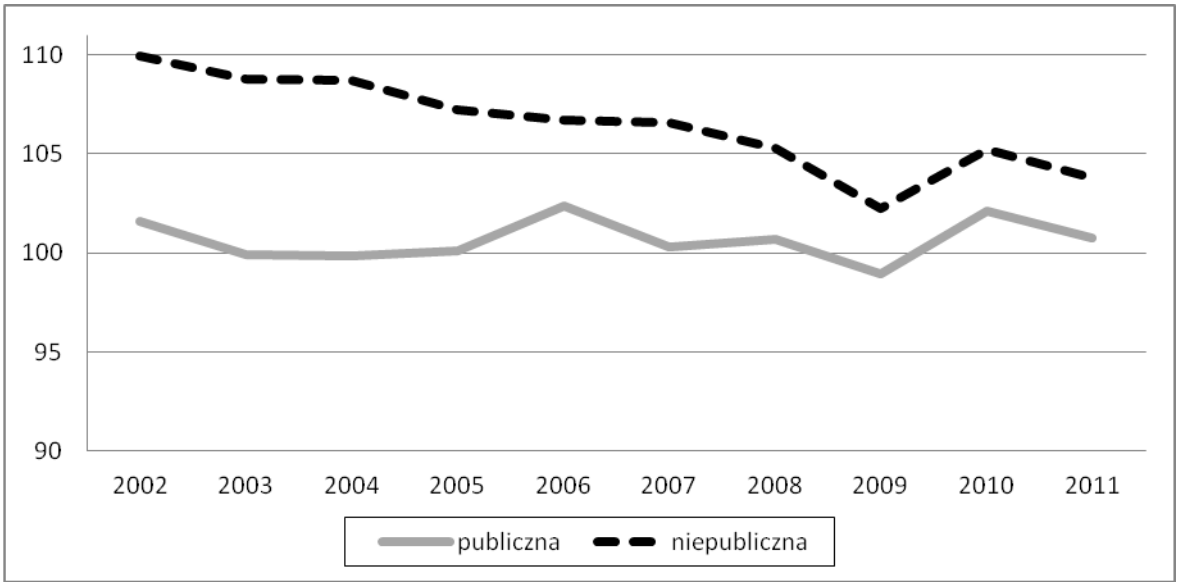


Powyższy rysunek przedstawia średnie wyniki uczniów w części matematyczno-przyrodniczej egzaminu gimnazjalnego w podziale na szkoły publiczne i niepubliczne. W roku 2002, gdy pierwsi uczniowie zdawali egzamin gimnazjalny, uczniowie ze szkół niepublicznych osiągnęli średnie wyniki o ponad 7 punktów wyższe, różnica ta stanowi 1/2 odchylenia standardowego. Aż do roku 2008 szkoły publiczne dzieli duży dystans od szkół niepublicznych, jednak od roku 2008 wyniki uczniów ze szkół niepublicznych zaczynają systematycznie spadać, przez co zmniejsza się różnica między szkołami publicznymi i niepublicznymi, a w roku 2011 wyniki uczniów z obu typów szkół niemal się zrównują.

Rysunek 7.19 przedstawia analogiczne dane dla części humanistycznej egzaminu gimnazjalnego. Podobnie jak w przypadku części matematyczno-przyrodniczej, uczniowie ze szkół niepublicznych osiągają średnio wyższe wyniki aniżeli uczniowie uczęszczający do szkół publicznych. Co istotne, na przestrzeni lat 2002-2011 dystans dzielący szkoły publiczne i niepubliczne systematycznie maleje, co jest spowodowane spadkiem średnich wyników gimnazjalistów ze szkół niepublicznych. Warto również zwrócić uwagę, iż różnica między szkołami publicznymi i niepublicznymi w pierwszym roku egzaminu gimnazjalnego wyniosła ponad 8 punktów, a w 2011 wynosiła już 3 punkty, czyli dystans między szkołami publicznymi i niepublicznymi zmniejszył się o 1/3 odchylenia standardowego. Z tej perspektywy interesujące będą wyniki zrównania części humanistycznej dla roku 2012 i pytanie, czy wówczas, wyniki szkół publicznych i niepublicznych zrównają się, tak jak miało to miejsce w roku 2011 dla części matematyczno-przyrodniczej egzaminu gimnazjalnego.

Analizy na zrównanych wynikach egzaminu gimnazjalnego wykazały, że w latach 2002-2011 nastąpił znaczny spadek wyników uczniów szkół niepublicznych. Pozostaje pytanie, co jest przyczyną takiego spadku wyników osiąganych przez gimnazjalistów ze szkół niepublicznych.

Rysunek 7.19. Średni poziom umiejętności uczniów – egzamin gimnazjalny, część humanistyczna, w podziale na szkoły publiczne i niepubliczne; lata 2002-2011; wyniki zrównane na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2003

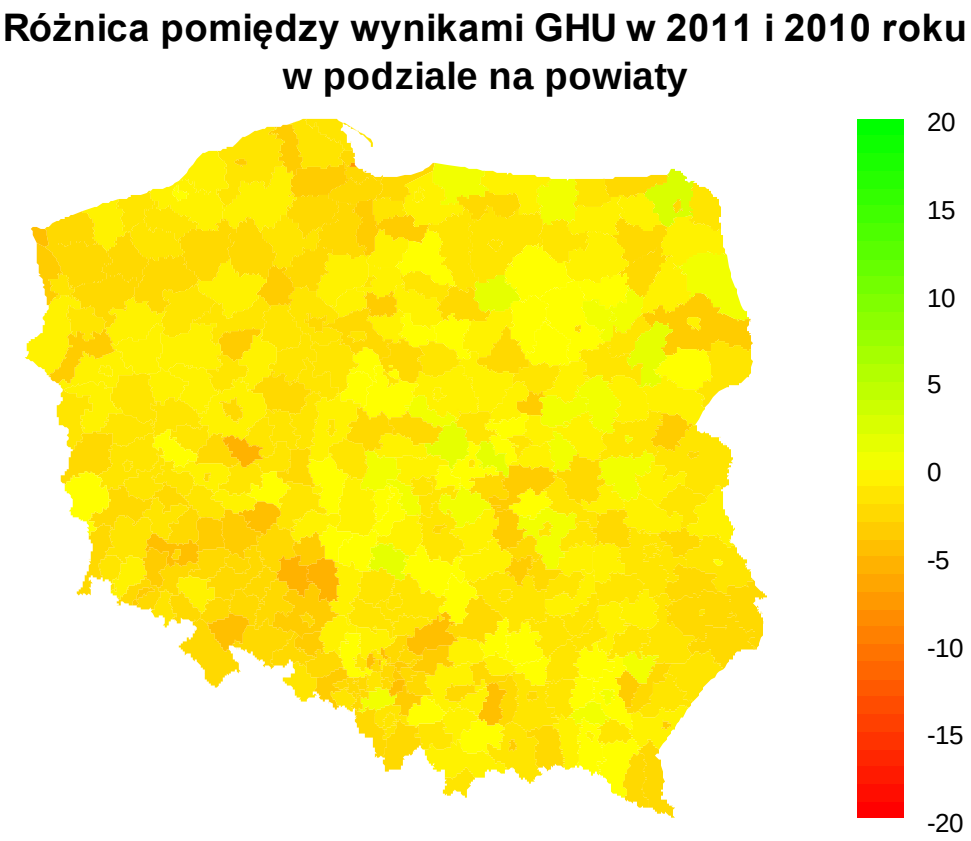


7.2.6. Różnice regionalne

Podczas badania zrównującego w 2012 roku, oprócz zrównania wyników sprawdzianu w latach 2002-2011, dokonano bieżącego zrównania wyników egzaminu gimnazjalnego z 2011 roku. Poniżej przedstawiona zostanie analiza zmian wyników egzaminu gimnazjalnego z roku 2011 w odniesieniu do roku 2010 z uwzględnieniem podziału terytorialnego kraju. Opis wcześniejszych edycji egzaminu w tym kontekście znaleźć można w raporcie z badań z 2011 roku. Wyniki zostały przedstawione na zrównanej skali o średniej 100 i odchyleniu standardowym 15 (zatem różnica rzędu 15 punktów oznacza różnicę jednego odchylenia standardowego). Paleta kolorów została ustalona tak, aby brak różnic (0) oznaczony był kolorem żółtym, różnice ujemne (czyli spadek wyników) rozciągał się do koloru czerwonego, a różnice dodatnie (czyli wzrost wyników) – do koloru zielonego.

W skali całego kraju średnie wyniki zarówno z części humanistycznej egzaminu gimnazjalnego, jak i części matematyczno-przyrodniczej w roku 2011 były nieco niższe niż w roku 2010. Jeśli przyjrzymy się średnim wynikom w rozpatrywanym okresie w podziale na województwa okaże się, że w każdym z województw zanotowano spadek średnich wyników z obydwu części egzaminu. Jednakże w części humanistycznej zaledwie dla dwóch województw, dolnośląskiego i opolskiego, spadek ten jest istotny statystycznie i wynosi około 2,5 punktu. W części matematyczno-przyrodniczej nastąpił statystycznie istotny spadek umiejętności we wszystkich województwach – najmniejsze spadki (poniżej 3 punktów) zanotowano w województwach: świętokrzyskim, podkarpackim i wielkopolskim, najwyższe (powyżej 4 punktów) w województwach: pomorskim i opolskim.

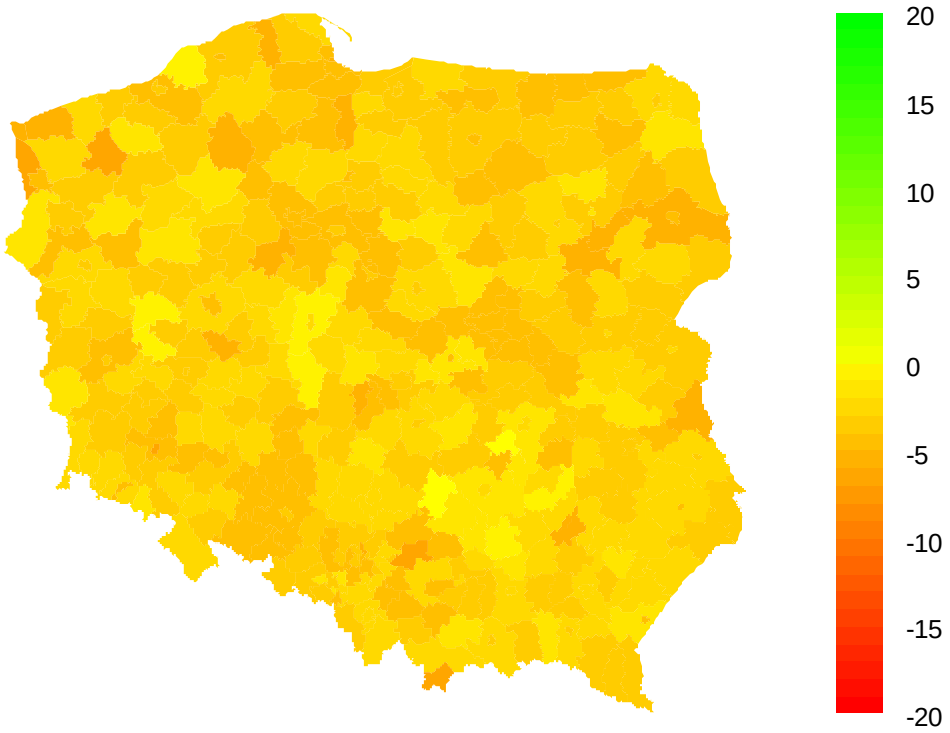
Rysunek 7.20. Różnica średniego poziomu umiejętności z części humanistycznej egzaminu gimnazjalnego pomiędzy rokiem 2011 i 2010 w podziale na powiaty na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2003



Analiza średnich wyników egzaminu w podziale na powiaty pokazuje, że zarówno w części humanistycznej, jak i matematyczno-przyrodniczej, w żadnym z powiatów nie zanotowano statystycznie istotnego wzrostu wyników. W części humanistycznej spadek średnich wyników powiatów w poszczególnych województwach był dość zróżnicowany. W pięciu województwach: lubelskim, lubuskim, łódzkim, świętokrzyskim i warmińsko-mazurskim nie było spadków; w pięciu województwach: dolnośląskim, małopolskim, opolskim, pomorskim i śląskim wyniki spadły w około 20% powiatów (od 18-28%), a w pozostałych województwach spadki nastąpiły w 4-10% powiatów. Jeśli chodzi o wyniki części matematyczno-przyrodniczej, to wyniki w większości województw spadły w co najmniej 2/3 powiatów (w województwach: podlaskim i lubelskim było to około 60%). Jedynie w województwie świętokrzyskim stosunkowo mała część powiatów charakteryzuje się spadkiem średnich wyników – 36%.

Rysunek 7.21. Różnica średniego poziomu umiejętności z części matematyczno-przyrodniczej egzaminu gimnazjalnego pomiędzy rokiem 2011 i 2010 w podziale na powiaty na skali o średniej 100 i odchyleniu standardowym 15 zakotwiczonej w roku 2003

Różnica pomiędzy wynikami GMP w 2011 i 2010 roku w podziale na powiaty



8. Prezentacja porównywalnych wyników egzaminacyjnych – egzamin gimnazjalny

W celu udostępnienia *porównywalnych wyników egzaminacyjnych* szerokiemu gronu odbiorców przygotowany został specjalny serwis internetowy, dostępny pod adresem http://pwe.ibe.edu.pl/. Centralnym punktem prezentacji są wyniki wybranej szkoły, które można osadzić w kontekście wyników innych szkół, gminy, powiatu, województwa, jak również wyników ogólnopolskich. Możliwe jest także dokonywanie porównań pomiędzy jednostkami samorządu terytorialnego. Serwis zawiera gotowe narzędzia do wizualizacji wyników, tak by możliwe było przeprowadzenie najważniejszych analiz bezpośrednio w serwisie. Umożliwia on również pobranie wyników zrównanych dla interesującej użytkownika grupy szkół i/lub jednostek samorządu terytorialnego w postaci zbioru danych, np. w celu wykorzystania ich do analiz w innych programach.

Podczas tworzenia serwisu starano się tam, gdzie to możliwe i celowe, zachować zgodność budowy i funkcjonalności z serwisami prezentującymi wyniki Edukacyjnej Wartości Dodanej[26]. Miało to na celu ułatwienie korzystania z serwisu porównywalnych wyników tym użytkownikom, którzy zetknęli się już z serwisami EWD. Dodatkowo umożliwi to w przyszłości zintegrowanie obu serwisów z wynikami egzaminacyjnymi gimnazjów.

8.1. Budowa serwisu

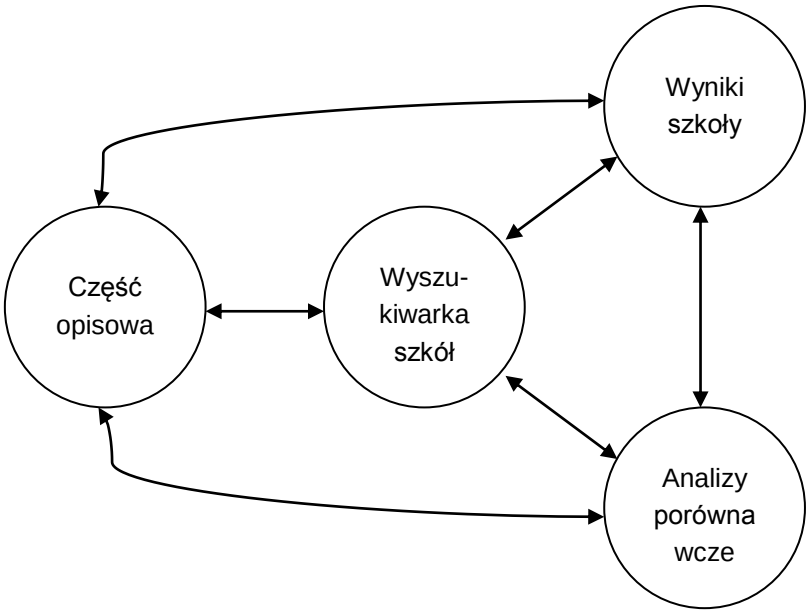
Serwis podzielony jest na cztery części:

1. część opisową,
2. wyszukiwarkę szkół,
3. moduł prezentacji wyników dla wybranej szkoły,
4. moduł analiz porównawczych.

Możliwości przechodzenia pomiędzy poszczególnymi modułami zaprezentowano na Rysunku 8.1.

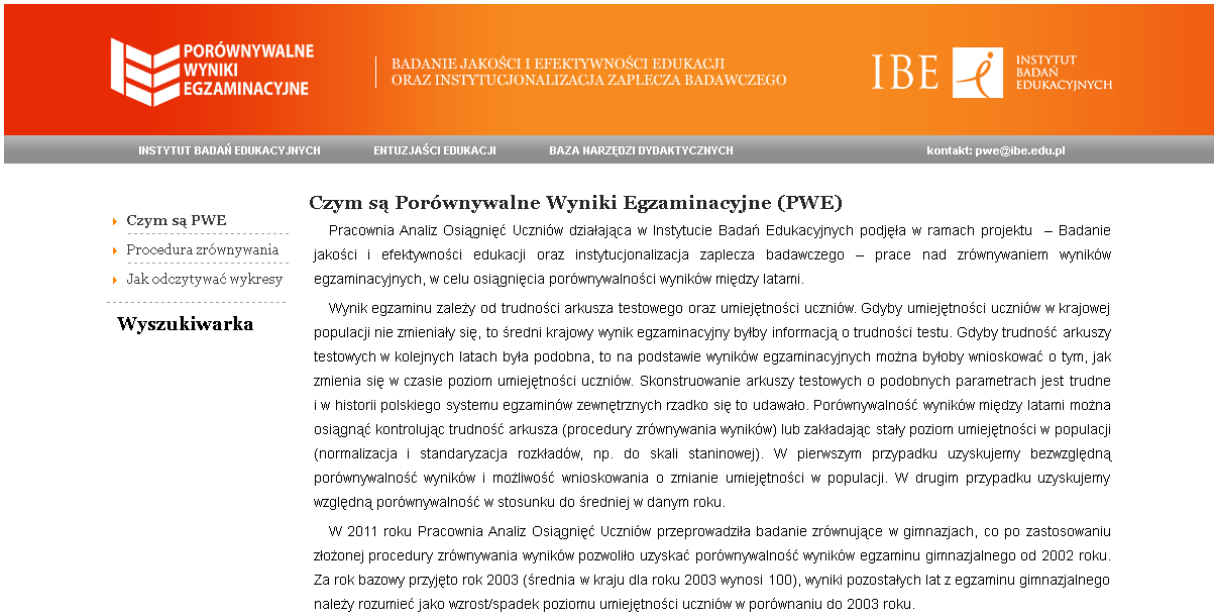
[26] http://gimnazjum.ewd.edu.pl/ oraz http://matura.ewd.edu.pl/

Rysunek 8.1. Powiązania między modułami serwisu PWE



Część opisowa w syntetyczny sposób przybliża użytkownikowi informacje o celu i metodologii zrównywania wyników egzaminacyjnych, opisuje również w przystępnej, ilustrowanej przykładami formie sposób interpretacji prezentowanych wyników. W tym celu podzielona została ona na trzy podstrony: *Czym są PWE*, *Procedura zrównywania* oraz *Jak odczytywać wykresy?*

Rysunek 8.2. Część opisowa serwisu pwe.ibe.edu.pl



Wyszukiwarka szkół umożliwia wybranie szkoły, której wyniki mają zostać zaprezentowane i/lub szkół, których wyniki mają być ze sobą porównane. Szkoły wyszukiwać można wg kilku kryteriów:

- jednostki samorządu terytorialnego, w której znajduje się szkoła – województwa, a później powiaty i gminy wybiera się z list rozwijalnych, które umożliwiają wybór dowolnej jednostki terytorialnej wg obowiązującego obecnie podziału Polski do poziomu gminy włącznie;

- wpisywanego ręcznie fragmentu nazwy i/lub adresu szkoły;

- dostępności wyników zrównanych danej szkoły dla wskazanych lat (jeśli nie podano zakresu lat wyszukiwane są szkoły, dla których dostępne są wyniki zrównane z ostatniego roku).

Powyższe kryteria można łączyć. Z poziomu wyszukiwarki szkół możliwe jest przejście do prezentacji wyników wybranej szkoły (poprzez kliknięcie na jej nazwę), dodanie wszystkich wyszukanych szkół do modułu analiz porównawczych, jak również przejście do modułu analiz porównawczych.

Rysunek 8.3. Przykład działania wyszukiwarki – wyszukano szkoły, które mają w nazwie słowo „im."

Rysunek 8.4. Przykład działania wyszukiwarki – wyszukano gimnazja województwa wielkopolskiego, które mieszczą się na ul. Szkolnej

Samo przeglądanie porównywalnych wyników egzaminacyjnych dostępne jest na dwa sposoby:

1) Poprzez moduł prezentacji wyników dla wybranej szkoły. Zrównane wyniki gimnazjum prezentowane są w tym module w podziale na części egzaminu gimnazjalnego[27]. Wynik danej szkoły można obejrzeć w kontekście wyników gminy i/lub powiatu i/lub województwa, w którym szkoła się znajduje, jak również w kontekście wyników ogólnopolskich. Z poziomu modułu prezentacji wyników wybranego gimnazjum możliwe jest dodanie tej szkoły do modułu analiz porównawczych, jak również przejście do serwisu EWD prezentującego wyniki edukacyjnej wartości dodanej dla danej szkoły.

2) Poprzez moduł analiz porównawczych. W tym wypadku prezentowane są wyniki wybranych szkół i/lub wybranych jednostek samorządu terytorialnego i/lub wyniki ogólnopolskie ze wskazanej części egzaminu gimnazjalnego. Szkoły dodaje się do modułu analiz porównawczych poprzez wyszukiwarkę lub moduł prezentacji wyników wybranej szkoły. Jednostki samorządu terytorialnego oraz prezentację wyników ogólnopolskich dodaje się do modułu analiz porównawczych w samym module. Poprzez kliknięcie na nazwie dowolnej szkoły znajdującej się w module analiz porównawczych możliwe jest przejście do modułu prezentacji wyników danej szkoły.

W ramach obydwu opisanych wyżej modułów prezentacja porównywalnych wyników egzaminacyjnych jest możliwa na trzy sposoby:

1) w postaci wykresu liniowego (tzw. *wykres podstawowy*) prezentującego zmianę wartości wyników zrównanych w czasie;

[27] W latach 2002-2011 są to część *humanistyczna* oraz część *matematyczno-przyrodnicza*.

2) w postaci wykresu skrzynkowego (tzw. *wykres zaawansowany*) prezentującego zmianę wybranych statystyk wyników zrównanych w czasie;

3) w postaci tabeli danych zawierającej dane prezentowane graficznie na wykresie liniowym bądź skrzynkowym w postaci tabelarycznej.

Rysunek 8.5. Przykład działania modułu porównywania szkół

8.2. Prezentacja wyników

8.2.1. Skala prezentacji wyników

Wyniki zrównane prezentowane są na skali, dla której wartość 100 odpowiada średniemu wynikowi uczniów piszących arkusz standardowy danej części egzaminu gimnazjalnego (humanistycznej lub matematyczno-przyrodniczej) w 2003 roku, natomiast różnica 15 punktów na skali odpowiada jednemu odchyleniu standardowemu wyników uczniów z danej części tego egzaminu. Skala ta będzie w dalszej części nazywana skrótowo skalą *(100; 15)*. Rok 2003 został w tym wypadku wybrany

arbitralnie – z metodologicznego punktu widzenia mógłby to być dowolny rok spośród lat, za które zrównywane były wyniki. Wyniki na skali *(100; 15)* należy więc interpretować jako *wynik, jaki osiągnąłby dany uczeń/szkoła/itd., gdyby napisał egzamin gimnazjalny z roku 2003, w odniesieniu do wyników uczniów, którzy faktycznie pisali egzamin w roku 2003.* Interpretacja ta dobrze oddaje istotę zrównywania wyników egzaminacyjnych – wyrażamy wyniki z różnych egzaminów za pomocą skali jednego, wybranego z nich.

To, co w sposób istotny różni skalę *(100; 15)* od skal, na jakich prezentowane są niezrównane wyniki egzaminu gimnazjalnego (liczba punktów uzyskanych na egzaminie i/lub skala procentowa od 0 do 100%) to brak minimum i maksimum. O ile np. na skali procentowej nie da się uzyskać wyniku niższego od 0% (minimum) ani wyższego niż 100% (maksimum), o tyle skala *(100; 15)* rozciąga się od minus nieskończoności do plus nieskończoności. Stąd dla skali *(100; 15)* możemy jedynie mówić o tym, jak prawdopodobne jest uzyskanie wskazanego wyniku, np. wynik mniejszy od 40 (niższy o ponad 4 odchylenia standardowe od średniej) spodziewamy się zaobserwować tylko dla ok. 0,003% uczniów, możemy więc go uznać za bardzo mało prawdopodobny. Nie oznacza to jednak ani że nie może on wystąpić, ani że musi wystąpić. To, jaki odsetek uczniów spodziewamy się zaobserwować w jakim przedziale wyników odczytywane jest z rozkładu normalnego o średniej 100 i odchyleniu standardowym 15.

8.2.2. Wykres liniowy

Wykres liniowy jest domyślnym sposobem prezentacji porównywalnych wyników egzaminacyjnych zarówno w module prezentacji wyników wybranej szkoły, jak i module analiz porównawczych. Przedstawia on średni zrównany wynik egzaminacyjny z wybranej części egzaminu gimnazjalnego wraz z 95% przedziałem ufności dla tej średniej. Prezentacja na wykresie przedziałów ufności pozwala w łatwy sposób stwierdzić, które różnice są istotne statystycznie – istotnie statystycznie różne są od siebie te wartości średnich, których przedziały ufności się nie przecinają. Korzystanie z tej reguły podczas porównywania ze sobą szkół, gmin, itp. jest kluczowe do tego, aby formułowane wnioski były poprawne metodologicznie.

Szerokość przedziału ufności zależy od dwóch czynników:

1) tego, jak spójne były wyniki egzaminacyjne osiągnięte przez uczniów w danej szkole, gminie, itp., – im wyniki w danej grupie mniej oddalone od średniej danej grupy, tym węższy jest przedział ufności;
2) tego, jak dużo uczniów przystąpiło do egzaminu w danej szkole, gminie, itp.– im większa liczba osób w danej grupie, tym węższy jest przedział ufności.

Szczególnie wyraźnie widoczny jest wpływ drugiego z wymienionych czynników – to on powoduje, że przedziały ufności dla gmin są na ogół[28] węższe od przedziałów ufności szkół, przedziały ufności powiatów są węższe od przedziałów ufności gmin, itd., najwęższy jest przedział ufności dla średniego wyniku ogólnopolskiego. Natomiast wśród szkół, gmin, itp., w których do egzaminu gimnazjalnego w danym roku przystąpiła zbliżona liczba uczniów, o tym, która z nich będzie posiadać węższy, a która szerszy przedział ufności, decydować będzie zróżnicowanie wyników egzaminacyjnych w ramach

[28] Wyjątek stanowią gminy, w których znajduje się tylko jedno gimnazjum – w takim wypadku wykres dla gminy pokrywa się z wykresem dla szkoły.

danej szkoły, gminy, itp. Aby móc rozróżnić od siebie wpływ obydwu wymienionych czynników w tabeli danych pod wykresem prezentowana jest liczba uczniów danej szkoły, gminy, itp., natomiast na analizie zróżnicowania wyników w ramach danej szkoły, gminy, itp. skupia się wykres skrzynkowy.

8.2.3. Wykres skrzynkowy

Wykres skrzynkowy, nazywany w serwisie wykresem zaawansowanym, jest alternatywną formą prezentacji wyników porównywalnych wyników egzaminacyjnych. Prezentuje on bardziej szczegółowe informacje niż wykres liniowy, przez co daje możliwość przeprowadzenia bardziej wnikliwych analiz, lecz również jest trudniejszy w interpretacji. Na wykresie pokazywane są:

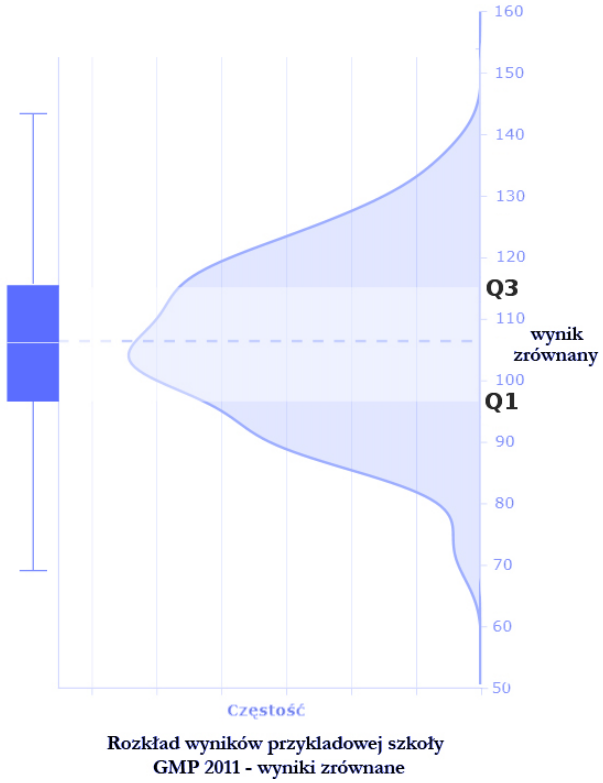
- pierwszy (Q1), drugi (mediana) i trzeci kwartyl (Q3), oznaczane odpowiednio przez dolną krawędź skrzynki, linię w środku skrzynki oraz górną krawędź skrzynki na wykresie;

- wynik minimalny i maksymalny z pominięciem obserwacji odstających, oznaczone odpowiednio przez końcówki dolnego i górnego z „wąsów" odchodzących od skrzynki. Za obserwacje odstające uznawane są te wyniki egzaminacyjne, które są niższe od wartości pierwszego kwartyla pomniejszonej o półtorakrotność rozstępu międzyćwiartkowego[29] lub wyższe od wartości trzeciego kwartyla powiększonego o półtorakrotność rozstępu międzyćwiartkowego.

O ile wykres podstawowy skupia się na prezentacji średniego zrównanego wyniku egzaminacyjnego dla danej szkoły, gminy, itp., o tyle wykres zaawansowany koncentruje się na tym, w jaki sposób rozkładają się wyniki poszczególnych grup uczniów w ramach danej szkoły, gminy, itp. Uczniowie najsłabsi reprezentowani są przez koniec dolnego „wąsa", słabi przez pozycję dolnego końca skrzynki, przeciętni przez medianę, dobrzy przez pozycję górnego końca skrzynki, najlepsi przez koniec górnego „wąsa". Im większa wysokość skrzynki i długość „wąsów", tym bardziej zróżnicowane wyniki uczniów. Wykres skrzynkowy można interpretować jako inną prezentację rozkładu wyników, co ilustruje poniższy rysunek.

[29] Rozstęp międzyćwiartkowy jest równy różnicy wartości trzeciego i pierwszego kwartyla.

Rysunek 8.6. Rozkład wyników przykładowej szkoły



8.2.4. Tabela danych

Prezentacji wykresu liniowego i skrzynkowego towarzyszy zawsze tabela danych. Zawiera ona te same informacje, które prezentowane są na danym wykresie oraz dodatkowo liczbę uczniów uwzględnionych przy obliczaniu wyników zrównanych dla danej szkoły, gminy, itp. za dany rok. Tabela danych pozwala odczytać dokładne wartości liczbowe prezentowanych na wykresie statystyk. Tabelę danych można również pobrać jako plik CSV w celu wykonania samodzielnych analiz i/lub wykresów, np. w arkuszu kalkulacyjnym lub programie statystycznym.

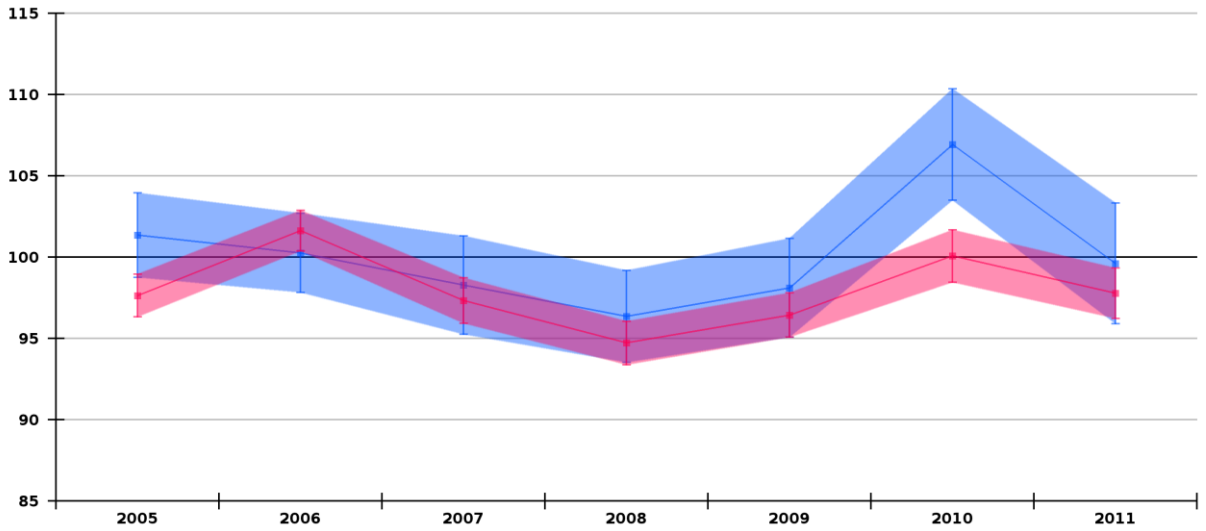
8.3. Przykłady

Na Rysunkach 8.7, 8.8 i 8.9 przedstawiono prezentacje porównywalnych wyników dla przykładowych szkół na dwóch typach wykresów: wykresie podstawowym (Rysunki 8.7 i 8.8) i wykresie zaawansowanym (Rysunek 8.9). Wyniki przykładowej szkoły na wykresie podstawowym pokazane są na tle powiatu, natomiast na wykresie zaawansowanym porównano dwa przykładowe gimnazja. Przedstawione wykresy zostały opatrzone komentarzem interpretacyjnym.

8.3.1. Przykład 1.

Gimnazjum nr 2 (kolor niebieski) oraz powiat, w którym się ono znajduje (kolor różowy), część humanistyczna egzaminu. Wykres podstawowy.

Rysunek 8.7. Prezentacja wyników dla przykładowej szkoły – część humanistyczna



rok		2005	2006	2007	2008	2009	2010	2011
Publiczne Gimnazjum nr 2	min. przedz. ufności	98,73	97,80	95,23	93,52	95,05	103,48	95,88
	średnia	101,34	100,24	98,26	96,34	98,09	106,91	99,60
	maks. przedz. ufności	103,95	102,68	101,28	99,16	101,14	110,34	103,31
	liczba uczniów	122	142	135	117	118	101	105
powiat	min. przedz. ufności	96,31	100,36	95,92	93,36	95,06	98,44	96,20
	średnia	97,62	101,60	97,31	94,70	96,42	100,06	97,75
	maks. przedz. ufności	98,93	102,85	98,70	96,05	97,78	101,67	99,30
	liczba uczniów	600	627	588	534	547	497	495

Wyniki szkoły w latach 2005-2007 oraz 2009 i 2011 nie różnią się w sposób istotny statystycznie od średniej ogólnopolskiej z roku 2003 (wartość 100 na osi Y), gdyż linia odpowiadająca wartości 100 przecina w tych latach przedział ufności średniego wyniku szkoły.

W roku 2008 szkoła zanotowała średni wynik egzaminacyjny istotnie statystycznie niższy od średniej ogólnopolskiej z roku 2003, gdyż linia odpowiadająca wartości 100 nie przecina przedziału ufności wyniku szkoły w tym roku. Rok 2008 jest jednocześnie rokiem, w którym szkoła zanotowała najsłabszy wynik w całym rozważanym okresie 7 lat, będący kulminacją czteroletniego trendu pogarszania się wyników egzaminacyjnych trzecioklasistów tej szkoły.

Można jednak zauważyć, że wspomniany powyżej trend spadkowy jest w latach 2006-2008 zgodny z trendem wyników egzaminacyjnych dla całego powiatu, dotyczył on więc również (choć zapewne w różnym stopniu – wymagałoby to dalszych analiz) także innych szkół w tym powiecie.

Od 2008 roku obserwowana jest poprawa wyników egzaminacyjnych osiąganych przez szkołę. Poprawa ta pozwoliła w 2009 roku powrócić szkole do grupy szkół osiągających wyniki zbliżone do średniej z 2003 roku (linia odpowiadająca wartości 100 przecina przedział ufności wyników szkoły), natomiast w roku 2010 osiągnąć zrównane wyniki istotnie statystycznie lepsze od średniego wyniku

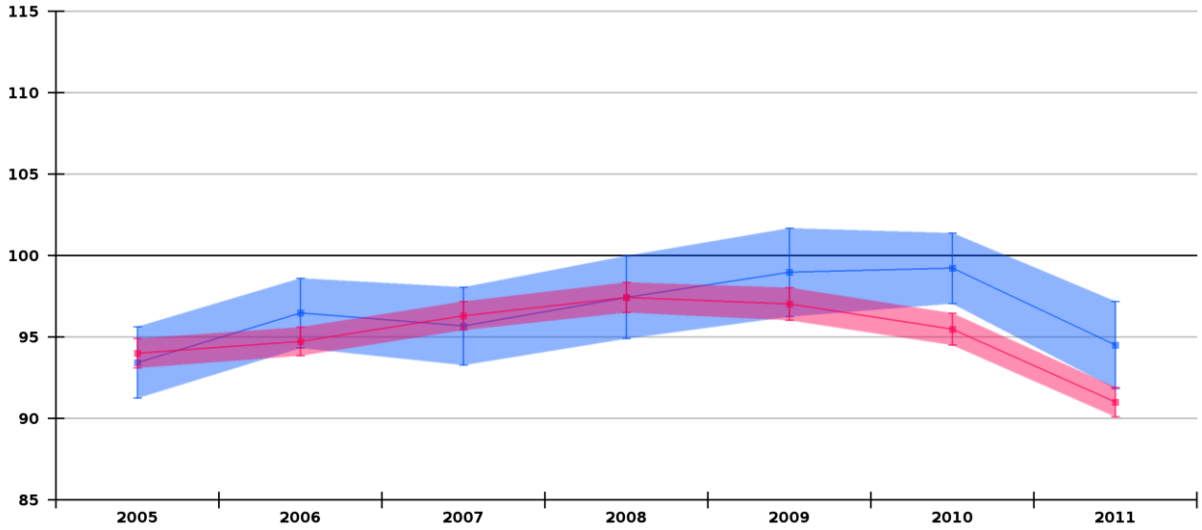
egzaminacyjnego z 2003 roku (linia odpowiadająca wartości 100 znajduje się poniżej przedziału ufności wyniku szkoły).

Jednocześnie 2010 rok jest jedynym rokiem w rozpatrywanym okresie, dla którego szkoła osiągnęła średni wynik egzaminacyjny istotnie statystycznie różny (w tym wypadku lepszy) od wyników dla całego powiatu. Nie był to jednak typowy rok dla szkoły. W 2011 roku ponownie wynik szkoły nie różni się statystycznie od średniego poziomu osiągnięć polskich gimnazjalistów w 2003 roku.

8.3.2. Przykład 2.

Gimnazjum nr 1 (kolor niebieski) oraz powiat, w którym się ono znajduje (kolor czerwony), część matematyczno-przyrodnicza egzaminu. Wykres podstawowy.

Rysunek 8.8. Prezentacja wyników dla przykładowej szkoły – część matematyczno-przyrodnicza



rok		2005	2006	2007	2008	2009	2010	2011
Gimnazjum nr 1	min. przedz. ufności	91,24	94,32	93,27	94,88	96,23	97,03	91,82
	średnia	93,42	96,45	95,65	97,42	98,95	99,20	94,49
	maks. przedz. ufności	95,60	98,58	98,04	99,97	101,67	101,37	97,16
	liczba uczniów	208	217	178	168	162	185	153
powiat	min. przedz. ufności	93,10	93,83	95,41	96,49	96,00	94,50	90,08
	średnia	93,99	94,71	96,29	97,41	97,00	95,47	90,99
	maks. przedz. ufności	94,88	95,58	97,17	98,33	98,00	96,45	91,89
	liczba uczniów	1454	1429	1357	1279	1252	1194	1081

Wyniki za lata 2005-2008 pokazują, że w tych latach uczniowie tej szkoły osiągali na egzaminie matematyczno-przyrodniczym wyniki słabsze od przeciętnego wyniku w roku 2003 (linia 100 obrazująca średni wynik egzaminu w 2003 roku znajduje się ponad przedziałem ufności wyniku szkoły).

Jednocześnie warto zauważyć, że wyniki szkoły ulegają od roku 2007 poprawie, co pozwala w latach 2009-2010 zakwalifikować szkołę do grupy szkół osiągających wyniki bliskie średniej z roku 2003 (linia

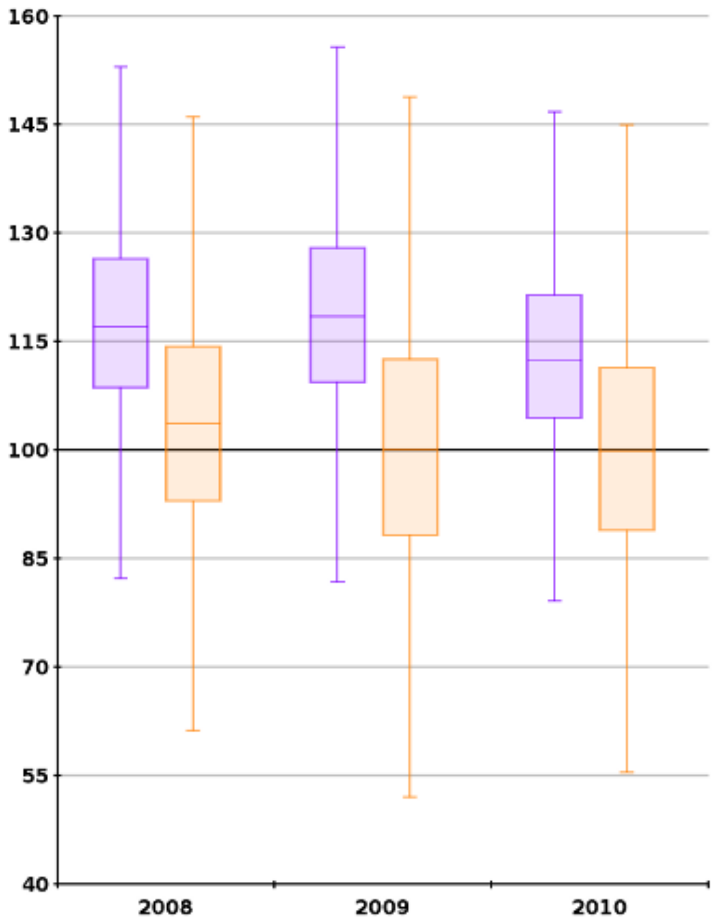
100 przecina w 2009 i 2010 roku przedział ufności wyniku szkoły). Rok 2011 okazał się słabszy – uczniowie osiągnęli wyniki na poziomie wyników z 2007 roku. Spadek wyników obserwujemy również w powiecie, gdzie występuje już od roku 2008. W związku ze spadkiem wyników w powiecie w roku 2010 i utrzymaniem przez szkołę zbliżonych wyników w latach 2009-2010, w roku 2010 szkole udało się uzyskać wynik istotnie statystycznie wyższy od średniej dla powiatu. Niestety w roku 2011 nie udało się utrzymać tej przewagi – zauważalny spadek wyników szkoły w tym roku połączony z proporcjonalnie niższym spadkiem wyników w całym powiecie spowodował, że średni wynik szkoły w roku 2011 ponownie nie różni się statystycznie istotnie od średniej dla powiatu.

Analizując sytuację powiatu można stwierdzić, że mamy do czynienia z powiatem, w którym uczniowie osiągają w sposób systematyczny wyniki słabsze od średniego wyniku egzaminu z 2003 roku. W latach 2005-2010 było to ok. 1/3 odchylenia standardowego, w 2011 wynik jest niższy od średniej 2003 roku o 2/3 odchylenia standardowego.

8.3.3. Przykład 3.

Gimnazjum C (kolor fioletowy) oraz gimnazjum D (kolor pomarańczowy), część matematyczno-przyrodnicza egzaminu. Wykres zaawansowany.

Rysunek 8.9. Prezentacja wyników dla przykładowej szkoły – część matematyczno-przyrodnicza (wykres zaawansowany)



Gimnazja C i D dość znacznie się różnią:

- Gimnazjum D osiąga wyniki bliskie średniemu wynikowi z roku 2003 (mediana w latach 2009-2010 równa 100 oznacza, że w tym okresie równo połowa uczniów osiągnęła wyniki lepsze, jak i słabsze od średniego wyniku z 2003 roku), podczas gdy w gimnazjum C w całym analizowanym okresie ponad 75% uczniów osiągało wyniki lepsze od średniego wyniku z 2003 roku (dolna granica skrzynki, a więc 1 kwartyl, powyżej linii 100), a mediana w latach 2008-2009 jest o ponad 15 (a więc ponad jedno odchylenie standardowe) wyższa od średniego wyniku ogólnopolskiego z roku 2003.

- Zróżnicowanie wyników osiąganych przez uczniów wewnątrz szkoły jest w gimnazjum C mniejsze (różnica między 1. a 3. kwartylem wynosi ok. 18) niż w gimnazjum D (tu różnica między 1. a 3. kwartylem wynosi ok. 22).
- O ile w obydwu szkołach wyniki najlepszych uczniów są do siebie zbliżone (górne „wąsy" skrzynek kończą się na podobnym poziomie – ok. 145-150 punktów na skali wykresu), o tyle najsłabsi uczniowie w gimnazjum C osiągają znacznie lepsze wyniki od najsłabszych uczniów w gimnazjum D – dolne „wąsy" skrzynek gimnazjum C kończą się na poziomie ok. 77-80 (ok. 1,5 odchylenia standardowego poniżej średniego wyniku z 2003 roku), podczas gdy gimnazjum D na poziomie ok. 50-65 (2,3-3,3 odchylenia standardowego poniżej średniego wyniku z 2003 roku).

8.4. Wyczyszczenie bazy danych wyników egzaminacyjnych oraz bazy szkół

Aby umożliwić analizę wyników zrównanych w podziale na regiony kraju, a także prezentację wyników zrównanych w serwisie internetowym niezbędne było:

- Stworzenie bazy danych adresowych szkół zawierającej dla każdej szkoły jej adres i kod TERYT gminy, w której dana szkoła się znajduje, dla każdego z lat 2002-2011.
    - W wypadku zrównanych wyników egzaminu gimnazjalnego jako bazę szkół wykorzystano bazę prowadzoną przez zespół Edukacyjnej Wartości Dodanej. Baza ta zawierała wszystkie niezbędne informacje.
    - W wypadku zrównanych wyników sprawdzianu w szóstej klasie szkoły podstawowej jako bazę szkół wykorzystano zestawienia zamówień na arkusze egzaminacyjne z lat 2004-2011 udostępnione przez Centralną Komisję Egzaminacyjną. Baza ta wymagała uzupełnienia o kody TERYT gmin.
- Połączenie bazy wyników zrównanych z bazą adresową szkół.
- Powiązanie szkół w bazie szkół między latami.
    - W wypadku bazy gimnazjów prowadzonej przez zespół Edukacyjnej Wartości Dodanej, szkoły były już ze sobą powiązane między latami.

W trakcie wykonywania wymienionych wyżej czynności napotkano na wiele problemów, które szczegółowo opisane zostały poniżej.

8.4.1. Określanie kodu TERYT gminy, w której znajduje się szkoła

Kod TERYT gminy, w której w danym roku znajduje się dana szkoła, można na podstawie bazy zamówień na arkusze egzaminacyjne ustalić na trzy sposoby:

1) na podstawie pierwszego członu kodu szkoły nadanego jej przez Okręgową Komisję Egzaminacyjną (co do zasady pierwsze sześć cyfr kodu szkoły to kod TERYT gminy, w której znajduje się szkoła w momencie nadania kodu OKE – od zasady tej istnieją jedynie nieliczne odstępstwa);
2) na podstawie adresu szkoły w danym roku zestawionego z bazą danych kodów pocztowych udostępnianą przez Pocztę Polską;
3) na podstawie kodu TERYT gminy odczytanego dla danej szkoły w innych latach.

Możliwość stosowania pierwszej z wymienionych metod ograniczają z jednej strony wyjątki od jej stosowania (na szczęście nieliczne, nie będą dokładniej omawiane) oraz zmiany w podziale terytorialnym Polski. Jeśli dana Okręgowa Komisja Egzaminacyjna aktualizuje kod OKE szkoły wraz ze zmianami podziału administracyjnego kraju, wtedy pierwszy człon kodu OKE będzie wskazywał na właściwą gminę. Jeśli natomiast kod OKE szkoły jest niezmienny w czasie, wtedy zmiana kodu TERYT gminy, w której znajduje się szkoła, spowoduje, że pierwszy człon kodu OKE szkoły przestanie wskazywać na tę gminę. Ponieważ Główny Urząd Statystyczny, odpowiadający za nadawanie kodów TERYT, nie wykorzystuje ponownie już raz użytych kodów TERYT, mamy gwarancję, że w wypadku takiej zmiany kod OKE szkoły nie będzie wskazywał na żadną gminę. Jeśli jednak wykryta zostanie taka sytuacja:

- Należy spróbować odnieść pierwszy człon kodu OKE szkoły do wiedzy na temat zmian w strukturze administracyjnej Polski. Na przestrzeni lat 2002-2012 miało miejsce 36 takich zmian, w wypadku których daje się jednoznacznie określić, jaki TERYT gminy przed zmianą odpowiada jakiemu kodowi TERYT po zmianie.

- Jeśli kod TERYT będący pierwszym członem kodu OKE szkoły nie pasuje do żadnej z tych zmian, pozostaje ustalić kod TERYT gminy, w której znajduje się szkoła, na podstawie adresu szkoły.

W praktyce kod OKE szkoły okazał się być dobrym źródłem danych na temat kodu TERYT gminy, w której znajduje się dana szkoła w danym roku – na 118 023 rekordy, odpowiadające szkołom podstawowym, które co najmniej raz w latach 2002-2011 przeprowadzały sprawdzian, dla 115 563 z nich (97,9%) udało się ustalić kod TERYT gminy wprost z kodu OKE szkoły, a po uwzględnieniu wspomnianych wyżej 36 zmian powodujących jednoznaczną zamianę kodu TERYT gminy kod TERYT gminy udało się odczytać dla 115 693 rekordów (98,0%).

Możliwości stosowania drugiej z wymienionych metod ograniczane są błędami w danych adresowych szkół. W większości były to literówki w kodach pocztowych, nazwach miejscowości i ulic, niekiedy jednak błędy te występowały systematycznie (np. szkoła konsekwentnie, rok po roku, podaje na zamówieniach na arkusze egzaminacyjne nieprawidłowy kod pocztowy). Skala występowania tego typu błędów była znaczna – na 118 023 rekordy adresów szkół w poszczególnych latach kod TERYT gminy udało się przypisać 95 475 (80,9%). Szczęśliwie, błędy te były w większości rozłączne z przypadkami, w których kodu TERYT nie daje się dopasować na podstawie kodu OKE szkoły – na 118 023 rekordy adresów szkół odnotowano tylko 96 sytuacji (mniej niż 0,1%), gdy obydwa problemy występowały łącznie. Zostały one poprawione ręcznie na podstawie analizy danych adresowych szkoły w bazie zamówień na arkusze oraz kodów pocztowych i adresów występujących w wykazie kodów pocztowych Poczty Polskiej.

Dla szkół, których kod TERYT udało się ustalić zarówno na podstawie kodu OKE szkoły, jak i adresu szkoły, możliwa była weryfikacja zgodności tych dwóch źródeł. Zgodność okazała się wysoka –

wyniosła 98,8%. W wypadku napotkania niezgodności za bardziej wiarygodny uznawany był kod TERYT odczytany na podstawie adresu szkoły. Decyzja ta wynika z tego, że w wypadku popełnienia literówki w pierwszym członie kodu OKE szkoły prawdopodobne jest, że, mimo błędu, pozostanie on poprawnym w danym roku kodem TERYT. Tymczasem popełnienie literówki w danych adresowych, gdzie kod TERYT dopasowywany jest zawsze co najmniej na podstawie kodu pocztowego i miejscowości, prawie zawsze skutkuje niedopasowaniem kodu TERYT na podstawie adresu[30].

Analizując liczbę niezgodności pomiędzy kodami TERYT odczytanymi z pierwszego członu kodów OKE szkół i odnosząc to do liczby szkół, w których nie udało się ustalić kodu TERYT na podstawie adresu można oszacować, jak dużo takich konfliktów zostałoby dodatkowo wykrytych, gdyby kod TERYT gminy, w której znajduje się szkoła, udało się na podstawie adresu ustalić dla wszystkich szkół. Przewidywanie wykazuje, że wśród 22 548 szkół błąd zostałby wykryty wśród ok. 280. Wyszukiwanie tych błędów nie zostało jednak przeprowadzone z uwagi na pracochłonność procesu – aby wykryć szacowaną ok. 280 błędów należałoby poprawić dane adresowe 22 548 szkół, co stanowi wyzwanie ponad siły zespołu zajmującego się badaniem.

Podsumowując, dzięki zastosowaniu dwóch uzupełniających się metod przypisywania szkołom kodu TERYT udało się uzyskać bazę szkół podstawowych, w której każda szkoła przyporządkowana jest do jakiejś gminy. Co prawda można się spodziewać, że pewien odsetek tych przypisań (ok. 0,2%) jest błędny, jednak znalezienie tych błędów jest zbyt czasochłonne, by móc je przeprowadzić.

8.4.2. Łączenie bazy porównywalnych wyników egzaminacyjnych z bazą szkół

Złączenie zbioru porównywalnych wyników egzaminacyjnych z uzupełnioną bazą szkół podstawowych oraz bazą gimnazjów odbywało się na podstawie kodu OKE szkoły. Podczas złączania odnotowano pewne rozbieżności:

- kody OKE szkoły, które występują w bazach zrównanych wyników za dany rok, ale nie występują w bazie szkół:
  - dla lat, których nie obejmuje baza szkół (2002-2003 dla szkół podstawowych, 2002-2004 dla gimnazjów) – w żadnym z lat objętych bazą szkół;
  - dla lat objętych bazą szkół (2004-2011 dla szkół podstawowych, 2005-2011 dla gimnazjów) – w bazie szkół z tego samego roku, co wyniki;
- szkoły, dla których z bazy szkół wynika, że zamawiały w danym roku arkusze egzaminacyjne, jednak nie ma ich w zbiorach porównywalnych wyników egzaminacyjnych za ten rok.

Tabela 8.2. Zestawienie częstości występowania rozbieżności w bazie szkół i bazie wyników dla szkół

	Szkoły podstawowe		Gimnazja	
	Lata 2002-2003	Lata 2004-2011	Lata 2002-2004	Lata 2005-2011

[30] Aby było inaczej, powstały w wyniku pomyłki kod pocztowy musiałby leżeć w mieście o takiej samej nazwie – nie jest to niemożliwe, ale jest bardzo mało prawdopodobne.

	(nieobjęte bazą szkół)[31]	(objęte bazą szkół)[32]	(nieobjęte bazą szkół)[33]	(objęte bazą szkół)[34]
brak szkoły w bazie szkół	1 207 szkół (4,45%) 3 621 wyników (0,34%)	360 szkół (0,34%) 14 113 wyników (0,41%)	849 szkół (4,77%) 130 780 wyników (3,96%)	226 szkół (0,40%) 37 308 wyników (0,56%)
brak wyników dla szkoły	n.d. (brak bazy szkół)	3 988 szkół (3,78%)	n.d. (brak bazy szkół)	n.d. (w bazie szkół EWD brak stosownych informacji)

W zależności od rodzaju rozbieżności podjęto różne kroki:

- Szkoły, które w bazie porównywalnych wyników egzaminacyjnych występują jedynie w latach, których nie obejmuje zasięgiem baza szkół.

  Szkół takich nie da się połączyć z bazą szkół. Jedyną możliwość, jaka pozostaje w ich wypadku, to dodanie ich do bazy szkół. Aby było to możliwe, niezbędne jest jednak przypisanie każdej z nich kodu TERYT gminy, w której się znajdują, przy czym w tym wypadku jedynym źródłem informacji o kodzie TERYT pozostaje kod OKE szkoły (patrz punkt 8.4.1). Dla szkół podstawowych na podstawie kodu OKE szkoły nie udało się dopasować kodu TERYT gminy 355 szkołom (29,4% szkół w tej grupie), w wypadku gimnazjów jedynie 2 (0,9%). Szkoły, dla których nie udało się ustalić kodu TERYT musiały zostać usunięte z prezentacji. Szkoły, którym udało się przyporządkować kod TERYT, zostały dodane do bazy szkół i uwzględnione w prezentacji, choć brak danych adresowych, jak również nazwy szkoły w praktyce uniemożliwia ich identyfikację przez użytkownika strony (zostały jednak uwzględnione np. przy obliczaniu danych dla gmin, powiatów i województw).

- Szkoły, które w danym roku występują w bazie wyników zrównanych, jednak nie występują w danym roku w bazie szkół.

[31] Procent liczby szkół względem liczby unikalnych identyfikatorów OKE w zbiorach wyników zrównanych z lat 2002-2003, procent liczby uczniów względem liczby wyników egzaminacyjnych z lat 2002-2003.

[32] Procent liczby szkół względem liczby szkół podstawowych w bazie szkół, procent liczby uczniów względem liczby wyników egzaminacyjnych z lat 2004-2011.

[33] Procent liczby szkół względem liczby unikalnych identyfikatorów OKE w zbiorach wyników zrównanych z lat 2002-2004, procent liczby uczniów względem liczby wyników egzaminacyjnych z lat 2002-2004.

[34] Procent liczby szkół względem liczby gimnazjów w bazie szkół, procent liczby uczniów względem liczby wyników egzaminacyjnych z lat 2004-2012.

Po bliższym przyjrzeniu się rozkładowi tego typu rozbieżności pomiędzy lata oraz okręgowe komisje okazało się, że zdecydowana większość błędów występuje w OKE Kraków (90% w przypadku gimnazjów i 85% w przypadku szkół podstawowych). Z pomocą OKE Kraków udało się wyjaśnić większość spośród rozbieżności odnotowanych w pochodzących z niej danych. Wyniki ze szkół w OKE Kraków, których nie udało się wyjaśnić (30 szkół podstawowych), jak również rozbieżności w pozostałych OKE, 53 szkoły podstawowe na przestrzeni 8 lat (0,05% ogółu szkół podstawowych w tym okresie) oraz 24 gimnazja na przestrzeni 7 lat (0,04% ogółu gimnazjów w tym okresie) zostały arbitralnie uznane za błędy w danych i pominięte w prezentacji.

- Szkoły podstawowe, które zamawiały w danym roku arkusze egzaminacyjne, jednak nie występują w zbiorach porównywalnych wyników egzaminacyjnych.

  Ten typ rozbieżności został przebadany jedynie dla OKE Kraków (przy okazji wyjaśniania różnic opisanych w poprzednim punkcie). Dla ponad 90% rozbieżności, którym przyjrzało się OKE Kraków stwierdzono, że kod OKE szkoły został przesłany do CKE z błędem i podano poprawne kody OKE tych szkół. Paradoksalnie jednak żaden spośród kodów OKE szkół wskazanych przez OKE Kraków jako poprawny nie występował w bazie wyników egzaminacyjnych CKE. Jednoznacznie wskazuje to na niespójność danych pomiędzy OKE i CKE i problem ten rozwiązać mogą jedynie CKE w porozumieniu z OKE. W związku z tym błędy tego typu nie były wyjaśniane.

8.4.3. Łączenie ze sobą tych samych szkół między latami

Aby móc prezentować zmianę porównywalnych wyników egzaminacyjnych w czasie niezbędne było powiązanie wyników poszczególnych szkół między latami. Złączenie zostało wykonane na podstawie identyfikatorów OKE szkół, jednak metoda ta ma pewne ograniczenia. Podstawowym z nich są zmiany kodu OKE szkoły w momencie zajścia zmian w położeniu administracyjnym szkoły (np. wchłonięciu obszaru, na którym znajduje się szkoła do miasta albo przeniesienie gminy z jednego powiatu do drugiego) lub statusu szkoły (np. włączenie do zespołu szkół). Niestety nie istnieją spójne dla wszystkich okręgowych komisji egzaminacyjnych reguły opisujące, kiedy dokonywana jest zmiana kodu OKE szkoły, a kiedy nie. Powoduje to, że szkoły z niektórych rejonów kraju prezentowane są od momentu wystąpienia którejś z opisanych wyżej zmian jako odrębne jednostki, podczas gdy w innych regionach pozostają identyfikowane jako ta sama szkoła. Trudno przy tym powiedzieć, które spośród stosowanych w różnych OKE reguł zmiany/pozostawiania bez zmian kodu OKE szkoły są najbardziej trafne. Bez wątpienia można jednak stwierdzić, że z punktu widzenia prezentacji wyników egzaminów najlepiej by było, by jedne, spójne zasady obowiązywały w całym kraju. W roku 2012 weszła w życie nowelizacja ustawy o Systemie Informacji Oświatowej, która wprowadza definicję szkoły jako bytu trwającego w czasie oraz ustanawia Rejestr Szkół i Placówek Oświatowych, w którym mają być gromadzone z zachowaniem danych historycznych informacje o szkołach, w tym także ich kodach OKE. Pozwala to mieć nadzieję, że dla danych egzaminacyjnych od roku 2013 możliwe będzie łączenie wyników szkół między latami na tych samych zasadach dla całej Polski.

9. Rekomendacje

W kontekście przeprowadzonych badań zrównujących analizy podłużne wyników egzaminu w wieloletnim przedziale czasowym, jako swego rodzaju „monitorowanie egzaminów", są w naszym kraju możliwe i wydają się niezbędne. Istotne jest, aby odróżniać fluktuacje trudności zadań tworzących kolejne arkusze egzaminacyjne od wahań poziomu osiągnięć uczniów i od długofalowych trendów *skuteczności edukacyjnej*. Jakość zrównywania zależy jednak od wielu czynników nie tylko związanych z rzetelnością przeprowadzonych badań zrównujących, ale także od tego, jak dobre są wszystkie ogniwa systemu egzaminacyjnego począwszy od przygotowania arkuszy egzaminacyjnych poprzez przeprowadzenie egzaminu, ocenę prac aż do interpretacji i prezentowania wyników. Polski system egzaminacyjny wymaga modernizacji w zakresie kilku obszarów:

- kontrola jakości arkuszy egzaminacyjnych stosowanych podczas egzaminów
- nowoczesny sposób skalowania
- zrównywanie wyników egzaminacyjnych między latami
- poprawa jakości przetwarzania i przechowywania wyników egzaminacyjnych

9.1. Kontrola jakości arkuszy egzaminacyjnych stosowanych podczas egzaminów

Polski system egzaminacyjny wciąż zostaje w tyle w zakresie procedur i standardów konstrukcji testów egzaminacyjnych zgodnych ze współczesnym w tej dziedzinie stanem wiedzy na świecie. Pewne zmiany są już wprowadzane przez CKE, ale wciąż nie mają one charakteru systemowego. Obecnie dla sprawdzianu standardowe arkusze egzaminacyjne przygotowywane są przez zespół rekrutujący się ze specjalistów z poszczególnych OKE i koordynowany przez OKE we Wrocławiu. Podobnie w przypadku matematyki na poziomie maturalnym arkusze przygotowywane są przez ogólnopolski zespół. W pozostałych egzaminach nadal propozycje arkuszy są przygotowywane przez OKE i wybierane, kompilowane i modyfikowane w CKE. Projekt dotyczący banków zadań współfinansowany z EFS i koordynowany przez CKE ciągle jeszcze nie wszedł w fazę, która umożliwiłaby konstruowanie testów z wykorzystaniem wykalibrowanych zadań z banku, zadań o znanych parametrach psychometrycznych i trafnych do założonych celów pomiaru zdefiniowanych obszarów umiejętności. Podczas analizy zadań z egzaminu gimnazjalnego w części humanistycznej i matematyczno-przyrodniczej w latach 2002-2010 zespół wytypował 66 zadań o bardzo słabych właściwościach psychometrycznych, które musiały być wykluczone z badań zrównujących.

Potrzebne jest wprowadzenie takich procedur, które zagwarantują wysoką jakość zadań egzaminacyjnych i całych arkuszy egzaminacyjnych. Warto skorzystać z doświadczeń innych krajów, gdzie w zespołach odpowiedzialnych za tworzenie narzędzi egzaminacyjnych oprócz ekspertów przedmiotowych istotną rolę odgrywają eksperci z zakresu psychometrii. Potrzebne jest również przyspieszenie prac w zakresie tworzenia profesjonalnych banków zadań, a także systematyczne badania pilotażowe zadań i arkuszy.

Zasadne wydaje się także uwzględnienie w trakcie prac nad arkuszami egzaminacyjnymi problemu precyzji pomiaru, jaką powinien zapewnić arkusz egzaminacyjny. W przypadku sprawdzianu wskaźnik wewnętrznej zgodności testu zawierającego różne formaty zadań (szacowany jako $alfa_{Feldt\text{-}Raju}$) waha się od 0,82 (2003 rok) do 0,86 (w latach 2006, 2007, 2009). Zwiększenie precyzji pomiaru można osiągnąć dwoma sposobami. Po pierwsze, poprzez poprawę jakości zadań, po drugie, poprzez wydłużenie testu, a co za tym idzie wydłużenie czasu trwania sprawdzianu. Ten problem podnoszony był wielokrotnie przez pracowników okręgowych komisji egzaminacyjnych.

9.2. Nowoczesny sposób skalowania

W większości rozwiniętych systemów egzaminacyjnych stosuje się procedury skalowania, które zwiększają precyzję szacowania poziomu umiejętności, umożliwiają przedstawianie wyników na wspólnych („stałych") skalach oraz ułatwiają analizy i interpretację wyników.

Zastosowanie w polskim systemie egzaminów zewnętrznych do prezentacji rezultatów egzaminacyjnych skali standardowej o średniej 100 i odchyleniu 15 znacznie ułatwiłoby porównywanie wyników i uprościło procedury rekrutacyjne do szkół kolejnego szczebla (szkoły ponadgimnazjalne i uczelnie wyższe). Skale te obecne są już w polskim systemie egzaminacyjnym, w projektach EWD (Edukacyjna Wartość Dodana) i OBUT (Ogólnopolskie Badanie Umiejętności Trzecioklasistów).

9.3. Zrównywanie wyników egzaminacyjnych pomiędzy latami

Rezultaty egzaminów komunikowane w postaci obserwowalnych punktów będących sumą rezultatów osiągniętych za wszystkie zadania, czy też punktów procentowych podlegają fluktuacjom w zależności od zmienności poziomu trudności zastosowanych w danym roku arkuszy egzaminacyjnych. Dlatego też niezbędne jest wprowadzenie procedur zrównywania wyników egzaminacyjnych. Warto rozważyć wprowadzenie wewnętrznego kotwiczenia arkuszy egzaminacyjnych stosowanych podczas sesji egzaminacyjnych.

Porównywalne wyniki egzaminacyjne wyrażone w skali roku przyjętego za bazowy (referencyjny) pozwalają na:

- monitorowanie wyników egzaminacyjnych,
- analizę zmian rezultatów egzaminacyjnych w czasie, dla szkół gmin, powiatów województw etc.,
- sprawiedliwy system rekrutacyjny.

9.4. Poprawa jakości przetwarzania i przechowywania wyników egzaminów

Podstawowym problemem napotykanym w momencie, gdy zachodzi potrzeba wykorzystania wyników egzaminów zewnętrznych, jest mnogość i niespójność źródeł danych, z których pochodzą wyniki. Te same dane gromadzone bywają w kilku miejscach (np. zarówno w komisjach egzaminacyjnych, jak i w Centralnej Komisji Egzaminacyjnej), jednak brak jest skutecznych procedur zapewniających ich spójność (np. przekazywania do CKE korekt wyników egzaminacyjnych naniesionych w OKE po okresie wglądów zdających do swoich ocenionych prac). Brak jest wspólnych dla wszystkich OKE

standardów weryfikacji i przechowywania wyników egzaminacyjnych, który pozwalałby na łatwe i rzetelne łączenie wyników egzaminacyjnych z różnych OKE oraz łączenie ich z bazami szkół. Brak jest jednolitych dla wszystkich OKE zasad prowadzenia bazy szkół i nadawania szkołom identyfikatorów egzaminacyjnych (w szczególności zasad regulujących sytuacje, w których identyfikator egzaminacyjny szkoły powinien ulec zmianie, a kiedy pozostać niezmieniony). Brak w końcu jednolitych zasad udostępniania wyników egzaminacyjnych. Niedociągnięcia te powodują, że podstawowe czynności wykonywane na danych egzaminacyjnych, jak zebranie ogólnopolskich wyników wybranego egzaminu i połączenie ich z bazą szkół, napotykają poważne trudności (patrz rozdział 8.4).

Niewątpliwie krokiem w dobrą stronę jest znowelizowana w 2011 r. Ustawa o Systemie Informacji Oświatowej, która:

- wprowadza definicję szkoły jako bytu trwającego w czasie;
- powołuje Rejestr Szkół i Placówek Oświatowych (dalej RSPO), w którym przechowywane są nie tylko aktualne, ale także historyczne informacje adresowe szkół i ich identyfikatorów egzaminacyjnych;
- zapewnia przechowywanie wyników egzaminacyjnych w Systemie Informacji Oświatowej, niestety jedynie zagregowanych.

Na potrzeby analiz wykorzystujących wyniki egzaminacyjne (np. zrównywania wyników egzaminacyjnych, czy obliczania Edukacyjnej Wartości Dodanej) niezwykle przydatny byłyby jednak dalsze kroki:

- udostępnienie dla upoważnionych podmiotów (np. zespołu zrównywania wyników egzaminacyjnych, zespołu Edukacyjnej Wartości Dodanej, zespołu badania OBUT) interfejsu do RSPO umożliwiającego automatyczne pobieranie danych o identyfikatorach egzaminacyjnych szkół (informacje te nie będą bowiem dostępne w portalu SIO);
- zbudowanie centralnej bazy danych wyników egzaminacyjnych na poziomie odpowiedzi udzielonych przez ucznia na poszczególne pytania egzaminu, utrzymywanej i aktualizowanej na zasadach analogicznych do reguł zapisanych w znowelizowanej ustawie o SIO w odniesieniu do zagregowanego wyniku punktowego ucznia.

Zmiany te pozwoliłyby znacznie podnieść efektywność i rzetelność projektów wykorzystujących wyniki egzaminacyjne.

9.5. Perspektywy wdrożenia zrównywania wyników egzaminacyjnych w polskim systemie egzaminów

Opracowana i zastosowana podczas prezentowanych badań metodologia zrównywania może stanowić podstawę do przygotowania projektu do wdrożenia w systemie polskich egzaminów zewnętrznych.

Najlepszym rozwiązaniem jest wprowadzenie oraz zastosowanie planu nierównoważnych grup z testem kotwiczącym podczas sesji egzaminacyjnej.

Należy zbadać, czy takie rozwiązanie wymaga zmian legislacyjnych w egzaminach zewnętrznych prowadzonych przez CKE i komisje egzaminacyjne.

Możliwe jest także włączenie zrównywania do praktyki CKE bez wprowadzenia zadań kotwiczących do egzaminu podczas sesji. Rozwiązanie takie wymaga jednak stosowania dodatkowego studium zrównującego tydzień przed lub/i tydzień po sesji egzaminacyjnej. W najprostszym przypadku można skorzystać z rozwiązań stosowanych w Australii czy w amerykańskim teście ACT z zastosowaniem zewnętrznych testów kotwiczących z równoważnymi (równoważonymi) grupami (ang. e*quivalent groups design)*.

Zrównywanie w pierwszej kolejności powinno być wdrożone dla egzaminów, które zostały dostosowane do nowej podstawy programowej i ich forma oraz zakres sprawdzanych umiejętności mogą być stabilne co najmniej przez kilka lat.

Interpretacja wzrostu czy spadku wyników zrównanych i prawdopodobnego wzrostu lub spadku osiągnięć zdających egzaminy gimnazjalistów i szóstoklasistów wykracza poza problematykę statystycznego zrównywania wyników. Konieczne są dalsze badania, oparte na bardziej szczegółowych analizach programowych i praktyki szkolnej. Szczególnej uwagi wymaga problem nauczania matematyki i przedmiotów przyrodniczych w Polsce. Na podstawie analizy osiągnięć gimnazjalistów z wykorzystaniem wyników zrównanych możemy przypuszczać, że osiągnięcia uczniów w kolejnych latach w części humanistycznej były podobne na przestrzeni 10 lat (2002-2011). Natomiast w części matematyczno-przyrodniczej począwszy od roku 2008 obserwowany jest niewielki trend spadkowy. Obserwacja ta jest szczególnie istotna w kontekście wyników badań międzynarodowych TIMSS na poziomie trzecioklasistów w szkole podstawowej.

10.Bibliografia

1. Allalouf, A. i G. Ben Shakhar (1998). *The effect of coaching on the predictive validity of scholastic aptitude tests*. *Journal of Educational Measurement* 35 (1): 31-47.

2. Balázsi, I . *National Assessment of Basic Competencies in Hungary* dostępny na stronie: http://www.iaea2006.seab.gov.sg/conference/download/papers/National%20assessment%20of%20basic%20competencies%20in%20Hungary.pdf.

3. Béguin, A. A. (2000). *Robustness of equating high-stakes tests* (Doctoral thesis). University of Twente, Enschede.

4. Beller, M. (1994). *Psychometric and social issues in admissions to Israeli universities*. *Educational Measurement: issues and practice* 13 (2): 12-20.

5. Brookhart S.M., (2004). *Grading*, Pearson Merrill Prentice Hall.

6. Bland, J. M., Altman, D. G., (1999) Measurement Agreement in Method Comparison Studies, *Statistical Methods in Medical Research*, 8(2), 135-160.

7. Cohen, J.,(1960) A Coefficient of Agreement for Nominal Scales, *Educational and Psychological Measurement*, 20(1), ss. 37-46.

8. Cook, J. (2009). *An event start: innovative resources to support teachers to better monitor and better support students measured below benchmark*. ACER Research Conference series 3.

9. de la Torre, J. (2009). Improving the Quality of Ability Estimates Through Multidimensional Scoring and Incorporation of Ancillary Variables. *Applied Psychological Measurement* 33.

10. Dolata R. Pokropek A. *Motywacja a wynik testu z nauk przyrodniczych. Studium na przykładzie PISA 2006* [w:] Niemierko B, Szmigel M.K., *Teraźniejszość i przyszłość oceniania szkolnego*, TOMAMI, Toruń 2010, 86-97

11. Domański H., Pokropek A. (2011), *Podziały terytorialne, globalizacja a nierówności społeczne, Wprowadzenie do modeli wielopoziomowych,* Warszawa: Wydawnictwo IFiS PAN

12. Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37(4), 281-306.

13. EQAO (2011). *EQAO's Technical report for 2009 – 2010 Assessments.* Toronto.

14. Ellis, J. L. & van der Woldenberg, A. L. (1993). Local homogeneity in latent trait models. A characterization of the homogenous monotone IRT model. *Psychometrika*, 58 (3), 417–429.

15. Glas C. A. (2010). *Preliminary Manual of the software program Multidimensional Item Response Theory (MIRT)*. (University of Twente)

16. Glas C. A. W. & Béguin A. A. (1996). *Appropriateness of IRT Observed-Score Equating* (Research Report 1996-2)

17. Glas C. A. W. & Béguin A. A. (2011). Robustness of IRT Observed-Score Equating. In von Davier, A. A. (Ed.), *Statistical Models for Test Equating, Scaling, and Linking* (pp. 21-42). New York, NY: Springer-Verlag.

18. Freeman, C. (2009). *First national literacy and numeracy tests introduced*. *Research Developments* 20 (20).

19. Gipps C.V., *Beyond Testing, Towards a theory of educational assessment,* The Falmer Press, London 1995.

20. Hanson, B. A. & Béguin A. A. (1999). *Separate Versus Concurrent Estimation of IRT Item Parameters in the Common Item Equating Design* (ACT Research Report Series, 1999-8). Iowa City, IA: ACT, Inc.

21. Holland, P. W., Dorans N. J., & Petersen N. S. (2007). Equating test scores. In Rao C. R. & Sinharay S. (Eds.). *Handbook of Statistics, Vol. 26. Psychometrics* (pp. 169–204). NY: Elsevier.

22. Kang, T., Petersen N. (2009). *Linking Item Parameters to a Base Scale* (ACT Research Report Series, 2009-2). Iowa City, IA: ACT, Inc.

23. Kolen, M. J. (1984). *Effectiveness of analytic smoothing in equipercentile equating. Journal of Educational Stati*). *The Swedish Scholastic Assessment Test (SweSAT)*. Department of Educational Measurement, Ume Univ.*stics* 9, 25–44.

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

25. Li, Y. & Lissitz, R. W. (2000) An Evaluation of the Accuracy of Multidimensional IRT Linking. *Applied Psychological Measurement*, 24(2), 115-138.

26. Liu, J. & Walker M. E. (2007). Score Linking Issues Related to Test Content Changes. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109–134). New York, NY: Springer-Verlag.

27. Livingston S.A. *Equating test stores* http://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf.

28. Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

29. Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley.

30. Madous G. (1988). *The influence of testing on the curriculum* [in] Tanner L.N. (ed) *Critical Issues in Curriculum (National Society for the Study of Education Yearbooks (Pt. 1)* , Chicago.

31. Niemierko B. (2002). *Ocenianie szkolne bez tajemnic,* WSiP, Warszawa.

32. Pawłowska, B. (2008). *Teorie motywacji*, http://www.soc-org.edu.pl/PL/emp_Pawlowska/res/-proces_motywacji.pdf.

33. Patz R. J. & 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.

34. PISA 2003 – Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2003 w POLSCE .

35. PISA 2006 – Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2006 w POLSCE.

36. PISA 2009 – Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2009 w POLSCE.

37. Pokropek A. (2011) *Zrównywanie wyników egzaminów zewnętrznych w kontekście międzynarodowym* [w:] Niemierko B., Szmigel M.K. (red.) *Ewaluacja w edukacji: koncepcje metody, perspektywy*, PTDE, Kraków.

38. Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer-Verlag.

39. Standardy 2001 – ROZPORZĄDZENIE MINISTRA EDUKACJI NARODOWEJ z dnia 10 sierpnia 2001 r. w sprawie standardów wymagań będących podstawą przeprowadzania sprawdzianów i egzaminów. (Dz. U. z 2001 r. Nr 92, poz. 1020) [Załącznik nr 2].

40. Standardy 2007 – ROZPORZĄDZENIE MINISTRA EDUKACJI NARODOWEJ z dnia 28 sierpnia 2007 r. zmieniające rozporządzenie w sprawie standardów wymagań będących podstawą przeprowadzania sprawdzianów i egzaminów. (DZ.U. z dnia 31 sierpnia 2007 r. Nr 157, poz. 1102) [Załącznik].

41. Rao, C. R. i S. Sinharay. (2007). *Psychometrics.* 26-ed. North Holland.

42. Stage, C. (2004). *Notes from the Tenth International SweSAT Conference*. Umeå, June 1–3, 2004.Stage, C i G. Ígren (2002). *The Swedish Scholastic Assessment Test (SweSAT)*. Department of Educational Measurement, Ume Univ.

43. Szaleniec H., Grudniewska M., Kondratek B., Kulon F., Pokropek A, (2011). *Analiza porównawcza wyników egzaminów zewnętrznych – Gimnazjum*. Raport z badań. Instytut Badań Edukacyjnych, Warszawa.

44. Tyralska-Wojtycza E. (2010). *Nowa formuła egzaminu gimnazjalnego – strata czy zysk dla przedmiotów przyrodniczych*, [w:] Niemierko B., Szmigel M.K., (red) *Teraźniejszość i przyszłość oceniania szkolnego*, PTDE.

45. von Davier, A. A., (2011). A Statistical Perspective on Equating Test Scores. In von Davier, A. A. (Ed.), *Statistical Models for Test Equating, Scaling, and Linking* (pp. 1-17). New York, NY: Springer-Verlag.

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

47. von Davier, M., von Davier, A. A. (2011). A General Model for IRT Scale Linking and Scale Transformations. In von Davier, A. A. (Ed.), *Statistical Models for Test Equating, Scaling, and Linking* (pp. 1-17). New York, NY: Springer-Verlag.

48. van der Linden, W. J., (2011). Local Observed-Score Equating. In von Davier, A. A. (Ed.), *Statistical Models for Test Equating, Scaling, and Linking* (pp. 201-223). New York, NY: Springer-Verlag.

49. Węziak D., (2007). *Metody zrównywania wyników wykorzystywane w skalowaniu Rascha. Propozycja zastosowań w warunkach polskich. „Egamin* [w:] *Biuletyn Badawczy Centralnej Komisji Edukacyjnej*" 10/2007, s. 76-77.

50. Wu. M. (2005). The Role of Plausible Values in Large-Scale Surveys. Elsevier: *Studies in Educational Evaluation*. 31, s. 114-128.

51. Yao, L. & Boughton K. (2009). Multidimensional Linking for Tests with Mixed Item Types. *Journal of Educational Measurement*, 46(2) 177–197.

11. Aneksy

Aneks A – Psychometryczne właściwości zadań egzaminacyjnych

A1. Egzamin gimnazjalny

A2. Sprawdzian po szkole podstawowej

Aneks B – Średnie wyniki egzaminacyjne uczniów w podziale na jednostki samorządu terytorialnego

B1. podział na województwa

B2. podział na powiaty

B3. podział na powiaty w ramach województw (na jednej mapie zawarto obszar jednego województwa)