Egzaminy zewnętrzne w polityce i praktyce edukacyjnej

Raport o stanie edukacji 2014

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

MINISTERSTWO EDUKACJI NARODOWEJ

IBE entuzjaści edukacji

UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

MINISTERSTWO EDUKACJI NARODOWEJ

IBE entuzjaści edukacji

UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY

Egzaminy zewnętrzne w polityce i praktyce edukacyjnej

Raport o stanie edukacji 2014

Warszawa 2015

Redakcja merytoryczna:
dr hab. Roman Dolata
dr Michał Sitek

Redakcja serii:
dr hab. Michał Federowicz
dr Michał Sitek

Recenzenci:
prof. dr hab. Jarosław Górniak
prof. dr hab. Bolesław Niemierko

Autorzy:
dr hab. Roman Dolata
Anna Hawrot
Grzegorz Humenny
Aleksandra Jasińska-Maciążek
Bartosz Kondratek
Maciej Koniewski
Filip Kulon
Przemysław Majkut
Katarzyna Matuszczak
dr Artur Pokropek
Anna Rappe
dr Michał Sitek
Paulina Skórska
dr Ewa Stożek
Karolina Świst
dr Henryk Szaleniec
Olga Wasilewska
Tymoteusz Wołodźko
Mateusz Żółtak
Tomasz Żółtak

Opracowanie językowe i korekta:
Beata Dąbrowska

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

ISBN- 978-83-65115-55-3

Wzór cytowania:
Dolata, R. i Sitek M. (red.) (2015). *Raport o stanie edukacji 2014. Egzaminy zewnętrzne w polityce i praktyce edukacyjnej.* Warszawa: Instytut Badań Edukacyjnych.

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

Egzemplarz bezpłatny

Publikacja została wydrukowana na papierze ekologicznym

Spis treści

Wprowadzenie

W dyskusjach o edukacji coraz częściej mówi się o jakości i coraz ważniejszą rolę odgrywają różne sposoby jej mierzenia. Jeszcze kilkadziesiąt lat temu porównania systemów oświatowych ograniczały się do zestawiania odsetków absolwentów kończących poszczególne etapy edukacji czy wskaźników upowszechnienia edukacji, np. w przedszkolach czy w szkolnictwie wyższym. W nadzorze nad szkołami dominował model tradycyjnej, hierarchicznej kontroli, polegającej na sprawdzaniu, czy przestrzegane są przepisy i inne zewnętrznie zdefiniowane normy. W ostatnich dekadach zwiększyły się oczekiwania stawiane przed edukacją. W większym stopniu zaczęto też postrzegać edukację przez pryzmat poziomu wiedzy i umiejętności absolwentów. W nadzorze nad szkołami zaczęto zwracać uwagę nie tylko na procesy, ale też i efekty, w tym zwłaszcza efekty uczenia się. Międzynarodowe badania osiągnięć edukacyjnych pokazały problem zróżnicowania poziomu umiejętności uczniów, w tym istnienia sporej grupy uczniów, którzy kończą szkoły z zasobem umiejętności nieodpowiadającym wyzwaniom współczesnego społeczeństwa i potrzebom rynku pracy.

W odpowiedzi na te wyzwania systemy edukacyjne stały się bardziej złożone; w wielu krajach zwiększono rolę samorządów terytorialnych oraz autonomię szkół i swobodę w dobieraniu treści i narzędzi dydaktycznych przez nauczycieli. Rozbudowano lub utworzono nowe instytucje odpowiedzialne za egzaminy i nadzór nad jakością. Stworzono nowe narzędzia mierzenia jakości. Coraz większą rolę w mierzeniu jakości odgrywają egzaminy i inne sposoby sprawdzania umiejętności uczniów oraz wyliczane na ich podstawie wskaźniki, takie jak wskaźniki edukacyjnej wartości dodanej. Egzaminy nie wypełniają już wyłącznie swoich tradycyjnych funkcji związanych z nadawaniem kwalifikacji i selekcją do kolejnych szczebli kształcenia, ale też są wykorzystywane do innych celów, w tym zwłaszcza do oceny jakości pracy szkół i doskonalenia nauczania oraz monitorowania działania systemu edukacji.

W Polsce wprowadzenie egzaminów zewnętrznych ma istotne skutki dla całego systemu edukacji. Od samego początku egzaminy i inne formy standaryzowanych sposobów mierzenia osiągnięć uczniów miały zarówno zwolenników, jak i zagorzałych wrogów. Pierwsi uważają, że wyniki egzaminów są najlepszym i jedynym obiektywnym źródłem danych, który można wykorzystać do oceny jakości edukacji. Drudzy widzą w egzaminach przede wszystkim zagrożenie, podkreślając, że egzaminy niszczą istotę edukacji, instrumentalizując i zawężając nauczanie w szkołach. W ich przekonaniu testy niszczą kreatywność, wzmacniają niepotrzebną rywalizację uczniów i szkół, uniemożliwiając osiąganie szerokiego zakresu umiejętności i wartości, które powinna kształcić szkoła. Wprowadzenie egzaminów zewnętrznych spowodowało, że testy i ich miejsce w nauczaniu przestały być wyłącznie tematem dyskusji akademickich i sporów pedagogów i stały się tematem obecnym w mediach. Coroczne wyniki egzaminacyjne, w tym zwłaszcza wyniki maturalne, stały się tematem doniesień prasowych i sporów o to, czy jakość edukacji w kraju lub konkretnej szkole pogarsza się, czy poprawia, a dane z egzaminów umożliwiły produkowanie różnorodnych zestawień i rankingów. Łatwa dostępność wyników egzaminacyjnych rodzi też pokusę wykorzystywania ich do oceny jakości pracy szkół, pracy dyrektorów i nauczycieli i uwzględniania tych informacji w podejmowaniu decyzji finansowych czy personalnych. Po wyniki egzaminacyjne coraz częściej sięgają rodzice i uczniowie zastanawiający się nad wyborem szkoły, ale też władze samorządowe decydujące o kształcie sieci szkół i poziomie wydatków publicznych na oświatę.

Egzaminy są bez wątpienia jednym z najważniejszych elementów polskiego systemu edukacji. Są one jednym z kluczowych, obok podstaw programowych i systemu nadzoru pedagogicznego, narzędziem koordynacji, które stwarza bodźce, jakimi w mniej lub bardziej świadomy sposób kierują się uczniowie i rodzice, wybierając szkoły, samorządy oceniające jakość edukacji w prowadzonych przez siebie szkołach, dyrektorzy szkół oceniający pracę nadzorowanych przez siebie nauczycieli, czy też sami nauczyciele, którzy zastanawiają się, na jaki zakres treści położyć większy nacisk na

lekcjach. Można dyskutować, czy rola egzaminów w Polsce nie jest za duża, czy nie można systemu egzaminacyjnego zorganizować w lepszy sposób i wreszcie czy w sposób właściwy wykorzystujemy wyniki egzaminacyjne. Tym tematom postanowiliśmy poświęcić tegoroczny *Raport o stanie Edukacji*. Tym co odróżnia konstrukcję tegorocznego raportu jest brak części stałej, omawiającej dane statystyczne opisujące stan systemu edukacji i kształtujące się trendy. W czasie pisania tego raportów zdecydowana większość ustaleń przedstawionych w części raportu z 2014 r. zatytułowanej *Edukacja w liczbach* jest wciąż aktualna bądź zmieniła się na tyle nieznacznie, że opisywanie trendów dotyczących uczestnictwa w edukacji, zmian w strukturze wykształcenia czy finansów edukacji byłaby powtórzeniem. Niezmienna pozostaje natomiast zasada, zapoczątkowana w poprzednich dwóch raportach, zgodnie z którą raport zawiera pogłębione i syntetyczne ujęcie konkretnej, kluczowej z punktu widzenia systemu edukacji kwestii. W raporcie z 2012 r. były to efekty uczenia się, w ostatnim raporcie o stanie edukacji, z 2013 r. , byli to nauczyciele. W tegorocznym raporcie przyglądamy się egzaminom z różnorodnych perspektyw, starając się porównać miejsce egzaminów w Polsce z ich rolą w innych krajach, przypominając i porządkując kwestię jakości egzaminów, podsumowując doświadczenia związane z rozwijaniem analiz egzaminacyjnych oraz przypatrując się, w jaki sposób egzaminy są wykorzystywane w praktyce ewaluacji zewnętrznej i wewnętrznej.

Polskie dyskusje dotyczące roli testów i egzaminów zewnętrznych i stawiane w nich argumenty są bardzo podobne do dyskusji prowadzonych w innych krajach. Oczywiście doświadczenia innych krajów są osadzone w różnych kontekstach prawno-organizacyjnych i społeczno-kulturowych. Mimo to sądzimy, że przyjrzenie się zagranicznym doświadczeniom jest przydatne i potrzebne. Pozwala to nie tylko lepiej zrozumieć wady i zalety stosowania testów i egzaminów zewnętrznych, ale daje też szansę uczenia się na doświadczeniach innych. Z tego względu raport zaczynamy od krótkiego przeglądu doświadczeń międzynarodowych, zwracając uwagę przede wszystkim na dwa aspekty – sposób organizacji funkcji ewaluacyjnej egzaminów w różnych krajach i pozytywne oraz negatywne konsekwencje egzaminów, widziane w kontekście konkretnych rozwiązań instytucjonalnych. W raporcie chcemy się jednak przede wszystkim skupić na jakości egzaminów i sposobach ich wykorzystania w Polsce. Kluczowe znaczenie ma kwestia jakości testów egzaminacyjnych. Ważną częścią problemu dyskusji nad testami i egzaminami jest dość powszechny brak zrozumienia, w jaki sposób testy powstają, co oznaczają uzyskiwane wyniki i jakie są ich ograniczenia. Testy są jedynie narzędziem, które można wykorzystać do określonego celu. Tak jak inne narzędzia, testy mogą być wykorzystywane w pożyteczny sposób – można twierdzić, że są wręcz niezastąpione. Ale testy mogą też być wykorzystywane w sposób nieuprawniony, a nawet w sposób, który przynosi więcej szkody niż pożytku. Aby umiejętnie wykorzystywać wyniki egzaminów, trzeba dobrze znać podstawy metodologii tworzenia i wykorzystywania testów. Wiele problemów związanych z testami i ich wykorzystaniem ma skomplikowany, techniczny charakter. Wiedza o testach i sposobach tworzenia testów rozwija się od przeszło 100 lat i przekształciła się w odrębną dziedzinę wiedzy: psychometrię. Obecnie jesteśmy w stanie ocenić dość dobrze jakość samych testów i współczesna psychometria dysponuje rozbudowaną metodologią i standardami ich oceny. Kwestie te omawia drugi rozdział raportu, w którym w sposób syntetyczny opisujemy podstawowe zasady tworzenia testów i komunikowania ich wyników, podkreślając najnowsze standardy zapewniania jakości tego procesu.

W kolejnym rozdziale poruszamy kluczową, z punktu widzenia wykorzystania egzaminów jako źródła informacji o jakości edukacji, kwestię porównywalności wyników egzaminacyjnych w czasie. W Polsce problem ten nie znalazł jeszcze systemowego rozwiązania. W pewnym stopniu możliwość prowadzenia tego typu porównań daje przeprowadzony przez Instytut Badań Edukacyjnych kilkuletni projekt *Porównywalne Wyniki Egzaminacyjne*. W rozdziale 3 podsumowujemy nasze doświadczenia z tego projektu i jego główne wyniki.

Wykorzystywanie testów do innych celów niż pomiar indywidualnych cech, takich jak zdolności czy osiągnięcia, np. do oceny pracy szkół czy nauczycieli jest złożonym zagadnieniem. Polska należy do jednych z nielicznych krajów, w których rozwinięto metodologię edukacyjnej wartości dodanej. Jej zaletą jest odejście od prostego porównywania wyników egzaminacyjnych. Metodologię mierzenia

EWD w Polsce opisujemy w rozdziale 4. Podkreślamy w niej możliwości, jakie dają wskaźniki EWD w pracy szkoły i w doskonaleniu jakości nauczania.

Coraz większa dostępność danych i wskaźników otwiera nowe możliwości ich wykorzystywania na poziomie szkoły. Problemem, który zauważany jest nie tylko w Polsce, jest słabe wykorzystanie danych i informacji – niewspółmierne do zwiększającej się liczby danych i możliwości ich analizowania. Lepsze wykorzystanie danych wymaga posiadania wiedzy i umiejętności, a najlepsze efekty przynosi wtedy, gdy dane egzaminacyjne łączy się z innymi informacjami dostępnymi na poziomie szkoły. W końcowym rozdziale skupiamy się na mocnych i słabych stronach wykorzystania danych egzaminacyjnych w ramach ewaluacji wewnątrzszkolnej.

1. Miejsce egzaminów zewnętrznych w systemach edukacyjnych i polityce edukacyjnej

Michał Sitek

1.1. Dlaczego egzaminy stały się tak ważne?

Zwiększenie roli egzaminów i innych form mierzenia osiągnięć uczniów to jeden z najważniejszych trendów w zmianach systemów edukacyjnych na całym świecie. Przyczyn tego zjawiska można upatrywać w wielu uniwersalnych procesach. Najważniejsze z nich to wzrost znaczenia efektywności, równości i jakości w dyskursie dotyczącym edukacji oraz związany z nim kryzys tradycyjnego biurokratyczno-profesjonalnego modelu zarządzania edukacją, czyli takiego modelu, w którym kluczową rolę w zarządzaniu odgrywają urzędnicy i profesjonaliści: nauczyciele, organizacje nauczycielskie oraz eksperci oświatowi.

W większości krajów szkoły były historycznie częścią hierarchicznie zorganizowanych organizacji i miały niewielką swobodę decyzji dotyczących zarządzania nauczycielami czy kierunkiem wydatków. Egzaminy były wykorzystywane w rekrutacji, służyły też do zaświadczania ukończenia nauki, ale ich wyniki nie były gromadzone i analizowane. Często, jak to miało miejsce w Polsce lat 90, uczniowie z różnych województw rozwiązywali różne zadania maturalne, nie było też instrumentów zapewniania porównywalności oceniania prac egzaminacyjnych przez nauczycieli. Danych egzaminacyjnych nie wykorzystywano też do oceny jakości szkół, wspomagania szkół czy kształtowania polityki edukacyjnej. Nawet na poziomie szkoły w praktyce nie gromadzono i nie przetwarzano odpowiedzi uczniów na poszczególne pytania i nie analizowano słabych i mocnych stron ich umiejętności. Model biurokratyczny zarządzania edukacją był obecny nie tylko w systemach gospodarki nakazowo-rozdzielczej, ale dominował też w wielu liberalnych demokracjach. Z biegiem czasu coraz bardziej oczywiste było, że system ten słabo się sprawdza. Odpowiedzią było zwiększanie autonomii szkół i nauczycieli, decentralizacja zarządzania i finansowania edukacji i wprowadzanie elementów konkurencji między szkołami. Wierzono, że zwiększenie autonomii pozwala lepiej zaspokajać lokalne potrzeby i poprawiać efekty nauczania, bo to władze samorządowe, nauczyciele i dyrektorzy szkół wiedzą najlepiej, jakie są lokalne potrzeby i jakie działania są najbardziej efektywne w określonych warunkach. Odchodzi się więc od wpływania przez państwo na programy nauczania i sposoby nauczania, co pozwala skupić uwagę na bardziej precyzyjnym definiowaniu oczekiwanych efektów i ich kontroli. Decentralizacja daje też lepsze możliwości rozliczania szkół z działalności przez rodziców i lokalne społeczności.

Decentralizacja i zwiększająca się autonomia szkół stwarza jednak zapotrzebowanie na miękkie narzędzia koordynowania polityki i praktyki edukacyjnej oraz nowe formy ewaluacji działania szkół. Potrzebę zewnętrznej oceny jakości i monitorowania efektów doskonale wypełniają właśnie egzaminy. Sprzyjał temu także rozwój technologii tworzenia testów i coraz doskonalsze możliwości gromadzenia, przetwarzania i udostępnienia danych wykorzystujące rozwój technologii informacyjno-komunikacyjnych. Egzaminy doskonale wpisywały się też w nowe sposoby myślenia o poprawie jakości sektora publicznego rozwijane w latach 80. Koncepcja nowego zarządzania publicznego (NPM), która stała się teoretycznym uzasadnieniem wprowadzenia wielu ówczesnych zmian w sektorze publicznym, kładzie nacisk na potrzebę lepszej oceny efektów i efektywności działań organizacji sektora publicznego, w miejsce tradycyjnego budżetowania i zarządzania przez procedury. Postuluje ona jasne określenie celów i sposobów ich mierzenia, decentralizację zarządzania lub podział dużych organizacji na mniejsze oraz wiązanie oceny wykonywania zadań z wynagrodzeniem. Próbowano także włączyć w funkcjonowanie sektora publicznego elementy konkurencji,

możliwe dzięki zwiększeniu autonomii instytucji publicznych, oraz nowych idei nauk o zarządzaniu stosowanych w sektorze prywatnym. W oświacie idee te znalazły odzwierciedlenie w dwóch ideach: próbach zwiększenia możliwości wyboru szkoły i wzmocnieniu roli testów. Wierzono, że zwiększenie roli testów przyczyni się do przekształcania tradycyjnego modelu zarządzania oświatą, polegającego na administrowaniu szkołami, w model bardziej zorientowany na poprawę osiągnięć uczniów. Upowszechnienie testów służyło poprawie dostępności informacji o jakości szkół dla rodziców, co w sytuacji swobodnego wyboru szkoły miało zwiększać motywację szkół do osiągania lepszych wyników. Zmiany w sposobie funkcjonowania sektora publicznego wprowadzano pod koniec lat 70. I w latach 80, przede wszystkim w krajach anglosaskich: Wielkiej Brytanii, Stanach, Zjednoczonych i Nowej Zelandii, ale podobne idee obecnie były podstawą wprowadzania zmian w wielu krajach europejskich.

Efektami tych procesów były wzmocnienie i rozbudowa funkcji egzaminów, które przestały służyć wyłącznie do celów rekrutacyjnych czy potwierdzania indywidualnych kwalifikacji, ale stanowiły źródło informacji, które można wykorzystać do celów ewaluacyjnych. Zaobserwować można także też coraz szersze wykorzystywanie różnego rodzaju wskaźników i rozwój narzędzi zewnętrznej ewaluacji. Zjawiska te nie ograniczają się jedynie do oświaty, ale są widoczne w różnych obszarach życia społecznego. Drugą stroną tego procesu było tworzenie nowych instytucji, często niezależnych od tych demokratycznie wybieralnych, których zadaniem jest regulowanie i kontrolowanie podmiotów działających w określonym obszarze. Warto zauważyć, że ten trend nie dotyczy jedynie egzaminów, ale też innych form oceny działania szkół, np. inspekcji szkolnych. Rośnie znaczenie wskaźników, głównie opierających się na wynikach uczniów, ale coraz częściej wykorzystujących także inne źródła informacji. Pozytywnym aspektem tego procesu jest zwiększenie profesjonalizacji wykonywania funkcji ewaluacyjnej, a także widoczna w ostatnich latach, także i w Polsce, dywersyfikacja sposobów mierzenia jakości edukacji, co pozwala na bardziej wszechstronny ogląd systemu edukacyjnego. Przykładem tego jest coraz częstsze wykorzystywanie w ewaluacji szkół technik badań ilościowych, takich jak sondaże prowadzone wśród uczniów, rodziców i innych interesariuszy, czy technik jakościowych stosowanych w naukach społecznych i nawiązujących do tradycji badań ewaluacyjnych. Zaobserwować można także proces dookreślania standardów i wymagań, które są podstawą programów nauczania i standardów egzaminacyjnych oraz wymagań wykorzystywanych przez inspekcje szkolne. Wzrost znaczenia egzaminów ściśle wiąże się z coraz powszechniejszym myśleniem o edukacji w kategoriach efektów uczenia się i definiowania w tej formie celów i wymagań edukacyjnych. Głównym wnioskiem badań pedagogicznych ostatnich dziesięcioleci jest wykazanie, że efektywność nauczania w dużej mierze zależy od nauczycieli i praktyki nauczania. Dlatego w ostatnich latach nowym zjawiskiem są też próby wprowadzania zmian w systemach oceniania nauczycieli. Wiele wysiłków reformatorskich na świecie skupia się właśnie na poprawie efektywności pracy nauczycieli czy oceny pracy nauczycieli, do czego coraz częściej próbuje się też wykorzystywać dane o osiągnięciach uczniów.

Od początku ważnym uzasadnieniem wprowadzania i upowszechniania testów i egzaminów było przekonanie, że zwiększają one przejrzystość w edukacji i służą zwiększaniu równości i merytokratycznym decyzjom, w którym decydującą rolę mają umiejętności uczniów, a mniejszą inne cechy związane np. z płcią, pochodzeniem czy statusem społeczno-ekonomicznym. Ocenianie zewnętrzne, czyli przeprowadzane przez zewnętrzny podmiot w możliwie wystandaryzowany sposób, było postrzegane jako bardziej obiektywne i bezstronne od innych sposobów oceny osiągnięć, takich np. jak oceny wystawiane przez nauczycieli czy egzaminy wstępne, co miało znaczenie przede wszystkim w procesach selekcyjnych.

1.2. Funkcje egzaminów

Zmiany w sposobie postrzegania egzaminów i nowe nadzieje z nimi związane sprawiły, że zmienia się przypisywana im funkcja. W literaturze znaleźć można różne typologie tych funkcji. Tu zwracamy uwagę na cztery z nich.

Podstawową, i historycznie pierwotną, funkcją egzaminów jest funkcja selekcyjna, w której wynik egzaminu warunkuje dostęp do ścieżki edukacyjnej lub wyższego etapu kształcenia. Za pierwsze egzaminy uważa się procedury wykorzystywane w Chinach w II wieku p.n.e. w procesach rekrutacji urzędników. Egzaminy obejmowały umiejętności w zakresie muzyki, łucznictwa i jeździectwa. Przeprowadzano także pisemne egzaminy ze znajomości prawa, wiedzy rolniczej i geografii (Bowman, 1989). Szersze wykorzystanie testów w edukacji wiąże się z rozwojem badań nad inteligencją na początku XX wieku. Testy inteligencji tworzone przez psychologów stały się wzorem do testów osiągnięć wykorzystywanych w armii amerykańskiej w czasie I i II wojny światowej do oceny zdolności rekrutów i promocji na wyższe stanowiska. Zaprojektowany do tego celu test psychologiczny, składający się z pytań wielokrotnego wyboru i mający mierzyć ogólne zdolności poznawcze, był pierwowzorem testu SAT (*Scholastic Aptitude Test*), który, począwszy od lat 50. Coraz powszechniej jest stosowany w rekrutacji na uczelnie. Funkcja selekcyjna w wielu krajach odgrywa wciąż dominującą rolę. Egzaminy zewnętrzne są elementem systemów rekrutacyjnych na wyższe uczelnie w wielu krajach, będąc w większym (jak jest np. w Polsce) lub mniejszym stopniu (Stany Zjednoczone) powiązanymi z programami kształcenia w szkolnictwie. Coraz częściej są też wykorzystywane do rekrutacji do szkół średnich, zastępując lub uzupełniając tradycyjne egzaminy wstępne.

Z funkcją selekcyjną wiąże się funkcja certyfikacyjna: potwierdzania kompetencji lub kwalifikacji. W niektórych krajach egzaminy uprawniają do otrzymania świadectwa czy promocji uczniów do kolejnej klasy. Dobrym przykładem są tu polskie egzaminy potwierdzające kwalifikacje zawodowe. Egzamin zewnętrzny stanowi wówczas zewnętrzne potwierdzenie, że uczeń posiada określoną wiedzę i umiejętności. W ostatnich latach ta funkcja zyskuje na znaczeniu ze względu na coraz większą potrzebę uczenia się dorosłych i potrzebę potwierdzania kompetencji zdobytych w edukacji pozaformalnej.

Trzecią ważną funkcją egzaminów jest funkcja kontrolno-ewaluacyjna. Egzaminy są źródłem obiektywnej, niezależnej od szkół informacji o osiągnięciach szkolnych, które mogą być w różny sposób wykorzystane do innych celów niż sprawdzenie umiejętności konkretnego ucznia. Wyniki mogą być wykorzystywane przez szkoły do doskonalenia praktyki nauczania, nadzór pedagogiczny do oceny jakości nauczania w konkretnej szkole, organy prowadzące do oceny jakości nauczania w szkołach czy przez ministerstwo do oceny stopnia, w jakim system edukacji osiąga oczekiwane rezultaty. Wzrost znaczenia egzaminów doprowadził do znacznego przeobrażenia tej funkcji. W wielu krajach egzaminy stały się ważnym instrumentem rozliczania szkół z efektów ich działalności. Ale widoczny jest także inny trend: coraz częstsze wykorzystywanie testów i danych egzaminacyjnych do poprawy jakości edukacji. Napięcie między funkcją kontrolną i ewaluacyjną jest kluczowym elementem dyskusji o roli testowania w edukacji.

Obok funkcji selekcyjnej i ewaluacyjnej warto zwrócić uwagę na odrębną i rzadko wspominaną funkcję, którą w mniej lub bardziej otwarty spełniają egzaminy. Egzaminy motywują uczniów, nauczycieli i szkoły do osiągania pożądanych efektów. Egzaminy spełniają w ten sposób ważną funkcję regulacyjną. Egzaminy sygnalizują, które umiejętności są ważne i na które należy położyć szczególny nacisk w nauczaniu. Za pomocą testów zaczęto nie tylko weryfikować umiejętności, ale też kształtować to, czego uczy się w szkole i co uznaje się za niezbędne minimum. Nie chodzi tu wyłącznie o zwiększenie motywacji uczniów i nauczycieli. Tworząc system egzaminacyjny, decydenci muszą zdecydować się, które przedmioty i jakie obszary wiedzy i umiejętności powinny być testowane, a które z nich powinny być obowiązkowe. Wybór tych przedmiotów jest informacją, że są one uznawane za szczególnie ważne i na nich powinna skupiać się uwaga szkół, nauczycieli i uczniów. Funkcję tę egzaminy spełniają też w bardziej subtelny sposób. Na przykład, nowe formy zadań wprowadzone

w 2012 r. w polskim egzaminie gimnazjalnym podkreślają znaczenie kształtowania umiejętności rozumowania i argumentacji, co w założeniu miało wzmocnić wdrażanie zapisów nowej podstawy programowej kształcenia ogólnego. Regulacyjna funkcja jest też ważną funkcją z dydaktycznego punktu widzenia i wiąże się z samoregulacją uczenia się dzieci i młodzieży.

1.3. Rola egzaminów i mierzenia osiągnięć uczniów w perspektywie porównawczej

Egzaminy lub diagnozy pozwalające na wnioskowanie o umiejętnościach uczniów są wykorzystywane w niemal wszystkich krajach Unii Europejskiej. Do niedawna wyjątkiem były trzy kraje: Czechy, Hiszpania i Grecja. Ale w Hiszpanii w ostatnich latach Ministerstwo Edukacji uruchomiło ogólnokrajowe badania diagnostyczne na poziomie szkoły podstawowej i średniej. W chwili pisania tego tekstu, w Czechach trwa pilotażowe wdrożenie zewnętrznego egzaminu maturalnego. Także w Grecji podejmowano w ostatnich latach próby stworzenia centralnej komisji egzaminacyjnej odpowiedzialnej za przygotowywanie części pytań używanych w egzaminach szkolnych.

W zdecydowanej większości krajów Unii Europejskiej najpowszechniejszym, a w niektórych krajach jedynym egzaminem jest egzamin na zakończenie szkoły średniej II stopnia podobny do polskiej matury. Kraje różnią się zakresem treści, jakie są objęte egzaminami, i tym, czy są one obowiązkowe, czy dobrowolne, a także ich znaczeniem w procesie rekrutacji na studia. Przy czym szczegółowe rozwiązania dość często się zmieniają.

Bardziej zróżnicowana sytuacja występuje na wcześniejszym etapie kształcenia, odpowiadającym polskiemu gimnazjum. W wielu krajach prowadzi się diagnozy (ang. *assessments*), ale w części krajów przeprowadza się egzaminy zbliżone do polskiego egzaminu gimnazjalnego, którego wynik jest wykorzystywany w celach rekrutacyjnych do szkoły średniej. Zazwyczaj obejmują one główne przedmioty szkolne, takie jak matematyka, język ojczysty, przyroda. W niektórych systemach na tym etapie daje się uczniowi możliwość wyboru dodatkowego przedmiotu. W niektórych systemach (np. w Danii czy Norwegii) część przedmiotów jest obowiązkowa dla wszystkich uczniów, ale umiejętności z innych, dodatkowych przedmiotów bada się w losowej próbie szkół.

Na najniższych szczeblach kształcenia egzaminy, których wynik ma duże znaczenie dla kariery edukacyjnej ucznia, należą do rzadkości. W wielu krajach w pierwszych etapach kształcenia dominują dobrowolne diagnozy. Są one prowadzone wśród wszystkich bądź prawie wszystkich uczniów (w formule podobnej do polskiego sprawdzianu po szkole podstawowej), ale w części krajów są przeprowadzane na losowej próbie szkół (np. w Estonii, Finlandii, Niemczech, Irlandii czy Holandii). Tego rodzaju formy standaryzowanego pomiaru osiągnięć, niewiążące się ze znaczącymi konsekwencjami dla ucznia, są znacznie lepiej dostosowane do monitorowania zmian osiągnięć, a ich nieobowiązkowy charakter ułatwia ich wykorzystanie do analizowania uwarunkowania osiągnięć, np. poprzez zbieranie informacji kontekstowych lub odpowiednie zaprojektowanie testu.

Jak widać, nieostre jest samo pojęcie egzaminu. Zaciera się zwłaszcza granica między egzaminem, a różnego rodzaju diagnozami. Jest to naturalne w świetle dydaktycznego (nieadministracyjnego) pojmowania egzaminu, definiowanego jako „każde sprawdzanie i ocenianie osiągnięć uczniów wyodrębnione w procesie kształcenia" (Niemierko, 2009, s. 237). Egzaminy mogą być wykorzystywane do różnych celów, mogą przybierać różną postać i obejmować różny zakres przedmiotów. Warto więc zwrócić uwagę na przyjęte w literaturze rodzaje testów egzaminacyjnych. Ze względu na sposób konstrukcji i interpretacji wyników wyróżnia się pomiar różnicujący (odniesiony do norm ilościowych, ang. *norm-referenced*), który pozwala wypowiadać się jedynie o wyniku na tle określonej grupy odniesienia. Wynik takiego pomiaru jest wskaźnikiem względnym – punktem odniesienia jest analizowana grupa. Tego rodzaju testami są sprawdzian po szkole podstawowej i egzamin gimnazjalny. Innym rodzajem jest pomiar sprawdzający (odniesiony do kryterium, ang. *criterion-referenced)*, który umożliwia wypowiadanie się o poziomie biegłości – zwykle odniesionym do zewnętrznego standardu, np. poziomu opanowania wymagań programowych (Niemierko, 2009, 6). Przykładem

wykorzystania w egzaminach pomiaru sprawdzającego są używane w latach 70. w Stanach Zjednoczonych testy minimum kompetencji, wprowadzone w celu zapewnienia osiągania przez absolwentów określonego progu kompetencji, np. jako warunku uzyskania świadectwa (Niemierko, 2009, s. 248–249). W ramach tej kategorii wyróżnia się ostatnio testy odniesione do wymagań (ang. *standards-referenced tests*), w których poziomy wykonania (ang. *performance levels*) odnoszą się do określonych efektów uczenia ujętych w dokumentach programowych dla konkretnej klasy czy etapu edukacyjnego.

Obecność egzaminów na poszczególnych etapach edukacyjnych i ich znaczenie dla ucznia to nie jedyne cechy, które charakteryzują poszczególne systemy egzaminacyjne. Sposób organizacji oraz zakres egzaminów i innych form mierzenia osiągnięć uczniów bardzo się różni. O ile np. w Polsce istnieje ścisły podział na ocenianie zewnętrzne i wewnętrzne, to w niektórych krajach podział nie jest tak wyraźny. Na przykład, w niektórych krajach prace egzaminacyjne są oceniane w szkołach przez nauczycieli, a porównywalność oceniania zapewnia się poprzez tzw. podwójne ocenianie, z wykorzystaniem powtórnej oceny przez zewnętrzną osobę. Bardziej fundamentalne różnice wiążą się z podziałem odpowiedzialności w systemie edukacyjnym i specyfiką organizacji systemu ewaluacji i oceniania. W zdecydowanej większości europejskich systemów edukacyjnych istnieją wyspecjalizowane instytucje zajmujące się oceną jakości edukacji. W większości krajów są to instytucje centralne - w niektórych, podobnie jak kuratoria oświaty w Polsce, są to instytucje działające na poziomie regionalnym lub lokalnym. To one odgrywały tradycyjnie dominującą rolę w ocenie jakości pracy szkoły, a wyniki egzaminacyjne były przez nie wykorzystywane jedynie pomocniczo, np. w celu identyfikacji szkół, które należało obejmować dodatkową lub częstszą kontrolą. W niektórych krajach to właśnie tym instytucjom przypisano rolę organizacji egzaminów. W innych, tak jak to miało miejsce w Polsce, utworzono odrębne instytucje. W nielicznych krajach (m.in. Holandia i Stany Zjednoczone) przygotowanie i przeprowadzanie egzaminów jest zadaniem instytucji spoza sektora publicznego.

Rolę egzaminów można lepiej zrozumieć, jeśli spojrzymy na nie z perspektywy poziomu wykorzystywania danych egzaminacyjnych i ich funkcji. Dane egzaminacyjne można wykorzystywać na poziomie ucznia lub absolwenta, na poziomie nauczyciela lub szkoły. Można je też wykorzystywać na poziomie jednostki terytorialnej lub całego systemu oświaty. Jeśli spojrzeć na egzaminy zewnętrzne z tej perspektywy, to stają się one elementem złożonego systemu ewaluacji i oceniania, w którym wyniki można wykorzystywać w różnych celach i w różny sposób. Podstawowym dylematem jest napięcie między funkcją rozliczania (ang. *accountability*) i funkcją poprawy jakości (ang. *improvement*): to pierwsze – obecne np. w krajach anglosaskich i nastawione na zwiększanie odpowiedzialności szkół i nauczycieli przed rodzicami i władzami oświatowymi, a to drugie – silniej obecne w systemach europejskich.

Różnorodność istniejących rozwiązań warto zilustrować pokazując kilka przykładów. Celowo dobraliśmy do tego porównania przykłady skrajne: Finlandię, Holandię, Anglię i Stany Zjednoczone. Pomoże to też umiejscowić opisane w kolejnym podrozdziale polskie rozwiązania.

Finlandia jest często wskazywana jako kraj, w którym unika się testów i nie przeprowadza się ogólnokrajowych egzaminów. Nie oznacza to jednak, że testy nie odgrywają w Finlandii żadnej roli. Sposobem ogólnokrajowego monitorowania osiągnięć uczniów są testy przeprowadzane na losowych próbach szkół, zwłaszcza na końcu poszczególnych etapów edukacyjnych. Testy tego rodzaju są organizowane pod nadzorem Ministerstwa Edukacji co kilka lat. Szkoły, które nie zostały do nich wylosowane, mogą dobrowolnie wziąć w nich udział, aby porównać osiągnięcia swoich uczniów do średniej krajowej. Pod koniec szkoły średniej uczniowie uczący się w średnich szkołach ogólnokształcących przystępują do egzaminu końcowego, będącego odpowiednikiem polskiej matury (fin. *ylioppilastutkinto*). Zdanie tego egzaminu, organizowanego przez komisję powoływaną przez Ministerstwo Edukacji, umożliwia rozpoczęcie studiów wyższych. Obowiązkowe przedmioty obejmują język ojczysty oraz kilka przedmiotów, które można zdawać na poziomie podstawowym lub rozszerzonym (drugi język narodowy, język obcy, matematyka, inny przedmiot ogólny do wyboru).

Egzaminy są rozłożone w czasie: sesje są organizowane 2 razy w roku, a zdanie egzaminu wymaga skompletowania poszczególnych z nich najpóźniej w trzeciej z kolejnych sesji egzaminu. Wynik egzaminu jest przedstawiany na 7-punktowej skali (0,2–7). Prace są najpierw oceniane przez nauczycieli, i, niezależnie przez egzaminatorów współpracujących z krajową komisją egzaminacyjną. Podobnie jak w Polsce, dokument potwierdzający wynik egzaminacyjny nie jest częścią świadectwa ukończenia szkoły. Obecnie system jest modernizowany, tak by wykorzystać możliwości, jakie dają nowe technologie informacyjno-komunikacyjne na różnych etapach przygotowania, przeprowadzania, zdawania i oceniania testów.

Przykładem ciekawego i złożonego systemu jest Holandia. Specyfiką holenderskiego systemu edukacyjnego jest spora autonomia szkół i duży odsetek szkół niepublicznych. Decyzje związane z ocenianiem są w dużej mierze podejmowane przez szkoły. Szkoły są zobowiązane do prowadzenia ewaluacji wewnętrznej oraz publikowania, co kilka lat, wyników osiąganych w testach. Same testy są jednak nieobowiązkowe i są opłacane przez szkoły. W organizowanym przez Centralny Instytut Tworzenia Testów (CITO) teście na zakończenie szkoły podstawowej uczestniczy ok. 85% szkół. Wyniki testów mają w założeniu być informacją pomocną dla uczniów w wyborze szkoły średniej, ale zagregowane wyniki są też wykorzystywane w ewaluacji wewnątrzszkolnej i ocenie funkcjonowania systemu edukacji na poziomie kraju. Dodatkowo, większość szkół korzysta w testów monitorujących (ang. *Leerling Volg Systeem*, LVS) udostępnianych odpłatnie przez CITO dla szkół podstawowych i średnich. Na poziomie szkoły podstawowej system ten składa się z wzajemnie powiązanych testów, umożliwiających podłużne śledzenie wyników konkretnego ucznia. Uczniowie rozwiązują testy dwa razy w roku: w styczniu i pod koniec roku szkolnego. System obejmuje szerokie spektrum umiejętności (językowe, matematyczne, przyrodnicze, historyczne) oraz rozwój społeczno-emocjonalny. Informacje te mają przede wszystkim walor dydaktyczny i służą wspomaganiu pracy nauczyciela, w tym zwłaszcza indywidualizacji nauczania. Umożliwiają one nie tylko porównanie w czasie, ale też odniesienie wyniku ucznia, klasy lub szkoły do rozkładu umiejętności w kraju i grupie porównawczej (np. innej szkoły lub klasy). Daje to bardzo duże możliwości wykorzystywania informacji w ewaluacji wewnątrzszkolnej. Egzamin na zakończenie szkoły średniej składa się z dwóch części: ogólnokrajowej i szkolnej. Za określenie zakresu egzaminu i odpowiadających mu kwalifikacji odpowiada Ministerstwo Edukacji, a za jego przygotowanie i przeprowadzenie odpowiada College Voor Examens (CVE), instytucja niezależna od ministerstwa. Szkoły mają dużą swobodę w określaniu zawartości części egzaminu: może to być osobny egzamin, ale istnieje też możliwość uwzględnienia wcześniejszych testów cząstkowych czy innych osiągnięć w trakcie nauki. Dane z testów są także wykorzystywane w nadzorze pedagogicznym, np. częściej kontroluje się szkoły, które osiągają słabe wyniki w testach.

Zasadniczo inną rolę odgrywają egzaminy w krajach anglosaskich. W Anglii osiągnięcia uczniów mierzy się na koniec każdego z etapów edukacyjnych (*Key Stage* 1–3, a więc po klasie 2, 6 i 11). Na koniec etapu 2 uczniowie rozwiązują testy z czytania, matematyki, gramatyki i ortografii. Prace pisemne uczniów są oceniane przez nauczycieli. W niektórych, wylosowanych szkołach przeprowadza się także testy z przyrody. Wyniki są podawane z użyciem poziomów osiągnięć (2–6, gdzie wynik poniżej poziomu 4 jest uważany za wynik niższy oczekiwanego, a poziom 5 za wynik wykraczający poza poziom oczekiwany). Pierwszym poważnym, standaryzowanym egzaminem jest egzamin kończący nauczanie obowiązkowe: General Certificate of Secondary Education (GCSE) na koniec *Key Stage* 4 (11 klasa), w którym testy obejmują przedmioty obowiązkowe: angielski, matematykę i nauki przyrodnicze, oraz przedmioty dodatkowe. Znaczenie wyników testów wzrosło po wprowadzeniu krajowych programów nauczania w 1988 r. W założeniu wyniki testów miały być wykorzystywane przez szkoły do poprawy jakości nauczania, ale ich upublicznianie miało też wzmacniać pozycję rodziców. Podkreślała to ogłoszona w 1991 r. Karta Rodziców (ang. *Parents' Charter*). W 1992 r. opublikowano tabele z wynikami szkół średnich (odsetki uczniów osiągających poszczególne poziomy GCSE), a w 1997 r. wyniki testów ze szkół podstawowych (obliczone na podstawie *Key Stage* 2). Obok wyników testów Ministerstwo Edukacji zaczęło latach 2001–2003 publikować wskaźniki edukacyjnej

wartości dodanej, z czasem uzupełnione o tzw. kontekstowe modele wartości dodanej, w których uwzględnia się nie tylko wyniki uczniów w testach, ale też różnego rodzaju charakterystyki uczniów i szkół. W założeniu informacje te mają sprzyjać zwiększeniu wykorzystania danych w ewaluacji wewnętrznej (temu służy m.in. portal RAISEonline zawierający dane o wynikach uczniów i informacje kontekstowe) odpowiedzialności szkół przed ich interesariuszami i dostarczać informacji przydatnych dla rodziców w wyborze szkoły (np. wprowadzony w 2013 r. portal School Data Dashboard). Zaczęto także definiować oczekiwania dotyczące minimalnych osiągnięć szkół, np. odsetka uczniów osiągających określony poziom osiągnięć (tzw. *school floor standards*). Pojawiły się także wskaźniki publikowane przez inne, niezależne od rządu organizacje, np. Fisher Family Trust, wykorzystujące nieco odmienną metodologię obliczania wskaźników EWD. Ważną częścią systemu ewaluacji i oceniania jest nadzór pedagogiczny prowadzony przez utworzony w 1992 Urząd ds. Standardów w Edukacji, Opieki nad Dziećmi i Umiejętności – OFSTED (Office for Standards in Education, Children's Services and Skills), którego głównym zadaniem jest nadzór nad realizacją krajowego programu nauczania (ang. *curriculum*). OFSTED publikuje jakościowe raporty z ewaluacji szkół, obejmującej trzy obszary: efekty, wyrównywanie szans i działania na rzecz bezpieczeństwa.

Krajem, w którym wyniki egzaminacyjne odgrywają bodaj największą rolę, są Stany Zjednoczone. W Stanach oceny szkół dokonuje się niemal wyłącznie na podstawie wyników testów. O ile początkowo istotą systemu było uzależnienie zdobycia przez ucznia świadectwa lub promocji do następnej klasy, to w wielu stanach USA wyniki osiągane przez uczniów mają bezpośrednie konsekwencje dla szkół, a nawet dla konkretnych nauczycieli, w postaci nagród lub różnego rodzaju sankcji. Testy są przeprowadzane co roku, obejmując każdy z roczników uczniów od klasy 3 do 8. Ich powszechność ma w założeniu zwiększyć odpowiedzialność za wyniki wszystkich uczniów. Wyniki testów można powiązać z charakterystykami ucznia i przypisać do konkretnego nauczyciela. Zarówno podstawy programowe, jak i testy są określane i organizowane przez poszczególne stany, co prowadzi do nieco różnych rozwiązań w poszczególnych częściach kraju. Przyjęta w 2001 r. ustawa *No Child Left Behind* (tłum. nie pozostawimy żadnego dziecka w tyle, NCLB) uzależniła dostęp do funduszy federalnych od wprowadzenia przez poszczególne stany podstaw programowych (standardów) i określenia progu wymagań uznanych za konieczne do osiągnięcia przez uczniów. Rozwiązania zwiększające rozliczalność (ang. *accountability*) były wprowadzane przez niektóre stany od początku lat 90. i w większości z nich z wynikami osiąganymi przez uczniów w testach związane były konkretne sankcje lub nagrody. Na przykład w stanie Floryda w 1999 r. wprowadzono tzw. „Plan A+", w ramach którego szkołom przyznawano oceny (od A do F) zależnie od wyników testów i przyrostu wyników. Ocena szkoły niosła za sobą sankcje (np. zapewnienie przez szkołę uczniom możliwości zmiany szkoły), jak i nagrody (np. środki na działania naprawcze, premie dla nauczycieli). W ramach rozwiązań określonych w NCLB ujednolicono i wzmocniono tego rodzaju rozwiązania, skupiając się przede wszystkim na identyfikowaniu szkół osiągających słabe wyniki. Stany zostały zobowiązane do publikowania wyników w podziale na pochodzenie uczniów, status społeczno-ekonomiczny i specjalne potrzeby edukacyjne, stawiając im konkretne cele do osiągnięcia do 2014 r. i zobowiązując je do monitorowania rocznych wskaźników dochodzenia do tych standardów (tzw. *adequate yearly progres*, AYP). Słabe wyniki określane są zwykle przez porównanie z poprzednim rocznikiem i z wykorzystaniem odniesienia do odsetka uczniów osiągających poszczególne poziomy wykonania (ang. *performance levels*). Szkoły uzyskujące słabe wyniki zostały zobowiązane do sporządzenia planów naprawczych, zaoferowania rodzicom możliwości zmiany szkoły z wyższymi wynikami, dodatkowych zajęć da uczniów z niskimi wynikami lub restrukturyzacji szkoły w przypadku braku poprawy wyników w kolejnych 5 latach. Wprowadzony w Stanach Zjednoczonych system opiera się na czterech elementach: standardach określających, co każdy uczeń powinien umieć, testach, które mierzą poziom umiejętności każdego ucznia, precyzyjnie określonych celach dotyczących umiejętności oraz jasno sformułowanych konsekwentnych wyniku osiąganego przez uczniów danej szkoły lub rejonu. W większości stanów publikuje się wyniki uzyskiwane przez poszczególne kategorie uczniów szkoły (tzw. ang. *report cards*) Wprowadzenie takich rozwiązań wiązało się z przekonaniem,

że uda się zwiększyć efektywność nauczania i wymusić na nauczycielach i szkołach koncentrowanie się na tym, co najważniejsze, czyli na umiejętnościach uczniów. Innym ważnym celem, który jest wyraźnie akcentowany w programie NCLB, jest wyrównywanie szans. Służyło temu monitorowanie nie tylko efektów uzyskiwanych przez wszystkich uczniów, ale też wyników w podziale na pochodzenie etniczne, status materialny czy specjalne potrzeby edukacyjne. Konsekwencje słabych wyników wiązały się nie tylko z ogólnym wynikiem, ale też słabym wynikiem którejkolwiek z tych grup. Nowym trendem jest wykorzystywanie wyników uczniów do oceny pracy nauczycieli. Zdecydowana większość stanów zobowiązała szkoły do uwzględniania wskaźników wartości dodanej lub innych danych opartych o wyniki testów do oceny pracy nauczyciela. System amerykański zasadniczo różni się od systemów europejskich, gdzie oceny jakości edukacji są formułowane na podstawie bardziej wieloaspektowych ewaluacji zewnętrznych i w których nie ma ściśle określonych sankcji, które wiązałyby się z uzyskiwaniem słabych wyników. Od krajów europejskich system amerykański odróżnia też to, że wyniki egzaminów na niższych szczeblach są rzadko wykorzystywane w rekrutacji – ze względu na istnienie osobnego systemu prywatnych egzaminów (np. SAT lub ACT) – niepowiązanych z podstawami programowymi czy programami nauczania.

1.4. Egzaminy w systemie oceniania i ewaluacji w Polsce

Polska jest krajem, w którym wyniki sprawdzianów i egzaminów odgrywają dużą rolę, choć podobnie jak w Anglii, egzaminy nie są jedyną metodą oceny jakości nauczania i wiodącą rolę w ocenie jakości szkół odgrywa system nadzoru pedagogicznego. W przeciwieństwie do rozwiązań angielskich polskie ministerstwo edukacji nie posługuje się oficjalnie wskaźnikami opartymi na danych egzaminacyjnych ani nie wyznacza ilościowych celów. W szczególności wyniki egzaminacyjne nie są oficjalnie wykorzystywane w postaci zagregowanej, np. na poziomie szkoły, gminy czy województwa, jak to ma miejsce w Anglii czy Stanach Zjednoczonych. Wprawdzie w ostatnich latach w Polsce rozwinięto szereg narzędzi opierających się na wynikach egzaminacyjnych: wskaźniki edukacyjnej wartości dodanej czy prowadzone w ostatnich latach badania monitorująco-diagnostyczne (np. badanie OBUT, które objęło zdecydowaną większość szkół w Polsce), to nie stanowią one elementu wbudowanego w formalne rozwiązania i nie są, jak to ma miejsce w Anglii, publikowane przez ministerstwo edukacji. Także polskie egzaminy zewnętrzne, które są jedynie pomiarem różnicującym ,utrudniają szersze wykorzystanie egzaminów na poziomie krajowym, np. do określania konkretnych celów, takich jak ograniczanie odsetka uczniów osiągających bardzo słabe wyniki. Tym niemniej znacząca rola egzaminów zewnętrznych wyróżnia Polskę na tle innych państw europejskich.

Wprowadzenie egzaminów zewnętrznych było jednym z kluczowych elementów przewidzianych w kompleksowej reformie edukacji z 1998 r. Prace przygotowawcze związane z wprowadzeniem krajowych egzaminów jednakowych w całym kraju, przeprowadzanych i ocenianych w taki sam sposób, zaczęły się znacznie wcześniej, wraz z uruchomieniem w 1994 roku programu Nowa Matura. Pierwsze zewnętrzne egzaminy przeprowadzono w 2002 roku (sprawdzian, egzamin gimnazjalny). Począwszy od 2004 roku, formę egzaminu zewnętrznego przyjęły także egzaminy potwierdzające kwalifikacje zawodowe (po szkole zawodowej), a od 2005 roku egzamin maturalny. Egzaminy były ściśle związane ze zwiększaniem autonomii szkół i nauczycieli oraz decentralizacją systemu oświaty. System egzaminacyjny był pomyślany jako gwarancja jednolitości kształcenia. Razem ze wspólną dla wszystkich szkół podstawą programową miał on przeciwdziałać różnicowaniu się kształcenia, zarówno pod względem zakresu treści kształcenia, jak i poziomu umiejętności uczniów. Tworzeniu polskiego systemu egzaminów zewnętrznych towarzyszyło wiele nadziei, w tym przede wszystkim przekonanie, że możliwość monitorowania w skali kraju wyników egzaminacyjnych i dostarczenie szkołom w pełni porównywalnych wyników zapewni lepsze wyniki uzyskiwane przez uczniów.

Poprzez powiązanie z podstawą programową egzaminy w Polsce pełnią funkcję regulacyjną, ale z formalnego punktu widzenia realizują przede wszystkim funkcję selekcyjną. Wyniki egzaminu

gimnazjalnego są uwzględniane jako kryterium rekrutacyjne w szkołach ponadgimnazjalnych, wyniki matury są warunkiem rozpoczęcia studiów i są używane przez uczelnie do rekrutacji przyszłych studentów. Funkcję certyfikacyjną mają z kolei egzaminy zawodowe. Znaczenie wyniku egzaminacyjnego wzrasta w kolejnych etapach kształcenia. Sprawdzian po szkole podstawowej w założeniu nie powinien być egzaminem, którego wynik ma istotne konsekwencje dla ucznia. W przypadku rekrutacji do szkół ponadgimnazjalnych wyniki egzaminacyjne są tylko jedną z informacji uwzględnianych w rekrutacji, obok ocen szkolnych, sukcesów w olimpiadach i konkursach przedmiotowych oraz innych osiągnięć. W praktyce, faktyczna doniosłość egzaminów i sprawdzianów różni się zależnie od stopnia, w jakim szkoły i uczelnie konkurują o absolwentów, czy raczej absolwenci konkurują w staraniu się o przyjęcie na wymarzony kierunek, uczelnię lub do wymarzonej szkoły. W większości dużych miast spory odsetek uczniów gimnazjów uczy się poza rejonem, co sprawia, że już wyniki ze sprawdzianu po szkole podstawowej wpływają na szanse dostania się do niektórych gimnazjów. Funkcja selekcyjna nie jest jedyną funkcją przypisaną polskim egzaminom. W założeniach polskich rozwiązań prawnych ocenianie zewnętrzne przynosi nie tylko informacje dotyczące wiedzy i kompetencji ucznia po zakończonym etapie kształcenia, ale też generuje informacje ważne dla szkoły, organów prowadzących, nadzoru pedagogicznego czy ministerstwa edukacji narodowej, dostarczając wiarygodnych i porównywalnych danych dotyczących jakości kształcenia. O tych założeniach świadczy wyraźnie katalog zadań Centralnej Komisji Edukacyjnej i okręgowych komisji egzaminacyjnych, które nie ograniczają się jedynie do przeprowadzania sprawdzianów i egzaminów. Do zadań okręgowych komisji egzaminacyjnych należy m.in. analizowanie wyników sprawdzianów i egzaminów, w tym eksternistycznych, a także formułowanie wniosków w tym zakresie oraz opracowywanie i przekazywanie dyrektorom szkół, organom prowadzącym szkoły, kuratorom oświaty i Centralnej Komisji Egzaminacyjnej sprawozdań z przeprowadzonych sprawdzianów i egzaminów (art. 9c ust. 2 pkt 3 i 4). Z kolei do zadań Centralnej Komisji Egzaminacyjnej należy m.in. dokonywanie analizy wyników sprawdzianu i egzaminów, w tym eksternistycznych, a także składanie ministrowi właściwemu do spraw oświaty i wychowania sprawozdań odpowiednio o poziomie osiągnięć uczniów na poszczególnych etapach kształcenia oraz dotyczących wyników egzaminów eksternistycznych (art. 9a ust. 2 pkt 2). Dane egzaminacyjne i ich interpretacja powinny więc służyć dyrektorom szkół, organom prowadzącym oraz nadzorowi pedagogicznemu do kształtowania polityki edukacyjnej.

Zadanie analizy wyników egzaminacyjnych nałożono także na jednostki samorządu terytorialnego, które są organami prowadzącymi większości szkół w Polsce. Od 2009 r. jednostki samorządu terytorialnego są zobowiązane do uwzględniania wyników egzaminacyjnych w informacjach o wykonywaniu zadań oświatowych. Wyniki egzaminacyjne nie są oczywiście jedyną informacją, którą można znaleźć w sprawozdaniach, ale omówienie wyników sprawdzianu i egzaminów w szkołach podległych danej jednostce samorządu, jest jedynym obligatoryjnym elementem tej informacji (art. 5a ust. 4 usos). Wprowadzenie tego przepisu motywowano potrzebą zwiększenia kontroli społecznej nad lokalną polityką oświatową i uczynienie jej przedmiotem debaty publicznej. Świadczy to o przekonaniu ustawodawcy, że wyniki egzaminacyjne są ważną i wiarygodną informacją o jakości edukacji.

Warto także zwrócić uwagę na wykorzystanie egzaminów w nadzorze pedagogicznym. Analizowanie i ocenianie „efektów działalności dydaktycznej, wychowawczej i opiekuńczej oraz innej działalności statutowej szkół i placówek" jest jednym z ustawowych zadań nadzoru pedagogicznego (art. 33 usos), którego zakres precyzuje obecnie rozporządzenie Ministra Edukacji Narodowej z dnia 27 sierpnia 2015 r. w sprawie nadzoru pedagogicznego. Efekty, o których mowa są tylko jednym z obszarów nadzoru pedagogicznego. Specyfiką ewaluacji zewnętrznej jest też to, że dotyczy ona szkoły jako całości – ma ona w założeniu kompleksowy, jakościowy charakter obejmujący różne aspekty działania szkoły. Punktem wyjścia ewaluacji zewnętrznej są powszechne, sprecyzowane w rozporządzeniu, wymagania wobec szkół i placówek. Obejmują one szeroko rozumiane efekty, nie tylko dydaktyczne, ale też efekty w zakresie działalności opiekuńczej i wychowawczej, a także procesy zachodzące w szkole, współpracę szkoły ze środowiskiem lokalnym oraz zarządzanie szkołą. W obszarze

dotyczącym efektów wymagania, określone w rozporządzeniu Ministra Edukacji Narodowej z dnia 6 sierpnia 2015 w sprawie wymagań wobec szkół i placówek, informacji ze sprawdzianu oraz egzaminów. Według wymagania 11 „szkoła lub placówka, organizując procesy edukacyjne, uwzględnia wnioski z analizy wyników sprawdzianu, egzaminu gimnazjalnego, egzaminu potwierdzającego kwalifikacje zawodowe i egzaminu potwierdzającego kwalifikacje w zawodzie oraz innych badań zewnętrznych i wewnętrznych". W tym przypadku wyniki egzaminacyjne są traktowane jako dane, które mogą być pomocne w doskonaleniu jakości nauczania. Wyraźnie widoczne jest też założenie, że z wyników egzaminacyjnych można, w powiązaniu z innymi danymi wyciągać wnioski o efektywności działań szkoły lub nauczycieli.

Ustawa o systemie oświaty daje także możliwość zastosowania środków prawnych w razie poważnych uchybień, które mogą polegać na działalności z naruszeniem prawa, ale też stwierdzeniu „niedostatecznych efektów kształcenia lub wychowania" (art. 34. ust 2). Dyrektor szkoły lub placówki, która osiąga „niedostateczne efekty", może zostać zobowiązany przez organ nadzoru do opracowania, w uzgodnieniu z organem prowadzącym, programu i harmonogramu poprawy efektywności kształcenia lub wychowania. Do stwierdzenia zaistnienia niedostatecznych efektów często wykorzystuje się wyniki egzaminacyjne.

Egzaminy zewnętrzne odgrywają także rolę w doskonaleniu zawodowym nauczycieli. Zgodnie z przepisami znowelizowanego w 2012 r. *Rozporządzenia ministra edukacji narodowej z dnia 19 listopada 2009 r. w sprawie placówek doskonalenia nauczycieli* (Dz. U. Nr 200, poz. 1537, z późn. zm.) publiczne placówki doskonalenia nauczycieli mają obowiązek organizowania i prowadzenia doskonalenia zawodowego nauczycieli m.in. w zakresie analizy wyników egzaminów zewnętrznych oraz potrzeb zdiagnozowanych na podstawie analizy wyników egzaminów zewnętrznych. System egzaminacyjny bezpośrednio angażuje nauczycieli w proces oceniania zadań egzaminacyjnych. Ponieważ w Polsce egzaminatorem może zostać wyłącznie nauczyciel, co roku kilkadziesiąt tysięcy nauczycieli uczestniczy w ocenianiu prac egzaminacyjnych uczniów. Jak można sądzić, jest to doświadczenie, które znajduje przełożenie na lepsze rozumienie istoty zadań egzaminacyjnych i może wpływać na proces nauczania.

Nowym zjawiskiem jest wykorzystywanie wyników egzaminacyjnych w wydatkowaniu funduszy europejskich. Wytyczne dotyczące wydatkowania środków na kształcenie kompetencji kluczowych niezbędnych na rynku pracy oraz właściwych postaw i umiejętności umożliwiły regionom w latach 2014–2020 uwzględnianie w kryteriach dostępu „wyników edukacyjnych osiąganych przez szkołę lub placówkę systemu oświaty w ramach konkretnego przedmiotu, (kwalifikacja na podstawie danych globalnych osiąganych przez szkoły lub placówki systemu oświaty) oraz sposobach ich weryfikacji (m.in. pomiaru za pomocą wyników egzaminów zewnętrznych, wskaźnika EWD czy badań osiągnięć edukacyjnych)" . W niektórych postępowaniach konkursowych kryteria dostępu wiążą się ze zagregowanymi wynikami szkół, np. relacją średniej szkoły do średniej w województwie, za czym kryje się przekonanie, że np. egzaminacyjna średnia szkoły czy wskaźnik EWD trafnie identyfikuje szkoły wymagające wsparcia. Także efektywność wydatkowania środków ocenia się coraz częściej przez pryzmat wyników egzaminacyjnych.

Podsumowując polskie rozwiązania, warto podkreślić przemyślaną koncepcję oceniania zewnętrznego, którego znaczenie rośnie wraz z kolejnymi etapami edukacyjnymi. Powoduje to, że polski sprawdzian po szkole podstawowej ma, przynajmniej w założeniu, w większym stopniu charakter badania monitorującego niż egzaminu ważnego z punktu ucznia. W polskich rozwiązaniach zwraca jednak też uwagę brak precyzyjnego zdefiniowania funkcji egzaminów. Szereg zapisów prawnych zakłada, że egzaminy mają nie tylko służyć potwierdzaniu kompetencji i rekrutacji do szkół średnich i wyższych, ale też mają być pomocne w doskonaleniu praktyki nauczania czy kształtowaniu polityki edukacyjnej. Zapisy te są jednak bardzo ogólne i nie są w pełni odzwierciedlone przez kształt egzaminów, które nie pozwalają np. porównywać wyników egzaminacyjnych w czasie. Warto także zwrócić uwagęna istotną rolę różnego rodzaju diagnoz i próbnych egzaminów, które nie są, jak to

ma miejsce w innych krajach, elementem strategii rządowych. Także wskaźniki edukacyjnej wartości dodanej nie pojawiają się jako oficjalne narzędzie polityki edukacyjnej.

1.5. Zamierzony i niezamierzony wpływ egzaminów na jakość edukacji

Rola testów w edukacji jest jedną z najczęściej dyskutowanych kwestii w badaniach edukacyjnych. Dotyczy to zarówno Polski, jak i praktycznie wszystkich innych krajów, nawet tak „wolnego od testów" kraju jak Finlandia. Większość zarzutów dotyczących testów nie odnosi się do testów samych w sobie, ale celów i sposobów, w jakich są one wykorzystywane. Dyskusje dotyczące testów dotyczą przede wszystkim skutków, jakie wywołują, w zamierzony lub niezamierzony sposób, w oświacie. Chodzi tu nie tylko o osławione „uczenie pod testy", ale też wpływ testów na postrzeganie celów edukacyjnych. Krytycy szerokiego stosowania standaryzowanych testów (w tym krajowych egzaminów), uważają, że takie działania nie służą doskonaleniu edukacji, a wręcz mają destrukcyjny wpływ na dobrą edukację. Ich zdaniem stymulowanie rywalizacji uczniów i szkół „konserwuje" przestarzały model edukacji, prowadzący do pogłębiania nierówności społecznych. Zdaniem krytyków wykorzystywanie wyników egzaminów do ewaluacji zewnętrznej i rozliczania szkół umożliwia stosowanie polityki edukacyjnej represji przez nadzór pedagogiczny oraz administrację szkolną.
Ścieranie się poglądów zwolenników i przeciwników nie jest polską specyfiką i znajduje odzwierciedlenie w piśmiennictwie i debatach publicznych na całym świecie. Większość badań dotyczących skutków testów prowadzona była w krajach anglosaskich, w tym zwłaszcza w Stanach Zjednoczonych, gdzie znaczenie testów zasadniczo różni się od wykorzystywania testów w krajach europejskich i w Polsce. Warto je jednak przywołać, bo w dużym stopniu obrazują one zarówno pozytywne, jak i negatywne skutki, do których może prowadzić zbyt duży nacisk na wykorzystanie testów jako narzędzia polityki oświatowej.

Znaczenie kontekstu instytucjonalnego

Wpływ egzaminów zależy przede wszystkim od funkcji, jakie wypełniają w krajowym systemie ewaluacji i oceniania. W literaturze wyróżnia się testy doniosłe, nazywane też testami wysokich stawek (ang. *high-stakes tests*) i testy niskiej stawki (ang. *low-stake tests*). Te pierwsze mają znaczące konsekwencje dla uczniów, nauczycieli lub szkół. Zamierzone i niezamierzone skutki testów są większe w tych systemach edukacyjnych, w których testy odgrywają większą rolę, ponieważ dyrektorzy i nauczyciele dążą do poprawy wyniku testów ze względu na pozytywne lub negatywne konsekwencje wyniku w postaci np. zwiększonego wynagrodzenia lub ryzyka utraty pracy. W przypadku testów niskiej stawki dyrektorzy i nauczyciele mogą wykorzystać wynik do poprawy jakości swojej pracy, nie obawiając się innych konsekwencji, ale nie będąc też specjalnie do tego zmotywowanymi. Warto zwrócić uwagę, że to pożyteczne rozróżnienie jest relatywne: test może mieć niewielkie znaczenie z punktu widzenia ucznia, ale duże znaczenie z punktu widzenia szkoły, np. jeśli niskie wyniki uzyskane przez uczniów wiążą się z konsekwencjami finansowymi dla szkoły. Różne może być postrzeganie znaczenia wyniku egzaminacyjnego przez dyrektora jedynej szkoły w niedużej gminie, a inne dyrektora szkoły, który konkuruje o uczniów z innymi szkołami – jak to ma miejsce w wielu dużych miastach w Polsce.
Kluczowe pytanie dotyczy tego, czy wykorzystywanie testów i egzaminów wpływa na umiejętności uczniów. Najprostszym sposobem sprawdzenia tego jest porównanie wyników uzyskiwanych przez uczniów z poszczególnych krajów w badaniach międzynarodowych, takich jak PISA, TIMSS czy PIRLS. Tego rodzaju porównania pokazują, że uczniowie z krajów, w których istnieją egzaminy zewnętrzne uzyskują lepsze wyniki niż uczniowie w krajach, w których nie ma egzaminów (zob. np. Juerges, Schneider i Büchel, 2005; Bishop, 2006; Fuchs i Woessmann, 2007; Woessmann, Luedemann, Schuetz i West, 2009). Pozytywny wpływ wprowadzania egzaminów zewnętrznych potwierdzają także analizy prowadzone w krajach federalnych, gdzie w różnych częściach państwa obowiązują

inne rozwiązania. Wyższe wyniki uczniów którzy uczą się w tych częściach kraju, w których wprowadzono egzaminy wykazano dla Stanów Zjednoczonych (Bishop, Moriarty i Mane, 2000), Kanady (Bishop, 1997, 1999) i Niemiec (Jürges i in., 2005; Jürges i Schneider, 2010).

Czy oznacza to, że wprowadzenie egzaminów poprawia umiejętności uczniów? Niekoniecznie. Egzaminy są tylko jednym z wielu aspektów organizacji systemu edukacyjnego i ich wpływ jest uzależniony także od innych zmiennych. Wykorzystując dane z 29 krajów OECD uczestniczących w badaniu PISA 2003, Hanushek i Woessmann (2014) pokazali, że sporą część międzynarodowych różnic w wynikach można wyjaśnić różnicami w rozwiązaniach instytucjonalnych, takich jak możliwość wyboru szkoły, istnienie egzaminów zewnętrznych, autonomia szkół, sposób monitorowania pracy nauczycieli, istnienie ewaluacji zewnętrznej czy wykorzystanie wyników egzaminacyjnych do porównywania szkół. Ważnym wnioskiem z tych analiz jest stwierdzenie, że większa autonomia szkół jest ujemnie skorelowana z wynikami uczniów w krajach, w których nie ma zewnętrznych egzaminów. Natomiast większa autonomia szkół jest pozytywnie skorelowana z wynikami uczniów w krajach, w których istnieją egzaminy. W innych analizach wykazano, że zwiększanie autonomii szkół przynosi negatywne efekty w krajach rozwijających się i pozytywne w krajach rozwiniętych (Hanushek, Link i Woessmann, 2011). Oznacza to, że egzaminy mogą pomagać w poprawie umiejętności uczniów, ale tylko wtedy, gdy nacisk na egzaminy jest skojarzony z innymi elementami zmian w systemie edukacji.

Zawężanie treści nauczania i uczenie pod test

Badania edukacyjne dają dobry obraz negatywnych i pozytywnych skutków egzaminów na poziomie uczniów, szkół i nauczycieli. Badania zamierzonych i niezamierzonych efektów egzaminów, prowadzone przede wszystkim w Stanach Zjednoczonych, pokazują, że szkoły i nauczyciele reagują na egzaminy i ich zawartość (zwłaszcza egzaminy wysokiej stawki). Wyróżnia się trzy rodzaje oddziaływań: zawężanie treści nauczania, uczenie pod test i koncentrowanie się na wybranych kategoriach uczniów.

Pierwszym z tych zjawisk jest zwiększanie czasu, który przeznacza się na nauczanie (nauczyciele) i uczenie się (uczniowie) i większy wysiłek wkładany przez nauczycieli w nauczanie tych przedmiotów, które są uwzględniane w egzaminach. Efekt ten można interpretować jako pozytywny, bo może prowadzić do zwiększenia efektywności nauczania tych przedmiotów szkolnych, które uznaje się za najważniejsze. Ale negatywną stroną jest ograniczanie zakresu nauczanych treści. Nauczyciele w większym stopniu skupiają się na tych treściach, które będą uwzględnione w egzaminach, i redukują czas przeznaczany na nauczanie tych obszarów podstawy programowej, które nie są nimi objęte, co prowadzi do zawężenia treści nauczania (ang. *narrowing curriculum*) (Abrams, 2004; Au 2008). W Stanach Zjednoczonych zaobserwowano nawet zwiększanie godzin nauczania przedmiotów objętych egzaminami kosztem innych przedmiotów, takich jak wychowanie fizyczne, historia czy zajęcia artystyczne (Rothstein, Jacobsen i Wilder, 2008, s. 45-52). Nawet w Finlandii, gdzie rola testów jest bardzo ograniczona, dość powszechnie narzeka się, że istnienie egzaminu maturalnego na koniec szkoły średniej powoduje uczenie pod testy (Sahlberg, 2014, s. 31). Wiele wskazuje na to, że zawężanie treści nauczania występuje także w Polsce czego wyrazistym przykładem może być mała uwaga przykładana do kształtowania umiejętności mówienia na lekcjach języka angielskiego – umiejętność ta nie jest sprawdzana w trakcie sprawdzianu po szkole podstawowej i egzaminu gimnazjalnegoi pojawia się dopiero na maturze. Także w niektórych podsumowaniach z ewaluacji wewnętrznych wnioski są formułowane na podstawie wyników egzaminów i zalecenia skupiają się na osiągnięciu poprawy w konkretnych przedmiotach egzaminacyjnych, nie dostrzegając szerszego spektrum umiejętności zapisanych w podstawie programowej (zob. rozdział 5 raportu).

Z zawężaniem treści nauczania wiąże się inne zjawisko – nauczanie pod konkretny test (ang. *teaching to the test*). Polega ono na przygotowaniu uczniów do uzyskania jak najwyższego wyniku w konkretnym teście, np. ćwiczeniu rozwiązywania zadań o podobnej treści i formacie co te występujące w egzaminie – ćwiczenie uczniów w rozwiązywaniu zadań zamkniętych, które najpowszechniej

występują na egzaminach. Z tego rodzaju zjawiskiem mamy też do czynienia w Polsce (Szaleniec, 2011). W naszym kraju arkusze egzaminacyjne wraz ze schematami oceniania, a często przykładami rozwiązań są dostępne powszechnie na stronie internetowej CKE. W rezultacie stanowią one dostępny materiał nie tylko dla uczniów, ale także dla nauczycieli. Takie repozytorium zadań wykorzystywane jest przez niektórych nauczycieli jako łatwy sposób „treningu" uczniów przed egzaminem. Podporządkowywanie nauczania konkretnemu sposobowi mierzenia umiejętności może powodować, że uczeń świetnie radzi sobie z zadaniami określonego rodzaju, np. z matematyki, ale uzyskałby dużo słabszy wynik w inaczej skonstruowanym teście mierzącym ten sam zakres umiejętności.
Zaobserwowano także zjawisko koncentrowania uwagi nauczycieli na poprawie wyników konkretnych kategorii uczniów, których poziom umiejętności ma największe znaczenie na egzaminie. Z tym zjawiskiem mamy do czynienia zwłaszcza w tych systemach, gdzie wymagane jest osiągnięcie określonego poziomu wykonania, jak to ma miejsce w Anglii czy Stanach Zjednoczonych. W amerykańskich warunkach oznacza to koncentrowanie większej uwagi na słabszych uczniach, tuż poniżej progu definiującego poziom umiejętności uznawany za wystarczający.

Oszukiwanie systemu

Skrajnym przykładem negatywnego wpływu na praktykę szkolną są próby oszukiwania systemu. Dyrektorzy i nauczyciele mogą próbować wpływać na to, którzy uczniowie przystępują do egzaminów. Ma to znaczenie zwłaszcza w tych systemach, w których wyniki egzaminacyjne są ważne dla oceny pracy szkoły lub nauczyciela. Jak pokazują badania amerykańskie, działania tego rodzaju mogą polegać na tym, że uczniowie powtarzają klasę, są wyrzucani ze szkoły lub celowo są nakłaniani do zdobycia zaświadczeń o specjalnych potrzebach edukacyjnych (Ravitch, 2010). Szkoły mogą też unikać przyjmowania uczniów mających słabe wyniki lub mających inne cechy, które są skorelowane z wynikami (np. uczniów z gorszym pochodzeniem społecznym). Według doniesień medialnych przejawy tego rodzaju zjawiska można znaleźć także w Polsce w przypadku egzaminu maturalnego. Ciekawym przykładem oszukiwania systemu są też działania służące obniżaniu wymagań na egzaminach. W Stanach Zjednoczonych oczekiwane przez rząd federalny zmniejszenie odsetka uczniów o niezadowalających wynikach, w wielu stanach okazało się efektem zmian w testach i sposobach punktowania zadań (Ravitch, 2010).

Konsekwencje nagród i sankcji dla szkół

Konsekwencje różnego rodzaju nagród i sankcji dla szkół wiążących się z uzyskaniem słabego lub dobrego wyniku egzaminacyjnego dobrze obrazują kilkunastoletnie doświadczenia amerykańskie z systemem rozliczania szkół wprowadzonym w 2001 r. ustawą NCLB. Zwraca się w nich uwagę na zróżnicowane oddziaływanie poszczególnych sankcji. Najsłabsze z nich, takie jak konieczność wprowadzenia zajęć dodatkowych, nie przynoszą oczekiwanego efektu. Lepsze efekty przynoszą mocniejsze sankcje przewidziane w ustawie: danie rodzicom uczniów uczących się w szkołach osiągających słabe wyniki możliwości zmiany szkoły. Zagrożenie taką sankcją sprzyja poprawie nauczania w szkołach (West i Peterson, 2006; Chiang, 2009). Poprawie wyników uczniów sprzyja też najbardziej radykalna konsekwencja: konieczność przeprowadzenia programu naprawczego, zmiany kadry nauczycielskiej, a nawet likwidacja szkoły. Głównym czynnikiem jest tu usprawnienie przywództwa i zarządzania w szkołach. Wielu autorów podchodzi krytycznie do tak radykalnego wykorzystywania danych egzaminacyjnych. Wskazuje się, że ryzyko niezamierzonych efektów w tym zwłaszcza skupianie uwagi na niektórych kategoriach uczniów: poniżej i w okolicach wymaganego progu umiejętności przynosi więcej kosztów niż korzyści (Jacob, 2005; Neal i Schanzenach, 2010; Ann i Vidgor, 2014).

Satysfakcja z pracy i motywacja nauczycieli

Szereg badań wskazuje, że egzaminy wpływają na postrzeganie szkoły przez nauczycieli i uczniów. Amerykańskie badania pokazują, że egzaminy obniżają poczucie satysfakcji zawodowej nauczycieli, którzy często uważają, że zmuszają ich do stosowania metod sprzecznych z ich wizją dobrego nauczania (Pedulla i in., 2003). Często wskazywanym przez nauczycieli problemem jest poczucie niesprawiedliwej oceny (np. niedocenienie niektórych kompetencji, brak wpływu na niektóre czynniki wpływające na wyniki uczniów, takie np. jak zróżnicowane pochodzenie społeczne uczniów, ale też zróżnicowanie wyposażenia szkół) i nierealistyczne oczekiwania (np. że uda się w krótkim terminie poprawić wyniki uczniów). Krytycy rozwiązań amerykańskich wskazują też, że rozliczanie szkół i nauczycieli z wyników uczniów powoduje zmniejszenie poczucia odpowiedzialności uczniów i ich rodziców za własne uczenie się. Wykorzystywanie egzaminów do oceny pracy szkół i nauczycieli jest też postrzegane jako zagrożenie autonomii i profesjonalizmu nauczycieli i wyraz braku zaufania władz do ich kompetencji. Dla niektórych nauczycieli rosnąca rola testów jest wymieniana jako jeden z ważnych powodów rezygnacji z zawodu (Ingersoll, Merrill i Stuckey 2014). Warto przy tym podkreślić, że w kontekście amerykańskim, skąd pochodzą powyższe wyniki, opinie wiążą się także z coraz częstszymi próbami wiązania informacji o wynikach osiąganych przez uczniów z oceną pracy nauczycieli, np. wyliczania indywidualnych wskaźników edukacyjnej wartości dodanej, co nawet w warunkach amerykańskich jest dużą zmianą w stosunku do tradycyjnych metod oceny pracy nauczyciela (Danielson, 2007; Rockoff i Speroni, 2010; Fundacja Gatesa, 2013). Zachęty w systemie rozliczania szkół nie muszą mieć wyłącznie charakteru negatywnego. Badania prowadzone w Stanach Zjednoczonych i innych krajach (Figlio i Kenny, 2007) pokazują, że wynagradzanie najlepszych nauczycieli dodatkami płacowymi sprzyja osiąganiu lepszych wyników oraz przeciwdziała odchodzeniu nauczycieli ze szkół działających w środowiskach defaworyzowanych.

Różnicowanie się szkół

Publikowanie i korzystanie z wyników testów może także wpływać na różnicowanie się szkół. Efekt ten jest wywołany, według badaczy, mechanizmem rywalizacji szkół wyrosłym na podstawie publikowania wyników uzyskiwanych przez uczniów z poszczególnych szkół (Ball, 2008; Howe, Eisenhart i Betebenner, 2001) i polega na skupianiu się w szkołach i klasach uczniów o niższych wynikach i jednocześnie niższym statusie społeczno-ekonomicznym. Dobrzy i jednocześnie zamożni uczniowie idą do dobrych klas i szkół (a to, które klasy i szkoły są dobre, można m.in. poznać po wynikach egzaminów, które są upubliczniane), a słabi i mniej zamożni – odwrotnie. W Polsce analizy prowadzone przez Romana Dolatę (Dolata, Jasińska i Modelewski, 2012) pokazały, że wprowadzenie systemu egzaminów zewnętrznych uruchomiło w dużych miastach proces selekcji i autoselekcji na progu gimnazjum. W konsekwencji wielkomiejskie gimnazja coraz bardziej różnią się od siebie pod względem wyników egzaminacyjnych. Procesy te można wiązać z lepszą dostępnością danych egzaminacyjnych, choć związek ten nie jest oczywisty. Zwiększanie dostępności informacji o jakości szkół poprawia sytuację mniej zamożnych, którzy zwykle mają słabszy dostęp do tego rodzaju informacji. Ale jednocześnie bardziej zamożni rodzice są bardziej skłonni korzystać z tego rodzaju informacji i są lepiej przygotowani do jej interpretacji.

Problemy metodologiczne wykorzystywania testów jako narzędzia polityki

Podstawową zaletą egzaminów zewnętrznych jest umożliwienie wiarygodnego porównania umiejętności i wiedzy uczniów kończących różne szkoły oraz możliwość porównywania wyników w różnych podziałach terytorialnych. Łatwa dostępność wyników egzaminacyjnych sprzyja także zwróceniu uwagi na jakość nauczania i efekty uczenia się uczniów. Wykorzystanie wyników egzaminacyjnych ograniczają słabości metodologiczne standaryzowanych form mierzenia osiągnięć. Podstawowy problem wiąże się z tym, że testy różnią się jakością. Nawet najlepsze testy nie są precyzyjnym instrumentem pomiarowym i ich wynik wiąże się z błędem pomiaru – czego często nie dostrzegają ich użytkownicy oraz opinia publiczna. Wagę tego problemu trudno przecenić – dlatego w tym

raporcie temu tematowi poświęcamy osobny rozdział. Podkreślić należy zwłaszcza kwestię trafności, w Polsce słabo obecną w dyskusjach dotyczących korzystania z wyników egzaminacyjnych, a coraz mocniej podkreślaną w standardach profesjonalnych. Trafność testu (stopień, w jakim dane i teoria uzasadniają interpretację wyników) różni się ze względu na sposób wykorzystania wyniku. Na przykład, można oczekiwać, że wynik maturalny odzwierciedla przyswojenie umiejętności opisanych w podstawie programowej i można pod tym kątem oceniać jakość egzaminu. Ale ten sam wynik można też traktować jako wskaźnik gotowości do podjęcia studiów wyższych – takie wykorzystanie wyniku, wymaga innych dowodów pokazujących, że konstrukcja egzaminów sprzyja wypełnianiu tej funkcji. Podobnie pracodawca zatrudniający absolwenta szkoły zawodowej z potwierdzonymi egzaminem kwalifikacjami spodziewa się, że absolwent będzie w stanie wykonywać szereg zadań zawodowych, a niekoniecznie tylko te, które były uwzględnione na egzaminie – nawet jeśli dość dobrze egzamin odzwierciedla zapisy podstawy programowej dla konkretnego zawodu. Problem ten ściśle wiąże się z funkcją testów, ponieważ każdy nowy sposób wykorzystania wyników wymaga dostarczenia nowych argumentów uzasadniających trafność danego wykorzystania wyników.
Duże trudności sprawia także podjęcie decyzji o tym, w jaki sposób powinno się agregować wyniki na poziomie szkoły lub gminy. Istnieje wiele możliwych metod komunikowania zagregowanych wyników – co powoduje konieczność dokonywania arbitralnych wyborów. Na przykład, licea ogólnokształcące można oceniać na podstawie odsetka uczniów, którzy zdali egzamin maturalny, używając do tego średniego wyniku maturalnego lub wskaźników edukacyjnej wartości dodanej. Każdy z tych sposobów ma swoje zalety i ograniczenia. Generalnie odchodzi się od prostego porównania wyników egzaminacyjnych, które faworyzuje szkoły działające w bogatszych gminach, gdzie uczą się dzieci dobrze wykształconych rodziców. Znaczenie ma także wielkość szkoły – w dużych szkołach wyniki uczniów mają w przybliżeniu rozkład normalny: najwięcej jest uczniów przeciętnych, mniej uczniów bardzo słabych i bardzo dobrych. W szkołach mających niewielu uczniów wyniki będą się bardziej różnić, co powoduje, że małe szkoły łatwiej mogą znaleźć się w grupie szkół osiągających najlepsze i najgorsze wyniki. Pożyteczne wykorzystanie wyników wymaga często sięgania po rezultaty, które są odpowiednio przetworzone, a ich interpretacja wymaga wiedzy o ograniczeniach przyjętej metodologii.

1.6. Wnioski

Odpowiedzialne wykorzystywanie egzaminów do określonych celów wymaga prowadzenia badań pokazujących, że wykorzystanie egzaminów przynosi określone korzyści, a nie ma negatywnych skutków ubocznych. W standardach środowiskowych dotyczących testowania podkreśla się konieczność gromadzenia i przedstawiania takich dowodów (szerzej omawiamy tą kwestię w kolejnym rozdziale). Dotyczy to przede wszystkim tych egzaminów, których wynik ma duże znaczenie dla kariery zdającego lub ma konsekwencje dla instytucji (np. ocena przez organ prowadzący). Z samej natury egzaminowania nie da się uniknąć konsekwencji testowania, w tym negatywnych. Szkoły i nauczyciele reagują na zachęty wbudowane w system oceniania szkół i nauczycieli. Dzieje się tak zwłaszcza wtedy, gdy z wynikami tej oceny wiążą się istotne konsekwencje. Trzeba pamiętać o tego rodzaju konsekwencjach, badać je i dokumentować, a także próbować minimalizować negatywne ich skutki, gdyż mogą one powodować efekt odwrotny od zamierzonego. Doświadczenia krajów, które zdecydowały się określać ilościowe cele i wskaźniki, często przypominają o zasadzie sformułowanej przez amerykańskiego socjologa Donalda T. Campbella: im większą wagę przywiązuje się do danego wskaźnika w podejmowaniu decyzji, tym bardziej będzie on zniekształcany i tym gorzej będzie on mierzył to, co miał mierzyć.
Ale trzeba też pamiętać o pozytywnej roli danych egzaminacyjnych. Egzaminy także mogą wspierać pozytywne zmiany w procesie uczenia się, jeżeli:

- zarówno treść zadań zastosowanych w egzaminie, jak i ich format pozwalają sprawdzać szerokie spektrum umiejętności wymagających kompetencji rozwijanych w szkole;
- wyniki egzaminu wykorzystywane są tylko do tych celów, do których dany egzamin został zaplanowany;
- komunikowanie wyników jest zoptymalizowane w taki sposób, aby szkoła mogła je wykorzystać do doskonalenia środowiska uczenia się uczniów;
- nauczyciele mają możliwość uzyskania wsparcia, jak interpretować i wykorzystywać wyniki egzaminów – w tym zakresie jest szerokie pole do działania dla placówek doskonalenia nauczycieli;
- wyniki egzaminów są przedmiotem analiz efektów szkoły, ale ważne decyzje są podejmowane w kontekście szerokiej gamy różnorodnej działalności szkoły i osiągnięć jej uczniów.

Egzaminy mogą mierzyć szerokie spektrum umiejętności. Wprawdzie najczęściej spotykane testy mierzą ogóle zdolności czy wiedzę i umiejętności z określonej dziedziny, np. matematyki czy biologii, jednak nic nie stoi na przeszkodzie, by metodą testową mierzyć także umiejętność rozwiązywania problemów, umiejętność pracy w grupie czy postawy patriotyczne. Różnorodność celów edukacyjnych może być odzwierciedlana w wykorzystywanych testach, co jest pożądane także z punktu widzenia dostrzegania tych celów edukacyjnych czy poprawy jakości nauczania. Zauważalnym trendem w wielu krajach jest poprawa jakości testów egzaminacyjnych oraz wykorzystywanie nowych możliwości, jakie dają technologie informacyjno-komunikacyjne. Minimalizowanie negatywnych skutków związanych z testowaniem częściowo jest możliwe poprzez zmiany w systemie egzaminacyjnym.

Egzaminy służą zazwyczaj realizowaniu różnych funkcji: założenie, że są one wykorzystywane tylko do jednej funkcji, jest nierealistyczne. Konstrukcja egzaminów i badań diagnostycznych powinna być jednak ukierunkowana na potrzeby głównej funkcji testu. Tworzenie testu, którego celem jest zweryfikowanie umiejętności pojedynczego ucznia lub porównywanie umiejętności konkretnych uczniów, różni się od tworzenia testu z myślą o ocenie poziomu umiejętności w szkole, regionie czy kraju. W drugim przypadku mniej istotne są odpowiedzi konkretnego ucznia i więcej uwagi można poświęcić zróżnicowaniu zadań testowych. Na przykład, w międzynarodowych badaniach (np. PISA) wykorzystuje się wiele zadań o bardzo zróżnicowanym poziomie trudności, które są losowo przydzielane uczniom. Dzięki temu uzyskujemy bardziej dokładny obraz rozkładu umiejętności wszystkich uczniów. Dylematy te dobrze ilustruje polski egzamin maturalny: gdzie trudność zadań powinna być tak dobrana, by możliwie precyzyjnie wypowiadać się o uczniach, których poziom umiejętności jest bliski progowi zaliczenia egzaminu, a jednocześnie potrzeba precyzyjnej informacji o poziomie umiejętności ucznia w celu realizacji funkcji rekrutacyjnej. Prowadzi to do sytuacji, w której test dobrze wypełniający określoną funkcję, może słabo nadawać się do wypełniania innej funkcji. Z tego względu każdy nowy sposób użycia testu wymaga przedstawiania danych dotyczących trafności wykorzystania wyników do określonego celu.

Doświadczenia międzynarodowe pokazują także, że często przecenia się wartość informacyjną wyników egzaminacyjnych. Egzaminy są obarczone błędem pomiaru. W Polsce przyjęło się podawać wyniki sondaży wraz informacją o możliwym błędzie, ale nie ma zwyczaju uwzględniania tego rodzaju informacji w komunikacji wyników egzaminacyjnych, mimo że w tym przypadku mamy do czynienia z podobnym zjawiskiem. O ile są tego świadomi eksperci tworzący test i badacze, którzy z nich korzystają, o tyle świadomość tego problemu jest dużo mniejsza wśród decydentów i innych użytkowników wyników egzaminacyjnych. Z tego względu wyników testu nie powinno się wykorzystywać jako jedynej informacji w decyzjach mających konsekwencje dla osób i instytucji. Największy pożytek daje łączenie informacji pochodzących z testów z innymi informacjami. Ważne są np. dane demograficzne, w systemach informacyjnych (w Polsce w systemie informacji oświatowej), dane gromadzone w szkołach odnoszące się do procesów edukacyjnych oraz dane dotyczące postaw i opinii, ważne z punktu widzenia oceny klimatu szkoły, aspiracji i oczekiwań edukacyjnych i postaw wobec uczenia się. Wiele badań wskazuje także na to, że cennym źródłem informacji,

nadspodziewanie dobrze przewidującym dalsze sukcesy ucznia, są oceny szkolne. Ważne są z punktu widzenia oceny efektywności nauczania, mają także dane dotyczące losów absolwentów. Tego rodzaju dane mogą być analizowane na różnych poziomach. Kluczowe znaczenie ma ich wykorzystanie na poziomie szkoły, ale są też ważne w zarządzaniu lokalnymi systemami edukacyjnymi czy kształtowaniu polityki edukacyjnej na poziomie krajowym.

Bibliografia

Abrams, L. M. (2004). *Teachers' Views on High-Stakes Testing: Implications for the Classroom. Policy Brief*. Education Policy Studies Laboratory, Arizona State University College of Education.

Ahn, T., i Vigdor, J. (2014). *The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina* (Working Paper no. 20511). Cambridge, MA: National Bureau of Economic Research.

Au, W. (2008). Between education and the economy: high-stakes testing and the contradictory location of the new middle class. *Journal of Education Policy*, 23(5), 501-513.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ball, S. J. (2008), *The education debate*, Bristol: The Policy Press.

Bernhardt, V. (2004). *Data analysis for continuous school improvement*. Larchmont, NY: Eye on Education.

Bird, S., Cox, D., Farewell, V., Goldstein, H., Holt, T., i Smith, P. (2005). Performance indicators: Good, bad, ugly. *Journal of the Royal Statistical Society*, 168(1), 1–27.

Bishop, J. H. (1997). The effect of national standards and curriculum-based exams on achievement. *The American Economic Review*, 260–264.

Bishop, J. H. (1999). Are national exit examinations important for educational efficiency? *Swedish Economic Policy Review*, 6 (2), 349–398.

Bishop, J. H. (2006). Drinking from the Fountain of Knowledge: Student Incentive to Study and Learn Externalities, Information Problems and Peer Pressure. w: E. A. Hanushek, F. Welch (red.), *Handbook of the Economics of Education*, Volume 2, 909–944. Amsterdam: North-Holland.

Bishop, J. H., Moriarty, J. Y., i Mane, F. (2000). Diplomas for learning, not seat time: the impacts of New York Regents examinations. *Economics of Education Review,* 19(4), 333-349.

Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist*, 44(3), 576b.

Chiang, H. (2009), How Accountability Pressures on Failing Schools Affects Student Achievement. *Journal of Public Economics,* t. 93, 1045–1057.

Danielson, C. (2007). *Enhancing professional practice: a framework for teaching* (wyd. 2). Alexandria: Association for Supervision and Curriculum Development.

Dolata, R., Jasińska, A., i Modzelewski, M. (2012). Wykorzystanie krajowych egzaminów jako instrumentu polityki oświatowej na przykładzie procesu różnicowania się gimnazjów w dużych miastach. *Polityka Społeczna*, 1, 41–46.

Federowicz, M. i Sitek, M. (red.). (2011). *Społeczeństwo w drodze do wiedzy. Raport o stanie edukacji 2010*. Warszawa: instytut Badań Edukacyjnych

Figlio, D. i Getzler, L. (2002). *Accountability, ability and disability: Gaming the system?* (National Bureau for Economic Research Working Paper 9307). Cambridge, MA: National Bureau for Economic Research.

Figlio, D. N., i Kenny, L. W. (2007). Individual teacher incentives and student performance. *Journal of Public Economics*, 91(5), 901-914.

Figlio, D. i Loeb, S, (2011). School Accountability. w: E. A. Hanushek, S. Machin i L. Woessmann, (red.). *Handbook of the Economics of Education*, t. 3, 383-421. Amsterdam: North-Holland.

Fuchs, T., i Woessmann, L. (2007). What Accounts for International Differences in Student Performance? A Re-examination using PISA Data. *Empirical Economics* 32 (2–3), 433–464.

Fundacja Gatesa (2013). *Measures of effective teaching* (MET). http://www.gatesfoundation.org/united-states/Pages/measures-of-effective-teaching-fact-sheet.aspx.

Hanushek, E. A., i Woessmann, L. (2014). Institutional Structures of the Education System and Student Achievement: A Review of Cross-country Economic Research*. *Educational Policy Evaluation through International Comparative Assessments*, 145.

Hanushek, E. A., Link, S. i Woessmann, L. 2011. *Does School Autonomy Make Sense Everywhere?* Panel Estimates from PISA. NBER Working Paper 17591. Cambridge, MA: National Bureau of Economic Research.

Howe, K., Eisenhart, M., i Betebenner, D 2001. *School Choice Crucible: A Case Study of Boulder Valley*. Phi Delta Kappan, 83(2): 137-146.

Ingersoll, R. M., Merrill, L. i Stuckey, D. (2014). Seven Trends: The Transformation of the Teaching Force. *CPRE Research Report* nr. 80.Philadelphia: Consortium for Policy Research in Education.

Jacob, B. (2005). Testing, accountability, and incentives: The impact of high-stakes testing in Chicago public schools. *Journal of Public Economics*, 89(5/6), 761–796.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of public Economics,* 89(5), 761-796.

Juerges, H., Schneider, K. i F. Buechel. (2005). The Effect of Central Exit Examinations on Student Achievement: Quasi-Experimental Evidence from TIMSS Germany. *Journal of the European Economic Association* 3 (5), 1134-1155.

Jürges, H., i Schneider, K. (2010). Central exit examinations increase performance... but take the fun out of mathematics. *Journal of Population Economics,* 23(2), 497-517.

Jürges, H., Schneider, K., i Büchel, F. (2005). The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany. *Journal of the European Economic Association*, 1134-1155.

Kane, T.J. i Staiger, D.O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16 (4), 91–114

Konarzewski, K. (2008) *Przygotowanie uczniów do egzaminu: pokusa łatwego zysku. Raport badawczy*, Warszawa.

Levitas, A. (2012). Wokół strategii oświatowych. Polskie i amerykańskie debaty o strategicznych problemach edukacji. w: A. Levitas (red.) *Strategie oświatowe*.. Warszawa: ORE-ICM

Neal, D. i D. Schanzenbach (2010) Left Behind By Design: Proficiency Counts and Test-Based Accountability. *Review of Economics and Statistics*. 92(2): 263-283.

Niemierko, B. (2009). *Diagnostyka edukacyjna*. Warszawa: Wydawnictwo Naukowe PWN.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., i Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Chestnut Hill, MA: National Board on Educational Testing and Public Policy, Boston College.

Ravitch, D. (2010). *The Death and Life of the Great American School System: How Testing and Choice are Undermining Education*, Perseus Books, Philadelphia

Rivkin, S. G., Hanushek, E. A. i J Kain, F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica* 73 (2) (March), 417-458.

Rockoff, J., i Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review*, 100(2), 261–66.

Rothstein, R., Jacobsen, R., i Wilder, T. (2008). *Grading education: Getting accountability right*. Washington, DC: Economic Policy Institute.

Sahlberg, P. (2014). *Finnish Lessons 2.0: What Can the World Learn from Educational Change in Finland?*. Teachers College Press.

Szaleniec, H. i Dolata, R. (2012). *Funkcje krajowych egzaminów w systemie edukacji. Polityka Społeczna*, (Nr tematyczny 1), 37–41.

West, Martin and Paul Peterson (2006), The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments, The Economic Journal 116 (510), C46-C62.

Woessmann, L. (2005). The Effect Heterogeneity of Central Exams: Evidence from TIMSS, TIMSS Repeat and PISA. *Education Economics* 13 (2), 143-169.

Woessmann, L., Luedemann, E., Schuetz, G., West, M. R. (2009). *Schoolaccountability, autonomy and choice around the world*. Cheltenham, UK:Edward Elgar.

2. Jakość testów egzaminacyjnych

Henryk Szaleniec, Paulina Skórska, Maciej Koniewski, Przemysław Majkut, Filip Kulon

2.1. Wstęp

Testy egzaminacyjne[1] są podstawowym narzędziem pomiaru osiągnięć uczniów. Żaden test nie jest jednak doskonały, w związku z czym wyniki testów nigdy nie odzwierciedlają w pełni umiejętności uczniów. Cechą każdego pomiaru jest występowanie różnego rodzaju błędów. Błędów tych nie da się całkowicie wyeliminować, ale można je minimalizować, by zapewnić użyteczność (przydatność w praktyce) uzyskanych wyników. Ponieważ wyniki testów egzaminacyjnych mogą mieć konsekwencje dla zdających uczniów lub dla szkół, od twórców testów i osób, które podejmują na ich podstawie decyzje wymaga się refleksji nad ograniczeniami pomiarowymi i jakością testów. Zrozumienie zasad i kryteriów oceny jakości testów jest też ważne dla dyskusji publicznej dotyczącej wyników egzaminacyjnych.

Rozdział ten rozpoczniemy od wyjaśnienia założeń pomiaru w edukacji. Następnie wprowadzimy kilka kluczowych pojęć Klasycznej Teorii Testów (KTT; *Classical Test Theory*, CTT), które będą potrzebne do zrozumienia istoty problemu rzetelności pomiaru. Drugą właściwością testowania w edukacji, omówioną w tym rozdziale, będzie trafność wyników egzaminów. W następnej kolejności przedstawimy szeroką dyskusję na temat standardów opracowywania wysokiej jakości testów egzaminacyjnych, które zostały przygotowane dla autorów testów, celem zapewnienia rzetelności i trafności pomiaru w edukacji. W kontekście tej dyskusji przedstawimy dotychczasową praktykę przygotowania i przeprowadzania egzaminów zewnętrznych w Polsce. W dalszej części rozdziału zajmiemy się przedstawieniem praktyki komunikowania wyników egzaminacyjnych przez okręgowe komisje egzaminacyjne dla różnych odbiorców. Rozdział kończy rzut oka na na obszary systemu egzaminacyjnego wymagajace doskonalenia.

Kwestia zapewniania jakości testów jest szeroko dyskutowana w środowisku badaczy i praktyków prowadzących badania i diagnozy w psychologii, pedagogice i edukacji. Od wielu lat tworzone i doskonalone są profesjonalne standardy tworzenia i korzystania z testów. Uznanym i szeroko wykorzystywanym zestawem standardów są Standardy dla testów stosowanych w psychologii i pedagogice („*Standards for Educational and Psychological Testing*"). Standardy zostały opracowane i opublikowane przez amerykańskie stowarzyszenia profesjonalne z dziedziny psychologii i edukacji (jest to wspólna publikacja *American Educational Research Association* (AERA), *American Psychological Association* (APA), oraz *National Council on Measurement in Education* (NCME)). Pod obecnym tytułem Standardy zostały po raz pierwszy opublikowane w 1966 r. i od tego czasu były czterokrotnie aktualizowane (1974, 1985, 1999, 2014). Najdłużej obowiązujące wydanie z 1999 roku doczekało się polskiego tłumaczenia pod redakcją Elżbiety Hornowskiej (Hornowska, 2007). W lipcu 2014 ukazała się najnowsza anglojęzyczna wersja Standardów, uwzględniająca znaczące zmiany, jakie miały miejsce w dziedzinie testowania na przestrzeni ostatnich 15 lat. Do Standardów – będących wyznacznikiem dobrych praktyk w zakresie jakości egzaminów – będziemy się w tym rozdziale odwoływać wielokrotnie.

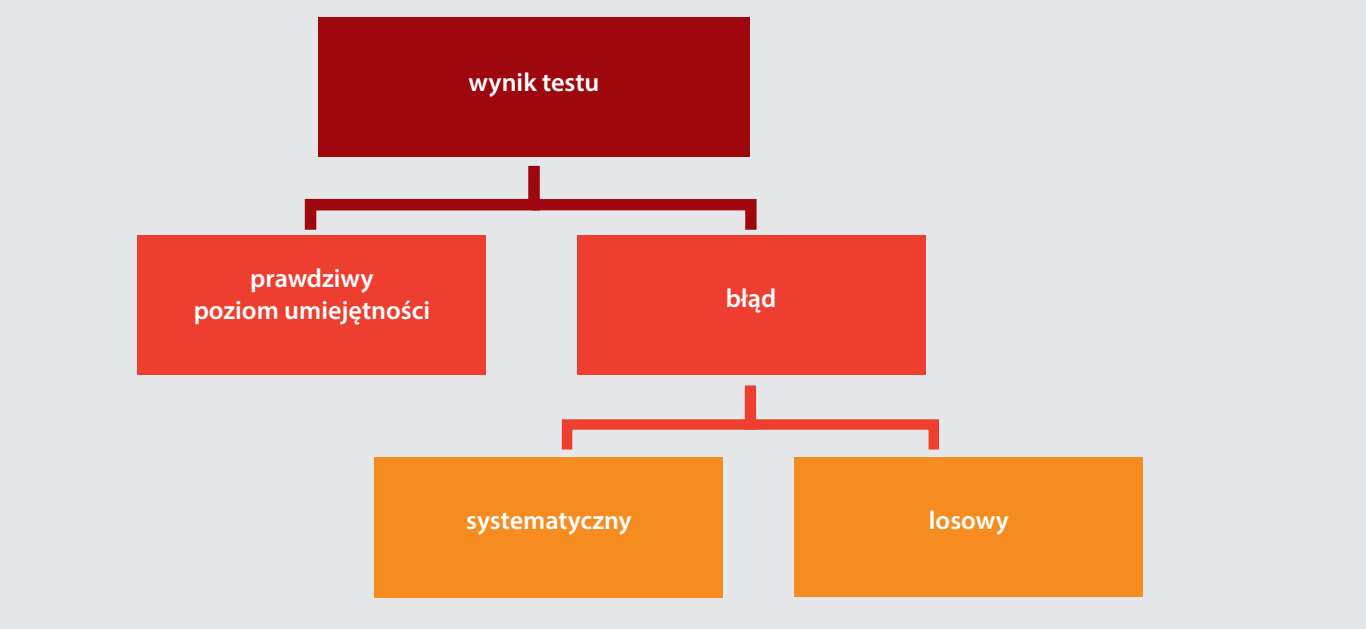
[1] W całym rozdziale dla uproszczenia zamiennie używamy określeń test i egzamin. Oczywiście istnieją testy, które nie są wykorzystywane w procedurze egzaminów zewnętrznych, np. nauczyciel w szkole może używać testu osiągnięć, który sam skonstruuje lub skorzysta z istniejących materiałów. W tym rozdziale mówiąc o testach chodzi nam za każdym razem o testy będące podstawą egzaminów zewnętrznych w Polsce.

2.2. Założenia pomiaru w edukacji

Większość cech (atrybutów) ludzi nie poddaje się bezpośrednio pomiarowi. Na przykład w edukacji jedną z najważniejszych cech uczniów, które chcielibyśmy mierzyć i na tej podstawie podejmować decyzje, jest poziom umiejętności/zdolności uczniów. Takie cechy są nazywane konstruktami, ponieważ istnieją realnie (wiemy, że ludzie posiadają określony poziom kompetencji matematycznych, językowych, inteligencji itd.), ale nie dają się bezpośrednio zaobserwować. Inaczej jest w przypadku cech fizycznych: np. wzrost czy waga, które dają się bezpośrednio obserwować, a to, na ile dokładny (bliski rzeczywistości) będzie ich pomiar, zależy od dokładności narzędzia pomiarowego i sprawności w posługiwaniu się nim. W psychologii i pedagogice o występowaniu i poziomie cech ludzi wnioskuje się na podstawie obserwowalnych zachowań, o których zakładamy, że są przejawem nieobserwowalnych cech (konstruktów). Na przykład, jeśli dana osoba uczestniczy w niedzielnym nabożeństwie, powiemy, że jest religijna. Jeśli uczeń prawidłowo rozwiąże trudny problem matematyczny, powiemy, że ma wysoki poziom umiejętności matematycznych. W oparciu o obserwacje pojedynczych zachowań nie można jednak w sposób pewny orzekać o stopniu religijności czy umiejętności matematycznych. Potrzebnych jest więcej wskaźników, czyli obserwacji zachowań ludzi w różnych sytuacjach i kontekstach. Zmierzenie poziomu interesującej nas cechy wymaga pomiaru zachowań ludzi, powtarzanego w czasie. Dlatego w edukacji poziom umiejętności ucznia określa się poprzez obserwację zachowań ucznia: wynik uzyskiwany w teście egzaminacyjnym, oceny szkolne, średnia ocen szkolnych, ocena dodatkowych osiągnięć ucznia itd. Aby pomiar umiejętności np. matematycznych był adekwatny, konstruuje się test, złożony z wielu zadań matematycznych, o różnym stopniu trudności i z różnych subdyscyplin matematyki.

Za każdym razem, kiedy uczeń rozwiąże test, uzyskany wynik reprezentuje tylko pewien zakres zachowań mogących wskazywać na poziom umiejętności. Uczeń udziela odpowiedzi na wybrane kilkanaście do kilkudziesięciu zadań, które są tylko częścią wszystkich zadań, które mogłyby mierzyć posiadaną umiejętność (kompetencję). W związku z tym, że żaden test nie jest skonstruowany idealnie, uzyskane z testu wyniki są obarczone błędem pomiaru. Błąd pomiaru może mieć charakter systematyczny lub losowy. Strukturę wyniku uzyskanego przez ucznia w teście umiejętności prezentuje rysunek 2.1.

Rysunek 2.1. Składniki pomiaru (wyniku w teście osiągnięć)



Źródło: Opracowanie własne na podstawie Haladyna i Downing (2004)

Błąd systematyczny (*systematic error*) występuje, gdy cechy uczniów lub samego testu w taki sam sposób (w tym samym kierunku) i powtarzalnie wpływają na wyniki wszystkich uczniów rozwiązujących test. Na przykład, jeśli test jest źle skonstruowany, to wpływa to na wyniki wszystkich uczniów, którzy go rozwiązują. Jeśli test z przyrody zawiera zadania łatwiejsze dla dzieci mieszkających na terenach wiejskich, to test może być obciążony błędem systematycznym: wyniki takiego testu będą systematycznie niższe dla dzieci z miasta (w porównaniu do dzieci ze wsi), niż powinny być, gdyby test nie zawierał zadań o stronniczym charakterze. Błąd taki byłby powtarzalny: gdyby jeszcze raz użyć takiego testu na innej grupie uczniów, to kierunek obciążenia wyników byłby taki sam (niedoszacowanie przez test wyników uczniów z miasta w stosunku do ich poziomu umiejętności przyrodniczych). Z kolei błąd losowy (*random error*) wpływa na różnych uczniów w różny sposób: u jednych uczniów efekt oddziaływania tego błędu jest ujemny, a u innych dodatni. Nie da się jednoznacznie zidentyfikować źródeł występowania tego błędu. Autorzy testu i osoby odpowiedzialne za przeprowadzenie testu nie mogą go kontrolować. Dla przykładu w dniu zdawania egzaminu niektórzy uczniowie mogą czuć się gorzej lub być chorzy. Choroba nie oddziałuje na wszystkich tak samo. Innymi przykładami warunków testowania, które są źródłem błędów losowych, są wpływ temperatury, wentylacji, inne czynniki mogące rozpraszać uczniów. Tego rodzaju błędy powinny być brane pod uwagę przez autorów i użytkowników testów. Określenie i minimalizowanie zagrożenia błędem systematycznym jest elementem podnoszenia trafności testowania, zaś oszacowanie błędu losowego elementem podnoszenia rzetelności pomiaru.

2.3. Rzetelność testu egzaminacyjnego

W polskim systemie egzaminacyjnym od samego początku jego powstania (zarówno w Centralnej, jak i okręgowych komisjach egzaminacyjnych (CKE i OKE)) wykorzystuje się klasyczną teoria testu (KTT), jako przewodnik dla autorów zadań i całych arkuszy egzaminacyjnych. Korzystają z niej także statystycy OKE przeprowadzając podstawowe analizy wyników egzaminacyjnych. Stanowi ona również podstawę pojęciową dla komunikowania i interpretacji wyników obok rozwijającej się w ostatnich latach teorii odpowiedzi na pozycje testowe - IRT (*Item Response Theory*)[2].

Podstawowym elementem testu jest zadanie lub pytanie, na które uczeń udziela odpowiedzi[3]. W arkuszach egzaminacyjnych stosowanych w polskich egzaminach zewnętrznych stosowane są zarówno zadania zamknięte, wymagające wyboru odpowiedzi, jak i zadania nazywane otwartymi, które wymagają krótkiej odpowiedzi, lub dłuższej (np. zapisania rozwiązania zadania z matematyki, fizyki, chemii lub napisania rozprawki, opowiadania, analizy krytycznej). W niektórych egzaminach stosowane są tylko zadania zamknięte, jak w egzaminie gimnazjalnym z historii i przedmiotów przyrodniczych (począwszy od roku 2012) lub tylko zadania otwarte - matura z języka polskiego i matematyki na poziomie rozszerzonym. W obrębie zadań zamkniętych i otwartych funkcjonuje kilka szczegółowych rozróżnień w zależności od typu sformułowania pytania i sposobu udzielania odpowiedzi (Niemierko, 1975; Hornowska, 2001; Downing, 2009; Skórska, Świst i Szaleniec, 2014a), o czym więcej w kolejnym podrozdziale.

Zadania często powiązane są z materiałami źródłowymi, którymi mogą być fragmenty tekstu, tabele, wykresy, mapy, rysunki, obrazy czy zdjęcia. Do jednego materiału źródłowego może być przypisane jedno zadanie lub kilka. I tak, w zestawie zadań ze sprawdzianu 2014 (S-1-142) do pierwszego tekstu źródłowego zatytułowanego „Zaklęty dźwięk" odnosiło się pięć zadań zamkniętych (wielokrotnego

[2] Krótkie wprowadzenie do teorii odpowiedzi na pozycje testowe przedstawione jest w rozdziale 3.1 tego raportu. Szersze wprowadzenie można znaleźć w artykule opublikowanym w Edukacji autorstwa Bartosza Kondratka i Artura Pokropka (Kondratek i Pokropek, 2013).

[3] W literaturze dotyczącej pomiaru psychologicznego i edukacyjnego taki element testu nazywany jest też pozycją testową (Hornowska, 2001). Ponieważ w Polsce częściej używa się terminu zadanie/pytanie testowe, pozostajemy przy nim w tym rozdziale. Pojęcie pozycja testowa zostało zachowane tylko dla nazwy własnej teorii odpowiedzi na pozycje testowe (IRT) oraz w przytaczanych cytatach.

wyboru – zdający wybierał jedną odpowiedź z czterech proponowanych), kolejne pięć zadań odnosiło się do fragmentu wiersza pod tytułem „Muzyka", a cztery ostatnie z 20 zadań zamkniętych powiązane były z tabelą. Sześć zadań otwartych, to w tym zestawie cztery zadania z matematyki, w których uczeń zapisuje cały tok rozwiązywania i dwa zadania z języka polskiego polegające na napisaniu ogłoszenia i opowiadania. Dla każdego zadania na etapie konstrukcyjnym przypisana jest skala punktacji. W cytowanym arkuszu każde z dwudziestu zadań zamkniętych punktowane było 0 lub 1, a kolejnych sześć otwartych zadań odpowiednio 0-1, 0-4, 0-3, 0-2, 0-1 i ostatnie badające umiejętność pisania własnego tekstu (opowiadanie) 0-9.

W rezultacie egzaminu każdemu zdającemu uczniowi przypisywany jest wynik w postaci liczby (w Polsce to suma punktów lub procent maksymalnej liczby punktów[4] za udzielenie odpowiedzi za zadania w arkuszu egzaminacyjnym). Taki wynik nazywamy zaobserwowanym, punktowym wynikiem z egzaminu. Wyniki punktowe zadań i wyniki całego egzaminu są wielkościami, na których w KTT przeprowadza się analizy. Wyniki obserwowane, zwane też wynikami surowymi, stanowią najczęściej stosowaną postać rezultatu egzaminacyjnego komunikowanego uczniom i szkołom w polskich egzaminach.

W odróżnieniu od wyników obserwowanych możemy zdefiniować wyniki prawdziwe egzaminu. Jednym z fundamentalnych założeń KTT jest to, że powtarzając wielokrotnie pomiar tym samym testem, na tej samej osobie wcale nie uzyskalibyśmy takich samych ocen, a wręcz cały indywidualny wachlarz (rozkład) wyników. Ponieważ podczas egzaminu przy rozwiązywaniu zadań następuje uczenie się i zapamiętywanie zadań, nie można empirycznie sprawdzić w prosty sposób tego założenia, jak można to w prosty sposób zrobić dla pomiaru fizycznego np. wagi ciała, wysokości itp.

W KTT wynikiem prawdziwym jest więc wynik średni z całego rozkładu poszczególnych wyników pomiaru dla tego samego ucznia – będziemy go oznaczać grecką literą τ. Wynik prawdziwy jest pojęciem statystycznym i nie można myśleć o nim jako o wyniku idealnym, czy wyniku, na który uczeń rzeczywiście zasłużył rozwiązując dany test. Wyniku prawdziwego się nie obserwuje, dlatego też nazywamy go często zmienną ukrytą (konstruktem). Wynik obserwowany i wynik prawdziwy to pojęcia, które umożliwiają zdefiniowanie błędu pomiaru, który będziemy oznaczać grecką literą ε.

Błąd pomiaru, o którym wspomnieliśmy już na początku rozdziału, definiujemy jako różnicę pomiędzy wynikiem obserwowanym X i wynikiem prawdziwym τ.

$$\varepsilon = X - \tau$$

Kolejnym ważnym pojęciem klasycznej teorii testu jest miara zmienności (zróżnicowania) wyników nazywana wariancją. Jest ona średnią arytmetyczną kwadratów odchyleń od średniej arytmetycznej wyniku. Pozwala ona określić, jakie jest rozproszenie wyników wokół średniej. Ponieważ wariancja jest miarą kwadratową, to w wielu wypadkach do opisu zróżnicowania wyników pomiaru wygodniej jest stosować pierwiastek kwadratowy z wariancji nazywany odchyleniem standardowym. Odchylenie standardowe (podawane przy wszystkich wynikach średnich w sprawozdaniach z egzaminów CKE i OKE) jest łatwe w interpretacji, gdyż jest wyrażane w takich samych jednostkach co wynik. Na przykład w Tabeli 2.1. średni wynik surowy dla kraju wynosił w 2014 r. 25,8 punktu a odchylenie standardowe 8 punktów[5].

[4] Przy czym w egzaminie gimnazjalnym od 2012, a w sprawdzianie i egzaminie maturalnym od 2015, do komunikowania wyników wykorzystywana jest skala centylowa i procent maksymalnej liczby punktów.

[5] Jak można zauważyć, analizując średnie wyniki surowe dla kraju, ich zróżnicowanie jest bardzo duże, a pomiędzy wynikiem z roku 2002 i roku 2009 różnica wynosi prawie 7 punktów – tyle co średnie odchylenie standardowe w ciągu 13 lat. Przyczyną takiego zróżnicowania może być różnica w poziomie umiejętności badanych sprawdzianem w poszczególnych rocznikach szóstoklasistów lub różnica w trudności testów zastosowanych na sprawdzianie albo jedno i drugie. Do problemu porównywalności wyników wrócimy w rozdziale 3 raportu.

Tabela 2.1. Średnie i odchylenia standardowe dla całej populacji wyników obserwowanych ze sprawdzianu (arkusz standardowy)

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
$X_{\text{średnie}}$	29,49	28,61	25,55	29,50	25,32	26,57	25,77	22,64	24,56	25,27	22,75	24,03	25,80
σ	6,83	6,73	7,83	7,43	8,56	7,82	7,52	7,63	8,03	7,50	7,63	8,38	8,00

Źródło: opracowanie własne na podstawie Szaleniec i in. (2015)

Powróćmy jeszcze do losowego błędu pomiaru. Potocznie, gdy rozmawiamy o błędzie, to mamy często na myśli pomyłkę i sytuację, w której możemy dokonać korekty polegającej na usunięciu przyczyny błędu i samego błędu. W teorii i praktyce pomiaru pojęcie błędu pomiaru ma bardziej złożone znaczenie. Każdy, kto uczestniczył w zawodach sportowych, zdaje sobie sprawę ze zmienności wyników, pomimo posiadania takiego samego rzeczywistego poziomu danej umiejętności przy kolejnych próbach, np. skoku wzwyż. Nikt z nas nie funkcjonuje w sposób ciągły na najwyższym poziomie swoich umiejętności, co dotyczy zarówno fizycznej, jak i intelektualnej aktywności. Ta fluktuacja wyników spowodowana jest przez wiele czynników zależnych od natury samego pomiaru.
W przypadku egzaminów do takich czynników możemy między innymi zaliczyć fluktuacje w zakresie fizycznej i intelektualnej wydajności uczniów piszących egzamin, niekontrolowaną zmienność warunków, w których przeprowadzany jest egzamin, różnice poszczególnych uczniów w percepcji konkretnych zadań w warunkach stresu egzaminacyjnego, różnice w interpretowaniu schematu oceniania przez oceniających zadania. Sumaryczny efekt tych wszystkich czynników składa się na to, co nazywamy ogólnie losowym błędem pomiaru. Występowanie losowego błędu pomiaru stanowi poważne zagrożenie dla rzetelności testów będących podstawą egzaminów zewnętrznych.
Rzetelność odnosi się zarówno zastosowanego narzędzia (testu egzaminacyjnego), jak i organizacji przeprowadzenia oraz oceniania egzaminu. Kiedy mówimy o rzetelności, zawsze mamy na myśli losowe błędy pomiaru. Rzetelność wyników testu definiuje się jako stosunek wariancji prawdziwej do wariancji całkowitej obserwowanego wyniku testowego. Minimalna wartość rzetelności wynosi zero. Jeżeli rzetelność egzaminu wynosi zero, oznacza to, że cała zmienność uzyskanych wyników pochodzi z błędu pomiaru. Na drugim biegunie mamy maksymalną rzetelność wynoszącą 1, co oznaczałoby brak błędu pomiaru, a cała zmienność wyników pochodziłaby od rzeczywistego zróżnicowania poziomu umiejętności uczniów. Jeżeli weźmiemy na przykład wskaźnik rzetelności sprawdzianu z 2012 roku, który wynosił 0,81, to moglibyśmy przypuszczać, że 81% zmienności zaobserwowanych wyników pochodzi ze zróżnicowania wyników prawdziwych, a 19% z błędu pomiaru.
Rzetelność obok trafności jest jednym z najważniejszych pojęć w pomiarze edukacyjnym. W klasycznej teorii pomiaru stosuje się różne podejścia do szacowania wskaźnika rzetelności np. korelację wyników dwóch testów równoległych zastosowanych w różnych terminach lub metodę połówkową. Dla pojedynczego testowania, jakim jest egzamin, w zasadzie nie da się oszacować dokładnie wskaźnika rzetelności. Możemy tylko oszacować dolną granicę rzetelności. W systemie egzaminów zewnętrznych stosuje się w tym celu wskaźnik alfa Cronbacha[6], który jest miarą wewnętrznej zgodności testu: wskaźnikiem pokazującym, w jakim stopniu wszystkie zadania w teście sprawdzają tę

[6] W ramach KTT stosuje się obok wskaźnika alfa Cronbacha jeszcze trzy metody badania rzetelności. Z praktycznych względów są trudne do zastosowania w polskim systemie egzaminów zewnętrznych. Metoda retestu daje współczynnik rzetelności zdefiniowany jako korelacja wyników tego samego testu przeprowadzonych na tych samych osobach w dwóch różnych punktach czasu. Metoda form alternatywnych dostarcza wskaźnika rzetelności jako współczynnika korelacji wyników testu z innym równoważnym testem. W Polsce można by go szacować, jeśli zadania do testu byłyby losowane z większej puli przygotowanych wcześniej zadań. Wtedy rzetelność byłaby korelacją między wynikami z dwóch testów złożonych z losowo wybranych zadań z całej grupy zadań mierzących umiejętności uczniów kończących dany etap edukacyjny. Metoda połówkowa polega na podziale testu na dwie losowe grupy zadań i skorelowaniu wyników z tych dwóch części testu. W praktyce jednak wskaźnik alfa można traktować jako uśrednione wartości współczynników alfa dla wszystkich możliwych kombinacji zadań w połówki testu (Novick i Lewis, 1967).

samą grupę umiejętności. Jeżeli sprawdzian byłby testem homogenicznym tzn. mierzącym jeden rodzaj umiejętności i wszystkie zadania byłyby w taki sam sposób punktowane, to wartość wskaźnika alfa byłaby bliska rzetelności. Ponieważ jednak sprawdzian nie jest takim testem, gdyż obejmuje umiejętności z różnych przedmiotów sprawdzane zadaniami o różnej długości skali punktów, wskaźniki alfa podane w tabeli 2.2. mogą być znacznie niższe od rzeczywistej rzetelności sprawdzianu. Tak więc jeżeli dla sprawdzianu w 2012 roku alfa Cronbacha wynosiła 0,81, to jedynie możemy powiedzieć, że rzetelność tego egzaminu nie była niższa od 0,81.

Tabela 2.2. Wskaźniki rzetelności α Cronbacha dla sprawdzianu z wykorzystaniem arkusza standardowego S1

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
α Cronbacha	0,77	0,77	0,79	0,78	0,90	0,89	0,79	0,77	0,79	0,80	0,81	0,83	0,81

Źródło: opracowanie własne na podstawie Szaleniec i in. (2015)

Jak już wspomniano rzetelność egzaminu zależy zarówno od zastosowanego narzędzia (testu egzaminacyjnego), jak i organizacji przeprowadzenia oraz oceniania egzaminu. Jeżeli skupimy się tylko na samym teście egzaminacyjnym to dominujący wpływ na wartość wskaźnika rzetelności będzie miała jakość zadań, z których składa się test, ich dobór (zgodnie z wcześniej przygotowanym planem w odniesieniu do zakresu umiejętności będących przedmiotem pomiaru na egzaminie) oraz liczba zadań. Zadania wchodzące w skład testu powinny pozwolić, na podstawie odpowiedzi na nie, dobrze odróżnić uczniów o niskich i wysokich umiejętnościach z zakresu danego egzaminu. Właściwość tę nazywamy mocą różnicującą zadań. W klasycznej teorii testu moc różnicująca zadań opisywana jest jedną liczbą i szacowana jest jako korelacja zadania z całym testem albo korelacja danego zadania z resztą testu (po wykluczeniu tego zadania) i przyjmuje wartości od minus jeden do plus jeden. W rzetelnym teście moc różnicująca zadań powinna być dodatnia i jak najwyższa. Im wyższa moc różnicująca zadań, tym większa wartość współczynnika rzetelności. Współczynnik rzetelności wzrasta także wraz ze wzrostem liczby zadań w teście. Liczba zadań w teście wynika z jednej strony z koncepcji egzaminu, a z drugiej z możliwości psychofizycznych zdających dany egzamin w trakcie jednej sesji. Jest ona zawsze przedmiotem kompromisu pomiędzy czasem przeznaczonym na test, a dążeniem do jak najwyższej rzetelności pomiaru i jest określana w procedurach przygotowywania zadań i testów stanowionych przez Centralną Komisję Egzaminacyjną.

2.4. Trafność

Trafność jest uważana za nadrzędną właściwość pomiaru (Crooks i in., 2008; Skorupiński, 2013). Jej centralne znaczenie dla jakości narzędzi pomiaru, w tym testów edukacyjnych zostało potwierdzone w profesjonalnych standardach dotyczących pomiaru w psychologii i edukacji (AERA, APA i NCME, 1999) oraz w większości wpływowych książek i artykułów z zakresu testowania (Crooks i in., 2008).
Tradycyjne koncepcje trafności testowania wskazywały na istnienie przynajmniej kilku różnych rodzajów trafności narzędzi pomiaru. Wyróżniano:

- trafność teoretyczną (*construct validity*) – odnosi się ona do stopnia, w jakim test mierzy umiejętność (umiejętności), do pomiaru której został zaprojektowany. Więcej miejsca temu rodzajowi trafności poświęcono w dalszej części rozdziału.
- trafność treściową/wewnętrzną (*content validity*) – odnosi się do stopnia, w jakim zawartość testu pokrywa się z zakresem umiejętności, które test ma mierzyć. Innymi słowy odnosi się do zakresu, w jakim zagadnienia zawarte w danym teście są reprezentatywne dla całego zbioru zadań mierzących daną umiejętność. Na przykład test z matematyki przeznaczony dla szóstoklasisty

powinien zawierać zadania, które będą w sposób reprezentatywny sprawdzać wiedzę i umiejętności, które zgodnie z podstawą programową powinien posiadać uczeń w szóstej klasie szkoły podstawowej. Określenie trafności treściowej testu opiera się na osądzie eksperckim i powinno być dokonane na etapie konstrukcji testu (Cronbach, 1980).

- trafność kryterialną/zewnętrzną (*criterion-related validity*) – jest określana poprzez porównanie wyników danego testu z określoną zmienną/cechą zewnętrzną wobec testu (nie mierzoną w teście), nazywaną kryterium. W zależności od tego, czy to kryterium zewnętrzne jest zdefiniowane jako mierzone wcześniej lub później w stosunku do ocenianego testu, w ramach trafności kryterialnej wyróżniono:
  - trafność prognostyczną (*predictive validity*) – która odpowiada na pytanie, na ile wyniki danego testu edukacyjnego pozwalają przewidywać późniejsze osiągnięcia ucznia. Na przykład, czy wyniki ucznia uzyskane w maturze pozwalają przewidywać jego osiągnięcia na studiach. Jeśli dany test (lub inne narzędzie pomiaru osiągnięć uczniów, np. oceny szkolne) charakteryzują się wysoką trafnością prognostyczną to na podstawie jego wyników, można stwierdzić, w jakich obszarach wiedzy i umiejętności uczeń może mieć problemy w przyszłości i na tej podstawie projektować pracę z uczniem na kolejnych etapach edukacyjnych. O wynikach badań nad trafnością prognostyczną wskaźników osiągnięć uczniów można przeczytać w ramce 2.1.
  - trafność diagnostyczną (*concurrent validity*) - odnoszącą się do tego, w jakim stopniu wyniki w teście korelują (są związane) z wynikami innego istniejącego testu, posiadającego sprawdzone właściwości psychometryczne (pomiarowe). Załóżmy, że naszym celem jest przygotowanie nowego testu z matematyki dla czwartoklasistów w szkole podstawowej. Aby ustalić trafność diagnostyczną tego testu, należy sprawdzić czy ma on porównywalną jakość do istniejących testów mierzących ten sam konstrukt. Jeśli wyniki nowego testu będą wysoko korelowały z wynikami istniejącego testu (np. TIMSS) to można uznać, że przygotowywany test cechuje się trafnością diagnostyczną.

Ramka 2.1. Trafność prognostyczna ocen szkolnych i wyników egzaminów zewnętrznych

CZY WIESZ, ŻE?

W powszechnej opinii oceny szkolne nie są uważane za całkowicie wiarygodny (porównywalny) wskaźnik osiągnięć ucznia, ze względu na brak stosowania jednolitych standardów i zasad oceniania we wszystkich szkołach, a nawet w ramach tych samych przedmiotów nauczanych w jednej i tej samej szkole (Camara i Michaelides, 2005; Zwick i Himelfarb, 2011). Wydawało by się, że bardziej wiarygodnym wskaźnikiem osiągnięć ucznia są wyniki egzaminów zewnętrznych, gdyż opierają się na obiektywnych, standaryzowanych testach. Wyniki badań empirycznych od wielu lat wskazują jednak, że to oceny szkolne i średnia tych ocen charakteryzują się wyższą trafnością prognostyczną niż wyniki egzaminów zewnętrznych (np. Geiser i Santalices, 2007). Według badań prowadzonych w Stanach Zjednoczonych średnia ocen z przedmiotów w szkole średniej dobrze przewiduje osiągnięcia na studiach (zob. np. Atkinson i Geiser, 2009). Średnia ocen ucznia uzyskana w szkole średniej pozwala aż w 30% przewidzieć jego/jej sukces lub porażkę na I roku studiów (Atkinson, 2001, Kobrin i in., 2008), a co więcej po I roku studiów zyskuje jeszcze wyższą moc prognostyczną (Geiser i Santelices, 2007). W Polsce badania na ten temat prowadzili dla egzaminu gimnazjalnego 2012-2013 Skórska, Świst i Szaleniec (2014b). Badania te potwierdziły bardzo wysoką moc prognostyczną średniej ocen uzyskanych w pierwszym semestrze ostatniego roku nauki w gimnazjum dla wyniku ucznia na egzaminie gimnazjalnym (35,1-48,5% mocy prognostycznej w zależności od przedmiotu egzaminowania). Zgodnie z wynikami badań na świecie średnia stanowi lepszy predyktor sukcesu na egzaminie niż pojedyncze oceny szkolne. Wyjątkiem w polskich badaniach okazała się ocena semestralna z matematyki, która

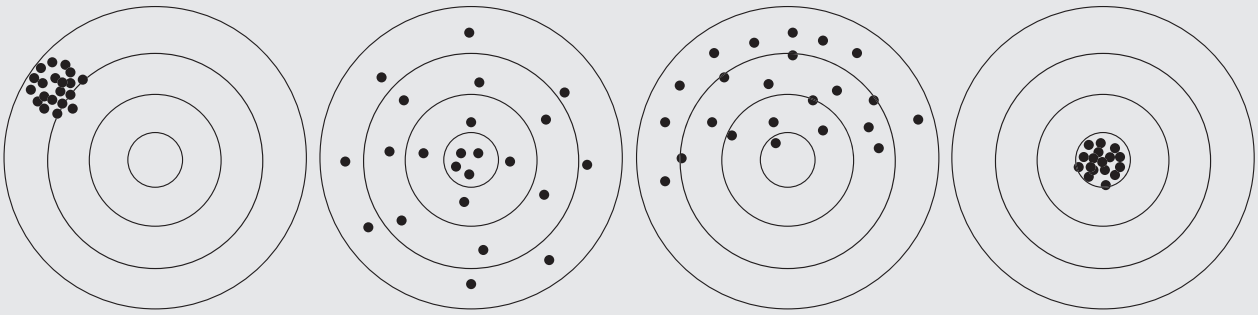
pozwala bardzo dobrze przewidywać wyniki w egzaminie gimnazjalnym (46-47,8%), lepiej niż średnia ocen.

W badaniach przeprowadzonych w 2012 i 2013 r. przez IBE, w których rejestrowano oceny semestralne uczniów i ich wynik z egzaminu gimnazjalnego z języka polskiego, historii i wiedzy o społeczeństwie, przedmiotów przyrodniczych i matematyki, pojedyncze oceny wyjaśniały od 32,8 do 47,3% zróżnicowania wyników egzaminacyjnych na poziomie indywidualnym, natomiast średnia ocen wyjaśniała od 40,4 do 48,2% zróżnicowania (Szaleniec i inni, 2013; Skórska, Świst i Szaleniec, 2014b).

Wyniki powyższych badań są mocnym argumentem przemawiającym za utrzymaniem obowiązujących rozwiązań rekrutacyjnych do szkół ponadgimnazjalnych, w których wynik egzaminu gimnazjalnego ma wagę 50%, a pozostałą liczbę punktów, w proporcjach zależnych od województwa, zapewniają oceny szkolne i szczególne osiągnięcia ucznia (np. sukcesy w konkursach przedmiotowych i artystycznych, sportowych oraz udokumentowana aktywność społeczna). Co ciekawe, znaczenie ocen szkolnych różni się między województwami, bo ich waga w rekrutacji zależy od decyzji kuratorów oświaty. W roku szkolnym 2014/2015 ich waga wynosiła, zależnie od województwa, od 25 do 40% (pozostałe 10-25% stanowiły inne osiągnięcia).

Ze względu na wielość rodzajów trafności i nie zawsze sprecyzowane procedury jej sprawdzania, trafność narzędzi pomiaru była częściej traktowana jako pożądana właściwość testowania w standardach i podręcznikach, niż sprawdzana w praktyce (Kane i American College Testing Program, 1990). W większym stopniu skupiano się na rzetelności narzędzi ze względu na matematyczny rygor teorii błędu pomiaru w analizie rzetelności oraz fakt, że rzetelność bazuje tylko i wyłącznie na wynikach testu, a trafność w dużej mierze zależy od subiektywnych ocen ekspertów i użytkowników wyników testu. Relacje tych dwóch najistotniejszych cech pomiaru w prosty sposób ilustruje rysunek 2.2.

Rysunek 2.2. Trafność i rzetelność jako dwie podstawowe cechy pomiaru



A – pomiar rzetelny (zgodne, wysoce precyzyjne wyniki), ale błędny. Wyniki rozmijają się z celem pomiaru - pomiar nietrafny.

B – pomiar trafny (zgodność z celem pomiaru), ale nieprecyzyjny, kolejne wyniki są rozbieżne – brak rzetelności.

C – niska trafność i rzetelność.

D – pomiar trafny i rzetelny.

Rzetelność jest warunkiem koniecznym, ale niewystarczającym trafnego pomiaru.

Możliwy jest jednak rzetelny pomiar, który jest nietrafny.

2.4.1. W kierunku współczesnego definiowania i badania trafności

Obecnie w teorii i praktyce pomiaru w edukacji nabiera coraz większego znaczenia holistyczna teoria trafności Messicka (1989; 1990; 1995; 2000). Wprowadziła ona dwie zasadnicze różnice wobec tradycyjnego rozumienia trafności pomiaru.

Po pierwsze dostrzeżono, że dotychczasowe koncepcje trafności były zorientowane zbyt wąsko i głównie w kierunku narzędzi i procedur przeprowadzania pomiaru. Po drugie, zwrócono uwagę na doniosłą rolę społecznych konsekwencji wyniku testu edukacyjnego i samego procesu testowania. Przeniosło to akcent rozważań o trafności. Ocenie pod kątem trafności podlegać powinno nie tyle samo narzędzie badawcze (np. test osiągnięć), co poprawność i adekwatność wniosków (interpretacji) wyciąganych na podstawie wyników testu, a co za tym idzie decyzji i działań podejmowanych na tej podstawie.

Ważne w tym kontekście są więc konsekwencje przeprowadzenia i wykorzystania wyników testu, zwłaszcza konsekwencje społeczne (Messick, 1980, 1989, 2000). Jak podkreślaliśmy w rozdziale 1 zasadniczym pytaniem w zakresie trafności testowania staje się kwestia tego, na ile dobrze test realizuje funkcje, do których został zaprojektowany. Wyniki tego samego testu mogą być wykorzystywane do skrajnie różnych celów i w odniesieniu do jednych konsekwencji test może być trafny, a stosunku do innych nie. Trafność narzędzia pomiarowego nie jest dana raz na zawsze i we wszystkich kontekstach (Stobart, 2001). Cronbach już w latach 70-tych (1971) podkreślał, że trafność nie tkwi tylko w samym teście, a każdorazowo odnosi się do zastosowania danego testu. Na przykład uczelnia wyższa przyjmuje lub odrzuca konkretnego kandydata na studia, szkoła ponadgimnazjalna klasyfikuje ucznia w zakresie matematyki jako słabego, średniego lub dobrego. W szkole podstawowej decyduje się, czy uczeń potrzebuje wsparcia w zakresie umiejętności czytania. Uzasadnienie każdej z tych decyzji opiera się na prognozie, że dany wynik testu będzie bardziej satysfakcjonujący w każdym z tych obszarów niż inny (Cronbach, 1971: s. 448). Ten sposób myślenia o trafności, rozwijany przez Messicka znalazł odzwierciedlenie w Standardach dla testów stosowanych w psychologii i pedagogice (AERA, APA i NCME, 1985) już w latach osiemdziesiątych i do dziś jest rozwijany. Już wtedy Standardy podkreślały (AERA, APA i NCME, 1985: s.13), że potrzebujemy dowodów, pozwalających ocenić przydatność użycia testu do klasyfikowania uczniów, oceny jednego ucznia względem drugiego lub udzielenia danemu uczniowi dodatkowego wsparcia dydaktycznego, a drugiemu nie. Nowoczesna teoria trafności jest holistyczna, ponieważ uznaje, że trafność jest jednolitym pojęciem, z nadrzędnym znaczeniem trafności teoretycznej, która jednocześnie podsumowuje wszystkie inne rodzaje trafności.

Ze wskazanych powyżej zmian wynika przyjęta współcześnie, zunifikowana definicja trafności, jak i określone podejście do procesu jej oceny. Współczesna definicja zakłada, że trafność odnosi się do stopnia w jakim, zarówno argumenty teoretyczne, jak i dowody empiryczne wspierają poprawność i adekwatność interpretacji uzyskanego wyniku testowego i podejmowanych na tej podstawie działań (Messick, 1989: s. 13). Trafność teoretyczna obejmuje wszelkie dowody wspierające wiarygodność wniosków wyciąganych na podstawie wyniku testowego. Dotyczy to zarówno dowodów i argumentów teoretycznych, jak i empirycznych (związki wyniku testu z innymi zmiennymi, np. poziomem inteligencji). Mówiąc inaczej trafność teoretyczna jest podstawą interpretacji wyniku testu. W Polsce na znaczenie trafności teoretycznej zwracano uwagę w zasadzie od momentu wprowadzenia systemu egzaminów zewnętrznych (Skorupiński, 2003a).

2.4.2. Proces oceny trafności pomiaru (walidacja testu)

Walidacja (ocena trafności) zgodnie z definicją samej trafności bazuje zarówno na argumentach natury teoretycznej, jak i dowodach empirycznych. Wnioski wyciągane na podstawie wyników testu należy traktować jak hipotezy, które w toku procesu walidacji będą weryfikowane. W procesie walidacji powinno się poszukiwać dowodów, które wspierają bądź obalają te hipotezy (proponowane

interpretacje i użycie wyników testu). Różnorodne dowody w tym zakresie nie są rożnymi alternatywami, lecz są komplementarne wobec siebie i wnoszą wiedzę do oceny trafności, jako całości. W tym sensie trafność jest zunifikowanym pojęciem, a historycznie wyróżniane różne rodzaje trafności mają ograniczenia i pojedynczo wykazywane nie dowodzą trafności pomiaru. Messick (1995, s.7) uważał, że powinniśmy poszukiwać dowodów i przesłanek oceny zamierzonych i niezamierzonych konsekwencji interpretacji i wniosków wyciąganych na podstawie wyników testu. Chodzi tu o wykorzystanie wyników w długo i krótko terminowej perspektywie, zwłaszcza w kontekście obciążenia wyników i interpretacji oraz niesprawiedliwego używania testu. W procesie walidacji potrzebujemy przekonujących dowodów uzasadniających wnioski wyciągane na podstawie wyników testu (a więc sposób użycia testu). W tym kontekście walidacja jest procesem podejmowania decyzji, a kluczowe pytanie brzmi: jaki jest rezultat, kiedy podejmuje się decyzję określonego typu (na podstawie wyników testu) i jaki byłby rezultat, gdyby ta decyzja była podejmowana bez tych konkretnych wyników testu (Cronbach, 1971). To co jest wymagane dla procesu walidacji to wiarygodne argumenty, że dostępne dowody uprawomocniają daną interpretację wyników testu i ich wykorzystanie. Proces walidacji wymaga skonstruowania, a następnie ewaluacji logicznej i spójnej argumentacji za i przeciw proponowanej interpretacji wyników testu i wykorzystania jego wyników (Cronbach, 1971, 1988; Messick, 1989; Kane, 1992), dlatego też proces walidacji jest de facto procesem ewaluacji *ciągłym* (Cronbach, 1988; Zumbo, 2009). Ewaluacja dotyczy tego, czy osoby interpretujące wyniki testu rozumieją je i są świadome ich ograniczeń (Stobart, 2001).

Opierając się na współczesnej definicji trafności, najnowsza edycja Standardów wymienia sześć źródeł dowodów trafności:

- treściowe (*content-oriented evidence*) – Standard 1.11. określa: „Jeśli uzasadnienie danej interpretacji i wykorzystania wyników testu opiera się na zawartości testu, to wszystkie procedury zastosowane do specyfikacji i tworzenia zawartości testu powinny zostać opisane i uzasadnione w odniesieniu do populacji (dla której test został przeznaczony) oraz cechy (umiejętności), którą test ma mierzyć (AERA, APA, NCME, 2014, s. 26). Na przykład, w ocenie trafności, autorzy testu mogą przygotować mapy, które będą pokazywały powiązania pomiędzy zadaniami testu a odpowiednimi elementami podstawy programowej. Elementy podstawy programowej, które nie znalazły odzwierciedlenia w zadaniach testu powinny być jasno wskazane.
- odnoszące się do procesów umysłowych ucznia (*evidence regarding cognitive processes*) – Standard 1.12. precyzuje: „Jeśli uzasadnienie interpretacji wyniku testowego w określonym jego użyciu opiera się na przesłankach dotyczących procesów umysłowych ucznia wykorzystywanych przy rozwiązywaniu określonych typów zadań, należy dostarczyć dowodów teoretycznych lub empirycznych, uzasadniających te przesłanki." (AERA, APA, NCME, 2014, s. 26). Jeśli więc plan testu stwierdza, że w zadaniach testowych będą sprawdzane określone procesy umysłowe ucznia, autorzy testu powinni to wykazać.
- odnoszące się do wewnętrznej struktury testu (*evidence regarding internal structure*) – Standard 1.13. określa, że „Jeśli uzasadnienie interpretacji wyniku testowego w określonym jego użyciu opiera się na założeniach o istnieniu związku między zadaniami testu, albo częściami testu, autorzy testu powinni dostarczyć dowodów na tę założoną strukturę wewnętrzną testu". (AERA, APA, NCME, 2014, s. 26-27). Na przykład autorzy testu powinni dostarczyć dowodów na to, że test jest jednowymiarowy, np. test z matematyki mierzy głównie umiejętności matematyczne, a nie inne, dodatkowe umiejętności, np. biegłość czytania treści zadań przez uczniów. Jeśli poza sumarycznym wynikiem, z testu otrzymywane są też punkty cząstkowe, powinno się wykazać ich rzetelność, a także wskazać na ich relacje między sobą.
- odnoszące się do relacji z innymi, powiązanymi cechami (*evidence regarding relationships with conceptually related constructs*) – Standard 1.16 podkreśla, że „Kiedy dowody na trafność testu zawierają analizy empiryczne odpowiedzi ucznia na zadania testowe oraz danych na temat innych cech ucznia, powinno się dostarczyć uzasadnienia wyboru takich, a nie innych cech" (AERA, APA, NCME, 2014, s. 27). Na przykład badania wskazują, że umiejętności matematyczne

mogą korelować z wynikami w testach inteligencji płynnej. W związku z tym dowody na trafność mogą obejmować korelację wyniku ucznia w teście umiejętności matematycznych z poziomem inteligencji. Dodatkowymi cechami, które można uwzględnić w analizie są cechy społeczno-demograficzne uczniów. W Polsce Centralna Komisja Egzaminacyjna gromadzi dane o zdających dany egzamin: np. płeć czy wielkość miejscowości zamieszkania ucznia.

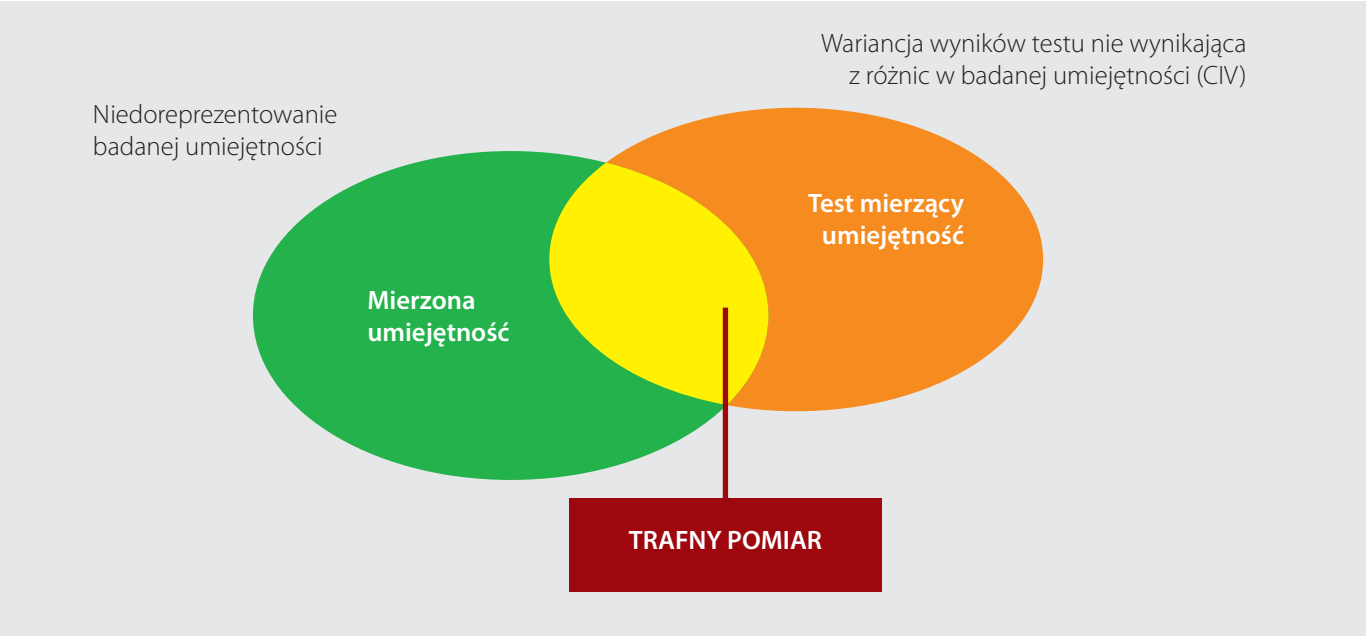
- odnoszące się do związku wyniku testu z kryterium (*evidence regarding relationships with criteria*) – Standard 1.17 podkreśla, że „Kiedy proces walidacji (oceny trafności) opiera się na dowodach, że wyniki testu są związane z jednym lub więcej kryterium, powinno się zaprezentować informacje wskazujące na adekwatny wybór kryterium, jak i jego techniczną jakość." (AERA, APA, NCME, 2014, s. 28). Na przykład takim kryterium mogą być przyszłe osiągnięcia ucznia w pracy lub na dalszych etapach edukacyjnych. Ten typ dowodów na trafność wywodzi się z klasycznej definicji trafności prognostycznej (o której pisano powyżej).
- odnoszące się do konsekwencji testu (*evidence based on consequences of tests*) – Standard 1.25 wskazuje, że „Kiedy zastosowanie testu powoduje nieplanowane konsekwencje, to należy sprawdzić, czy konsekwencje te nie wynikają z wrażliwości testu na inne cechy niż te, które z założenia podlegają ocenie, lub też z tego, że test nie reprezentuje w pełni założonego konstruktu" (AERA, APA, NCME, 2014, s. 30). Upewnienie się, że nieplanowane konsekwencje zostały ocenione pod kątem ich zakresu i przyczyn, jest obowiązkiem osób bądź instytucji odpowiedzialnych za podejmowanie decyzji o tym, czy i do czego użyć wyników testu.

W Polsce w sposób systemowy określa się wyłącznie treściowe dowody na rzecz trafności egzaminów. Inne typy dowodów, jeśli są gromadzone - to poprzez niezależne badania empiryczne.

2.4.3. Zagrożenia dla trafności pomiaru

Znaczenie procesu walidacji rośnie wraz ze wzrostem doniosłości danego testu, a więc wraz ze wzrostem liczby i wagi konsekwencji, jakie wiążą się z wynikiem testu dla osób i instytucji. Zagrożenia dla trafności można podzielić na dwie grupy (Messick, 1995). Pierwsza sprowadza się do tzw. niedoreprezentowania badanej umiejętności w teście (*construct underrepresentation*). W takiej sytuacji pomiar (np. za pomocą testu) ma zbyt wąski zakres i nie obejmuje wszystkich ważnych wymiarów, czy aspektów mierzonej umiejętności. W praktyce często problem ten objawia się zbyt małą liczbą zadań w teście, co powoduje, że na podstawie odpowiedzi uczniów na zadania nie można wnioskować (z odpowiednią precyzją) o poziomie mierzonej umiejętności. Druga kategoria wiąże się z występowaniem wariancji wyników testu nie pochodzącej z różnic w badanej umiejętności (*construct irrelevant variance, CIV*). Oznacza to, że wyniki będą odzwierciedlać nie tylko poziom umiejętności uczniów, które w zamierzeniu test miał mierzyć. Wyniki będą odzwierciedlać także inne umiejętności (niezamierzone w pomiarze) oraz niedoskonałości konstrukcyjne samego testu (np. występowanie zadań podatnych na zgadywanie poprawnej odpowiedzi). Wszystkie te elementy, które są niezwiązane z umiejętnością, którą miał w zamierzeniu mierzyć test, będą obniżać trafność wyników. Jest to naturą każdego pomiaru za pomocą testu (w tym edukacyjnego), że odzwierciedla nie tylko poziom umiejętności, do pomiaru której został stworzony, ale także wpływ czynników zakłócających. Wariancja wyników nie wynikająca z różnic w badanej umiejętności jest skutkiem występowania błędów systematycznych. Strukturę relacji pomiędzy pomiarem a prawdziwie występującym poziomem umiejętności reprezentuje rysunek 2.3. Aby zapewnić trafność testowania, w pomiarze zależy nam na tym, by zakres testu pokrywał się z zakresem mierzonej umiejętności ucznia. Jeśli wyniki testu odzwierciedlają inne czynniki, nie związane z umiejętnością ucznia, to w wynikach występuje wariancja nie związana z konstruktem (CIV). Te aspekty umiejętności ucznia, które nie znajdują odzwierciedlenia w teście obrazują problem niedoreprezentowania badanej umiejętności.

Rysunek 2.3. Relacje pomiędzy badaną umiejętnością (konstruktem) a testem



Źródło: opracowanie własne

Błąd wynikający z tego, że test mierzy coś więcej niż umiejętności ucznia, może być stały (taki sam) dla wszystkich uczniów z danej grupy piszących danych test. W takiej sytuacji wszyscy uczniowie będą mieli albo przeszacowany, albo niedoszacowany poziom umiejętności. Oznacza to, że wyniki testu dla wszystkich osób będą wskazywać albo za niski, albo za wysoki poziom umiejętności. Przykładem[7] takiego stałego błędu jest efekt egzaminatora. Jeśli dany egzaminator jest surowy, to ta surowość będzie wpływać na wszystkie oceniane przez niego prace egzaminacyjne. Prawdopodobnie wszyscy uczniowie oceniani przez tego egzaminatora będą mieć niższe wyniki testu niż powinni otrzymać przy swoim poziomie umiejętności (niedoszacowanie poziomu umiejętności przez wyniki testu). W drugim przypadku niedoszacowanie lub przeszacowanie poziomu umiejętności uczniów na podstawie wyników testu, dotyka różnych uczniów w różny sposób. Wyobraźmy sobie dwóch uczniów, którzy mają taki sam poziom umiejętności w zakresie przedmiotów przyrodniczych. Test z przyrody, który rozwiązują jest jednak tak skonstruowany, że treść zadań jest długa i ich rozwiązanie wymaga wysokiej umiejętności czytania. Choć dwaj przywołani w przykładzie uczniowie mają porównywalny poziom wiedzy i umiejętności z przyrody, jeden uzyskuje wyższe wyniki w teście niż drugi, gdyż jest bieglejszy w zakresie czytania. W tym wypadku test jest obciążony, gdyż mierzy nie tylko poziom umiejętności przyrodniczych, ale także w zakresie czytania, a problem nieprawidłowego oszacowania poziomu umiejętności przyrodniczych dotyczy uczniów w różny sposób (z zależności od ich umiejętności czytania). Innymi przykładami tego rodzaju błędu systematycznego są m.in. różny poziom motywacji testowej uczniów, różny poziom lęku testowego czy różne tempo męczenia się ucznia w trakcie rozwiązywania testu.

Standardy dla testów stosowanych w psychologii i pedagogice (AERA, APA i NCME, 1999; 2014) wskazują na znaczenie i konsekwencje występowania wariancji wyników testu nie wynikającej z różnic w umiejętności, jednak nie precyzują w sposób systematyczny, występujących w praktyce źródeł problemu. Taksonomia źródeł wariancji wyników testu nie pochodzącej z różnic w badanej umiejętności została zaproponowana przez Haladynę i Downinga (2004) i została zaprezentowana w tabeli 2.3.

[7] Wszystkie przykłady pochodzą z artykułu *Haladyny i Downinga (2004)*.

Tabela 2.3. Taksonomia źródeł CIV (construct irrelevant variance) w testach edukacyjnych

Kategoria żródeł	Źródło	Typ źródła
Przygotowanie do testu	1. Sposobność opanowania umiejętności badanych testem (opportunity to learn)	grupowe
	2. Zakres przygotowania uczniów do testu	grupowe
	3. Nieetyczne przygotowanie uczniów do testu	grupowe
Konstrukcja testu	1. Jakość zadań testowych	grupowe
	2. Format zadań testowych	grupowe
	3. Zróżnicowane funkcjonowanie zadań testowych	grupowe
Przeprowadzanie (administrowanie) testu	1. Lokalizacja (miejsce rozwiązywania) testu	grupowe
	2. Modyfikacje w sposobie przeprowadzenia testu	grupowe
	3. Partycypacja i wykluczenie z testu	grupowe
	4. Przeprowadzanie testu na komputerze	grupowe
	5. Używanie kalkulatorów w teście	grupowe
Punktowanie (ocenianie) testu	1. Błędy w punktowaniu	grupowe
	2. Weryfikacja arkusza odpowiedzi	grupowe
	3. Porównywalność wersji testu	grupowe
	4. Efekt egzaminatora i podpowiadanie przez nauczycieli	grupowe
	5. Adekwatność progów zdawalności/wykonania	grupowe
Uczniowie	1. Wpływ zdolności werbalnych na wykonanie w teście	indywidualne
	2. Lęk testowy, motywacja testowa i zmęczenie testowe	indywidualne
	3. Przystosowanie testu do określonych populacji uczniów	indywidualne
Oszukiwanie	1. Instytucjonalne	grupowe
	2. Indywidualne	grupowe

Źródło: opracowanie własne na podstawie Haladyna i Downing (2004)

W przypadku przygotowania się uczniów do egzaminu zewnętrznego podstawowym standardem jest założenie, że wszyscy uczniowie mieli zapewnione możliwości uczenia się treści objętych testem (*opportunity to learn*) oraz że w razie problemów z opanowaniem materiału zostały zastosowane środki zaradcze mające wyrównać poziom umiejętności słabszych uczniów. W Polsce punktem wyjścia dla realizacji tej zasady jest podstawa programowa, która precyzuje zakres wiedzy i umiejętności, które powinien opanować uczeń na określonym etapie edukacyjnym.

Zagrożeniem dla trafności może być także niepoprawnie przygotowany test, który zawiera obciążone zadania. Źródłem obciążenia mogą być błędy konstrukcyjne zadania, np. stosowanie negacji w treści, nielogiczna struktura możliwych odpowiedzi itd. Zadanie może źle funkcjonować psychometrycznie, być bardzo łatwe lub zbyt trudne, słabo różnicować uczniów lub być podatne na zgadywanie prawidłowej odpowiedzi przez uczniów, którzy mają bardzo niski poziom wiedzy i umiejętności. Więcej o konstrukcji i funkcjonowaniu zadań w teście można przeczytać w dalszej części rozdziału.

Sposób przeprowadzenia testu również może obniżać trafność jego wyników. Znaczenie może mieć miejsce rozwiązywania testu, przykładowo można rozważyć, czy wyniki egzaminu byłyby takie same, gdyby był on przeprowadzany w klasach, które są bardziej naturalnym środowiskiem dla uczniów, zwłaszcza najmłodszych. W niektórych przypadkach dopuszcza się możliwości przeprowadzenia

testu w sposób zmodyfikowany w stosunku do jego standardowej wersji, np. wydłużenie czasu rozwiązywania testu określonym grupom uczniów. W takich sytuacjach każdorazowo należy zastanowić się, czy takie a nie inne odstępstwa od standardowej wersji przeprowadzenia testu są sprawiedliwe dla wszystkich uczniów i nie występują nadużycia w tym obszarze. Wykluczanie lub zniechęcanie uczniów do przystąpienia do egzaminu (np. matury), zwłaszcza o niskich osiągnięciach (celem utrzymania wysokiego średniego wyniku klasy lub szkoły) może znacząco obniżać trafność wyników testu.

Powodem obniżenia trafności wyników testu mogą być błędy w punktowaniu i ocenianiu prac egzaminacyjnych, w tym efekt egzaminatora odnoszący się do różnic w łagodności/surowości oceniania zadań otwartych. Więcej o efekcie egzaminatora można przeczytać w ramce 2.3. Problemy z trafnością mogą pojawiać się też w związku z przenoszeniem odpowiedzi z testu na arkusze odpowiedzi. Porównywalność wersji egzaminu, zwłaszcza w różnych latach jego przeprowadzania powinna być zapewniona poprzez stosowanie procedur zrównywania wyników. W Polsce nie jest to zagwarantowane systemowo, lecz realizowane w ramach projektu badawczego, o czym więcej można przeczytać w rozdziale 3. Ostatnim elementem mogącym być źródłem obniżenia trafności wyników jest nieadekwatność progu zdawalności, w egzaminach, w których występuje. Więcej o konsekwencjach wyboru wysokości progu zdawalności można przeczytać w poprzednim rozdziale.

Ramka 2.3. Efekt egzaminatora w Polsce

CZY WIESZ, ŻE?

Egzamin gimnazjalny z historii i WOS (w części humanistycznej) oraz matematyki (w części matematyczno-przyrodniczej) zawiera tylko zadania zamknięte, oceniane automatycznie. W pozostałych egzaminach stosowane są również zadania otwarte, oceniane przez zewnętrznych egzaminatorów. Egzaminator posługując się instrukcją jednolitą w całym kraju dla danego egzaminu przypisuje odpowiedziom na zadania poszczególnych zdających pozycje na skali, w zależności od jakości rozwiązania tych zadań. Pozycje te, jak podkreśla Popham (1990) mogą być obciążone trzema potencjalnymi rodzajami błędów, których źródłem mogą być:

- schemat oceniania,
- procedura oceniania,
- osoba kodującego odpowiedzi w zadaniach otwartych (czyli egzaminator).

Źródła błędów związane z osobą egzaminatora określane są ogólnie mianem efektu egzaminatora. Dotyczą one szerokiej gamy efektów generujących zmienność (wariancję) punktacji za poszczególne zadania niezwiązaną z rzeczywistą jakością odpowiedzi na zadania poszczególnych zdających, ale związaną z cechami egzaminatora (Scullen, Mount i Goff, 2000). Wśród tych efektów można wyróżnić kilka najczęściej badanych i opisywanych (Saal i in., 1980):

- efekt halo - z którym mamy do czynienia, gdy oceniający przypisuje ocenianemu ten sam poziom umiejętności w różnych wymiarach (kryteriach) na podstawie ogólnego wrażenia, zamiast oceniać poszczególne wymiary niezależnie,
- łagodność i surowość - polegające na systematycznym przypisywaniu wszystkim ocenianym ocen niższych lub wyższych niż odpowiednie do ich poziomu umiejętności,
- ograniczenia skali (tendencja centralna, przydzielanie ocen skrajnych) - polegają na przypisywaniu oceniającym ocen blisko środka skali niezależnie od poziomu ich umiejętności.,
- zgodność - występuje, gdy dwóch lub więcej oceniających przypisuje tym samym ocenianym takie same oceny.

W egzaminach z przedmiotów humanistycznych, w których stosowane są zadania wymagające napisania opowiadania, rozprawki, analizy krytycznej itp. część zmienności wyników generowana przez efekty związane z cechami egzaminatora jest istotnie wyższa niż to ma

miejsce odnośnie matematyki. Jak wykazały badania prowadzone w IBE (Szaleniec i in., 2015) różnice w łagodności oceniania występują nie tylko na poziomie indywidualnym – poszczególnych egzaminatorów, ale są obserwowane również na poziomie zespołów egzaminatorów, w tym okręgowych komisji egzaminacyjnych.
Warto zauważyć, że efekt egzaminatora występuje zarówno w ocenianiu zewnętrznym, jak i w ocenianiu wewnątrzszkolnym, przy czym w ocenianiu na egzaminach mogą i powinny być monitorowane w celu minimalizacji ich wpływu na wynik.

Wyniki testu nie odzwierciedlają wyłącznie umiejętności, które test ma mierzyć, jeśli mają na nie wpływ określone cechy ucznia. Na przykład, niski wynik w teście nie musi świadczyć o niskim poziomie umiejętności ucznia, a np. o obniżonej motywacji ucznia, zawyżonym lęku testowym lub zmęczeniu ucznia w czasie rozwiązywania testu. Test może, wbrew założeniom jego autorów, mierzyć różne typy umiejętności ucznia. Na przykład test matematyczny może mierzyć też sprawność czytania. Także cechy uczniów takie jak dysleksja, problemy z koncentracją, czy niski status społeczno-ekonomiczny mogą wpływać na uzyskiwane w egzaminach rezultaty. Cechą nieuwzględnioną w klasyfikacji Haladyny i Downinga (2004), mogącą prowadzić do obciążeń, jest tzw. obycie testowe *(test-wiseness)*. Termin ten oznacza umiejętność ucznia wykorzystania wiedzy o charakterystykach i formacie testu oraz sytuacji egzaminacyjnej, w celu uzyskania wyższego wyniku (Millman, Bishop i Ebel, 1965, s. 707). Wiedzę taką uczeń może nabyć zarówno podczas lekcji w klasie szkolnej jak i na specjalnych zajęciach przygotowujących do testu. W ostatnich latach w naszym kraju jest dostępnych coraz więcej testów do zastosowania jako egzamin próbny oferowanych przez wydawnictwa i firmy komercyjne. Podczas badań prowadzonych przez IBE w gimnazjach w 2011 roku na reprezentatywnej próbie szkół, tylko 26 procent nauczycieli stwierdziło, że nie korzystało na lekcjach z komercyjnych i innych ogólnodostępnych testów. Jednocześnie 90 procent nauczycieli języka polskiego i 88 procent nauczycieli matematyki stwierdziło, że korzystało na lekcjach z arkuszy egzaminacyjnych z poprzednich lat (Szaleniec, 2011).
Ostatnią kategorią czynników obniżających trafność wyników egzaminu jest oszukiwanie. Oszukiwanie instytucjonalne odnosi się do sytuacji, w których wskutek łamania procedur egzaminacyjnych nauczyciele pomagają uczniom w rozwiązywaniu testu. Przykłady takich sytuacji to pomoc w wybraniu prawidłowej odpowiedzi na egzaminie, dawanie wskazówek, poprawianie odpowiedzi po uczniu w arkuszu. Oszukiwanie indywidualne odnosi się do uczniów, którzy podczas egzaminu podejmują próby odpisywania (ściągania).

2.5. Standardy jakości testów egzaminacyjnych

Celem podrozdziału jest syntetyczny opis standardów postępowania, które umożliwiają przygotowanie wysokiej jakości testów, stanowiących podstawę egzaminów zewnętrznych. Opis sposobu przygotowania testów zostanie przedstawiony zgodnie z kolejnością etapów przygotowania testów, zaproponowaną przez Roberta Linna (2006).

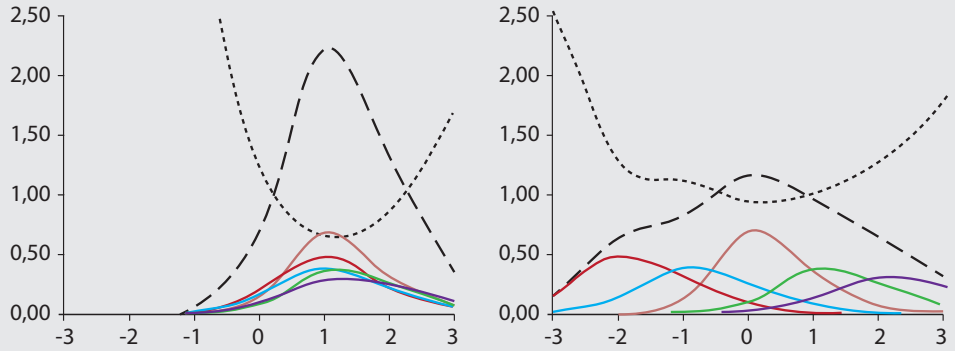
2.5.1. Określenie celów testu

Najważniejszym elementem przygotowania testu jest określenie jego celów, czyli planowanego sposobu wykorzystania wyników (Linn, 2006).
Podkreśla to standard 1.2: „Autor testu powinien wyraźnie określić, jaki jest pożądany kierunek interpretacji i sposób wykorzystania wyników testowych. (...)" (Hornowska, 2007, s. 45).
Precyzyjnie zdefiniowane cele są kierunkowskazem, do którego należy się odnosić podczas kolejnych etapów przygotowania testu, jego walidacji i interpretacji wyników.

Ramka 2.4. Określenie celu testu na przykładzie egzaminu maturalnego

CZY WIESZ, ŻE?

Dobry test nie może realizować wielu celów jednocześnie. Inaczej musi być skonstruowany test mający pełnić funkcję selekcyjną (np. rekrutacja na studia), a inaczej test diagnostyczny (np. pomiar umiejętności w populacji). Poniżej prezentujemy wykresy tzw. funkcji informacyjnej zadania testowego (*item information function, IIF*) – linie ciągłe, funkcji informacyjnej testu (*test information function, TIF*) – linie kreskowe przerywane, oraz błędu oszacowania poziomu umiejętności ucznia (standard error of the estimate, SEE) – linie punktowe przerywane. Na osi poziomej wyrażono skalę umiejętności od -3 do +3. Krzywa funkcji informacyjnej testu lub zadania testowego dostarcza wiedzy, w jakim przedziale umiejętności, test lub zadanie, jest najbardziej precyzyjnym narzędziem pomiaru tej umiejętności. Im bardziej stroma i smukła, tym bardziej precyzyjne narzędzie (test lub zadanie). Krzywa informacyjna testu stanowi agregat krzywych informacyjnych zadań, na owy test się składających.



Oba wykresy stanowią graficzne narzędzie oceny testu, który w tym przykładzie składa się z pięciu zadań o takich samych parametrach dyskryminacji i „zgadywania", ale różnych parametrach trudności. Wykres po lewej to przykład dobrego testu selekcyjnego, gdzie próg selekcji został wyznaczony o jedno odchylenie standardowe powyżej średniej, która wynosi zero. W okolicach kryterium, błąd oszacowania umiejętności ucznia jest najniższy, co oznacza że test jest w tym zakresie umiejętności najbardziej precyzyjny. Wykres po prawej to przykład dobrego testu diagnostycznego. Znajdują się w nim zadania testowe dedykowane dla uczniów o różnym poziomie umiejętności. Błąd oszacowania jest niewielki dla szerokiej grupy uczniów o poziomie umiejętności między -2 a 2 odchylenia standardowe, czyli dla 95,4% populacji uczniów.

2.5.2. Zdefiniowanie konstruktu i określenie podstawy testu

Drugim ważnym etapem w procesie opracowania testu jest zdefiniowanie zakresu konstruktu (definicja i przykłady konstruktów psychologicznych znajdują się na początku rozdziału), który test ma mierzyć. Tak zdefiniowany zakres nazywany jest „podstawą testu" *(test framework)*. Podstawa testu jest rozszerzeniem wyjściowych celów i powinna precyzować, jaka wiedza i jakie procesy umysłowe są wymagane, aby móc stwierdzić określony poziom umiejętności ucznia. Precyzyjnie zdefiniowana podstawa testu ułatwia podejmowanie decyzji co do uwzględniania w teście lub nie konkretnych pytań i zadań.

Twórcy testu powinni, przygotować opis zaleceń tworzenia zadań testowych, które będą jasno wskazywały na związek pomiędzy tymi zadaniami a modelem umiejętności, który ma być mierzony, zarówno pod kątem wiedzy, jak i procesów umysłowych ucznia, składających się na tę umiejętność.

Wszystkie te elementy muszą wynikać logicznie zarówno z ustalonych celów (funkcji) testu, jak i przyjętej definicji (modelu) badanej umiejętności.
Szczególnie ważne jest określenie procesów umysłowych wymaganych w ramach danej umiejętności. Na przykład, jeśli celem nauczania jest przygotowanie ucznia do radzenia sobie w społeczeństwie w dorosłym życiu, to powinno się kształtować umiejętność rozwiązywania złożonych problemów *(complex problem solving)*. Dlatego określenie treści testu nie powinno ograniczać się wyłącznie do wyznaczenia oczekiwanej od ucznia wiedzy faktograficznej.
Na przykład, podstawa testu matematycznego, używanego w amerykańskim badaniu NAEP (*National Assessment of Educational Progress*) definiuje kilka obszarów wiedzy (liczby i operacje na nich, pomiar, geometria, analiza danych, statystyka i prawdopodobieństwo, algebra i funkcje) oraz kilka rodzajów procesów umysłowych składających się na zdolności matematyczne (rozumienie, wiedza proceduralna i rozwiązywanie problemów).
W Polsce umiejętności, które powinien posiadać uczeń po ukończeniu kolejnych etapów edukacyjnych zostały sprecyzowane w podstawie programowej. Do tej pory np. w sprawdzianie należały do nich: czytanie, pisanie, rozumowanie, korzystanie z informacji, wykorzystanie wiedzy w praktyce. W odniesieniu do każdej z tych umiejętności, specyfikowane były rodzaje czynności umysłowych, które powinien potrafić przeprowadzić uczeń kończący kolejny etap kształcenia.

2.5.3. Opracowanie planu testu

Trzecim kluczowym etapem w przygotowaniu testu, wskazanym przez Linna (2006), jest opracowanie planu testu. Plan testu powinien określać jego cechy formalne, do których należą: format zadań (np. zadania wielokrotnego wyboru, zadania otwarte – więcej o formatach zadań w ramce 2.5), format odpowiedzi albo kryteria formułowania odpowiedzi, oraz rodzaj procedury oceniania. Opis cech formalnych może obejmować rozkłady cech psychometrycznych (np. trudność i moc dyskryminacyjna) zadań testowych, cechy testu takie jak jego trudność, moc informacyjna, rzetelność. Plan testu obejmować może także: ograniczenie czasu, charakterystyczne cechy populacji, które mają być badane testem, oraz procedury badania testem (Hornowska, 2007, s. 77).
W praktycznych aplikacjach algorytmów automatycznego/optymalnego tworzenia testów stosuje się zestawy kilkuset a nawet kilku tysięcy wymogów *(constraints)* dotyczących atrybutów zadań testowych i testu zapisanych na różnych poziomach (van der Linden, 2005).

Ramka 2. 5. Znaczenie formatu zadań egzaminacyjnych

CZY WIESZ, ŻE?

W Polsce prowadzono badania na temat klasyfikowania formatu zadań egzaminacyjnych. Skórska, Świst i Szaleniec (2014) sprawdzali, czy zadania wykorzystywane w arkuszach egzaminów gimnazjalnych w latach 2002-2011 są poprawnie klasyfikowane pod kątem formatu. Co najmniej 9 zadań z objętych analizą arkuszy zostało uznanych przez CKE jako otwarte, choć zgodnie z przyjętymi na świecie i akceptowanymi definicjami powinny one zostać uznane za zamknięte.
Błędna klasyfikacja zadań zamkniętych jako otwarte ma konsekwencje w wyliczaniu końcowego wyniku egzaminu. Dokument „Przygotowanie propozycji pytań, zadań i testów do przeprowadzenia sprawdzianu i egzaminu" (CKE, 2005) zakłada ściśle określoną wagę punktów za zadania otwarte i zamknięte dla całkowitego wyniku testu. W przypadku sprawdzianu uczeń może uzyskać po 20 punktów za zadania otwarte i 20 za zamknięte, a więc waga punktów za zadania o obu typach formatu wynosi (20:20). W przypadku egzaminu gimnazjalnego uczeń może uzyskać 20 punktów za zadania zamknięte i 30 za otwarte w części humanistycznej (waga 20:30) oraz po 25 punktów za oba typy zadań w części matematyczno-przyrodniczej

(waga 25:25). Jeśli zadanie jest określone jako otwarte, a jest tak naprawdę zadaniem zamkniętym, to zaburzona jest założona struktura testu i przyjęte proporcje punktacji.

2.5.4. Walidacja planu testu

Zauważyć można, że kolejne etapy tworzenia testu są rozwinięciem i doprecyzowaniem poprzednich. Od sformułowań ogólnych i teoretycznych przechodzi się stopniowo do sformułowań operacyjnych. Odpowiada to procedurze wszelkich badań empirycznych, gdzie od konceptualizacji problemu badawczego przechodzi się do jego operacjonalizacji, czyli opracowania procedur przeprowadzenia badania. Zachowanie ciągłości i logicznego powiązania kolejnych etapów badania, gwarantuje efektywne wykorzystanie wyników testu, zgodnie z założonymi celami.

Plan testu stanowi przewodnik dla opracowania i wyboru zadań do ostatecznej formy testu. Nie można założyć, że raz opracowany jest zadowalającej jakości. Plan testu powinien być poddany walidacji. Standardy jasno wskazują, że fundamentalną kwestią w procesie opracowania i walidacji testu jest jego trafność, czyli poparta argumentami zgodność testu z zamierzoną interpretacją jego wyników.

Dowody na trafność wg Standardów powinny opierać się na co najmniej pięciu źródłach, do których należą: (1) treści objęte testem, (2) procesy umysłowe zachodzące podczas rozwiązywania testu, (3) struktura testu, (4) związek z innymi kryteriami zewnętrznymi, (5) konsekwencje wykorzystania wyników testu.

Na etapie walidacji planu testu kluczowy zestaw argumentów wspierających trafność przyszłej interpretacji wyników testu powinien wywodzić się z analizy zdefiniowanego w podstawie testu modelu umiejętności, mierzonego przez test, na który składają się treści testu (wiedza) oraz procesy umysłowe zachodzące podczas rozwiązywania testu. Analiza taka odnosi się do trafności treściowej *(content validity)*. Trafność treściowa to stopień w jakim test odzwierciedla zakres materiału, który powinien mierzyć *(domain of content)* (Carmines i Zeller, 1979, s. 20).

Według Standardów dowody wywiedzione z analizy treści testu *(content-based evidence)* mogą zawierać logiczną, jak i empiryczną analizę stopnia odwzorowania (związku) mierzonej domeny (konstruktu) w treści (zawartości) testu. Ponieważ domena mierzona przez testy edukacyjne jest zwykle określona przez podstawę programową, przygotowywaną na poziomie krajowym, walidacja planu testu powinien opierać się na wypracowaniu argumentów na rzecz wsparcia tezy o adekwatnym odwzorowaniu podstawy programowej w treści testu.

Na konieczność dostarczenia dowodów powiązania testu z docelową domeną w zakresie wiedzy i procesów myślowych wskazuje standard 13.3: „Gdy test jest wykorzystywany jako wskaźnik osiągnięć w ramach określonego programu nauczania lub oceny stopnia opanowania określonych standardów nauczania, to należy dostarczyć danych świadczących o zakresie, w jakim test reprezentuje docelową domenę wiedzy i aktywizuje procesy umysłowe odpowiadające celom nauczania. Trzeba też wystarczająco dokładnie opisać zarówno obszar poddawany testowaniu, jak i obszar będący celem badania, tak by można było ocenić stopień ich powiązania. W takiej analizie powinno się wyraźnie wskazać te aspekty docelowo badanej domeny, które test odzwierciedla, jak i te aspekty, których nie odzwierciedla (Hornowska, 2007, s. 245).

Ponieważ zagadnienie trafności testów jest szerzej omówione w innym miejscu tego rozdziału, tu jedynie przypominamy, że dwoma największymi zagrożeniami dla trafności treściowej, które należy mieć na uwadze dokonując walidacji planu testu są niedoreprezentacja treściowa konstruktu *(construct underrepresentation)* oraz możliwość generowania podczas rozwiązywania testu wariancji niezwiązanej z konstruktem *(construct-irrelevant variance)*.

2.5.5. Budowanie zadań i pytań testowych

Kolejnym kluczowym etapem przygotowania testu (po wyznaczeniu celów, ustaleniu zakresu konstruktu poddanego pomiarowi, wypracowaniu planu testu i jego walidacji z perspektywy trafności treściowej), jest budowanie pytań i zadań testowych. Osoby odpowiedzialne za budowanie pytań i zadań testowych powinny być odpowiednio przeszkolone zarówno w zakresie ogólnych wytycznych do przygotowywania pytań, jak również w zakresie planu konkretnego testu, do którego ma powstać zestaw pytań.

Formaty zadań testowych dzielą się na najbardziej ogólnym poziomie na dwa rodzaje (Downing, 2009): zadania otwarte *(constructed-response,* CR) oraz zadania zamknięte *(selected-response,* SR). Zadania otwarte wymagają od ucznia samodzielnego wytworzenia odpowiedzi. Zadania zamknięte wymagają wyboru lepszej według ucznia odpowiedzi z ustalonej, skończonej listy możliwych odpowiedzi. W grupie zadań otwartych można wyróżnić zadania:

- krótkiej odpowiedzi *(short answer constructed-response,* KO) dla odpowiedzi nie dłuższych niż trzy zdania;
- rozszerzonej odpowiedzi *(long answer constructed-response,* RO) z tekstem nie dłuższym niż pięć stron.

W grupie zadań zamkniętych można wyróżnić:

- tradycyjne zadania zamknięte wielokrotnego wyboru *(multiple-choice item,* MCQ). Są to najczęściej zadania z jedną prawidłową lub najlepszą z możliwych odpowiedzią.
- złożone zadania zamknięte wielokrotnego wyboru *(complex multiple-choice,* Type K). W zadaniach tych dostępne są alternatywy typu „wszystkie podane odpowiedzi są poprawne", „żadna z podanych odpowiedzi nie jest poprawna", „dwie lub więcej z podanych odpowiedzi są poprawne".
- zadania zamknięte typu prawda/fałsz (*true-false*, TF).
- wielokrotne zadania zamknięte typu prawda/fałsz *(multiple true-false,* MTF). Ocena na skali prawda/fałsz nie następuje dla pojedynczego stwierdzenia, ale dla zestawu stwierdzeń powiązanych tematycznie.

W Polsce przyjęła się klasyfikacja proponowana przez Bolesława Niemierko, podsumowuje ją tabela 2.4.

Tabela 2.4. Formaty zadań testowych wg B. Niemierki

Rodzaje zadań	Zalety zadań	Typ zadania	Forma odpowiedzi	Czas rozwiązywania
Otwarte	- Nie sugerują odpowiedzi. - Pozwalają uczniowi na większą samodzielność.	zadanie krótkiej odpowiedzi	Zadanie wyraźnie ogranicza odpowiedź ucznia do niezbędnych informacji. Odpowiedź ta musi być zwięzła, konkretna, zgodna z poleceniem.	3 min.
		zadanie z luką	Zadanie wymaga wpisania w określonej kolejności i określonym miejscu terminów lub sformułowań.	1 min.
		zadanie dłuższej odpowiedzi	Zadanie wymaga dłuższej, logicznie skonstruowanej wypowiedzi.	10-15 min.
Zamknięte	- Stawiają ucznia w sytuacji wyboru, zmuszając do myślenia. - Wszechstronnie sprawdzają różne obszary wiedzy.	zadanie na dobieranie	Uczeń z podanych propozycji konstruuje odpowiedź.	4-5 min.
		zadanie wielokrotnego wyboru	Uczeń z podanych propozycji odpowiedzi wybiera właściwą.	1,5 min.
		zadanie typu „prawda-fałsz"	Uczeń określa prawdziwość stwierdzenia.	0,5 min.

Mimo, że pytania wielokrotnego wyboru są wciąż najbardziej popularne, ponieważ są efektywne na etapie prowadzenia i oceniania testu, to coraz częściej korzysta się z alternatywnych form pytań testowych. Badania nad nowymi formami pytań są obecnie prowadzone bardzo intensywnie. Ciekawe propozycje alternatywnych form zadań testowych zostały zaprezentowane np. w pracy pod redakcją Stevena Downinga "*Handbook of Test Development*" (Downing, 2006).

Do tworzenia zadań składających się na test odnosi się standard 3.6: „Rodzaj zadań testowych, format odpowiedzi, procedury oceny wyników oraz procedury badania testem powinny być zgodne z celami testu, mierzoną dziedziną oraz populacją, na której test ma być stosowany. Na ile to możliwe, treść testu powinna zostać tak dobrana, by wnioski wyciągane na podstawie wyników testowych były jednakowo trafne dla różnych kategorii osób, dla których test jest przeznaczony. Proces oceny testu powinien obejmować analizy empiryczne i – gdy to właściwe – skorzystanie z techniki sędziów kompetentnych do oceny zadań testowych i formatów odpowiedzi. Należy także podać udokumentowane kwalifikacje oraz opisać doświadczenie oraz cechy demograficzne sędziów kompetentnych" (Hornowska, 2007, s. 87).

W literaturze można znaleźć wiele przewodników dotyczących tworzenia zadań testowych. Ogólne wytyczne do opracowywania zadań testowych są dobrze opracowane, np. w pracach Thomasa Haladyny; zwłaszcza w publikacji, która zdobyła szerokie uznanie, a dotyczy tworzenia zadań wielokrotnego wyboru: "*Developing and Validating Multiple-choice Test Items*" (Haladyna, 2004). W innej pracy Haladyna i Downing (1989) podają klasyfikację zasad, według których powinny być konstruowane zadania zamknięte w teście. Oto najważniejsze z nich:

1. Każde zadanie powinno bazować na umiejętności i jej elementach, zidentyfikowanych na etapie definicji konstruktu (modelu umiejętności), który test ma mierzyć.
2. Należy wybrać jeden z dwóch typów formatów odpowiedzi – prośba o wskazanie odpowiedzi poprawnej lub najlepszej z możliwych.
3. Należy, jeśli to możliwe, unikać używania skomplikowanych formatów odpowiedzi, np. złożonych zadań zamkniętych wielokrotnego wyboru *(Type K)*.

4. Format odpowiedzi powinien być przedstawiony pionowo, a nie poziomo.
5. Czas czytania danego zadania powinien być w miarę możliwości minimalizowany.
6. Należy unikać „zadań-pułapek", których treść ma na celu zmylenie ucznia, co zwiększa prawdopodobieństwo udzielenia niepoprawnej odpowiedzi.
7. Słownictwo wykorzystane przy pisaniu zadania musi odpowiadać poziomowi edukacyjnemu uczniów, dla których przeznaczony jest test.
8. Należy unikać sytuacji, w której jedno zadanie zawiera wskazówki do poprawnego rozwiązania innego zadania. Każde zadanie powinno być możliwie niezależne od pozostałych zadań.
9. Należy unikać książkowych, dosłownych sformułowań przy pisaniu treści zadań.
10. Jeżeli zadanie zawiera w sobie tzw. „puste miejsce", które wymaga uzupełnienia, to nie należy tego „pustego miejsca" umieszczać na początku, albo w środku zdania, składającego się na treść zadania.
11. Należy upewnić się, że instrukcja w treści zadania jest jasna i słownictwo dokładnie wskazuje, co uczeń ma wykonać w ramach zadania.
12. Treść zadania, ani opcje odpowiedzi nie powinny zawierać negacji.
13. Opcje odpowiedzi powinny być niezależne i nie powinny się pokrywać.
14. Należy unikać opcji odpowiedzi „wszystkie z powyższych" i „żadna z powyższych".
15. Należy unikać sformułowań „nigdy" i „zawsze".
16. Opcja zawierająca poprawną odpowiedź powinna pojawiać się porównywalnie często między różnymi pozycjami (nie należy umieszczać poprawnej odpowiedzi na tej samej pozycji, np. zawsze/często opcja „C" jest prawidłowa).

Podobnie jak w przypadku opracowania planu testu, tak w przypadku budowania zadań testowych, nie należy zakładać, że są one idealne. Wymagają one rewizji i oceny. Nawet doświadczeni autorzy mogą tworzyć obciążone zadania testowe, czy popełniać błędy logiczne i formalne, np. brak poprawnej odpowiedzi wśród opcji odpowiedzi. Powstałe pytania i zadania testowe powinny zostać zweryfikowane zarówno pod kątem ich formalnych cech, które powinny być zgodne z ogólnymi wytycznymi opracowywania zadań testowych, jak również powinny dobrze wpisywać się i odpowiadać na wymogi planu testu.
Powiązanie określonych wymagań podstawy programowej, a w przeszłości standardów wymagań egzaminacyjnych, z konkretnymi zadaniami danego testu znaleźć można w tzw. sprawozdaniach udostępnianych na stronach CKE po przeprowadzeniu danego typu egzaminu w określonym roku.

2.5.6. Testowanie zadań

Mimo tego, że zadania wchodzące w skład testu zostały przygotowane zgodnie z rygorem metodologicznym i poddane ewaluacji przez ekspertów, nie można być pewnym, że dane zadanie prawidłowo funkcjonuje w realnym teście, jeśli nie zostanie ono przetestowane na próbie pochodzącej z populacji uczniów, tej samej do której docelowo kierowany będzie test. Dlatego też kolejnym etapem na drodze tworzenia testu jest badanie pilotażowe zadań testowych. Możliwość wyciągania wiarygodnych wniosków z takiego badania oczywiście w największym stopniu zależy od poprawnego i adekwatnego doboru próby uczniów.
Standard 3.8 stwierdza wprost, że: „Jeśli prowadzi się badania pilotażowe dotyczące zadań testowych czy testów, to należy podać kryteria doboru próbki/prób osób badanych oraz dane demograficzne charakteryzujące te próbki. Jeśli jest to możliwe, to próbki te powinny być reprezentatywne dla populacji osób, dla których test jest przeznaczony." (Hornowska, 2007, s. 88).

2.5.7. Zdefiniowanie procedur oceniania

W kolejnym kroku konstrukcji testu należy zdefiniować procedury punktowania odpowiedzi. Procedury punktowania zadań testowych powinny być określone i przedstawione osobom rozwiązującym test. W przypadku zadań zamkniętych punktowanie może sprowadzać się wyłącznie do zliczenia zadań z poprawnie udzieloną odpowiedzią. Wynik całkowity za zadanie może zawierać w sobie oczywiście punkty cząstkowe. Może wystąpić także sytuacja, w której różnym zadaniom jest nadawana różna waga, w związku z tym będą mieć różne znaczenie dla wyniku testowego. Na przykład, jeśli zadanie punktowane 0 (niepoprawna odpowiedź) lub 1 (poprawna odpowiedź) zostanie uznane za szczególnie ważne dla testu w kontekście badanej umiejętności można mu nadać wagę wyższą niż 1, np. 1,5. Przy poprawnej odpowiedzi zadanie to wniesie do całkowitego wyniku ucznia na teście 1,5 punktu zamiast 1 punktu. Niezależnie od tego, jaką procedurę punktowania zadań i całego testu wybrano, powinna być ona jasno określona.

Odnosi się do tego standard 3.22., który stwierdza: „Autor testu powinien jasno i wyraźnie przedstawić procedurę obliczania wyników testowych oraz – gdy to potrzebne – podać kryteria oceny, po to, by zwiększyć jej dokładność. Należy też dać wyraźne wskazówki na temat posługiwania się skalami szacunkowymi czy też dotyczące obliczania wyników ustrukturyzowanych wypowiedzi za pomocą kodowania, skalowania czy klasyfikowania. Jest to szczególnie ważne, gdy testy są oceniane lokalnie." (Hornowska, 2007, s. 93).

W przypadku zadań otwartych, większa część z nich wymaga oceny przez człowieka, opierającej się na osądzie spełnienia przez ucznia określonych kryteriów w zadaniach. W związku z tym dobór oceniających, ich trening, kwalifikacje i monitorowanie ich działań jest kluczowe w kontekście uzyskania nieobciążonych wyników.

Standard 3.23 precyzuje: „Autor testu powinien opisać proces wyboru, ćwiczenia oraz niezbędne kwalifikacje osób oceniających wyniki testowe. Materiały ćwiczeniowe, takie jak schematy oceniania, przykłady ocenionych na różnym poziomie odpowiedzi oraz procedury treningowe dla osób oceniających, powinny zwiększać stopień zgodności między sędziami i w ten sposób przyczyniać się do interpretacji wyników zgodnie z założeniami autora testu. Osoby odpowiedzialne za trening w zakresie stosowania testu powinny określić rzetelność sędziujących oraz ewentualne odchylenia od standardów oceniania." (Hornowska, 2007, s. 94).

Procedury punktowania zadań w teście nabierają szczególnego znaczenia w przypadku egzaminów, które mają określone progi zdawalności. Ramka 2.6 zawiera informację o sposobach definiowania progów zdawalności i uwagi dotyczące progu funkcjonującego w Polsce w ramach egzaminu maturalnego.

Ramka 2.6. Sposoby definiowania progów zdawalności i opis progu w polskiej maturze

CZY WIESZ, ŻE?

W Polsce próg zdawalności jest wyznaczony dla egzaminu maturalnego w części podstawowej i wynosi 30% punktów możliwych do uzyskania w egzaminie. Wysokość progu nie jest ustalana w odniesieniu do określonego zakresu wymagań: nie można uznać, że próg zdawalności matury wynika z podstawy programowej. Choć z założenia egzamin maturalny był pomyślany jako forma testowania opartego na kryterium (minimalne kompetencje określone w podstawie programowej, które powinien posiadać absolwent szkoły ponadgimnazjalnej). Próg zdawalności jest jednak arbitralny, gdyż jego wysokość wynosząc 30%, niezmienna między poszczególnymi latami, nie jest uzasadniona w aktach prawnych i planu testu oraz nie bierze pod uwagę zmieniającej się trudności arkuszy (30% punktów nie jest porównywalne między latami).

Warto zauważyć, że dodatkowo próg ten zmieniał się w czasie, np. wskutek obniżenia go z 40% w 2002.

2.5.8. Opracowanie instrukcji testowania

Integralnym elementem przygotowania testu jest specyfikacja procedur przeprowadzenia (administracji) testu. Stosowanie jednolitych procedur przeprowadzenia testu jest kluczowe dla przeprowadzenia standaryzowanego pomiaru.

Na tym etapie należy jasno określić czas przeznaczony na rozwiązywanie testu oraz określić warunki stanowiące o wyjątkowym traktowaniu osób rozwiązujących test (np. zwiększony czas na rozwiązywanie testu dla osób o konkretnym typie zaburzeń). Test powinien zawierać jasną instrukcję, która wskazuje uczniowi czego się od niego/niej oczekuje. Warto przedstawić także informacje o zawartości testu, formacie zadań, procedurze oceniania.

Standard 3.20 określa „Instrukcja przedstawiona osobom badanym powinna być na tyle szczegółowa, aby odpowiadały one w sposób założony przez autora testu. Jeśli jest to konieczne, to respondenci zanim zaczną odpowiadać na pozycje testowe, powinni się zapoznać z przykładami materiałów, ćwiczeń zadań czy pytań, kryteriami oceny oraz przykładem zadania testowego typowego dla podstawowego obszaru, czy dziedziny mierzonej przez test. Przykłady tego typu mogą się również znaleźć w materiałach testowych jako element standardowej instrukcji badania testem." (Hornowska, 2007, s. 92).

Standard 3.21 dodaje, że „Jeżeli autor testu dopuszcza możliwość wprowadzenia zmian do procedury badania testem, zarówno w wypadku jednostek, jak i grup osób badanych, to powinien określić, jaki jest zakres dopuszczalnych zmian w procedurze badania oraz uzasadnić swoje stanowisko." (Hornowska, 2007, s. 93).

2.5.9. Analiza zadań testowych

Następnym etapem na drodze przygotowania testu, jest analiza zadań testowych, na podstawie wyników uzyskanych w badaniu pilotażowym testu. Analiza zadań testowych spełnia kilka różnych funkcji. Pozwala zidentyfikować zadania, które najwięcej wnoszą dla testu z punktu widzenia jego rzetelności. Umożliwia także wykrycie zadań, które w jakikolwiek sposób dyskryminują określonych kategorii zdających, np. chłopców względem dziewcząt lub odwrotnie.

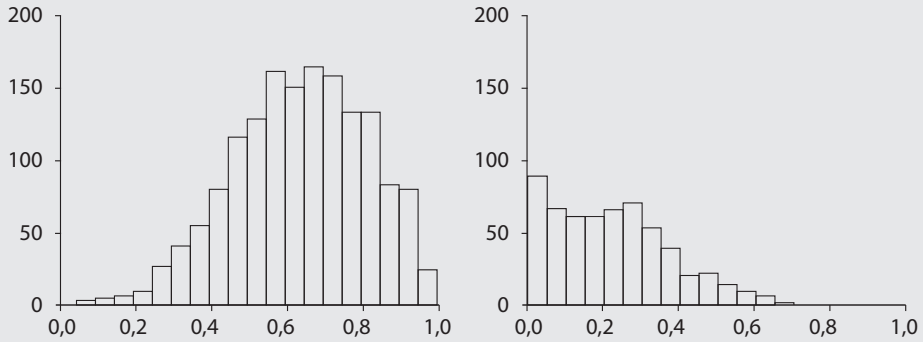
Standardy szczegółowo precyzują, że identyfikacja zadań potencjalnie obciążonych lub nie wykazujących się pożądanymi właściwościami psychometrycznymi powinna zostać szczegółowo opisana.

Standard 3.9 określa, że „Jeśli autor testu dokonuje oceny właściwości psychometrycznych zadań testowych, to należy opisać, jaki wykorzystuje w tym celu model klasycznej teorii testów czy model teorii odpowiedzi na pozycje testowe (IRT). Należy także opisać badane próby oraz podać, czy ich wielkość i zróżnicowanie były odpowiednie do zastosowanych procedur. Trzeba także przedstawić procedurę selekcji zadań testowych oraz wykorzystane dane, takie jak trudność zadań, ich moc dyskryminacyjna i/lub funkcje informacji o zadaniach testowych. Jeżeli do wyznaczenia psychometrycznych właściwości zadań testowych wykorzystuje się teorię odpowiedzi na pozycje testowe, to należy opisać model IRT, procedury szacowania wartości oraz stopień dopasowania modelu do danych." (Hornowska, 2007, s. 88-89). Ramka 2.7. wskazuje na wyniki analizy jednej z właściwości zadania: łatwości/trudności.

Ramka 2.7. Łatwość zadań wielokrotnego wyboru w sprawdzianie i egzaminie gimnazjalnym

CZY WIESZ, ŻE?

Prawie co piąte pytanie wielokrotnego wyboru w polskich egzaminach zewnętrznych jest na tyle proste, że prawidłowej odpowiedzi jest w stanie udzielić 4 na 5 zdających. W analizie uwzględniono wszystkie pytania wielokrotnego wyboru ze sprawdzianów oraz egzaminów gimnazjalnych od 2002 do 2014 (łącznie ponad 1500 pytań). Wykres po lewej stronie prezentuje liczbę pytań wielokrotnego wyboru (oś pionowa) wraz z odsetkiem zdających, którzy udzielili poprawnej odpowiedzi (oś pozioma). Jak widać w polskich egzaminach zewnętrznych dominują łatwe pytania.



Wykres po prawej stronie prezentuje liczbę pytań wielokrotnego wyboru (oś pionowa), które pojawiły się na egzaminach gimnazjalnych od 2002 do 2014 roku (łącznie ponad 750 pytań) wraz z wartościami parametrów „zgadywania" (oś pozioma). Im wyższa wartość parametru „zgadywania" tym łatwiej można odganąć prawidłową odpowiedź mimo niskiego poziomu umiejętności. Najczęściej liczba możliwych odpowiedzi, spośród których wybiera uczeń, wynosi 4; tak więc prawdopodobieństwo odgadnięcia prawidłowej odpowiedzi nie powinno przekraczać 1/4. Odsetek zadań wielokrotnego wyboru w polskich egzminach gimnazjalnych, dla których wartość parametru „zgadywania" przekracza 1/4 wynosi ponad 40%.

Jednym z kluczowych elementów analizy danych z badań pilotażowych powinna być analiza zróżnicowanego funkcjonowania zadania (*differential item functioning, DIF*). Analiza ta wynika z potrzeby respektowania i zapewnienia tzw. sprawiedliwości testowania (*test fairness*). W pomiarze edukacyjnym osiągnięć uczniów warunkiem koniecznym, ale niewystarczającym jest stosowanie narzędzi sprawiedliwych dla wszystkich uczniów poddawanych tej procedurze (Langenfeld, 1997). Pomiar w edukacji jest niesprawiedliwy, jeśli jest stronniczy, a więc systematycznie nie reprezentuje niektórych osób czy grup w zakresie mierzonego konstruktu (Kane, 2010). Zainteresowanie sprawiedliwością testowania zaczęło wzrastać wraz z wpływami ruchów na rzecz praw obywatelskich w latach 60-tych i ruchami na rzecz praw kobiet (Cole i Zieky, 2001). Wczesna definicja sprawiedliwości testowania Cleary (1968) zakładała, że test jest niesprawiedliwy, jeśli systematycznie przeszacowuje, lub niedoszacowuje umiejętności uczniów pochodzących z różnych grup społecznych. Możliwe obciążenie testu osiągnięć, polegać może na tym, że różnice w wyniku testu pomiędzy różnymi sub-populacjami, np. kobietami i mężczyznami będą odzwierciedlały nie tylko różnice w poziomie mierzonej testem umiejętności, ale będą też wynikiem obciążenia kulturowego, nieodłącznie związanego z pomiarem (Angoff, 1993).

Oczywiście zapewnienie sprawiedliwego, tj. bezstronnego testowania jest zasadą przekrojową i odnosi się do wszystkich etapów tworzenia testu. W kontekście analizy danych z badań pilotażowych testu odnosi się jednak do detekcji zadań funkcjonujących w sposób zróżnicowany w porównywanych grupach. Zadanie funkcjonuje w sposób zróżnicowany w różnych grupach, jeśli to czy uczeń poprawnie na nie odpowie (prawdopodobieństwo poprawnej odpowiedzi na zadanie),

zależy nie tylko od poziomu jego/jej umiejętności, ale od innych cech ucznia, np. płci, pochodzenia etnicznego, pochodzenia społecznego itd. (Ironson,1982; Linn, Levine, Hastings i Wardrop, 1981). Na przykład, statystyczna analiza danych może wykazać, że określone zadanie funkcjonuje różnie w grupie dziewcząt i chłopców. Jeśli zadanie jest trudniejsze dla dziewcząt niż chłopców, to odsetek poprawnych odpowiedzi na to zadanie będzie wyższy wśród chłopców niż dziewczynek. Decyzja o wykluczeniu danego zadania z testu nie może jednak opierać się na wniosku, że zadanie różnie funkcjonuje w porównywanych grupach. Konieczne jest uzupełnienie analizy statystycznej, ewaluacją i oceną ekspercką, która odpowie na pytanie co jest przyczyną innego funkcjonowania zadania w porównywanych grupach. Jeśli przykładowo zadanie z matematyki jest łatwiejsze dla chłopców, ponieważ treść zadania jest bliższa doświadczeniom chłopców (np. treść odnosi się do obliczenia pojemności silnika samochodowego) to zadanie może być obciążone. W tym przykładzie odpowiedź na zadanie wymagające przeprowadzenia obliczeń zależy nie tylko od umiejętności matematycznych, ale także od doświadczeń specyficznych dla grupy chłopców. Innym przykładem może być zadanie przyrodnicze, wymagające np. identyfikacji na podstawie rysunku gatunku zboża. Możliwe, że zadanie takie będzie łatwiejsze dla dzieci pochodzących z terenów wiejskich, w porównaniu do dzieci pochodzących z miasta. W tym wypadku zadanie nie tylko odzwierciedla poziom umiejętności przyrodniczych uczniów, ale także czynnik miejsca pochodzenia ucznia.

Potrzeba prowadzenia analiz zróżnicowanego funkcjonowania zadania, jako podstawowego elementu i pierwszego kroku dla zapewnienia bezstronności testowania znajduje dzisiaj odzwierciedlenie nie tylko w Standardach, ale także odrębnych wytycznych dotyczących zapewnienia sprawiedliwości testowania opracowanych przez Joint Committee on Testing Practices (1998, 2004) i Educational Testing Service (2002).

Standard 7.3 jasno wskazuje, że „Jeśli wiarygodne wyniki badań wskazują na to, że w grupach zdefiniowanych ze względu na wiek, płeć, rasę, pochodzenie etniczne, kulturę, niepełnosprawność czy język mamy do czynienia ze zróżnicowanym funkcjonowaniem zadań testowych spowodowanym zakresem treści mierzonym przez test, to autorzy testów powinni – w miarę możliwości – przeprowadzić odpowiednie badania w tym obszarze. Badania takie powinny umożliwić wykrycie i wyeliminowanie tych elementów związanych z tworzeniem testu, jego trafnością i formatem, które mogą obciążać wyniki testowe otrzymane przez określone grupy osób badanych." (Hornowska, 2007, s. 149).

Zagadnienie sprawiedliwości testowej ze względu na przynależność do różnych grup mniejszościowych jest w centrum zainteresowania w krajach o zróżnicowanej populacji, takich jak Stany Zjednoczone. Zagadnienie sprawiedliwej administracji testu dla różnych grup mniejszościowych jest także jednym z obszarów w najnowszym wydaniu Standardów (2014), które zostało najbardziej rozbudowane i zmodyfikowane. Ramka 2.8 prezentuje wyniki polskich badań dotyczących zróżnicowanego funkcjonowania zadań.

Ramka 2.8. Wyniki badań nad zróżnicowanym funkcjonowaniem zadań ze względu na wersję testu

CZY WIESZ, ŻE?

Analiza zróżnicowanego funkcjonowania zadań testowych (differential item functioning, DIF) nie jest w Polsce wbudowana w system oceny zadań egzaminacyjnych i wyników egzaminów zewnętrznych. Badania dotyczące potencjalnego obciążenia zadań testowych ze względu na wersję testu prowadzili Koniewski, Majkut i Skórska (2014). Arkusze A i B w polskim systemie egzaminacyjnym nie różnią się treścią zadań testowych a kolejnością możliwych odpowiedzi, co ma zapobiegać ściąganiu (Szaleniec, 2006). Wyniki badań pokazały, że zmiana kolejności możliwych odpowiedzi między arkuszami może obniżać prawdopodobieństwo udzielenia prawidłowej odpowiedzi, w sytuacji gdy prawidłowa odpowiedź dla analizowanego zadania oraz zadań go poprzedzającego i następującego jest oznaczona zawsze tym samym symbolem, np. A, A, A. Taka konstrukcja arkusza jest niezgodna z ostatnim zaleceniem tworzenia zadań wskazanym wcześniej w tym rozdziale, pochodzącym z publikacji Haladyny i Downinga (1989).

Możliwe wyjaśnienie tego faktu zakłada, że uczniowie widząc, że w kilku zadaniach pod rząd poprawna jest zawsze odpowiedź np. A, uznają taką sytuację za mało prawdopodobną. Widząc taki wzorzec odpowiedzi mogą wracać do zadania trudniejszego, w którym nie mieli pewności co do odpowiedzi i zmieniać odpowiedź na błędną, choć w ich ocenie bardziej prawdopodobną.

W IBE prowadzono także analizy zróżnicowanego funkcjonowania zadania ze względu na płeć ucznia. Analizy zadań ze sprawdzianu z 2011 roku ujawniły, że część zadań funkcjonuje inaczej w grupie chłopców i dziewcząt. Po wstępnej diagnostyce statystycznej DIF, należy ekspercko zweryfikować treść „podejrzanych" zadań testowych. W ramach polskiego systemu egzaminów zewnętrznych, analizy takie można wykonywać ze względu na to, że systemowo gromadzone są informacje o uczniach, np. płeć, dysleksja.

2.5.10. Wybór zadań testowych do ostatecznej formy testu

Kolejny etap polega na ostatecznym wyborze zadań składających się na test. Decyzja o wyborze zadania do testu, powinna bazować na komplementarnych dowodach dwojakiego rodzaju:

- analizie statystycznej – która może wskazać zadania pozbawione odpowiednich właściwości psychometrycznych i funkcjonujące różnie w różnych sub-populacjach;
- analizie eksperckiej – która ocenia treściowe właściwości zadania, a więc bierze pod uwagę wyniki jednego z poprzednich etapów tworzenia testów, tj. rewizji, przeglądu i ewaluacji zadań testowych.

Decyzja o wykluczeniu zadania w teście nie powinna być podejmowana wyłącznie na podstawie dowodów jednego z tych dwóch rodzajów analiz.

Standard 3.10 stwierdza, że „Jeżeli pozycje testowe są dobierane praktycznie wyłącznie na podstawie związków empirycznych, a nie na podstawie założeń teoretycznych, to autorzy testów powinni przeprowadzić krzyżowe badania walidacyjne. Należy podać zakres, w jakim te same zbiory zadań testowych zostały otrzymane w różnych badaniach" (Hornowska, 2007, s.89).

2.5.11. Wybór metody skalowania, raportowania wyników, walidacja testu

Przy przeprowadzeniu testu konieczne jest określenie sposobu skalowania i sposobu raportowania wyników. Głównym celem skalowania wyników jest doprowadzenie wyników do takiej formy, która będzie umożliwiała ich interpretację. W teście różnicującym (*norm-referenced test*), którego wyniki są relatywne w ramach konkretnej zbiorowości, można zastosować percentyle, co umożliwia odniesienie wyniku testu konkretnego ucznia do wyników reszty uczniów. W przypadku testów

sprawdzających (*criterion-referenced test*) wynik ucznia może zostać sklasyfikowany do jednej z kilku kategorii, np. poziom podstawowy, poziom średni, poziom zaawansowany. Niezależnie od sposobu klasyfikowania wyniku ucznia, autorzy testu powinni jasno go opisać i uzasadnić. Mówi o tym standard 4.1: „Opis testu powinien zawierać zarówno jasne wyjaśnienie znaczenia przyjętej interpretacji skali wyników przeliczonych, jak i opis jej ograniczeń." (Hornowska, 2007, s. 105).

Standard 4.10 dodaje „Jeżeli się przyjmuje, że wyniki testowe otrzymane w różnych formach tego samego testu można stosować zamiennie, to założenie to należy jasno uzasadnić i przedstawić odpowiednie dane. W niektórych przypadkach istnieje możliwość dostarczenia danych wprost dotyczących równoważności wyników testowych. W innych sytuacjach podanie danych może polegać na wykazaniu, że założenia teoretyczne leżące u podstaw procedury określającej możliwość porównywania wyników zostały spełnione. Określone uzasadnienie oraz wymagane dane zależą częściowo od planowanych sposobów wykorzystania testu zakładających równoważność wyników testowych." (Hornowska, 2007, s. 109).

Ponieważ żadne narzędzie pomiaru nie jest całkiem pozbawione błędu pomiarowego, każdorazowo należy prowadzić badania rzetelności testu. Więcej o sposobach badania rzetelności można przeczytać w poprzednich częściach opracowania. Tu warto jedynie podkreślić, że określenie błędu pomiaru narzędzia stanowi integralną część procesu jego tworzenia.

Standard 2.1 podkreśla, że „Należy określić rzetelność, standardowy błąd pomiaru lub funkcję informacji dla testu w wypadku każdego wyniku ogólnego, wyniku cząstkowego lub wyniku złożonego, dla których jest przewidziana określona interpretacja." (Hornowska, 2007, s. 66).

Oczywiście wybór aspektów rzetelności do badania testu zależy od celów i przeznaczenia testu. Na przykład, jeśli celem będzie uzyskanie porównywalnych między latami wyników egzaminacyjnych autorom testu może zależeć zwłaszcza na badaniu spójności wyników uzyskiwanych z roku na rok. Niezależnie od tego, jak jest badana rzetelność testu, odpowiednie statystyki powinny być raportowane.

Standard 2.2. podkreśla, że „Należy przedstawić dane dotyczące standardowego błędu pomiaru, zarówno całkowitego, jak i warunkowego, jeżeli ma on znaczenie, zarówno w terminach wyników surowych (czy wyrażonych na wyjściowej skali), jak i na skali wyników przeliczonych, zalecanej dla określonej interpretacji tych wyników." (Hornowska, 2007, s. 66).

Ważnym aspektem walidacji testu jest dostarczenie dowodów jego trafności, zwłaszcza wywiedzionych z oceny struktury testu, związku z innymi kryteriami zewnętrznymi, konsekwencji wykorzystania wyników testu. W sytuacji, kiedy raportuje się punktację nie tylko za cały test, ale także za rozwiązanie wiązek zadań testowych (grup zadań dotyczących konkretnej umiejętności), konieczne jest dostarczenie dowodów na rzecz trafności interpretacji takich wyników cząstkowych.

Standard 1.3. stwierdza, że „Jeżeli trafność typowej lub prawdopodobnej interpretacji wyników testowych nie została potwierdzona lub proponowana interpretacja nie jest zgodna z istniejącymi danymi, to fakt ten powinien zostać jasno wskazany, a potencjalnych użytkowników należy ostrzec przed nieuprawnionymi interpretacjami." (Hornowska, 2007, s. 45).

Ostatnim etapem przygotowania testu jest dostarczenie procedur interpretacji wyniku. Wynik testu nie może być prezentowany w izolacji. Należy wskazać właściwości określonego wyniku testowego łącznie z wielkością błędu pomiaru. Do wyniku testowego należy przygotować i dostarczyć materiały informacyjne wskazujące na to, w jaki sposób można interpretować wynik i w jaki sposób prawidłowo go używać.

Standard 5.10 podkreśla, że „Jeżeli wyniki testowe są udostępniane studentom, rodzicom, urzędnikom, nauczycielom, klientom czy mediom, to osoby odpowiedzialne za badania testami powinny opracować odpowiednie ich interpretacje. Tworząc te interpretacje, należy używać prostego języka, opisywać dokładnie to, co test mierzy, co oznaczają wyniki testowe, dokładność wyników testowych, jakie są najczęściej popełniane błędy interpretacyjne oraz w jaki sposób wykorzystuje się wyniki testowe." (Hornowska, 2007, s. 123).

Kierunki rozwoju pomiaru edukacyjnego

Między czwartym (1999) a piątym (2014) wydaniem standardów minęło 15 lat. Najnowsze wydanie Standardów uwzględnia ważne aspekty rozwoju dziedziny testów, które dokonały się na przestrzeni tych lat. Najważniejsze zmiany dotyczą następujących obszarów związanych z pomiarem edukacyjnym:

- Rosnące znaczenie testów w tworzeniu polityki edukacyjnej państwa (możliwość traktowania wyników jako dowodów będących podstawą podejmowanych decyzji);
- Upowszechnianie idei dostępności testów dla wszystkich egzaminowanych (zwłaszcza tych o specjalnych potrzebach, np. niewidomi);
- Konieczność uwzględnienia rosnącej roli technologii informacyjno-komunikacyjnej.

Najważniejsze osiągnięcia w rozwoju dyscypliny mimo, że upowszechnione dopiero wraz z najnowszym wydaniem Standardów, już dawno są codzienną praktyką wiodących instytucji i firm świadczących usługi testowania. Ich początki sięgają jeszcze wcześniej. Ronald Hambleton we wstępie do książki *„Linear Models For Optimal Test Design"* Wima van der Lindena (2005) pisze o czterech krokach milowych w rozwoju dziedziny testów. Pierwszym był rozwój pomiaru opartego o kryterium *(criterion-reference testing)* zapoczątkowany w późnych latach 60. XX w. w pracach Roberta Glasera i Jamesa Pophama. Drugim krokiem milowym na przestrzeni lat 40. i 70. był rozwój teorii odpowiedzi na pozycje testowe (IRT) i zastąpienie przez nią klasycznej teorii testu (KTT). Trzecim ważnym momentem rozwojowym dyscypliny było zapoczątkowane w latach 70. odejście od przeprowadzania testów papierowych na rzecz komputerowych. Najnowszym wspomnianym przez Hambletona momentem zwrotnym w rozwoju dziedziny testów było wprowadzenie w latach 80. pierwszych prób automatycznego tworzenia testów, za pomocą algorytmów, które dobierają zadania testowe z banków zadań według pożądanych przez badacza własności testowych.

Dostępność w systemie egzaminacyjnym sprawnie administrowanego banku zadań daje nie tylko możliwości implementacji automatycznego generowania testów, ale także porównywalność pionową i poziomą wyników, generowanie wielu form równoległych tych samych testów, administrowanie komputerowe testów (w tym testy automatycznie adaptujące się do poziomu umiejętności uczniów).

2.6. Praktyka przygotowywania i przeprowadzania egzaminów w polskim systemie egzaminacyjnym

2.6.1. Procedury tworzenia arkuszy

Praktyka przygotowania i przeprowadzania egzaminu zmieniała się w ciągu kilkunastu lat obowiązywania w Polsce zewnętrznego systemu egzaminów. W CKE w latach 2009-2014 obowiązywało dziewięć różnych procedur przygotowywania materiałów egzaminacyjnych. W początkowym okresie przygotowaniem arkuszy egzaminacyjnych zajmowały się autorskie zespoły ze wszystkich OKE do każdego egzaminu. CKE dokonywała wartościowania i wyboru arkuszy z poszczególnych komisji do sesji głównej, dodatkowej i zapasowych zestawów do stosowania w sytuacjach krytycznych[8]. W pierwszych latach egzaminów zewnętrznych zbiór zasad przygotowania narzędzi egzaminacyjnych zapisany był bardzo ogólnie i daleko odbiegał od wymagań stosowanych na świecie i opisanych w Standardach dla testów stosowanych w psychologii i pedagogice (Hornowska, 2007). Główny nacisk położony był na procedury administracyjne dotyczące zachowania tajemnicy egzaminacyjnej podczas próbnego zastosowania zadań, recenzji nauczycielskich i akademickich. Przed zmianą podstawy programowej kształcenia ogólnego w 2008 r. wyprowadzone z podstawy programowej standardy wymagań egzaminacyjnych dla każdego egzaminu określały konstrukt

[8] Na przykład w wypadku unieważnienia egzaminu dla całej szkoły lub grupy uczniów z powodu uchybień organizacyjnych w przeprowadzeniu egzaminu.

definiujący umiejętności, które były przedmiotem sprawdzania na egzaminie[9]. Pierwsze sformułowanie Standardów wymagań egzaminacyjnych wprowadzone zostało rozporządzeniem ministra edukacji narodowej[10] i stanowiły one układ odniesienia do analizy trafności dla autorów testów, jak i dla recenzentów (ekspert akademicki i ekspert nauczycielski). W kolejnych latach procedury były doprecyzowane zarówno na poziomie ogólnym, obowiązującym dla wszystkich egzaminów, jak i w części szczegółowej odnoszącej się do konkretnych egzaminów. Kolejne edycje procedur ustalane były w zespole dyrektorów OKE i CKE. Opracowane i przyjęte przez wymieniony zespół procedury w 2005 roku obejmowały już stosunkowo szeroki zbiór zasad ze specyfikacją parametrów psychometrycznych określonych na podstawie zastosowania zestawów egzaminacyjnych nazywanych w OKE i CKE standaryzacją. Podane były też specyfikacje dla przygotowania planu i kartoteki testu. Kolejne uszczegółowienie procedur przygotowywania narzędzi egzaminacyjnych to procedury obowiązujące od 2007 roku[11].

W procedurze określona była liczba zadań i maksymalna liczba punktów z podziałem na zadania zamknięte i otwarte oraz z podziałem na poszczególne standardy wymagań egzaminacyjnych, przedział łatwości dla całego testu i podtestów sprawdzających umiejętności w zakresie poszczególnych standardów wymagań egzaminacyjnych.

W kolejnych latach następowały zmiany (ustalane dla każdego roku) polegające na ograniczeniu liczby przygotowywanych arkuszy egzaminacyjnych w OKE. Poszczególne komisje w parach specjalizowały się w pracach nad arkuszami egzaminacyjnymi do poszczególnych egzaminów. Stosowane były też rozwiązania polegające na przygotowywaniu testów w ogólnopolskich zespołach. Od 2007 roku do matematyki arkusze egzaminacyjne przygotowywane były przez Centralny Zespół Ekspertów Matematycznych (CZEM) pracujący w CKE w ramach projektu EFS, złożony zarówno z przedstawicieli poszczególnych OKE, pracowników wyższych uczelni oraz nauczycieli praktyków. Podobnie w latach 2010-2013 do sprawdzianu arkusze przygotowywane były przez ogólnopolski zespół złożony z przedstawicieli poszczególnych okręgowych komisji, którym koordynował dyrektor OKE we Wrocławiu.

Kolejnym, już bardzo rozbudowanym dokumentem są wdrożone procedury przygotowania testów egzaminacyjnych do egzaminu począwszy od 2015 roku[12]. Zostały w nim rozszerzone i uszczegółowione zasady standaryzacji (próbnego zastosowania), wymagania dotyczące recenzowania przez ekspertów w zakresie trafności testów oraz zakres analiz statystycznych. W tym dokumencie zawarto też szczegółowy harmonogram pracy nad testem w ciągu dwóch lat obejmujący wszystkie etapy przygotowywania i walidacji testu. Założono też, że do przygotowania arkuszy egzaminacyjnych mogą być wykorzystywane zadania stworzone w ramach projektu CKE-EFS Budowa banków zadań. Zasady te, zastosowane już do egzaminu w 2015 roku, stanowią istotny krok w doskonaleniu zewnętrznego systemu egzaminów w Polsce w zakresie profesjonalizacji pracy nad testami egzaminacyjnymi.

Warto jednak podkreślić, że pomimo wprowadzanych zmian, stosowane w polskim systemie egzaminacyjnym rozwiązania nadal odbiegają od standardów obowiązujących na świecie i wymagają dalszego doskonalenia a może nawet radykalnych przedsięwzięć. Konieczne jest wzmocnienie CKE o ekspertów w dziedzinie psychometrii i współczesnej statystyki wykorzystującej teorię odpowiedzi na pozycje testowe - IRT (Item Response Theory) a także wprowadzenie rozwiązań, które przy zachowaniu tajemnicy zadań egzaminacyjnych umożliwiłyby jednak pełną standaryzację testów

---

[9] Standardy wymagań egzaminacyjnych będące podstawą przeprowadzenia sprawdzianu w ostatnim roku nauki w szkole podstawowej. http://archiwum.cke.edu.pl/images/stories/Standardy/masowe_spr.pdf.

[10] Rozporządzenie Ministra Edukacji Narodowej z dnia 21 lutego 2000 r. w sprawie standardów wymagań będących podstawą przeprowadzania sprawdzianów i egzaminów, http://bip.men.gov.pl/men_bip/akty_pr_1997-2006/rozp_52.php?wrapper=test

[11] Przygotowanie propozycji pytań, zadań i testów do przeprowadzenia sprawdzianu i egzaminu gimnazjalnego. CKE, Warszawa 2007.

[12] Procedury przygotowania zadań i testów oraz ustalania zasad do przeprowadzenia sprawdzianu i egzaminów: gimnazjalnego maturalnego oraz eksternistycznych w 2015 roku. CKE, Warszawa 2013.

egzaminacyjnych, co przy obecnej praktyce jest nadal niemożliwe. Obiecujące są zamierzenia CKE wprowadzenia tworzenia arkuszy z wykorzystaniem banków zadań.

2.6.2. Przeprowadzenie egzaminu

Egzaminy przeprowadzane są w terminach ustalonych przez dyrektora Centralnej Komisji Egzaminacyjnej odrębnie dla każdego roku i ogłaszane na stronie internetowej CKE. Egzaminy odbywają się w podobnych terminach w kolejnych latach. Przestrzeganie stałych terminów zapewnia porównywalny między latami czas przeznaczony na edukację w poszczególnych latach, w miesiącach poprzedzających egzamin.

W przypadku sprawdzianu, egzaminu gimnazjalnego i maturalnego uczniowie zasiadają do egzaminu w szkole, do której uczęszczali. Dystrybucja materiałów egzaminacyjnych do szkoły organizowana jest centralnie, zgodnie z zamówieniami na arkusze egzaminacyjne przesyłanymi wcześniej przez szkoły do OKE. W zespole nadzorującym egzamin oprócz zespołu nadzorującego powołanego przez przewodniczącego uczestniczy także obserwator zewnętrzny (począwszy od roku 2003). Sposób przeprowadzenia, nadzorowania egzaminu i zabezpieczenia arkuszy przed nieuprawnionym ujawnieniem jest regulowany jednolitymi procedurami CKE. Przekazywanie prac do OKE i do oceny regulowane jest procedurami na poziomie OKE i w tym zakresie rozwiązania się różnią. Wstępują dwa rozwiązania. Przewodniczący szkolnego zespołu egzaminacyjnego przekazuje (w ustalony przez dyrektora OKE sposób) prace do okręgowej komisji egzaminacyjnej, gdzie następuje wprowadzenie ich do ewidencji i rozdział do ośrodków oceniania lub odbiór i ewidencjonowanie prac po egzaminie prowadzone jest przez pracowników OKE bezpośrednio w ośrodkach, gdzie prowadzone jest ocenianie.

Organizacja oceniania

Wyłączając egzamin z historii i WOS egzaminu gimnazjalnego oraz z przedmiotów przyrodniczych, dla wszystkich pozostałych, arkusze zawierają zadania otwarte wymagające przydzielenia określonej punktacji przez egzaminatora zewnętrznego. 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 do 2014 roku (włącznie) uczniowskie prace egzaminacyjne oceniane były w całym kraju wyłącznie w ośrodkach egzaminacyjnych. W 2015 roku w czterech OKE egzamin gimnazjalny z matematyki oceniany był w technologii e-oceniania z wykorzystaniem Internetu i własnych (domowych) komputerów egzaminatorów. W e-ocenianiu egzaminatorzy pracujący w rozproszeniu tworzą wirtualne zespoły nadzorowane przez PZE z wykorzystaniem dedykowanego oprogramowania.

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[13] 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

[13] Obecnie przesyłają elektronicznie kopie prac.

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 w tym za organizację podwójnego oceniania około 10 procent prac przydzielonych do oceniania w danym zespole.

2.7. Komunikowanie wyników

2.7.1. Elementy składowe procesu komunikowania wyników

Komunikowanie wyników[14] jest jednym z kluczowych etapów procesu egzaminacyjnego. Najistotniejszym rezultatem egzaminu dla wszystkich nim zainteresowanych, choć przede wszystkim dla samych zdających i szkół, w których się on odbywał, jest właśnie wynik komunikowany po egzaminie. Najbardziej istotne dla procesu komunikowania wyników wydają się być cztery elementy dotyczące przekazywanych informacji:

1. odbiorca;
2. zakres;
3. forma;
4. cel.

Każdy komunikat o wynikach egzaminacyjnych może być rozpatrywany w odniesieniu do tych elementów. Pozwala to na identyfikację najważniejszych składowych całego procesu komunikowania wyników. Co istotne, poszczególne elementy tego procesu są ze sobą w ścisłej relacji. Myślenie o informacji o wynikach egzaminacyjnych można sprowadzić zasadniczo do pytań o to, komu, jaką, w jakiej formie i do czego potrzebną informację należy przekazać. Wskazanie odbiorcy informacji pozwala na określenie zakresu komunikatu oraz formy, w jakiej zostanie ona przekazana. Konieczna do tego jest także wiedza na temat sposobu wykorzystania danej informacji. W potocznej świadomości wynik egzaminu, podawany w surowych punktach jest całkowicie tożsamy z wiedzą ucznia, a średnie wyniki szkoły z danego egzaminu są doskonałym sposobem na mierzenie jakości pracy szkoły. Teorie pomiaru pokazują, że wyciąganie wniosków o wiedzy i umiejętnościach ucznia jest bardziej skomplikowane niż utożsamianie wyniku egzaminu czy testu z umiejętnościami. Z tego

[14] W części rozdziału poświęconego komunikacji wyników przez okręgowe komisje egzaminacyjne i CKE wykorzystano dane zebrane podczas projektu „Modernizacja systemów informatycznych do obsługi systemu egzaminów zewnętrznych i nadzoru pedagogicznego I etap". Priorytet III Wysoka jakość oświaty, Działanie 3.1. Modernizacja systemu zarządzania i nadzoru w oświacie, Poddziałanie 3.1.1 Tworzenie warunków i narzędzi do monitorowania, ewaluacji i badań systemu oświaty.

względu informacja o wynikach egzaminacyjnych musi być podana w sposób, który umożliwia jej zrozumienie oraz wykorzystanie. Przygotowując ją, trzeba zatem ciągle balansować między szczegółowością podawanych informacji a przydatnością do poprawnego i adekwatnego formułowania wniosków (interpretacji) wyciąganych na podstawie wyników egzaminu przez poszczególne grupy potencjalnych odbiorców (por. podrozdział 2.4). Z tego względu, niemożliwe jest przygotowanie jednego komunikatu o wynikach w taki sposób, by spełniał oczekiwania wszystkich grup odbiorców. Dokładne określenie wskazanych powyżej elementów oraz relacji między nimi umożliwia przygotowanie komunikatu o wynikach egzaminacyjnych optymalnie dostosowanego do potrzeb danego odbiorcy. Myślenie w tych kategoriach pozwala także na systematyzację tego zagadnienia i porównanie praktyk poszczególnych OKE oraz CKE w tym zakresie.

Warto podkreślić, że komisje egzaminacyjne (czyli instytucje komunikujące wyniki egzaminacyjne) mają pełną kontrolę jedynie nad zakresem informacji i ich formą. Z komunikatu, zwłaszcza jeśli jest on udostępniony publicznie, może skorzystać ktoś inny, niż docelowy odbiorca. Także cel wykorzystania danej informacji może być różny od planowanego, czego przykładem może być nagminne wykorzystywanie zestawień wyników egzaminacyjnych szkół do tworzenia na tej podstawie rankingów. W kontekście wyników egzaminacyjnych jest to niezwykle istotne, ponieważ informacje o wynikach są często bardzo ważnym elementem w procesie oceny funkcjonowania szkół.

Większość informacji o wynikach egzaminacyjnych jest przygotowywana w okręgowych komisjach, jednak proces ten jest nadzorowany i koordynowany przez CKE. To okręgowe komisje egzaminacyjne przekazują zdającym zaświadczenia i świadectwa, poświadczające zdany egzamin. One także udostępniają szczegółowe wyniki egzaminów szkołom, w których się one odbywały oraz publikują ogólnodostępne zestawienia średnich wyników szkół. Na przestrzeni lat nastąpiło zróżnicowanie między komisjami w tym zakresie. Ramy prawne systemu egzaminów zewnętrznych określają minimalny zakres informacji oraz odbiorców, którym ta informacja musi zostać przekazana (zostało to opisane dokładnie w następnej części rozdziału). Jednak rzeczywisty zakres i forma, w jakiej są przygotowywane te zestawienia, różni się między komisjami egzaminacyjnymi. Wynika to z różnych przyczyn, z których jako najważniejszą można wskazać istniejące w komisjach egzaminacyjnych zróżnicowane rozwiązania informatyczne, umożliwiające przekazywanie wyników egzaminacyjnych, służących tym samym celom, w różnym zakresie i w różnej formie. W ostatnich latach obserwujemy tendencję do ujednolicenia rozwiązań w zakresie komunikowania wyników. Centralna Komisja Egzaminacyjna prowadzi obecnie intensywne działania w zakresie dokładnego określenia zakresu i formy informacji o wynikach kierowanej do różnych odbiorców.

2.7.2. Wynik egzaminacyjny w przestrzeni publicznej i w szkolnej rzeczywistości

Wyniki egzaminacyjne szkół stały się w Polsce informacją publiczną[15]. W praktyce oznacza to, że są dostępne na stronach internetowych okręgowych komisji egzaminacyjnych a niektóre szkoły prezentują swoje wyniki na własnych stronach www. Każda zainteresowana osoba może skorzystać z tych informacji. Dodatkowo dostępne są sprawozdania z przebiegu egzaminu, których przygotowanie jest statutowym obowiązkiem poszczególnych okręgowych komisji egzaminacyjnych (na podstawie danych z obszaru, który obejmuje dana OKE) i CKE (na podstawie danych ogólnopolskich). Krótkie przedstawienie zawartości tych sprawozdań zostało zamieszone w dalszej części rozdziału.

Niektórym grupom odbiorców Komisje egzaminacyjne przekazują specjalnie przygotowane informacje o wynikach. Najważniejszymi odbiorcami takich komunikatów są:

1. uczniowie piszący dany egzamin;
2. dyrektorzy szkół, w których się on odbywa;
3. kuratoria oświaty (organy nadzorujące szkoły);

[15] Podstawą prawną jest w tym przypadku Ustawa z 6 września 2001 r. o dostępie do informacji publicznej (tekst jednolity Dz. U. z 2014 r. poz. 782, 1662).

4. organy prowadzące szkoły (dla szkół publicznych są to jednostki samorządu terytorialnego, dla szkół niepublicznych inne podmioty, takie jak stowarzyszenia, fundacje, prywatne firmy, związki wyznaniowe itp.).

Informacje o wynikach kierowane do tych grup odbiorców różnią się od siebie zakresem oraz formą. Na to nakładają się także różnice w komunikatach między poszczególnymi okręgowymi komisjami egzaminacyjnymi. Należy jednak podkreślić, że ogólne ramy komunikowania wyników egzaminacyjnych dla tych grup są określone ustawowo. Zwykle takie komunikaty są udostępniane za pomocą dedykowanych serwisów internetowych, do których odbiorcy muszą logować się za pomocą specjalnego klucza otrzymanego z okręgowej komisji egzaminacyjnej.

Uczniowie

Podstawowym sposobem, w jaki zdający otrzymują informacje o swoich wynikach jest przekazanie im papierowych świadectw maturalnych i zaświadczeń o wynikach sprawdzianu i egzaminu gimnazjalnego. Wygląd i zakres informacji na zaświadczeniach i świadectwach jest dokładnie określany przez odpowiednie rozporządzenia[16]. Zakres informacji o wynikach na zaświadczeniach i świadectwach maturalnych zmieniał się, co przede wszystkim wiąże się ze zmianami w strukturze egzaminu (np. zwiększenie lub zmniejszenie liczby zdawanych przedmiotów). W 2014 roku zaświadczenie o wynikach sprawdzianu, oprócz sumy punktów z całego sprawdzianu, wraz z maksymalną liczbą punktów możliwych do uzyskania, zawierało informacje o wynikach ucznia z poszczególnych standardów egzaminacyjnych[17]. Zaświadczenie z egzaminu gimnazjalnego z tego roku zawierało wyniki ucznia z poszczególnych części egzaminu gimnazjalnego[18] oraz języka obcego zdawanego na tym egzaminie, wyrażone w procentach punktów oraz pozycję na skali centylowej wyznaczaną przez dany wynik. Świadectwo maturalne w 2014 roku zawierało natomiast wyniki z przedmiotów, które zdawał uczeń, wyrażone w procentach punktów[19].

Widocznym trendem w przypadku zaświadczeń i świadectw jest umieszczanie na nich, oprócz wyników z poszczególnych egzaminów i przedmiotów, pozycji ucznia na skali centylowej. Pojawiła się ona po raz pierwszy w 2012 roku na zaświadczeniu z egzaminu gimnazjalnego, a od 2015 roku także na świadectwie maturalnym. Umożliwia ona ocenę przez ucznia, jak wypadł z danego egzaminu wśród wszystkich innych go piszących (skala centylowa zostanie opisana dokładniej w późniejszej części rozdziału).

Jak już wspomniano wcześniej, oprócz papierowych zaświadczeń i świadectw, zdający egzaminy zewnętrzne w Polsce mają w większości dostęp do swoich wyników w formie elektronicznej za pomocą specjalnie przygotowanych serwisów internetowych. Prawie wszystkie okręgowe komisje egzaminacyjne wypracowały własny sposób udostępniania wyników. Elementem wspólnym dla nich jest dedykowany kanał dostępu – dzięki temu wgląd w swoje wyniki mają tylko osoby, których te wyniki dotyczą. Zakres prezentowanych w ten sposób informacji jest w przypadku większości OKE bardzo porównywalny. Zdający po zalogowaniu się do odpowiedniego serwisu ma przede wszystkim dostęp do informacji analogicznej do tej, którą uzyskuje na zaświadczeniu o wyniku lub świadectwie. Ponadto większość okręgowych komisji udostępnia także wyniki odpowiedzi na poszczególne zadania zamknięte (wraz z udzieloną przez daną osobę odpowiedzią) oraz otwarte, wraz z maksymalną liczbą punktów możliwą do uzyskania. Część komisji egzaminacyjnych prezentuje także wyniki zdającego z całego egzaminu oraz z poszczególnych przedmiotów na tle szkoły, gminy, powiatu lub województwa, dodatkowo umożliwiając zdającym pobieranie arkusza egzaminacyjnego, którego dotyczy dany wynik.

---

[16] Rozporządzenie Ministra Edukacji Narodowej z dnia 29 grudnia 2014 roku zmieniające rozporządzenie w sprawie świadectw, dyplomów państwowych i innych druków szkolnych.

[17] Są to: czytanie, pisanie, rozumowanie, korzystanie z informacji, wykorzystanie wiedzy w praktyce.

[18] Są to: w części humanistycznej: język polski, historia i wiedza o społeczeństwie; w części matematyczno-przyrodniczej: matematyka, przedmioty przyrodnicze (biologia, chemia, fizyka, geografia).

[19] Osoby, które nie zdały egzaminu maturalnego otrzymują zaświadczenie o wynikach, zawierające wyniki z poszczególnych zdawanych przedmiotów.

Rysunek 2.4. Wyniki egzaminacyjne dostępne dla ucznia po zalogowaniu się do specjalnego serwisu OKE w Krakowie w 2014 roku (przykład)

Szczegółowe wyniki egzaminu

Imię i nazwisko:
Sesja egzaminacyjna: Sesja egzaminacyjna 2013/2014
Typ egzaminu: sprawdzian
Egzamin: sprawdzian (A1) Wybierz
Wynik: 83%(33,0/40,0)

Wyniki w poszczególnych kategoriach umiejętności

Lp.	Kategoria umiejętności	Pkt	Maks. pkt	%
1.	Czytanie	9,0	10,0	90%
2.	Pisanie	8,0	10,0	80%
3.	Rozumowanie	7,0	8,0	88%
4.	Korzystanie z informacji	2,0	4,0	50%
5.	Wykorzystywanie wiedzy w praktyce	7,0	8,0	88%

Arkusz S-B1-142

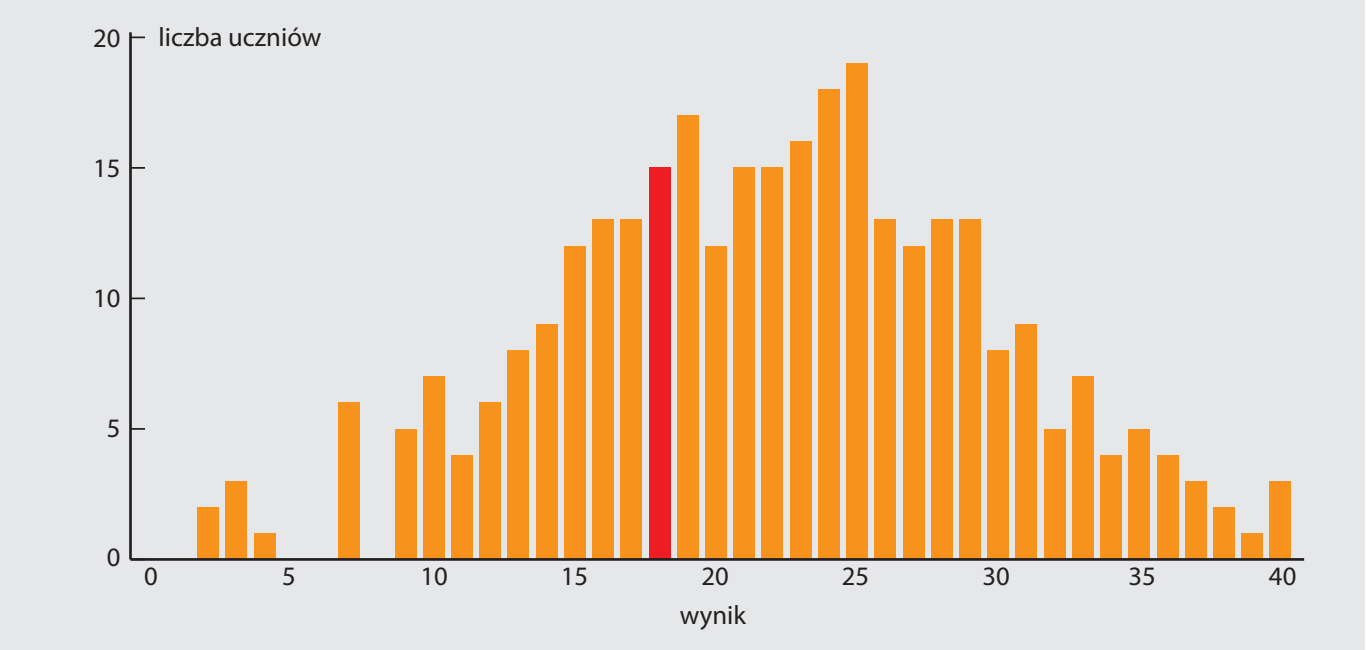
Wyniki dla poszczególnych zadań egzaminacyjnych

Nr zadania	Odpowiedź	Poprawna	Pkt	Maks. pkt
1	B	A	0,0	1,0
2	D	D	1,0	1,0
3	C	C	1,0	1,0
4	A	A	1,0	1,0
5	D	D	1,0	1,0
6	C	C	1,0	1,0
7	B	B	1,0	1,0
8	B	B	1,0	1,0
9	A	A	1,0	1,0
10	A	A	1,0	1,0
11	B	B	1,0	1,0
12	C	B	0,0	1,0
13	C	C	1,0	1,0
14	B	C	0,0	1,0
15	B	D	0,0	1,0
16	B	B	1,0	1,0
17	B	D	0,0	1,0
18	D	D	1,0	1,0
19	C	C	1,0	1,0
20	D	D	1,0	1,0
21			1,0	1,0
22			3,0	3,0
23			2,0	2,0
24			4,0	4,0
25			1,0	2,0
26			7,0	8,0

Zobacz treść zadań egzaminacyjnych z tego arkusza

1 z 2

Rysunek 2.5. Rozkład wyników w całej szkole (kolorem czerwonym zaznaczono wynik ucznia) – dostępne dla ucznia po zalogowaniu się do specjalnego serwisu OKE w Krakowie w 2013/2014 roku (przykład)



Szkoła

W dedykowanych portalach internetowych jest dostępna szczegółowa informacja o wynikach egzaminu także dla dyrektorów i nauczycieli. Po zalogowaniu się dyrektor ma dostęp do plików z informacją o wynikach egzaminu. Zakres tej informacji jest szeroki. Zwykle zawiera ona średnie wyniki dla całej szkoły oraz poszczególnych klas, często na tle wyników gminy, powiatu i województwa. Dodatkowo szkoły otrzymują dostęp do szczegółowych wyników uczniów uzyskanych za poszczególne zadania, a w przypadku matury także wskaźnik zdawalności egzaminu[20]. Część okręgowych komisji egzaminacyjnych udostępnia także dodatkowo informacje o poziomach wykonania zadań (łatwości zadań w populacji zdających egzamin) w poszczególnych klasach i całej szkole, także na tle gminy, powiatu i województwa. Informacje przekazywane są szkołom w formie tabelarycznej (jako pliki arkusza kalkulacyjnego Excel lub PDF) (por. tabela 2.5, tabela 2.6 i tabela 2.7).

Tabela 2.5. Zbiorcze wyniki egzaminu przekazywane szkole dostępne po zalogowaniu się do specjalnego serwisu OKE w Krakowie w 2014 roku (przykład).

	Standardy											
	Czytanie		Pisanie		Rozumowanie		Korzystanie z informacji		Wykorzystywanie wiedzy w praktyce		Wynik	
	pkt	% pkt	pkt	% pkt	pkt	% pkt	pkt	% pkt	pkt	% pkt	pkt	% pkt
Klasa A	7,1	71	7,5	75	5,4	68	3,3	81	5,7	71	28,9	72
Klasa B	6,8	68	6,6	66	5,2	64	2,9	72	4,8	60	26,2	66
Szkoła	6,3	63	6,1	61	4,5	57	2,9	73	4,6	57	24,4	61
Gmina	6,5	65	5,8	58	4,3	53	2,9	73	4,0	50	23,5	59
Powiat	6,5	65	5,8	58	4,2	53	2,9	72	4,0	50	23,4	59
Województwo	6,4	64	6,2	62	4,2	53	2,9	71	4,0	50	23,7	59

[20] Zdawalność egzaminu maturalnego to procent osób, które otrzymały świadectwo maturalne spośród zdających egzamin.

Przekazanie w plikach szczegółowych wyników uczniów za poszczególne zadania daje szkołom możliwość przeprowadzenia dodatkowych analiz za pomocą narzędzi statystycznych. Natomiast podanie szkołom średnich wyników uzyskanych przez uczniów np. w danej gminie jest przydatne do porównania wyników uczniów na tle innych szkół. Takie dane dają szkołom możliwości prowadzenia różnorodnych analiz z wykorzystaniem wyników egzaminacyjnych uczniów według swoich potrzeb.

Tabela 2.6. Szczegółowe wyniki egzaminu (za poszczególne zadania) przekazywane szkole dostępne po zalogowaniu się do specjalnego serwisu OKE w Krakowie w 2014 roku (przykład)

Przedmiot	**Sprawdzian**
Poziom	**nie dotyczy**
Typ wymagań	**standardowy**
Język zdawania	**polski**
Maksymalna liczba punktów	**40**

Oddział	Nr dziennika	Imię	Drugie imię	Nazwisko	PESEL	1	2	3	4	5	6	7	8	9	10
A	1					1	1	1	1	1	1	1	1	1	1
A	2					1	1	1	1	1	1	1	1	1	1
A	3					0	1	1	1	1	0	1	1	1	1
A	4					0	0	0	0	0	0	0	0	0	0
A	5					1	1	0	0	1	0	1	0	1	0
A	6					1	1	1	1	1	0	1	0	1	1
A	8					1	1	1	1	1	0	0	1	1	1
A	11					1	0	0	0	0	0	1	0	0	0
A	12					0	1	1	1	1	0	1	1	1	1
A	13					1	1	1	1	0	0	1	1	1	1
A	14					1	1	0	1	1	0	1	1	1	1

Tabela 2.7. Wskaźniki wykonania zadań przekazywane szkole dostępne po zalogowaniu się do specjalnego serwisu OKE w Krakowie w 2014 roku (przykład)

Wykonanie zadań w procentach	1	2	3	4	5	6	7	8	9	10
Klasa A	69	88	75	75	75	19	88	75	81	81
Klasa B	76	96	68	92	84	28	72	68	64	84
Szkoła	74	85	51	74	70	19	75	64	70	82
Gmina	74	86	52	81	75	25	82	59	71	86
Powiat	75	85	49	82	73	27	84	59	68	87
Województwo	75	84	47	83	72	26	86	58	65	83

Kuratoria oświaty oraz organy prowadzące szkoły

Informacje kierowane do kuratoriów oświaty oraz organów prowadzących szkoły są bardzo zbliżone do siebie w formie i zakresie. Informacja zawiera zwykle średnie wyniki uzyskane z poszczególnych przedmiotów dla każdej ze szkół. W niektórych komisjach egzaminacyjnych przekazuje się dodatkowe informacje, takie jak: liczba uczniów, procent uczniów z dysleksją, informacja o tym, czy dana szkoła jest publiczna lub niepubliczna. Dodatkowo w przypadku sprawdzianu i egzaminu gimnazjalnego zamieszczana jest informacja o pozycji szkoły na skali staninowej, natomiast w przypadku egzaminu maturalnego także informacja o zdawalności uczniów w szkole.

Informacja kierowana do kuratoriów i organów prowadzących szkoły jest obecnie w większości przypadków tożsama z informacjami o szkołach prezentowanymi publicznie na stronach okręgowych komisji egzaminacyjnych. W przeszłości sytuacja ta wyglądała odmiennie – wyniki szkół na danym terenie były dostępne dla organów nadzorujących i prowadzących szkoły za pomocą dedykowanych serwisów internetowych lub wręcz w formie papierowych raportów przesyłanych do nich pocztą. Społeczne oczekiwanie dostępności wyników egzaminacyjnych szkół sprawiło, że większość informacji o szkołach kierowanych do organów prowadzących i nadzorujących szkoły pokrywa się z tymi, które są dostępne publicznie. Przykład informacji egzaminacyjnej dla organów nadzorujących i prowadzących szkoły ilustruje tabela 2.8.

Wyniki w ogólnodostępnych serwisach

Informacje dostępne publicznie są kierowane do wszystkich zainteresowanych wynikami egzaminów. Najczęściej wykorzystywane są one przez dziennikarzy oraz rodziców w celu sprawdzenia wyników szkół, do których uczęszczają lub potencjalnie mogą uczęszczać ich dzieci.

Informacje o wynikach egzaminacyjnych w ogólnie dostępnych serwisach można podzielić na dwie kategorie: informacje o wynikach uczniów w poszczególnych szkołach, gminach czy też powiatach, oraz ogólne sprowadzania OKE i CKE o poszczególnych egzaminach.

Tabela 2.8. Wyniki egzaminacyjne szkół dostępne powszechnie na portalu internetowym OKE w Łodzi w 2014 roku (przykład)

Powiat	Gmina	Szkoła	Adres		Egzamin	Wyniki egzaminów w % punktów					
						Liczba zdających	W szkole	W gminie	W powiecie	W województwie	W kraju
bełchatowski	m. Bełchatów	Publiczne Gimnazjum nr 3 w Bełchatowie	ul. Edwardów 5	Bełchatów	Język polski	113	66	69	67	68	63
bełchatowski	m. Bełchatów	Publiczne Gimnazjum nr 3 w Bełchatowie	ul. Edwardów 5	Bełchatów	Historia i WOS	113	56	58	58	59	58
bełchatowski	m. Bełchatów	Publiczne Gimnazjum nr 3 w Bełchatowie	ul. Edwardów 5	Bełchatów	Matematyka	113	42	4$	4S	&	47
bełchatowski	m. Bełchatów	Publiczne Gimnazjum nr 3 w Bełchatowie	ul. Edwardów 5	Bełchatów	Przedmioty przyrodnicze	113	51	54	52	53	52
bełchatowski	m. Bełchatów	Publiczne Gimnazjum nr 3 w Bełchatowie	ul. Edwardów 5	Bełchatów	Język angielski PP	110	61	72	67	66	67
bełchatowski	m. Bełchatów	Publiczne Gimnazjum nr 3 w Bełchatowie	ul. Edwardów 5	Bełchatów	Język angielski PR	110	39	51	48	46	46
bełchatowski	m. Bełchatów	Publiczne Gimnazjum nr 3 w Bełchatowie	ul. Edwardów 5	Bełchatów	Język niemiecki PP	3	43	53	51	55	54
bełchatowski	m. Bełchatów	Publiczne Gimnazjum nr 2 w Samorządowym Zespole Szkół Ogólnokształcących nr 2 im. Adama Mickiewicza w Bełchatowie	ul. Edwardów 5	Bełchatów	Język polski	53	61	69	67	66	68
bełchatowski	m. Bełchatów	Publiczne Gimnazjum nr 2 w Samorządowym Zespole Szkół Ogólnokształcących nr 2 im. Adama Mickiewicza w Bełchatowie	ul. Edwardów 5	Bełchatów	Historia i WOS	53	48	58	58	59	59
bełchatowski	m. Bełchatów	Publiczne Gimnazjum nr 2 w Samorządowym Zespole Szkół Ogólnokształcących nr 2 im. Adama Mickiewicza w Bełchatowie	ul. Edwardów 5	Bełchatów	Matematyka	52	33	43	46	48	47

Tabela 2.9. Dane dotyczące zdawalności egzaminu maturalnego dostępne powszechnie na portalu internetowym OKE we Wrocławiu w 2014 roku (przykład)

Typ szkoły	Liczba zdających						Liczba wydanych dokumentów				Zdawalność w %		
	po raz pierwszy*		poprawia-jących wyniki	podwyższ-ających wyniki	pozostali	RAZEM (kol. 2+4+5+6)	świadectw dla kol. 2 i 3	świadectw dla kol. 4	świadectw dla kol. 6	aneksów	przystępujących po raz perwszy (dla kol. 3)	poprawiają-cych wyniki (dla kol. 4)	pozostał-ych (dla kol. 6)
	a)	b)											
1	2	3	4	5	6	7	8	9	10	11	12	13	14
						OKE WROCŁAW							
LO	16743	16402	1383	2096	118	20340	13085	174	31	1589	79,8	12,6	26,3
LP	556	521	301	31	9	897	184	34	1	27	35,3	11,3	11 1
LU	114	93	190	1	12	317	10	15	9	1	10,8	7,9	16,7
T	9082	8664	1822	224	60	11188	4685	278	17	193	54,1	15,3	28,3
TU	142	118	62	1	9	214	13	5	1	1	11,0	6,1	11,1
RAZEM	26637	25798	3758	2353	208	32956	17977	506	52	1811	69,7	13,5	25,0

Jak już wcześniej wspominano, obecnie zdecydowana większość informacji o szkołach jest dostępna dla każdego zainteresowanego za pomocą serwisów internetowych poszczególnych komisji egzaminacyjnych. Informacja ta zawiera zwykle średnie wyniki szkoły, wraz z liczbą uczniów zdających egzamin. Czasem dodatkowo w zestawieniach zawarte są informacje analogiczne do tych, które otrzymują organy nadzorujące i prowadzące szkoły.

Tabela 2.10. Wyniki uczniów z egzaminu gimnazjalnego w powiatach dostępne powszechnie na portalu internetowym OKE w Gdańsku w 2014 roku (przykład)

Powiat/miasto na prawach powiatu	Część humanistyczna z zakresu					
	języka polskiego			historii i wiedzy o społeczeństwie		
	Liczba zdających	Wynik średni w procentach	Odchylenie standardowe w procentach	Liczba zdających	Wynik średni w procentach	Odchylenie standardowe w procentach
aleksandrowski	509	64,37	17,16	509	59,91	16,15
brodnicki	855	61,82	17,39	855	57,32	15,4
bydgoski	1 018	64,71	16,82	1 019	58,1	15,95
chełmiński	486	64,99	17,08	486	57,91	14,49
golubsko-dobrzyński	529	62,75	18,72	529	55,46	15,46
grudziądzki	401	63,73	17,66	401	54,87	15,54
inowrocławski	1 554	63,73	19,05	1 554	56,47	15,6
lipnowski	749	60,89	18,21	749	55,24	15,09
Bydgoszcz	2906	68,63	17,75	2906	61,1	16,34
Grudziądz	855	64,23	18,99	855	56,86	16,86
Toruń	1 838	70,18	18,64	1 838	61,83	16,39
Włocławek	1 125	65,27	19,38	1 125	58,5	16,1

Sprawozdania o przebiegu egzaminów są przygotowywane przez CKE dla całej Polski i przez poszczególne OKE dla województw, których egzaminy przeprowadza dana komisja. Zawierają one bardzo szczegółowe informacje dotyczące przebiegu egzaminu oraz wyników egzaminacyjnych, często w podziale na różne grupy zdających. Informacje w tych raportach są ujęte zarówno w formie tabelarycznej, jak i graficznej. Szczegółowe informacje o podstawowych parametrach statystycznych wyników egzaminów, takich jak średnia, modalna, mediana, odchylenie standardowe, można odnaleźć właśnie tam. W niektórych sprawozdaniach podawany jest także wskaźnik rzetelności danego egzaminu (alfa Cronbacha). Oprócz zestawień statystycznych zawierają one opis zadań sprawdzających umiejętności, które wypadły najlepiej i najgorzej na danym egzaminie.

Sprawozdania w pierwszych latach funkcjonowania systemu egzaminacyjnego znacznie różniły się między komisjami pod wieloma względami, takimi jak stopnień szczegółowości, stosowane parametry statystyczne, czy też bogactwo opisu poszczególnych tabel i wykresów. Od 2014 roku szablon sprawozdań jest przygotowywany przez CKE, zatem forma i zakres danych prezentowanych przez każdą OKE taki sam. Ujednolicenie raportów ma swoje zalety, gdyż każdy zainteresowany znajdzie dokładnie ten sam zakres informacji w sprawozdaniach z różnych OKE, co ułatwia ich wykorzystywanie. Pewnym problemem może być natomiast kwestia dostosowywania ich treści do lokalnych uwarunkowań. Niektóre egzaminy są zdawane rzadko w poszczególnych OKE. Z tego względu zastosowywanie w niektórych przypadkach parametrów statystycznych w opisie rozkładu wyników wydaje się być wątpliwe. Wymóg podawania tych samych informacji w każdym przypadku może zatem prowadzić do przedstawienia zestawień statystycznych, obarczonych dużym błędem.

Prześledźmy przykładowo sprawozdanie ze sprawdzianu 2014 zatytułowane *Osiągnięcia uczniów kończących szkołę podstawową w roku 2011* (Centralna Komisja Egzaminacyjna, 2014a). Obejmuje ono dwie główne grupy informacji: organizacja i przebieg sprawdzianu oraz krajowe wyniki egzaminacyjne. W rozdziale o organizacji sprawdzianu możemy zapoznać się ze skrótowym opisem arkusza egzaminacyjnego i z opisem populacji przystępujących do sprawdzianu. Także można poznać główne dane dotyczące organizacji sprawdzianu takie jak liczba szkół, liczba obserwatorów, liczba egzaminatorów oceniających zadania otwarte w sprawdzianie, liczba unieważnień z powodu niesamodzielnego rozwiązywania zadań, naruszenia przepisów przeprowadzenia sprawdzianu oraz liczba wglądów do prac egzaminacyjnych.

W rozdziałach o wynikach można zapoznać się rezultatami egzaminacyjnymi w podziale na grupy uczniów piszących sprawdzian z wykorzystaniem arkuszy przeznaczonych dla uczniów bez dysfunkcji i uczniów ze specyficznymi trudnościami w uczeniu się (arkusz standardowy), wyniki uczniów z autyzmem, w tym z zespołem Aspergera, wyniki uczniów słabo widzących i niewidomych, wyniki uczniów słabosłyszących i niesłyszących, wyniki uczniów z upośledzeniem umysłowym w stopniu lekkim, wyniki uczniów piszących sprawdzian w języku litewskim.

Dla każdej z wymienionych grup przedstawiany jest rozkład wyników uczniów i szkół w skali surowych wyników (liczba punktów uzyskanych za wszystkie zadania) oraz komunikowane są podstawowe statystyki dotyczące rozkładu wyników, takie jak: miary tendencji centralnej (wynik średni, mediana i modalna), miary rozproszenia (odchylenie standardowe oraz rozstęp), wynik minimalny i maksymalny. Przedstawiane jest także zróżnicowanie wyników ze względu na płeć ucznia, lokalizację oraz typ szkoły (publiczne vs niepubliczne). Zarówno dla wyników uczniów, jak i dla wyników zagregowanych do średnich dla szkół tworzona jest także dziewięciostopniowa standardowa skala staninowa, która jest normalizowana na populacji zdających w danym roku. Dla każdego stopnia tej skali CKE podaje przedziały wyników surowych (patrz podrozdział 2.8). Odrębną grupę stanowią szczegółowe informacje i komentarze dotyczące umiejętności sprawdzanych przez poszczególne zadania.

Odnosząc się do sprawozdań, należy wspomnieć o publikacji przez OKE i CKE wstępnej informacji o wynikach danego egzaminu. Jest ona opublikowana w momencie ogłoszenia wyników danego egzaminu. Ten materiał zawiera podstawowe dane statystyczne o egzaminie i zadaniach, takie jak: podstawowe parametry statystyczne z poszczególnego egzaminu bądź przedmiotu (średnia,

mediana, dominanta, odchylenie standardowe), liczbę zdających egzamin, rozkłady wyników oraz krótką informację o najprostszych i najtrudniejszych zadaniach. Materiał ten jest zwykle wykorzystywany przez media do przygotowania informacji prasowej o wynikach danego egzaminu. Ta sama informacja stanowi także część sprawozdania o przebiegu egzaminu, który jest publikowany w późniejszym terminie.

2.7.3. Możliwości komunikowania wyników

Obecny sposób prezentowania wyników egzaminacyjnych jest wypadkową uwarunkowań prawnych systemu egzaminów zewnętrznych oraz potrzeb odbiorców tych informacji. Pozytywnie należy ocenić możliwość powszechnego dostępu do informacji o wynikach za pomocą stron internetowych komisji egzaminacyjnych. Szczegółowe dane są dostępne jedynie dla bezpośrednio zainteresowanych, tj. zdających egzaminy oraz dyrektorów i nauczycieli szkół poprzez dedykowane im serwisy. Zakres informacji wydaje się być wystarczający dla potrzeb odbiorców, którzy nie są zainteresowani prowadzeniem na tych wynikach bardziej zawansowanych analiz (często też nie posiadają odpowiednich kompetencji w tym zakresie).

Ważną kwestią jest zagadnienie interpretacji otrzymanych wyników. Praktycznie rzecz biorąc, jedynie w sprawozdaniach o przebiegu danego egzaminu wyniki egzaminu są opatrzone komentarzem. W przypadku wyników uczniów i szkół nie ma to miejsca. Przyczyna jest prosta – interpretacja wyników egzaminacyjnych wymaga czasu i sporego nakładu pracy. System egzaminacyjny nie posiada obecnie zasobów umożliwiających przygotowanie zindywidualizowanych interpretacji wyników uczniów czy szkół. W zamian komisje egzaminacyjne starają się przekazywać wiedzę o zasadach analizy i interpretacji wyników egzaminacyjnych na spotkaniach i szkoleniach kierowanych do dyrektorów i nauczycieli szkół, a także przygotowując poradniki w tym zakresie. Wydaje się jednak, że zastosowanie nowoczesnych technologii informatycznych mogłoby pozwolić na przygotowywanie zautomatyzowanych raportów dla szkół, które także zawierałyby proste interpretacje otrzymanych rezultatów.

Osobnym problemem jest komunikowanie wyników uczniów z danego egzaminu szkołom, w których uczniowie rozpoczynają kolejny etap edukacyjny. Dla przykładu, gimnazjum ma bezpośredni dostęp jedynie do informacji o wynikach sprawdzianu, które uczniowie mają na zaświadczeniach o wynikach. Wyniki ucznia z poprzedniego egzaminu nie są dostępne elektronicznie dla szkół, w których rozpoczynają oni naukę po napisaniu danego egzaminu. Oznacza to, że szkoły, jeśli chcą wykorzystać informacje z wcześniejszego egzaminu w procesie dydaktycznym, muszą te informacje same przetworzyć, przepisując je z papierowych zaświadczeń. Problem dostępu do danych z wcześniejszego etapu edukacyjnego wynika głównie stąd, że system egzaminów zewnętrznych nie posiada informacji, do jakiej szkoły trafia uczeń po zakończeniu kształcenia na danym etapie. Informacja o tym pojawia się w systemie dopiero w momencie, gdy nowa szkoła zgłasza ucznia do następnego egzaminu. Oczywiście, szkoły mogą po prostu zgłosić się z prośbą o te dane do odpowiedniej OKE, podając informacje pozwalające na identyfikację uczniów. Znacznie lepsze byłoby tutaj jednak rozwiązanie systemowe. Pewną propozycją jest w tym zakresie tak zwana „diagnoza na wejściu", wdrożona przez OKE w Krakowie. Szkoła, za pomocą specjalnie dedykowanego serwisu o ograniczonym dostępie ma możliwość wpisania do formularza danych identyfikacyjnych uczniów, którzy rozpoczęli w niej naukę. Następnie informacja o wynikach tych uczniów jest automatycznie pobierana z baz danych komisji i przekazywana do szkoły za pomocą raportu, dostępnego w wyżej wymienionym serwisie. Dzięki temu szkoła ma dostęp do podstawowych danych egzaminacyjnych i może wykorzystywać je do planowania pracy dydaktycznej z uczniami. Zestawienie rodzajów informacji o wynikach egzaminów przekazywanych przez okręgowe komisje egzaminacyjne zawiera tabela 2.11.

Tabela 2.11. Zestawienie dedykowanych informacji o wynikach egzaminacyjnych przekazywanych przez okręgowe komisje egzaminacyjne[21]

Rodzaj komunikatu	Odbiorca	Zakres	Forma	Cel wykorzystania	Dostęp do informacji
Wyniki uzyskane przez zdającego.	Zdający dany egzamin	Suma punktów uzyskana na danym egzaminie (informacja analogiczna do otrzymanej na świadectwie lub zaświadczeniu o wyniku); dodatkowo część komisji udostępnia wyniki za poszczególne umiejętności (sprawdzian): wyniki i odpowiedzi na poszczególne pytania zamknięte (odpowiedź jaka została udzielona, poprawna odpowiedź, uzyskane punkty, maks. pkt.) oraz pytania otwarte (suma punktów za zadanie, maks. pkt).	Zestawienie tabelaryczne; dodatkowo niektóre komisje prezentują histogram zawierający wynik zdającego na tle szkoły, gminy, powiatu i województwa oraz umożliwiają pobranie arkusza egzaminacyjnego, który rozwiązywał uczeń (w formacie PDF).	Umożliwienie zdającym dany egzamin sprawdzenie swoich wyników. W niektórych komisjach zdający ma możliwość porównania swojego wyniku z wynikami uczniów w swojej szkole, gminie lub powiecie oraz województwie.	Ograniczony – zdający egzamin po zalogowaniu do dedykowanego portalu.
Wyniki uzyskane przez zdającego	Dyrektor szkoły	Średnie wyniki szkoły i klas na tle gminy lub powiatu oraz województwa; szczegółowe wyniki uczniów uzyskane za poszczególne zadania; wyniki uczniów uzyskane z poszczególnych przedmiotów/standardów; zdawalność matury; dodatkowo część komisji udostępnia poziomy wykonania zadań (łatwość zadań), na tle klasy, szkoły, gminy lub powiatu oraz województwa.	Zestawienie tabelaryczne; dane dostępne w formie plików xls oraz (lub) PDF.	Umożliwienie zapoznania się dyrektorom szkół z pełną informacją o wynikach egzaminacyjnych uczniów. Zestawienie zawiera podsumowanie egzaminu w szkole – od średnich wyników w szkole, w klasach po wyniki pojedynczego ucznia. W celu umożliwienia analizy wyników dyrektorzy zwykle mają także podane średnie wyniki z danego egzaminu w gminie, powiecie, województwie, niekiedy w całym OKE.	Ograniczony – dyrektor danej szkoły po zalogowaniu do dedykowanego portalu.
Wyniki uzyskane w szkole					
Wyniki uzyskane w szkołach	Kuratoria oświaty (organy nadzorujące szkoły)	Średnie wyniki uzyskane z poszczególnych przedmiotów/standardów każdej ze szkół, działającej na terenie obejmowanym nadzorem pedagogicznym przez dane kuratorium; dodatkowo prezentowane są: liczba uczniów, procent uczniów z dysleksją, informacja o tym, czy dane szkoła jest publiczna lub niepubliczna; w przypadku sprawdzianu i egzaminu gimnazjalnego dodatkowo wynik szkoły w skali staninowej, dla egzaminu maturalnego także informacja o zdawalności.	Zestawienie tabelaryczne; dane dostępne w formie plików Excela oraz (lub) PDF.	Umożliwienie prowadzącym nadzór nad oświatą na danym terenie zapoznania się z wynikami, uzyskiwanymi przez szkoły im podlegające.	Większość OKE udostępnia informację o wynikach szkół publicznie, nie kierując specjalnej informacji do KO. Część komisji udostępnia te informacje także za pomocą portali o dedykowanym dostępie.
Wyniki uzyskane w szkołach	Jednostki samorządu terytorialnego oraz podmioty prowadzące szkoły niepubliczne (organy prowadzące szkoły)	Średnie wyniki uzyskane z poszczególnych przedmiotów/standardów każdej ze szkół, dodatkowo prezentowane są: liczba uczniów, procent uczniów z dysleksją, informacja o tym, czy dane szkoła jest publiczna lub niepubliczna; w przypadku sprawdzianu i egzaminu gimnazjalnego dodatkowo wynik szkoły w skali staninowej, dla egzaminu maturalnego także informacja o zdawalności.	Zestawienie tabelaryczne; dane dostępne w formie plików Excela oraz (lub) PDF). Część komisji przekazuje te informacje w postaci wydruków bezpośrednio pod adres organów prowadzących szkoły.	Umożliwienie zapoznanie się organom prowadzącym szkoły z wynikami, uzyskiwanymi przez szkoły im podlegające.	Większość OKE udostępnia informację o wynikach szkół publicznie, nie kierując specjalnej informacji do organów prowadzących szkoły. Część komisji udostępnia te informacje także za pomocą portali o dedykowanym dostępie.

[21] Opracowane na podstawie „Raport z analizy systemów funkcjonujących w obszarze oświaty" z 2014 roku, zrealizowane w ramach projektu „Modernizacja systemów informatycznych do obsługi systemu egzaminów zewnętrznych i nadzoru pedagogicznego I etap". Priorytet III Wysoka jakość oświaty, Działanie 3.1. Modernizacja systemu zarządzania i nadzoru w oświacie, Poddziałanie 3.1.1 Tworzenie warunków i narzędzi do monitorowania, ewaluacji i badań systemu oświaty.

2.8. Skale stosowane w komunikowaniu wyników w Polsce

Naturalne wydaje się komunikowanie wyników egzaminów jako sumy punktów przyznanych uczniom za poprawnie wykonane zadania. Dla uczniów jest to najbardziej zrozumiała forma, gdyż odnosi się bezpośrednio do arkusza egzaminacyjnego. Również dla innych odbiorców nie nastręcza ona trudności w interpretacji. W takiej właśnie formie prezentowane są wyniki od początku istnienia sytemu egzaminów zewnętrznych w Polsce, np. na zaświadczeniach dla sprawdzianu po szóstej klasie szkoły podstawowej. W przypadku odbiorców informacji dotyczących kilku egzaminów suma punktów nie jest jednak zbyt wygodna ze względu na różną liczbę punktów możliwych do zdobycia w zależności od typu egzaminu. Czym innym są 24 punkty zdobyte podczas sprawdzianu, a czym innym podczas matury z języka polskiego. Maksymalnie na sprawdzianie (do 2014 roku włącznie) uczeń mógł uzyskać 40 punktów, a na poziomie podstawowym egzaminu maturalnego z języka polskiego 70 punktów. Sumę punktów za poprawnie rozwiązane zadania zawsze odnosi się do maksymalnej możliwej do zdobycia punktacji, zatem bardzo wygodne jest przedstawianie jej w formie procentowej. Gdyby podana przykładowo punktacja dotyczyła tego samego ucznia – wynik w postaci 60% ze sprawdzianu i 34% z matury niósłby za sobą więcej informacji niż tzw. wynik surowy (prosta suma punktów). Od razu widoczna jest różnica w poziomie wykonania dla obydwu egzaminów, co wymaga dodatkowych operacji przy wcześniejszej formie przedstawienia wyników. W formie procentowej wyniki są prezentowane np. na świadectwach maturalnych i od 2012 roku także na zaświadczeniach z egzaminu gimnazjalnego.

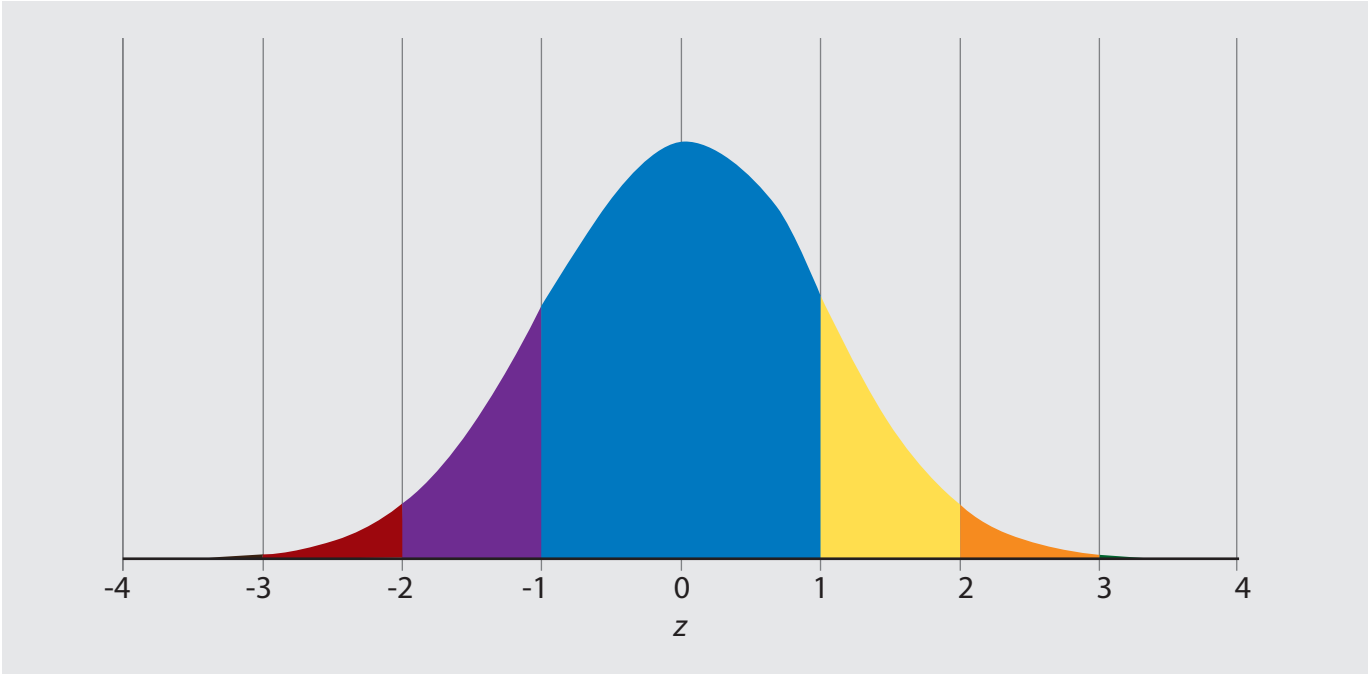
Zaprezentowane powyżej formy komunikowania wyników w gruncie rzeczy niewiele mówią o poziomie badanych umiejętności uczniów. Również w postaci wyników uśrednionych np. dla szkół czy województw nie możemy określić, czy są one wysokie czy niskie bez dodatkowych informacji. Mogłoby się zdawać, że przykładowy wynik rzędu 60% na egzaminie to wynik przyzwoity, zakładając, że średni wynik to 50% (co nie jest bezzasadne). Do oceny tej informacji potrzebne są dane o rozkładzie wyników wszystkich uczniów biorących udział w egzaminie. W 2014 roku średni wynik sprawdzianu po szóstej klasie szkoły podstawowej dla uczniów piszących arkusz standardowy wyniósł 25,8 punktu (Centralna Komisja Egzaminacyjna, 2014a) czyli 65%. Okazuje się, że podany jako przykład wynik wynoszący 60% plasuje się poniżej średniego wyniku. Gdybyśmy nie sprawdzili rzeczywistego rozkładu wyników i opierali się wyłącznie na naszym przekonaniu, że 50% powinno oznaczać średni wynik popełnilibyśmy dość istotny błąd w interpretacji. Podobne problemy z interpretacją, bez informacji o rzeczywistym rozkładzie wyników, będziemy mieli z przykładowym wynikiem z egzaminu maturalnego z języka polskiego. Na poziomie podstawowym próg zdawalności wyznaczony został na 30%, zatem nasz przykładowy wynik wynoszący 34% gwarantuje uczniowi zdanie tego egzaminu, choć możemy odnosić wrażenie, że jest to wynik raczej słaby. Średnia z egzaminu maturalnego z poziomu podstawowego dla języka polskiego w 2014 roku wynosi 51% (Centralna Komisja Egzaminacyjna, 2014b), co odpowiada przyjętemu wcześniej założeniu, zatem w tym przypadku nasza intuicyjna interpretacja zostaje potwierdzona. Wahania średnich wyników egzaminacyjnych pomiędzy latami (por. rozdział 3) mogą jednak powodować, że nie zawsze tak będzie i podobnie jak w przykładzie ze sprawdzianem możemy popełnić błąd.

Jak widać na powyższych przykładach, wyniki surowe czy ich przekształcenie do wyników procentowych nie są dobrą formą prezentacji wyników, jeśli jej celem jest dokonywanie porównań (zwłaszcza między latami). Są to formy proste w przygotowaniu i odbiorze, lecz stosunkowo łatwo mogą powodować błędy w interpretacji u niezbyt doświadczonych odbiorców. Konieczność uwzględniania rozkładu wyników w ich przedstawianiu prowadzi do zastosowania skal standardowych.

Najbardziej znaną skalą standardową jest skala nazywana skalą *z*, w której jednostką jest odchylenie standardowe (miara zróżnicowania wyników), a wartość średnia wynosi 0. Skale standardowe opierają się na pewnych właściwościach rozkładu normalnego:

1. zarówno poniżej, jak i powyżej średniej znajduje się 50% wyników,
2. w odległości jednego odchylenia standardowego od średniej mieści się około 68% wyników (symetrycznie po 34% poniżej, jak i powyżej średniej),
3. w odległości dwóch odchyleń standardowych od średniej mieści się około 95% wyników (również symetrycznie wokół średniej),
4. w odległości trzech odchyleń standardowych od średniej mieści się ponad 99% wyników (symetrycznie wokół średniej),
5. pozostałe obserwacje leżą dalej niż trzy odchylenia standardowe od średniej.

Rysunek 2.6. Skala standardowa *z*



Ilustrację powyższych stwierdzeń przedstawia rysunek 2.6. Przeliczenia dowolnego wyniku surowego na skalę standardową *z* można dokonać bardzo prosto – odejmując od niego wynik średni, a następnie dzieląc przez odchylenie standardowe. Odchylenie standardowe dla sprawdzianu 2014 wynosi 8 (Centralna Komisja Egzaminacyjna, 2014a), zatem nasz przykładowy wynik ze sprawdzianu w skali *z* wynosi: (24-25,8)/8=-0,225. Znajduje się on w obszarze jednego odchylenia standardowego poniżej średniej, czyli w grupie około 34% podobnych wyników[22]. Odchylenie standardowe opisywanego egzaminu maturalnego wynosi 17% (Centralna Komisja Egzaminacyjna, 2014b), zatem na skali *z* przykładowy wynik z matury z języka polskiego na poziomie podstawowym wynosi: (34%-51%)/17%=-1. Taki wynik uzyskany przez ucznia oznaczałby, że tylko około 16% uczniów uzyskało wynik gorszy od niego (od połowy wszystkich wyników poniżej średniej równej 0 należy odjąć 34% mieszczące się w odległości jednego odchylenia standardowego poniżej średniej; por. rysunek 2.6). Od początku istnienia systemu egzaminów zewnętrznych jedną z form komunikacji wyników jest skala staninowa, o której wspomniano już we wcześniejszej części tego rozdziału. Jest to skala standardowa, podobnie jak skala *z*, ma ona jednak nieco inne jednostki, a co za tym idzie, inaczej wyznaczone przedziały wyników. Jej nazwa pochodzi od angielskiego określenia *standard nine*, czyli

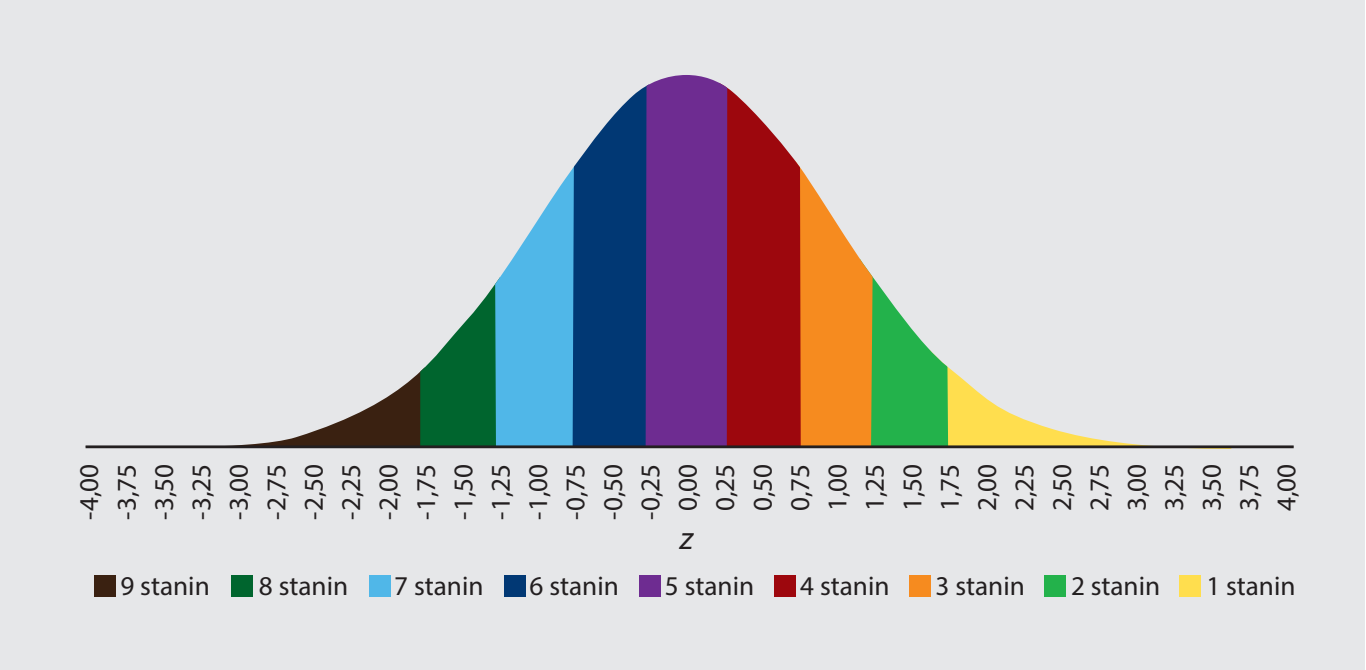
[22] Dla celów dydaktycznych przykłady zawarte w tej części opierają się na założeniu, że wyniki egzaminacyjne mają rozkład normalny.

standardowa dziewiątka, co odnosi się do wyznaczonych dziewięciu przedziałów wyników, tzw. staninów. Średni wynik na tej skali przypada na piąty stanin (który odpowiada wartości 0 na skali *z*), a każdy stanin (za wyjątkiem skrajnych: pierwszego i dziewiątego) zawiera w sobie wyniki z obszaru 0,5 odchylenia standardowego (co jest równoznaczne z tym, że odchylenie standardowe skali staninowej wynosi 2). Taka konstrukcja skali powoduje, że:

- w piątym staninie mieści się 20% wyników,
- w czwartym i szóstym staninie – po 17% wyników,
- w trzecim i siódmym staninie – po 12% wyników,
- w drugim i ósmym staninie – po 7% wyników,
- w pierwszym i dziewiątym staninie po 4% wyników.

Przedziały na skali staninowej i ich odniesienie do skali *z* ilustruje rysunek 2.7.

Rysunek 2.7 Skala staninowa w odniesieniu do skali *z*



Skal standardowych opartych na podobnych założeniach jak skala staninowa można stworzyć wiele, wymienić można choćby skalę stenową (Hornowska, 2007). Ma ona dziesięć przedziałów (stenów), jej średnia wynosi 5,5, a odchylenie standardowe 2, a nazwa pochodzi od angielskiego *standard ten*, czyli standardowa dziesiątka. Ważną cechą tego typu skal jest to, że wartości interpretuje się tylko w odniesieniu do wyznaczonych na nich przedziałów, co w przypadku danych egzaminacyjnych skutkuje utratą informacji, gdyż przedziałów jest mniej niż możliwych do zdobycia surowych punktów.

Centralna Komisja Egzaminacyjna wyznacza co roku przedziały skali staninowej np. dla sprawdzianu i publikuje je w sprawozdaniach. Tabela 2.12. zawiera przedziały wyników dla sprawdzianu 2014 i odpowiadające im staniny (Centralna Komisja Egzaminacyjna, 2014a). Już pierwszy rzut oka na jej zawartość ujawnia pewne problemy – procent wyników w poszczególnych staninach odbiega od wartości teoretycznych i nie jest symetryczny względem środkowego staninu. Jest to skutkiem tego, że sprawdzian jest punktowany na skali od 0 do 40 punktów więc nie możemy wystarczająco dobrze odróżnić od siebie uczniów, aby przeprowadzić podział na przedziały wyników w określonych miejscach. Ponadto w jednym staninie mieszczą się uczniowie osiągający różną sumę punktów, czyli redukujemy w ten sposób informacje zawarte w skali surowych wyników. Skala staninowa upraszcza zatem nieco rzeczywistość i nie jest tak samo przydatna do różnych celów. Gdybyśmy chcieli za jej

pomocą uszeregować uczniów w klasie, to musielibyśmy przyznawać miejsca *ex aequo* uczniom, którzy otrzymali np. 20 i 24 punkty ze sprawdzianu (4 stanin; por. tabela 2.12.).

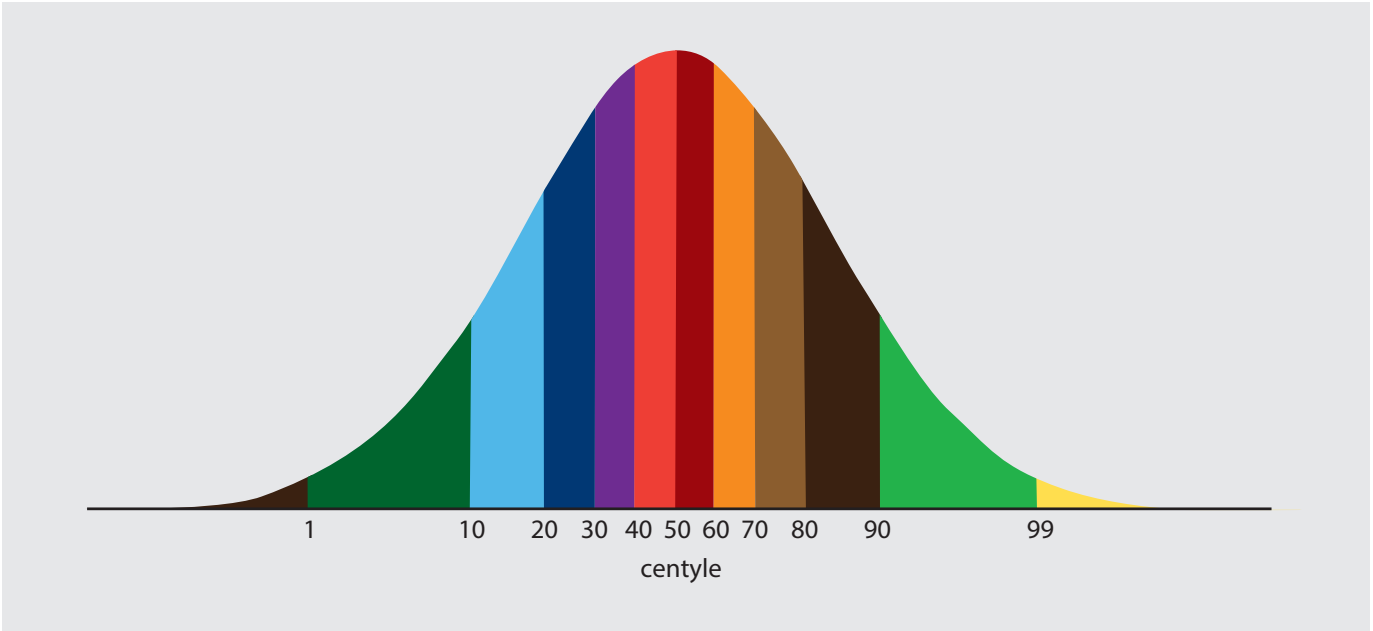
Tabela 2.12 Przedziały skali staninowej dla wyników sprawdzianu 2014

Stanin	Procent wyników	Przedział wyników
1	4,3	0–10
2	6,2	11–14
3	12,0	15–19
4	17,3	20–24
5	22,5	25–29
6	14,6	30–32
7	13,0	33–35
8	6,0	36–37
9	4,1	38–40

Źródło: Centralna Komisja Egzaminacyjna (2014a)

Innym typem skali stosowanej do komunikowania wyników egzaminacyjnych jest skala centylowa, o której wspomniano we wcześniejszej części rozdziału. Została ona wprowadzona do komunikowania wyników egzaminu gimnazjalnego w 2012 roku i wskazuje, jaki procent obserwacji (wyników) jest mniejszy lub równy niż dana wartość. Zasadę jej konstrukcji można opisać przy pomocy tzw. kwartyli, które dzielą wyniki na cztery grupy (stąd ich nazwa). Pierwszy kwartyl określa wartość, gdzie 25% wyników jest mniejsze lub jej równe, drugi kwartyl to mediana, czyli wartość, gdzie 50% wyników jest mniejsze lub jej równe, natomiast trzeci kwartyl to wartość, gdzie 75% wyników jest mniejsze lub jej równe. Podobnie kwintyle dzielą wyniki na pięć, a decyle na dziesięć grup. Łatwo zatem wywnioskować, że centyle (zwane też percentylami) dzielą wyniki na sto grup.

Rysunek 2.8 Przedziały wyników na skali centylowej



We wcześniejszych przykładach skal (staninowa i stenowa) grupy były wyznaczane przy użyciu odległości od średniej w jednostkach odchylenia standardowego skali *z*. Skutkowało to różnym udziałem procentowym wyników w poszczególnych przedziałach. W przypadku skali centylowej mamy do czynienia z sytuacja odwrotną – przedziały pomiędzy poszczególnymi centylami mają zawierać taką samą liczbę obserwacji, czyli 1%. Skutkuje to tym, że im bliżej średniej, tym pomiędzy centylami zawiera się coraz mniejszy przedział wyników. Ilustrację tej cechy skali centylowej przedstawia rysunek 2.8.

Choć skala centylowa ma niewątpliwie swoje zalety, ma też niestety wady. Zaletą tej skali jest to, że nie redukuje ona informacji o wynikach w tak dużym stopniu jak skala staninowa. Gdybyśmy mieli jednak do czynienia z egzaminem, na którym możliwe byłoby osiągnięcie dużej liczby punktów, to mogłoby dojść do sytuacji, kiedy pomiędzy dwoma centylami znajdowałyby się wyniki różniące się pomiędzy sobą na surowej skali. Wadą skali centylowej jest natomiast niemożliwość dokonywania na niej pewnych operacji matematycznych, które są naturalne w przypadku porównywania wyników (choć wynika to wprost z konstrukcji tej skali). O ile w przypadku surowych wyników możliwe jest określenie, że uczeń, który otrzymał ze sprawdzianu 20 punktów (50%) uzyskał ich dwukrotnie mniej niż uczeń, którego punktację ustalono na 40 punktów (100%), o tyle w przypadku centyli jest to niemożliwe. Za pomocą skali centylowej nie możemy nawet określić tak prostej zależności, jak to, o ile punktów więcej lub mniej od średniej uzyskał dany uczeń. Musimy przyjąć zupełnie inny sposób interpretacji wyników i odnosić się do proporcji uczniów posiadających wynik niższy lub taki sam (lub wyższy).

Dokonując podsumowania cech skal zaprezentowanych powyżej można dojść do (słusznego) wniosku, że wybór skali, na której prezentowane są wyniki zależny jest od celu, któremu mają one służyć. Każda skala znajdzie swoim zwolenników i przeciwników i nie ma skali, która spełniałaby oczekiwania wszystkich odbiorców informacji o wynikach egzaminacyjnych. Stosowane przez Centralną Komisję Egzaminacyjną do komunikowania skale są szeroko używane także w innych krajach przez instytucje zajmujące się edukacją. Za przykład może posłużyć chociażby amerykańska organizacja o nazwie Educational Records Bureau[23] oferująca szereg testów osiągnięć czy egzaminów wstępnych, która w swoim niezależnym egzaminie wstępnym dla szkół (Independent School Entrance Exam) prezentuje uczniom wyniki zarówno na skali staninowej, jak i centylowej[24]. Przykładem z drugiej strony globu może być nowozelandzka organizacja New Zealand Council for Educational Research[25] zajmująca się badaniami edukacyjnymi, która ich wyniki komunikuje również przy użyciu tych dwóch skal[26].

Skale stosowane w edukacji nie ograniczają się do zaprezentowanych przykładów. Poza skalą staninową bardzo często można spotkać inne skale standardowe. W międzynarodowych badaniach PISA, PIRLS, TIMSS wykorzystywana jest skala o średniej 500 i odchyleniu standardowych równym 100 (zwana skalą CEEB[27]). Popularną i dobrze znaną skalą jest też standardowa skala o średniej 100 i odchyleniu standardowym równym 15. Wielość skal standardowych wynika z możliwości arbitralnego ustalenia wartości średniej i wielkości odchylenia standardowego. Można w ten sposób tworzyć skale dostosowane do specyficznych potrzeb danego egzaminu, badania czy testu. Przeliczania wyników surowych na skale standardowe dokonuje się poprzez użycie skali *z* – wynik w skali *z* mnoży się przez odchylenie standardowe używanej skali i do wyniku tej operacji dodaje jej wartość średnią. Dowolność wyboru parametrów skal standardowych rodzi pewne konsekwencje dotyczące ich precyzji. Wspomniana skala staninowa ma zbyt mały zakres, aby dobrze reprezentować zróżnicowanie

---

[23] https://www.erblearn.org/

[24] Opis komunikowania wyników ISEE można znaleźć pod adresem http://www.ehow.com/how_7801376_understand-scores-isee-exam.html http://www.erblearn.org/sites/default/files/images/parents/Understanding%20the%20ISEE%20Report_r3.pdf.

[25] http://www.nzcer.org.nz/

[26] Por. http://www.nzcersupport.org.nz/marking/?p=75

[27] College Entrance Examination Board (CEEB) to dawna nazwa obecnej College Board, czyli instytucji odpowiedzialnej głównie za przeprowadzanie egzaminów wstępnych na studia w Stanach Zjednoczonych: Scholastic Assessment Test (SAT) i Advance Placement (AP).

wyników opisywanych egzaminów, przez co następowała utrata części informacji. Oczywiście nic nie stoi na przeszkodzie, aby zamiast używania tylko liczb całkowitych wyrażać wyniki również z częścią ułamkową. Jeśli średnia skali staninowej wynosi 5, a jej odchylenie standardowe 2, to podany wcześniej jako przykład wynik ze sprawdzianu z 2014 wynoszący 24 punkty (z=-0,225) roku można według opisanej wyżej reguły obliczyć jako: -0,225*2+5, co daje 4,55. Nie jest to już jednak wynik na skali staninowej, a na innej skali o średniej 5 i odchyleniu standardowym 2, gdyż na skali staninowej wynik należy zaokrąglić do liczby całkowitej (por. tabela 2.12). Pomimo dodania części ułamkowej nadal jednak mamy do czynienia ze „ściśnięciem" wyników z zakresu 0–40 do zakresu 0–9, co może być niewygodne. Z drugiej strony moglibyśmy użyć skali CEEB, w której wspomniany wynik miałby wartość 478[28]. Tak duża liczba może powodować iluzję, że nasz egzamin jest bardzo precyzyjny – potencjalnie możemy przecież odróżnić kogoś, kto uzyskał 477 punktów od kogoś, kto uzyskał 476 czy 478 punktów. Dopiero obliczenie, że jeden punkt na skali surowej sprawdzianu 2014, czyli 1/8 odchylenia standardowego, jest równy 12,5 punktu w skali CEEB uświadamia nam, że w rzeczywistości jest to niemożliwe. Za sprawą takich wskaźników, jak Edukacyjna Wartość Dodana (EWD) czy Porównywalne Wyniki Egzaminacyjne (PWE) w komunikowaniu wyników egzaminów zewnętrznych w Polsce zaczęto używać skali o średniej równej 100 i odchyleniu standardowym równym 15 (zob. rozdział 3 i 4 raportu)

2.9. Możliwe kierunki rozwoju systemu egzaminacyjnego w kontekście zmian formuły egzaminów w 2015 roku

Zmiany formuły egzaminu wprowadzane w ostatnich latach wiążą się ze zmianą podstawy programowej i odchodzeniem od oceniania analitycznego (kryterialnego) do holistycznego. Ocenianie analityczne było stosowane w naszych egzaminach od początku ich wprowadzenia. Wybór takiego rozwiązania w chwili, kiedy był przygotowywany i wdrażany polski system egzaminacyjny związany był ze światowym trendem pozytywistycznej psychometrii, której jednym z założeń jest maksymalizacja rzetelności pomiaru edukacyjnego. Zaletą oceniania analitycznego jest większa niż w przypadku holistycznego zgodność oceniania pomiędzy oceniającymi. W praktyce kilkunastu lat doświadczeń okazało się, że ocenianie analitycznie nie umożliwia jednak osiągnięcia takiej zgodności, jakiej byśmy oczekiwali.

Od roku 1990 następują na świecie radykalne zmiany w dziedzinie psychometrii i oceny zadań otwartych rozszerzonej odpowiedzi, takich jak wypracowania pisemne z przedmiotów humanistycznych, czy zadania wymagające od zdających rozwiniętego zapisu rozwiązania w przedmiotach matematyczno-przyrodniczych. Zmiany te mają miejsce szczególnie odnośnie trafności oceniania tego typu prac. W tym nowym kontekście teoretycznym, po raz pierwszy otwarły się możliwości krytycznego spojrzenia na analityczne ocenianie prac humanistycznych. Szczególne znaczenie dla tego nurtu miały prace Huota (1996), Mossa (1992), Messicka (1990). Ocenianie stało się publicznym i edukacyjnym problemem, a nie tylko dziedziną, którą zajmowały się instytucje przygotowujące egzaminy, dbające o jakość w kontekście wymagań pozytywistycznej psychometrii i wysokiej zgodności oceniania. Zaczęto poszukiwać odpowiedzi na pytanie, czy ocenianie (nie tylko na egzaminach) pomaga, czy też może w pewnych warunkach mieć negatywny wpływ na uczenie się i nauczanie. W praktyce egzaminacyjnej akcent zaczął się przesuwać także i w naszych egzaminach od oceniania analitycznego w kierunku oceniania holistycznego.

W egzaminie gimnazjalnym z języka polskiego i matematyki ocenianie holistyczne zadań otwartych wprowadzono w 2012 roku. W przypadku egzaminu maturalnego z matematyki już od kilku lat w wyniku działania Centralnego Zespołu Ekspertów Matematycznych (CZEM) ocenianie holistyczne praktykowane jest w coraz to szerszym zakresie. Natomiast w pozostałych egzaminach maturalnych

[28] W skali CEEB, podobnie jak w skali staninowej, wynik również należy zaokrąglić do liczby całkowitej. Dokładnie wartość ta wynosi 477,5.

i w sprawdzanie w szóstej klasie szkoły podstawowej ocenianie holistyczne stanie się obowiązującą praktyką od wiosennej sesji 2015 roku.

Główne kierunki rozwoju systemu egzaminacyjnego, także w kontekście wprowadzanych od 2015 zmian powinny dotyczyć przede wszystkim lepszego przygotowywania zadań egzaminacyjnych i monitorowania trafności arkuszy egzaminacyjnych w różnych jej aspektach. Bardzo ważnym obszarem wymagającym poprawy jest proces oceniania zadań rozszerzonej odpowiedzi przez egzaminatorów (w czym pomocne będzie wdrażanie e-oceniania), a także udoskonalanie i uspójnianie rozwiązań organizacyjnych i informatycznych w zakresie przetwarzania i komunikowania wyników.

Aby diametralnie podnieść jakość arkuszy egzaminacyjnych konieczne jest przejście od autorsko tworzonych testów do budowania narzędzi egzaminacyjnych bazując na wykalibrowanych bankach zadań. Przygotowywanie egzaminów powinno przebiegać podobnie jak przygotowanie rzetelnego badania naukowego. Do tego potrzebni są nie tylko eksperci znający się na dydaktyce poszczególnych obszarów edukacyjnych ale także eksperci w dziedzinie psychometrii i sztuki tworzenia egzaminacyjnych testów. Wątek ten przewija się od samego początku funkcjonowania polskiego systemu egzaminów zewnętrznych (Szaleniec i Szmigel, 2001). Dziś po 13 latach od pierwszych egzaminów prowadzonych przez OKE i CKE testy egzaminacyjne przygotowywane są w formie autorskich propozycji przez OKE i wybierane, kompilowane i modyfikowane w CKE. Projekt dotyczący banków zadań współfinansowany z EFS i koordynowany przez CKE jest dopiero pierwszym krokiem w stronę docelowych rozwiązań, umożliwiających konstruowanie testów z wykorzystaniem tych banków, tak by w arkuszach egzaminacyjnych znalazły się zadania o znanych parametrach psychometrycznych i trafnych do założonych celów określonego egzaminu. O pilnej potrzebie przyspieszenia prac nad bankami zadań może świadczyć fakt, że podczas analizy zadań z egzaminu gimnazjalnego w części humanistycznej i matematyczno-przyrodniczej w latach 2002-2010 Zespół Analiz Osiągnięć Uczniów IBE wytypował 66 zadań o słabych właściwościach psychometrycznych, które nie powinny znaleźć się w testach egzaminacyjnych (Szaleniec i in., 2013). Potrzebne są też zmiany strukturalne dotyczące profesjonalizacji Centralnej Komisji Egzaminacyjnej jaki i całego spektrum oferowanych egzaminów.

Bibliografia

American Educational Research Association (AERA), American Psychological Association (APA) i National Council on Measurement in Education (NCME). (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association (AERA), American Psychological Association (APA) i National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association (AERA), American Psychological Association (APA) i National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. W: P.W Holland i H. Wainer (red.), *Differential item functioning* (3-23), Hillsdale, N J Erlbaum.

Ban, E., Chamczyk, R., Czarnotta-Mączyńska, J., Murawska, M., Raczkowska, M. i Ślęzakowska, R. (2014). *Osiągnięcia uczniów kończących szkołę podstawową w roku 2014*, Sprawozdanie ze sprawdzianu 2014, Centralna Komisja Egzaminacyjna, Warszawa.

Brookhart, S.M. (2009) *How To Give Effective Feedback to Your Students*. Virginia: ASCD USA.

Camara, W. i Michaelides, M. (2005). *AP use in admissions: a response to Geiser and Santelices.* College Board Research Note, May 11, 2005. Dostępne **http://www.collegeboard.com/research/pdf/051425Geiser_050406.pdf**

Carmines, E.G. i Zeller, R.A. (1979). *Reliability and Validity Assessment*. Thousand Oaks, CA: SAGE Publications.

Centralna Komisja Egzaminacyjna. (2005). *Przygotowanie propozycji pytań, zadań i testów do przeprowadzenia sprawdzianu i egzaminu gimnazjalnego* [Niepublikowane procedury ustalone na zebraniu dyrektorów CKE i OKE w dniu 24 listopada 2005 r.].

Kozak, W., Kosińska-Pułka, M. i Spilkowski, A. (2014). *Sprawozdanie z egzaminu maturalnego 2014. Język polski*. Warszawa: Centralna Komisja Egzaminacyjna.

Cleary, A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

Cole, N.S. i Zieky, M.J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*, 369–382.

Crooks, T.J. (1988). The impact of classrooms evaluation practices on students. *Review of Educational Research,* 58, 438-481.

Crooks, T.J., Kane, M.T. i Cohen, A.S. (2008). Threats to the valid use of assessments. W: H. Wynne (red.), *Student assessment and testing*: Vol. 2 (151-171). Thousand Oaks, CA: Sage.

Cronbach, L.J. (1971). Test validation. W: R.L. Thorndike (red.), *Educational measurement* (wyd.2 , 443-507). Washington, DC: American Council on Education.

Cronbach, L.J. (1980). *Validity on parole: How can we go straight? New directions for testing and measurement -- Measuring achievement over a decade* -- Proceedings of the 1979 ETS Invitational Conference (99-108). San Francisco: Jossey-Bass.

Cronbach, L.J. (1988). Five perspectives on validity argument. W: H. Wainer i H. Braun (red.), *Test validity* (3-17). Hillsdale, NJ: Erlbaum.

Downing, S.M. (2006). Twelve Steps for Effective Test Development. W: S.M. Downing i T.M. Haladyna (red.), *Handbook of Test Development* (s. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S.M. (2009). Written tests: constructed-response and selected-response formats. W: S. M. Downing i R. Yudkowsky (red.). *Assessment in Health Professions Education* (s. 149–185). New York, NY: Routledge.

Educational Testing Service (2002). *ETS standards for quality and fairness*. Princeton, NJ: ETS.

Geiser, S., i Santelices, M. V. (2007). *Validity of high-school grades in predicting student success beyond the freshman year: high-school record vs standardized tests as indicators of four-year college outcomes.* Berkeley, CA: Center for Studies in Higher Education, University of California.

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T.M. i Downing, S.M. (1989). A taxonomy of multiple choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37-50.

Haladyna, T.M. i Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27.

Hornowska, E. (2007). *Standardy dla testów stosowanych w psychologii i pedagogice*. Gdańsk: GWP.

Hornowska, E. (2001). *Testy psychologiczne: teoria i praktyka*. Warszawa: Scholar.

Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias. W: R. Berk (red.), *Handbook of methods for detecting test bias* (s. 117–155). Baltimore: Johns Hopkins University Press.

Joint Committee on Testing Practices (1988). *Code of fair testing practices in education*. Washington, DC: Author.

Joint Committee on Testing Practices (2004). *Code of fair testing practices in education*. Washington, DC: Author.

Kane, M.T. i American College Testing Program. (1990). *An argument-based approach to validation*. Iowa City, Iowa: American College Testing Program.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.

Kane, M. (2010). Validity and fairness. *Language Testing,* 27(2), 177-182.

Kondratek, B. i Pokropek, A. (2013). IRT i pomiar edukacyjny. *Edukacja, 4*(124), 42–66.

Koniewski, M., Majkut, P. i Skórska, P. (2014). Zróżnicowane funkcjonowanie zadań testowych ze względu na wersję testu. *Edukacja, 1*(126), 79–94.

Langenfeld, T.E. (1997). Test fairness: Internal and external investigations of gender bias in mathematics testing. *Educational Measurement: Issues and Practice, 16*, 20-26.

Linden van der, W.J. (2005). *Linear Models for Optimal Test Design*. New York, NY: Springer.

Linn, R.L. (2006). The Standards for Educational and Psychological Testing: Guidance in Test Development. W: S.M. Downing i T.M. Haladyna (red.), *Handbook of test development* , 27-38. Mahwah, NJ: Lawrence Erlbaum Associates.

Linn, R. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Linn, R.L., Levine, M.V., Hastings, C.N. i Wardrop, J.L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*(2), 159–173.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Messick, S. (1989). Validity. W: R.L. Linn (red.), *Educational measurement* (wyd. 3., 13-103). New York: American Council on Education, Macmillan.

Messick, S. (1990). *Validity of test interpretation and use*. Princeton, N.J: Educational Testing Service. ETS-RR-90-11

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Messick, S. (2000). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. W: R.D. Goffin i E. Helmes (red.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (3–20). Boston: Kluwer Academic Publishers.

Millman, J., Bishop, C. H. i Ebel, R. L. (1965). An analysis of test wiseness. *Educational and Psychological Measurement, 25*(3), 707-726.

Najwyższa Izba Kontroli (2015). *System egzaminów zewnętrznych w oświacie.* Dostępne https://www.nik.gov.pl/plik/id,8001,vp,10018.pdf

Niemierko, B. (1975). *Testy osiągnięć szkolnych. Podstawowe pojęcia i techniki obliczeniowe.* Warszawa, WSiP.

Niemierko, B. (2001). Chłodne oblicze egzaminu zewnętrznego W: B. Niemierko i M.K. Szmigel (red.), *Teoria i praktyka oceniania zewnętrznego.* PANDID, Kraków.

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

Niemierko, B. (2009). Diagnostyka edukacyjna. Warszawa: Wydawnictwo Naukowe PWN.

Novick, M. i Lewis, G. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1-13.

OECD (2006), *Ocenianie kształtujące. Doskonalenie kształcenia w szkole średniej*, Centralny Ośrodek Doskonalenia Nauczycieli.

Popham W.J. (1990). *Modern Educational Measurement. A practitioner's perspective*. Englewood Cliffs, New York, Prentice Hall.

Scullen, S.E., Mount, M.K. i Goff, M. (2000). Understanding the Latent Structure of Job Performance Ratings. *Journal of Applied Psychology, 85*(6), 956–970.

Skorupiński, P.M. (2003a). Geneza pojęcia trafności teoretycznej. W: B. Niemierko (red.), *Trafność pomiaru jako podstawa obiektywizacji egzaminów szkolnych* (61-68). Łódź: Wydawnictwo Wyższej Szkoły Humanistyczno-Ekonomicznej w Łodzi.

Skorupiński, P.M. (2003b). Kryterialny wymiar trafności interpretacji wyników egzaminu gimnazjalnego. W: B. Niemierko (red.), *Trafność pomiaru jako podstawa obiektywizacji egzaminów szkolnych* (219-228). Łódź: Wydawnictwo Wyższej Szkoły Humanistyczno-Ekonomicznej w Łodzi.

Skorupiński, P.M. (2013). Modele trafności pomiaru. W: M. Karwowski (red.), *Ścieżki rozwoju edukacyjnego młodzieży – szkoły pogimnazjalne. Trafność wskaźników edukacyjnej wartości dodanej dla szkół maturalnych* (13–26). Warszawa: Wydawnictwo IFiS PAN.

Skórska, P., Świst, K. i Szaleniec, H. (2014a). Konsekwencje błędnego określenia rodzaju zadania testowego. *Edukacja, 2*(127), 67-84.

Skórska, P., Świst, K. i Szaleniec, H. (2014b). Szacowanie trafności predykcyjnej ocen szkolnych z wykorzystaniem hierarchicznego modelowania liniowego. *Edukacja, 3*(128), 77-96.

Stobart, G. (2001). The Validity of National Curriculum Assessment. *British Journal of Educational Studies*, 49(1), 26-39.

Szaleniec, H. i Szmigel, M.K. (2001). *Egzaminy zewnętrzne. Podnoszenie kompetencji nauczycieli w zakresie oceniania zewnętrznego*. Wydawnictwo Zamiast Korepetycji, Kraków.

Szaleniec, H. (2006). Oszukiwanie na egzaminie istotnym źródłem majowej porażki. W: B. Niemierko i M.K. Szmigel (red.), *O wyższą jakość egzaminów szkolnych*. Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Szaleniec, H. (2011). Wyniki egzaminu a ewaluacja zewnętrzna szkoły. W: B. Niemierko i K. Szmigel (red). *Ewaluacja w edukacji: koncepcje, metody, perspektywy* (39-46). Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Szaleniec, H., Grudniewska, M., Kondratek, B., Kulon, F., Pokropek, A., Stożek, E. i Żółtak, M. (2013). *Analiza Porównawcza Wyników Egzaminów Zewnętrznych - Sprawdzian w Szóstej Klasie Szkoły Podstawowej i Egzamin Gimnazjalny*. Warszawa: Instytut Badań Edukacyjnych. http://ibe.edu.pl/

Szaleniec, H., Kondratek, B., Kulon, F., Pokropek, A., Skórska, P., Świst, K., Wołodźko, T. i Żółtak, M. (2015). *Porównywalne wyniki egzaminacyjne*. Warszawa: Instytut Badań Edukacyjnych.

Zumbo, B.D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. W: R.W. Lissitz (red.), *The concept of validity: Revisions, new directions and applications* (65–82). Charlotte, NC: IAP—Information Age Publishing, Inc.

Zwick, R., i Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement, 48*(2), 101–121.

3. Porównywalne wyniki egzaminacyjne

Artur Pokropek, Henryk Szaleniec, Bartosz Kondratek, Filip Kulon, Paulina Skórska, Karolina Świst, Tymoteusz Wołodźko, Mateusz Żółtak

Wstęp

Od momentu wprowadzenia w Polsce sytemu egzaminów zewnętrznych uzyskiwane co roku przez uczniów wyniki ze sprawdzianu, egzaminu gimnazjalnego oraz egzaminu maturalnego są przedmiotem różnorakich debat i analiz. Częstokroć rozważaniom na temat wyników egzaminacyjnych towarzyszą próby zinterpretowania ich w terminach zmiany w poziomie umiejętności uczniów poprzez porównywanie średnich wyników między latami bądź, w przypadku egzaminu maturalnego, odsetka uczniów zdających egzamin. Tego typu działania są odzwierciedleniem zapotrzebowania na informację, która umożliwiałaby monitorowanie efektywności systemu edukacyjnego na poziomie całego kraju jak i wybranych rejonów czy nawet pojedynczych szkół.

Niestety, korzystanie z surowych wyników dostarczanych przez system egzaminów zewnętrznych nie uprawnia do przeprowadzania wnioskowania na temat zmian umiejętności uczniów na przestrzeni lat ze względu na brak procedur umożliwiających kontrolę różnic w trudnościach poszczególnych edycji egzaminów. Problem braku porównywalności wyników egzaminacyjnych dotyczy nie tylko prowadzenia polityki edukacyjnej, ale także pojawia się, gdy wyniki egzaminacyjne z różnych lat lub różnych sesji służą do celów diagnostycznych lub selekcyjnych.

W kontekście decyzji na poziomie jednostkowym najbardziej jaskrawym oraz społecznie i prawnie istotnym przykładem dla zilustrowania problemu braku porównywalności egzaminów jest zadanie porównania wyników maturalnych kandydatów na studia, którzy zdawali egzamin maturalny w różnych latach. Zadanie to, przed którym co roku stoją uczelnie wyższe, przeprowadzając nabór, jest obecnie niewykonalne bez poczynienia pewnych bardzo silnych założeń. Jednym rozwiązaniem jest założenie, że nie następują żadne zmiany w poziomie umiejętności uczniów podchodzących do różnych edycji egzaminów i selekcja kandydatów na podstawie wyników centylowych. Innym rozwiązaniem jest założenie braku zmian w trudności egzaminów i selekcja kandydatów na podstawie procentu zdobytych na maturze punktów. Jednak ze względu na obserwowane z roku na rok zmiany w rozkładach wyników egzaminacyjnych wiadomo, że przynajmniej jedno z tych założeń musi być błędne. Jeżeli uwzględnić wyniki różnych badań i doświadczeń edukacyjnych innych krajów (Donlon, 1984; OECD, 2014; Szaleniec i in., 2013), należy stwierdzić, że najpewniej oba wymienione założenia są w jakimś stopniu błędne, co stawia pod znakiem zapytania trafność decyzji rekrutacyjnych podejmowanych w opisanej sytuacji.

Nakreślona powyżej diagnoza stanowi wyzwanie dla polskiego systemu egzaminacyjnego. Na brak porównywalności egzaminów z różnych lat również zwróciła wyraźnie uwagę Najwyższa Izba Kontroli w informacji o wynikach z kontroli działania systemu egzaminacyjnego (NIK, 2015), rekomendując wprowadzenie systemowych zmian takową umożliwiających.

Niniejszy rozdział ma na celu przybliżenie tematu porównywalności wyników egzaminacyjnych. Oprócz ogólnego wprowadzenia do zagadnienia (podrozdziału 3.1) opisane zostaną badania nad porównywalnością polskich egzaminów zewnętrznych przeprowadzone w Instytucie Badań Edukacyjnych w latach 2011–2015. Zadaniem projektu badawczego realizowanego w IBE było umieszczenie na wspólnej skali archiwalnych wyników zgromadzonych przez system egzaminów zewnętrznych poprzez wykorzystanie informacji o relatywnej trudności zadań z różnych edycji egzaminów zebranej podczas specjalnie zaprojektowanych i przeprowadzonych badań. Sposób, w jaki przeprowadzono badania, zostanie przybliżony w podrozdziale 3.2.

W podrozdziale 3.3 zostanie natomiast przedstawiony szereg analiz wykorzystujących wyliczone w efekcie prac projektu porównywalne wyniki egzaminacyjne. Analizy te ukażą, jak wyrażone na wspólnej skali oszacowania poziomu umiejętności uczniów ze sprawdzianu, egzaminu gimnazjalnego oraz dwóch egzaminów maturalnych zmieniają się na przestrzeni lat. Podrozdział ten stanowi również ilustrację tego, jakie korzyści wynikają z możliwości przeprowadzania analiz na wynikach porównywalnych. W podrozdziale 3.4 temat przeprowadzania analiz na porównywalnych wynikach egzaminacyjnych będzie kontynuowany. Ogólnodostępna strona internetowa poświęcona porównywalnym między latami wynikom egzaminów (http://pwe.ibe.edu.pl/) pozwala na samodzielne przeprowadzanie analiz z wykorzystaniem bazy danych z porównywalnymi wynikami egzaminacyjnymi, dzięki czemu stanowi elastyczne narzędzie do monitorowania procesów edukacyjnych na wybranym poziomie analizy (szkoła, jednostki podziału terytorialnego).

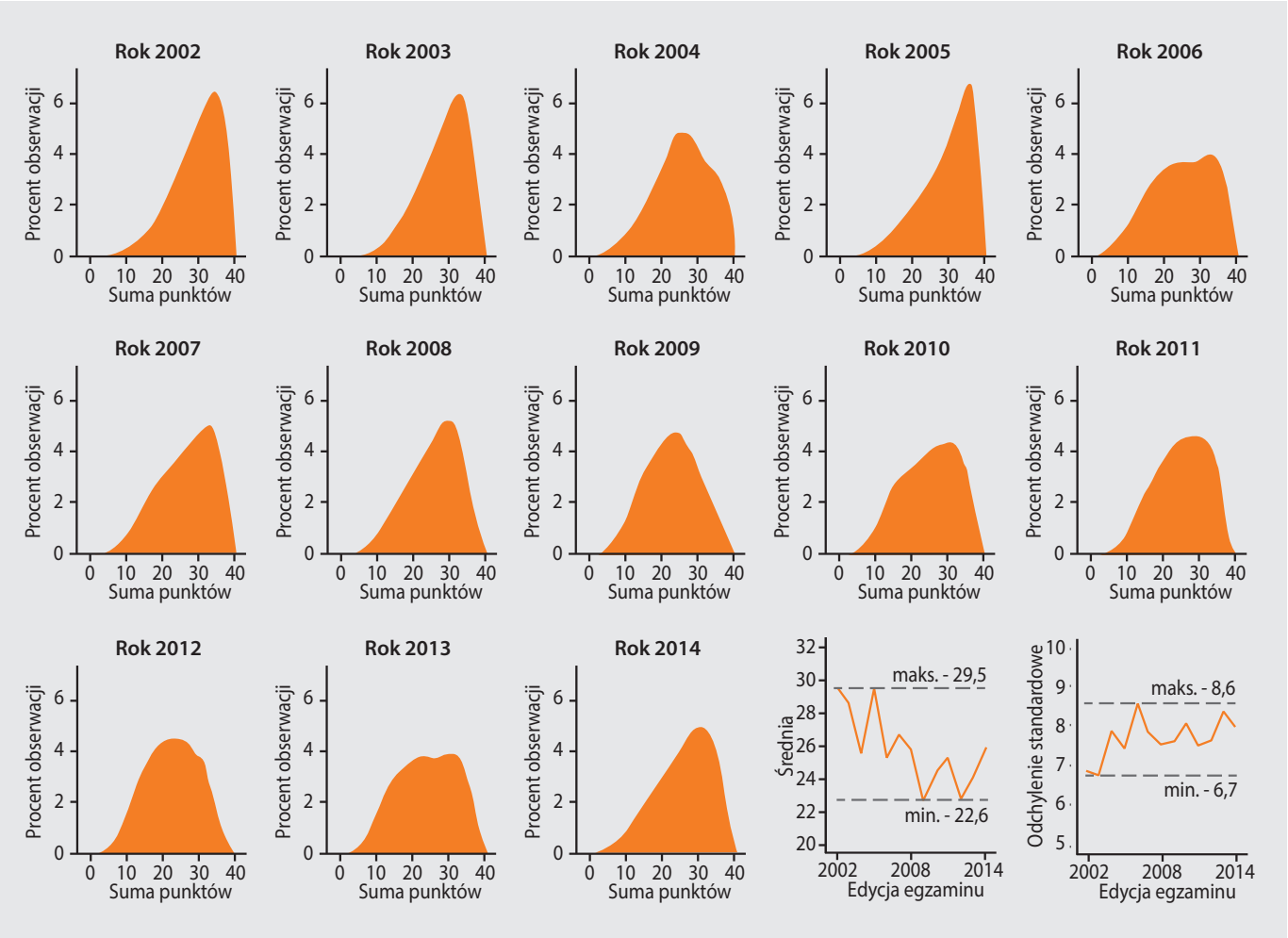
Na koniec w podrozdziale 3.5 rozważone zostaną trzy różne scenariusze, zgodnie z którymi można wprowadzić rozwiązania umożliwiające porównywalność wyników egzaminacyjnych do polskiego systemu egzaminów zewnętrznych. Opisany w pierwszej części rozdziału projekt badawczy, wraz ze swoimi rezultatami, stanowi próbę wykorzystania informacji zebranej przez system egzaminacyjny do wnioskowania o zmianach poziomu umiejętności uczniów na przestrzeni lat, jednak nie może on zastąpić rozwiązań systemowych. Przyjęta w badaniach przeprowadzonych przez IBE metodologia badania jest obarczona kilkoma istotnymi wadami (o których więcej w podrozdziale 3.2), od których właściwie zaimplementowane rozwiązania systemowe powinny być wolne, tak aby oferowane przez system egzaminów zewnętrznych porównywalne wyniki stanowiły możliwie najlepsze rozwiązanie problemu.

3.1. Porównywalność wyników egzaminacyjnych – wprowadzenie

3.1.1. Rozróżnienie między właściwościami egzaminu a właściwościami grupy uczniów podchodzących do egzaminu

Aby skonkretyzować nakreślony we wstępie problem zmieniających się właściwości egzaminu w zależności od jego edycji, na rysunku 3.1. zebrano rozkłady wyników uzyskiwanych przez uczniów podchodzących do sprawdzianu na przestrzeni lat 2002–2014. Możemy zaobserwować istotne zmiany kształtu rozkładu wyników: od zbliżonych do normalnego, np. w 2009 roku, do silnie skośnych, jak np. w 2005 roku. Zmianom kształtów rozkładów towarzyszyły również zmiany średniej oraz odchylenia standardowego rozkładu wyników z tego egzaminu (prawy dolny róg rysunku 3.1). W przypadku średniej wahania obejmowały zakres od 22,6 do 29,5 punktu z 40 punktowej skali testu, co odpowiada różnicy przeszło 17 punktów procentowych na skali odsetka maksymalnego możliwego do uzyskania wyniku, zatem mamy do czynienia z fluktuacją wyników o bardzo istotnych rozmiarach.

Rysunek 3.1. Rozkłady wyników ze sprawdzianu w latach 2002–2014. Dla zwiększenia czytelności skale dla średniej oraz odchylenia standardowego nie obejmują pełnego zakresu możliwych do uzyskania w teście punktów (0–40)

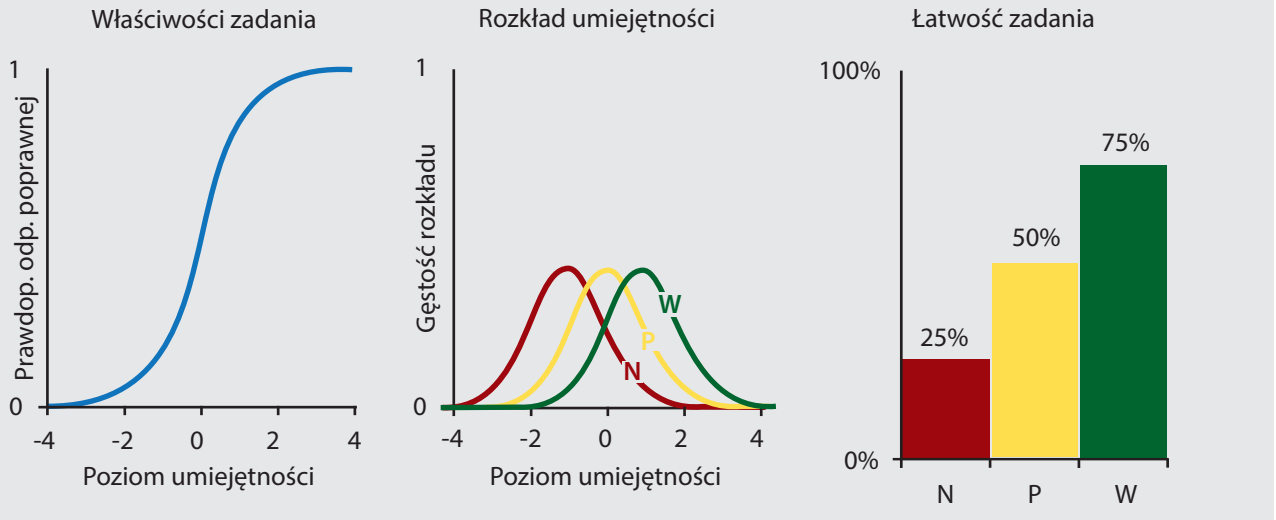


Źródłem widocznych na rysunku 3.1 zmian w rozkładach wyników sprawdzianu mogą być dwa niewykluczające się czynniki – zmiany w trudności zadań wchodzących w skład różnych arkuszy egzaminacyjnych albo zmiany w poziomie umiejętności uczniów między latami. Pierwszy czynnik z punktu widzenia oceny jakości egzaminów odgrywa negatywną rolę, gdyż oznacza brak stałości w zakresie kompetencji koniecznych do osiągnięcia sukcesu na egzaminie. Ponadto z pomiarowego

punktu widzenia trudność egzaminu powinna być dobrana optymalnie do celu, jakiemu egzamin służy (np. sprawdzian, jako test w zamierzeniu diagnostyczny, powinien mieć trudność na poziomie 50% – wartość, jakiej nigdy w historii nie osiągnął). O ile zjawiska zmian w trudności różnych arkuszy tego samego egzaminu nie da się wyeliminować całkowicie, należy jednak dążyć do jego świadomej minimalizacji, np. poprzez korzystanie przy tworzeniu egzaminów z tzw. banków zadań, stanowiących pulę zadań o wcześniej oszacowanych właściwościach psychometrycznych na podstawie starannie przeprowadzonych badań pilotażowych. Drugi z potencjalnych czynników odpowiedzialnych za zmiany w rozkładach wyników na rysunku 3.1, czyli zmiany w poziomie umiejętności uczniów między latami, jest bardzo interesujący np. z punktu widzenia polityki edukacyjnej. Bez zebrania pewnych dodatkowych danych oraz przeprowadzenia stosownej analizy rozdzielenie wkładu tych dwóch czynników w obserwowane zmiany wyników jest jednak niemożliwe.

Rozdzielenie źródeł zmienności rozkładów wyników egzaminu między latami na zmiany w ich trudności oraz na zmiany poziomu umiejętności uczniów ma kluczowe znaczenie zarówno dla zrozumienia, dlaczego surowe wyniki egzaminacyjne nie są porównywalne między latami, jak i dla zrozumienia, jak porównywalność egzaminów może zostać zapewniona. Na rysunku 3.2 oraz 3.3 zilustrowano problem rozdzielenia wkładu właściwości narzędzia pomiarowego od właściwości populacji uczniów przy interpretacji obserwowanych w teście wyników. Opisany w kolejnych akapitach przykład jednocześnie stanowi bardzo przystępne, nieformalne wprowadzenie do teorii odpowiedzi na pozycję testową (*item response theory,* IRT), będącej podstawowym współczesnym narzędziem statystycznym wykorzystywanym do analiz danych testowych, z której także korzystano przy konstrukcji porównywalnych wyników w projekcie badawczym IBE opisanym w dalszej części rozdziału. Czytelnik bliżej zainteresowany tematem IRT może sięgnąć do artykułu Bartosza Kondratka i Artura Pokropka (2013).

Rysunek 3.2. Zależność między rozkładem umiejętności w populacji a obserwowaną łatwością przy stałych właściwościach zadania

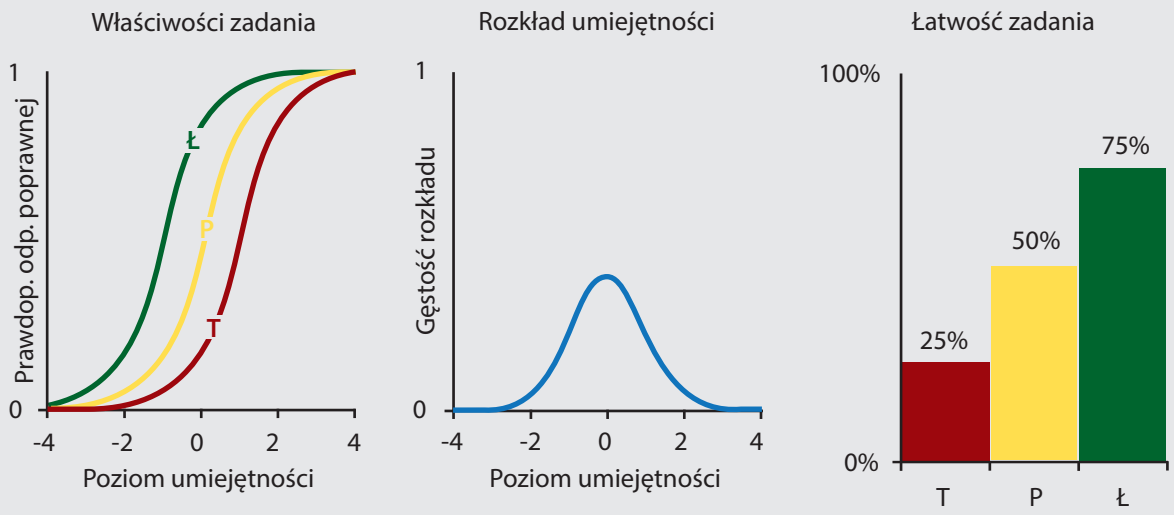


Przykład ograniczono do przypadku interpretacji obserwowanej łatwości pojedynczych zadań ocenianych dwukategorialnie, tj. takich, w których można udzielić odpowiedzi ocenianej jako poprawna lub błędna. Z lewej strony na rysunku 3.2 przedstawiono przykładową krzywą, która określa, jakie jest prawdopodobieństwo udzielenia w zadaniu poprawnej odpowiedzi w zależności od mierzonego za pomocą testu poziomu umiejętności ucznia. Taka s-kształtna krzywa nosi nazwę „krzywej charakterystycznej zadania" i określa w pełni jego właściwości psychometryczne – informuje o tym,

jak z rozwiązaniem danego zadania będą sobie radzić sobie uczniowie o różnych poziomach umiejętności. Sposób, w jaki rozumiany jest tu poziom umiejętności, jest specyficzny i wymaga komentarza. Mianowicie umiejętność jest tu tzw. „ukrytą zmienną", której skala jest niezależna od zadań, jakimi może być mierzona, jest ciągła oraz przyjmuje w populacji rozkład o kształcie zbliżonym do normalnego. Skala, na której umiejętność jest wyrażona, ma arbitralny charakter (umiejscowienie średniej oraz odchylenia standardowego) i jej wybór jest podyktowany głównie wygodą interpretacyjną. W omawianych przykładach dobrano rozkłady o odchyleniu standardowym równym jeden oraz średniej znajdującej się w punktach -1, 0 lub 1.

W przykładzie na rysunku 3.2 rozpatrywane jest pojedyncze zadanie o określonych właściwościach (z lewej) oraz trzy hipotetyczne grupy uczniów, których rozkłady umiejętności (w środku), różnią się tylko położeniem średniej – uczniowie o niskim (N), przeciętnym (P) oraz wysokim (W) poziomie umiejętności. Z prawej strony rysunku 3.2 ukazano natomiast, jaką zaobserwujemy łatwość takiego zadania (procent poprawnych odpowiedzi) w zależności od tego, w której z trzech populacji zadanie będzie rozwiązywane. W opisanym scenariuszu stałe właściwości zadania (może to być to samo zadanie, ale również trzy różne zadania o takich samych właściwościach) przekładają się na bardzo odmienną obserwowaną jego łatwość, wyłącznie z powodu różnicy w rozkładzie umiejętności uczniów je rozwiązujących.

Rysunek 3.3 Zależność między właściwościami zadania a obserwowaną łatwością przy stałym rozkładzie umiejętności w populacji



Na rysunku 3.3 przedstawiono przykład, w którym obserwowane różnice w łatwości zadania (z prawej) mają identyczne wartości jak na rysunku 3.2, jednak interpretacja przyczyn leżących u podstaw takiego rozkładu poprawnych odpowiedzi jest diametralnie różna. Mianowicie poziom umiejętności uczniów odpowiadajacych na zadanie (środek rysunku jest taki sam, ale właściwości zadania (z lewej) są odmienne. Zadanie o krzywej charakterystycznej najbardziej przesuniętej w lewą stronę jest zadaniem łatwym (Ł), zadanie o krzywej przesuniętej w prawą stronę jest zadaniem trudnym (T), a pomiędzy nimi znajduje się krzywa dla zadania o trudności przeciętnej (P). W tym przypadku zróżicowana obserwowana łatwość zadań odzwierciedla wyłącznie różnice w ich właściwościach.

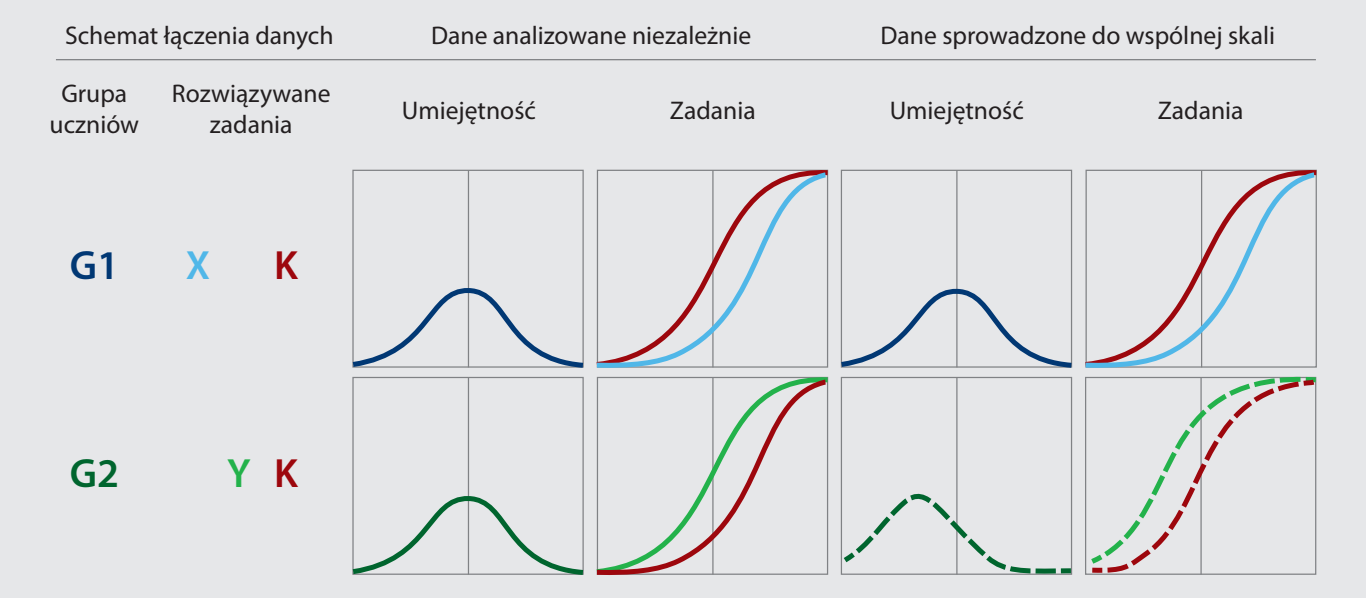
Z omówionych przykładów można łatwo przejść na poziom całego testu egzaminacyjnego. Sytuacja przedstawiona na rysunku 3.2 oznaczałby, że zadania z wszystkich egzaminów mają identyczne krzywe charakterystyczne, przez co różne edycje egzaminów są swoimi wiernymi psychometrycznymi kopiami, tzn. są równoważnymi i w pełni wymiennymi narzędziami pomiarowymi,

a obserwowane różnice w wynikach uczniów są wyłącznie odzwierciedleniem różnic w poziomie umiejętności. Natomiast rysunek 3.3 przedstawiałby sytuację, w której umiejętność uczniów między latami pozostaje bez zmian, a obserwowane różnice w rozkładach odpowiedzi są wyłącznie konsekwencją różnic we właściwościach egzaminu zastosowanego w danym roku. Najważniejszym wnioskiem z przykładów 3.2 oraz 3.3 jest jednak to, że analizując wyłącznie obserwowane wyniki niezależnie skonstruowanych wersji egzaminów – jak to ma miejsce w przykładzie sprawdzianu na rysunku 3.1 – nie jesteśmy w stanie rozstrzygnąć przyczyny obserwowanych różnic w rozkładach.
Aby móc rozdzielić efekt właściwości zadań od efektu właściwości populacji uczniów, konieczne jest stworzenie odpowiednich „połączeń" między danymi egzaminacyjnymi z różnych lat. W dalszej części opisane zostaną dwie podstawowe strategie zbierania danych, jakie można zastosować, aby rozdzielenie efektu trudności od różnic w poziomie umiejętności było możliwe.

3.1.2. Schematy łączenia wyników egzaminacyjnych

W celu wyrażenia wyników z różnych edycji egzaminu na wspólnej skali umiejętności należy zebrać informacje umożliwiające oszacowanie relatywnej trudności zadań występujących w róznych arkuszach egzaminacyjnych. Schematy zbierania danych, które służą realizacji takiego zadania, noszą nazwę schematów łączenia wyników (ang. *linking designs*). Opisane zostaną dwa schematy zbierania danych dla dwóch grup uczniów rozwiązujących różne egzaminy. W pierwszym schemacie informacja łącząca jest zbierana poprzez wykorzystanie w dwóch różnych egzaminach wspólnych zadań (tzw. „zadań kotwiczących"). W drugim informacja łącząca jest zbierana w grupie uczniów, którzy wykonują test składający się z wybranych zadań wykorzystanych w różnych egzaminach, których wyniki chcemy sprowadzić do wspólnej skali (tzw. łączenie z wykorzystaniem „wspólnych osób").
Omawiając schematy łączenia wyników, skorzystamy także z wprowadzonego w poprzednim podrozdziale sposobu rozróżnienia na właściwości grupy uczniów oraz właściwości zadań z wykorzystaniem ukrytej zmiennej umiejętności oraz krzywych charakterystycznych zadań. Dzięki temu będzie możliwe obrazowe zilustrowanie, w jaki sposób zaimplementowanie danego schematu umożliwia kontrolę różnic w trudności egzaminów i sprowadzenie uczniowskich wyników do wspólnej skali. Należy zauważyć, że dokonany wybór schematów zrównywania nie jest przypadkowy, gdyż pierwszy schemat stanowi najbardziej preferowane w podrozdziale 3.5 rozwiązanie systemowe dla polskiego systemu egzaminów zewnętrznych, a drugi schemat został wykorzystany w badaniach nad porównywalnością wyników egzaminacyjnych przeprowadzonych w IBE, o których będzie mowa w podrozdziałach 3.2, 3.3 oraz 3.4.
Schemat wykorzystujący łączenie poprzez zadania kotwiczące został przedstawiony z lewej strony na Rysunku 3.4. Grupa uczniów G1 rozwiązuje arkusz egzaminacyjny X, natomiast grupa G2 rozwiązuje inny arkusz egzaminacyjny Y, oprócz tego obie grupy rozwiązują pewną pulę dodatkowych zadań kotwiczących, oznaczonych literą K. W środkowej części rysunku 3.4 pokazano, jak mogłoby wyglądać, przy niezależnej analizie danych z dwóch grup, rozbicie na rozkłady umiejętności w tych grupach oraz na właściwości trzech zadań – po jednym z każdego egzaminu oraz jednego wspólnego zadania kotwiczącego. Widzimy, że przyjmując założenie o takim samym poziomie umiejętności, uzyskaliśmy różniące się kształtem krzywe charakterystyczne dla tego samego zadania kotwiczącego (kolor czerwony). Krzywa dla zadania kotwiczącego jest w grupie G2 przesunięta w prawo w porównaniu do grupy G1, co oznacza, że wyniki takiej analizy nie są wyrażone na wspólnej skali – to samo zadanie powinno tak samo się zachowywać względem tego samego poziomu umiejętności.

Rysunek 3.4. Schemat łączenia wyników z wykorzystaniem zadań kotwiczących wraz z ilustracją sprowadzania do wspólnej skali. Kolorami zakodowano informacje odnoszące się do tych samych grup zadań oraz grup uczniów



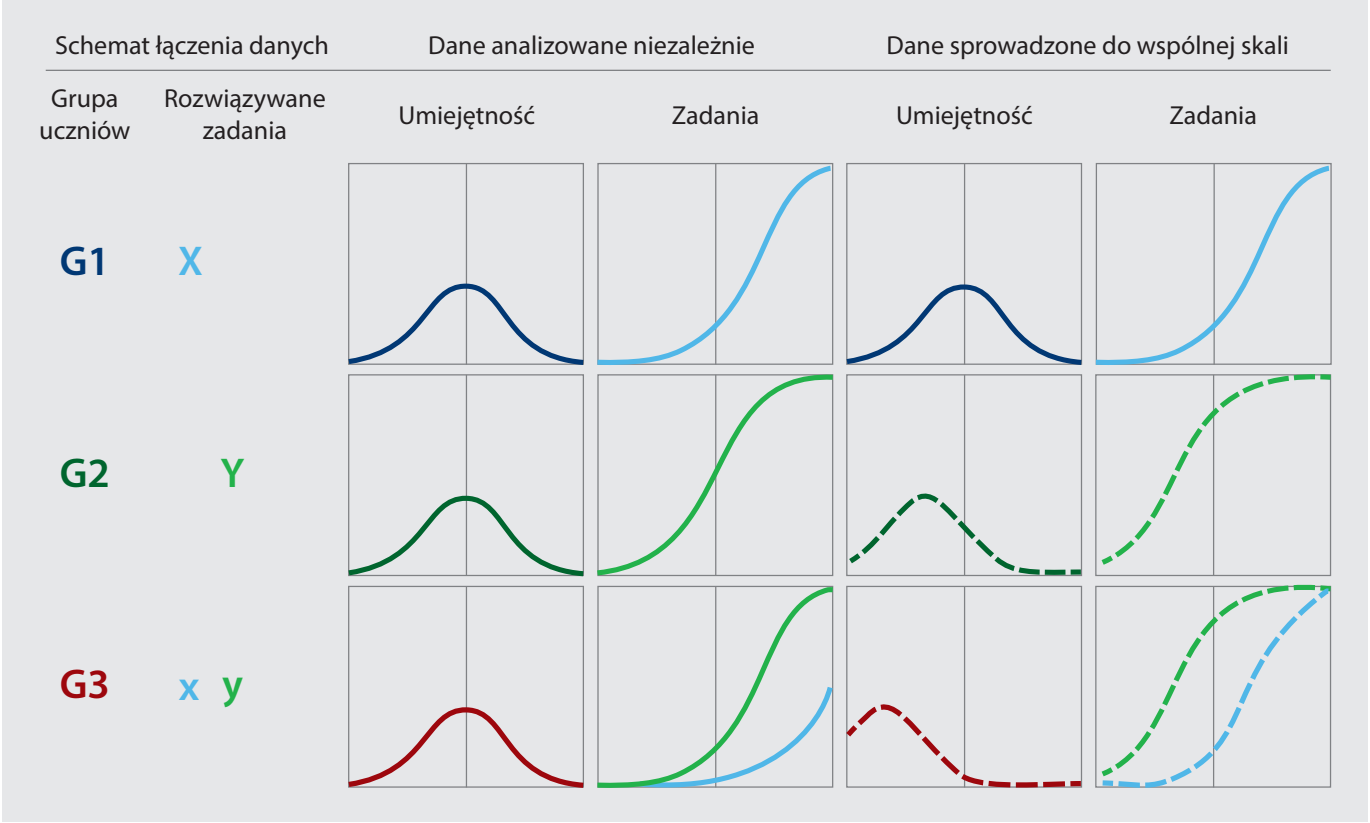
Przyjmijmy, że pragnęlibyśmy wyrazić poziom umiejętności uczniów z obu grup na skali umiejętności grupy G1. Na naszym uproszczonym, trzyzadaniowym przykładzie będzie to oznaczało przesunięcie rozkładu umiejętności uczniów z grupy G2 w lewo tak, aby uzyskać pokrywanie się krzywej dla zadania kotwiczącego w tej grupie z jej kształtem w grupie G1. Rezultat takiego sprowadzającego do wspólnej skali przekształcenia widać po prawej stronie rysunku 3.4 – w tym hipotetycznym przykładzie uczniowie G2 mają niższy poziom umiejętności niż G1.

Widzimy zatem, że zebrane poprzez zastosowanie wspólnych zadań dane dostarczają brakującej informacji o relatywnym położeniu rozkładów umiejętności dwóch różnych grup uczniów rozwiązujących różne testy. Jednocześnie oznacza to również możliwość porównania trudności dwóch egzaminów. W przykładzie z rysunku 3.4, po sprowadzeniu właściwości grup uczniów i zadań do wspólnej skali, okazuje się, że analizowane zadanie z testu Y jest znacznie łatwiejsze od zadania z testu X, gdyż krzywa charakterystyczna tego zadania (zielona) jest przesunięta w lewo w porównaniu z krzywą charakterystyczną zadania z testu X (niebieska). Należy przy tym zauważyć, że sprowadzona do wspólnej skali różnica w łatwościach tych dwóch zadań, rozumiana jako odległość między krzywymi, jest większa niż wynikałoby to z niezależnej analizy egzaminów.

Nic nie stoi na przeszkodzie, żeby porównanie łatwości zadań przeprowadzić również na poziomie obserwowanego odsetka odpowiedzi poprawnych w całych grupach uczniów – zastosowane w przykładzie zadania i rozkłady umiejętności wykorzystano już wcześniej przy okazji rysunków 3.2 oraz 3.3, dzięki czemu znamy ich łatwość w klasycznym rozumieniu. Wybrane zadanie z egzaminu X ma w grupie G1 łatwość 25%, natomiast wybrane zadanie z egzaminu Y ma w grupie G2 łatwość 50%. Dzięki sprowadzeniu do wspólnej skali możemy natomiast stwierdzić, że gdyby uczniowie z grupy G1 rozwiązywali wybrane zadanie z testu Y (którego faktycznie nie rozwiązywali), to uzyskaliby w tym zadaniu 75% poprawnych odpowiedzi.

W analogiczny sposób na rysunku 3.5 przedstawiono schemat łączenia danych z wykorzystaniem dodatkowej grupy uczniów rozwiązujących test składający się jednocześnie z pewnej puli zadań dobranych z egzaminu X (x) oraz z egzaminu Y (y). Przy ilustracji sprowadzania do wspólnej skali wykorzystano takie same rozkłady umiejętności i właściwości zadań wybranych z testów X oraz Y na rysunku 3.4. Różnica w tym schemacie łączenia polega na tym, że nie ma zadań kotwiczących, wykorzystanych w obu grupach piszących egzamin. Informacja łącząca, konieczna do sprowadzenia do wspólnej skali, jest dostarczana przez dodatkową grupę uczniów rozwiązujących zadania z dwóch różnych egzaminów.

Rysunek 3.5. Schemat łączenia wyników z wykorzystaniem grupy uczniów rozwiązującej zadania z obu egzaminów wraz z ilustracją sprowadzania do wspólnej skali. Kolorami zakodowano informacje odnoszące się do tych samych grup zadań oraz grup uczniów



Na dole, w środkowej części rysunku 3.5 widać, że analiza wyników uczniów z G3 pozwala na umiejscowienie na tej samej skali krzywych charakterystycznych dwóch zadań z różnych egzaminów, tym samym dostarczając informacji o ich relatywnej trudności. Jeżeli pragnęlibyśmy ponownie sprowadzić wszystkie wyniki do wspólnej skali właściwej grupie G1, to trzeba byłoby najpierw przesunąć krzywe i rozkład w grupie G3 tak, aby krzywa charakterystyczna zadania z egzaminu X (niebieska) pokrywała się z kształtem uzyskanym w G1, a następnie tak przesunąć krzywe i rozkład w grupie G2, aby krzywa dla zadania z egzaminu Y (zielona) pokrywała się z wcześniejszym przekształceniem dla grupy G3. Rezultat opisanych operacji jest przedstawiony z prawej strony rysunku 3.5. Uzyskane sprowadzenie do wspólnej skali jest dla grup G1 oraz G2 identyczne jak w przykładzie schematu łączenia z wspólnymi zadaniami, bez zmian pozostaje zatem również interpretacja. Nową informacją, jaką uzyskujemy z rysunku 3.5, jest wniosek o poziomie umiejętności uczniów z grupy G3, którzy w dobranym przykładzie okazują się grupą o najniższym poziomie umiejętności. Jednak z praktycznego punktu widzenia poziom umiejętności uczniów w grupie G3 sam w sobie jednak rzadko ma istotne znaczenie, gdyż uczniowie ci odgrywają w tym schemacie zazwyczaj jedynie techniczną rolę „dawców informacji" o relatywnej trudności zadań.

Mimo iż na przedstawionej prezentacji schematu łączenia wyników za pomocą wspólnych zadań jak i za pomocą wspólnych osób, uzyskano identyczne wnioski na temat różnic w rozkładach umiejętności uczniów oraz łatwości przykładowych zadań, należy zaznaczyć, że te dwa schematy różnią się istotnie w kilku ważnych aspektach. Schemat wykorzystujący zadania kotwiczące charakteryzuje się metodologiczną wyższością, gdyż uczniowie, rozwiązując zadania dostarczające informacji łączącej, robią to w tym samym momencie i w tej samej sytuacji motywacyjnej, w której rozwiązują właściwe zadania egzaminu. W kontekście uzyskiwania porównywalności egzaminów przeprowadzanie badania na dodatkowej grupie uczniów w celu uzyskania informacji łączącej będzie się zawsze wiązało z niebezpieczeństwem trudnych do skontrolowania efektów motywacyjnych bądź brakiem

pewności co do zakresu opanowanej przez tych uczniów wiedzy. Przeprowadzenie dodatkowego badania przed właściwym egzaminem na wystarczająco dużej próbie (np. przy okazji standaryzacji) niesie ryzyko ujawnienia treści zadań, natomiast przeprowadzenie po egzaminie wprowadza problemy z potencjalną znajomością zadań. Dodatkowo schemat z zadaniami kotwiczącymi charakteryzuje się większą mocą wnioskowania statystycznego. Do wątku porównania tych strategii zbierania danych łączących powrócimy pod koniec rozdziału, przy okazji rekomendacji dla systemu egzaminów zewnętrznych.

3.1.3. Sprowadzanie wyników egzaminów do wspólnej skali a zrównywanie wyników egzaminacyjnych

Wymagającą wyjaśnienia kwestią jest rozróżnie między dwoma spokrewnionymi, ale nie tożsamymi, pojęciami, które mogą się pojawić w kontekście uzyskiwania porównywalnych wyników egzaminacyjnych. Pierwszym terminem jest stosowane już wielokrotnie „sprowadzanie wyników do wspólnej skali", a drugim jest „zrównywanie wyników" (ang. *test equating*).

Poprzez sprowadzanie wyników do wspólnej skali rozumie się ogólnie procedury, które umożliwiają wyrażenie wyników z różnych egzaminów w taki sposób, aby nie były obciążone efektem różnic w ich trudnościach. W szczególności przybliżone wcześniej modelowanie tak zwanej ukrytej zmiennej umiejętności, przy wykorzystaniu informacji łączącej wyniki egzaminów z różnych edycji, stanowi przykład sprowadzania do wspólnej skali, w wyniku którego osiągana jest porównywalność wyników egzaminacyjnych. Innym przykładem sprowadzania do wspólnej skali, w wyniku którego można otrzymać porównywalne wyniki, będzie właśnie ich zrównywanie.

Zrównywanie wyników jest procedurą, której celem jest uzyskanie porównywalnych wyników z dwóch różnych narzędzi pomiarowych na skali wyników obserwowanych, tzn. na dobrze znanej skali sumy punktów uzyskanych w teście. Celem zrównywania jest wyznaczenie przekształceń, które przeliczałyby wynik surowy w jednej edycji egzaminu (X) na wynik surowy w innej edycji egzaminu (Y) w taki sposób, aby nie miało znaczenia, czy posługujemy się wynikiem X, czy też zrównanym do skali X wynikiem egzaminu Y (i na odwrót). Zasadnicza różnica między zrównywaniem wyników a sprowadzaniem do wspólnej skali poprzez modelowanie zmiennej ukrytej sprowadza się własnie do skali, na której wyniki są prezentowane.

W omawianych w dalszej części rozdziału badaniach IBE dotyczących porównywalnych egzaminów zewnętrznych prezentowane będą przede wszystkim wyniki na skali zmiennej ukrytej, gdyż korzystanie z niej znacznie ułatwia interpretację wyników oraz umożliwia łatwe budowanie przedziałów ufności niosących informację o zakresie pewności statystycznej prezentowanych zależności. W ramach projektu badawczego zostało przeprowadzone również zrównywanie wyników, tj. zbudowanie porównywalnych wyników na skali surowych punktów uzyskiwanych w teście, jednak nie będzie tu szerzej analizowane – zainteresowanego Czytelnika odsyłamy do raportu z badania (Szaleniec i in.,2013).

3.2. Jak powstają porównywalne wyniki egzaminacyjne w Polsce?

W przeciwieństwie do dużej liczby krajów stosujących standaryzowane testy egzaminacyjne (Pokropek i Kondratek, 2012), zrównywanie wyników egzaminacyjnych w Polsce nie zostało zintegrowane z systemem egzaminacyjnym. Porównywalność wyników z roku na rok i z sesji na sesję nie jest zatem bezpośrednio zapewniona. System egzaminacyjny nie daje możliwości określenia, czy dwa egzaminy z różnych lat były tak samo trudne. Nie można też na jego podstawie stwierdzić, czy uczniowie z jednego rocznika posiadają wyższy poziom umiejętności, czy niższy, niż ich rówieśnicy z innego rocznika. Jednym ze sposobów na uzyskanie porównywalnych wyników, poza ramami systemu egzaminacyjnego, jest przeprowadzenie oddzielnych badań zrównujących. Takie badania zaplanowane i przeprowadzone zostały przez IBE.

Badania zrównujące zaplanowane zostały na kilka lat i obejmowały kolejno egzamin gimnazjalny, sprawdzian na zakończenie szkoły podstawowej i wybrane przedmioty na poziomie maturalnym. Pierwszy etap badań przeprowadzono w 2011 roku. W tym etapie zebrano dane pozwalające na zrównanie egzaminu gimnazjalnego dla lat 2002–2010. Drugi etap badań, przeprowadzony w 2012 roku, dotyczył głównie sprawdzianu i jego podstawowym celem było zrównanie wyników sprawdzianu przeprowadzonego w latach 2002–2011. W tym etapie zebrano również dane pozwalające na kontynuację zrównywania egzaminu gimnazjalnego. Trzeci etap, przeprowadzony w roku 2013, dotyczył gównie egzaminu maturalnego z matematyki, ale zebrano również dane pozwalające na kontynuację zrównania egzaminów z niższych szczebli edukacji. Czwarty etap, którego realizacja przypadła na 2014 rok, skoncentrowany był na wybranych egzaminach maturalnych – języku polskim oraz angielskim. Dodatkowym elementem czwartego etapu badań było zebranie danych umożliwiających kontynuację wcześniejszych etapów zrównywania.

Badania zrównujące przeprowadzone przez IBE miały charakter zrównywania post factum, czyli losowa próba uczniów z danego rocznika (2011, 2012, 2013 lub 2014) rozwiązywała zadania egzaminacyjne z wcześniejszych lat. Następie na podstawie informacji o rozwiązanych zadaniach z różnych edycji egzaminu pierwotne wyniki były przekształcane w taki sposób, aby zapewnić porównywalność. Wykorzystano w tym celu metodę kalibracji łącznej, zgodnie z którą w jednym kroku, na wspólnej skali, szacowano parametry wielogrupowego modelu IRT dla próby zrównującej i zbiorów danych egzaminacyjnych.

W celu prezentacji wyników przyjęto skalę, dla której wartość 100 odpowiada średniemu wynikowi uczniów piszących arkusz standardowy egzaminu w 2012 roku, natomiast różnica 15 punktów na skali odpowiada jednemu odchyleniu standardowemu wyników uczniów z tego egzaminu. Rok 2012 został w tym wypadku wybrany arbitralnie – z metodologicznego punktu widzenia mógłby to być dowolny rok spośród lat, za które zrównywane były wyniki.

Skala (100; 15), na której raportowane są porównywalne wyniki egzaminacyjne w badaniu, różni się w sposób istotny od skal, na jakich prezentowane są niezrównane wyniki egzaminów, gdyż odnosi się do tak zwanej „ukrytej zmiennej" umiejętności. Raportowane przez CKE wyniki egzaminów odnoszą się natomiast do skali opartej na sumie punktów uzyskanych za egzamin (liczba punktów uzyskanych na egzaminie i/lub skala procentowa od 0 do 100%), którą nazywa się skalą „wyników obserwowanych". Dodatkową korzyścią z korzystania ze skali zmiennej ukrytej jest fakt, iż jej rozkład dla tak dużych grup jak kohorta uczniów piszących powszechny egzamin w danym roku będzie miał kształt zbliżony do rozkładu normalnego, co bardzo ułatwia interpretację wyników – skala (100;15) przy założeniu normalności staje się jedną z popularniejszych skal standardowych, wykorzystywaną np. przy pomiarze inteligencji (skala IQ).

Główna zaleta przyjętego planu i metodologii badania związana jest z tym, że dotyczyło ono zrównywania wyników egzaminów, które już się odbyły. Nie wymagało to wprowadzenia żadnych zmian w organizacji prawdziwego egzaminu ani żadnych zmian prawnych, które wiązałyby się na przykład z utajnieniem części zadań (czyli w rozwiązaniu stosowanym np. w amerykańskim teście SAT). Przyjęty plan obarczony jest jednak kilkoma istotnymi wadami. Nie jest on odporny na dwa potencjalne czynniki zakłócające:

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

Wymienione dwie, potencjalnie zakłócające, zmienne są względem siebie w opozycji. Czynnik motywacyjny powinien obniżać wyniki podczas sesji zrównującej w porównaniu z warunkami testu doniosłego. Natomiast czynnik ujawnienia zadań stawia w uprzywilejowanej pozycji uczniów z sesji zrównującej w porównaniu z ich kolegami i koleżankami widzącymi zadania po raz pierwszy podczas egzaminu. Jeżeli czynniki motywacji oraz ujawnienia zadań będą równomiernie działały

pomiędzy zeszytami zrównującymi, to nie powinny one wpłynąć na wyniki zrównania w sposób systematyczny, w innym przypadku oszacowanie zrównanych wyników może być systematycznie obciążone. Wskaźniki z badań pilotażowych wskazują na brak istotnych różnic między poziomem motywacji i znajomością zadań między różnymi zeszytami, co zwiększa wiarygodność otrzymanych wyników. Niestety trzeba pamiętać, iż wykorzystane wskaźniki kontrolowane były tylko deklaratywnymi miarami i nie zapewniały pełnej walidacji spełnienia założeń niezbędnych do otrzymania nieobarczonych rezultatów.

3.3. Zastosowanie porównywalnych wyników egzaminacyjnych – przykładowe analizy

Wprowadzenie porównywalnych wyników egzaminacyjnych (PWE) pozwala na dokonanie szeregu analiz istotnych z perspektywy polskiej oświaty, które w pełnym kształcie nie mogły zostać przeprowadzone wcześniej. W tym rozdziale podjętych zostanie kilka tematów poruszanych w badaniach edukacyjnych, dotyczących zróżnicowania wyników egzaminacyjnych ze względu na:

- płeć uczniów (por. EACEA, 2010),
- diagnozę dysleksji u uczniów (por. Wejner, 2009),
- lokalizację szkoły (por. Dolata, 2008; Żółtak, 2011; Dolata, Jasińska i Modzelewski, 2012),
- typ szkoły (niepubliczna vs publiczna) (por. Putkiewicz i Wiłkomirska, 2004).

Choć zmienne te są ważne dla opisu systemu edukacji, należy pamiętać, że nie mamy podstaw dla jednoznacznego zidentyfikowania przyczyn występującego zróżnicowania lub jego braku, a także zmienności tych trendów w czasie. Niemniej jednak trafny i rzetelny opis zmian poziomu umiejętności uczniów w czasie jest niezbędnym punktem wyjścia do dalszych analiz.
Analizowane wyniki egzaminów przedstawiono w podziale na trzy etapy edukacyjne (szkoła podstawowa, gimnazjum i szkoła ponadgimnazjalna). Oprócz zrównanych wyników przedstawiamy również analizę zróżnicowania poziomu umiejętności uczniów ze względu na przynależność do szkół, z uwzględnieniem podziału na jednostki samorządu terytorialnego (powiaty i województwa). Wskaźnik międzyszkolnego zróżnicowania wyników uczniów mówi o tym, w jakim stopniu można przewidywać wynik ucznia na egzaminie na podstawie tego, którą skończył szkołę (Dolata, 2012). Analogicznie, wskaźnik zróżnicowania wyników między powiatami oraz między województwami mówi o możliwości przewidywania wyników uczniów na podstawie informacji o jednostkach administracyjnych, w których znajduje się szkoła. Zróżnicowanie międzyszkolne jest szczególnie ważne dla szkół podstawowych i gimnazjów, które z założenia powinny być jednolite i uczęszczanie do danej szkoły nie powinno różnicować szans poszczególnych uczniów na sukces edukacyjny (Dolata, 2010). Z rozwinięciem niniejszych analiz oraz szczegółowym przedstawieniem ich metodologii Czytelnik może się zapoznać w publikacjach Zespołu Analiz Osiągnięć Uczniów (Szaleniec i in., 2012; Szaleniec i in., 2013).

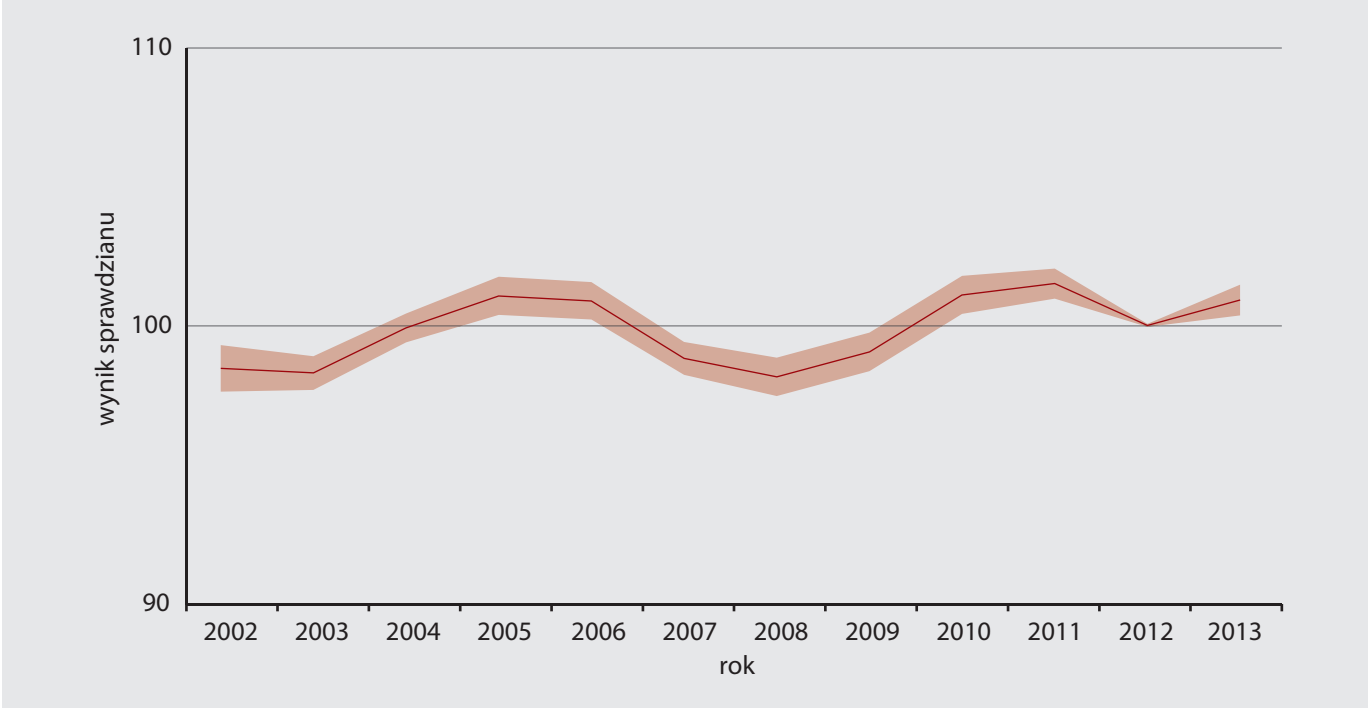
3.3.1. Sprawdzian

Rysunek 3.6 przedstawia porównywalne wyniki sprawdzianu w latach 2002–2013. Na osi pionowej prezentowane są zrównane wyniki (zrównanie przeprowadzono do roku 2012), a na osi poziomej kolejne edycje egzaminu. Wartość 100 odpowiada średniemu wynikowi sprawdzianu w 2012 roku, a różnica 15 punktów na skali odpowiada jednemu odchyleniu standardowemu wyników egzaminu w 2012 roku. Średni zrównany wynik sprawdzianu dla kraju oznaczono ciągłą linią – jaśniejszy pasek dookoła niej oznacza 95% przedział ufności[29]. Przedział ufności służy do zobrazowania niepewności, jaką obarczone jest szacowanie średniego zrównanego wyniku egzaminu. Szerszy przedział ufności

[29] Kolejne wykresy porównywalnych wyników egzaminacyjnych będą prezentowane w analogiczny sposób.

świadczy o większej niepewności, natomiast jeśli przedziały ufności dla dwóch wyników nachodzą na siebie, oznacza to, że nie mamy pewności co do tego, czy w rzeczywistości różnią się one między sobą. Średni porównywalny wynik sprawdzianu mówi o tym, jaki rezultat osiągnęliby uczniowie rozwiązujący sprawdzian w danym roku, gdyby rozwiązywali sprawdzian w roku bazowym (czyli w roku 2012). Więcej informacji na temat sposobu wizualizacji porównywalnych wyników egzaminacyjnych można znaleźć w rozdziale 3.4, w serwisie Porównywalnych Wyników Egzaminacyjnych pod adresem http://pwe.ibe.edu.pl/ oraz w publikacji Szaleńca i współpracowników (2013).

Rysunek 3.6. Porównywalne wyniki sprawdzianu w latach 2002–2013

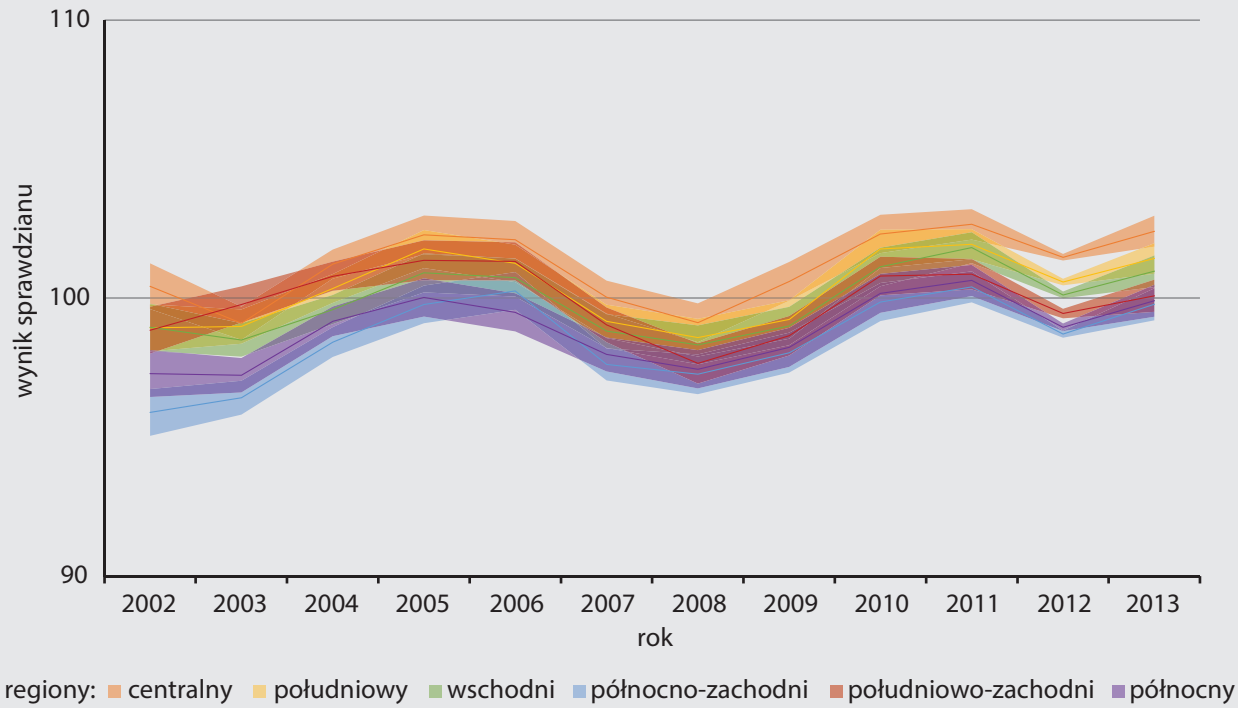


Wyniki sprawdzianu dla całego kraju podlegają niewielkim fluktuacjom na przestrzeni czasu, lecz nie przedstawiają wyraźnego trendu wzrostowego lub spadkowego. Przeciętne wyniki sprawdzianu statystycznie znacząco różnią się między latami, ale różnice te nie są większe niż 5 punktów.

W analizach dotyczących funkcjonowania systemu edukacji może interesować nas zróżnicowanie regionalne wyników, zwłaszcza w kontekście potencjalnego rozwoju regionów. W związku z tym, że zaobserwowano zależność na poziomie makroekonomicznym pomiędzy wynikami w testach umiejętności oraz poziomem rozwoju gospodarczego (Hanushek i Kim, 1995), można przypuszczać, że im lepsza edukacja w danym regionie, tym lepszy jego rozwój gospodarczy, konkurencyjność oraz poziom życia mieszkańców (Herbst, 2004; Kurek, 2010). Rysunek 3.7 przedstawia porównywalne wyniki sprawdzianu w rozbiciu na regiony w klasyfikacji zgodnej z Nomenklaturą Jednostek Terytorialnych do Celów Statystycznych (NTS)[30]. tabela 3.1 zawiera nazwy regionów wraz z wyszczególnieniem województw wchodzących w ich skład.

[30] *Rozporządzenie Rady Ministrów z dnia 13 lipca 2000 r. w sprawie wprowadzenia Nomenklatury Jednostek Terytorialnych do Celów Statystycznych* (NTS) (Dz. U. z 2000 r. Nr 58, poz. 685).

Rysunek 3.7. Porównywalne wyniki sprawdzianu w latach 2002–2013 w podziale na regiony wg NTS



W regionach centralnej, południowej i wschodniej Polski zaobserwowano wyższe wyniki egzaminacyjne niż w północnej i zachodniej części kraju. Wyniki te są zgodne z ustaleniami Herczyńskiego i Herbsta (2002) oraz Herbsta (2004), którzy zauważyli, że wyniki sprawdzianu i egzaminu gimnazjalnego uzyskiwane przez uczniów z zachodniej Polski są niższe niż we wschodniej części kraju. Zróżnicowanie to można tłumaczyć szeregiem uwarunkowań społeczno-ekonomicznych (np. stopa bezrobocia, poziom urbanizacji) i historycznych (obszary będące pod różnymi zaborami oraz w różnym stopniu obciążone rolnictwem opartym na PGR-ach). Ciekawym przypadkiem jest region południowo-zachodni, w którym wyniki w latach 2002–2007 były bliższe wynikom regionów: centralnego, południowego i wschodniego, a w latach 2008–2013 stały się relatywnie niższe.

Tabela 3.1. Regiony i województwa wchodzące w ich skład (wg NTS)

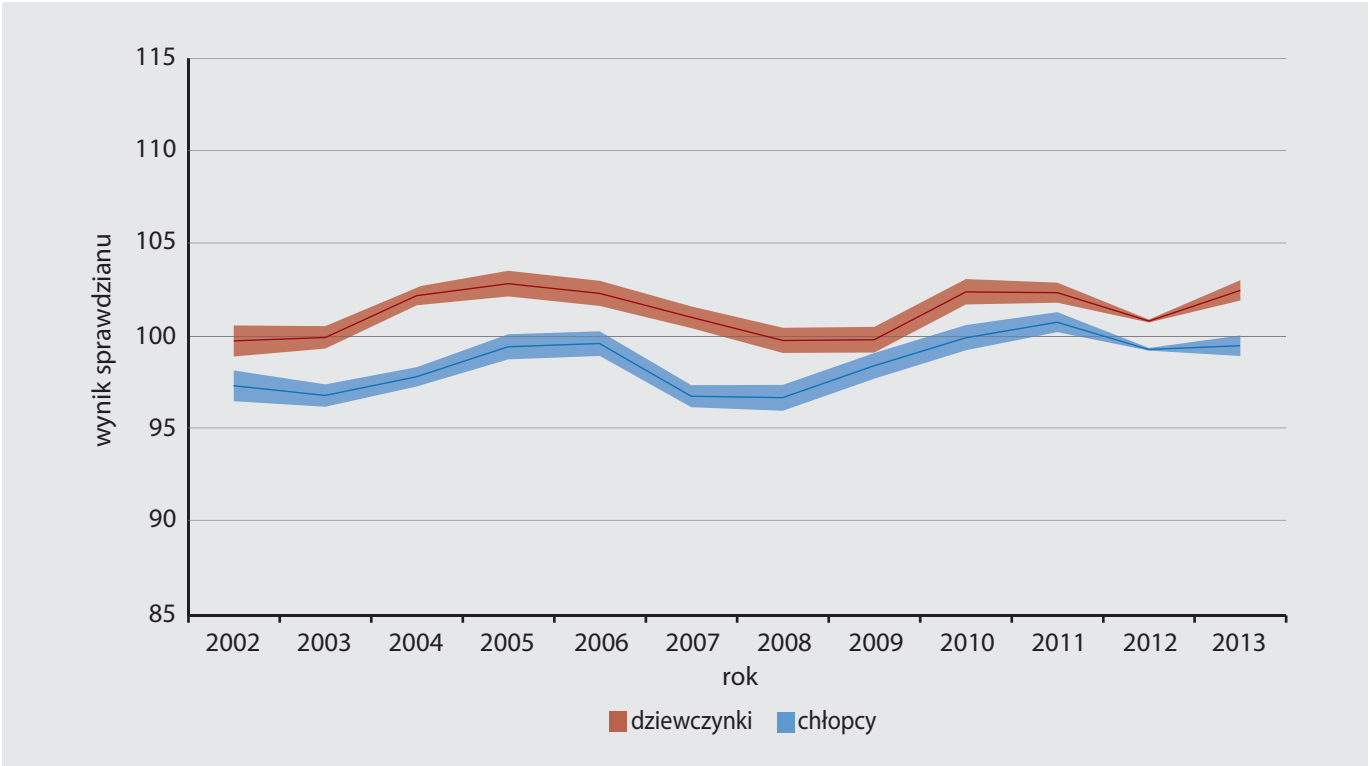
Region	Województwa
centralny	łódzkie i mazowieckie
południowy	małopolskie i śląskie
wschodni	lubelskie, podkarpackie, świętokrzyskie i podlaskie
północno-zachodni	wielkopolskie, zachodniopomorskie i lubuskie
południowo-zachodni	dolnośląskie i opolskie
północny	kujawsko-pomorskie, warmińsko-mazurskie i pomorskie

Aspekt równego dostępu do edukacji dla obu płci jest priorytetowy dla polityki równościowej (zob. EACEA, 2010; World Economic Forum, 2013). Badania dowodzą, że kobiety osiągają zazwyczaj wyższe oceny w szkole niż mężczyźni (m.in. Konarzewski, 1996), natomiast ich wyniki w standaryzowanych egzaminach zewnętrznych są niższe (np. Perkins, Kleiner, Roey i Brown, 2004; Noftle i Robins,

2007). Jednocześnie wskazuje się, że iloraz inteligencji kobiet nie różni się w sposób istotny statystycznie od ilorazu inteligencji mężczyzn. Efekt ten określa się mianem niedoszacowania predykcji wyników egzaminacyjnych dla kobiet (ang. *female underprediction effect*, FUE) i przeszacowania predykcji wyników egzaminacyjnych dla mężczyzn (zob. Hyde i Kling, 2001). Dodatkowo wskazuje się, że chłopcy osiągają wyższe wyniki z matematyki, a dziewczęta w czytaniu i pisaniu, jednakże rozbieżność ta pojawia się głównie na dalszych etapach edukacyjnych (Willingham i Cole, 1997).

rysunek 3.8 obrazuje porównywalne wyniki egzaminacyjne w poszczególnych latach dla grup chłopców i dziewcząt. W przypadku sprawdzianu różnica w wynikach egzaminacyjnych chłopców i dziewczynek waha się nieznacznie na przestrzeni czasu i wynosi około trzech punktów na korzyść kobiet, z wyjątkiem lat 2009, 2011 i 2012, kiedy jest nieco mniejsza i wynosi mniej niż dwa punkty. Można zatem dojść do wniosku, że w przypadku sprawdzianu nie mamy do czynienia z przeszacowaniem poziomu umiejętności dla chłopców – ich wyniki są w analizowanym okresie systematycznie niższe niż wyniki dziewczynek. Trendy wyników egzaminacyjnych w grupach chłopców i dziewczynek przebiegają podobnie do siebie. Należy jednak pamiętać, że w tym okresie sprawdzian miał charakter ponadprzedmiotowy i mierzył zarówno umiejętności czytania i pisania, jak również umiejętności matematyczne. W związku z tym ocena zróżnicowania wyników dla poszczególnych umiejętności w grupach chłopców i dziewcząt nie jest możliwa.

Rysunek 3.8. Porównywalne wyniki sprawdzianu w latach 2002–2013 w podziale na płeć uczniów

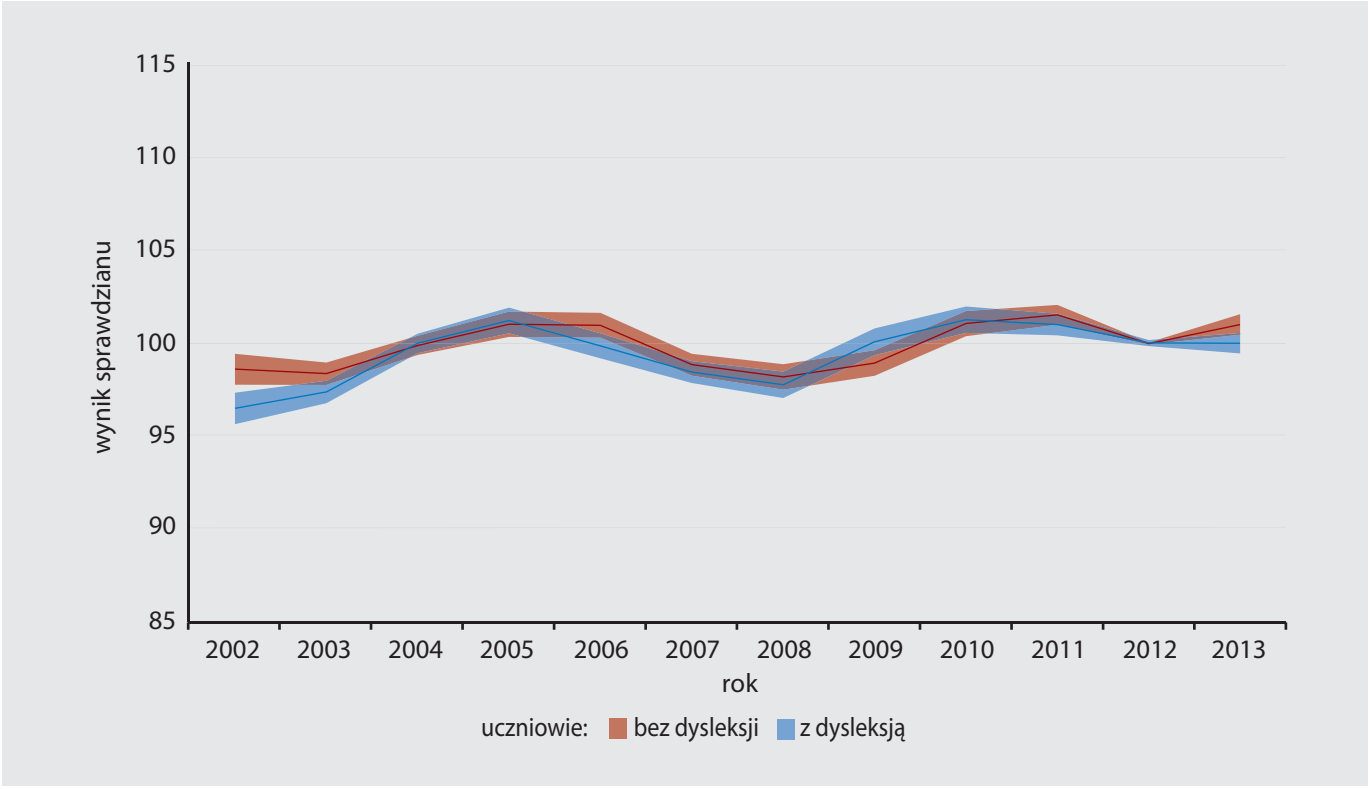


W związku z tym, że uczniowie ze zdiagnozowaną dysleksją rozwojową mogą korzystać ze specjalnie dostosowanych rozwiązań podczas sprawdzianu, należy zbadać efekty takiego postępowania. Wskazuje się (CKE, 2009), że uczniowie z specyficznymi trudnościami w uczeniu się czytania i pisania (z dysleksją rozwojową) mogą reagować zaburzeniami uwagi i pamięci podczas sytuacji stresującej (np. egzaminu zewnętrznego), a ich tempo pracy jest zdecydowanie wolniejsze. Mają oni także trudności z czytaniem, a ich pismo jest mało czytelne i zawiera wiele błędów. Wprowadzone dla nich dostosowania obejmują np. wydłużenie czasu pisania sprawdzianu do 50%, zgodnie z zaleceniami zawartymi w opinii wydanej przez poradnię psychologiczno-pedagogiczną lub poradnię specjalistyczną, odczytywanie zadań przez członków komisji czy specjalnie dostosowane kryteria

oceniania. Uczniowie ci nie muszą też przenosić odpowiedzi na karty odpowiedzi – robią to za nich egzaminatorzy (CKE, 2009).

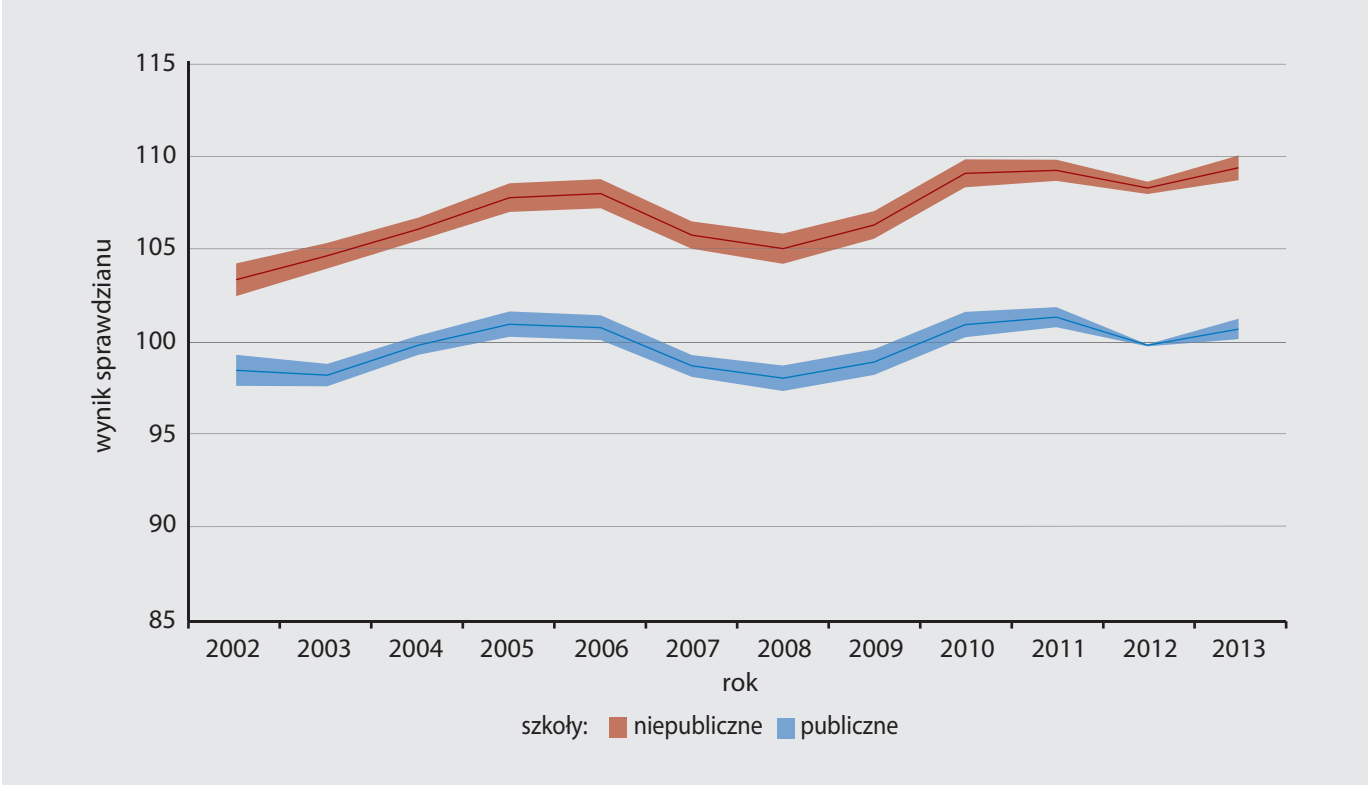
Porównywalne wyniki egzaminacyjne w poszczególnych latach dla grup uczniów z opinią o dysleksji i bez tej opinii ilustruje rysunek 3.9. Obserwujemy, że wyniki egzaminacyjne uczniów z dysleksją i bez dysleksji przeważnie nie różnią się od siebie w sposób istotny statystycznie (przedziały ufności na siebie nachodzą). Jedynie w roku 2002 można zaobserwować niewielkie różnice między wynikami tych dwóch grup, lecz z uwagi na to, że była to pierwsza edycja sprawdzianu, do jego wyników należy podchodzić bardzo ostrożnie. W związku z brakiem różnic wyników uczniów z dostosowaniami i bez nich można stwierdzić, że w odniesieniu do zjawiska dysleksji sprawdzian w poszczególnych edycjach zachowuje się stabilnie.

Rysunek 3.9. Porównywalne wyniki sprawdzianu w latach 2002–2013 w podziale na grupy bez dysleksji rozwojowej i z dysleksją rozwojową



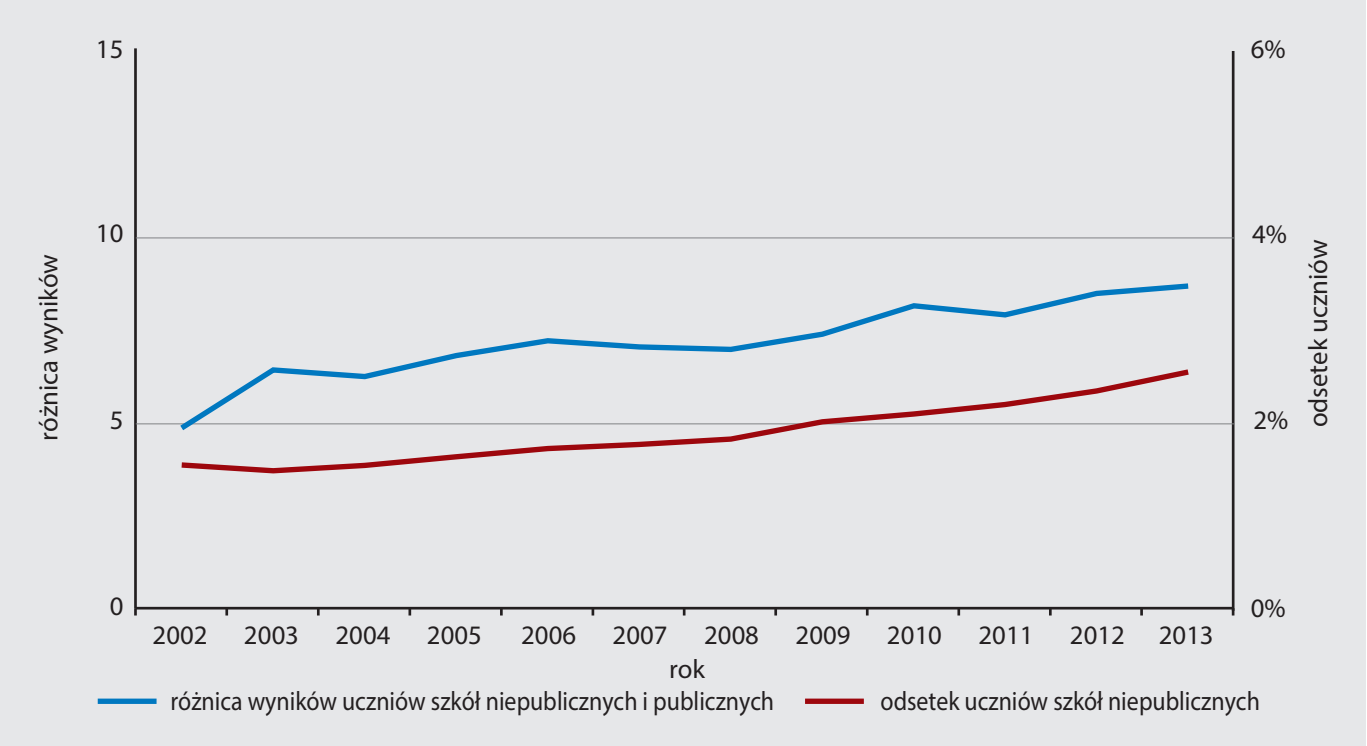
Kolejnym aspektem, który może różnicować wyniki uczniów, jest typ szkoły, do której uczęszczają (szkoły publiczne i niepubliczne). Jak wskazują Elżbieta Putkiewicz i Anna Wiłkomirska (2004), w szkołach niepublicznych typ kształcenia jest specyficzny: oddziały są mniejsze, nacisk kładzie się na rozwój indywidualnych zainteresowań uczniów, a możliwość zaimplementowania autorskich programów nauczania jest zdecydowanie większa. Według Marty Piekarczyk (2014) szkoły niepubliczne mają także zagwarantować lepsze traktowanie uczniów i większe bezpieczeństwo. Uczniowie uczęszczający do szkół publicznych i niepublicznych mogą także różnić się ze względu na status społeczno-ekonomiczny (SES) bądź wykształcenie rodziców (Jung-Miklaszewska i Rusakowska, 1995; Zahorska-Bugaj, 1994).

Rysunek 3.10 Porównywalne wyniki sprawdzianu w latach 2002–2013 w podziale na typ szkoły (niepubliczna vs publiczna)



Dla wszystkich analizowanych lat porównywalne wyniki sprawdzianu dla szkół niepublicznych są wyższe niż wyniki dla szkół publicznych (zob. rysunek 3.10). W roku 2002 dla prawie siedmiu procent uczniów nie udało się określić typu szkoły, więc dane z tego roku należy traktować ostrożnie, gdyż nie wiadomo, czy braki te mają charakter losowy. Warto jednak zwrócić uwagę, że różnica ta staje się z roku na rok coraz większa, w roku 2003 wynosiła 6,4 punktu, a w roku 2013 już 8,7 punktu. Jednocześnie zmieniają się proporcje uczniów uczących się w szkołach niepublicznych i publicznych – uczniowie szkół niepublicznych stanowili w 2003 roku 1,5% wszystkich uczniów zdających sprawdzian, a w roku 2013 r. 2,6%. Dane te dla całego rozpatrywanego okresu prezentuje rysunek 3.11. W swoim raporcie dotyczącym zmian w sieci szkół podstawowych i gimnazjów w latach 2007–2012 Jan Herczyński i Aneta Sobotka (2014) wskazują na fakt przejmowania przez stowarzyszenia małych szkół zamykanych przez gminy. Często wiąże się to ze zmianą statusu szkoły z publicznej na niepubliczną. Wzrost liczby uczniów ze szkół niepublicznych jest również powodowany przez otwieranie szkół przez podmioty komercyjne, co jest jednak zjawiskiem dużo rzadszym niż przekazywanie szkół gminnych. Pomimo zwiększania się różnicy w średnich wynikach sprawdzianu pomiędzy uczniami ze szkół niepublicznych i publicznych, trend w obydwu typach szkół jest podobny. Można zatem uznać, że sprawdzian nie zmienia się na korzyść lub niekorzyść uczniów z któregoś typu szkoły, a prawdopodobnie szkoły niepubliczne osiągają lepszą efektywność kształcenia. Oznaczałoby to, że szkoły podstawowe przejęte przez stowarzyszenia osiągają wyższe wyniki kształcenia niż w sytuacji, gdy były prowadzone przez gminy. Ta hipoteza wymaga jednak zweryfikowania na podstawie danych dotyczących konkretnych szkół.

Rysunek 3.11. Różnica porównywalnych wyników sprawdzianu oraz odsetek uczniów szkół niepublicznych w latach 2002–2013



Aspektem potencjalnie różnicującym wyniki uczniów jest również lokalizacja szkoły. Możemy ją rozpatrywać, biorąc pod uwagę rodzaj jednostki samorządu terytorialnego, na terenie której leży szkoła. W podziale administracyjnym Polski wyróżnia się trzy typy gmin: gminy wiejskie – nie zawierające na swym terytorium miasta, miejsko-wiejskie – w ich skład wchodzi zarówno miasto, jak i wsie oraz miejskie – zawierające na swym terytorium miasto. Niestety wadą tego podziału jest to, iż niezbyt dobrze odzwierciedla on zróżnicowanie gmin w zakresie składu społecznego mieszkańców. W szczególności gminami wiejskimi są gminy ościenne wielkich aglomeracji miejskich jak i gminy peryferyjne bez ośrodka miejskiego (Dolata i in., 2014). Lokalizacja szkoły może wpływać na wyniki osiągane przez uczniów poprzez następujące czynniki: poziom bezrobocia (związany także ze średnim poziomem zarobków rodziców oraz relacjami w rodzinach), struktura gospodarki (liczba podmiotów gospodarczych zatrudniających wykwalifikowanych pracowników) czy powszechność edukacji przedszkolnej (zob. Dolata i in., 2014).

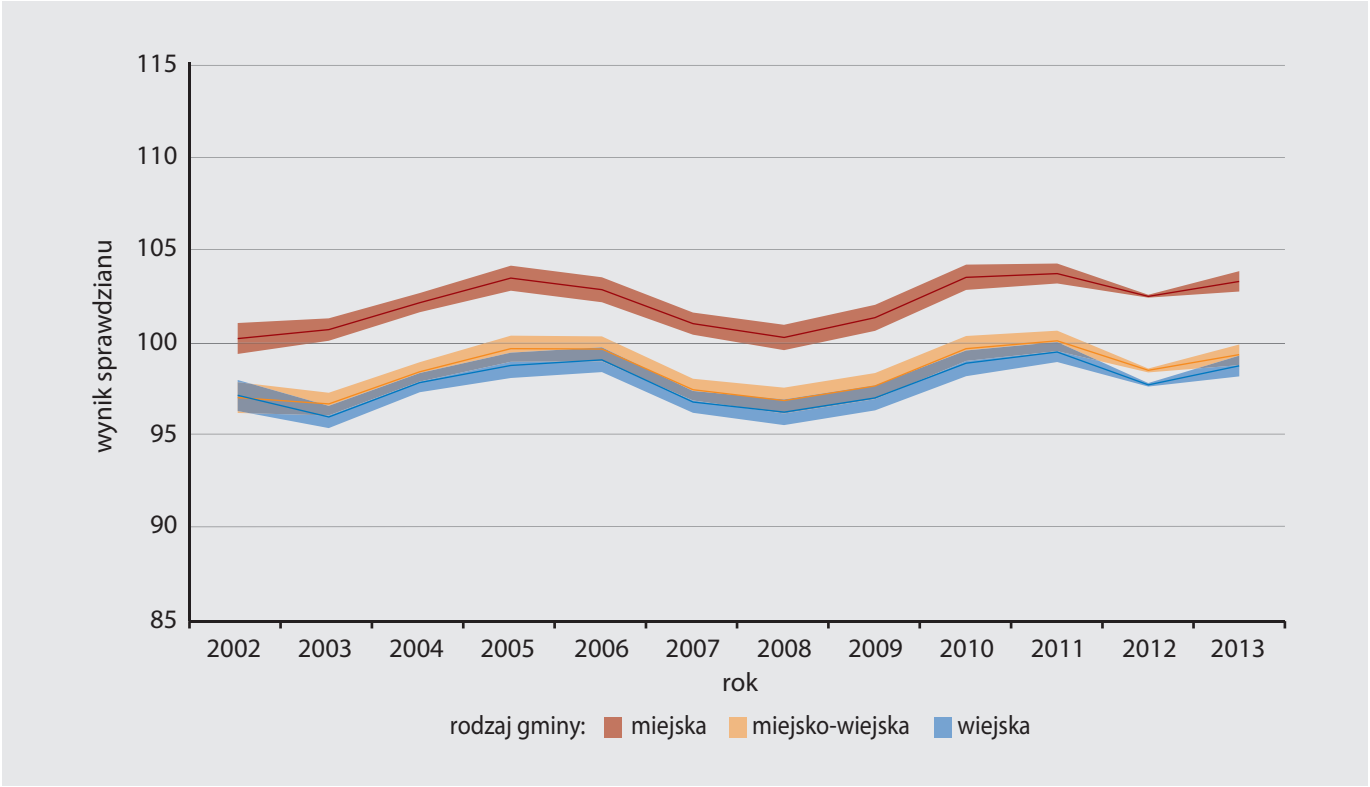
Herbst (2004), analizując wyniki sprawdzianu 2002, zwraca uwagę, że choć przeciętne wyniki testu osiągane w gminach miejskich są wyższe niż w gminach wiejskich (różnica pomiędzy średnimi wynosi 3,5%), to wskaźnik wartości dodanej szkoły jest wyższy w gminach wiejskich niż miejskich[31]. Dolata (2008) wskazuje, że choć różnica między średnimi wynikami uzyskiwanymi przez szkoły wiejskie i miejskie jest znacząca statystycznie, to wewnątrzgrupowe zróżnicowanie szkół wiejskich oraz miejskich może być bardzo różne, o czym należy pamiętać, analizując zbiorczo wyniki szkół miejskich i wiejskich.

Uczniowie ze szkół usytuowanych w gminach miejskich uzyskują wyższe wyniki sprawdzianu niż uczniowie ze szkół z gmin wiejskich i miejsko-wiejskich dla wszystkich analizowanych lat, przy czym trendy w wynikach są zbliżone. Podobnie jak w przypadku typu szkoły, dla części uczniów z 2002 roku nie było możliwe określenie rodzaju gminy, w której znajdowała się szkoła, do jakiej uczęszczali.

[31] W publikacji tej wartość dodaną zdefiniowano jako iloraz standaryzowanego przeciętnego wyniku testu matematyczno-przyrodniczego i standaryzowanego średniego wyniku humanistycznej części egzaminu, zakładając, że o jakości pracy szkół podstawowych w danym regionie można wnioskować na podstawie jakości pracy szkół gimnazjalnych (Herbst, 2004; s. 91–92).

Powoduje to, że wyniki z 2002 roku mogą być nieco inne niż w rzeczywistości[32]. Różnica między szkołami w gminach miejsko-wiejskich i wiejskich nie przekracza jednego punktu i nie jest istotna statystycznie. Zależności te ilustruje rysunek 3.12. Wyższe wyniki uczniów ze szkół miejskich mogą jednak nie być zasługą samych szkół, gdyż, jak zauważa Konarzewski (2012, s.7), „nie wynikają one z bardziej skutecznych metod kształcenia, lecz wyłącznie stąd, że w rejonie tych szkół mieszka więcej rodzin zamożnych, z wykształconymi i ustabilizowanymi zawodowo rodzicami. Rzekome upośledzenie szkół wiejskich to w istocie upośledzenie polskiej wsi – relatywnie biednej i źle wykształconej".

Rysunek 3.12. Porównywalne wyniki sprawdzianu w latach 2002–2013 w podziale na rodzaj gminy, w której znajduje się szkoła



Miary zróżnicowania wyników sprawdzianu świadczą o stopniu jednolitości systemu edukacji, co, jak zauważają Jasińska i Modzelewski (2013, s. 165), oznacza, że „w przypadku idealnie jednolitego systemu szkolnego szkoły osiągałyby te same średnie wyniki w testach, a wskaźnik zróżnicowania osiągałby wartość 0% (czyli, inaczej mówiąc, uczniowie każdej ze szkół osiągaliby średnio takie same wyniki). W przypadku odwrotnym całe zróżnicowanie wyników indywidualnych uczniów sprowadzałoby się do różnic między szkołami – wszyscy uczniowie w danej szkole osiągaliby te same wyniki, a wskaźnik przyjąłby wartość 100% (to, do jakiej szkoły uczęszczałby uczeń, całkowicie determinowałoby jego wynik)". Podczas interpretacji wyników należy jednak pamiętać, że zróżnicowanie wyników zależy zarówno od charakterystyk uczniów: ich uprzednich osiągnięć, zmiennych indywidualnych oraz społeczno-demograficznych, ale także efektywności pracy szkoły (Jasińska i Modzelewski, 2013; por. też rozdział 4 niniejszego raportu).

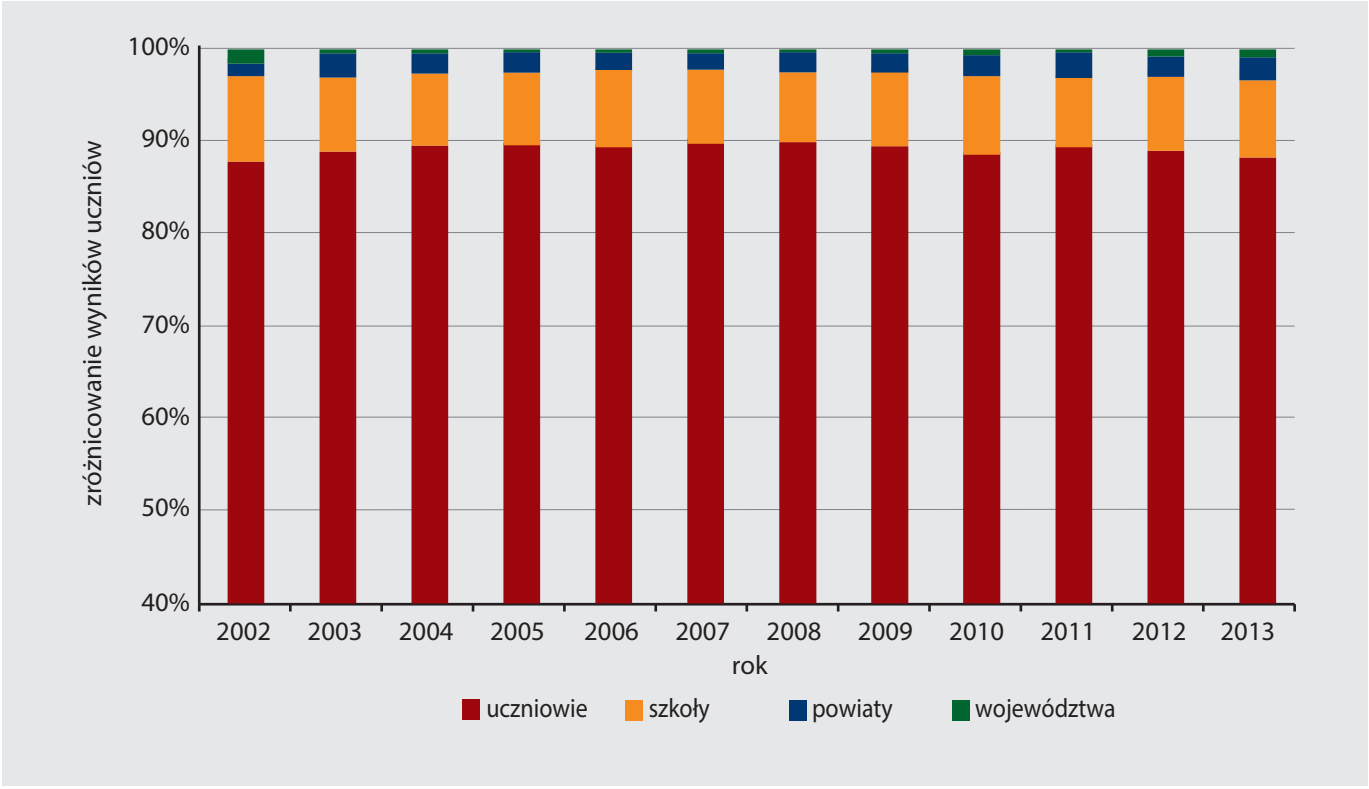
Rysunek 3.13 przedstawia całkowite zróżnicowanie wyników uczniów na sprawdzianie rozbite na zróżnicowanie związane z indywidualnym poziomem umiejętności uczniów oraz z przynależnością do konkretnej szkoły, powiatu i województwa dla poszczególnych lat. W przypadku sprawdzianu największe zróżnicowanie wyników obserwujemy między uczniami (około 90%), nieco mniejsze

[32] Była to ponadto pierwsza edycja sprawdzianu, więc jej przebieg mógł nieco odbiegać od ustalonych procedur.

między szkołami (około 7–8%) oraz zdecydowanie mniejsze na poziomie powiatów (około 2%) i województw (mniej niż 1%). Oznacza to, że zdecydowanie większe różnice w wynikach egzaminacyjnych obserwujemy między pojedynczymi uczniami niż między szkołami. Zróżnicowanie związane z zamieszkaniem w konkretnych jednostkach podziału administracyjnego jest wręcz pomijane. Jest to zjawiskiem pozytywnym – wynik uczniów w największym stopniu zależy od ich własnego poziomu umiejętności. Można więc przypuszczać, że na tym etapie nie dochodzi jeszcze do międzyszkolnej segregacji uczniów (zob. Dolata, 2008; 2012), co jest zgodne z założeniem o jednolitości i powszechności szkolnictwa.

Udział różnic międzyszkolnych i między jednostkami podziału administracyjnego w całkowitym zróżnicowaniu wyników sprawdzianu jest względnie stały. Jedynie dla roku 2002 (pierwszej edycji sprawdzianu szóstoklasisty) zróżnicowanie dla województw jest niewiele wyższe niż dla powiatów, do wyjaśnienia tego jednorazowego zjawiska potrzebne byłyby pogłębione analizy. Analiza zróżnicowania wyników sprawdzianu jest spójna z analizą zróżnicowania przeprowadzoną w badaniu PIRLS. W badaniu tym zróżnicowanie wynikające z indywidualnego poziomu umiejętności uczniów szacuje się na około 85%, natomiast zróżnicowanie wynikające z przynależności do konkretnych szkół na około 7–10% (Konarzewski, 2012). Podsumowując, wyniki analiz zróżnicowania porównywalnych wyników sprawdzianu są zgodne z wnioskami Dolaty (2012, s. 14) – „w skali kraju system szkół podstawowych jest dość bliski zakładanej jednolitości".

Rysunek 3.13. Całkowite zróżnicowanie wyników sprawdzianu w podziale na wpływ indywidualnych umiejętności ucznia oraz przynależności do szkoły, powiatu i województwa



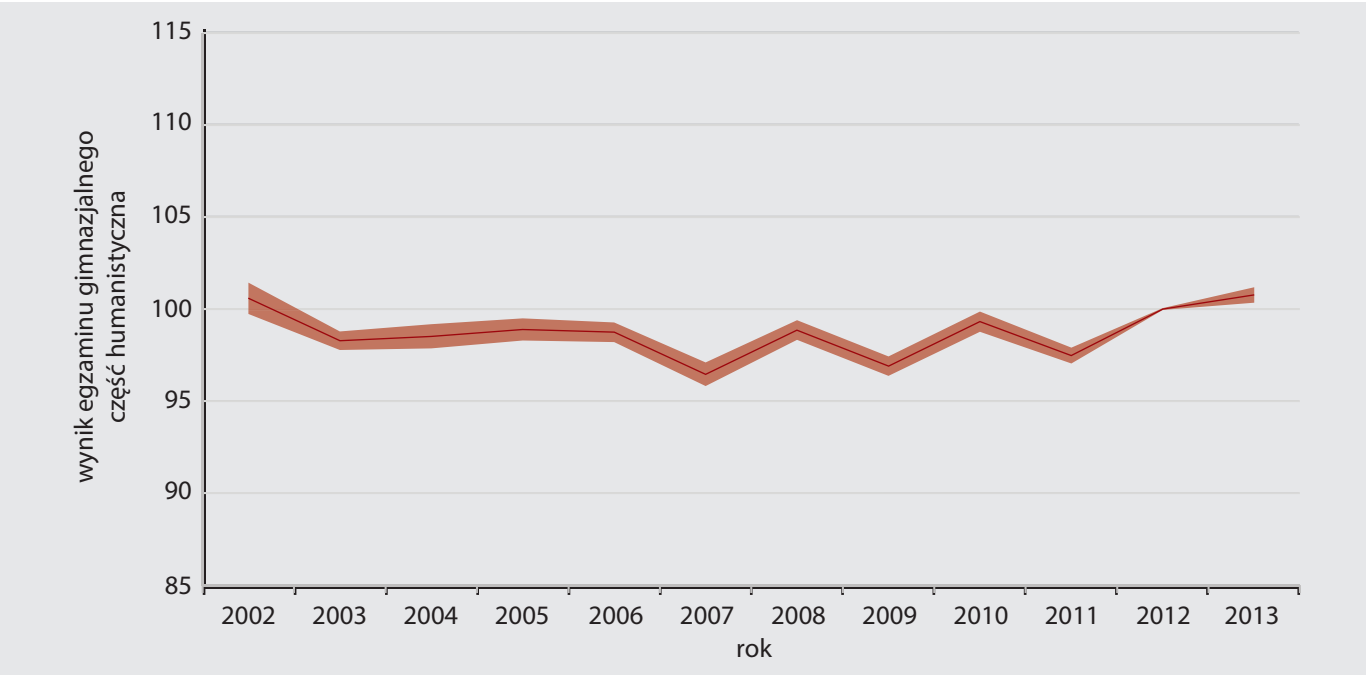
W porównaniu do egzaminu gimnazjalnego i egzaminu maturalnego zróżnicowanie wyników sprawdzianu związane z uczęszczaniem do konkretnej szkoły jest najmniejsze. Jest to zgodne z założeniem dotyczącym powszechności i jednolitości systemu edukacyjnego – wynik uczniów w największym stopniu zależy od nich samych, nie zaś od uczęszczania do „dobrej" szkoły. Zaobserwowano jednak niewielkie różnice na poziomie regionalnym – wyższe wyniki obserwujemy w centralnym, południowym i wschodnim regionie Polski, co może być związane z uwarunkowaniami

gospodarczymi i historycznymi poszczególnych regionów. Różnice międzypłciowe w osiąganych wynikach są niewielkie na przestrzeni czasu – wynoszą średnio około trzech punktów (na korzyść dziewczynek). Nie występuje więc opisywany w literaturze efekt niedoszacowania wyników egzaminacyjnych dla kobiet. Ze względu na charakter sprawdzianu podczas analizy sumarycznych wyników niemożliwe jest stwierdzenie, dla których umiejętności dana płeć osiąga lepsze wyniki. Można zaobserwować zwiększającą się różnicę w wynikach osiąganych przez uczniów uczęszczających do szkół publicznych i niepublicznych, jednak trend wyników jest podobny. Rozstrzygnięcie o przyczynie tej różnicy wymagałoby porównania efektywności kształcenia pomiędzy szkołami publicznymi i niepublicznymi. W odniesieniu do zjawiska dysleksji sprawdzian funkcjonuje stabilnie – można przypuszczać, że dostosowania wprowadzone dla uczniów z dysleksją pozwalają na wyrównanie szans powodzenia na sprawdzianie. W przypadku typu gminy można zaobserwować różnice na korzyść gmin miejskich, natomiast nie występują różnice pomiędzy gminami wiejskimi i miejsko-wiejskimi. Nie można natomiast wnioskować o lepszej efektywności pracy szkół w miastach, gdyż wyniki te mogą być związane np. z wyższym statusem społeczno-ekonomicznym rodzin w miastach.

3.3.2. Egzamin gimnazjalny

W przeciwieństwie do sprawdzianu egzamin gimnazjalny od momentu jego wprowadzenia podzielony był na część humanistyczną i matematyczno-przyrodniczą. Dzięki temu podziałowi mamy możliwość sprawdzenia zmian w czasie w zakresie poszczególnych rodzajów umiejętności (np. względem płci). W roku 2012 części te podzielono na osobne egzaminy z: języka polskiego, historii i wiedzy o społeczeństwie, matematyki oraz przedmiotów przyrodniczych.
Podobnie jak w przypadku sprawdzianu, zrównanie wyników obydwu części egzaminu gimnazjalnego[33] przeprowadzono do roku 2012. Wartość 100 odpowiada średniemu wynikowi egzaminu w 2012 roku, a różnica 15 punktów na skali odpowiada jednemu odchyleniu standardowemu wyników egzaminu w 2012 roku. Wyniki na wykresach prezentowane są wraz z 95% przedziałami ufności.

Rysunek 3.14. Porównywalne wyniki części humanistycznej egzaminu gimnazjalnego w latach 2002–2013

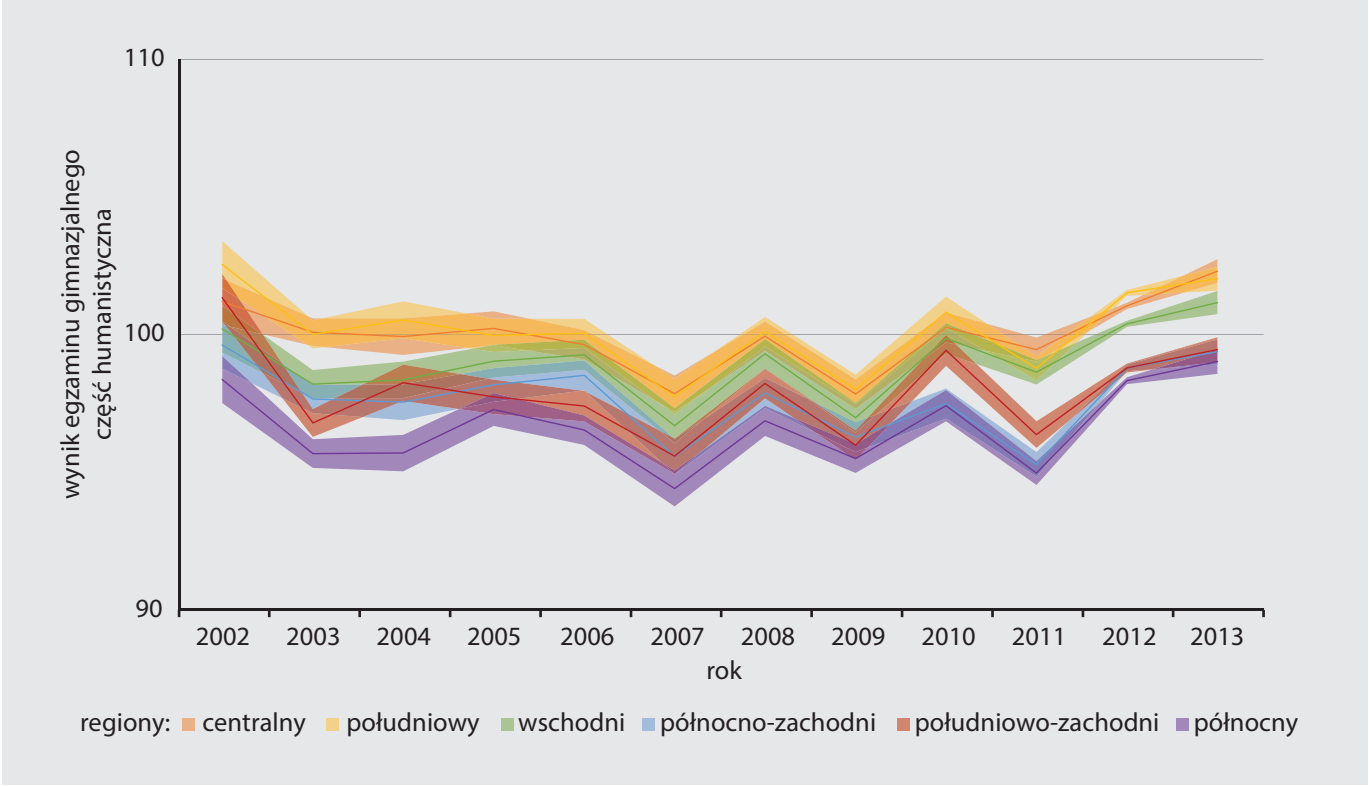


[33] Dla zachowania porównywalności wyników między latami od roku 2012 wyniki egzaminu z poszczególnych przedmiotów łączone są w części: humanistyczną (język polski oraz historia i wiedza o społeczeństwie) i matematyczno-przyrodniczą (matematyka oraz przedmioty przyrodnicze).

Część humanistyczna

Na poziomie kraju wyniki części humanistycznej egzaminu gimnazjalnego, które ilustruje rysunek 3.14, ulegają niewielkim fluktuacjom na przestrzeni czasu, analogicznie jak wyniki sprawdzianu. Początkowy spadek wyników między 2002 i 2003 rokiem prawdopodobnie jest spowodowany zawyżonymi wynikami egzaminu w 2002 roku. Był to pierwszy rok wprowadzenia egzaminu na tym etapie edukacyjnym, stąd do jego wyników należy podchodzić ostrożnie. Prawdopodobnie usprawnienie procedur przeprowadzania egzaminu spowodowało bardziej rzetelne szacowanie poziomu umiejętności uczniów w następnych latach. Obserwowane są także coroczne wahania wyników, różnice pomiędzy latami w okresie 2006–2012 wynosiły od dwóch do dwóch i pół punktu. Nie należy spodziewać się znaczących różnic w wynikach populacji uczniów pomiędzy sąsiednimi latami w sytuacji, gdy zakres treściowy egzaminu nie ulega znaczącym zmianom. Pomimo statystycznej istotności różnice wyników części humanistycznej egzaminu gimnazjalnego w latach 2006–2012 są na tyle małe, że nie powinno to budzić niepokoju.

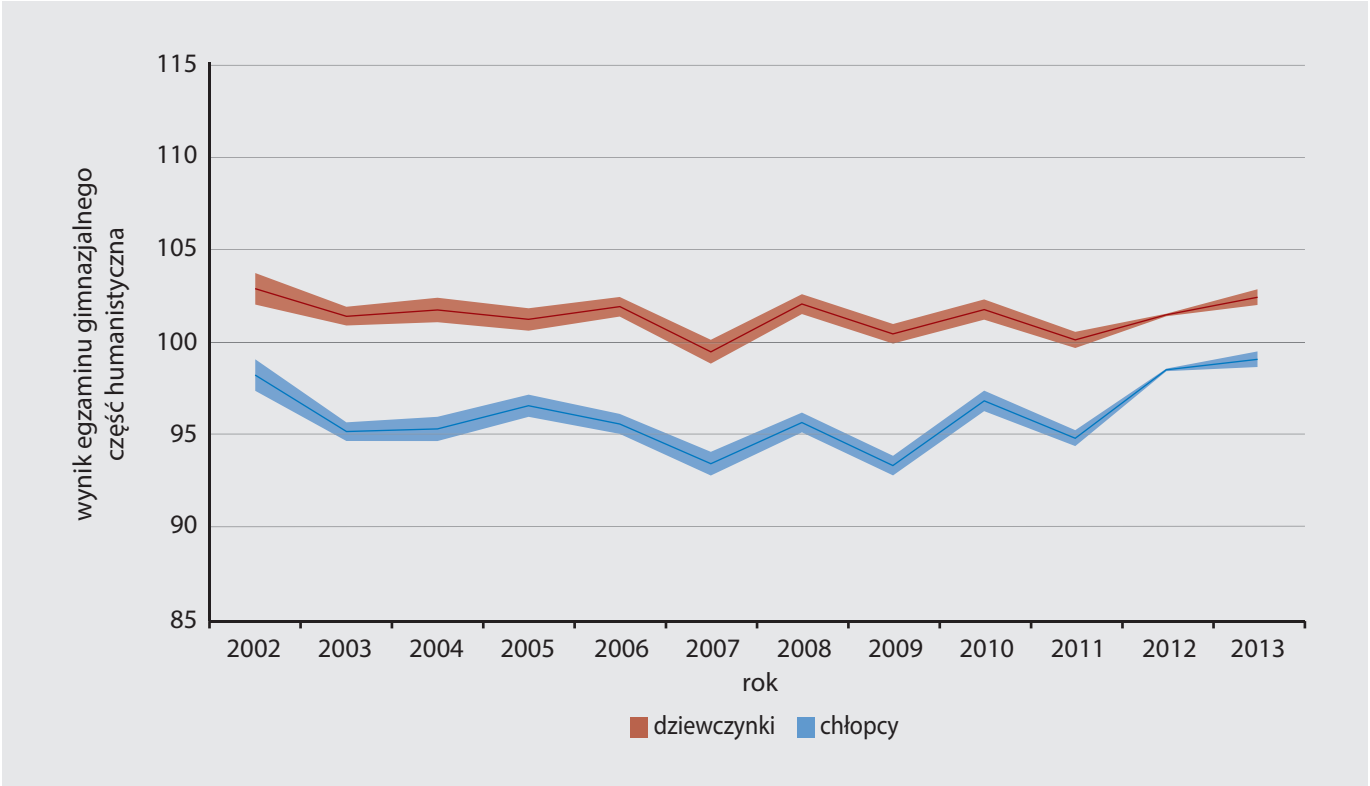
Rysunek 3.15. Porównywalne wyniki części humanistycznej egzaminu gimnazjalnego w latach 2002–2013 w podziale na regiony wg NTS



Na etapie sprawdzianu zaobserwowano zróżnicowanie wyników ze względu na analizowany region, dlatego też sprawdzono, czy tendencja ta utrzymuje się na dalszych etapach edukacyjnych. W przypadku analizy części humanistycznej egzaminu gimnazjalnego można dostrzec bardziej wyraźne w ostatnich latach zróżnicowanie pomiędzy centralną, południową i wschodnią częścią kraju a północną i zachodnią. Od roku 2011 występuje wyraźna polaryzacja średnich wyników pomiędzy tymi obszarami Polski (zob. rysunek 3.15). Wyższe wyniki egzaminacyjne zaobserwowano w regionach centralnej, południowej i wschodniej części kraju, natomiast niższe wyniki w północnej i zachodniej części Polski. Odzwierciedla to tendencję zauważalną już na etapie sprawdzianu, przy czym w odniesieniu do wyników egzaminu gimnazjalnego zachodnia część kraju osiąga systematycznie niższe wyniki. Podobnie jak w przypadku sprawdzianu, w regionie południowo-zachodnim uczniowie osiągali relatywnie wyższe wyniki w pierwszych edycjach egzaminu, niż miało to miejsce

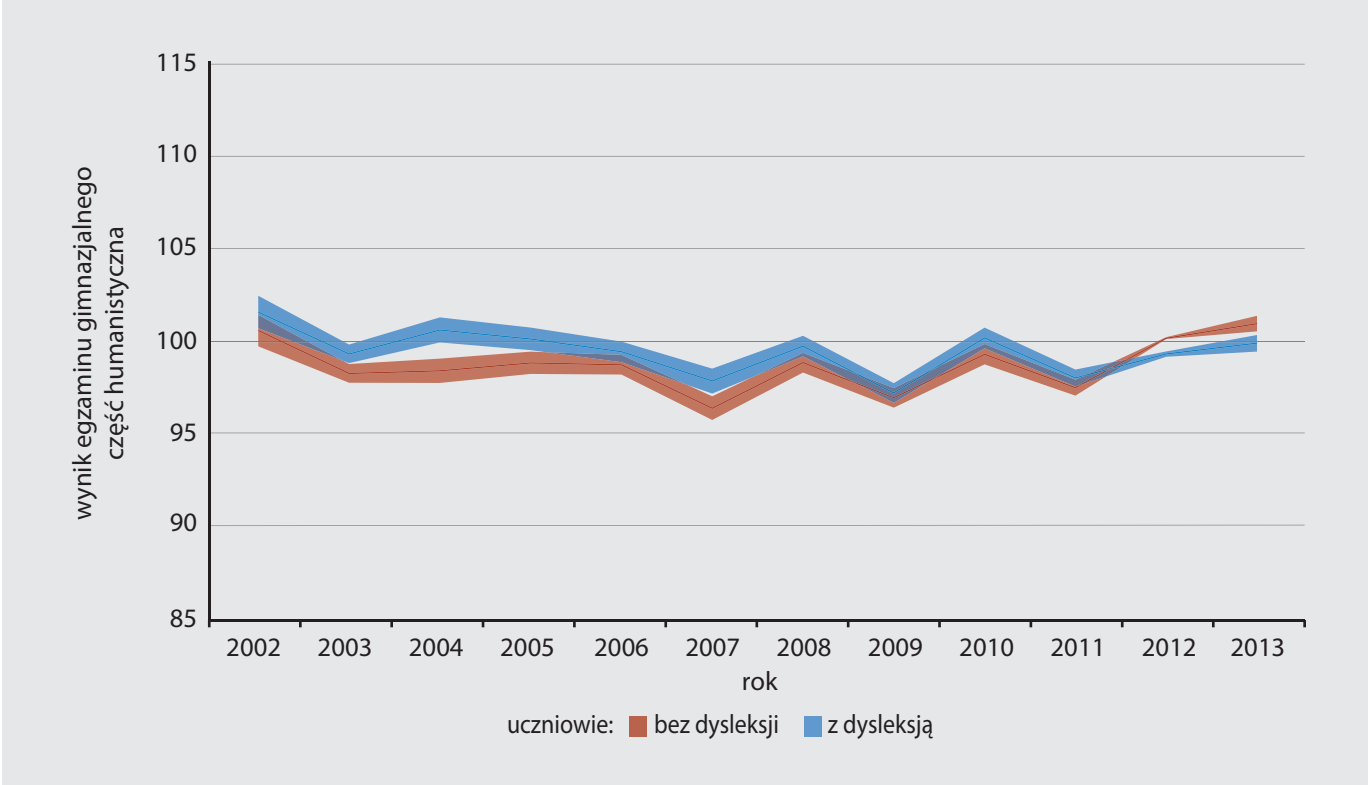
w ostatnich latach. Jest to jedyny region, w którym trend odbiega nieco od trendu na poziomie kraju. Pomimo tego, że obserwowane różnice są istotne statystycznie, są one jednak niewielkie i w latach 2011–2013 wynoszą nieco ponad dwa punkty. W okresie tych trzech lat są one też względnie stałe, lecz trudno prognozować, jak będą kształtowały się w przyszłości.

Rysunek 3.16. Porównywalne wyniki części humanistycznej egzaminu gimnazjalnego w latach 2002–2013 w podziale na płeć uczniów



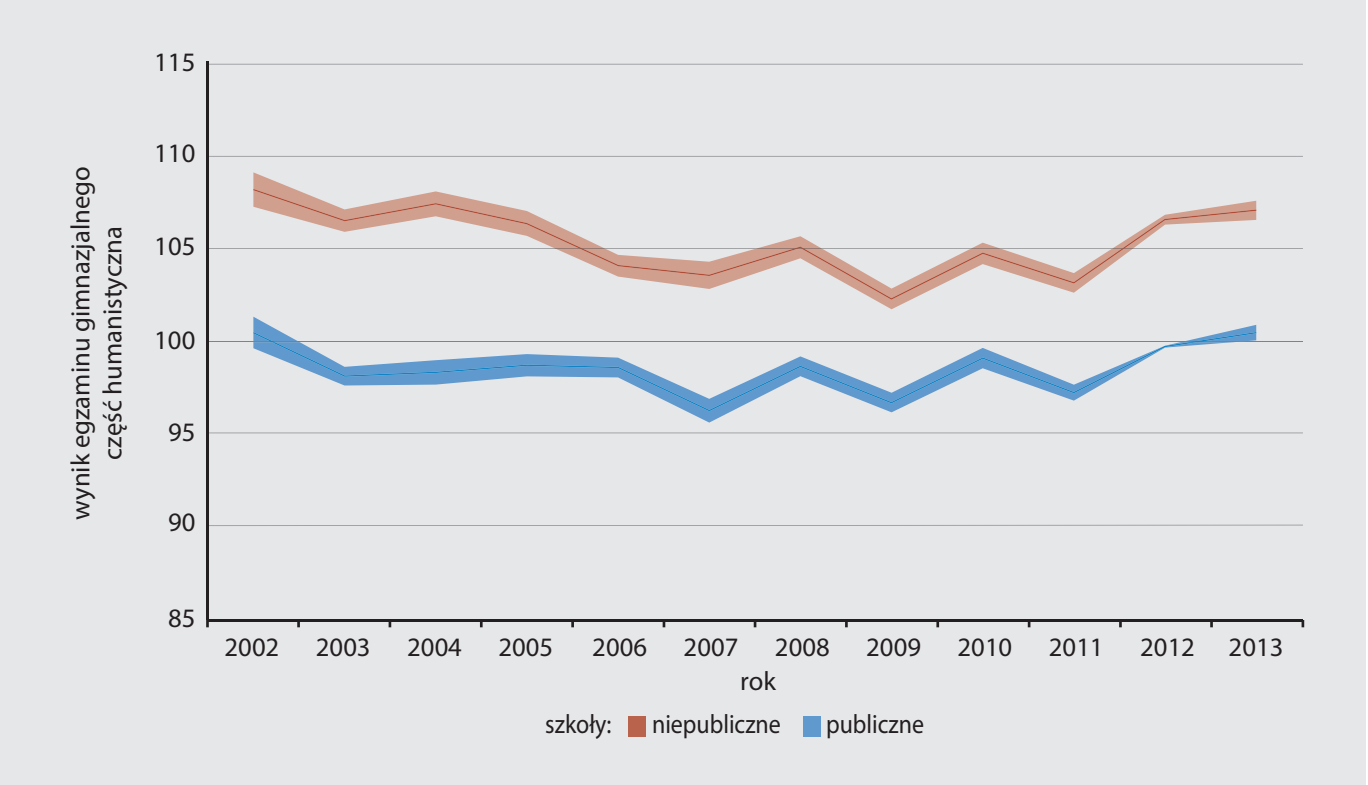
W związku z tym, że różnice w wynikach uczniów pomiędzy płciami mogą mieć charakter rozwojowy, tj. pogłębiać się w czasie (Willingham i Cole, 1997), należy sprawdzić, czy zjawiska zaobserwowane na etapie sprawdzianu utrzymują się w grupach starszych uczniów. Różnica w wynikach części humanistycznej egzaminu gimnazjalnego pomiędzy chłopcami i dziewczynkami utrzymuje się na zbliżonym poziomie na przestrzeni lat (zob. rysunek 3.16), podobnie jak w przypadku sprawdzianu, dziewczynki uzyskują więcej punktów niż chłopcy. Należy jednak zauważyć, że różnica w wynikach pomiędzy dziewczynkami i chłopcami zmniejsza się w ostatnich latach (2011–2013), co może być pocieszającym zjawiskiem. W latach 2002–2011 różnice wynosiły między pięć a siedem punktów, a w latach 2012 i 2013 spadły do około trzech punktów. W literaturze zwraca się uwagę na to, że dziewczynki uzyskują wyższe wyniki z przedmiotów humanistycznych, natomiast chłopcy z przedmiotów matematycznych i ścisłych (np. Willingham i Cole, 1997; Von Schrader i Ansley, 2006; Skórska i Świst, 2014) – w przypadku przedmiotów humanistycznych zależność ta znajduje odzwierciedlenie w analizowanych danych. Wskazuje się (OECD, 2009; Zasacka, 2014), że różnicę w wynikach z czytania wśród chłopców i dziewcząt należy powiązać z zainteresowaniem czytaniem. Indeks zainteresowania czytaniem w badaniach PISA (OECD, 2009) jest zdecydowanie niższy w przypadku chłopców we wszystkich analizowanych krajach (około pół odchylenia standardowego różnicy w stosunku do dziewczynek).

Rysunek 3.17. Porównywalne wyniki części humanistycznej egzaminu gimnazjalnego w latach 2002–2013 w podziale na grupy uczniów bez dysleksji rozwojowej i z dysleksją rozwojową



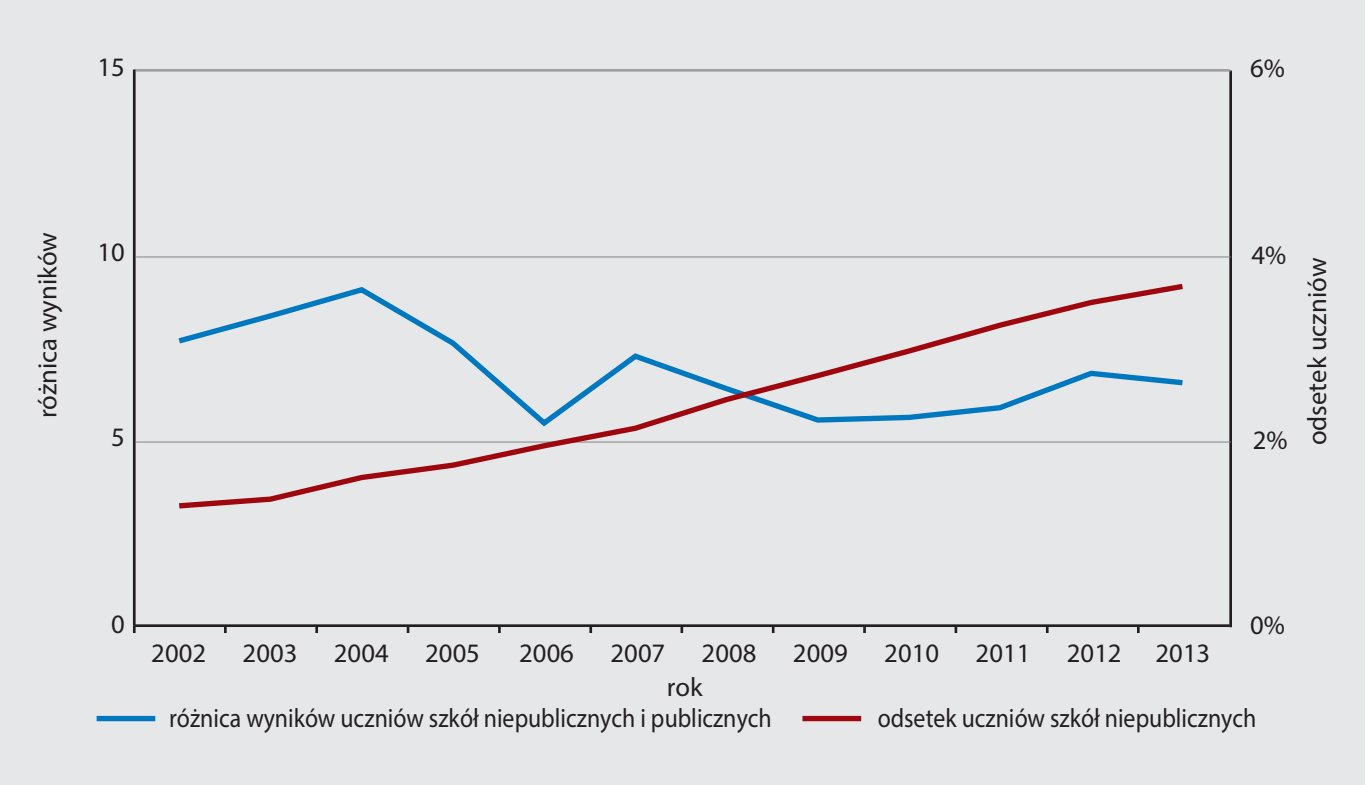
Podobnie jak w przypadku sprawdzianu, uczniowie z opinią o dysleksji mogą korzystać z szeregu dostosowań również podczas pisania egzaminu gimnazjalnego, np. zaznaczania odpowiedzi w zadaniach zamkniętych bez konieczności przenoszenia ich na kartę, wydłużenia czasu pisania, korzystania z pomocy nauczyciela wspomagającego czy zastosowania szczególnych kryteriów podczas oceny zadań otwartych (CKE, 2013). W początkowych latach istnienia egzaminu gimnazjalnego częściej niż w latach późniejszych można dostrzec różnice w wynikach uczniów ze zdiagnozowaną dysleksją i bez dysleksji przy czym ci pierwsi osiągali wyższe wyniki (zob. rysunek 3.17). Różnice te nie mają jednak systematycznego charakteru. Od 2012 roku to uczniowie bez zdiagnozowanej dysleksji rozwojowej osiągają średnio minimalnie wyższe wyniki niż uczniowie z dysleksją, lecz różnice te nie są istotne statystycznie. Utrzymanie się w następnych latach braku różnic powinno świadczyć o stabilności konstrukcji egzaminu i kryteriów oceniania w stosunku do zjawiska dysleksji rozwojowej.

Rysunek 3.18. Porównywalne wyniki części humanistycznej egzaminu gimnazjalnego w latach 2002–2013 w podziale na typ szkoły (niepubliczna vs publiczna)



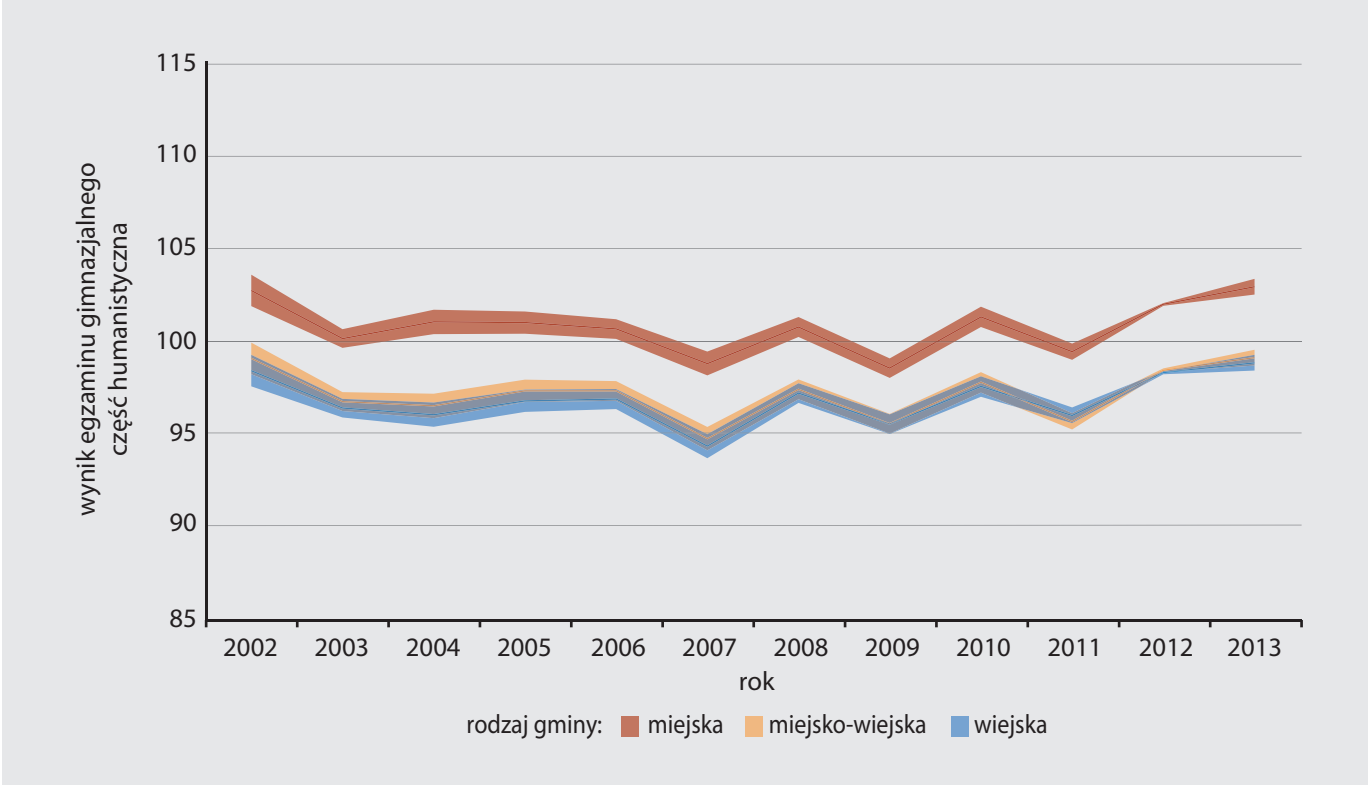
Zaobserwowaną różnicę w wynikach sprawdzianu pomiędzy szkołami publicznymi i niepublicznymi warto zbadać również na następnym etapie edukacyjnym – w gimnazjum. Analizując wyniki w podziale na szkoły niepubliczne i publiczne, należy mieć na uwadze, że liczba szkół niepublicznych jest o wiele mniejsza niż liczba szkół publicznych. Raporty CKE (CKE, 2010; 2011; 2012; 2013) wskazują, że w skali całego kraju uczniowie gimnazjów niepublicznych uzyskali w latach 2010–2013 wyższe wyniki z egzaminu gimnazjalnego niż uczniowie z gimnazjów publicznych. Podobnie jak w przypadku sprawdzianu, tendencja ta jest widoczna w przypadku analiz porównywalnych wyników egzaminacyjnych na przestrzeni wszystkich analizowanych lat. Jeśli pominiemy dwie pierwsze edycje egzaminu (wcześniej wskazano, że do ich wyników należy podchodzić z ostrożnością), to od 2004 roku można obserwować niewielkie zmniejszenie się różnic średnich wyników pomiędzy gimnazjami niepublicznymi i publicznymi (zob. rysunek 3.18). Zmiana ta nie ma jednak systematycznego charakteru, różnice w rozpatrywanym okresie stawały się raz większe, raz mniejsze. Jednocześnie, podobnie jak w przypadku szkół podstawowych, cały czas wzrasta odsetek uczniów z gimnazjów niepublicznych (zob. rysunek 3.19). Herczyński i Sobotka (2014) wskazują, że w przypadku gimnazjów, inaczej niż w przypadku szkół podstawowych, przekazywanie przez gminy w latach 2007–2012 szkół innym organom prowadzącym było zjawiskiem rzadkim. Powstało w tym okresie jednak sporo nowych szkół, których organem prowadzącym był inny podmiot niż gmina. Zmiany zachodzące w sieci szkół niewątpliwie mają wpływ na obserwowane wyniki. Trudno jednak wyjaśniać obserwowane różnice i ich okresowe wahania bez dodatkowych analiz (np. EWD w podziale na typy szkół).

Rysunek 3.19. Różnica porównywalnych wyników części humanistycznej egzaminu gimnazjalnego oraz odsetek uczniów szkół niepublicznych w latach 2002–2013



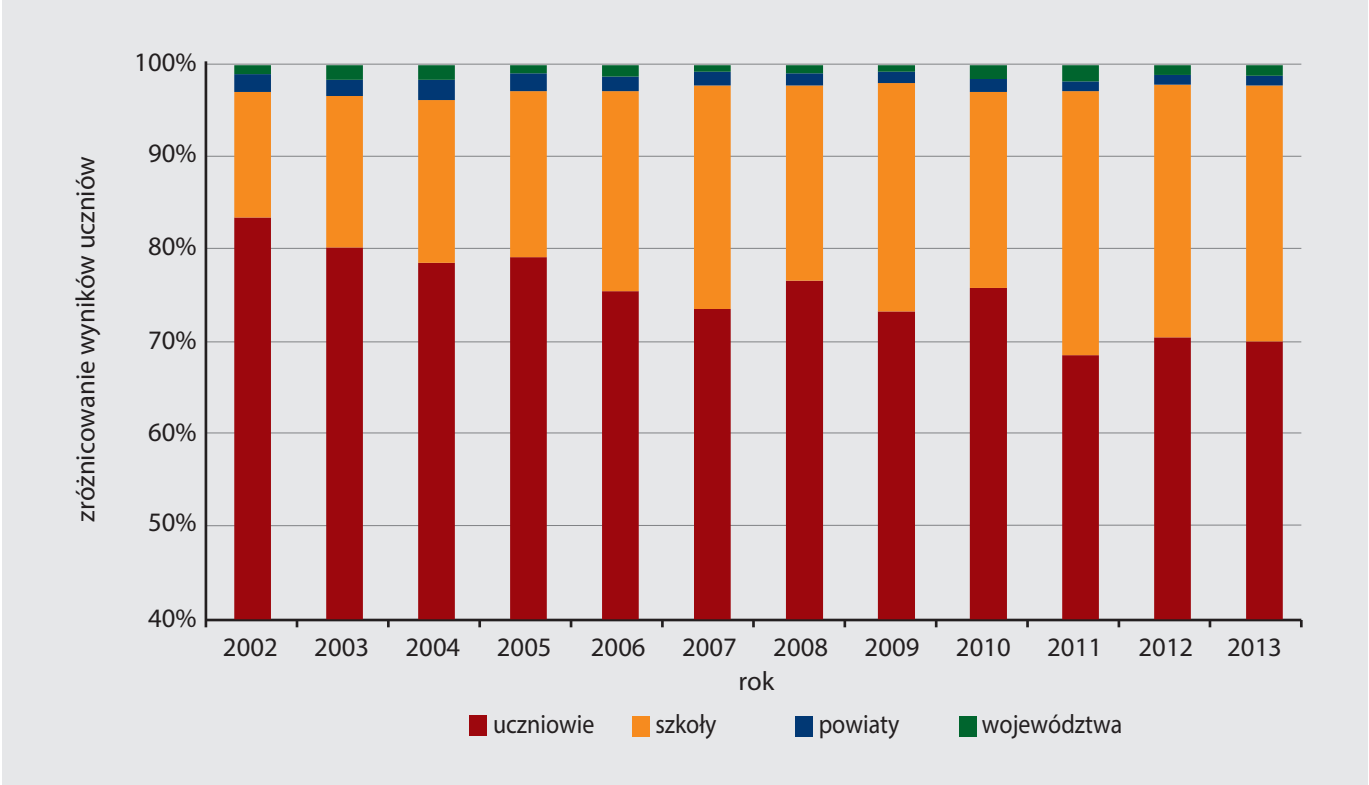
Obserwowane na etapie sprawdzianu zróżnicowanie wyników uczniów w zależności od lokalizacji szkół utrzymuje się w przypadku części humanistycznej egzaminu gimnazjalnego. Podobnie jak w przypadku sprawdzianu, różnice pomiędzy uczniami z gmin miejsko-wiejskich i wiejskich są pomijalne. W przypadku uczniów z gmin miejskich widać podobne zależności jak w przypadku sprawdzianu – uzyskują oni wyniki wyższe średnio o trzy do pięciu punktów niż uczniowie z pozostałych typów gmin. Zmiany średnich wyników pomiędzy latami przebiegają podobnie, niezależnie od typu gminy, w której mieściła się szkoła.

Rysunek 3.20. Porównywalne wyniki części humanistycznej egzaminu gimnazjalnego w latach 2002–2013 w podziale na rodzaj gminy, w której znajduje się szkoła



Przejdźmy do problemu międzyszkolnego zróżnicowania wyników nauczania – jego potencjalny wpływ na wyniki uczniów został opisany w poprzednich częściach rozdziału. Największe zróżnicowanie obserwujemy między uczniami, mniejsze między szkołami, a bardzo małe na poziomie powiatów i województw (zob. rysunek 3.21). Zróżnicowanie między szkołami w przypadku części humanistycznej egzaminu gimnazjalnego jest nieco większe niż w przypadku sprawdzianu. Może to oznaczać, że na etapie gimnazjum rozpoczyna się wzrost poziomu segregacji uczniów – lepsi uczniowie idą do „lepszych" szkół, słabsi idą do „gorszych" szkół (por. Dolata, 2008, 2010). Dziwić może fakt, że na poziomie gimnazjów obserwujemy większe zróżnicowanie międzyszkolne niż w przypadku szkół podstawowych, mimo iż o przyjęciu do obydwu typów szkół decyduje zamieszkanie w wyznaczonym obwodzie szkolnym. Jak jednak zauważa R. Dolata (2012, s. 10): „nawet w sytuacji rygorystycznej rejonizacji najaktywniejsze jednostki potrafią obejść biurokratyczne zasady i np. poprzez fikcyjny meldunek lub faktyczną zmianę miejsca zamieszkania aktywnie wybrać szkołę". W polskich realiach trudno mówić o rygorystycznym trzymaniu się zasady rejonizacji, szczególnie w dużych miastach. Coraz częściej wielkomiejskie gimnazja publiczne znajdują sposób na selekcję uczniów (np. szkoły dwujęzyczne) lub w całkowicie już jawny sposób funkcjonują jak szkoły selekcyjne. Zjawisko to wiąże się z funkcjonowaniem quasi-rynków edukacyjnych, które prowadzą do autoselekcji i selekcji, a tym samym wpływają na procesy różnicowania się szkół (Dolata, 2010). Wśród polskich badań na ten temat wymienić można pracę Herbsta i Herczyńskiego (2005), którzy pokazali, że poziom koncentracji lokalnego rynku oświatowego jest negatywnie skorelowany ze średnimi wynikami egzaminu gimnazjalnego w gminach miejskich. Zjawisko konkurowania szkół o uczniów można tłumaczyć strachem dyrektorów o zamykanie szkół, potęgowanym przy tym niżem demograficznym lub konkurowaniem szkół o dotacje (Herbst i Herczyński, 2005).

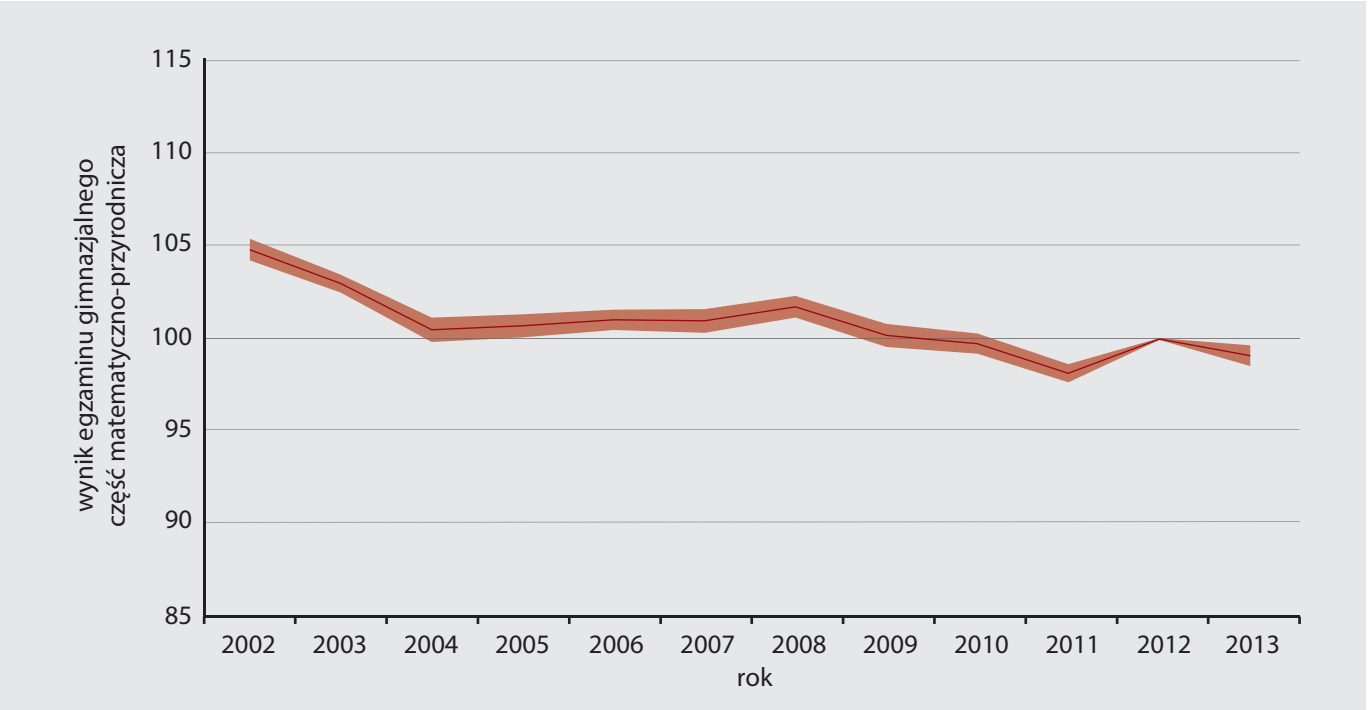
Rysunek 3.21. Całkowite zróżnicowanie wyników egzaminu gimnazjalnego w części humanistycznej w podziale na wpływ indywidualnych umiejętności ucznia oraz przynależności do szkoły, powiatu i województwa



Część matematyczno-przyrodnicza

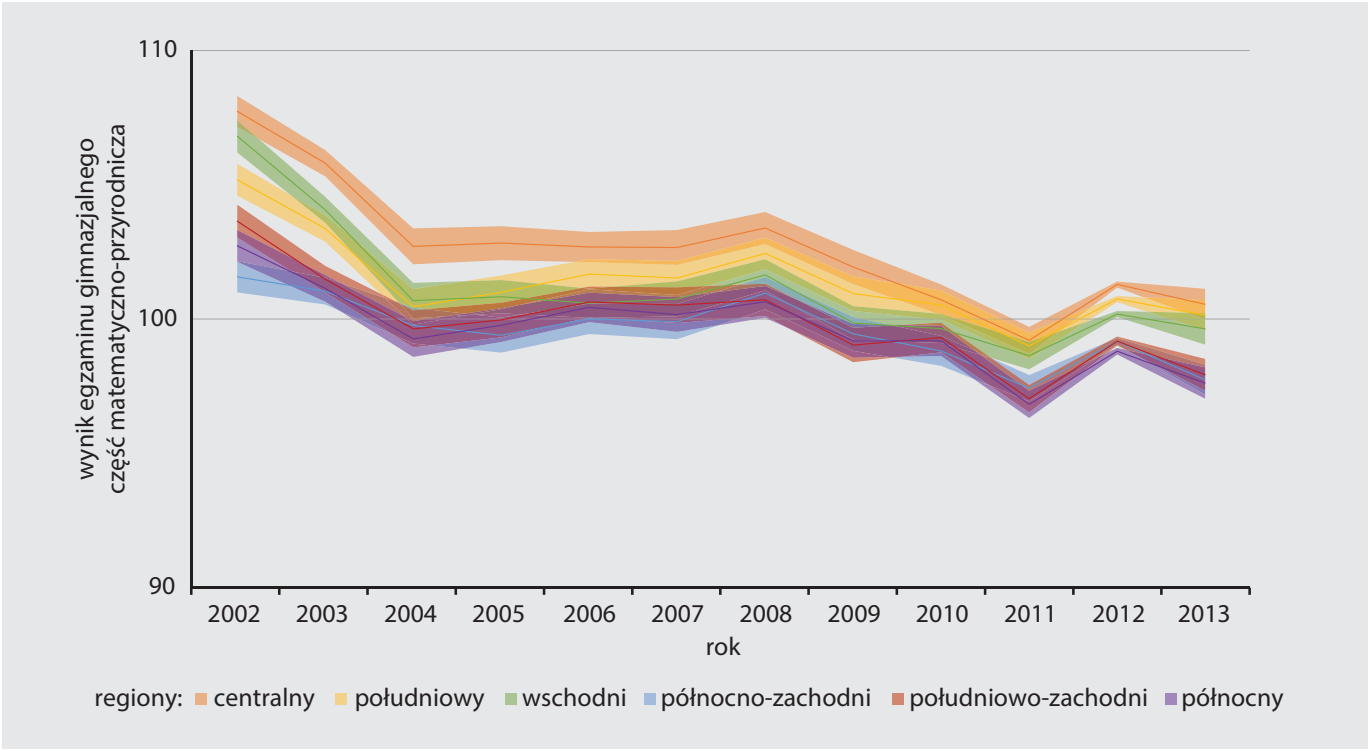
W pierwszej kolejności przeanalizujmy ogólny trend porównywalnych wyników dla części matematyczno-przyrodniczej egzaminu gimnazjalnego. Na poziomie kraju wyniki części matematyczno-przyrodniczej egzaminu gimnazjalnego są bardziej stabilne w czasie niż w przypadku części humanistycznej. Oprócz znaczącego spadku w latach 2002–2004 mamy do czynienia z kilkuletnimi okresami stopniowego spadku lub wzrostu wyników (zob. rysunek 3.22). Silne zmiany w pierwszych latach egzaminu są zapewne związane ze stopniowym wdrażaniem procedur nowego egzaminu i nie powinny być alarmujące.

Rysunek 3.22. Porównywalne wyniki części matematyczno-przyrodniczej egzaminu gimnazjalnego w latach 2002–2013



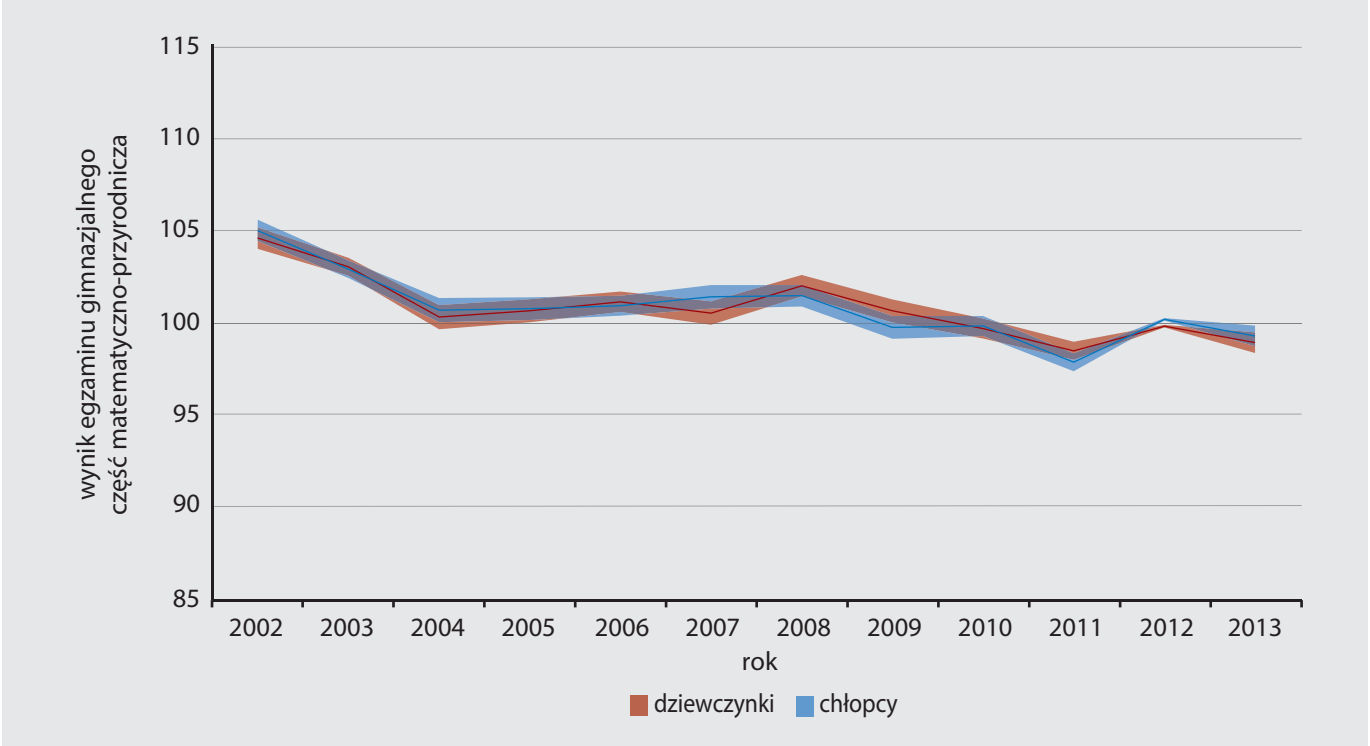
Zaobserwowane wcześniej zróżnicowanie wyników pomiędzy regionami kraju jest obecne także w przypadku części matematyczno-przyrodniczej egzaminu gimnazjalnego. Analiza porównywalnych wyników po raz kolejny uwidacznia polaryzację kraju na część centralną i południowo-wschodnią oraz północno-zachodnią (zob. rysunek 3.23). Przy czym, odmiennie niż w części humanistycznej egzaminu, w części matematyczno-przyrodniczej podział ten uwidaczniał się już w pierwszych edycjach egzaminu gimnazjalnego w latach 2002–2003, a nie tylko w latach 2011–2013. Wyższe wyniki egzaminacyjne w tych okresach zaobserwowano w regionach: centralnym, południowym i wschodnim, natomiast niższe w regionach: północnym, północno-zachodnim i południowo-zachodnim. Wyraźnie widać też przewagę uczniów z regionu centralnego w osiąganych wynikach nad pozostałymi uczniami – ich średni wynik jest w każdym roku najwyższy. Pomimo zmniejszenia się różnic pomiędzy poszczególnymi regionami w latach 2004–2010, obserwowany wcześniej terytorialny podział kraju w odniesieniu do średnich wyników utrzymuje się. Odmiennie niż wcześniej kształtują się też wyniki regionu południowo-zachodniego – od wprowadzenia egzaminu gimnazjalnego uczniowie z jego obszaru osiągają wyniki podobne do wyników uczniów z regionu północno-zachodniego i północnego.

Rysunek 3.23. Porównywalne wyniki części matematyczno-przyrodniczej egzaminu gimnazjalnego w latach 2002–2013 w podziale na regiony wg NTS



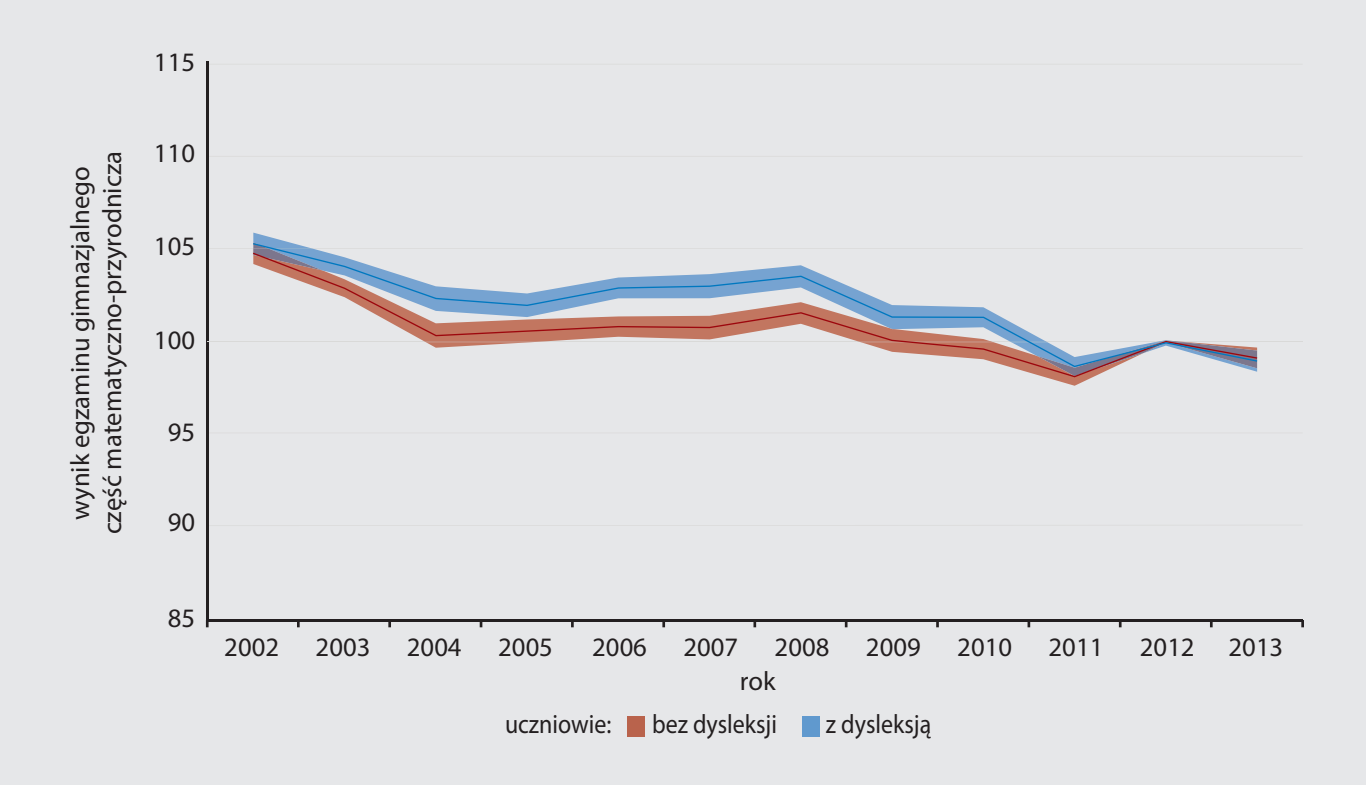
Choć w literaturze wskazuje się, że chłopcy mogą uzyskiwać wyższe wyniki z matematyki i nauk ścisłych (Willingham i Cole, 1997), nie znaleziono potwierdzenia tej tezy w przypadku analizowanych danych dotyczących części matematyczno-przyrodniczej egzaminu gimnazjalnego. Różnice w wynikach dziewczynek i chłopców w poszczególnych latach, co przedstawia rysunek 3.24, są nieistotne statystycznie. Jest to sytuacja odmienna niż w przypadku części humanistycznej, gdzie rokrocznie dziewczynki osiągały wyższe wyniki niż chłopcy (podobnie jak podczas sprawdzianu). Nie znajdujemy zatem dla części matematyczno-przyrodniczej egzaminu gimnazjalnego potwierdzenia wspomnianego wcześniej przy analizie wyników sprawdzianu zjawiska niedoszacowania predykcji wyników egzaminacyjnych dla kobiet (zob. Hyde i Kling, 2001). Również w badaniu PISA różnice między chłopcami a dziewczynkami w dziedzinie umiejętności matematycznych okazały się nieistotne statystycznie (OECD, 2014).

Rysunek 3.24. Porównywalne wyniki części matematyczno-przyrodniczej egzaminu gimnazjalnego w latach 2002–2013 w podziale na płeć uczniów



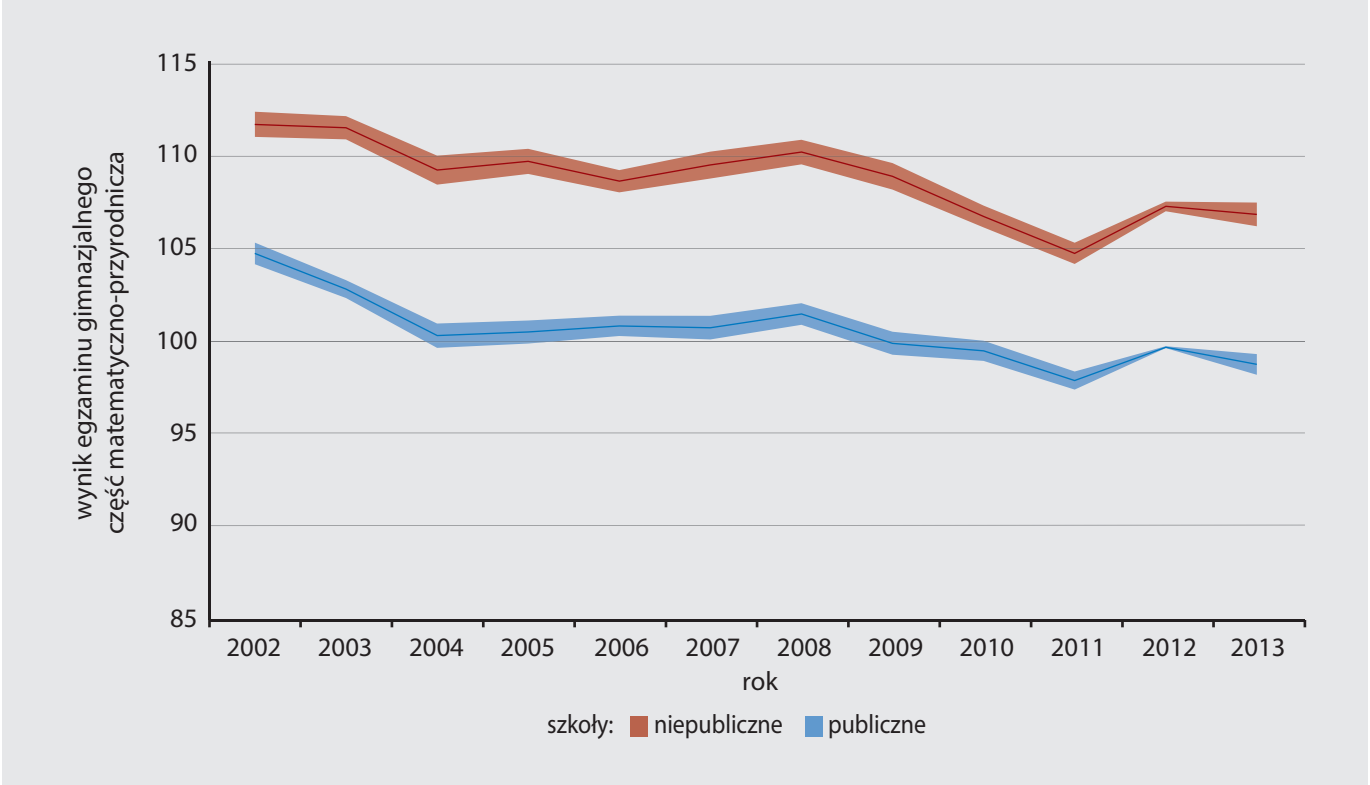
Specyfika egzaminu gimnazjalnego w części matematyczno-przyrodniczej zdecydowanie różni się od części humanistycznej – przede wszystkim nie wymaga formułowania dłuższych wypowiedzi, sprawiających szczególne problemy uczniom z diagnozą dysleksji. Pomijając dwie pierwsze edycje egzaminu gimnazjalnego w latach 2002 i 2003, uczniowie z diagnozą dysleksji rozwojowej osiągali do roku 2010 wyższe wyniki z części matematyczno-przyrodniczej tego egzaminu niż uczniowie bez takiej diagnozy. Od roku 2011 nie ma statystycznie istotnych różnic w wynikach obydwu tych grup. Biorąc pod uwagę to, że dla części humanistycznej zaobserwowano mniejsze różnice, można wnioskować, że dostosowania dla uczniów dyslektycznych mogły stwarzać tej grupie przewagę w części matematyczno-przyrodniczej (prawdopodobnie głównie ze względu na wydłużony czas na udzielanie odpowiedzi). Zadania otwarte z matematyki i przedmiotów przyrodniczych zapewne sprawiają uczniom z diagnozą dysleksji mniej trudności niż zadania z części humanistycznej, stąd zatem mogą wynikać różnice pomiędzy częściami egzaminu. Jeśli uznać, że celem wprowadzenia udogodnień dla uczniów z dysleksją rozwojową jest wyrównanie ich szans w stosunku do uczniów bez dysleksji, to ich efektem powinien być raczej brak różnic w średnich wynikach tych grup, który możemy obserwować od roku 2011. Omawiane wyniki przedstawia rysunek 3.25.

Rysunek 3.25. Porównywalne wyniki części matematyczno-przyrodniczej egzaminu gimnazjalnego w latach 2002–2013 w podziale na grupy uczniów bez dysleksji rozwojowej i z dysleksją rozwojową

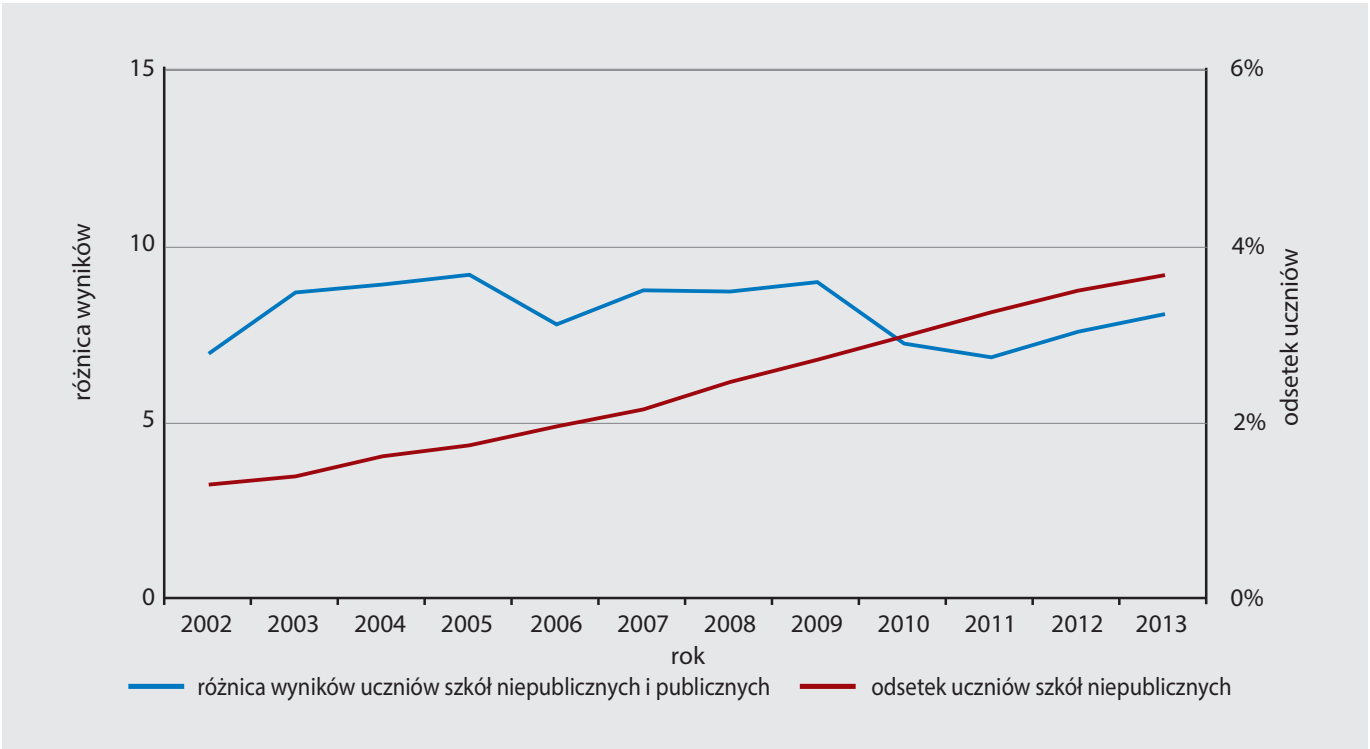


Analiza porównywalnych wyników części matematyczno-przyrodniczej egzaminu gimnazjalnego w podziale na szkoły publiczne i niepubliczne (zob. rysunek 3.26) prowadzi do podobnych wniosków, jak dla części humanistycznej. W każdym z analizowanych roczników, uczniowie gimnazjów niepublicznych osiągali średnio wyższe wyniki niż uczniowie gimnazjów publicznych. Różnica ta waha się w przedziale od siedmiu do dziewięciu punktów na korzyść szkół niepublicznych (zob. rysunek 3.27). Brak systematycznego spadku lub wzrostu tej różnicy świadczy o stałej przewadze w osiąganych wynikach uczniów z gimnazjów niepublicznych nad uczniami ze szkół publicznych.

Rysunek 3.26. Porównywalne wyniki części matematyczno-przyrodniczej egzaminu gimnazjalnego w latach 2002–2013 w podziale na typ szkoły (niepubliczna vs publiczna)



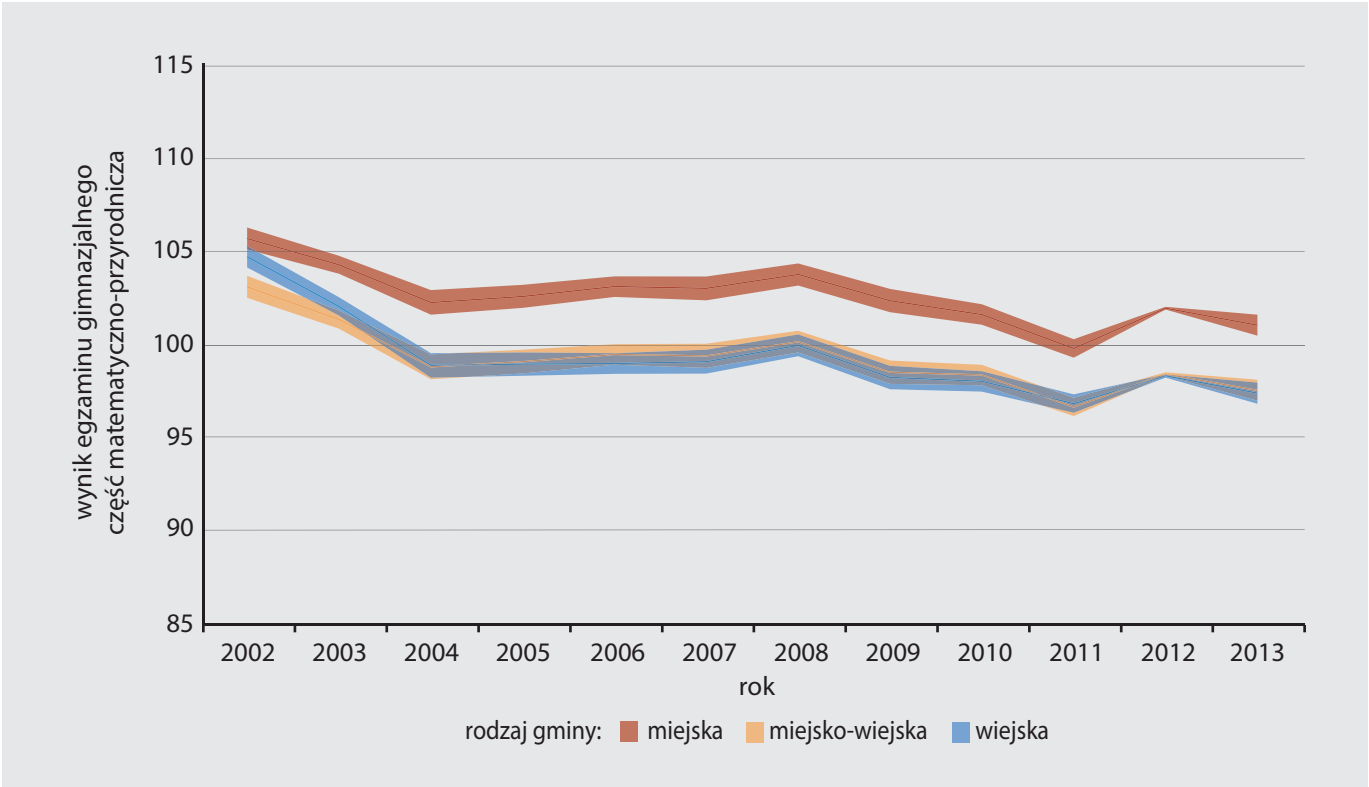
Rysunek 3.27. Różnica porównywalnych wyników części matematyczno-przyrodniczej egzaminu gimnazjalnego oraz odsetek uczniów szkół niepublicznych w latach 2002–2013



Lokalizacja szkół, podobnie jak wskazywano w odniesieniu do sprawdzianu i części humanistycznej egzaminu gimnazjalnego, wiąże się ze zróżnicowaniem wyników uczniów. Różnice w średnich porównywalnych wynikach części matematyczno-przyrodniczej egzaminu gimnazjalnego w podziale

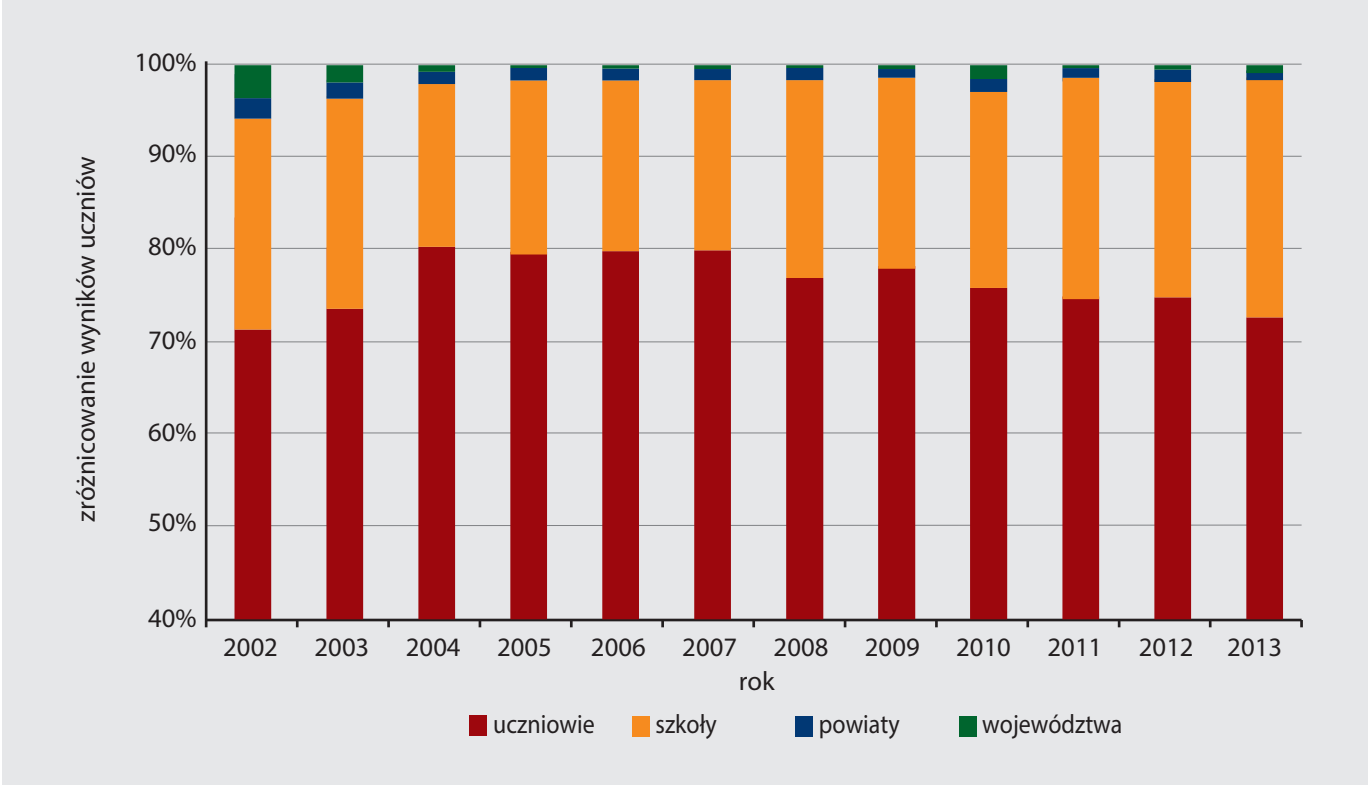
na typy gmin, które przedstawia rysunek 3.28, są zbliżone do różnic obserwowanych w części humanistycznej. Od roku 2004 uczniowie, którzy uczęszczali do szkół w gminach miejskich, uzyskują średnio około dwa do czterech punktów więcej niż uczniowie ze szkół w gminach wiejskich i miejsko-wiejskich, pomiędzy którymi nie ma istotnie statystycznych różnic. Pewną ciekawostką są średnie wyniki z gmin wiejskich, które w roku 2002 były bliższe średnim wynikom gmin miejskich, lecz do roku 2004 nastąpił ich spadek. Trudno jednak znaleźć wyjaśnienie tego fenomenu bez przeprowadzenia pogłębionych badań. Być może to właśnie w wynikach szkół z gmin wiejskich należy upatrywać głównej przyczyny spadku wyników w latach 2003 i 2004, wyraźnie widocznego w wynikach na poziomie kraju.

Rysunek 3.28. Porównywalne wyniki części matematyczno-przyrodniczej egzaminu gimnazjalnego w latach 2002–2013 w podziale na rodzaj gminy, w której znajduje się szkoła



Podobnie jak w przypadku części humanistycznej egzaminu gimnazjalnego, warto przyjrzeć się zróżnicowaniu wyników uczniów z części matematyczno-przyrodniczej, związanemu z ich poziomem umiejętności, uczęszczaniem do danej szkoły, zamieszkaniem w danym powiecie i województwie. Największe zróżnicowanie wyników obserwujemy między uczniami, mniejsze między szkołami oraz bardzo małe zróżnicowanie na poziomie powiatów i województw (zob. rysunek 3.29), jak miało to miejsce dla części humanistycznej. Dane te prowadzą do takiego samego wniosku, jak wcześniej, iż możemy zauważyć coraz większe zróżnicowanie wyników pomiędzy szkołami, co może świadczyć o większej selekcyjności gimnazjów niż szkół podstawowych (Dolata, 2008).

Rysunek 3.29. Całkowite zróżnicowanie wyników egzaminu gimnazjalnego w części matematyczno-przyrodniczej w podziale na wpływ indywidualnych umiejętności ucznia oraz przynależności do szkoły, powiatu i województwa



Podsumowując analizę porównywalnych wyników egzaminu gimnazjalnego, można zauważyć przede wszystkim pojawienie się znacznego zróżnicowania międzyszkolnego, rosnącego w części humanistycznej, bardziej stabilnego w części matematyczno-przyrodniczej. Oznacza to istnienie selekcji w gimnazjach – uczniowie z wyższym poziomem umiejętności są przyjmowani do innych szkół niż uczniowie z niższym poziomem umiejętności. Jest to spójne z wnioskami z badań prowadzonych przez Dolatę (2008; 2010; 2012). Wielkość współczynnika zróżnicowania międzyszkolnego określającą próg selekcji dla szkół (dla roku 2013 wynosi ona około 26% dla części matematyczno-przyrodniczej oraz 28% dla części humanistycznej) można porównać z wynikami PISA. Wskaźnik ten w krajach o najmniejszym zróżnicowaniu międzyszkolnym (takich jak Finlandia, Norwegia czy Dania) wynosi 10% (Dolata, 2012). W związku z tym, że etap szkolnictwa gimnazjalnego powinien być powszechny i jednolity, wyniki te mogą być niepokojące. Podobnie jak w przypadku sprawdzianu, można zaobserwować rosnące zróżnicowanie wyników ze względu na region – na korzyść regionu centralnego, południowego i wschodniego, co można powiązać z uwarunkowaniami gospodarczymi i historycznymi. Można też zaobserwować pewne zróżnicowanie wyników ze względu na płeć, pojawiające się w części humanistycznej egzaminu gimnazjalnego. Dziewczynki uzyskują wyższe wyniki od chłopców, co może być potencjalnie związane z ich wyższym poziomem czytelnictwa, raportowanym zarówno w polskich (Zasacka, 2014), jak i zagranicznych (OECD, 2009) badaniach. Może to być sugestią do wprowadzenia szczególnych inicjatyw zorientowanych na poprawę umiejętności w zakresie czytania wśród chłopców, podobnych do inicjatyw wprowadzonych w Wielkiej Brytanii (zob. EACEA, 2010). Różnica w wynikach dziewcząt i chłopców w części humanistycznej egzaminu gimnazjalnego zmniejsza się jednak wraz z upływem czasu. Dla części matematyczno-przyrodniczej różnica wyników pomiędzy płciami jest praktycznie pomijalna. Jest to pozytywnym wnioskiem, biorąc szczególnie pod uwagę szeroko zakrojone działania promujące karierę w nauce i technologii wśród dziewcząt. Wyniki dla szkół niepublicznych, podobnie jak w przypadku sprawdzianu, są wyższe niż dla szkół publicznych. Rozstrzygnięcie o przyczynie tego zjawiska wymagałoby porównania efektywności prac szkół publicznych i niepublicznych, na przykład stosując wskaźniki edukacyjnej

wartości dodanej. Dostosowania wprowadzone dla uczniów z dysleksją (szczególnie ważne w przypadku części humanistycznej wymagającej formułowania dłuższych wypowiedzi) pozwalają na wyrównanie szans oraz wyników egzaminacyjnych – można zaobserwować stopniowe zmniejszanie się różnic pomiędzy obiema grupami uczniów. Zróżnicowanie wyników ze względu na typ gminy wygląda podobnie jak w przypadku sprawdzianu – brak istotnych statystycznie różnic pomiędzy gminami wiejskimi i miejsko-wiejskimi oraz różnica na korzyść gmin miejskich. Jednakże przyczynowość tego zjawiska nie musi leżeć po stronie efektywności pracy szkoły. Dodatkowo może być to związane z różnicowaniem się szkół – jak wskazują Dolata, Jasińska i Modzelewski (2012), proces ten jest szczególnie silny w dużych miastach.

Egzamin gimnazjalny jest więc egzaminem dość silnie warunkowanym przez procesy selekcji do szkół oraz w pewnym stopniu związany z płcią uczniów (różnice w części humanistycznej). Jego wyniki są także zróżnicowane ze względu na regiony kraju. Choć można wyróżnić pewne społeczno-gospodarcze oraz historyczne korelaty tego zjawiska, określenie przyczyn oraz procesu prowadzącego do występowania tego zróżnicowania wymaga dalszych badań.

3.3.3. Matura

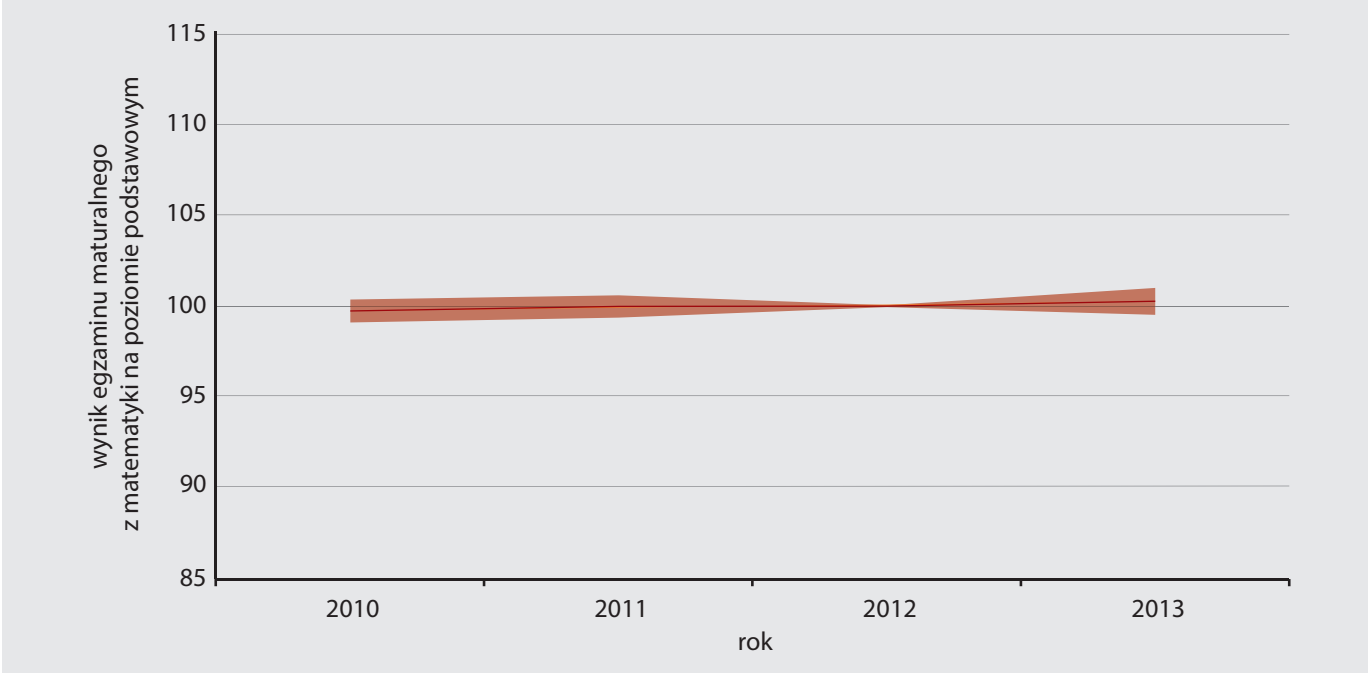
W niniejszej części przedstawimy porównywalne wyniki maturalne z matematyki i języka angielskiego dla poziomu podstawowego w latach 2010–2013[34]. Przy analizowaniu danych z tego etapu kształcenia szczególnie ważne jest określenie grupy uczniów, dla których przeprowadzane są analizy. Jednym z powodów takiej konieczności jest różnorodność typów szkół ponadgimnazjalnych. Kiedy myślimy o szkołach ponadgimnazjalnych, najczęściej mamy na myśli licea ogólnokształcące i technika dla młodzieży. Warto jednak pamiętać, że istnieją także licea profilowane, których uczniowie kształcą się nie tylko ogólnie, lecz także zawodowo. Poza tym oprócz szkół dla młodzieży istnieją szkoły dla dorosłych, w których specyfika nauczania różni się znacząco od szkół dla młodzieży. Trudno oczekiwać, że wyniki kształcenia w tak różnych typach szkół, utworzonych, aby spełniać różne cele, będą podobne. Większość uczniów na etapie kształcenia ponadgimnazjalnego uczy się w liceach ogólnokształcących, profilowanych i technikach. Uczniowie innych typów szkół stanowią niewielką grupę, zatem dokonywanie analiz w podziale na podgrupy staje się trudne ze względu na ich gwałtownie malejącą liczebność. Istotnym elementem jest też kompletność danych, którymi dysponujemy – niestety w przypadku szkół uzupełniających dla dorosłych częściej stykamy się z brakiem pewnych informacji niż w liceach i technikach dla młodzieży. W związku z powyższym, analizę porównywalnych wyników egzaminacyjnych egzaminu maturalnego przeprowadzono tylko dla uczniów z liceów ogólnokształcących, profilowanych i techników dla młodzieży. Bardziej istotny dla tego wyboru jest jednak fakt, że badania zrównujące przeprowadzono właśnie w takich szkołach – z powodów logistycznych i finansowych z badania wykluczono pozostałe typy szkół.

[34] W czasie powstawania tego raportu prace nad porównywalnymi wynikami maturalnymi z języka polskiego wciąż trwają.

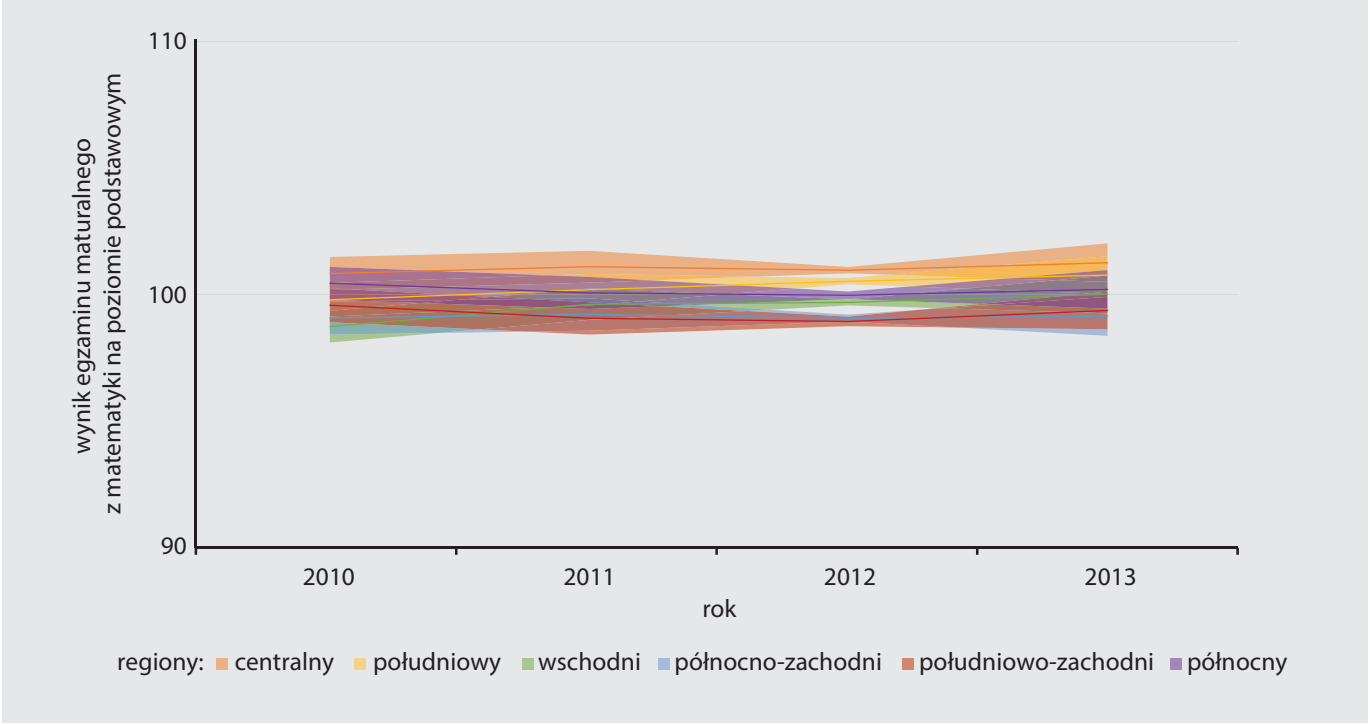
Matematyka

Omówienie porównywalnych wyników egzaminacyjnych dla matury rozpoczniemy od prezentacji ogólnego trendu dla egzaminu maturalnego z matematyki. rysunek 3.30 przedstawia średnie porównywalne wyniki dla poziomu podstawowego tego egzaminu. W analizowanych latach – od 2010 do 2013 roku, wyniki te są stabilne i nie wykazują fluktuacji. Jak już wspominano, jest to zgodne z oczekiwaniami w stosunku do pomiaru poziomu umiejętności populacji uczniów w krótkim okresie.

Rysunek 3.30. Porównywalne wyniki egzaminu maturalnego z matematyki w latach 2010–2013

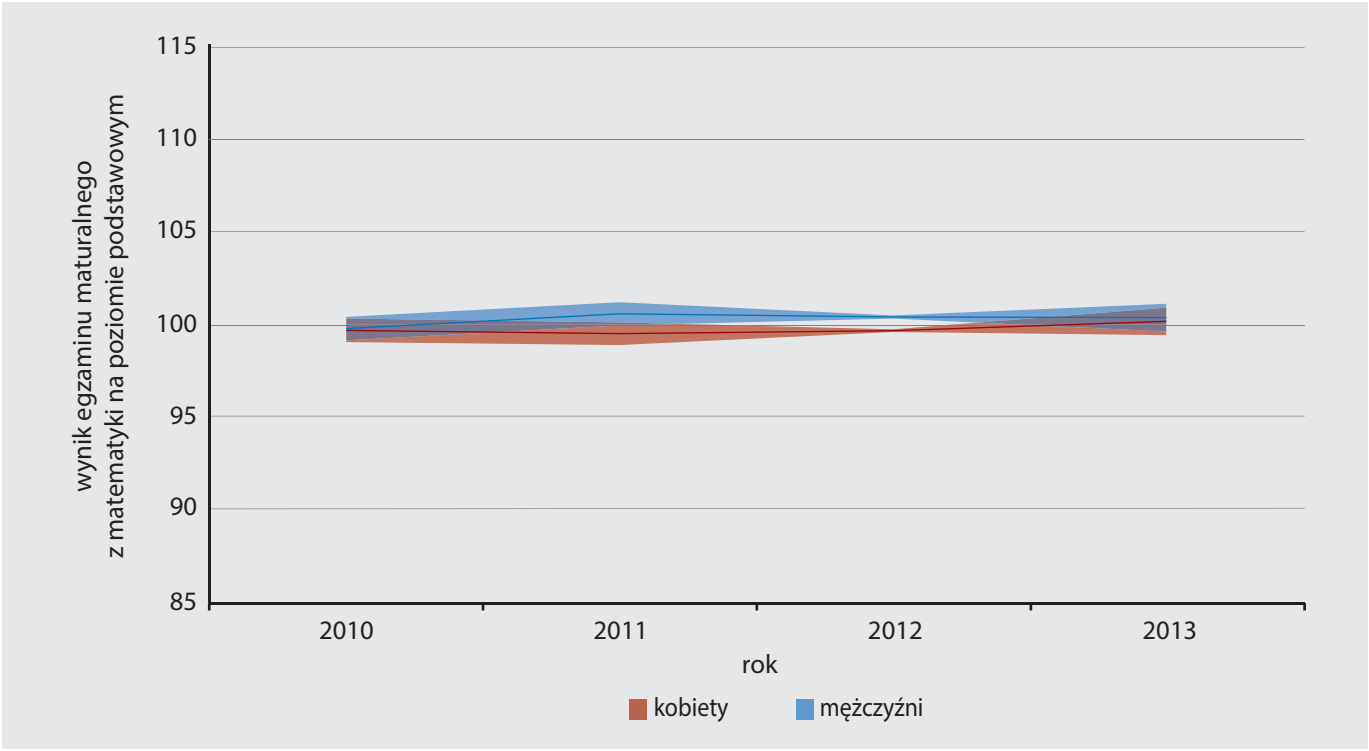


Rysunek 3.31. Porównywalne wyniki egzaminu maturalnego z matematyki w latach 2010–2013 w podziale na regiony wg NTS



Analiza porównywalnych wyników egzaminacyjnych ze względu na region ujawniła zróżnicowanie między zachodnią i wschodnią częścią kraju dla etapów sprawdzianu i egzaminu gimnazjalnego. Wyniki egzaminu maturalnego z matematyki w podziale na regiony kraju, które przedstawia rysunek 3.31, nie ujawniają znaczących różnic pomiędzy poszczególnymi częściami kraju, jak miało to miejsce na wcześniejszych etapach kształcenia. W takiej sytuacji można uznać, że poziom umiejętności uczniów mierzonych na egzaminie maturalnym z matematyki, w toku kształcenia w szkołach ponadgimnazjalnych, wyrównuje się w skali całego kraju. Jest to cecha pożądana w egalitarnym systemie edukacyjnym, bowiem brak jest oznak różnicowania szans w procesie rekrutacji na uczelnie wyższe w zależności od regionu, z którego pochodzi uczeń.

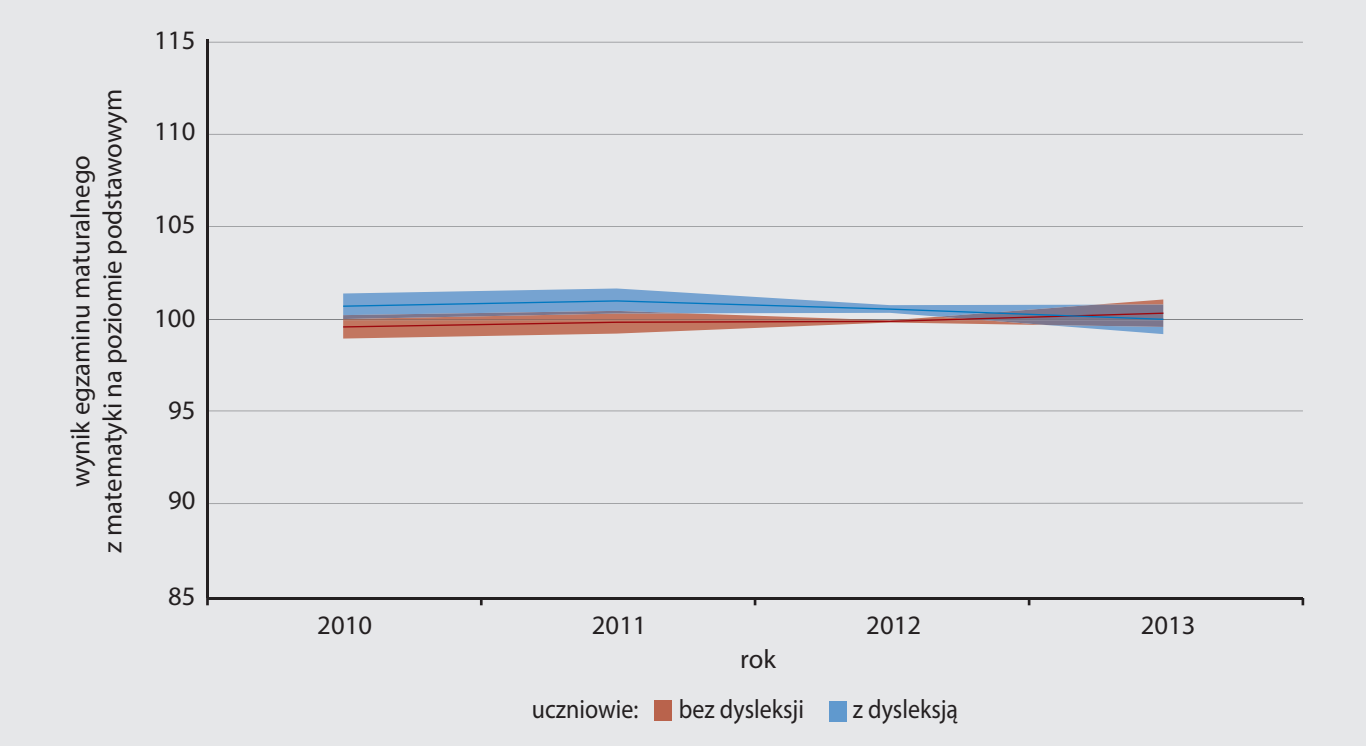
Rysunek 3.32. Porównywalne wyniki egzaminu maturalnego z matematyki w latach 2010–2013 w podziale na płeć uczniów



Zarówno płeć uczniów (zob. rysunek 3.32), jak też zdiagnozowana dysleksja rozwojowa (zob. rysunek 3.33) nie powodują znaczących różnic w wynikach osiąganych na egzaminie maturalnym z matematyki. Jest to zgodne z tym, co obserwujemy w ostatnich latach w części matematyczno-przyrodniczej egzaminu gimnazjalnego. Dostosowania dla uczniów z dysleksją rozwojową na egzaminie maturalnym obejmowały w analizowanym okresie wyłącznie zastosowanie szczegółowych kryteriów w ocenianiu[35], uczniowie ci nie dysponowali dodatkowym czasem na rozwiązanie zadań, jak miało to miejsce podczas wcześniejszych etapów edukacji.

[35] Pomijając możliwość pisania pracy na komputerze w szczególnych przypadkach głębokiego zaburzenia grafii.

Rysunek 3.33. Porównywalne wyniki egzaminu maturalnego z matematyki w latach 2010–2013 w podziale na grupy uczniów bez dysleksji rozwojowej i z dysleksją rozwojową



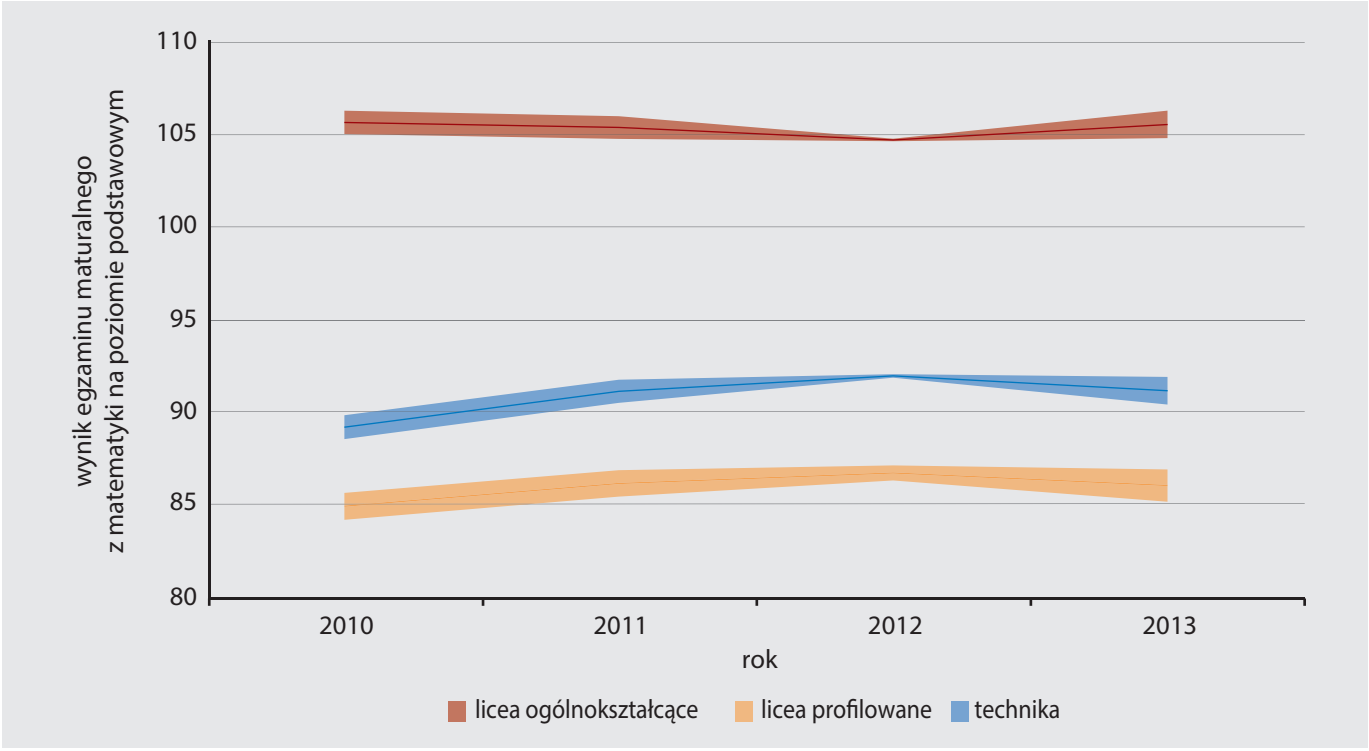
Specyfika szkół ponadgimnazjalnych powoduje, że trudno przeprowadzać na tym etapie edukacji analizy w podziale na szkoły publiczne i niepubliczne, czy też na rodzaj gminy, w której położona jest szkoła. Spośród wybranych do analizy uczniów (z liceów ogólnokształcących, profilowanych i techników dla młodzieży) zdecydowana większość uczy się w szkołach publicznych – w zależności od roku jest to 96–97%. W przypadku typu gminy, w której położona jest szkoła, mamy do czynienia z podobnymi problemami. Większość szkół ponadgimnazjalnych położona jest w gminach miejskich lub miejsko-wiejskich, zaledwie cztery procent uczniów z opisywanych szkół uczęszczało do szkół zlokalizowanych na terenie gmin wiejskich.

Właściwym sposobem analizy wyników w szkołach ponadgimnazjalnych wydaje się spojrzenie na wyniki w podziale na typy szkół: licea ogólnokształcące, profilowane i technika. Należy przy tym pamiętać, że zostały one utworzone w różnych celach, pomimo możliwości przystąpienia do egzaminu maturalnego w każdym z tych typów szkoły. Dopiero wewnątrz tych grup można rozważać różnice pomiędzy szkołami publicznymi i niepublicznymi, czy w zależności od typu gminy, w której mieści się szkoła.

Porównywalne wyniki z egzaminu maturalnego z matematyki dla lat 2010–2013 w podziale na typy szkół prezentuje rysunek 3.34. Najwyższe wyniki we wszystkich analizowanych latach osiągają uczniowie liceów ogólnokształcących, jednocześnie w tego typu szkołach uczą się prawie dwie trzecie uczniów. Uczniowie techników, w zależności od roku, osiągają wyniki niższe o 13–16 punktów od wyników uczniów liceów ogólnokształcących i stanowią drugą pod względem liczebności grupę. Najniższe wyniki, o ponad jedno odchylenie standardowe niższe niż uczniowie liceów ogólnokształcących (około 20 punktów), osiągają uczniowie liceów profilowanych. Są oni też najmniej liczną grupą uczniów, w roku 2010 stanowili cztery procent wszystkich uczniów, w 2011 trzy procent, a w latach 2012 i 2013 zaledwie dwa procent. Trendy w poszczególnych typach szkół nieco różnią się od siebie, lecz nie na tyle, aby mówić o systematycznych różnicach. W każdym z tych typów szkół uczniowie szkół niepublicznych uzyskiwali wyniki niższe niż uczniowie szkół publicznych. Należy

jednak mieć na uwadze małą liczbę ponadgimnazjalnych szkół niepublicznych dla młodzieży[36]. Podobnie dla każdego typu szkół uczniowie szkół z gmin miejskich uzyskiwali najwyższe wyniki, a uczniowie szkół z gmin wiejskich najniższe. Odmiennie niż w przypadku wyników sprawdzianu czy egzaminu gimnazjalnego, różnice pomiędzy gminami miejsko-wiejskimi i wiejskimi są istotne statystycznie. Zależność ta jest jednak mocno powiązana z typami szkół, które przeważają w danym rodzaju gmin. Procentowy udział liceów ogólnokształcących oraz uzupełniających liceów ogólnokształcących na wsi jest zdecydowanie niższy w porównaniu z technikami. Zgodnie z danymi GUS dla roku 2012/2013 na wsiach znajdowało się zaledwie 8% takich liceów (GUS, 2013, s. 214), jednakże dla tego samego roku technika na wsi stanowiły 14% ogółu tego typu szkół (GUS, 2013, s. 229).

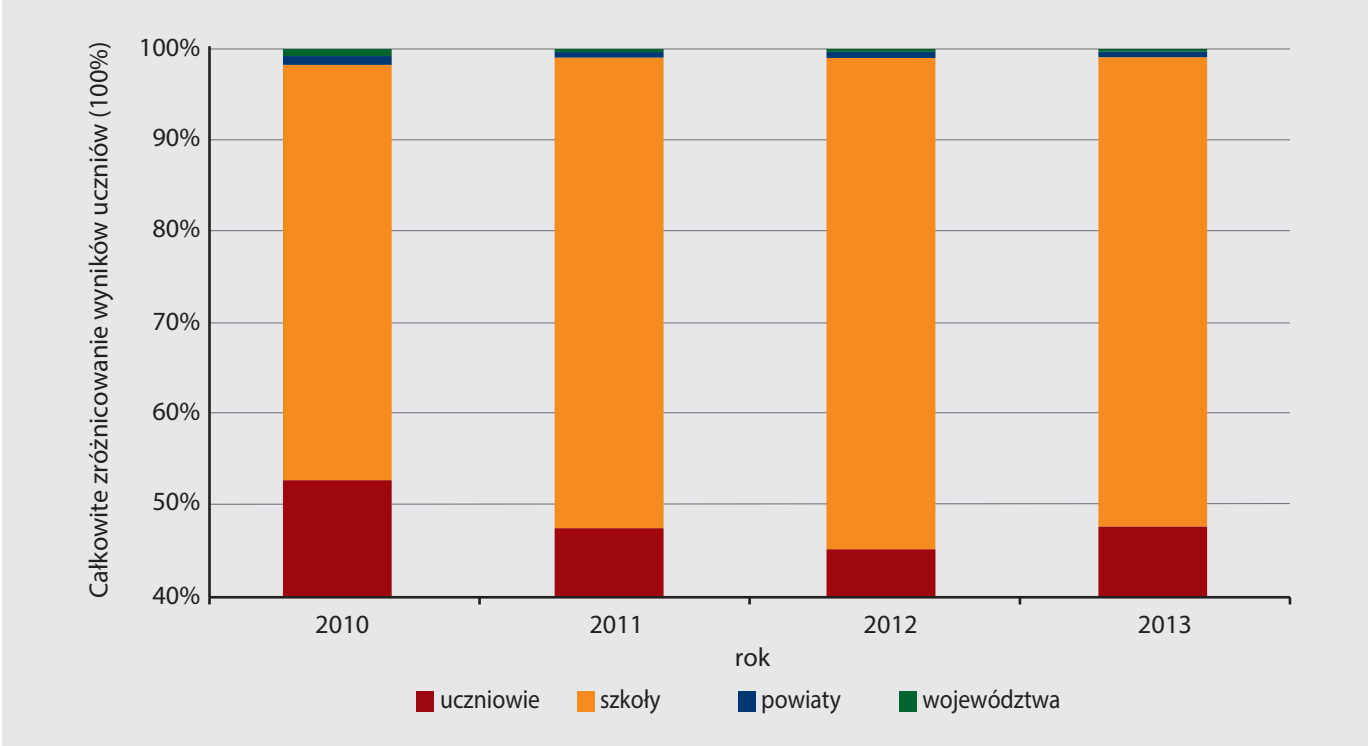
Rysunek 3.34. Porównywalne wyniki egzaminu maturalnego z matematyki w latach 2010–2013 w podziale na typy szkół



Zróżnicowanie wyników z matury z matematyki jest uwarunkowane w większości wpływem indywidualnych umiejętności ucznia oraz przynależności do danej szkoły, co ilustruje rysunek 3.35. Wpływ uczęszczania do szkoły w konkretnym powiecie lub województwie jest pomijalny – rzędu 1–2%. Co więcej, zróżnicowanie wyników ze względu na region zmniejsza się wraz z upływem czasu, natomiast zróżnicowanie wyników pomiędzy uczniami oraz zróżnicowanie międzyszkolne jest na porównywalnym poziomie. W latach 2011–2012 można zaobserwować wręcz wyższe zróżnicowanie międzyszkolne od zróżnicowania wewnątrzszkolnego. Zróżnicowanie międzyszkolne uwidaczniające się w wynikach matury z matematyki jest zdecydowanie wyższe (około 50%) niż to dla części matematyczno-przyrodniczej egzaminu gimnazjalnego (17–25%). Świadczy to o zdecydowanie rosnącej selekcyjności szkół ponadgimnazjalnych w stosunku do gimnazjów w zależności od poziomu umiejętności uczniów na wejściu.

[36] W roku szkolnym 2012/2013 w skali kraju były to 752 szkoły niepubliczne o uprawnieniach szkół publicznych oraz 4 szkoły niepubliczne w stosunku do 6383 ponadgimnazjalnych szkół publicznych (GUS, 2013, s.194–196).

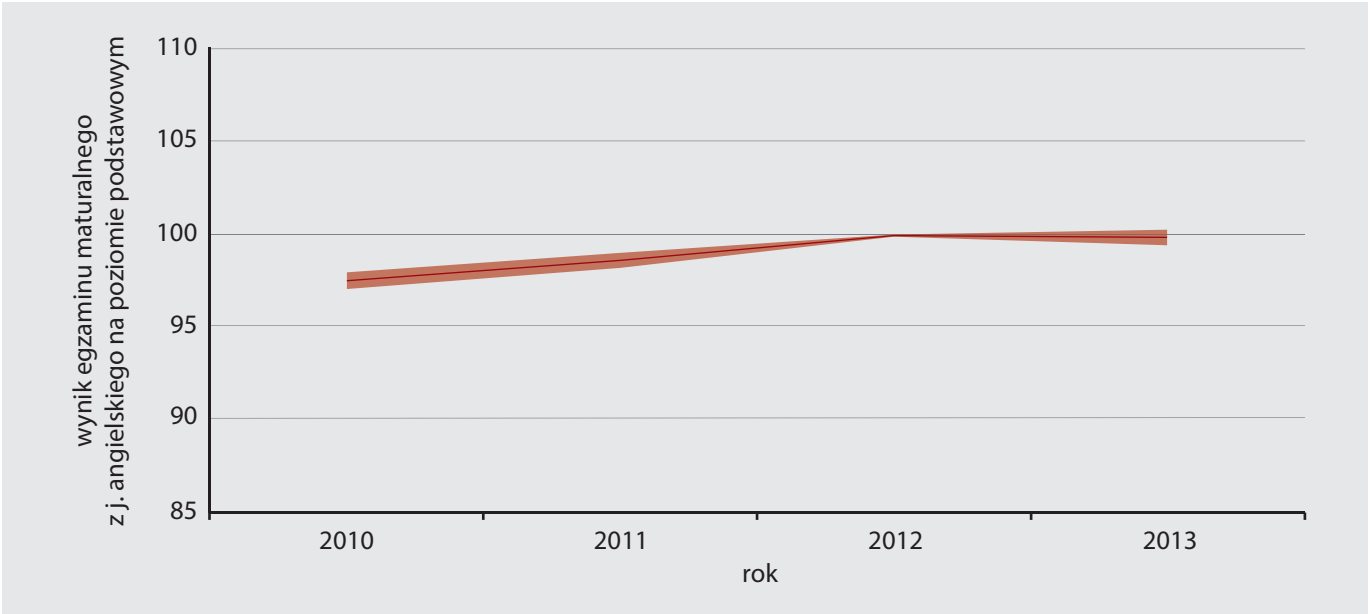
Rysunek 3.35. Całkowite zróżnicowanie wyników egzaminu maturalnego z matematyki w podziale na wpływ indywidualnych umiejętności ucznia oraz przynależności do szkoły, powiatu i województwa



Język angielski

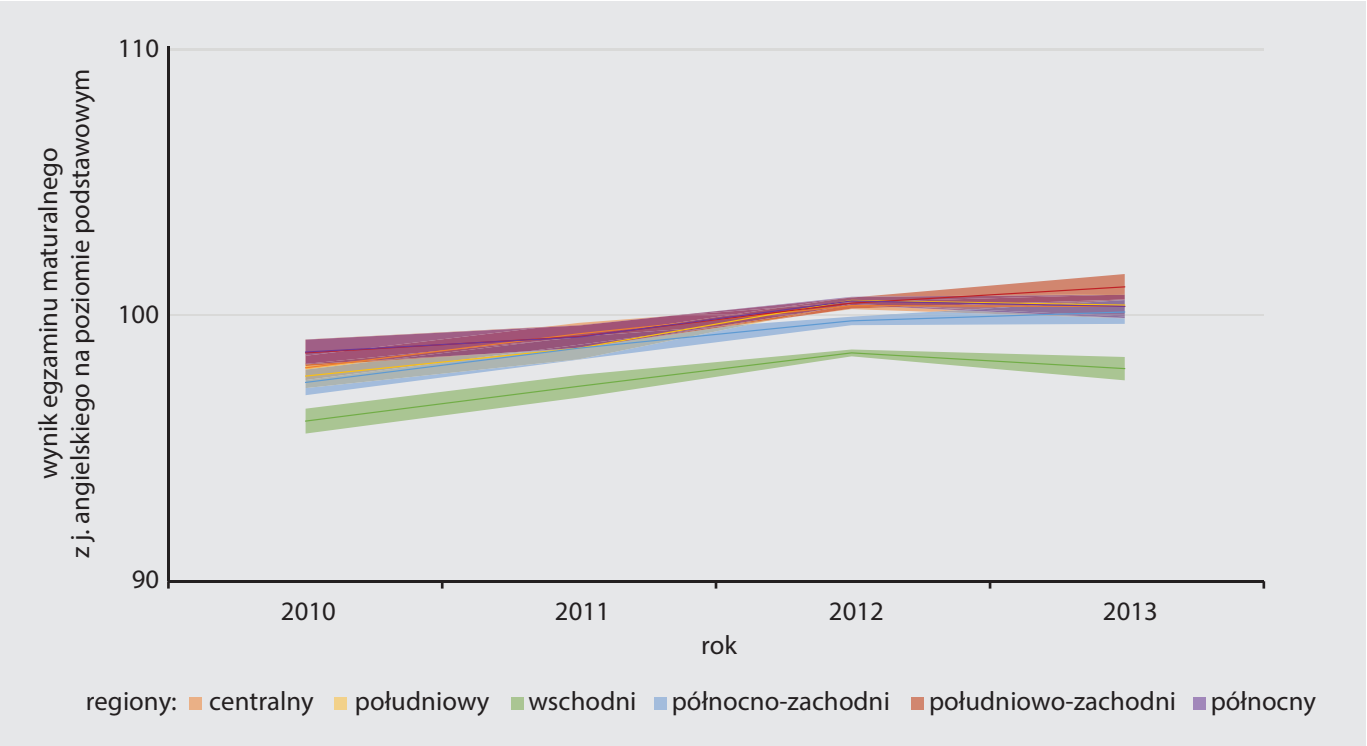
Język angielski jest najczęściej wybieranym przez uczniów językiem nowożytnym na egzaminie maturalnym. Według sprawozdań Centralnej Komisji Egzaminacyjnej z egzaminu maturalnego każdego roku język ten wybiera ponad 80% zdających. Jest to jednak tylko jeden z możliwych do wyboru języków, drugim najczęściej wybieranym jest język niemiecki, a trzecim rosyjski. Porównywalne wyniki z egzaminu maturalnego z języka angielskiego wykazują w latach 2010–2012 delikatny trend wzrostowy. Pomiędzy rokiem 2012 i 2013 poziom umiejętności uczniów mierzonych na tym egzaminie nie zmienia się. Omawiane wyniki ilustruje rysunek 3.36.

Rysunek 3.36. Porównywalne wyniki egzaminu maturalnego z języka angielskiego w latach 2010–2013



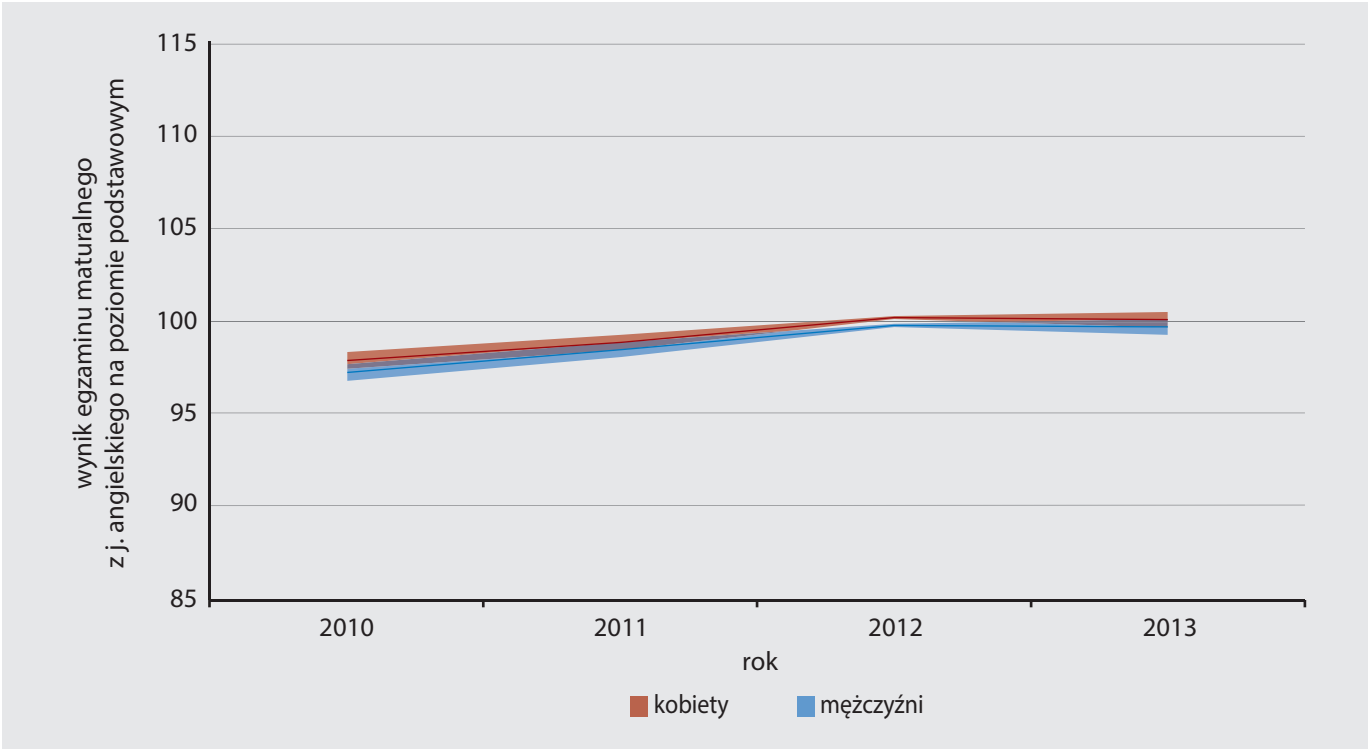
Analizując porównywalne wyniki egzaminu maturalnego z języka angielskiego w podziale na regiony kraju, można zauważyć, że region wschodni osiągał w opisywanych latach wyniki niższe niż pozostałe części Polski (zob. rysunek 3.37). Nie jest to różnica duża, wynosi około dwóch punktów w zależności od roku, jednak dla wszystkich lat jest istotna statystycznie. Można też zauważyć, że w regionie wschodnim wyniki pomiędzy rokiem 2012 i 2013 minimalnie spadają, podczas gdy w pozostałych regionach pozostają na tym samym poziomie lub nieznacznie rosną. Wybieralność języka angielskiego w regionie wschodnim nie różni się znacząco od wybieralności w regionie centralnym czy południowym, co niestety prowadzi do wniosku, że uczniowie w regionu wschodniego prawdopodobnie prezentują niższy poziom umiejętności z tego języka. Co dziwne, przy bliższym przyjrzeniu się wynikom w podziale na poszczególne województwa z regionu okazuje się, że najniższe wyniki z języka angielskiego osiągają uczniowie z województwa świętokrzyskiego. Zjawisko to zdaje się zatem nie mieć związku z geograficzną bliskością krajów posługujących się językami wschodniosłowiańskimi, gdyż wyniki z województw bezpośrednio sąsiadujących z Rosją, Białorusią, czy Ukrainą nie odbiegają od średniej ogólnopolskiej. Bliższe przyjrzenie się nauczaniu języków obcych w tej części Polski (a głównie w województwie świętokrzyskim) mogłoby rzucić nieco światła na ten fenomen.

Rysunek 3.37. Porównywalne wyniki egzaminu maturalnego z języka angielskiego w latach 2010–2013 w podziale na regiony wg NTS



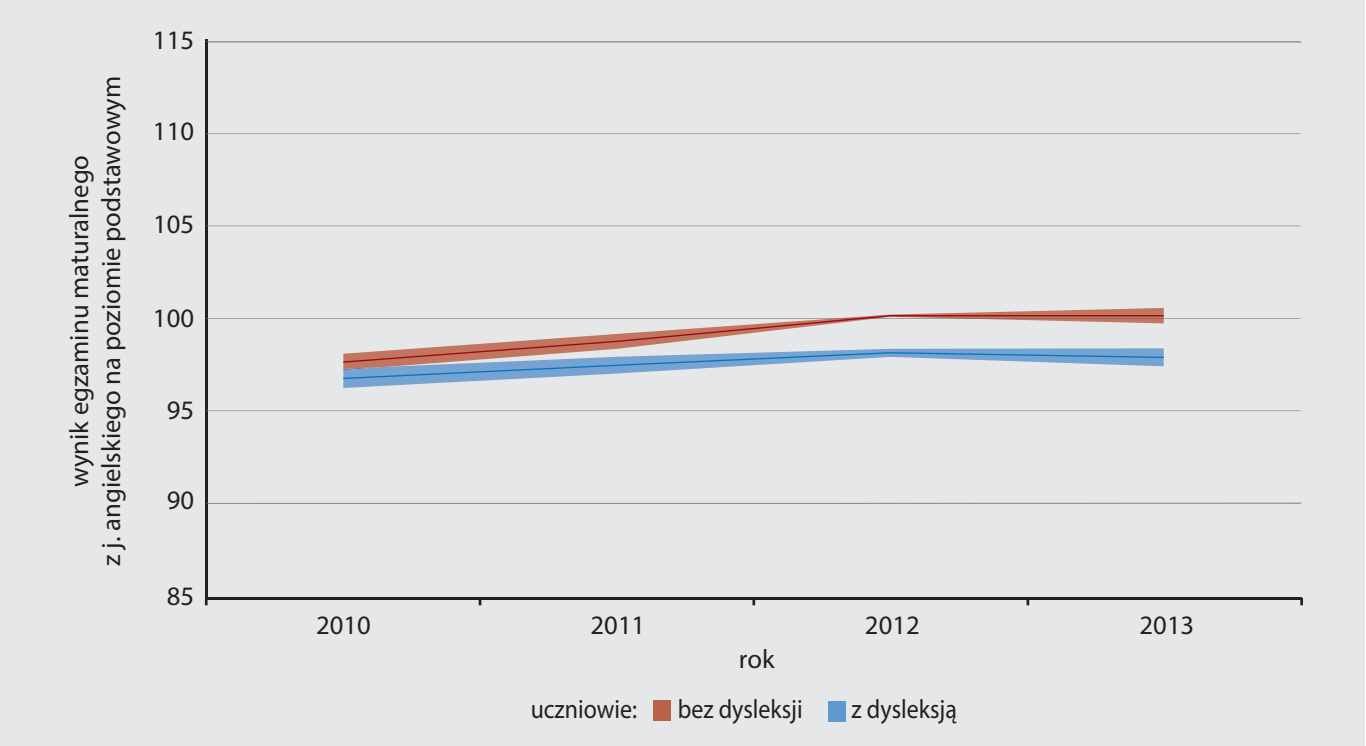
Analiza porównywalnych wyników egzaminu maturalnego z języka angielskiego w podziale na płeć uczniów (zob. rysunek 3.38) nie ujawnia różnic w poziomie umiejętności kobiet i mężczyzn. W okresie od 2010 do 2013 roku obydwie grupy osiągały, zaobserwowany dla całego kraju, minimalny wzrost wyników. Przyjmując zatem, że obydwie płcie nie różnią się między sobą pod kątem sprawdzanych umiejętności, można uznać, że egzamin maturalny z języka angielskiego nie dawał przewagi żadnej z nich.

Rysunek 3.38. Porównywalne wyniki egzaminu maturalnego z języka angielskiego w latach 2010–2013 w podziale na płeć uczniów



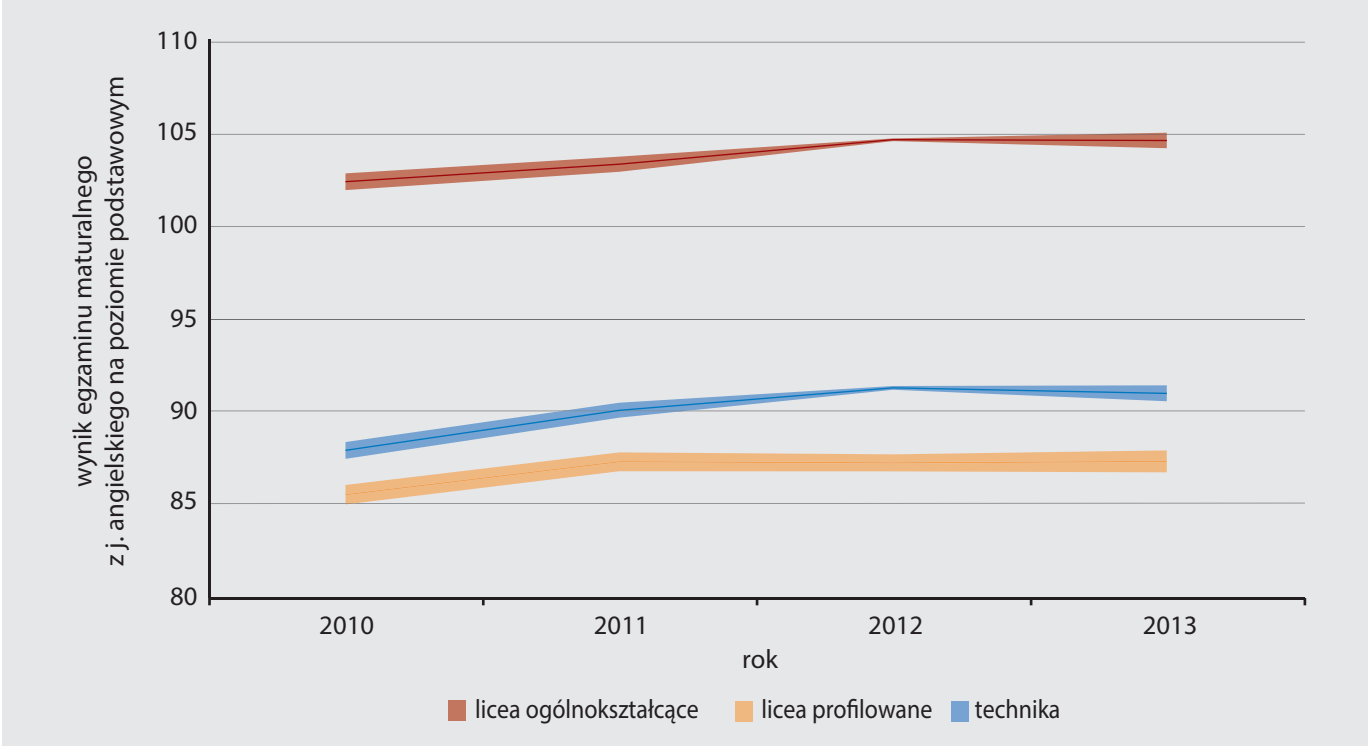
Odmiennie niż w przypadku egzaminu maturalnego z matematyki, w przypadku języka angielskiego możemy zaobserwować różnice w średnich wynikach pomiędzy uczniami z dysleksją rozwojową oraz bez niej. Niepokojące jest to, że choć w 2010 roku różnice te były znikome i nieistotne statystycznie, to w następnych latach systematycznie rosły (zob. rysunek 3.39). W roku 2013 uczniowie bez dysleksji rozwojowej osiągali średnio dwa punkty więcej niż ci, u których zdiagnozowano to zaburzenie. Może to świadczyć o tym, że dostosowania dla uczniów z dysleksją stosowane podczas egzaminu z języka angielskiego nie są wystarczające do wyrównania szans tej grupy w stosunku do uczniów bez zaburzeń. Wskazuje się, że dysleksja może przysparzać uczniowi szczególne problemy podczas nauki języków obcych – np. przez niemożność identyfikacji i rozróżniania fonemów, niezbędnych podczas rozumienia ze słuchu, oraz spowolnienie funkcjonowania pamięci długotrwałej, utrudniającej przyswojenie materiału (Zawadzka, 2010). U podłoża tych problemów może leżeć trudność w opanowaniu języka ojczystego, również spowodowana niską świadomością fonologiczną (Nijakowska, 2009).

Rysunek 3.39. Porównywalne wyniki egzaminu maturalnego z języka angielskiego w latach 2010–2013 w podziale na grupy uczniów bez dysleksji rozwojowej i z dysleksją rozwojową



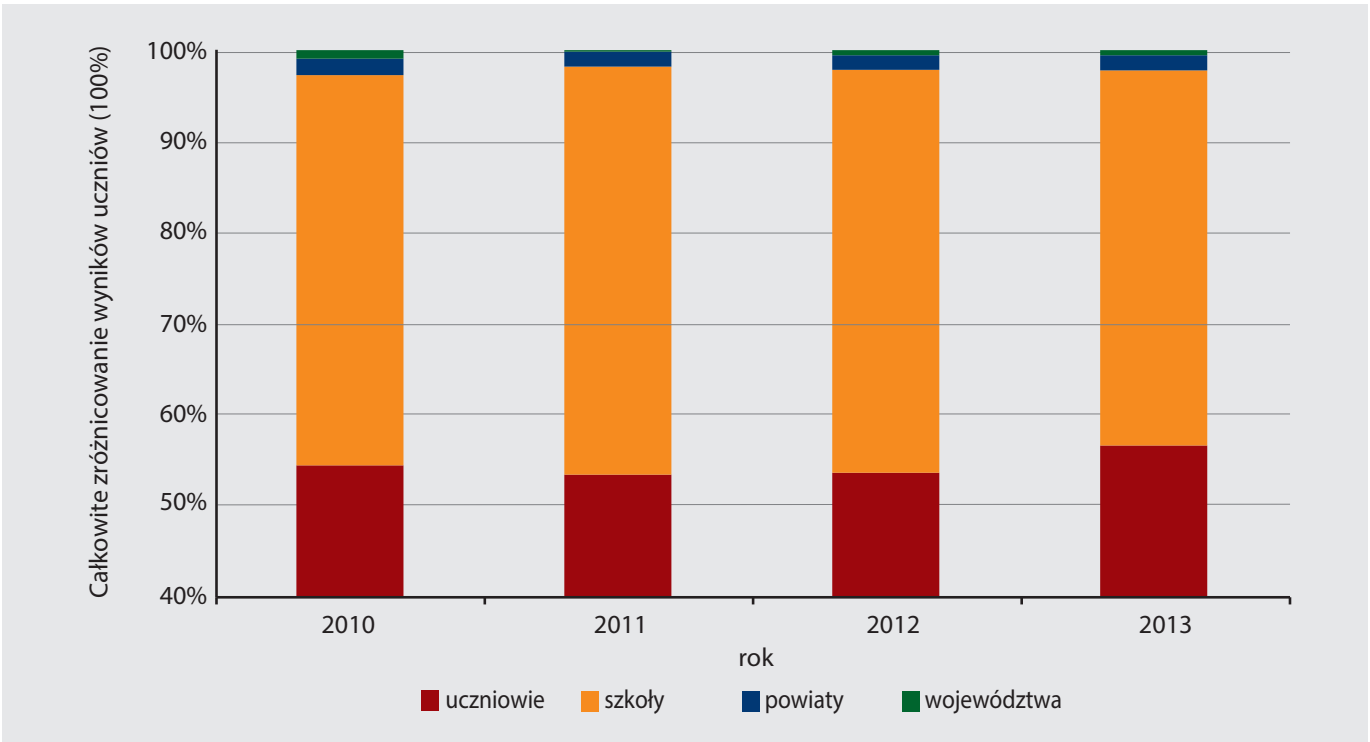
W przypadku analizy porównywalnych wyników egzaminu maturalnego z języka angielskiego w podziale na typy szkół (licea ogólnokształcące, profilowane i technika) dochodzimy do podobnych wniosków jak w przypadku matury z matematyki. Najwyższe wyniki uzyskują uczniowie liceów ogólnokształcących, o ponad dwie trzecie odchylenia standardowego niższe od nich uczniowie techników, a najniższe uczniowie liceów profilowanych (o ponad jedno odchylenie standardowe niższe niż uczniowie liceów ogólnokształcących) (zob. rysunek 3.40). Podobnie jak w przypadku egzaminu z matematyki, uczniowie szkół niepublicznych uzyskiwali wyniki niższe niż uczniowie szkół publicznych. Uczniowie ze szkół z gmin miejskich osiągali wyniki wyższe niż z gmin miejsko-wiejskich i wiejskich. Należy jednak pamiętać o związku pomiędzy typem szkoły i jej lokalizacją, na który zwracano uwagę w części poświęconej egzaminowi maturalnemu z matematyki.

Rysunek 3.40. Porównywalne wyniki egzaminu maturalnego z języka angielskiego w latach 2010–2013 w podziale na typy szkół



Zróżnicowanie wyników z matury z języka angielskiego, podobnie jak w przypadku matematyki, można przypisać w równym stopniu wpływowi indywidualnych umiejętności ucznia oraz przynależności do konkretnej szkoły. Natomiast zróżnicowanie związane z regionem (tj. powiatem i województwem), w którym znajduje się szkoła, jest praktycznie pomijalne. Trend ten jest stabilny (zob.

Rysunek 3.41. Całkowite zróżnicowanie wyników egzaminu maturalnego z języka angielskiego w podziale na wpływ indywidualnych umiejętności ucznia oraz przynależności do szkoły, powiatu i województwa



rysunek 3.41), choć należy pamiętać, że dysponujemy wynikami jedynie dla czterech lat. Porównując zróżnicowanie międzyszkolne dla matury z języka angielskiego (rzędu około 42–44%) ze zróżnicowaniem międzyszkolnym dla obu części egzaminu gimnazjalnego (17–25% dla części matematyczno-przyrodniczej oraz 13–28% dla części humanistycznej) możemy wnioskować o zdecydowanie wyższej selekcyjności szkół na etapie ponadgimnazjalnym niż gimnazjalnym. Warto jednak zauważyć, że zróżnicowanie międzyszkolne dla wyników matury z języka angielskiego jest niższe niż dla egzaminu maturalnego z matematyki.

Wyniki egzaminu maturalnego są w bardzo dużym stopniu związane z procesami zróżnicowania międzyszkolnego. Proces ten w szkołach ponadgimnazjalnych jest dużo silniejszy niż w gimnazjach, choć należy pamiętać, że szkoła maturalna nie jest powszechna i z założenia jest zróżnicowana (a nie jednolita). Konsekwencje tego wniosku mogą więc nie być aż tak znaczące jak w przypadku procesów selekcji na etapie gimnazjum. Zarówno w przypadku egzaminu maturalnego z matematyki jak i egzaminu maturalnego z języka angielskiego nie obserwujemy znaczących różnic w wynikach osiąganych przez obie płcie. Dla wyników maturalnych zanika wyraźna różnica pomiędzy regionem północnym i zachodnim a regionami centralnym, wschodnim i południowym, obserwowalna na wcześniejszym etapie edukacyjnym. Można jednak zaobserwować gorsze wyniki uzyskiwane z matury z języka angielskiego w regionie wschodnim – choć różnica nie jest duża, jest istotna statystycznie dla wszystkich lat. Zjawisko dysleksji nie różnicuje wyników matury z matematyki, jednakże wyniki osiągane z języka angielskiego przez uczniów posiadających opinię o dysleksji rozwojowej są niższe niż uczniów bez takiej diagnozy. W związku z tym, że nauka języków obcych może być wyjątkowo trudna dla osób z dysleksją rozwojową, dostosowania dla egzaminu z tego przedmiotu powinny w sposób szczególny odpowiadać ich potrzebom. Zarówno dla egzaminu maturalnego z matematyki, jak i z języka angielskiego najwyższe wyniki uzyskiwali uczniowie liceów ogólnokształcących, a uczniowie techników uzyskiwali wyższe wyniki niż uczniowie liceów profilowanych.

3.3.4. Podsumowanie

Porównywalne wyniki egzaminacyjne pozwalają na analizę zmian poziomu umiejętności uczniów na przestrzeni czasu. Zarówno dla sprawdzianu, części humanistycznej egzaminu gimnazjalnego, części matematyczno-przyrodniczej egzaminu gimnazjalnego, jak i matury z matematyki i matury z języka angielskiego wyniki egzaminacyjne przedstawione na wspólnej skali nie podlegają dużym wahaniom. W przypadku sprawdzianu i części humanistycznej egzaminu gimnazjalnego obserwujemy w poszczególnych latach niewielkie fluktuacje wyników, ale w dłuższej perspektywie czasowej wyniki tych egzaminów utrzymują się na podobnym poziomie. Poza spadkiem wyników w latach 2002–2004 również wyniki części matematyczno-przyrodniczej egzaminu gimnazjalnego z biegiem czasu utrzymują stabilny poziom. W przypadku matury z języka angielskiego obserwujemy niewielki wzrost wyników w latach 2010–2012. Krótkie przedziały czasowe, dla jakich dysponujemy porównywalnymi wynikami egzaminacyjnymi dla egzaminów maturalnych, nie pozwalają jednak na wnioskowanie o tym, czy te trendy będą się utrzymywać w kolejnych latach. Prezentowane wyniki wskazują na istnienie zróżnicowania terytorialnego umiejętności uczniów na drugim etapie edukacyjnym – uczniowie, którzy podchodzili do egzaminu gimnazjalnego w centralnej, południowej i wschodniej części kraju, uzyskiwali wyższe wyniki niż uczniowie z pozostałych regionów. W literaturze można spotkać się z historycznymi, ekonomicznymi i społecznymi korelatami tego zjawiska, ale prezentowane tu dane nie pozwalają na wnioskowanie o jego przyczynach. W przypadku matury z języka angielskiego tylko region wschodni charakteryzuje się niższymi wynikami (co jest powodowane głównie niskimi wynikami w województwie świętokrzyskim). Porównania wyników między płciami wskazały na przewagę dziewczynek w przypadku sprawdzianu i części humanistycznej egzaminu gimnazjalnego, natomiast pozostałe egzaminy nie różnicowały obu płci. W przypadku części humanistycznej egzaminu gimnazjalnego można próbować powiązać to zjawisko ze spadkiem czytelnictwa wśród chłopców. Zmienną, która generalnie nie różnicowała wyników egzaminacyjnych,

jest dysleksja. Jest to dobra wiadomość, ponieważ pośrednio potwierdza, że istniejące w systemie mechanizmy pozwalające na wyrównywanie szans dla uczniów z dysleksją spełniają swoją rolę. Jedynym egzaminem, którego wyniki różniły się na niekorzyść uczniów z diagnozą dysleksji, była matura z języka angielskiego. W związku z tym, że nauka języków obcych może być szczególnie trudna dla uczniów z dysleksją, dostosowania egzaminacyjne powinny dodatkowo uwzględniać tę specyfikę. Wyniki sprawdzianu i egzaminu gimnazjalnego są wyższe wśród uczniów ze szkół niepublicznych, jednak w przypadku części matematyczno-przyrodniczej egzaminu gimnazjalnego możemy obserwować, że z biegiem czasu różnica ta nieznacznie się zmniejsza.

Przedstawione analizy sugerują, że zróżnicowanie wyników znacznie różni się między poszczególnymi typami egzaminów. W przypadku sprawdzianu obserwujemy stosunkowo niewielkie różnice między szkołami, w przypadku egzaminu gimnazjalnego zróżnicowanie między szkołami w stosunku do zróżnicowania między uczniami zwiększa się, a w przypadku egzaminu maturalnego zróżnicowanie międzyszkolne dorównuje wewnątrzszkolnemu. Proces ten może odzwierciedlać nasilającą się selekcję międzyszkolną, gdzie w przypadku szkół podstawowych o wyborze szkoły decyduje przede wszystkim rejonizacja, podczas gdy wybór gimnazjum staje się w coraz większym stopniu pochodną indywidualnych preferencji uczniów i ich predyspozycji. Proces ten pogłębia się na etapie szkół ponadgimnazjalnych. Efektem selekcji są duże różnice między szkołami w przypadku matury, które są szczególnie wyraźne, jeśli porównamy wyniki liceów ogólnokształcących względem innych typów szkół. Dysponując jedynie danymi opisującymi te różnice, trudno jest jednak wnioskować o ich przyczynach i potencjalnych skutkach.

3.4. Możliwości wykorzystania porównywalnych wyników egzaminacyjnych

Możliwości wykorzystania PWE rozpatrywać można na kilku płaszczyznach:

- rodzajów analiz możliwych do wykonania z użyciem PWE;
- grup uczniów, dla których możliwe jest analizowanie PWE (np. szkoły, gminy, itd., ale też np. chłopcy i dziewczynki, dyslektycy i niedyslektycy, itp.);
- narzędzi ułatwiających przeprowadzanie analiz z wykorzystaniem PWE.

PWE umożliwiają przeprowadzanie wielu rodzajów analiz. Najważniejszym, unikalnym dla PWE rodzajem analiz jest porównywanie zmian w poziomie umiejętności uczniów w czasie. Możemy więc wypowiadać się na temat tego, czy uczniowie w jednym roku umieją więcej lub mniej niż w innym i jak duża jest to różnica. Drugi rodzaj możliwych porównań to śledzenie względnych relacji pomiędzy poziomem umiejętności uczniów oraz zmian tych relacji w czasie[37]. Może się np. okazać, że co prawda poziom umiejętności uczniów w naszej szkole, gminie, itp. wzrósł na skali bezwzględnej (a więc umieją więcej niż uczniowie w poprzednich latach), jednak w skali kraju wzrost ten był jeszcze większy (a więc relatywnie do innych uczniów w obecnym roczniku uczniowie danej szkoły, gminy, itp. wypadają gorzej niż w poprzednich latach). Istotnym zastrzeżeniem w wypadku obydwu opisanych rodzajów analiz jest, że o ile bardzo dobrze pozwalają one opisać obserwowane zmiany w poziomie umiejętności uczniów, o tyle nie pozwalają w jednoznaczny sposób wypowiadać się na temat tego, czym zmiany te zostały spowodowane. Wnioskowanie na temat obserwowanych zjawisk mówi tu raczej o współwystępowaniu (np. średnio dziewczynki osiągają na sprawdzianie wyższy wynik niż chłopcy) niż o tym, jaka jest przyczyna występowania tych zjawisk.

[37] Ten rodzaj analiz możliwy jest również na innych wystandaryzowanych postaciach wyników egzaminacyjnych, np. wynikach wystandaryzowanych ekwikwantylowo lub wyskalowanych z wykorzystaniem Item Response Theory.

Dane, na podstawie których policzone zostały PWE, umożliwiają dokonywanie analiz w podziale na grupy ze względu na:

- informacje o uczniu: płeć, rok urodzenia, posiadanie opinii o dysleksji podczas przystępowania do danego egzaminu oraz bycie laureatem konkursu zwalniającego z przystępowania do danego egzaminu, szkoła, w której uczeń przystępował do danego egzaminu;
- informacje o szkole: gminę, powiat i województwo, w których znajduje się szkoła, liczbę mieszkańców miejscowości, w której znajduje się szkoła, informację o tym, czy szkoła jest placówką publiczną, czy niepubliczną, czy szkoła jest placówką specjalną oraz czy jest to szkoła artystyczna.

Ponieważ PWE obejmują całą populację uczniów[38], nie stanowi problemu wykonywanie analiz w podziale na złożone grupy, np. ze względu na płeć i bycie dyslektykiem w poszczególnych powiatach województwa świętokrzyskiego.

Aby ułatwić korzystanie z PWE, powstał serwis internetowy http://pwe.ibe.edu.pl. Umożliwia on wykonywanie różnorodnych analiz wykorzystujących PWE obejmujących również wizualizację na wykresach oraz pobranie zbioru danych z wartościami PWE do dalszej analizy w innych programach (np. arkuszu kalkulacyjnym albo programach statystycznych). Podstawowym ograniczeniem serwisu internetowego jest ograniczenie dostępnych sposobów pogrupowania uczniów do szkół, gmin, powiatów, województw oraz wartości ogólnopolskich. Jeśli ktoś chciałby dokonać analiz PWE, grupując uczniów według innych kryteriów, jest to możliwe, jednak wymaga większej wiedzy i poświęcenia większej ilości czasu[39]. W takim wypadku należy skorzystać z pakietu ZPD dla programu statystycznego R, który opisany został na stronie http://zpd.ibe.edu.pl/doku.php?id=r_zpd.

W dalszej części rozdziału zamieszczone zostały przykłady analiz prezentujące możliwości wykorzystania PWE na wszystkich wyżej wspomnianych płaszczyznach. Jako główną oś podziału wybrano poziom agregacji – począwszy od makroskopowego, przez regionalny, do analiz na poziomie lokalnym.

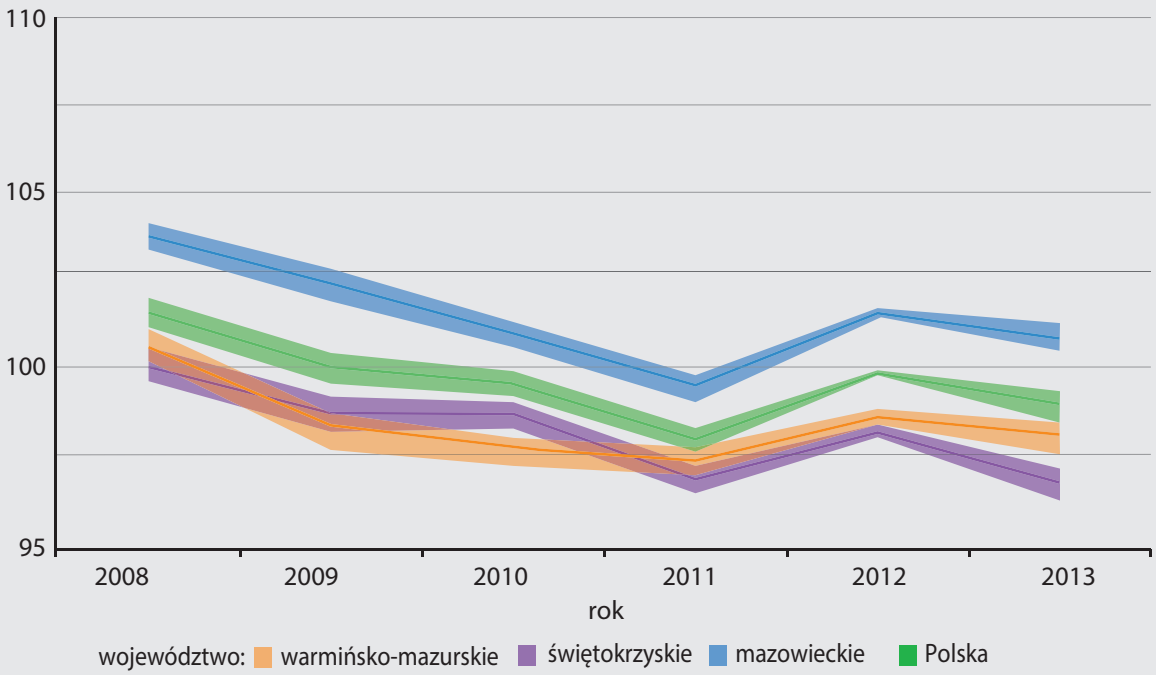
3.4.1. Analizy makroskopowe

Jako podstawowe zastosowanie PWE wskazać można analizy zmian wyników w czasie na poziomie makroskopowym. Tego typu analizy mogą być wartościowe szczególnie dla organów administracji państwowej, w celu ewaluacji efektów działań podejmowanych na szczeblu centralnym lub szczeblu województw. Wyniki różnorodnych analiz na poziomie makroskopowym zawarto w rozdziale 3.3, w tym miejscu przytoczona została się więc tylko jedna, przykładowa. Jest to analiza zmian w średnim poziomie umiejętności uczniów w części matematyczno-przyrodniczej egzaminu gimnazjalnego w latach 2003–2010 dla trzech wybranych województw (mazowieckiego, świętokrzyskiego i warmińsko-mazurskiego) oraz całej całej Polski (rysunek 3.43). Analiza przeprowadzona została z wykorzystaniem serwisu http://pwe.ibe.edu.pl, z niego też pochodzi poniższy wykres.

[38] Precyzyjnie – wszystkich uczniów, którzy rozwiązywali arkusz podstawowy danego egzaminu.

[39] Czasochłonność obliczania PWE dla dowolnego podziału uczniów na grupy jest podstawowym powodem, dla którego nie są one dostępne w serwisie http://pwe.ibe.edu.pl.

Rysunek 3.42. Średnie PWE z lat 2003–2010 dla trzech wybranych województw oraz Polski dla części matematyczno-przyrodniczej egzaminu gimnazjalnego



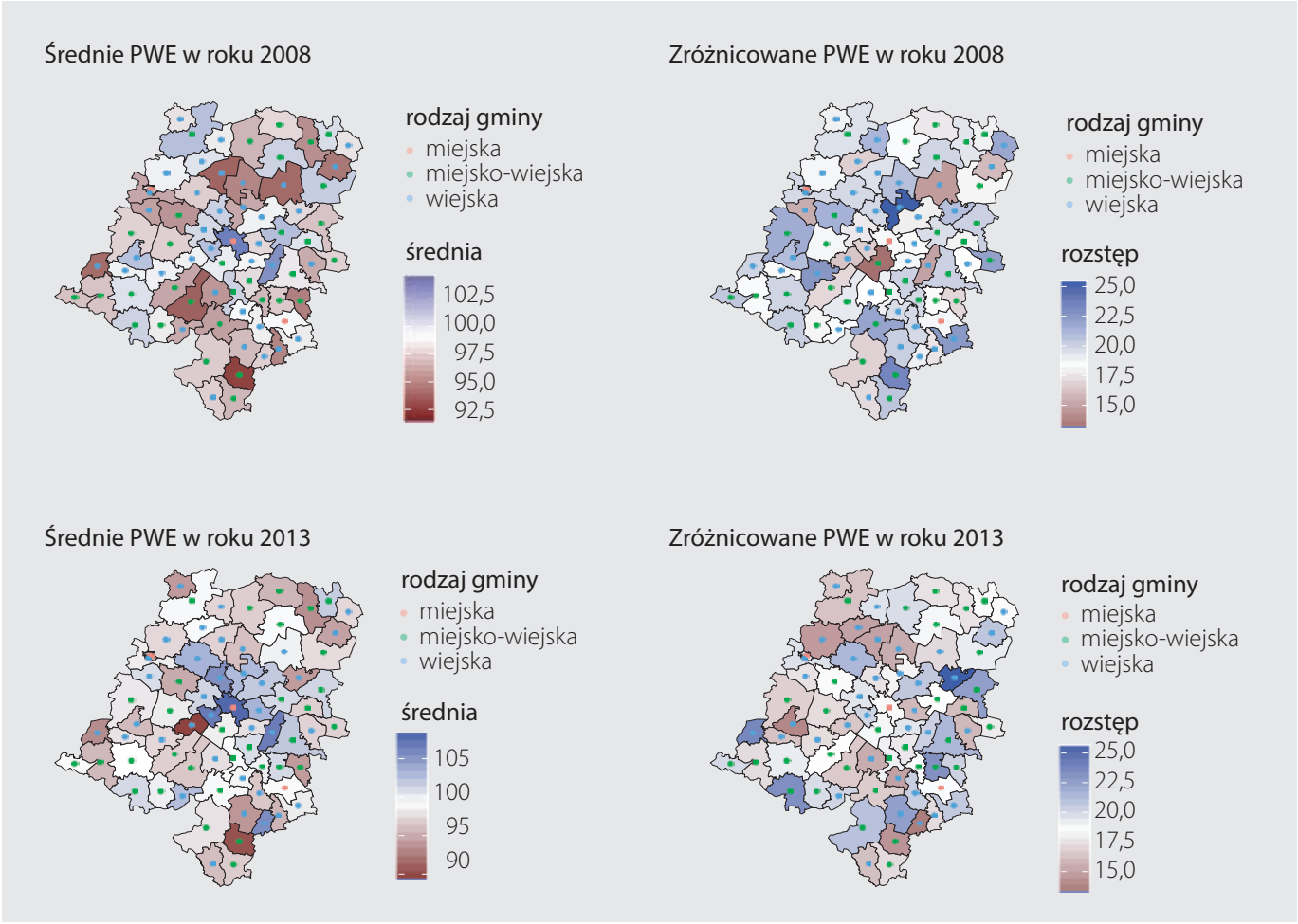
Wskazuje ona na istnienie znacznego zróżnicowania średnich PWE pomiędzy rozważanymi województwami, które jednak zmniejsza się z upływem lat. W roku 2003 różnica pomiędzy średnim wynikiem w województwie warmińsko-mazurskim i mazowieckim wynosiła 7,3, a więc niespełna połowę odchylenia standardowego, natomiast w latach 2007-2010 nie przekraczała już 3,7 punktów na skali, czyli zmalała prawie dwukrotnie. Widać też, że o ile województwa warmińsko-mazurskie oraz mazowieckie utrzymują w poszczególnych latach dość stabilną pozycję względem wyników ogólnopolskich, o tyle dla województwa świętokrzyskiego obserwowana jest wyraźna zmiana. W latach 2003-2004 średnie wyniki w województwie świętokrzyskim były wyraźnie wyższe niż średnie wyniki ogólnopolskie i tylko nieznacznie (nieistotnie statystycznie) niższe niż w województwie mazowieckim, jednak obserwowany w pozostałych województwach oraz całej Polsce trend spadkowy z lat 2003–2004 utrzymał się w województwie świętokrzyskim aż do roku 2006. Spowodowało to, że w latach 2007–2010 średni wynik w województwie świętokrzyskim był porównywalny już nie z województwem mazowieckim, lecz ze średnią krajową i województwem warmińsko-mazurskim, które w tym czasie zdążyło wyraźnie zmniejszyć dystans dzielący je od wyników ogólnopolskich. Co więcej, obserwowany we wszystkich rozważanych województwach oraz w całej Polsce trend spadkowy z lat 2008–2010 przebiegał z większą intensywnością dla województwa świętokrzyskiego i w 2010 roku osiągnęło ono najniższy (istotnie statystycznie niższy niż średnia ogólnopolska i nie różniący się istotnie statystycznie od średniej województwa warmińsko-mazurskiego) średni wynik spośród rozpatrywanych jednostek.

3.4.2. Analizy regionalne

Kolejną grupę analiz stanowią te na poziomie regionalnym, gdzie porównywane będą ze sobą wyniki uczniów w ramach powiatów czy gmin. Analizy takie mogą wspomagać prowadzenie polityki oświatowej na szczeblu kuratorium oświaty (a więc województwa) czy powiatu, jak również służyć do ewaluacji programów wspierania oświaty obejmujących swoim zasięgiem większe obszary.

Przykładem tego typu analiz może być np. ocena średnich PWE oraz średniego zróżnicowania PWE[40] na sprawdzianie w gminach województwa opolskiego w 2008 i 2013 roku (rysunek 3.43). Dane do tej analizy pobrane zostały z serwisu http://pwe.ibe.edu.pl, a następnie zwizualizowane w zewnętrznym programie.

Rysunek 3.43. Średnie PWE oraz rozstęp międzyćwiartkowy PWE na sprawdzianie szóstoklasisty w gminach województwa opolskiego w latach 2008 i 2013



W kontekście regionalnym interesować może nas odpowiedź na takie pytania, jak np.:

- Jak przebiega zróżnicowanie PWE pomiędzy gminami i czy zależności te zmieniają się pomiędzy rokiem 2008 i 2013?
- Czy daje się zaobserwować pewne wzorce przestrzenne (np. gminy miejskie o wyższych średnich wynikach otoczone przez gminy wiejskie o niższych)? Czy przebiegają one tak samo w roku 2008 i 2013?

Niewątpliwie wyróżniający się rejon stanowią Opole i jego bliskie okolice – średnie PWE tych gmin należą do najwyższych w województwie zarówno w roku 2008, jak i 2013. Warto przy tym zauważyć, że jednocześnie tereny te charakteryzują się przeciętnym zróżnicowaniem PWE uczniów w ramach gminy, a więc wyższym średnim wynikom nie towarzyszy większe rozwarstwienie poziomu uczniów. Pomiędzy rokiem 2008 a 2013 zaobserwować można, że gminy te zwiększyły nieco dystans, jaki dzieli je od reszty województwa, np. w 2008 roku średnie PWE województwa wynosiły 98,0

[40] Precyzyjnie rozstępu międzyćwiartkowego, czyli różnicy pomiędzy najniższym PWE spośród 25% najlepszych uczniów i najwyższym PWE spośród 25% najsłabszych uczniów.

wobec 104,9 dla Opola (różnica 6,9), a w 2013 roku 99,8 wobec 108,1 (różnica 8,3). Również gminy należące do powiatów strzeleckiego i krapkowickiego (na południe i południowy wschód od Opola) należą do tych rejonów, gdzie w obydwu rozważanych latach zaobserwować można wyniki wyższe niż przeciętne w województwie. W wypadku powiatu strzeleckiego wyższym wynikom średnim towarzyszy jednak często większe zróżnicowanie poziomu uczniów w ramach gminy. Z kolei większość gmin okalających wyżej wymienione tereny od południa, zachodu i północy konsekwentnie zarówno w roku 2008, jak i 2013 osiągają średnie PWE poniżej przeciętnej w województwie.

Na tle tych ogólnych zależności wyróżniają się niektóre gminy, jak np. gmina Lasowice Wielkie. W roku 2008 średnie PWE w tej gminie wynosiło 93,5 i należało do najniższych w województwie. Wynik w roku 2013 wynoszący 98,7 jest co prawda nadal nieco niższy od średniej w województwie (99,8), jednak różnica ta nie jest już istotna statystycznie, a w stosunku do roku 2008 zanotowano poprawę o 5,2 punktu (ponad 1/3 odchylenia standardowego). Przy formułowaniu wniosków trzeba być jednak ostrożnym, gdyż w gminie tej uczy się niewielka liczba uczniów (80 w 2008 roku, 54 w 2013 roku). W takim wypadku najbezpieczniej byłoby uwzględnić w analizie także dane z pozostałych lat i sprawdzić, czy opisana wyżej różnica wpisuje się w zmiany obserwowane na przestrzeni tych lat, czy jest raczej wynikiem ogólnej niestabilności średniego wyniku w tak małej gminie.

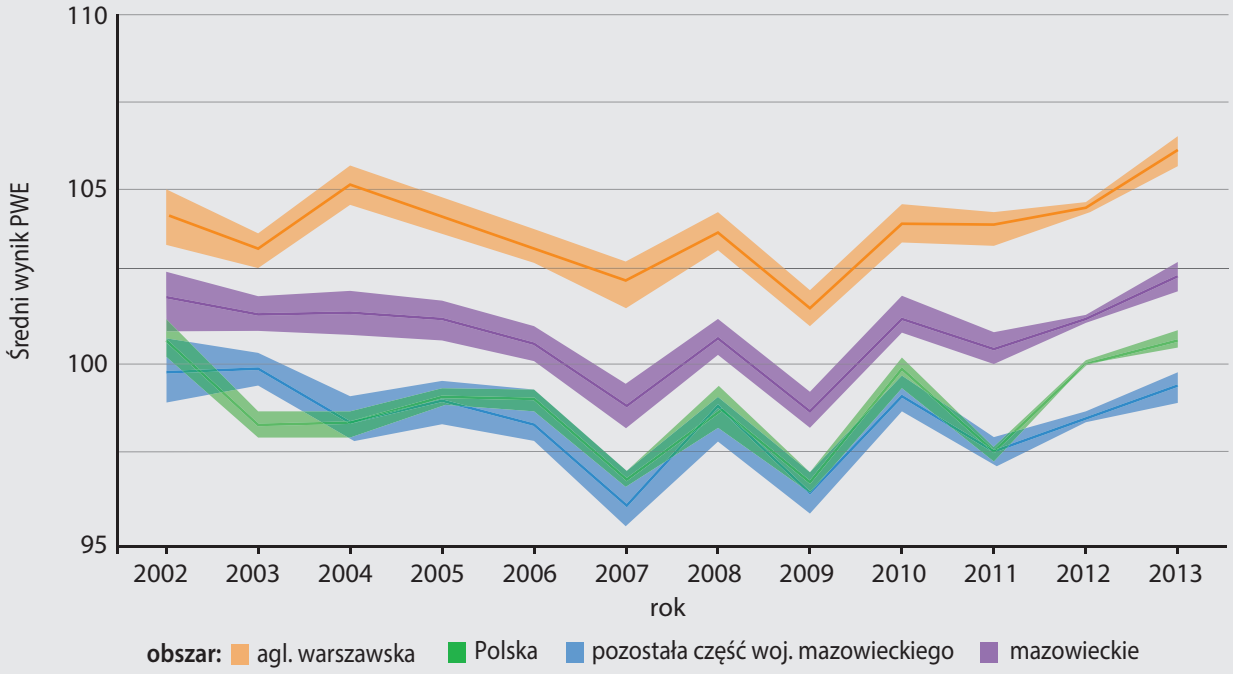
Drugim pytaniem, jakie postawiliśmy na początku, była kwestia występowania wzorców przestrzennych, np. odnośnie do średnich wyników w gminach miejskich i otaczających je gminach wiejskich. W województwie opolskim istnieją dwie gminy miejskie: Kędzierzyn-Koźle (na południowym wschodzie) i Brzeg (na północnym zachodzie), jak również powiat grodzki Opole, który potraktujemy w tym wypadku jako trzecią gminę miejską. Niewątpliwie powtarzalny jest wzorzec dla Opola – jego średnie wyniki są w obydwu latach wyższe od średnich wyników w otaczających je gminach wiejskich. Podobną zależność zaobserwować można dla Kędzierzyna-Koźla, przy czym warto zwrócić uwagę na to, że średnie wyniki PWE w gminach wiejskich otaczających Opole są wyraźnie wyższe od tych w gminach wiejskich otaczających Kędzierzyn-Koźle i porównywalne raczej ze średnimi wynikami w samym Kędzierzynie-Koźlu. Także gmina Brzeg w 2008 roku osiągnęła średnie wyniki PWE wyższe od otaczających ją gmin wiejskich, jednak w 2013 roku sytuacja uległa odwróceniu wobec obniżenia się średnich wyników w Brzegu i wzrostu w sąsiadujących z nim gminach wiejskich. Możemy więc stwierdzić, że rozważany przez nas wzorzec przestrzenny znajduje potwierdzenie w większości sytuacji.

Innym przykładem diagnozy regionalnej może być analiza PWE dla samodzielnie określonych grup jednostek samorządu terytorialnego. Jako przykład takiej analizy posłużyć może porównanie na przestrzeni lat średnich wyników PWE dla części humanistycznej egzaminu gimnazjalnego dla aglomeracji warszawskiej[41] oraz pozostałych terenów województwa mazowieckiego (rysunek 3.44). W odróżnieniu od prezentowanych wcześniej analiz tej nie da się przeprowadzić (ani nawet pobrać do niej danych) za pomocą serwisu http://pwe.ibe.edu.pl. Zamiast tego trzeba skorzystać z pakietu ZPD dla R[42].

[41] Warszawa, powiaty grodziski, legionowski, piaseczyński, warszawski zachodni, żyrardowski oraz gminy Chynów, Grójec, Pniewy, Dębe Wielkie, Halinów, Mińsk Mazowiecki (miejska i wiejska), Sulejówek, Czosnów, Leoncin, Nowy Dwór Mazowiecki, Pomiechówek, Zakroczym, Celestynów, Józefów, Karczew, Kołbiel, Otwock, Wiązowna, Brochów, Nowa Sucha, Sochaczew (miejska i wiejska), Teresin, Dąbrówka, Klembów, Kobyłka, Marki, Radzymin, Tłuszcz, Wołomin, Ząbki, Zielonka, Somianka, Wyszków, Zabrodzie.

[42] Patrz http://zpd.ibe.edu.pl/doku.php?id=r_zpd oraz https://github.com/zozlak/ZPD

Rysunek 3.44. Średnie PWE z części humanistycznej egzaminu gimnazjalnego wraz z 95% przedziałami ufności dla aglomeracji warszawskiej, pozostałej części województwa mazowieckiego oraz Polski



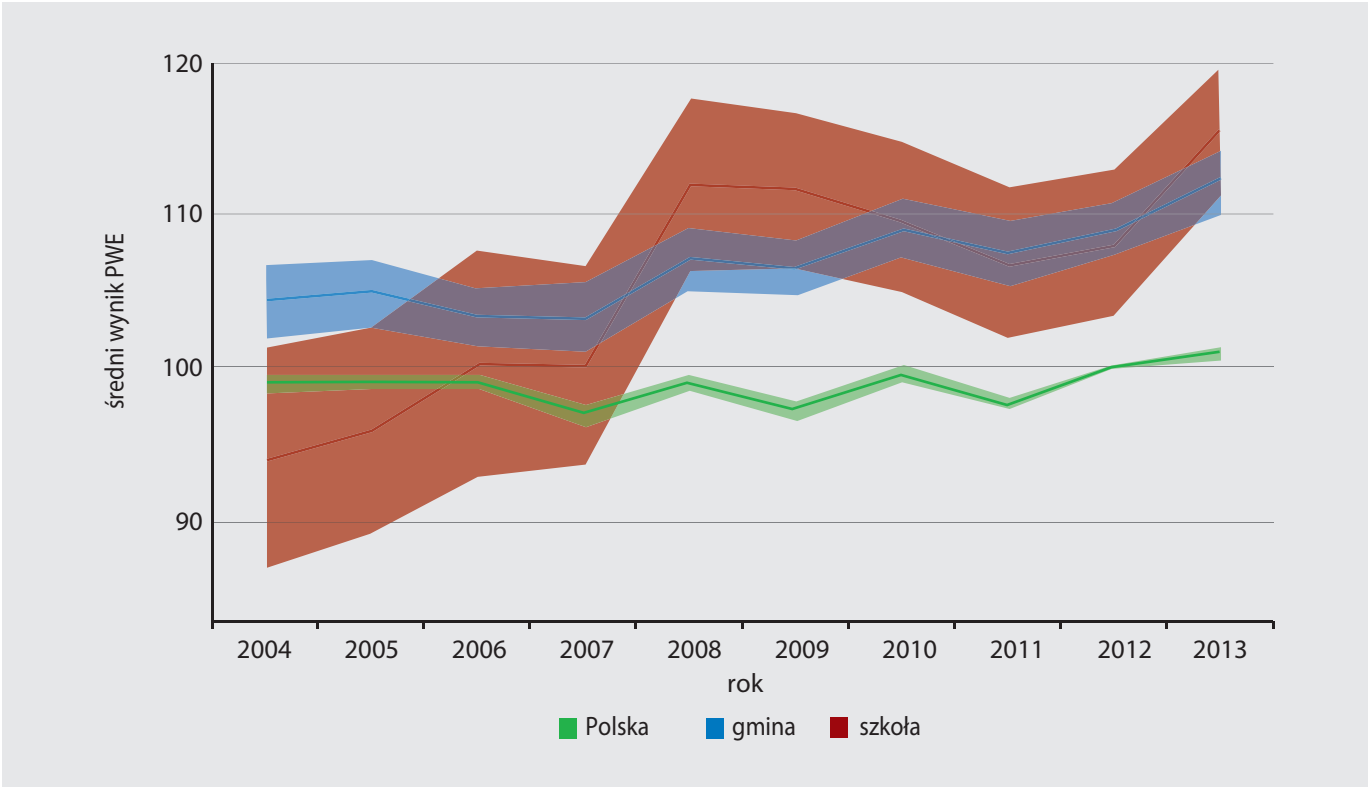
Porównując średnie PWE województwa mazowieckiego oraz Polski, odnotowujemy systematycznie i statystycznie istotnie (wyłączywszy rok 2002) wyższe wyniki osiągane przez uczniów w województwie mazowieckim. Jeśli jednak podzielić województwo mazowieckie na aglomerację warszawską oraz pozostałe obszary, wtedy widać, że przewaga ta bierze się z bardzo wysokich wyników osiąganych przez uczniów z Warszawy i okolic, podczas gdy średni wynik dla pozostałej części województwa jest bardzo zbliżony do średniej ogólnopolskiej. Dla lat 2005–2010 i 2012–2013 różnica pomiędzy aglomeracją warszawską i pozostałą częścią województwa mazowieckiego jest podobna i wynosi ok. 1/3 odchylenia standardowego (4,3–5,8 punktów na skali). W roku 2003 była z kolei nieco niższa (3,5 punktu), a w 2004 i 2011 nieco wyższa (odpowiednio 7,4 oraz 6,8 punktu). Obserwowane dla aglomeracji warszawskiej i pozostałej części województwa zmiany średniego PWE dla większości lat pokrywają się z trendami ogólnopolskimi. W tym wypadku wyjątki stanowią okresy 2002-2005 oraz 2012-2013. W wypadku aglomeracji warszawskiej obniżenie się średniego wyniku w roku 2003 było wyraźnie bardziej łagodne niż dla całej Polski, a już w 2004 roku zaobserwować można odbicie (warto odnotować, że różnica wyników pomiędzy rokiem 2003 i 2004 jest istotna statystycznie). Z kolei dla lat 2003-2006 na poziomie ogólnopolskim obserwujemy delikatny trend wzrostowy, podczas gdy średnie PWE w Warszawie i okolicach wykazują w tym czasie trend spadkowy. W pozostałej części województwa mazowieckiego obserwowane na poziomie ogólnopolskim w 2003 roku obniżenie średnich PWE opóźnione jest z kolei o rok. W roku 2012–2013 mamy natomiast do czynienia z wyraźną poprawą średniego wyniku na poziomie ogólnopolskim, podczas gdy wzrost średniego wyniku dla województwa mazowieckiego (zarówno aglomeracji warszawskiej, jak i pozostałej części województwa) jest wyraźnie niższy. Różnica ta spowodowała również, że średni wynik województwa mazowieckiego, z pominięciem aglomeracji warszawskiej, jest w tych latach, po raz pierwszy w całym obserwowanym okresie, istotnie statystycznie niższy od średniej ogólnopolskiej.

3.4.3. Analizy lokalne

Ostatnią omawianą grupę analiz stanowią analizy lokalne. Koncentrują się one na niewielkim obszarze, np. wszystkich szkołach w danej gminie lub nawet jednej, wybranej szkole. W wypadku tego typu analiz warto pamiętać o tym, że PWE dla stosunkowo niewielkich grup uczniów (jak np. uczniowie niedużej szkoły) obarczone są znacznym poziomem niepewności i wrażliwe na specyficzne czynniki lokalne, które w danym roku mogły wpływać na wyniki uczniów w danej zbiorowości. W związku z tym, dokonując analiz PWE na poziomie lokalnym, warto dysponować możliwie dużą liczbą informacji kontekstowych o badanej zbiorowości, a przy wyciąganiu wniosków pamiętać o kryterium istotności statystycznej (w wypadku porównywania średnich PWE jest to rozłączność przedziałów ufności). W odniesieniu do gimnazjów oraz szkół ponadgimnazjalnych warto także rozważyć, czy więcej informacji (z racji uwzględniania również wyników uczniów na wejściu do szkoły) nie przyniosą analizy EWD – patrz rozdział 4.

Analizy na poziomie szkół można łatwo przeprowadzić w serwisie internetowym pwe.ibe.edu.pl – z niego pochodzą zamieszczony poniżej wykres oraz przytaczane wartości liczbowe. Serwis umożliwia porównywanie ze sobą wyników różnych szkół, jak również odniesienie ich do wyników w gminie, powiecie, województwie lub całej Polsce. W poniższym przykładzie przeanalizowano wyniki przykładowej szkoły w części humanistycznej egzaminu gimnazjalnego na tle wyników gminy oraz ogólnopolskich. Wykorzystane zostały zarówno analizy wyników średnich (rysunek 3.45), jak i kwartyli (rysunek 3.46).

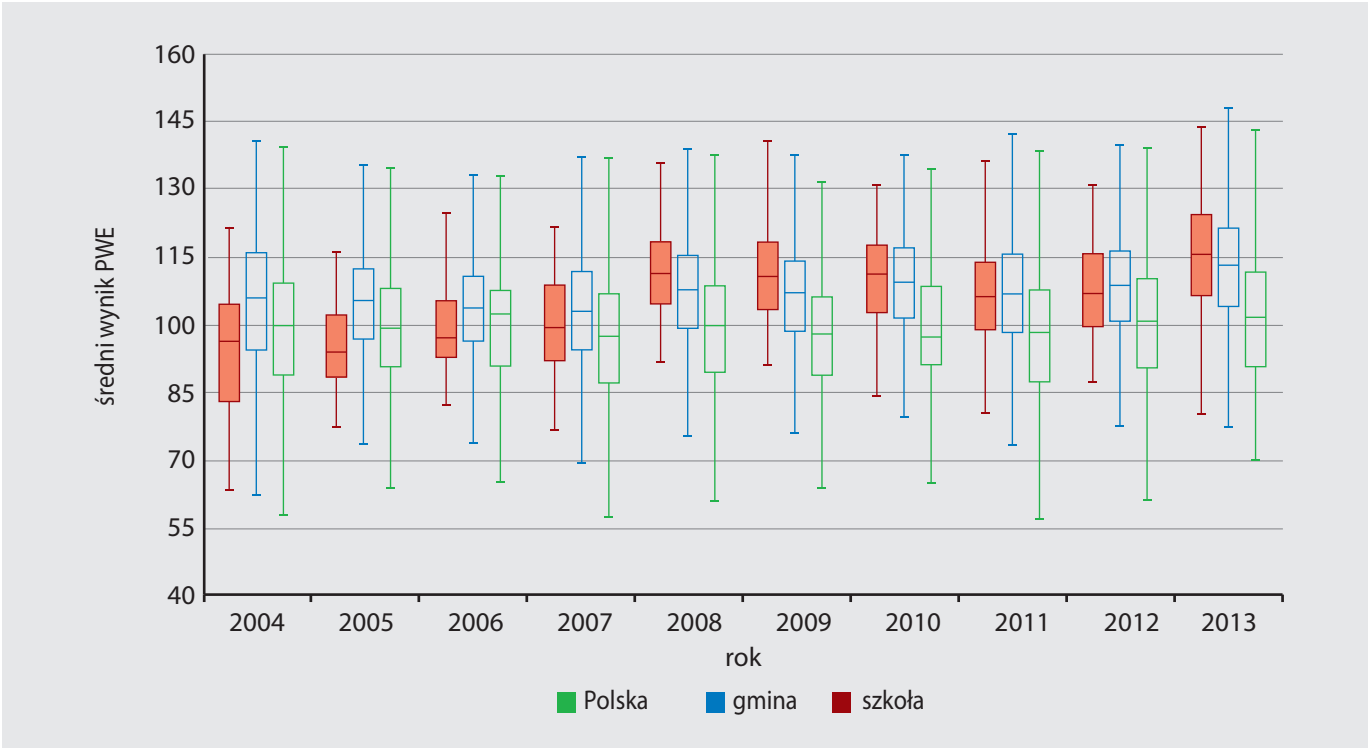
Rysunek 3.45. Średnie wyniki PWE w części humanistycznej egzaminu gimnazjalnego wraz z 95% przedziałami ufności dla przykładowej szkoły, na tle gminy oraz Polski



Obserwując powyższy wykres, łatwo zauważyć, że na przestrzeni lat szkoła bardzo wyraźnie poprawiła swoje średnie wyniki. O ile w latach 2004–2007 nie różniły się one w sposób statystycznie istotny od średniej ogólnopolskiej, jednocześnie do 2005 będąc istotnie statystycznie niższe od średniej dla gminy, o tyle w roku 2008 odnotowano znaczącą poprawę, która, mimo lokalnych (i nieistotnych

statystycznie) fluktuacji, utrzymuje się do końca obserwowanego okresu. Jednocześnie mniejsza szerokość przedziałów ufności dla średniej szkoły od roku 2008 sugeruje mniejsze zróżnicowanie wyników uczniów w ramach placówki i/lub zwiększenie się liczby nauczanych uczniów. Odwołanie się do danych na temat liczby uczniów potwierdza, że wzrosła ona na przestrzeni lat od 15–21 w latach 2004–2007 do 39–59 w latach 2011–2013. Warto zauważyć, że w związku z niżem demograficznym jest to zjawisko nietypowe. Może się ono wiązać np. z likwidacją innego gimnazjum w okolicy i/lub znaczną poprawą jakości kształcenia, która przyciągnęła uczniów do analizowanej szkoły. Przy okazji odnotować można istotnie wyższe od ogólnopolskich średnie wyniki w gminie, co można tłumaczyć tym, że jest to gmina podwarszawska.

Rysunek 3.46. Parametry rozkładu PWE (kwartyle, najniższy wynik nieodstający, najwyższy wynik nieodstający – objaśnienie w tekście) w części humanistycznej egzaminu gimnazjalnego dla przykładowej szkoły, na tle gminy oraz Polski

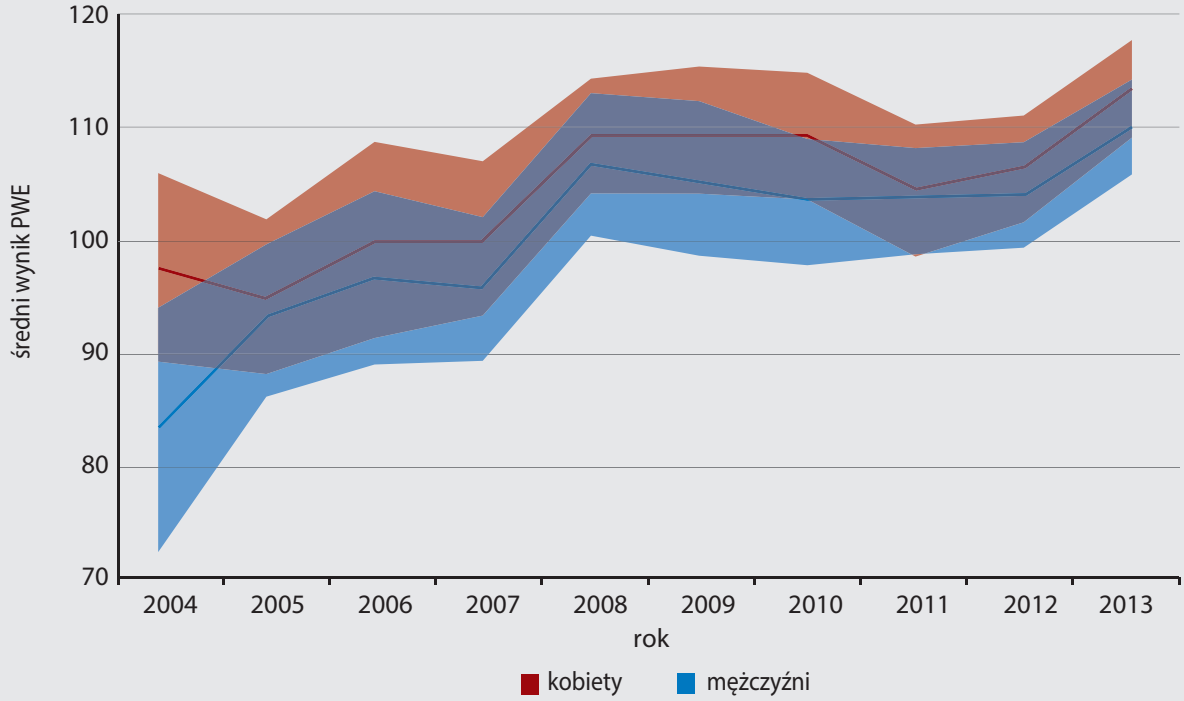


Kolejnych informacji dostarcza analiza parametrów rozkładów wyników PWE, takich jak kwartyle czy najmniejszy i największy wynik nieodstający. 1., 2. i 3. kwartyl są to najlepsze wyniki uzyskiwane, odpowiednio, w grupie 25%, 50% i 75% najsłabszych uczniów. 2. kwartyl nazywany jest także medianą. Za najmniejszy nieodstający wynik przyjmuje się najniższy wynik nie mniejszy jednak od wartości 1. kwartyla pomniejszonej o 1,5-krotność różnicy pomiędzy 3. i 1. kwartylem. Analogicznie największy nieodstający wynik to najwyższy wynik, nie większy jednak od wartości 3. kwartyla zwiększonego o 1,5-krotność różnicy pomiędzy 3. i 1. kwartylem.

Widzimy, że zróżnicowanie pomiędzy najlepszym i najsłabszym uczniem w szkole jest dla większości lat znacznie niższe niż dla gminy (a także kraju). W interesujący sposób wygląda rozkład wyników w roku 2008 – okazuje się, że wyższy w stosunku do gminy średni wynik szkoły brał się przede wszystkim z podniesienia wyników uczniów słabszych. Najwyższy nieodstający wynik w szkole jest nawet nieco niższy od analogicznego dla gminy, 3. kwartyl i mediana szkoły wypadły już trochę powyżej analogicznych wyników dla gminy, natomiast 1. kwartyl wyników szkoły jest tylko nieznacznie niższy od mediany dla gminy, a najniższy nieodstający wynik szkoły o blisko 25 punktów (ponad 1,5 odchylenia standardowego) przewyższa najniższy nieodstający wynik w gminie. Jednocześnie to właśnie w 2008 roku zaobserwowano w szkole znaczące podniesienie się średniego wyniku

egzaminacyjnego. Być może dobra znajomość lokalnego kontekstu pracy szkoły, którą niestety w tych analizach nie dysponujemy, pozwoliłaby powiązać ze sobą te obserwacje.
Możliwe jest również przeprowadzenie bardziej szczegółowych analiz, uwzględniających znane przy obliczaniu PWE cechy uczniów, jak rok urodzenia czy płeć. W tym celu niezbędne jest posłużenie się pakietem ZPD dla R[43]. Jako przykład takiej analizy prześledzimy wyniki rozważanej przed chwilą szkoły w podziale na płeć uczniów (rysunek 3.47).

Rysunek 3.47. Średnie PWE wraz z 95% przedziałami ufności dla części humanistycznej egzaminu gimnazjalnego w podziale na płeć dla przykładowej szkoły



Na wykresie widać, że średni wynik chłopców i dziewcząt nie różni się w sposób statystycznie istotny, chociaż dziewczynki osiągają systematycznie nieco lepsze wyniki od chłopców. Możemy również zaobserwować, że różnica pomiędzy chłopcami a dziewczętami była największa w 2004 roku, później natomiast uległa zmniejszeniu, zaś omawiana już wcześniej znaczna poprawa średniego wyniku szkoły w roku 2008 wzięła się między innymi ze zmniejszenia różnicy pomiędzy średnimi wynikami chłopców i dziewcząt.

3.5. Podsumowanie

Wprowadzenie do systemu egzaminacyjnego mechanizmów zapewniających porównywalność wyników kolejnych edycji egzaminów jest ważne zarówno dla funkcji monitorującej jak i selekcyjnej (szczególnie na progu szkoły wyższej). Zwróciła na to również uwagę Najwyższa Izba Kontroli, która w raporcie System egzaminów zewnętrznych wśród rekomendacji wnioskuje wprowadzenie na etapie tworzenia testów egzaminacyjnych metod, które nadadzą wynikom egzaminacyjnym walor międzyedycyjnej porównywalności.

[43] Patrz http://zpd.ibe.edu.pl/doku.php?id=r_zpd oraz https://github.com/zozlak/ZPD

Choć porównywalność wyników egzaminów miedzy latami jest najważniejsza w procesach selekcyjnych na progu szkoły wyższej, to możliwość analizy trendów może być wykorzystana w ewaluacji pracy szkół oraz monitowaniu realizacji zadań edukacyjnych w gminach, powiatach lub województwach. Co najważniejsze, inaczej niż inne dostępne wskaźniki (choćby EWD), porównywalne między latami wyniki egzaminacyjne mogą służyć do oceny efektywności nauczania całego systemu oświatowego.
Warto rozważyć wdrożenie jednego z trzech opisanych poniżej rozwiązań. Warianty te w mniejszym lub większym stopniu wiążą się z koniecznością zmian w systemie egzaminów zewnętrznych i związanych z tym zmian w prawie. Trzy warianty są przedstawione w kolejności pożądanych własności metodologicznych.

Wariant A – włączenie do egzaminu niejawnych zadań kotwiczących.
Z metodologicznego i pomiarowego punktu widzenia jest to rozwiązanie najlepsze, dające najbardziej precyzyjne oszacowanie funkcji zrównującej i najmniej narażone na potencjalne czynniki zakłócające. Rozwiązanie to wymaga jednak istotnych zmian w sposobie przeprowadzania egzaminów.

Wariant B – jednoczesna standaryzacja zadań z wielu edycji egzaminu.
Rozwiązanie jest mniej efektywne ze względu na wyznaczanie funkcji zrównującej. Wyniki mogą być obarczone wpływem niższego poziomu motywacji testowej uczniów na sesji standaryzacyjnej w porównaniu do sesji egzaminacyjnej.

Wariant C – dodatkowe badania na reprezentatywnej próbie.
Jest to rozwiązanie nie wprowadzające żadnych zmian w systemie egzaminacyjnym, przez co byłoby najprostsze we wdrożeniu. Jednak wariant C rodzi największą liczbą potencjalnych problemów prowadzących do zakłócenia procesu zrównywania, ponadto odznacza się najniższą efektywnością. Ze względu na znaczną niepewność statystyczną wyników, jakie przynosi ta metoda, jej przydatność do zrównywania wyników pojedynczych uczniów jest ograniczona. Zaletą tego rozwiązania jest możliwość śledzenia zmian w poziomie umiejętności populacji między latami oraz zbieranie informacji o stabilności trudności testów egzaminacyjnych miedzy latami.

Prezentowane powyżej rozwiązania różnią się trudnością wdrożenia. Wariant C nie ingeruje w funkcjonowanie systemu egzaminów zewnętrznych, jednak cechuje się bardzo dużym błędem statystycznym, ograniczeniem do zrównywania jedynie podstawowych części egzaminu oraz uzależnieniem pewności wyników od jakości (w tym wielkości) doboru próby. Wariant B, natomiast, nie kontroluje poziomu motywacji uczniów, wymaga stworzenia znacznej liczby zadań do standaryzacji i wiąże się z ryzykiem ujawnienia treści zadań przed właściwym egzaminem. Zalety wariantu A są bezsprzeczne: duża moc wnioskowania statystycznego, możliwość zrównywania wyników dla wszystkich arkuszy egzaminacyjnych, brak obciążenia wyników zróżnicowanym poziomem motywacji uczniów oraz zdecydowane niższe koszty niż dla wariantu C lub B. Jednakże jest to wariant najtrudniejszy od strony legislacyjnej - wymaga zmian prawnych i organizacyjnych, które pozwoliłyby na skuteczne utajnienie kotwiczącej części egzaminu. Mamy świadomość, że wariant ten może spotkać się z obawami opinii publicznej.
W wypadku testów egzaminacyjnych tworzonych dla niewielkich populacji zdających należałoby jednak ocenić, czy nakłady niezbędne do zapewnienia porównywalności wyników, są adekwatne do korzyści wdrożenia takich rozwiązań. Wydaje się, że w przypadku egzaminów rozwiązywanych przez nieliczne grupy uczniów zapewnienie porównywalności wyników może być uzyskane poprzez odwołanie się do opinii ekspertów na temat relatywnej trudności testów.
Szczegóły przedstawionych rozwiązań czytelnik może odnaleźć w Monografii Porównywalne wyniki egzaminacyjne (Szaleniec i inni 2015).

Bibliografia

Bogdanowicz, M. (1994). *O dysleksji, czyli specyficznych trudnościach w czytaniu i pisaniu*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.

Boughton, K. A. i Yamamoto, K. (2007). A hybrid model for test speededness. w: M. von Davier i C.H. Carstensen. *Multivariate and mixture distribution Rasch models: Extensions and applications* (s. 147–156). New York, NY: Springer.

CKE (2009). *Informator z aneksem dla uczniów ze specyficznymi trudnościami w uczeniu się. Sprawdzian w klasie szóstej szkoły podstawowej przeprowadzany od roku szkolnego 2009/2010*. Warszawa: Centralna Komisja Egzaminacyjna.

CKE (2010). *Osiągnięcia uczniów kończących gimnazjum w roku 2010*. Warszawa: Centralna Komisja Egzaminacyjna.

CKE (2011). *Osiągnięcia uczniów kończących gimnazjum w roku 2011*. Warszawa: Centralna Komisja Egzaminacyjna.

CKE (2012). *Osiągnięcia uczniów kończących gimnazjum w roku 2012*. Warszawa: Centralna Komisja Egzaminacyjna.

CKE (2013). *Osiągnięcia uczniów kończących gimnazjum w roku 2012*. Warszawa: Centralna Komisja Egzaminacyjna.

Dolata, R. (2008). *Szkoła – segregacje – nierówności*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Dolata, R. (2010). Cicha rewolucja w polskiej oświacie – proces różnicowania się gimnazjów w dużych miastach. *Edukacja. Studia, Badania, Innowacje, 109*(1), 51–60.

Dolata, R. (2012). *Międzyszkolne zróżnicowanie wyników nauczania na poziomie szkoły podstawowej i gimnazjum*. Warszawa: Instytut Badań Edukacyjnych.

Dolata, R., Hawrot, A., Humenny, G., Jasińska-Maciążek, A., Koniewski, M. i Majkut, P. (2014). *Kontekstowy model oceny efektywności nauczania po pierwszym etapie edukacyjnym*. Warszawa: Instytut Badań Edukacyjnych.

Dolata, R., Jasińska, A. i Modzelewski, M. (2012). Wykorzystanie krajowych egzaminów jako instrumentu polityki oświatowej na przykładzie procesu różnicowania się gimnazjów w dużych miastach. *Polityka Społeczna, 1*, 41–46.

Donlon, T. F. (1984). *The college board technical handbook for the Scholastic Aptitude Test and Achievement Tests*. College Board.

EACEA (2010). *Różnice w wynikach nauczania a płeć uczniów. Obecna sytuacja i działania podejmowane w Europie*. Warszawa: Fundacja Rozwoju Systemu Edukacji.

GUS (2013). *Oświata i wychowanie w roku szkolnym 2012/2013*. Warszawa: Zakład Wydawnictw Statystycznych.

Hanushek, E. A i Kim, D. (1995). *Schooling, labor force quality, and economic growth*. National Bureau of Economic Research Working Paper No. 5399. Pobrano z: http://www.nber.org/papers/w5399

Herbst, M. (2004). Zróżnicowanie jakości kapitału ludzkiego w Polsce. Od czego zależą wyniki edukacyjne? *Studia Regionalne i lokalne*, *3* (17), 93–104.

Herbst, M. i Herczyński, J. (2005). *School Choice and Student Achievement. Evidence from Poland*. MPRA Paper No. 6138. Pobrano z: http://mpra.ub.uni-muenchen.de/6138/

Herczyński, J. i Herbst, M. (2002). *Pierwsza odsłona: społeczne i terytorialne zróżnicowanie wyników sprawdzianu szóstoklasistów i egzaminu gimnazjalnego przeprowadzonych wiosną 2002 roku. Raport przygotowany na zlecenie Fundacji Klub Obywatelski*. Warszawa: Fundacja „Klub Obywatelski".

Herczyński, J. i Sobotka, A. (2014). *Diagnoza zmian w sieci szkół podstawowych i gimnazjów 2007–2012*. Warszawa: Instytut Badań Edukacyjnych.

Hyde, J. S. i Kling, K. C. (2001). Women, motivation, and achievement. *Psychology of Women Quarterly, 25,* 364-378.

Jasińska, A. i Modzelewski, M. (2013). Międzyszkolne zróżnicowanie wyników nauczania po pierwszym etapie kształcenia. w: B. Niemierko i M. K. Szmigel (red.). *Polska edukacja w świetle diagnoz prowadzonych z różnych perspektyw badawczych* (s. 165–178). Kraków: gRUPA TOMAMI.

Jung-Miklaszewska, J. i Rusakowska, D. (1995). *Szkoły społeczne in statu nascendi*. Warszawa: Instytut Badań Edukacyjnych.

Konarzewski, K. (1996). *Problemy i schematy. Pierwszy rok nauki szkolnej dziecka.* Warszawa: Wydawnictwo „Żak".

Konarzewski, K. (2012). *TIMSS i PIRLS 2011. Osiągnięcia szkolne polskich trzecioklasistów w perspektywie międzynarodowej*. Warszawa: Centralna Komisja Egzaminacyjna.

Kondratek, B., i Pokropek, A. (2013). IRT i pomiar edukacyjny. *Edukacja, 4 (124)*.

Kurek, S. (2010). Przestrzenne zróżnicowanie poziomu rozwoju regionalnego w Unii Europejskiej w świetle wybranych mierników. *Prace Komisji Geografii Przemysłu Polskiego Towarzystwa Geograficznego*, *16*, 87–104.

Najwyższa Izba Kontroli (2015). *System egzaminów zewnętrznych w oświacie.* Pobrano z https://www.nik.gov.pl/plik/id,8629,vp,10737.pdf

Nijakowska, J. (2009). Hipoteza różnic w kodowaniu językowym – próba wyjaśnienia trudności w uczeniu się języków obcych. w: M. Pawlak, M. Derenowski i B. Wolski (red.). *Problemy współczesnej dydaktyki języków obcych* (s. 41–55). Kalisz: Wydział Pedagogiczno-Artystyczny UAM w Kaliszu.

Noftle, E. E. i Robins, R. W. (2007). Personality Predictors of Academic Outcomes: Big Five Correlates of GPA and SAT Scores. *Journal of Personality and Social Psychology, 93*(1), 116–130.

OECD (2009). *Equally prepared for life? How 15-year-old boys and girls perform in school.* Pobrano z http://www.stemequitypipeline.org/_documents/OECD%20%282009%29%20gender%20diff%20similarities%20in%2015-year-old.pdf

OECD (2014). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science* (Volume I, Revised edition, February 2014), PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201118-en

OECD (2014). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science* (Volume I, Revised edition, February 2014). PISA, OECD Publishing.

Perkins, R., Kleiner, B., Roey, S. i Brown, J. (2004). *The High School Transcript Study: A Decade of Change in Curricula and Achievement, 1990-2000*. Washington, DC: National Center for Education Statistics.

Piekarczyk, M. (2014). *Jak Polacy postrzegają szkoły publiczne i niepubliczne: preferencje dotyczące szkolnictwa w Polsce*. Warszawa: Centrum Badań nad Uprzedzeniami.

Pokropek, A. (2011). Missing by design: planned missing-data designs in social science. *ASK. Research & Methods*, 20, 81–105.

Pokropek, A. i Kondratek, B. (2012). Zrównywanie wyników testowania. Definicje i przykłady zastosowania. *Edukacja*, *4 (120)*, 52–71.

Putkiewicz, E. i Wiłkomirska, A. (2004). *Szkoły publiczne i niepubliczne. Porównanie środowisk edukacyjnych*. Warszawa: Instytut Spraw Publicznych.

Skórska, P. i Świst, K. (2014). Wielkość efektu płci w wewnątrzszkolnych i zewnątrzszkolnych wskaźnikach osiągnięć ucznia. W B. Niemierko i M.K. Szmigel (red.). Diagnozy edukacyjne: dorobek i nowe zadania. (s. 89–103). Kraków: gRUPA TOMAMI.

Spionek, H. (1965). *Zaburzenia psychoruchowego rozwoju dziecka*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.

Szaleniec, H., Grudniewska, M., Kondratek, B., Kulon, F. i Pokropek, A. (2012). Wyniki egzaminu gimnazjalnego 2002–2010 na wspólnej skali. *Edukacja, 119*(3), 9–30.

Szaleniec, H., Grudniewska, M., Kondratek, B., Kulon, F., Pokropek, A., Stożek, E. i Żółtak, M. (2013). *Analiza porównawcza wyników egzaminów zewnętrznych – sprawdzian w szóstej klasie szkoły podstawowej i egzamin gimnazjalny.* Warszawa: Instytut Badań Edukacyjnych.

von Schrader, S. i Ansley, T. (2006). Sex differences in the tendency to omit items on multiple-choice tests: 1980–2000. *Applied Measurement in Education, 19*(1), 41–65.

Wejner, T. (2009). Ocenianie uczniów dyslektycznych w innych krajach. w: B. Niemierko i M. K. Szmigel (red.). *Badania międzynarodowe i wzory zagraniczne w diagnostyce edukacyjnej. r.* (s. 331-341). Kraków: gRUPA TOMAMI.

Willingham, W.W. i Cole, N.S. (1997). *Gender and fair assessment.* Mahwah NJ: Lawrence Erlbaum.

World Economic Forum (2013). The Global Gender Gap Report (2013). Pobrano z http://www3.weforum.org/docs/WEF_GenderGap_Report_2013.pdf

Zahorska-Bugaj, M. (1994). Szkoła prywatna czy państwowa. W: R. Siemieńska (red.), *Szkoły niepaństwowe w systemie edukacji w Polsce. Kwartalnik Pedagogiczny, 1–2*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Zasacka Z. (2014). *Czytelnictwo dzieci i młodzieży*. Warszawa: Instytut Badań Edukacyjnych.

Zawadzka, E. (2010) *Nauczyciel języków obcych i jego niepełnosprawni uczniowie (z zaburzeniami i dysfunkcjami)*. Kraków: Wydawnictwo Impuls.

Żółtak, T. (2011). Znaczenie informacji o średnim wyniku uczniów na wejściu dla własności jednorocznych wskaźników EWD gimnazjów. W B. Niemierko i M.K. Szmigel (red.). *Ewaluacja w edukacji: koncepcje, metody, perspektywy*. (s. 505–523). Kraków: gRUPA TOMAMI.

4. Metoda edukacyjnej wartości dodanej w Polsce

Roman Dolata, Anna Hawrot, Grzegorz Humenny, Aleksandra Jasińska-Maciążek, Anna Rappe, Ewa Stożek, Tomasz Żółtak

4.1. Wprowadzenie

Decentralizacja i zwiększanie autonomii szkół zwiększa zapotrzebowanie na informacje o jakości pracy poszczególnych placówek. Tego rodzaju informacje mogą być wykorzystywane zarówno do zewnętrznej oceny pracy szkoły, jak i do doskonalenia jakości nauczania. Egzaminy i informacje z testów osiągnięć szkolnych przeprowadzonych w sposób umożliwiający dokonywanie porównań między szkołami są podstawowym źródłem informacji, które można do tego celu wykorzystać. Jednak pełne wykorzystanie informacji płynących z systemu egzaminacyjnego wymaga odpowiednich narzędzi analitycznych. W poprzednim rozdziale opisane zostały porównywalne między latami wyniki egzaminacyjne, innym ważnym instrumentem analitycznym jest metoda edukacyjnej wartości dodanej (EWD).

Wyniki egzaminacyjne uczniów zależą od wielu czynników niezależnych od szkoły, w której uczyli się na danym etapie edukacyjnym, przede wszystkim od uprzednich osiągnięć. Metoda EWD to modele statystyczne pozwalające na wnioskowanie o efektywności szkoły, czyli o wkładzie szkoły w wyniki nauczania. By można stosować metodę EWD, potrzebujemy wyników przynajmniej dwóch pomiarów osiągnięć szkolnych: na początku nauki w danej szkole i na jej zakończenie. Istotą metody EWD jest odejście od patrzenia na szkoły jedynie przez pryzmat wyniku osiąganego przez uczniów na koniec nauki. Ponieważ wyniki nauczania zależą w znaczącej mierze od czynników, które są poza kontrolą szkoły, używanie rezultatów egzaminów końcowych jako miary efektywności musi prowadzić do niesprawiedliwych ocen. EWD jest zdecydowanie bardziej wartościową metodą oceny efektywności nauczania. Uwzględnienie wyników uczniów na początku rozpoczęcia nauki w danej szkole pozwala wskazać szkoły, które mimo relatywnie słabych wyników końcowych mogą pochwalić się wysoką efektywnością nauczania i odwrotnie, szkoły będące wysoko w rankingu wyników testów końcowych, ale niewiele przyczyniające się do postępów uczniów. Należy przy tym zauważyć, że wyniki „na wejściu" informują nie tylko o poziomie osiągnięć szkolnych ucznia na progu szkoły, ale są też „nośnikiem" wiedzy o środowisku rodzinnym dziecka, o jego poziomie zdolności i motywacji szkolnej. Dzieje się tak dlatego, że wyniki egzaminu „na wejściu" są uwarunkowane podobnymi czynnikami, co rezultaty egzaminu końcowego.

Pojęcie edukacyjnej wartości dodanej pojawiło się po raz pierwszy w połowie lat 70. Od początku lat 90. intensywne prace nad metodą edukacyjnej wartości dodanej prowadzone są głównie w Stanach Zjednoczonych i Wielkiej Brytanii, a od początku tego wieku w Australii. Niektórzy badacze uważają pojęcie edukacyjnej wartości dodanej za najważniejsze narzędzie analityczne, jakie w naukach pedagogicznych pojawiło się pod koniec XX wieku (Schagen i Hutchison, 2003). Nie oznacza to oczywiście, że metoda ta jest idealna i nie jest pozbawiona wad.

Na świecie istnieje wiele wariantów metody EWD, co częściowo odzwierciedla zróżnicowane możliwości korzystania z danych, a częściowo różne cele stosowania metody. Na przykład możliwości modelowania zmian umiejętności uczniów są większe w Stanach Zjednoczonych niż w Polsce, ze względu na fakt, że testy przeprowadzane są rokrocznie przez 8 lat nauki. Z drugiej strony, metody szacowania edukacyjnej wartości dodanej powinny być podporządkowane temu, czemu mają służyć i kto będzie odbiorcą informacji. Wskaźniki EWD mogą służyć ocenie efektywności nauczania na potrzeby ewaluacji wewnątrzszkolnej, mogą być wykorzystywane przez nadzór pedagogiczny i samorządy w ewaluacji zewnętrznej czy w końcu w ocenie pracy szkoły na potrzeby rodziców

i uczniów wybierających szkołę czy oceniających szkołę, z usług której korzystają (Raudenbush i Willms, 1995). Metoda może być stosowana w analizie efektywności nauczania szkół lub poszczególnych nauczycieli.

W wypadku szacowania EWD na potrzeby profesjonalnych procesów ewaluacji ważne jest uwzględnienie różnic nie tylko w poziomie uprzednich osiągnięć, ale też, w miarę możliwości i udokumentowanej badawczo potrzeby, innych czynników wpływających na postęp edukacyjny. Choć badania pokazują, że wpływ bazy materialnej jest znikomy (por. Hanushek, 2003), to już np. wpływ składu społecznego szkoły może być znaczący (Markman i in., 2003).

Innego sposobu szacowania EWD wymagałyby natomiast wskaźniki adresowane do rodziców i uczniów. W tym wypadku wartość dodana powinna być liczona dla całej placówki, bez wyłączania wpływu zasobów szkolnych (zarówno materialnych, jak i np. składu społecznego szkoły), bowiem z punktu widzenia rodziców interesujące jest całościowe oddziaływanie danej placówki na postęp w osiągnięciach szkolnych dziecka. W wypadku wskaźników EWD dla rodziców jest też bardzo ważne, czy efektywność nauczania w danej placówce zależy od wstępnego poziomu osiągnięć szkolnych. Szkoła może bardzo efektywnie nauczać uczniów dobrych, a być nieefektywna w pracy z uczniami słabszymi, albo odwrotnie (por. Meyer, 1997).

Z faktu, że EWD jest metodą statystyczną, wynika jej siła analityczna, ale jest to też powód jej ważnego ograniczenia. Metody statystyczne mogą być stosowane tylko do zbiorowości. Można się spierać, czy minimalna liczebność takiej zbiorowości to 10 czy 30 elementów, ale z pewnością dla bardzo małych grup metoda EWD nie może być stosowana. Jeżeli na przykład w danej szkole do egzaminu przystąpiło czworo uczniów, żadna statystyka nie pozwoli na dokonywanie na tej podstawie jakichkolwiek statystycznych uogólnień. Jest to problem w Polsce dość istotny, bo spora grupa placówek to szkoły bardzo małe. Można próbować radzić sobie częściowo z tym ograniczeniem poprzez gromadzenie danych z kolejnych lat i liczenie wskaźników wieloletnich.

Stosowanie modeli EWD do polskich danych egzaminacyjnych napotyka na inne jeszcze ograniczenie. Pomiar osiągnięć szkolnych na kolejnych etapach edukacyjnych i w kolejnych edycjach egzaminów na niezależnie tworzonych skalach – np. wynik na sprawdzianie i egzaminie gimnazjalnym z matematyki nie jest wyrażany na jednej skali; wynik z danego roku nie jest wyrażany na tej samej skali co w innych latach – sprawia, że miary EWD mogą mieć jedynie relatywny sens. Dodatni wskaźnik EWD szkoły mówi, że efektywność nauczania jest w niej w porównaniu ze szkołami w kraju w danym roku ponadprzeciętna, ujemna, lub poniżej przeciętnej. Gdy dysponujemy wynikami testowania porównywalnymi tylko w obrębie danego egzaminu i danego roku, to nie sposób odpowiedzieć na pytanie, czy EWD w skali kraju w kolejnych latach rośnie, czy maleje. Należy jednak wyraźnie powiedzieć, że na potrzeby procesu ewaluacji pracy szkół relatywne miary są bardzo przydatne i nie jest to mankament znacząco zmniejszający użyteczność wskaźników EWD.

Metoda EWD może być rozpatrywana w ogólnym kontekście polityki poprawiania jakości oświaty oraz w bardziej swoistym otoczeniu problemów pojawiających się wszędzie tam, gdzie szkoły zaczynają ze sobą konkurować o uczniów. Badania nad funkcjonowaniem systemów oświatowych, w których szkoły rywalizują o uczniów i są finansowane zależnie od ich liczby, wskazują, że poza ewentualnymi korzyściami płynącymi z tych rozwiązań pojawiają się zagrożenia. Najczęściej wskazuje się na silne różnicowanie systemu szkół. Różnice między szkołami zwiększają się. Najsłabsze szkoły nie są eliminowane, ale z różnych powodów z coraz mniejszą liczbą uczniów trwają na rynku. Najlepsze placówki, wykorzystując mechanizm zwany „spijaniem śmietanki", umacniają swoje pozycje. Różnicowanie to zagraża ważnemu celowi polityki oświatowej, jakim jest równość szans edukacyjnych. Z procesem różnicowania szkół mamy też do czynienia w Polsce (por. Dolata, 2008; Dolata, Jasińska i Modzelewski, 2012; Zawistowska, 2012). Ocena szkoły na podstawie wskaźnika EWD może zmniejszyć nacisk na selekcję na wejściu do szkoły, zaś ocena nauczycieli według tej miary może podnieść atrakcyjność pracy w szkołach działających w mniej korzystnych warunkach społecznych. Nie ma dowodów na to, że tak się dzieje, ale jest szansa, która może się zrealizować. Oczywiście EWD z pewnością nie zahamuje segregacji społecznych i ekonomicznych w skali makro, ale w dłuższej

perspektywie, jeżeli będzie elementem szerszej polityki edukacyjnej, w której równość szans będzie brana na serio pod uwagę, może przyczynić się do rewitalizacji szkół pracujących w środowiskach społecznie, kulturowo i ekonomicznie upośledzonych.

Rozwój metody edukacyjnej wartości dodanej w Polsce

W Polsce wykorzystanie modeli EWD w analizie wyników egzaminacyjnych stało się możliwe w 2005 roku. W tym właśnie roku gimnazja ukończył pierwszy rocznik absolwentów, dla których był dostępny zarówno wynik sprawdzianu w klasie VI, jak i egzaminu gimnazjalnego. W 2006 roku ukazały się pierwsze polskie publikacje pokazujące możliwości metody EWD w analizie wyników gimnazjalnych (Dolata, 2006a, 2006b). W tekstach tych wykorzystywano proste modele regresji, w których zmienną wyjaśnianą był wynik z danej części egzaminu gimnazjalnego, a zmienną wyjaśniającą wynik na sprawdzianie w klasie VI szkoły podstawowej. Wskaźniki EWD dla gimnazjów czy oddziałów klasowych szacowane były punktowo, jako średnie reszt z równania regresji. Ze względu na brak scalonych krajowych baz danych z połączonymi wynikami sprawdzianu i egzaminu gimnazjalnego analizy prowadzono na poziomie lokalnym (np. gimnazja w danym mieście). Najważniejszym osiągnięciem tego okresu było uświadomienie decydentom potencjału analitycznego tkwiącego w połączonych, podłużnie analizowanych wynikach egzaminacyjnych. Dzięki temu powołano przy Centralnej Komisji Egzaminacyjnej zespół badawczy ds. rozwoju metody EWD (pierwszy projekt EFS, 2005–2006).

Pierwszy, skromny budżetowo projekt, pozwolił rozpocząć prace nad rozwojem metody EWD na potrzeby polskiego systemu egzaminacyjnego. Pomoc okręgowych komisji egzaminacyjnych umożliwiła scalenie wyników sprawdzianu i egzaminu gimnazjalnego w jedną zintegrowaną bazę danych. Dzięki temu mogła powstać pierwsza edycja Kalkulatora EWD dla gimnazjów. Był to arkusz kalkulacyjny z zaimplementowanym algorytmem wyliczania wskaźników EWD dla szkół lub innych grup uczniów (oddziały klasowe, chłopcy, dziewczęta itp.). Algorytm wykorzystywał oszacowane dla danych krajowych wartości przewidywane dla uczniów o danym wyniku na sprawdzianie. W estymacji wartości przewidywanych wykorzystywano dodatkowo zmienne kontrolne dostępne w systemie egzaminacyjnym – informacje o płci i dysleksji. Wskaźniki EWD szacowane były punktowo i przedziałowo. Szkoły mogły bezpłatnie pobierać ze strony CKE Kalkulator EWD i po wprowadzeniu danych swoich absolwentów przeprowadzać analizy na potrzeby ewaluacji wewnątrzszkolnej. W ramach projektu odbyło się pierwsze, kaskadowe szkolnenie dyrektorów gimnazjów w wykorzystaniu metody EWD w ewaluacji wewnętrznej. W 2007 roku powstała kolejna edycja Kalkulatora EWD, równolegle testowano bardziej złożone statystycznie modele szacowania EWD. W roku tym uruchomiono też przy CKE kolejny zespół badawczy ds. rozwoju metody EWD (EFS, 2007–2013). Dysponował on już znacznym budżetem i możliwościami działania.

W 2008 roku powstała strona www.ewd.edu.pl. Z czasem strona przerodziła się w portal wiedzy nt. wykorzystania wyników egzaminacyjnych do ewaluacji pracy szkół. W tym samym roku odbyło się kolejne masywne szkolenie potencjalnych użytkowników metody EWD pod nazwą Wiosenna Szkoła EWD. Szkoły EWD stały się w następnych latach główną formą upowszechniania wiedzy o metodzie i możliwościach jej wykorzystania. W 2008 roku ukazała się też przygotowana przez członków zespołu pierwsza polska monografia naukowa poświęcona metodzie *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania* (Dolata, 2007). W 2008 roku miało miejsce jeszcze jedno ważne wydarzenie – opublikowano raport OECD opracowany przez międzynarodową grupę ekspertów poświęcony metodzie EWD *Measuring Improvements in Learning Outcomes* (2008). Jednym z ekspertów był Maciej Jakubowski, członek polskiego zespołu EWD. Następny rok zaowocował opracowaniem całkowicie nowego Kalkulatora EWD. Pod nazwą Kalkulator EWD Plus kryła się samodzielna aplikacja komputerowa do analiz wyników egzaminacyjnych dla gimnazjów. W tym samym roku po raz pierwszy zespół w trybie pilotażowym upublicznił na stronie projektu trzyletnie wskaźniki egzaminacyjne dla gimnazjów. Wskaźniki w nowatorski, graficzny sposób pokazywały pozycję danego gimnazjum w dwuwymiarowej przestrzeni – jedna oś to średni wynik egzaminu, druga

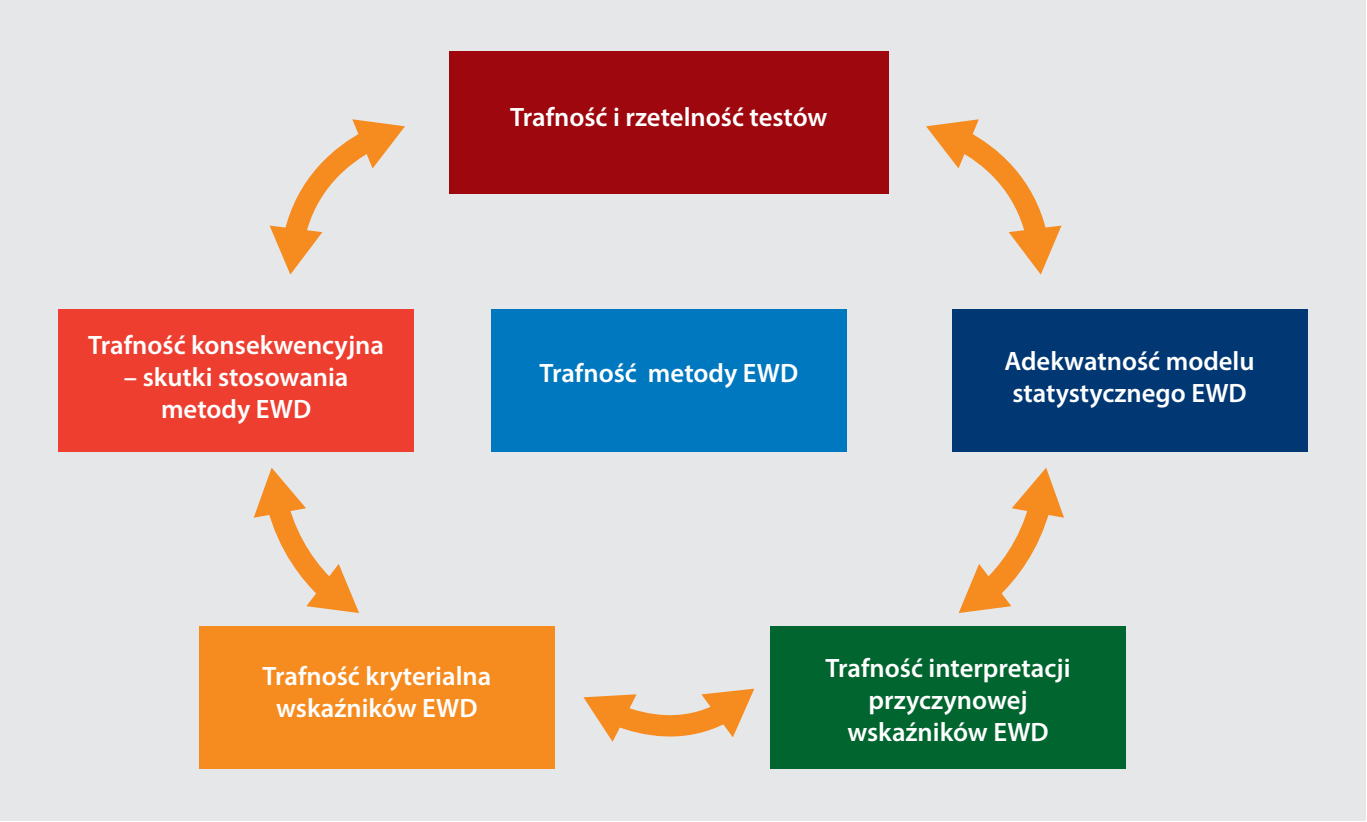
to miara EWD. O ile Kalkulator EWD z założenia był przeznaczony do wykorzystania w ewaluacji wewnątrzszkolnej efektywności nauczania, to publikowane w Internecie, powszechnie dostępne wskaźniki trzyletnie miały służyć szerszej grupie odbiorców. Założonymi użytkownikami tych wskaźników były oczywiście szkoły, ale również organy nadzoru pedagogicznego i organy prowadzące.
W końcu 2009 roku rozpoczęła się też realizacja trzech badań podłużnych mających dostarczyć informacji potrzebnych do rozwoju metody EWD, w tym do oceny jej trafności. Dwa z nich były – w szkołach podstawowych i gimnazjach - realizowane przez Zespół EWD, natomiast badanie w szkołach ponadgimnazjalnych na zlecenie CKE prowadził IFiS PAN (informacje o badaniach patrz: www.ewd.edu.pl oraz www.ifispan.waw.pl). Pokłosiem tych realizowanych w kolejnych latach studiów empirycznych są dostępne na stronie projektu bazy danych wraz z pełną dokumentacją badawczą oraz trzy ważne monografie: *Ścieżki rozwoju edukacyjnego młodzieży – szkoły pogimnazjalne* (Karwowski 2013), *Trafność metody edukacyjnej wartości dodanej dla gimnazjów* (Dolata i in., 2013) oraz *Kontekstowy model oceny efektywności nauczania po pierwszym etapie edukacyjnym* (Dolata i in., 2014). Wnioski z tych prac zostały wykorzystane w niniejszym rozdziale.
W 2010 roku upubliczniono po raz pierwszy wskaźniki egzaminacyjne dla liceów ogólnokształcących i techników w zakresie matematyki i języka polskiego. Graficzna postać wskaźników była analogiczna jak w wypadku gimnazjów, jednak proces ich wyliczania był bardziej złożony. Na podstawie wykonania zadań maturalnych na poziomie podstawowym i rozszerzonym szacowano za pomocą metody IRT poziom umiejętności matematycznych i polonistycznych maturzystów, następnie, analogicznie jak w wypadku wskaźników gimnazjalnych, szacowano średni wynik i EWD szkoły.
W 2011 roku udoskonalono modele EWD dla gimnazjów, uporządkowano bazę szkół, dzięki czemu można było lepiej przypisywać wyniki egzaminacyjne szkołom, co stało się palącą potrzebą w obliczu publikowania wskaźników EWD dla kilku okresów trzyletnich dla każdej szkoły. W tym samym roku znacząco zmodyfikowano wskaźniki dla LO i techników. Wprowadzono trzy syntetyczne wskaźniki: umiejętności w zakresie przedmiotów humanistycznych, matematyczno-przyrodniczych i oddzielnie matematyki. Wspólne skalowanie zadań z arkuszy egzaminacyjnych z różnych przedmiotów wymagało zastosowania odpowiednich modeli statystycznych, by rozwiązać problem różnic w poziomie umiejętności subpopulacji przystępujących do egzaminu maturalnego z różnych przedmiotów i na różnych poziomach.
Zmiana struktury egzaminu gimnazjalnego wymusiła w 2012 roku opracowanie nowej aplikacji komputerowej – Kalkulatora EWD 100. Narzędzie to, choć funkcjonalnie podobne do Kalkulatora EWD, ma inną architekturę, dzięki której łatwo można implementować nowe funkcjonalności i nowe typy wskaźników egzaminacyjnych. Ważną zmianą jest też przejście w Kalkulatorze EWD 100, analogicznie do wskaźników wieloletnich, na skale standaryzowane o średniej w skali kraju 100 i odchyleniu standardowym 15. W 2012 roku opublikowano również trzyletnie wskaźniki egzaminacyjne dla LO i techników w zakresie przedmiotów humanistycznych, matematyczno-przyrodniczych i oddzielnie polskiego i matematyki.
W roku 2013 rozpoczął się zakrojony na blisko trzy lata kolejny projekt finansowany ze środków unijnych, dedykowany rozwojowi metody EWD w Polsce. W latach 2013–2015 trwały prace rozwojowe dotyczące modeli szacowania EWD (między innymi jednoroczny model EWD dla maturalnej matematyki na potrzeby Kalkulatora EWD 100), porządkowano bazy danych, stworzoną nową stronę internetową oraz prowadzono działania upowszechniające i badawcze dotyczące wykorzystania metody EWD w ewaluacji nauczania.

Problem trafności wskaźników EWD

W dotychczasowych pracach nad rozwojem metody EWD w Polsce dużą wagę przykładano do weryfikacji trafności wskaźników EWD. Przed konstruktorami modeli EWD stoi wiele trudnych do pokonania przeszkód. Na podstawie współczesnych ujęć trafności i wymagań stawianych testom używanym w modelowaniu EWD można wskazać na pięć powiązanych z sobą, ale dających się

analitycznie rozdzielić aspektów badania trafności metody EWD. Schematycznie przedstawia je poniższy rysunek.

Rysunek 4.1. Aspekty badania trafności metody EWD



Pierwszy aspekt trafności wiąże się z faktem, że podstawowych danych do modelowania dostarczają testy egzaminacyjne. Jakość tych danych jest warunkiem koniecznym trafności metody EWD. W zakresie jakości danych testowych przede wszystkim stawia się pytanie o to, czy wykorzystane do modelowania EWD testy są dobrą miarą realizacji celów nauczania. Wiąże się to z problemem zakresu dozwolonej generalizacji treściowej wyników testu. Test jest zawsze tylko próbką wymagań programowych, ale jeżeli próbka została dobrze dobrana, to wyniki testu można z dającym się zaakceptować błędem uogólniać na całość wymagań programowych (przynajmniej tę część, która była podstawą budowy testu). Ważne jest zatem, by sprawdzić, czy wyniki testów, które wykorzystujemy w modelowania EWD, pozwalają na takie uogólnienie. Jednak sam fakt, że dwa testy dobrze reprezentują wymagania programowe na poziomie kolejnych lat nauczania, jest niewystarczający. Jeżeli wykorzystywane testy nie pozwalają śledzić zmian w poziomie umiejętności na kolejnych etapach edukacji, to należy sprawdzić, czy testy wykorzystane do szacowania danego wskaźnika EWD mierzą ten sam lub przynajmniej bardzo zbliżony konstrukt. Jest to konieczne, by móc interpretować uzyskane wskaźniki jako miary względnego przyrostu osiągnięć szkolnych. To założenie jest w praktyce trudne do spełnienia, bo nawet egzaminy z tego samego przedmiotu, np. egzamin gimnazjalny i matura z matematyki, obejmują różne treści matematyczne, np. składają się z różnych proporcji treści z algebry czy geometrii. Ważna jest również rzetelność pomiaru osiągnięć szkolnych i to, by jego wyniki nie były obciążone działaniem czynników niezwiązanych z poziomem osiągnięć. Ponadto skala, na której wyrażane są wyniki pomiaru, powinna być skalą co najmniej interwałową.
Drugi aspekt trafności metody EWD wiąże się ze statystycznym modelowaniem postępu. Jak właściwie uwzględnić w modelowaniu hierarchiczny charakter danych edukacyjnych: szkoła–oddział –uczeń? Jak uwzględnić zjawisko regresji ku średniej? Za pomocą jakiej statystyki opisywać

przeciętny postęp w zakresie osiągnięć szkolnych uczniów w danej szkole? Jak uwzględnić różne przedmioty egzaminacyjne? Jakie zmienne kontrolne uwzględniać w modelowaniu?

Ostatni problem – zmienne kontrolne – prowadzi nas bezpośrednio do trzeciego aspektu trafności metody EWD: przyczynowej interpretacji wskaźników EWD. Jeżeli wskaźniki EWD mają być użyteczne, to musimy mieć argumenty pozwalające oszacowany za pomocą modelu EWD postęp w zakresie osiągnięć szkolnych zasadnie przypisać działaniom szkoły, czyli wskaźniki efektywności nauczania powinny dawać się sensownie interpretować w kategoriach przyczynowo-skutkowych. Ponieważ wykorzystywane w EWD dane nie są zbierane w schemacie eksperymentalnym, w którym uczniowie byliby losowo przypisywani do szkół, nie możemy konkluzywnie wypowiadać się na podstawie modeli EWD o przyczynowo-skutkowych zależnościach między zmiennymi (Rubin, Stuart i Zanutto, 2004). Innymi słowy nie mamy pewności, że szkoła mająca wysokie wskaźniki EWD miałaby podobne wskaźniki, gdyby uczęszczali do niej inni uczniowie o tych samych wynikach w momencie rozpoczęcia nauki w szkole. W modelowaniu EWD możemy jedynie próbować kontrolować czynniki, które są niezależne od szkoły, a rzutują na postępy uczniów. W sytuacji nielosowego przydziału uczniów do szkół – co ma miejsce właściwie w każdym systemie szkolnym, choć z różnym nasileniem – część wpływu przypisywanego działaniom szkoły może być de facto wynikiem oddziaływania zmiennych zewnętrznych wobec szkoły, nieuwzględnianych w obliczaniu EWD. Możemy wprawdzie włączać tego rodzaju zmienne do modelu – np. zmienną wskazującą na to, czy szkoła działa na wsi, czy w mieście - ale utrudni to rozróżnianie od siebie szkół pracujących efektywnie i nieefektywnie. Działoby się tak dlatego, że włączenie do modelu danej zmiennej kontrolnej oznacza, że zakładamy, że nie jest ona powiązana z efektywnością nauczania. Założenie to ma sens w wypadku zmiennej płci ucznia, ale jest mocno problematyczne w wypadku lokalizacji szkoły. Jeżeli obiektywnie bardziej efektywnie nauczają szkoły miejskie (albo wiejskie), to włączenie tej zmiennej kontrolnej do modelu szacowania EWD szkoły zaniży szacunki dla tych szkół. Niestety, nawet najbardziej złożone modele statystyczne nie są w stanie w pełni rozwiązać problemu nielosowego przydziału uczniów do szkół i oddziałów klasowych. Interpretację wskaźnika jako informacji o efektywności szkoły utrudniać też mogą problemy z oddzieleniem efektu szkoły od wpływu innych czynników, takich np. dynamika procesu grupowego w klasie czy zmiany w środowisku społeczno-ekonomicznym, w którym działa szkoła, z jakich tylko część można przypisać działaniom szkoły (Braun i Weiner, 2007). Choć problem przyczynowej interpretacji wskaźników EWD nie może być w pełni rozwiązany, to badawcza weryfikacja trafności wskaźników EWD może rozwiać wiele wątpliwości.

Kolejnym aspektem trafności metody EWD jest skonfrontowanie miar postępu z zewnętrznymi w stosunku do wyników nauczania kryteriów. O tym, czy interpretacja przyczynowo-skutkowa jest zasadna, możemy wnioskować nie tylko na podstawie skuteczności kontroli zmiennych powiązanych z selekcją miedzyszkolną, ale również patrząc, czy szkoły o wysokim EWD to szkoły, w których zgodnie z naszą normatywną wiedzą o funkcjonowaniu szkoły dobrze się dzieje.

Ostatni wyróżniony na schemacie aspekt trafności metody EWD to konsekwencje stosowania tej metody z punktu widzenia ewaluacyjnej funkcji egzaminów zewnętrznych. Czy metoda służy skutecznej ewaluacji i autoewaluacji szkół, czy jest motorem doskonalenia nauczania? Ten problem jest analizowany w ostatnim rozdziale raportu.

Struktura rozdziału[44]

Pierwszy podrozdział poświęcony jest kontekstowym modelom oceny efektywności nauczania dla I etapu edukacyjnego. Dla tego etapu nie może być stosowana standardowa metoda EWD, ponieważ nie dysponujemy i nie będziemy dysponować pomiarem osiągnięć szkolnych na progu

[44] W rozdziale wykorzystano materiał faktograficzny i fragmenty dwóch prac, które powstały w zespole EWD: Dolata R., Hawrot A., Hummeny G., Jasińska A., Koniewski M., Majkut P. i Żółtak T., (2013) *Trafność metody edukacyjnej wartości dodanej dla gimnazjów*. Warszawa, Instytut Badań Edukacyjnych. Dolata, R., Hawrot, A., Humenny, G., Jasińska-Maciążek, A., Koniewski, M. i Majkut, P., (2014). *Kontekstowy model oceny efektywności nauczania po pierwszym etapie edukacyjnym*. Warszawa: Instytut Badań Edukacyjnych.

klasy I. By używać do oceny efektywności nauczania wyników pomiarów osiągnięć przeprowadzanych na zakończenie I etapu edukacji (np. takich jak *Ogólnopolskie Badanie Umiejętności Trzecioklasistów* OBUT), należy w inny sposób kontrolować wpływ pozaszkolnych czynników rzutujących na wyniki uczniów. Wykorzystując dane z przeprowadzonego badania, pokazujemy, że w takiej sytuacji możemy model EWD zastąpić statystyczną kontrolą statusu społecznego i innych cech środowiska rodzinnego ucznia. Drugi podrozdział dotyczy modelu EWD dla II etapu edukacyjnego zbudowanego na podstawie wyników testu diagnostycznego OBUT (miara osiągnięć na „wejściu") i rezultatów sprawdzianu w klasie VI (miara osiągnięć na „wyjściu"). Przedstawiony został model statystyczny oraz możliwości wykorzystania go przez szkoły. W trzecim podrozdziale opisane jest zastosowanie metody EWD dla gimnazjów. Po charakterystyce modeli statystycznych używanych do szacowania jednorocznych i trzyletnich wskaźników EWD, przedstawiono wyniki badań trafności wskaźników i na kilku przykładach wskazano na możliwości wykorzystania przez szkoły wyników egzaminacyjnych w procesie ewaluacji wewnętrznej. Czwarty, ostatni podrozdział traktuje o zastosowaniu metody EWD w liceach ogólnokształcących i technikach. Analogicznie jak w wypadku gimnazjów, opisano modele statystyczne wykorzystywane do obliczania jednorocznych i trzyletnich wskaźników EWD, wskazano na kluczowe wyniki badania trafności oraz opisano możliwości zastosowania metody przez szkoły. W wypadku wskaźników maturalnych bardzo ważnym i trudnym do rozwiązania problemem jest skalowanie wyników egzaminacyjnych, obszerny zatem fragment tego podrozdziału został poświęcony temu zagadnieniu. Każdy podrozdział wieńczy krótki opis ograniczeń metody i perspektyw jej rozwoju w Polsce.

4.2. Kontekstowy model oceny efektywności nauczania po pierwszym etapie edukacyjnym

Pierwsze lata życia uważane są za najważniejsze dla sukcesu edukacyjnego uczniów. W porównaniu do późniejszych okresów nauczania szkoły podstawowe w Polsce na pierwszym etapie edukacyjnym mają do dyspozycji niewiele narzędzi do diagnozy i monitorowania efektów kształcenia. Prowadzone w ramach projektu *Badania dotyczące rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej* (EWD) analizy podłużne w szkołach podstawowych (więcej o badaniu w ramce 4.1) pozwoliły opracować metodę oceny efektywności nauczania w klasach I–III, która jest przykładem tzw. kontekstowych modeli oceny nauczania. W tej części raportu przedstawione zostaną najważniejsze wyniki tych analiz. Kontekstowe modele oceny efektywności nauczania stosowane są od wielu lat w skali lokalnej w Polsce, a w skali krajowej np. w Australii.

Ramka 4.1. Opis podłużnego badania uwarunkowań wyników nauczania w szkołach podstawowych

Badanie podłużne w szkołach podstawowych rozpoczęto w roku szkolnym 2009/2010. Objęło ono reprezentatywną, ogólnopolską próbę losową 312 oddziałów klas pierwszych ze 180 szkół podstawowych (ok. 5 tys. uczniów). Do roku 2013 badanie było realizowane w ramach projektu *Badania dotyczące rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej* (EWD), a od 2014 roku w ramach projektu *Rozwój metody edukacyjnej wartości dodanej na potrzeby wzmocnienia ewaluacyjnej funkcji egzaminów zewnętrznych*. Badanie obejmie pełen cykl szkoły podstawowej (śledzone są losy tych samych uczniów) i skończy się w 2015 roku. Uczestniczą w nim uczniowie, ich rodzice, nauczyciele oraz dyrektorzy szkół.

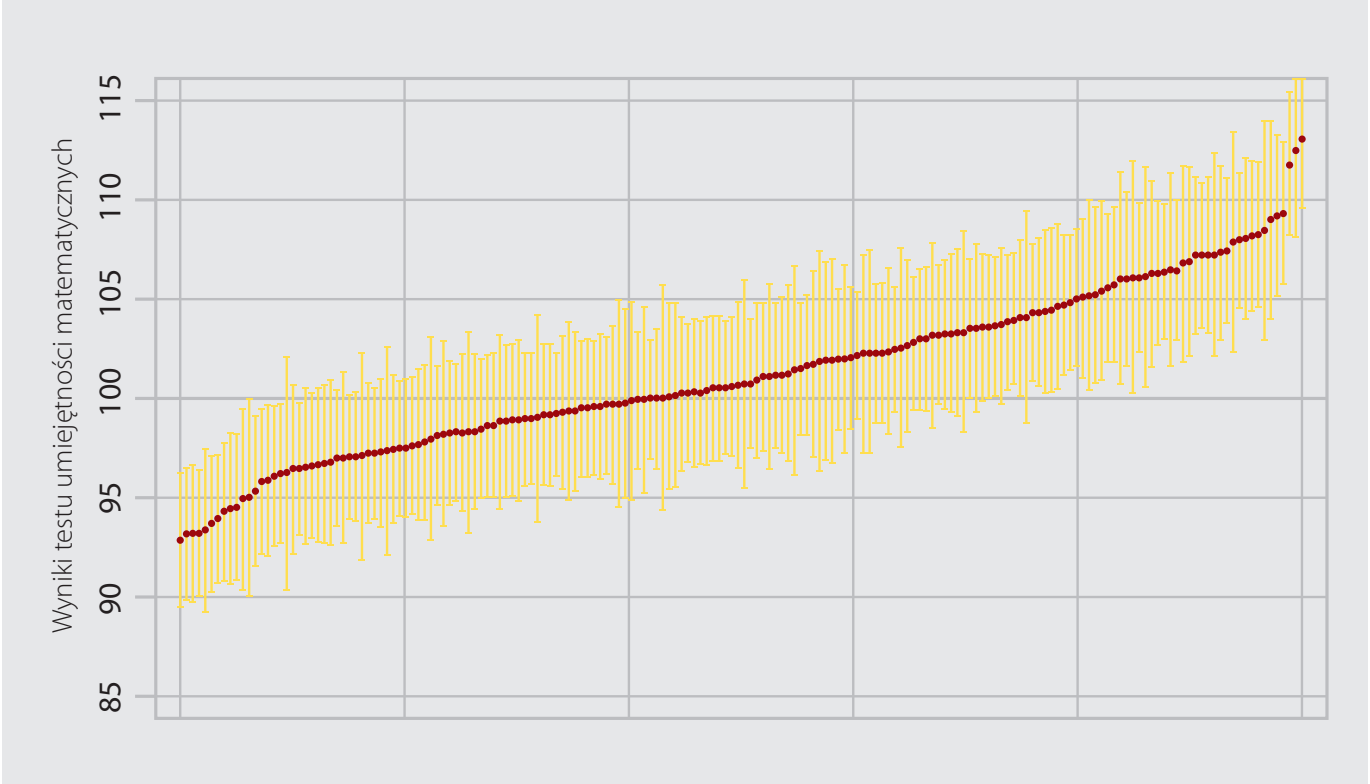
Głównym celem badania jest opracowanie metod oceny efektywności nauczania na I i II etapie edukacyjnym. Ponadto zebrane dane mają pozwolić na lepsze poznanie indywidualnych, rodzinnych i szkolnych czynników odpowiedzialnych za osiągnięcia szkolne uczniów.

Podczas pierwszego etapu badania, który odbył się, kiedy badani uczniowie uczęszczali do pierwszych klas, zebrano m.in. dane opisujące uczniów (płeć, data urodzenia, poziom inteligencji) i ich rodziny (różne miary statusu społeczno-ekonomicznego, aspiracje edukacyjne względem dzieci, struktura rodziny). W kolejnym etapie badania, który odbył się na przełomie III i IV klasy, przeprowadzono m.in. pomiar osiągnięć szkolnych z wykorzystaniem trzech standaryzowanych testów, o udokumentowanej wysokiej jakości (Jasińska i Modzelewski, 2014), mierzących osiągnięcia w zakresie umiejętności czytania, świadomości językowej[1] i umiejętności matematycznych. Więcej informacji na temat badania można znaleźć na stronie: www.ewd.edu.pl.

[1] Test świadomości językowej mierzył zasób słownikowy, elementy wiedzy o języku i umiejętności związane z pisaniem tekstów.

Szkoły podstawowe różnią się wynikami uzyskiwanymi przez uczniów w testach osiągnięć szkolnych. Zróżnicowanie to obserwujemy już po pierwszym etapie edukacyjnym, choć, jak pokazały przeprowadzone badania, nie jest ono bardzo duże. Podział na szkoły wyjaśnia 8,6% zróżnicowania wyników testu umiejętności czytania, 11,8% zróżnicowania wyników testu świadomości językowej oraz 10,5% zróżnicowania wyników testu matematycznego. Ilustrację, jak te odsetki przekładają się na realne różnice między szkołami, pokazano na rysunku 4.2. Zaprezentowano na nim średnie wyniki uzyskane przez uczniów w 180 badanych szkołach w teście matematycznym (czarne punkty) wraz z przedziałami ufności (pionowe kreski). Oś pionowa to wynik testu przedstawiony na standardowej skali o średniej krajowej 100 i odchyleniu standardowym 15. Wyniki szkół zostały uporządkowane rosnąco według wartości średniej.

Rysunek 4.2. Zróżnicowanie międzyszkolne wyników nauczania po III klasie na przykładzie testu matematycznego. Średnie wyniki szkół wraz z 90-procentowymi przedziałami ufności



Przedział ufności wskazuje na zakres wyników, w którym z 90-procentowym prawdopodobieństwem znajduje się prawdziwy wynik szkoły. Dwie szkoły istotnie różnią się wynikami, jeśli ich przedziały ufności są rozłączne. Średni wynik w populacji uczniów wynosi 100 punktów. Wobec tego, jeśli cały przedział ufności znajduje się powyżej wartości 100, z dużą pewnością można powiedzieć, że uczniowie w tej szkole osiągnęli przeciętnie wyższy wynik niż w kraju; jeśli poniżej – uzyskali średnio wynik niższy. Jeśli przedział ufności zawiera wartość 100, średnia szkoły na podstawie uzyskanych w badaniu wyników jest statystycznie nierozróżnialna z przeciętną dla kraju.

Rysunek ten pokazuje skalę zróżnicowania międzyszkolnego wyników uczniów w zakresie osiągnięć matematycznych. Mimo że jest duża grupa szkół, które nie różnią się między sobą pod względem średniego poziomu osiągnięć szkolnych swoich uczniów, to widzimy też sporą liczbę szkół o wynikach istotnie powyżej i poniżej przeciętnych. Natomiast różnica między średnimi wynikami szkoły o najniższej i najwyższej średniej wynosi prawie 20 punktów, czyli 1 i 1/3 odchylenia standardowego wyników uczniów. Międzyszkolne zróżnicowanie osiągnięć szkolnych uczniów ze świadomości językowej jest porównywalne, a z zakresu umiejętności czytania trochę mniejsze.

Czy oznacza to, że szkoły, które osiągnęły średnio wyższe wyniki, uczyły lepiej, efektywniej niż te, które uzyskały słabsze rezultaty? Niekoniecznie. Osiągnięcia szkolne uczniów są bowiem częściowo kształtowane przez czynniki, na które szkoła nie ma wpływu. Spośród pozaszkolnych czynników najsilniej determinuje poziom osiągnięć szkolnych uczniów po III klasie szkoły podstawowej inteligencja (Dolata i in., 2014). Ma ona nieco większe znaczenie dla osiągnięć z matematyki niż ze świadomości językowej i umiejętności czytania. Spośród czynników statusowych największe znaczenie ma zasobność gospodarstwa domowego (mierzona zestawem pytań o posiadane dobra mogące wspierać proces uczenia się, jak np. spokojne miejsce do nauki dla dziecka, dostęp do Internetu, aparat cyfrowy, globus, a także liczba książek w domu dziecka) oraz wykształcenie rodziców. Im wyższy poziom zasobności lub wyższy poziom wykształcenia rodziców, tym więcej punktów otrzymują uczniowie na testach osiągnięć. Pozytywny związek stwierdzono także dla aspiracji edukacyjnych rodziców rozumianych jako pożądany przez rodziców poziom formalnego wykształcenia dla ich dzieci.

Uczniowie, którzy poszli do szkoły rok później niż ich rówieśnicy (mieli więc odroczony start szkolny), uzyskali znacząco niższe wyniki niż średnio ich klasowi koledzy i koleżanki. Z kolei dzieci posłane do szkoły wcześniej – jako sześciolatki – uzyskały wyniki porównywalne z uczniami, którzy rozpoczęli naukę w wieku siedmiu lat. Należy jednak pamiętać, że w roku szkolnym 2009/2010 była to grupa nieliczna i silnie pozytywnie wyselekcjonowana. Podobnie opóźniony start szkolny jest bezpośrednio powiązany z czynnikami mającymi znaczenie dla późniejszych sukcesów szkolnych. Omawianych związków nie można zatem interpretować jako wpływu wieku. Analiza głównej grupie wiekowej wskazuje jednak, że wiek biologiczny uczniów ma znaczenie. Uczniowie starsi (urodzeni bliżej początku roku) uzyskali trochę wyższe wyniki w testach osiągnięć (Dolata i in., 2014).

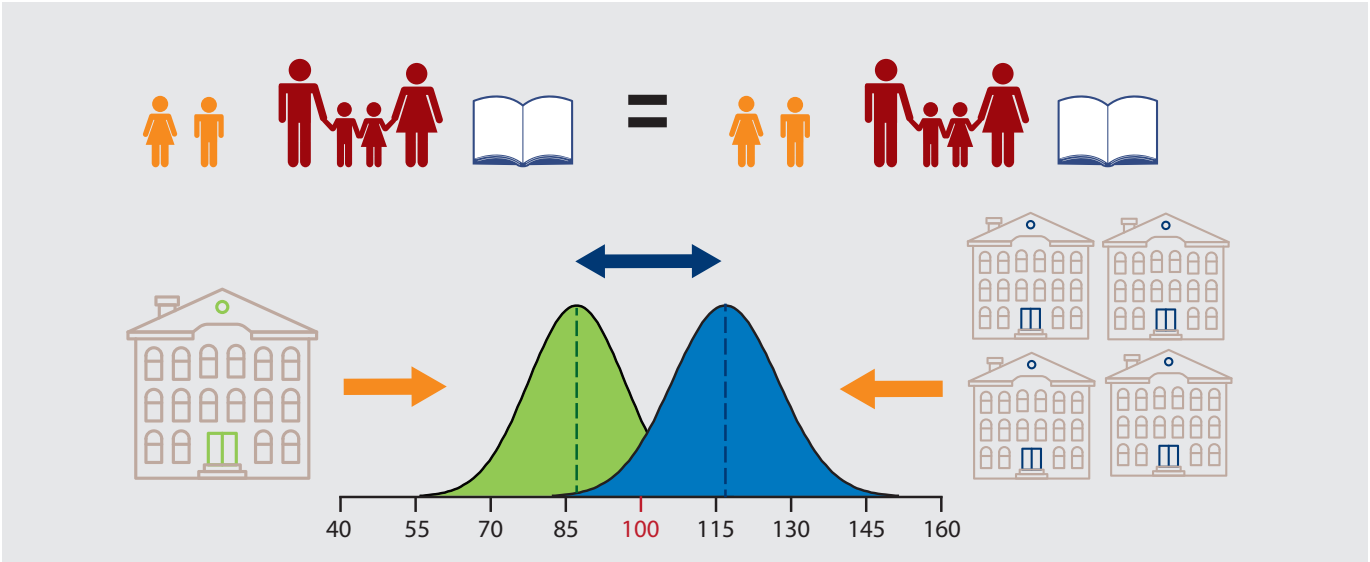
Gdyby szkoły nie różniły się poziomem wymienionych cech uczniów, każda z nich miałaby takie same szanse na uzyskanie podobnych średnich wyników. Tak jednak nie jest. Badania pokazały, że szkoły znacząco różnią się pod względem cech uczniów powiązanych z osiągnięciami szkolnymi. Największe zróżnicowanie stwierdzono dla inteligencji uczniów oraz rodzinnych czynników statusowych, takich jak wykształcenie rodziców, wskaźnik statusu społeczno-ekonomicznego, zasobność rodziny czy prestiż wykonywanego przez rodziców zawodu. Szkoły różniły się także pod względem aspiracji rodziców wobec wykształcenia ich dzieci. Najmniejsze, choć także istotne statystycznie, zróżnicowanie międzyszkolne zaobserwowano dla wieku uczniów.

Kontekstowy model oceny efektywności nauczania

Rezultaty przytaczanych badań potwierdziły, że to, jakie wyniki nauczania uzyskuje szkoła, zależy nie tylko od wysiłku nauczycieli czy oferty edukacyjnej, ale także od tego, jacy uczniowie do niej uczęszczają. Żeby więc móc ocenić wkład danej szkoły w wyniki nauczania, należy porównać je z wynikami szkół, które pracowały z podobnymi uczniami i w podobnym środowisku. Porównań

takich dokonuje się z wykorzystaniem złożonych metod statystycznych. W pewnym uproszczeniu ideę stosowanych modeli analizy ilustruje poniższy rysunek.

Rysunek 4.3. Idea kontekstowych wskaźników efektywności nauczania



Wyniki nauczania mierzone testem uzyskane w danej szkole A (przedstawione na rysunku za pomocą zielonego rozkładu) porównujemy z wynikami szkół pracujących z uczniami o podobnych cechach (niebieski rozkład). Średni wynik w tych szkołach (niebieska przerywana linia) jest punktem odniesienia, to najbardziej prawdopodobny wynik, jaki może uzyskać szkoła, dla której są one grupą odniesienia (takie samo nasilenie cech pozaszkolnych decydujących o osiągnięciach szkolnych). Nazywamy go wynikiem statystycznie przewidywanym, ponieważ – na podstawie danych uzyskanych w badaniu – taki właśnie średni w skali kraju wynik w teście uzyskano, pracując z uczniami o założonych cechach (założonych, czyli takich jak w szkole A).

Następnie wynik przez szkołę faktycznie uzyskany (średni wynik jej uczniów – oznaczony zieloną przerywaną linią) porównujemy z wynikiem dla niej przewidywanym (najbardziej prawdopodobnym). Różnica to wartość kontekstowego wskaźnika efektywności nauczania. Nazywamy go kontekstowym, bo pozwala na porównanie szkół podobnych do siebie ze względu na wiele ważnych cech opisujących kontekst pracy szkoły.

Jeśli w danej szkole uzyskano średnio wyniki niższe niż wartość statystycznie oczekiwana (tak jak na rysunku 4.3), wartość wskaźnika będzie ujemna. Będzie to świadczyć o poniżej przeciętnej efektywności nauczania. Jeśli szkoła uzyskała wynik wyższy niż przewidywany, to wartość wskaźnika efektywności nauczania będzie dodatnia. Będzie to oznaczało ponadprzeciętną efektywność. Jeśli uzyskany przez szkołę średni rezultat i wynik przewidywany będą podobne, wskaźnik efektywności będzie przyjmował wartość w okolicy zera. Będzie to świadczyć o tym, że szkoła naucza z efektywnością taką jak większość szkół w kraju.

Wskaźnik efektywności nauczania jest więc liczony w taki sposób, że szansa na jego wysoką wartość nie zależy od tego, z uczniami o jakich cechach (uwzględnionych w modelu) pracowała szkoła. Szkoła, która świetnie pracowała z uczniami z niekorzystnego środowiska społecznego, osiągnie dodatnią wartość wskaźnika, mimo że jej średnie wyniki w testach osiągnięć będą prawdopodobnie niższe niż szkoły, która słabo pracowała z uczniami pochodzącymi z „dobrych domów".

Prace nad kontekstowymi wskaźnikami efektywności nauczania były możliwe dzięki badaniu przeprowadzonemu z udziałem 180 szkół podstawowych (patrz ramka 4.1). Dane zebrane w jego toku wykorzystano do zbudowania oraz przetestowania kilku alternatywnych modeli i wyboru optymalnego. Obliczono także wskaźniki dla 180 szkół biorących udział w projekcie.

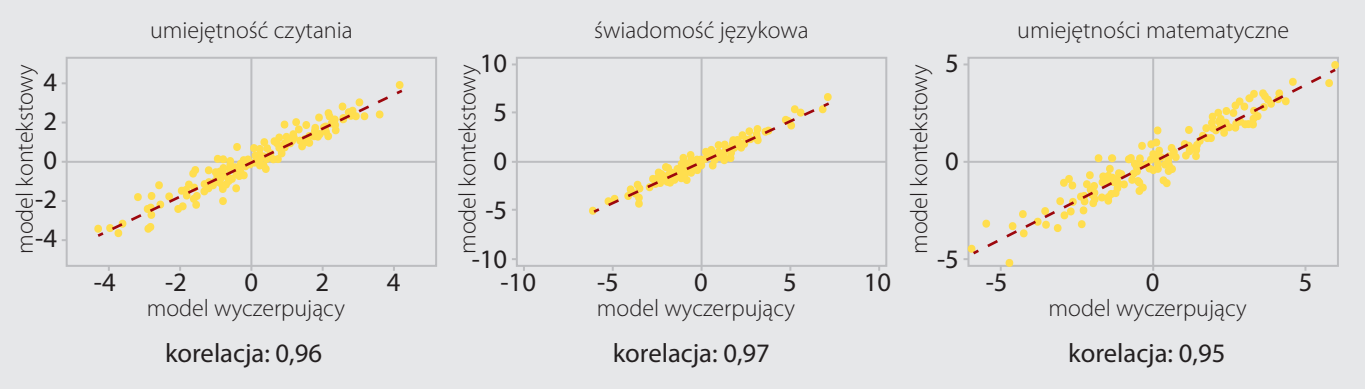
Jakie cechy uczniów uwzględniono w kontekstowych modelach oceny efektywności nauczania? Wybierając zmienne opisujące środowisko pracy szkoły, kierowano się dwoma kryteriami. Po pierwsze, wybrane charakterystyki miały wyjaśniać znaczącą część indywidualnego i międzyszkolnego zróżnicowania osiągnięć szkolnych. Po drugie, wykorzystane informacje musiały być możliwe do pozyskania przez szkoły. W sytuacji badawczej można było zebrać znacznie więcej informacji (np. o poziomie inteligencji uczniów), jednak gdyby wskaźniki kontekstowe miały być wyliczane na większą skalę, trzeba by opierać się przede wszystkim na takich danych, które są możliwe do zebrania w sposób wiarygodny przez szkoły. Przeprowadzone badanie pozwoliło jednak sprawdzić, czy taki zdawałoby się ograniczony zbiór informacji można uznać za wystarczający.

W kontekstowym modelu oceny efektywności nauczania po pierwszym etapie edukacyjnym spośród zmiennych opisujących rodzinę ucznia wykorzystano wskaźniki statusu społeczno-ekonomicznego rodziny ucznia (informację o wykształceniu rodziców i liczbie posiadanych książek, która jest wyrazem zarówno kapitału kulturowego rodziny, jak i jej zasobności) oraz informację o aspiracjach edukacyjnych rodziców względem ich dzieci. W modelu uwzględniono także informację o płci i wieku uczniów. Dodatkowo środowisko, w którym pracuje szkoła, opisano za pomocą średniej wskaźnika wykształcenia rodziców uczniów w szkole. Oznacza to, że kontekstowe wskaźniki efektywności nauczania pokazują wyniki szkół na tle innych, które pracują w takich samych warunkach, ze względu na wymienione charakterystyki. Wskaźniki takie wyliczono dla trzech obszarów osiągnięć szkolnych odpowiadającym trzem wykorzystanym w badaniu testom: umiejętności czytania, świadomości językowej i umiejętności matematycznych.

Trafność kontekstowych wskaźników efektywności nauczania

Skąd jednak wiemy, że uwzględnione w modelu zmienne w wystarczającym stopniu pozwalają uwzględnić to, że szkoły pracują z uczniami o różnym potencjale? Dzięki przeprowadzonemu badaniu, w którym zebrano informację o wielu innych cechach uczniów, ważnych dla ich przyszłych osiągnięć, możliwe było wyliczenie wskaźników uwzględniających szeroki wachlarz charakterystyk opisujących kontekst pracy szkoły i porównanie ich z zaproponowanymi wskaźnikami efektywności nauczania. W rozbudowanych modelach (nazwanych wyczerpującymi), poza zmiennymi uwzględnionymi podczas wyliczania omówionych kontekstowych wskaźników efektywności nauczania, wzięto dodatkowo pod uwagę informację o poziomie inteligencji uczniów (jako charakterystykę indywidualną uczniów oraz średnią w szkole), zasobności rodziny, prestiżu wykonywanych przez rodziców zawodów oraz strukturze rodziny. Wskaźniki obliczone z modeli wyczerpujących porównano z proponowanymi wskaźnikami kontekstowymi. Zestawienie to dla każdego z trzech obszarów osiągnięć pokazano na poniższym rysunku.

Rysunek 4.4. Porównanie kontekstowych wskaźników efektywności nauczania ze wskaźnikami obliczonymi z modeli wyczerpujących



Na osiach poziomych opisano wartości wskaźników obliczonych z modeli wyczerpujących, na osiach pionowych – wartości wskaźników z proponowanych modeli kontekstowych. Kropki pokazują pozycje badanych szkół na tych dwóch wymiarach. Pod wykresami podano współczynniki korelacji między dwoma wskaźnikami. Jeśli korelacja byłaby równa 1, oznaczałoby to, że wskaźniki wyliczone z obu modeli przyjmują takie same wartości, a punkty na wykresach ułożyłyby się na prostej linii. Jeśli współczynnik korelacji wyniósłby 0, oznaczałoby to, że nie ma żadnego związku między wartościami wskaźników obliczonych z obu modeli, więc punkty na wykresach przyjęłyby kształt rozproszonej chmury.

Prezentowane wyniki pokazały, że wskaźniki wyliczone z proponowanych modeli kontekstowych przyjmują wartości bardzo podobne do tych z modeli zawierających znacznie więcej charakterystyk opisujących uczniów i środowisko pracy szkoły. Oznacza to, że mimo iż w proponowanych modelach nie uwzględniamy np. informacji o inteligencji uczniów czy prestiżu zawodu rodziców, to pozostałe zmienne w wystarczającym stopniu pozwalają opisać międzyszkolne zróżnicowanie niezależnych od szkoły cech uczniów mających znaczenie dla osiągnięć uczniów. Jest to jeden z argumentów na rzecz trafności proponowanych wskaźników.

Innym spojrzeniem na problem trafności proponowanych wskaźników jest sprawdzenie, czy w stosunku do nieprzetworzonych wyników testów są one znacząco słabiej zależne od czynników pozaszkolnych wyznaczających osiągnięcia szkolne. Przeprowadzone badania potwierdziły, że średnie wyniki testów osiągnięć uczniów w szkołach są silnie determinowane przez kontekst pracy szkoły. Zależą one od charakterystyk takich jak: średni poziom inteligencji uczniów w szkole czy średni status społeczno-ekonomiczny (mierzony wykształceniem rodziców, wskaźnikiem prestiżu wykonywanego zawodu lub wskaźnikiem zasobności gospodarstwa domowego). Wartości kontekstowych wskaźników efektywności nauczania niemal trzykrotnie słabiej zależą od średniego poziomu inteligencji uczniów, a co więcej, nie są powiązane ze statusem społeczno-ekonomicznym rodzin uczniów w szkole. W tabeli 4.1 zaprezentowano współczynniki korelacji między wskaźnikami opisującymi kontekst pracy szkoły a średnimi wynikami w testach osiągnięć oraz kontekstowymi wskaźnikami efektywności nauczania. Pogrubione wartości wskazują, że korelacja jest istotna statystycznie (współczynnik korelacji dla związków pozytywnych może przyjmować wartość od 0, co oznacza brak związku, do 1 – liniowa zależność).

Tabela 4.1. Korelacje średnich wyników szkół oraz efektywności nauczania z miarami kontekstu pracy szkoły

Wskaźniki wyliczone jako średnia dla szkoły	Umiejętność czytania		Świadomość językowa		Umiejętności matematyczne	
	średni wynik testu	wsk. efektywności	średni wynik testu	wsk. efektywności	średni wynik testu	wsk. efektywności
Średnia inteligencja uczniów[1]	**0,563**	**0,208**	**0,547**	**0,175**	**0,569**	**0,197**
Średnie wykształcenie rodziców[2]	**0,731**	0,014	**0,737**	0,010	**0,720**	-0,005
Średnia wskaźnika prestiżu wykonywanego przez rodziców zawodu	**0,697**	0,015	**0,694**	0,002	**0,704**	0,027
Średnia zasobność gospodarstwa domowego[3]	**0,704**	0,049	**0,699**	0,035	**0,704**	0,034

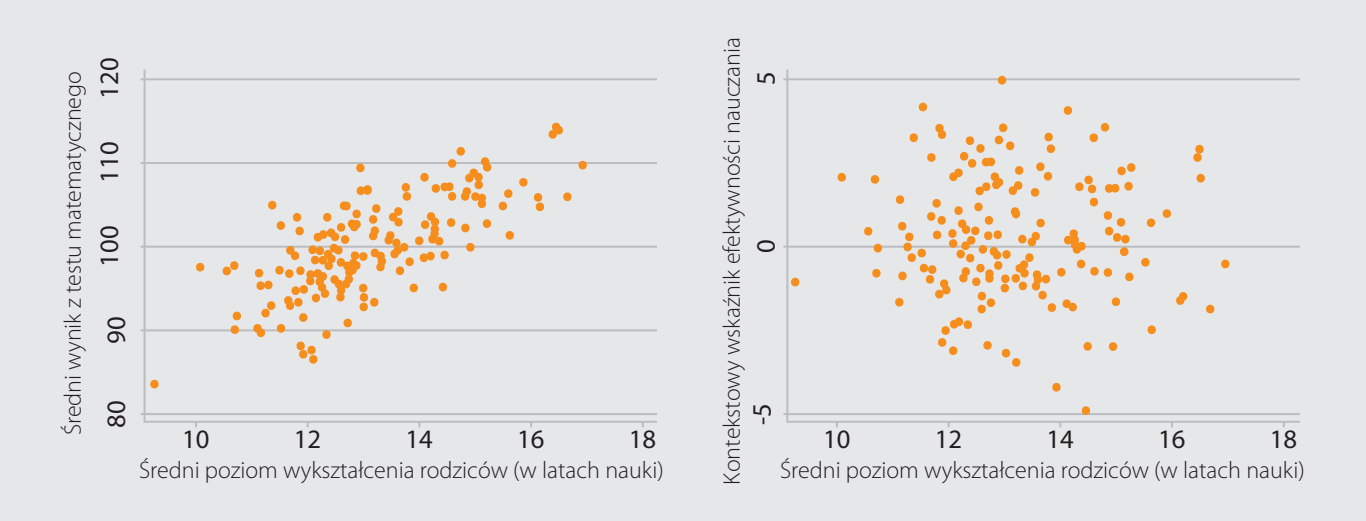
[1] Wykorzystano Test Matryc Ravena.

[2] Wykształcenie rodziców mierzone w latach nauki potrzebnych do osiągnięcia danego poziomu wykształcenia.

[3] Zastosowano skalę wykorzystywaną w badaniach PISA OECD.

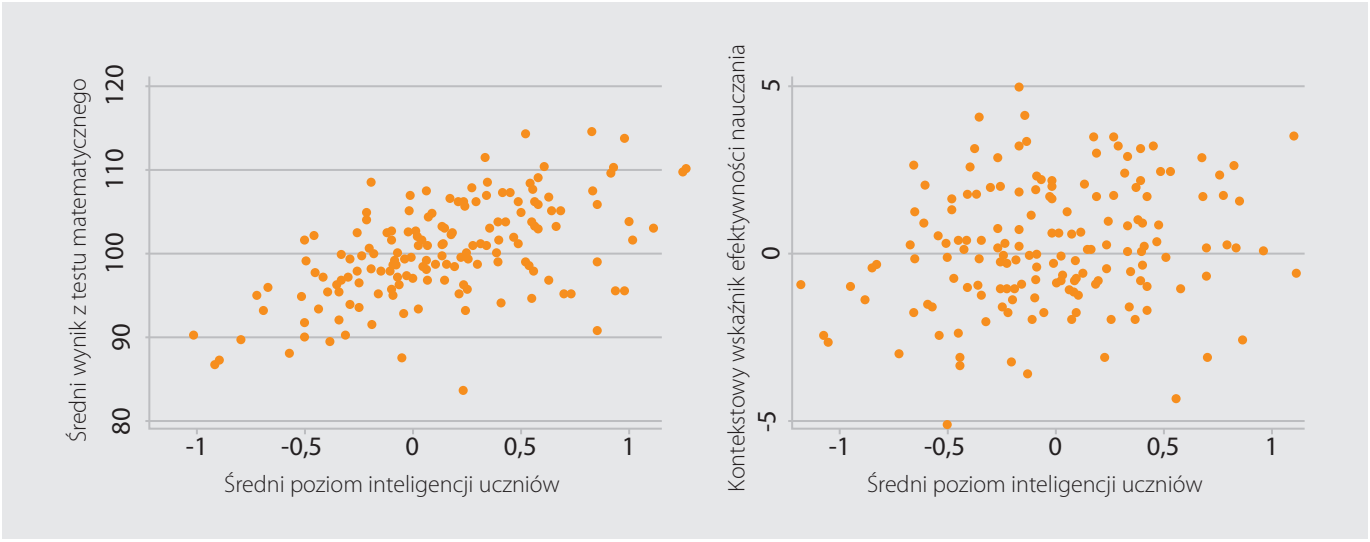
Ilustrację tego, co w praktyce oznaczają podane wartości, zaprezentujemy na przykładzie testu matematycznego. Na rysunku 4.5. przedstawiono zależność wyników testu od poziomu wykształcenia rodziców w szkole (z lewej) oraz relację między wskaźnikiem efektywności nauczania a wykształceniem rodziców uczniów (z prawej). Jedna kropka reprezentuje jedną szkołę.

Rysunek 4.5. Związek średniego poziomu wykształcenia rodziców uczniów ze średnimi wynikami szkół (z lewej) oraz z efektywnością nauczania mierzoną wskaźnikami kontekstowymi (z prawej) dla obszaru umiejętności matematycznych



Wykres z lewej strony bardzo dobrze pokazuje, jak silnie średni wynik w teście osiągnięć szkolnych zależy od środowiska, w którym pracuje szkoła (opisanego tu przez średni poziom wykształcenia rodziców). Szkoły uzyskujące wysokie wyniki to przede wszystkim szkoły pracujące z uczniami pochodzącymi z rodzin o wyższym poziomie wykształcenia rodziców. Szkoły uzyskujące niższe wyniki pracują z uczniami, których rodzice zdobyli średnio niższy poziom wykształcenia. Oceniając więc pracę szkoły tylko na podstawie wyników testowych, oceniamy ją w dużej mierze za to, na co nie ma ona wpływu. Kontekstowe wskaźniki efektywności nauczania natomiast nie zależą od statusu społeczno-ekonomicznego rodzin uczniów. Każda szkoła ma szansę na wysoką wartość wskaźnika, co wyraźnie widać na wykresie prawym, gdzie punkty na wykresie mają kształt rozproszonej chmury. Jak jednak ten obraz wygląda w przypadku inteligencji uczniów, gdzie nawet dla kontekstowej miary efektywności zaobserwowano pozytywną, choć słabą zależność? Przedstawiamy to na poniższym rysunku. Wykresy pokazują, że szkoły, w których średni poziom inteligencji uczniów jest poniżej średniej, uzyskują raczej przeciętne lub poniżej przeciętnej wyniki w testach osiągnięć. Mają one natomiast szansę na dodatnią wartość kontekstowego wskaźnika efektywności nauczania. Wskaźniki efektywności są bowiem znacznie słabiej powiązane ze średnim poziomem inteligencji uczniów.

Rysunek 4.6. Związek średniego poziomu inteligencji uczniów ze średnimi wynikami szkół (z lewej) oraz z efektywnością nauczania mierzoną wskaźnikami kontekstowymi (z prawej) dla obszaru umiejętności matematycznych



Za inny dowód trafności kontekstowych wskaźników efektywności nauczania mogą posłużyć dane przemawiające za tym, że szkoły określone przez proponowane miary jako efektywne, to placówki, w których obserwujemy korzystne z normatywnego punktu widzenia procesy i zjawiska, i nie dostrzegamy czegoś niepokojącego. Dotychczasowe analizy doprowadziły do kilku interesujących wniosków.

W szkołach czy oddziałach, które zostały określone mianem efektywnych (na podstawie proponowanych wskaźników), obserwujemy wyższy poziom integracji szkolnej uczniów w trzech wymiarach:

- integracji społecznej, związanej z satysfakcją z kontaktów z klasowymi kolegami, oceną możliwości zbudowania satysfakcjonujących więzi społecznych;
- integracji emocjonalnej, wyrażającej pozytywny stosunek uczniów do szkoły;
- integracji motywacyjnej, która odnosi się do obrazu siebie jako ucznia i znajduje wyraz w samoocenie szkolnych kompetencji; wyraża to, czy uczeń czuje, że jest w stanie sprostać wymaganiom stawianym mu w szkole i podejmować zadania z pozytywną motywacją.

Istotne zależności obserwuje się dla wszystkich analizowanych obszarów osiągnięć (umiejętności czytania, świadomości językowej i umiejętności matematycznych). Integracja szkolna opisuje wymiary środowiska szkolnego, które mogą być kształtowane przez nauczycieli i pozostałych pracowników szkoły. Pozytywny związek miar integracji z efektywnością nauczania, mierzoną omawianymi wskaźnikami, pokazuje, że w placówkach o wysokich wartościach wskaźnika kontekstowego uczniowie odczuwają ponadprzeciętną satysfakcję z relacji społecznych, mają lepsze nastawienie do szkoły, a także większą wiarę w swoje możliwości uczenia się.

Analiza czynników związanych z nauczycielem pozwoliła na wykrycie tylko części z zakładanych zależności. Stwierdzono związek wykształcenia nauczycieli z efektywnością nauczania w zakresie umiejętności czytania i świadomości językowej. Uczniowie nauczani przez osoby z wykształceniem wyższym magisterskim uzyskali trochę wyższe wyniki w tych obszarach w stosunku do uczniów nauczycieli z tytułem licencjata, jeśli porównujemy dzieci o takim samym poziomie cech uwzględnionych w kontekstowym modelu efektywności nauczania. Dla rozwoju osiągnięć szkolnych w zakresie świadomości językowej i umiejętności matematycznych znaczenie miało także to, czy nauczyciel potrafi utrzymać dyscyplinę na lekcjach. Nauczyciele, którzy lepiej panowali nad sytuacją w klasie, uzyskiwali lepsze wyniki nauczania. Nie stwierdzono natomiast związku efektywności nauczania ze stylem pracy nauczyciela na lekcjach (co może wynikać z deklaratywnego charakteru tych danych) ani ze stopniem awansu zawodowego.

Analiza znaczenia kultury organizacyjnej szkoły i stylów zarządzania przyniosła niejednoznaczne rezultaty; w większości przypadków stwierdzono brak istotnych związków. Przyczyną może być to, że ewentualne zależności są zbyt subtelne lub zastosowane w badaniu narzędzia okazały się za mało wrażliwe na to, co w zarządzaniu szkołą ma największe znaczenie dla osiągnięć szkolnych. Wykryto jednak m.in. pozytywny związek współpracy rodziców ze szkołą z efektywnością nauczania, ale tylko w obszarze świadomości językowej.

Analizę trafności proponowanych kontekstowych wskaźników efektywności nauczania na pewno trzeba pogłębić. Czytelników zainteresowanych szczegółowymi wynikami omawianych tu analiz odsyłamy do przytaczanej już publikacji (Dolata i in., 2014), dostępnej także na stronie projektu[45].

Komunikowanie kontekstowych wskaźników efektywności nauczania i możliwości wykorzystania ich przez szkoły

Szkoły, które uczestniczyły w badaniu, otrzymały zindywidualizowane raporty przedstawiające wyniki i efektywność nauczania w zakresie czytania, świadomości językowej i umiejętności matematycznych. Wyniki były omawiane z przedstawicielami szkół na specjalnej konferencji podsumowującej pierwszą część badania[46]. Poniżej prezentujemy przykładowe omówienie wyników dla jednego z obszarów umiejętności dla wybranej szkoły. Można je potraktować jako propozycję analiz, jakie mogłyby wykonywać szkoły, gdyby kontekstowe modele oceny efektywności nauczania były dostępne dla wszystkich szkół. Prezentowane analizy zostały zaczerpnięte z jednego z raportów dla szkół.

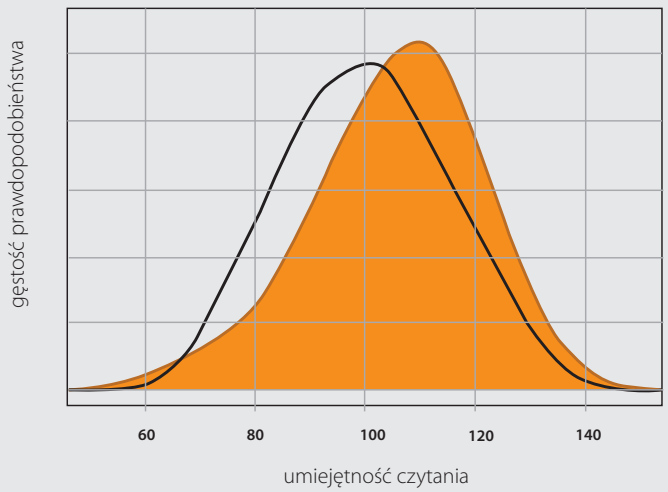
Wyniki uzyskane przez uczniów danej szkoły w każdym z testów osiągnięć zostały zaprezentowane na trzy sposoby. Po pierwsze, zaprezentowany został rozkład wyników w szkole na tle rozkładu wyników w populacji krajowej uczniów (rysunek 4.7). Pomarańczowym kolorem zaznaczono rozkład wyników w teście umiejętności czytania uczniów w szkole. Czarną linią wyrysowano rozkład wyników w populacji krajowej uczniów. Jego średnia wynosi 100 punktów, a odchylenie standardowe 15 punktów. Porównując wzajemne położenie obu tych rozkładów, można ocenić, czy uczniowie szkoły uzyskali lepsze czy słabsze wyniki niż średnio uczniowie w kraju. Jeśli rozkład wyników w szkole jest przesunięty w prawo w stosunku do rozkładu populacyjnego, oznacza to, że uczniowie uzyskali wyniki powyżej przeciętnych. Gdyby rozkład był przesunięty w lewo, oznaczałoby to, że uczniowie uzyskali wyniki poniżej średniej.

Dodatkowo można porównać kształty obu rozkładów. Jeśli rozkład wyników w szkole jest bardziej skupiony wokół średniej niż rozkład populacyjny, oznacza to, że wyniki uczniów w szkole są mniej zróżnicowane niż średnio w populacji. Jeśli jest on silniej rozproszony wokół średniej, oznacza to, że uczniowie w szkole uzyskali wyniki bardziej zróżnicowane niż średnio w populacji. W przypadku prezentowanej szkoły kształt rozkładu jest zbliżony do rozkładu populacyjnego, co oznacza, że zróżnicowanie wyników w szkole jest typowe.

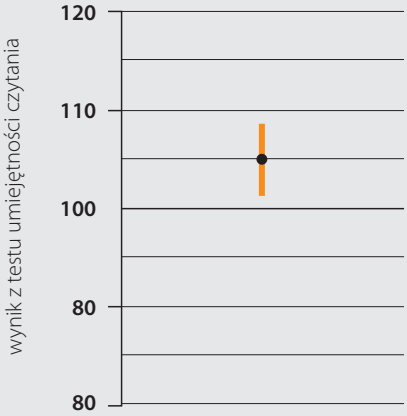
[45] http://ewd.edu.pl/publikacje/

[46] Informacja o konferencji oraz materiały pokonferencyjne znajdują się na stronie: http://2013.ewd.edu.pl/szkoly-ewd/sp-2013/

Rysunek 4.7. Rozkład wyników w teście umiejętności czytania w szkole na tle rozkładu wyników w populacji



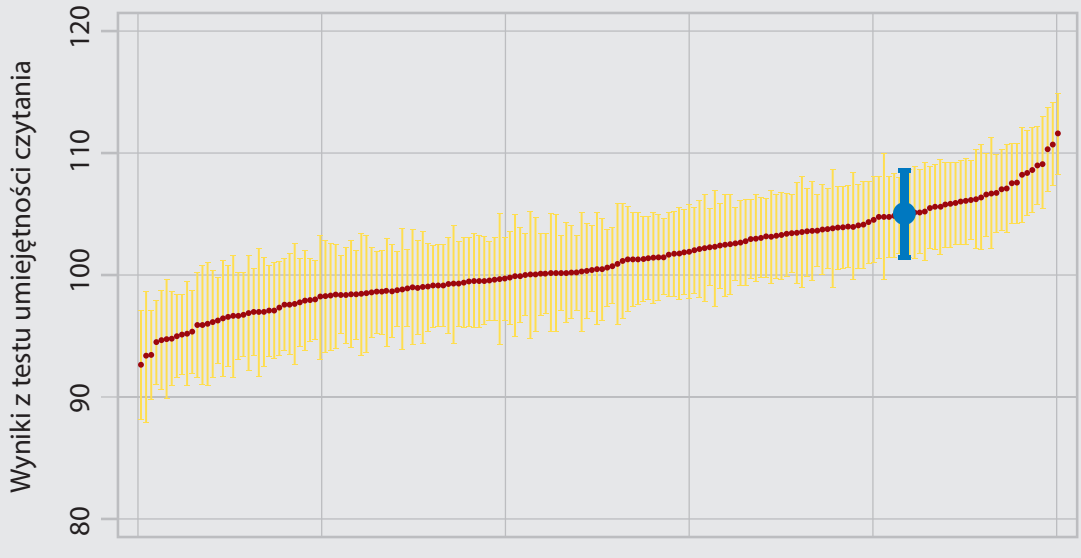
Rysunek 4.8. Średni wynik szkoły w teście umiejętności czytania wraz z przedziałem ufności



Drugi sposób pokazania wyników szkoły to prezentacja średniego wyniku w teście osiągnięć wraz z przedziałem ufności (rysunek 4.8). Pozwala on odpowiedzieć na pytanie, czy średni wynik testu w szkole jest istotnie statystycznie różny od przeciętnej dla kraju. Oś pionowa to skala wyników uzyskanych w teście umiejętności czytania. Średni wynik szkoły jest reprezentowany przez czarny punkt. Przedział ufności (przyjęto poziom ufności 90%) jest zaznaczony niebieskim odcinkiem. Jeśli cały przedział ufności znajduje się ponad wynikiem 100 (tak jak na prezentowanym wykresie), to można z bardzo dużą pewnością powiedzieć, że uczniowie uzyskali w teście wynik wyższy niż średni wynik w populacji. Gdyby cały przedział ufności znajdował się poniżej wyniku 100, można by poprawnie statystycznie lokować średni wynik szkoły poniżej średniej krajowej. Gdyby przedział ufności przecinał poziom 100 punktów, oznaczałoby to, że uczniowie uzyskali wynik porównywalny ze średnim wynikiem w populacji uczniów.

Średni wynik szkoły przedstawiono także na tle wyników wszystkich szkół uczestniczących w badaniu (wykres poniżej). Wyniki szkół zostały uporządkowane rosnąco według wartości średniej. Informacja ta miała pomóc szkołom lepiej zinterpretować położenie ich wskaźnika. Dzięki niej można zobaczyć, ile szkół ma istotnie różne wyniki od danej szkoły.

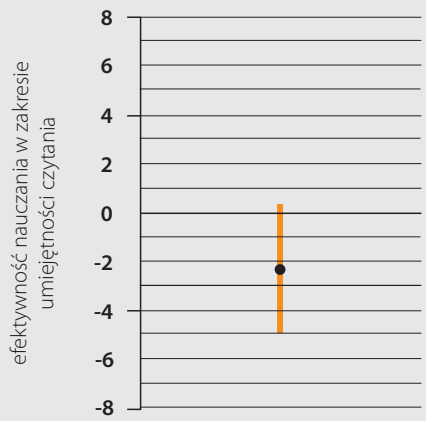
Rysunek 4.9. Średnie wyniki szkoły w teście umiejętności czytania na tle wyników wszystkich badanych szkół



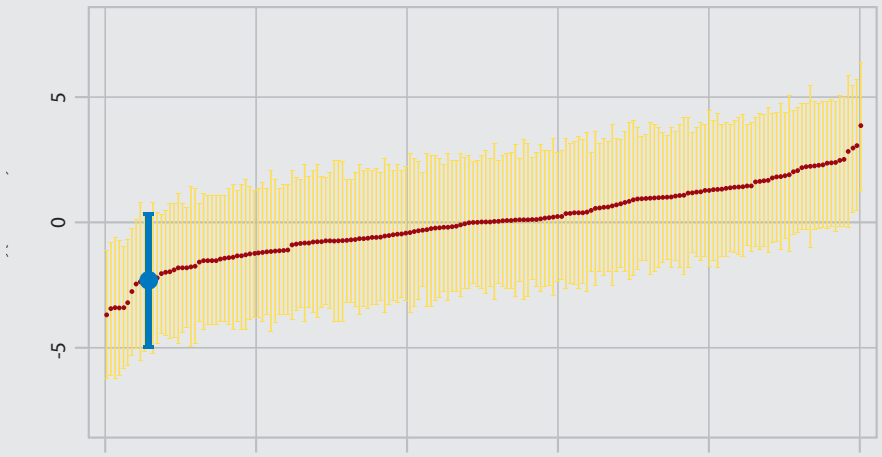
Kontekstowe wskaźniki efektywności nauczania zostały zaprezentowane graficznie wraz z przedziałem ufności. Czarnym punktem oznaczono średnią wartość wskaźnika, a pomarańczowym odcinkiem 90-procentowy przedział ufności. Oś pionowa to skala efektywności nauczania. Im wyższy na niej wynik, tym wyższa efektywność nauczania. Wynik 0 oznacza średnią efektywność nauczania w populacji. Jeśli cały przedział ufności wskaźnika znajduje się powyżej punktu 0, oznacza to, że jest wysoce prawdopodobne, że szkoła naucza z powyżej przeciętną efektywnością. Uzyskuje więc średnio wyniki nauczania wyższe niż inne szkoły pracujące z podobnymi uczniami. Jeśli przedział ufności znajduje się poniżej punktu 0, oznacza to, że szkoła z dużym prawdopodobieństwem naucza z poniżej przeciętną efektywnością, a inne szkoły pracujące z podobnymi uczniami uzyskały średnio lepsze rezultaty. Jeśli przedział ufności przecina wartość 0, oznacza to, że szkoła pracuje z przeciętną efektywnością, czyli taką jak większość szkół w kraju.

Informacja o efektywności nauczania także została przedstawiona na tle wskaźników innych szkół biorących udział w badaniu. Efektywność nauczania w zakresie umiejętności czytania w prezentowanej szkole należałoby ocenić jako przeciętną (przedział ufności zawiera zero), ale fakt, że wskaźnik „ciąży" w stronę wyników ujemnych, powinien skłonić szkołę do szukania przyczyn takiej sytuacji.

Rysunek 4.10. Efektywność nauczania w zakresie umiejętności czytania

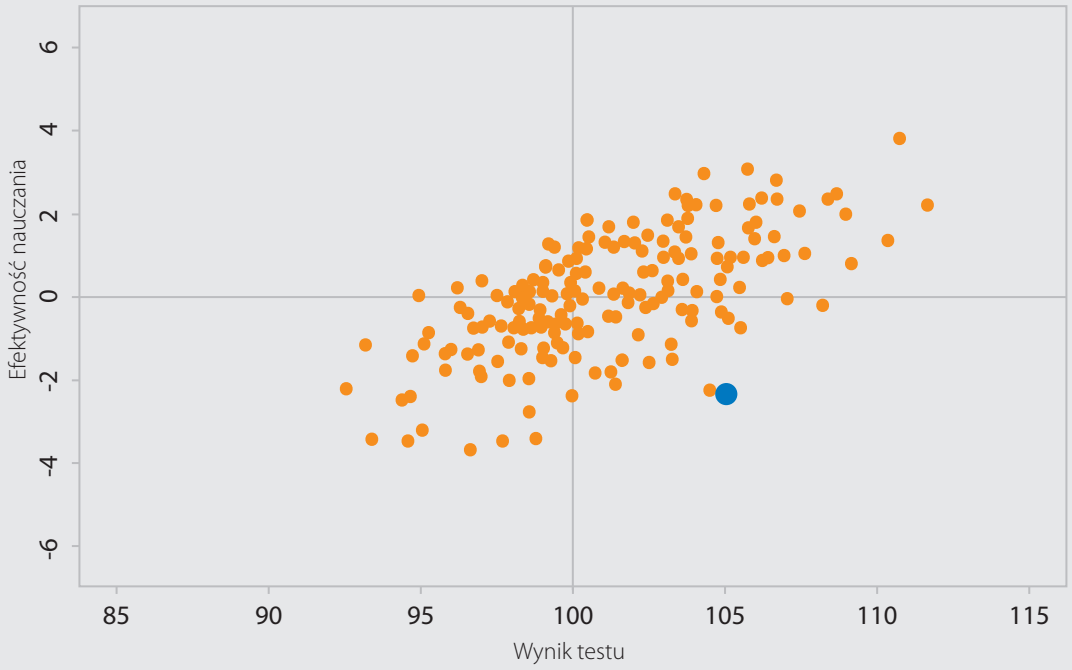


Rysunek 4.11. Efektywność nauczania w zakresie umiejętności czytania na tle wszystkich badanych szkół



Szkołom przedstawiono także łączną informację o średnim wyniku w teście i efektywności nauczania na tle wyników wszystkich szkół biorących udział w badaniu. Dzięki takiemu zestawieniu informacji można zobaczyć, gdzie ze względu na wartości obu wskaźników łącznie znajduje się dana szkoła w porównaniu do innych szkół. Na osi poziomej przedstawiono wyniki uzyskane w teście umiejętności czytania. Na osi pionowej przedstawiono wskaźnik efektywności nauczania w tym obszarze. Na tym wykresie wyniki przedstawiono bez przedziałów ufności, by zwiększyć jego czytelność. Trzeba jednak pamiętać, że są one obarczone niepewnością statystyczną, więc nie należy przeceniać małych różnic.

Rysunek 4.12. Łączna informacja o wyniku w teście osiągnięć i efektywności nauczania w zakresie umiejętności czytania



Szkoły otrzymały informację zarówno o średnich wynikach w testach, jak i efektywności nauczania. Obie te informacje wzajemnie się dopełniają. Wyniki testów są oczywiście w sposób naturalny powiązane z wartościami kontekstowych wskaźników efektywności nauczania, bowiem wyższa lub niższa efektywność nauczania przekłada się na wyższe lub niższe wyniki[47]. Niemniej możemy wyłonić takie szkoły, które osiągają podobne wyniki, ale pracują z różną efektywnością, jak i odwrotnie.

Ograniczenia i rozwój modeli kontekstowych

Przeprowadzone badania pokazały, że kontekstowe wskaźniki efektywności nauczania mogłyby być wartościowym narzędziem wykorzystywanym przez szkoły do ewaluacji własnej pracy na pierwszym etapie edukacyjnym. Jakie warunki musiałyby być spełnione, by można je było powszechnie stosować?

Po pierwsze, niezbędny byłby dobry pomiar osiągnięć szkolnych po trzeciej klasie szkoły podstawowej. Nie jest jednak konieczne, by zostały nim objęte wszyscy uczniowie w kraju. Wystarczające byłoby badanie na reprezentatywnej próbie szkół, którego wyniki posłużyłyby do wyliczenia modeli i ustalenia punktu odniesienia do porównań dla innych szkół. Pozostałe placówki, zainteresowane takimi wskaźnikami, mogłyby dobrowolnie przeprowadzić u siebie takie pomiary i porównać ich rezultaty z wynikami badania reprezentatywnego. W podobnej formule było realizowane badanie OBUT (*Ogólnopolskie Badanie Umiejętności Trzecioklasistów),* co pokazuje, że taki pomiar diagnostyczny jest możliwy i mógłby być podstawą do wyliczania takich wskaźników. Oczywiście, konieczna byłaby krytyczna analiza testów, które byłyby wykorzystane do wyliczenia wskaźników, bowiem od ich trafności i rzetelności zależałaby jakość szacowanych na ich podstawie miar[48]. Poza pomiarem osiągnięć potrzebne byłyby dane o uczniach, które musiałyby być zbierane i przechowywane przez szkoły, a następnie łączone z wynikami testów osiągnięć. Dla proponowanych wskaźników należałoby zebrać informację o wieku, płci dziecka, wykształceniu jego rodziców oraz przeprowadzić

[47] Korelacja pomiędzy kontekstowymi wskaźnikami a wynikami testów wynosi około 0,64.

[48] Testy OBUT nie były konstruowane na potrzeby uzyskania całościowego obrazu umiejętności uczniów. W kolejnych edycjach dobierano różne, punktowe obszary opisane w podstawie programowej, tak by rozwiązania poszczególnych zadań były dobrą informacją zwrotną dla szkół i nauczycieli uczestniczących w badaniu.

krótką ankietę wśród rodziców w celu pozyskania informacji o aspiracjach edukacyjnych względem dzieci i liczbie książek posiadanych w domu. W przypadku pytania o aspiracje edukacyjne ważne jest, by było ono skierowane do rodziców na początku edukacji ich dzieci w klasie pierwszej. Aspiracje rodziców mogą się bowiem zmieniać pod wpływem osiągnięć szkolnych dzieci oraz działań szkoły. Dlatego, jeśli chcemy w modelu uwzględnić czynniki niezależne od pracy szkoły, ważne jest, by pomiar aspiracji przeprowadzać już w pierwszej klasie. To jednak rodzi dodatkową komplikację w postaci konieczności zbierania danych kontekstowych dla badanego rocznika już na początku nauki oraz przechowywania ich i ochrony przez szkoły przez dwa kolejne lata. Oczywiście nie jest to nierealne, a ankieta kierowana do rodziców na początku drogi edukacyjnej ich dzieci mogłaby stanowić okazję dla szkół do poznawania środowiska rodzinnego uczniów. Wyniki analiz pokazały, że można by również rozważyć zbudowanie modeli prostszych, wymagających pozyskania tylko informacji o wieku, płci i wykształceniu rodziców. W takiej sytuacji wystarczyłoby zebranie tych danych w okolicach pomiaru osiągnięć. Wskaźniki wyliczone z takich modeli miały tylko trochę gorsze właściwości (szczegóły zostały omówione w przytaczanej już publikacji: Dolata i in., 2014).
I na koniec kwestia odpowiednich narzędzi wspomagających analizy z wykorzystaniem wskaźników kontekstowych. Po przeprowadzeniu pomiaru osiągnięć szkolnych na koniec pierwszego etapu edukacyjnego zebrane przez szkołę dane mogłyby być wprowadzane do specjalnie przygotowanej aplikacji (np. podobnej do funkcjonujących już Kalkulatora EWD 100 lub Kalkulatora EWD SP, omówionych w kolejnych częściach rozdziału), w której uprzednio zaimplementowano by modele efektywności nauczania, wyliczone na podstawie danych z badania na reprezentatywnej próbie. W ten sposób osoby zainteresowane mogłyby analizować efektywność nauczania w danej szkole czy w podgrupach uczniów w szkole, bez konieczności wysyłania danych do zarządzającej tym procesem centrali.

4.3. Wykorzystanie metody edukacyjnej wartości dodanej po drugim etapie edukacyjnym

Wskaźniki edukacyjnej wartości dodanej dla drugiego etapu edukacyjnego (klasy IV–VI szkoły podstawowej) są ważnym uzupełnieniem tego typu miar dostępnych już od kilku lat dla szkół gimnazjalnych oraz ponadgimnazjalnych. Dotychczasowe modele EWD wykorzystywały wyniki z systemu egzaminacyjnego. Miary EWD opracowane dla drugiego etapu edukacyjnego wykorzystują jako miarę uprzednich osiągnięć wyniki pomiaru diagnostycznego realizowanego w klasach III szkoły podstawowej w ramach *Ogólnopolskiego Badania Umiejętności Trzecioklasistów* (OBUT). W realizowanym od 2011 roku badaniu bierze udział większość szkół podstawowych w Polsce. Ze względu na diagnostyczny charakter OBUT metoda EWD opracowana dla II etapu edukacyjnego ma wspomagać ewaluację wewnątrzszkolną i nie powinna być używana w ewaluacji zewnętrznej.

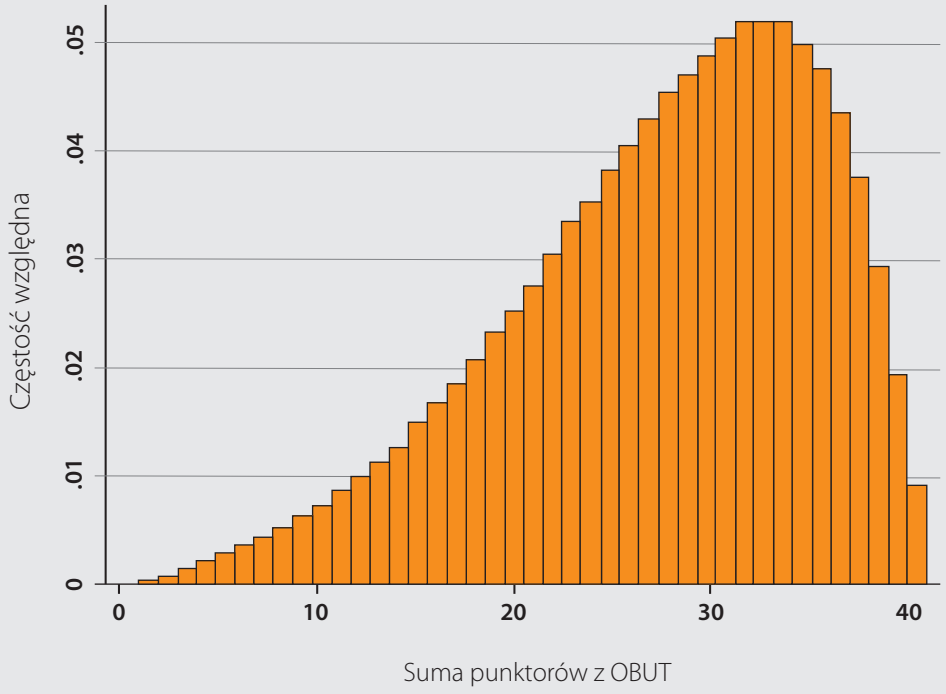
Model edukacyjnej wartości dodanej dla drugiego etapu edukacyjnego

Wyniki sprawdzianu w klasie VI nie są wystarczającą przesłanką do oceny efektywności nauczania w szkole podstawowej (Dolata i in., 2013). Aby ocenić nauczanie w danej placówce na podstawie wyników sprawdzianu, należy porównywać wyniki jej uczniów z rezultatami podobnych uczniów w grupie odniesienia. Dzięki danym z OBUT możliwe jest zastosowanie w tym celu metody EWD. Należy jednak pamiętać, że wskaźniki wyliczone na bazie wyników OBUT i sprawdzianu w klasie VI dotyczą tylko efektywności nauczania w klasach IV–VI i grupą odniesienia jest populacja uczniów uczęszczających do szkół, które brały udział w danej edycji OBUT. Wskaźniki EWD dla drugiego etapu edukacji są miarą relatywną, informującą o tym, czy dana szkoła wypracowała z uczniami więcej, mniej czy tyle samo, co inne, statystycznie podobne szkoły.
Aby można było wyznaczyć EWD dla szkoły, najpierw należy określić prognozę statystyczną mówiącą, jaki jest najbardziej prawdopodobny wynik ucznia na sprawdzianie w zależności od jego wcześniejszych osiągnięć. Taką prognozę nazywamy modelem EWD. Do obliczenia takiego modelu

potrzebne są dwa pomiary osiągnięć szkolnych: na „wejściu" i na „wyjściu" ze szkoły. W dalszej części podrozdziału będziemy zajmować się modelem EWD dla II etapu edukacyjnego, w którym pomiarem na wejściu jest wynik z OBUT 2011, a pomiarem na wyjściu przeprowadzony trzy lata później sprawdzian w klasie VI, czyli w 2014 r.

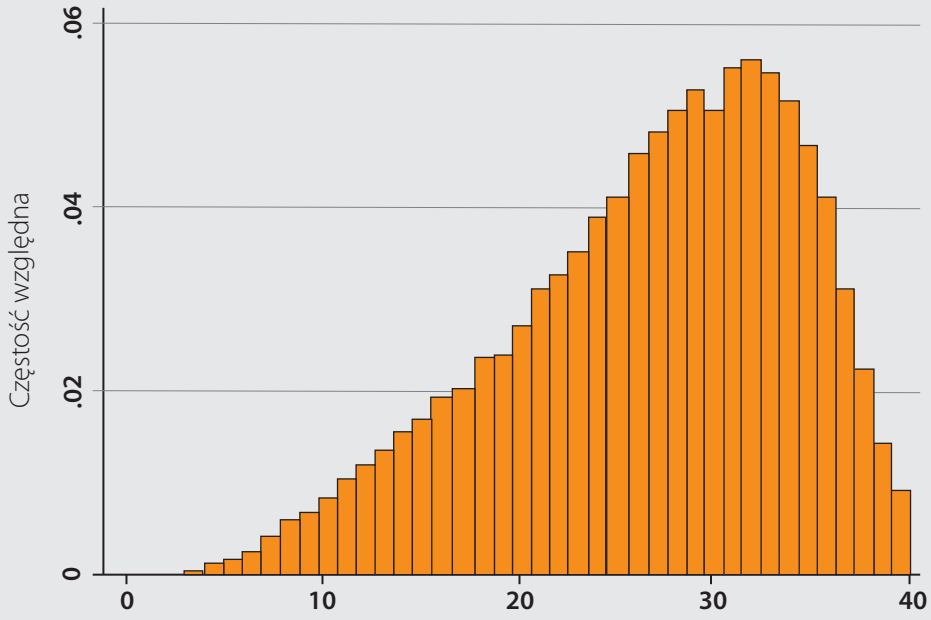
W 2011 roku w badaniu OBUT wzięło udział 9 756 placówek, tj. ponad 75% wszystkich szkół podstawowych. OBUT składał się z dwóch części: polonistycznej i matematycznej. W części z języka polskiego było 11 zadań, za które można było uzyskać łącznie 24 punkty. Część matematyczna składała się z 9 zadań, za ich rozwiązanie maksymalnie można było otrzymać 17 punktów. Łącznie za cały test można było uzyskać 41 punktów. Do głównej bazy badania OBUT przesłano łącznie 273 898 wyników testowania. Średnio badani uczniowie uzyskali 27,8 punktu (przy odchyleniu standardowym 8,1). Rozkład wyników był silnie lewowoskośny, co świadczy o tym, że test był relatywnie łatwy (por. rysunek 4.13).

Rysunek 4.13. Rozkład wyników badania OBUT 2011 w populacji badanych



Pomiarem końcowym osiągnięć szkolnych, a analizowanym modelu EWD jest sprawdzian w klasie VI. W 2014 roku standardowy arkusz egzaminacyjny zawierał 26 zadań, w tym 20 zadań zamkniętych wielokrotnego wyboru oraz 6 otwartych. Arkusz ten rozwiązywali uczniowie bez dysfunkcji i z dysleksją rozwojową. Za poprawne wykonanie wszystkich zadań można było maksymalnie otrzymać 40 punktów. Średni wynik w Polsce na skali wyników surowych to 25,8 punktu (przy odchyleniu standardowym 8,0). Podobnie jak w przypadku OBUT, sprawdzian był relatywnie łatwy, o czym świadczy lewoskośny rozkład wyników (por. rysunek 4.14).

Rysunek 4.14. Rozkład wyników sprawdzianu w 2014 roku w populacji uczniów VI klas szkół podstawowych

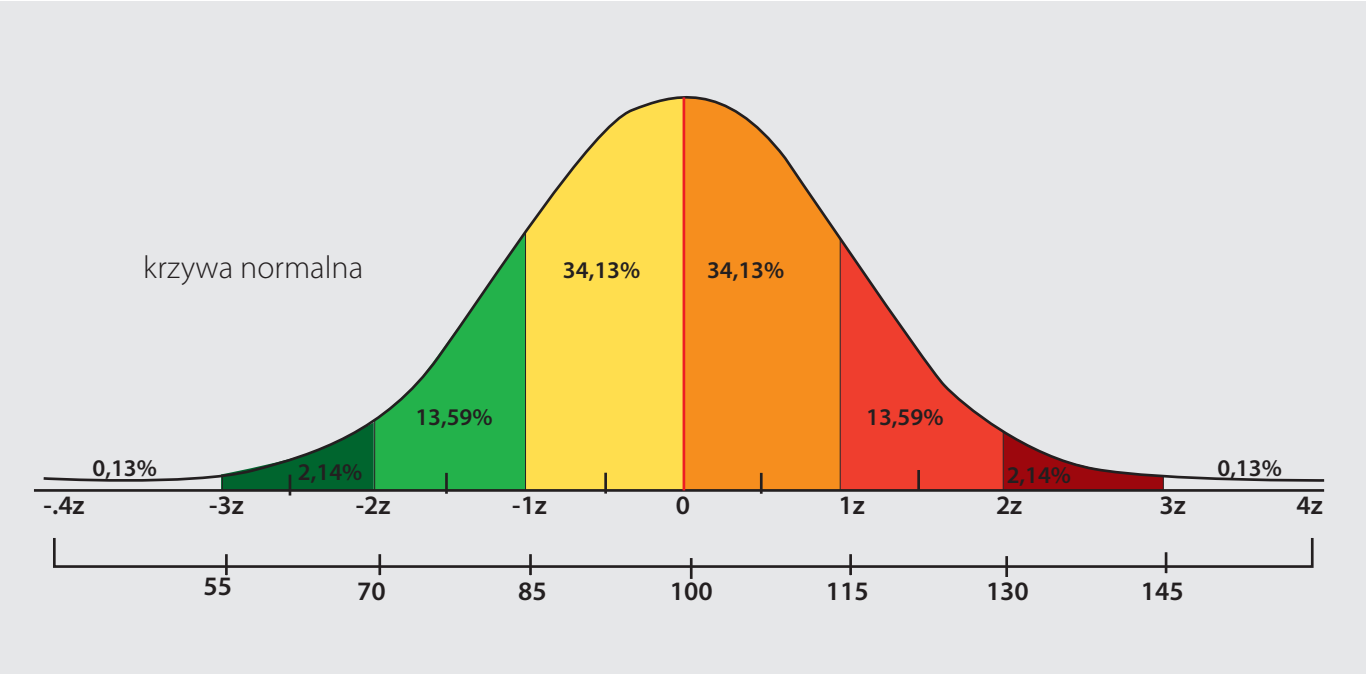


Aby statystyczne modelowanie EWD dało jak najlepsze rezultaty, a wyniki miały ustaloną interpretację ilościową, konieczne jest odpowiednie przygotowanie danych z testowania. Wyniki surowe zostały przekształcone tak, by rozkład był maksymalnie zbliżony do normalnego, i następnie przełożone na skalę standardową o średniej 100 i odchyleniu standardowym równym 15.

Jak widzieliśmy, rozkłady wyników testu OBUT 2011 i sprawdzianu 2014 bardzo odbiegają od normalnego. Konieczne jest zatem takie wyskalowanie obu testów, by maksymalnie zbliżyć rozkład do założonego. W tym celu wykorzystano jedną z nowoczesnych metod skalowania testów – model Rascha[49]. Dzięki temu można w prosty sposób przełożyć wynik surowy testów na skalę dającą rozkład bardziej zbliżony do normalnego. Zastosowanie nowoczesnej metody skalowania pozwoliło równocześnie sprawdzić pomiarową jakość testów użytych w OBUT 2011 i sprawdzianie 2014.

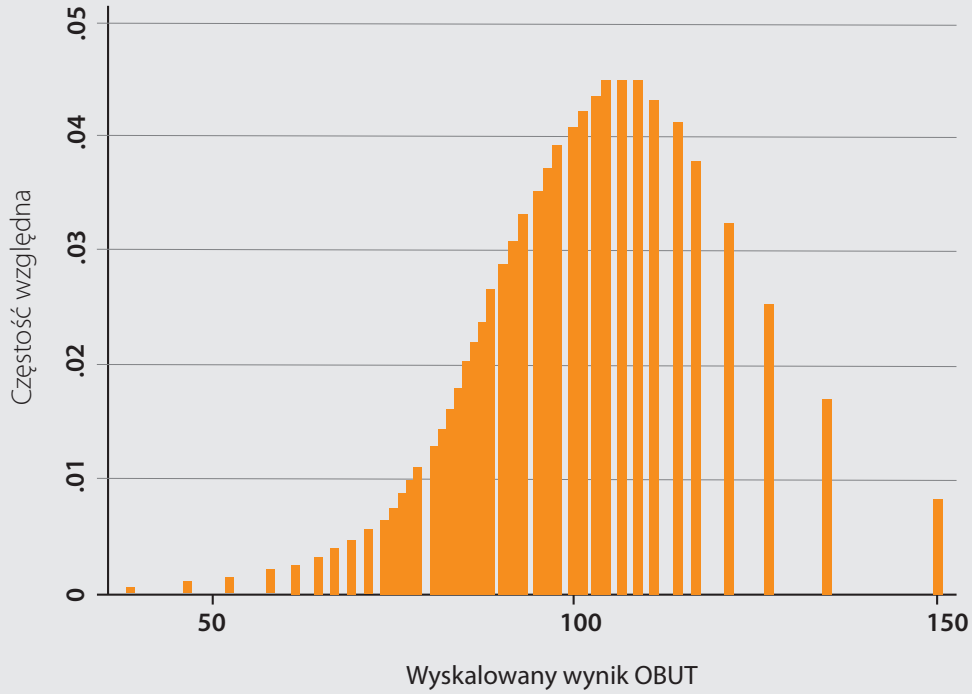
[49] Więcej informacji na temat tej metody skalowania testów zainteresowany Czytelnik może znaleźć w pracy Aleksandry Jasińskiej i Michała Modzelewskiego (2014) oraz w książce pod redakcją Artura Pokropka (2015).

Rysunek 4.15. Krzywa normalna i skala standardowa 100; 15



Jak przedstawia to rysunek 4.15. przekształcenie wyników do rozkładu normalnego i wyrażenie na skali o średniej 100 i odchyleniu standardowym 15 pozwala w łatwy sposób odnieść wartość na skali do odsetka osób posiadających wskazany poziom umiejętności. Dla przykładu, wiemy, że jeśli uczeń uzyskał wynik równy 100, to połowa uczniów w grupie odniesienia uzyskała wynik niższy, a druga połowa wyższy. Jeśli uczeń uzyskał wynik 115, to gorzej test napisało 84,13% rówieśników, a lepiej 15,87%. Taki sposób komunikowania wyników sprawia, iż nie musimy pamiętać, jaka była w danym roku średnia liczba punktów zdobytych na teście OBUT czy na sprawdzianie, co roku będzie to 100. Odpowiedzi wszystkich uczniów biorących udział w badaniu OBUT w 2011 roku posłużyły do wyskalowania wyników z użyciem modelu Rascha. Analiza dopasowania danych do modelu pozwala przyjąć, że uzyskane w efekcie skalowania wyniki wiarygodnie odzwierciedlają poziom umiejętności uczniów. Każdemu uzyskanemu w OBUT wynikowi sumarycznemu, można przyporządkować wynik na skali o średniej 100 i odchyleniu 15 (por. rysunek 4.16).

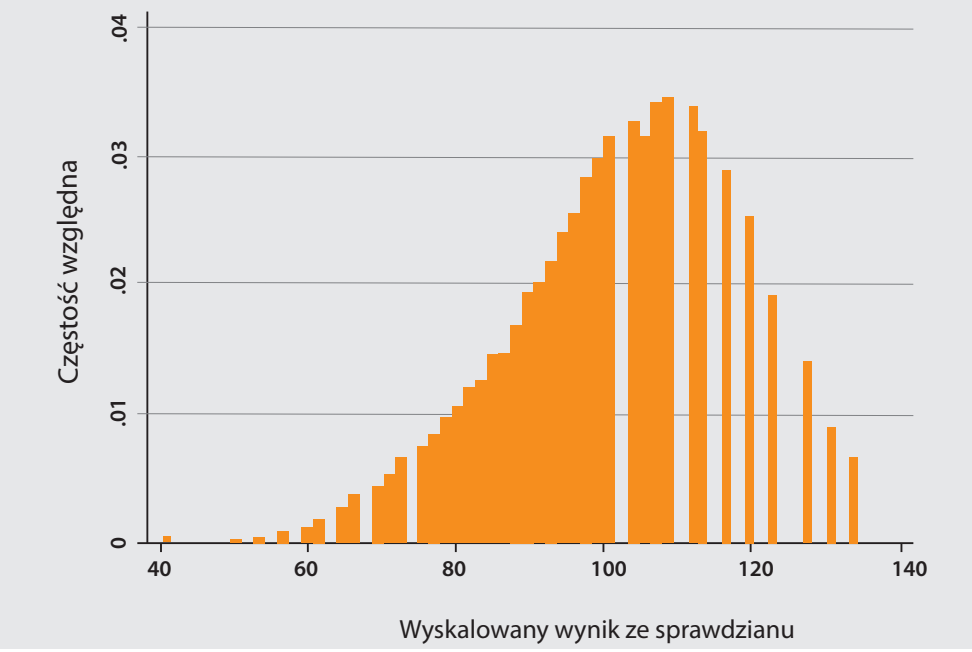
Rysunek 4.16. Przeskalowane wyniki testu OBUT wyrażone na skali 100; 15



Niestety silny efekt pułapu sprawił, że na skali 100;15 widzimy w strefie wyników wysokich dużą „nieciągłość" skali.

Podobnie jak miało to miejsce z wynikami testu OBUT, również wyniki sprawdzianu dla całej populacji uczniów rozwiązujących arkusz standardowy wyskalowany został z wykorzystaniem modelu Rasha. Dopasowanie danych do modelu okazało się dobre. Podobnie jak miało to miejsce w przypadku testów OBUT, oszacowania umiejętności uczniów przekształcono tak, aby w grupie odniesienia (wśród uczniów piszących arkusz standardowy sprawdzianu w 2014 roku) średnia równa była 100, a odchylenie standardowe 15. Przekształcone wyniki sprawdzianu na standardowej skali 100; 15 przedstawia rysunek 4.17.

Rysunek 4.17. Przeskalowane wyniki sprawdzianu wyrażone na skali 100; 15



Podobnie jak w wypadku testu OBUT, w wypadku wyskalowanych wyników sprawdzianu 2014 w strefie wyników wysokich widzimy pewną nieciągłość, ale luki są na szczęście mniejsze.

Na potrzeby wyliczenia modelu EWD potrzebne są połączone dla całej populacji uczniów (precyzyjnie mówiąc – grupy odniesienia) wyniki testu OBUT 2011 i sprawdzianu 2014. Ze względu na organizację badania OBUT i konieczność anonimizacji krajowej bazy danych z tego badania nie było jednak możliwości połączenia wyników wszystkich uczniów piszących OBUT z ich wynikami na sprawdzianie. Z tego względu szacowanie modelu EWD przeprowadzono na danych z ogólnopolskiej, reprezentatywnej dla populacji szkół biorących udział w OBUT 2011 próby 200 szkół podstawowych. Dyrektorzy tych placówek, jako osoby posiadające dostęp do niezanonimizowanych wyników uczniów dla obu testów, poproszeni zostali o połączenie tych wyników i przesłanie ich (po anonimizacji) w formie elektronicznej do IBE. Poza samym wynikiem dyrektorów poproszono o przekazanie informacji o płci ucznia oraz zaznaczenie, czy pisząc sprawdzian, miał on opinię o dysleksji rozwojowej. W wyznaczonym terminie udało się zebrać dane ze 181 szkół.

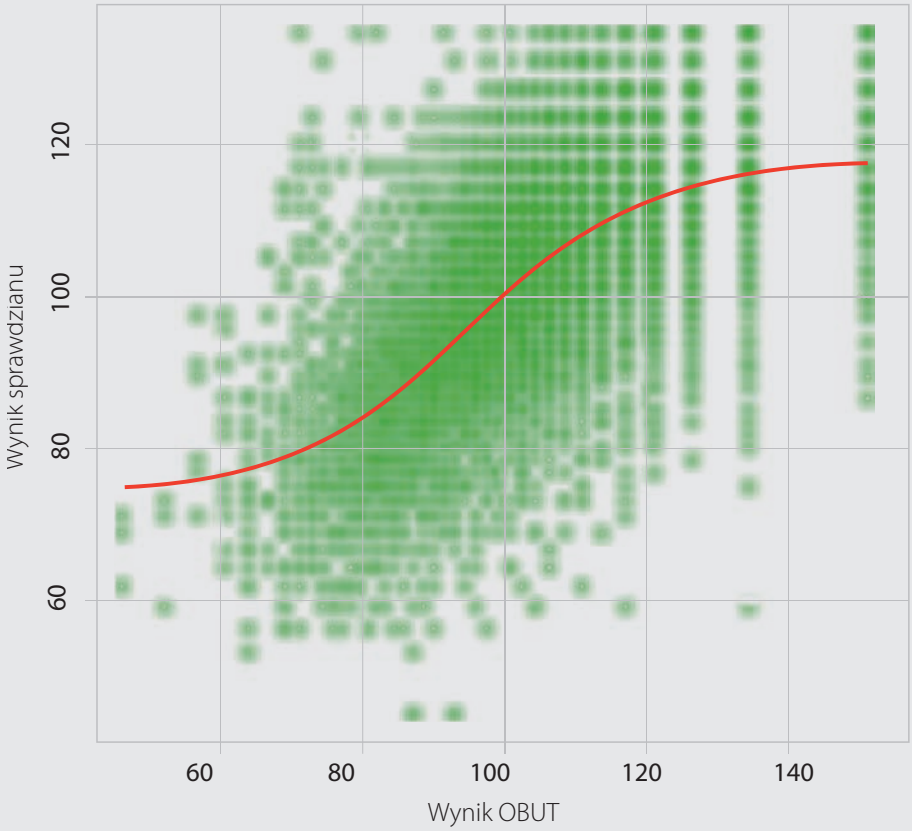
Dla oceny reprezentatywności próby zebrane dane porównano z charakterystykami populacyjnymi. Średnie wyniki uczniów w badanych szkołach okazały się nieco lepsze niż średnie w populacji. W przypadku OBUT uczniowie, których wyniki zostały przesłane przez szkoły, uzyskali średnio 101,1 punktu (na skali o średniej w populacji 100 i odchyleniu standardowym 15). Zróżnicowanie wyników, wyrażone odchyleniem standardowym wyniosło 15,4. Dla sprawdzianu średni wynik w próbie wyniósł 101,4 punktu (na skali 100; 15), a odchylenie standardowe 15,1 (por. tabela 4.2). Można zatem przyjąć, że próba dobrze odzwierciedla wyniki populacyjne.

Tabela 4.2. Porównanie wyników testu OBUT i sprawdzianu pomiędzy populacją a próbą wylosowaną do przygotowania modelu EWD.

Pomiar	Grupa	Średnia	Odchylenie st.
OBUT 2011	populacja	100,0	15,0
	próba	101,1	15,4
Sprawdzian 2014	populacja	100,0	15,0
	próba	101,4	15,1

Na danych z próby przygotowany został model EWD dla II etapu edukacyjnego, czyli określono, jaki jest najbardziej prawdopodobny wynik ucznia na sprawdzianie w zależności od tego, jaki wynik uczeń uzyskał na teście OBUT. Przebieg tej zależności można przedstawić graficznie za pomocą linii przewidywanego wyniku, jak to widać na poniższym wykresie.

Rysunek 4.18. Linia przewidywanego wyniku uczniów na sprawdzianie 2014 w zależności od wyniku na teście OBUT 2011



Znajdujące się na nim punkty reprezentują rzeczywiste wyniki uczniów na obu pomiarach: na osi X wynik w teście OBUT, na osi Y wynik na sprawdzianie. Punkty powyżej linii przewidywanego wyniku reprezentują uczniów, którzy na sprawdzianie uzyskali wyniki wyższe, niż można by się tego spodziewać na podstawie wyników z OBUT. I na odwrót, poniżej znajdują się wyniki niższe niż przewidywane przez model. Różnicę między wynikiem faktycznie uzyskanym przez ucznia a wynikiem przewidywanym z modelu EWD nazywamy resztą. Reszta to ważny termin, często wykorzystywany w analizach EWD. Należy zwrócić uwagę na zjawisko zwane regresją ku średniej. Im wynik w pierwszym pomiarze bardziej odbiega od średniej, in minus lub in plus, tym w drugim pomiarze przewidywany statystycznie wynik bardzie ulegał „ściągnięciu" do średniej. Na przykład, dla wyniku na teście OBUT równym 100, model regresji przewiduje na sprawdzianie trzy lata później też 100 punktów. Ale jeżeli uczeń w klasie III uzyskał 140 punktów, to w klasie VI model nie przewiduje rezultatu 140, ale znacznie bliższy średniej (z wykresu można odczytać, że ok 118). Analogiczne zjawisko obserwujemy oczywiście w wypadku bardzo niskich wyników na „wejściu".

Model EWD jest obliczany ex post i co roku może wyglądać trochę inaczej (trochę inny przebieg linii przewidywanego wyniku). Proces obliczania wskaźników EWD dla szkoły lub jakiejś innej grupy uczniów można w pewnym uproszczeniu (uproszczenie polega na nieuwzględnieniu zmiennych kontrolnych, czyli informacji o płci i dysleksji ucznia) przedstawić następująco:

(1) na podstawie wyników testu OBUT oraz modelu EWD dla każdego ucznia w szkole/grupie wyznaczamy przewidywany wynik sprawdzianu;

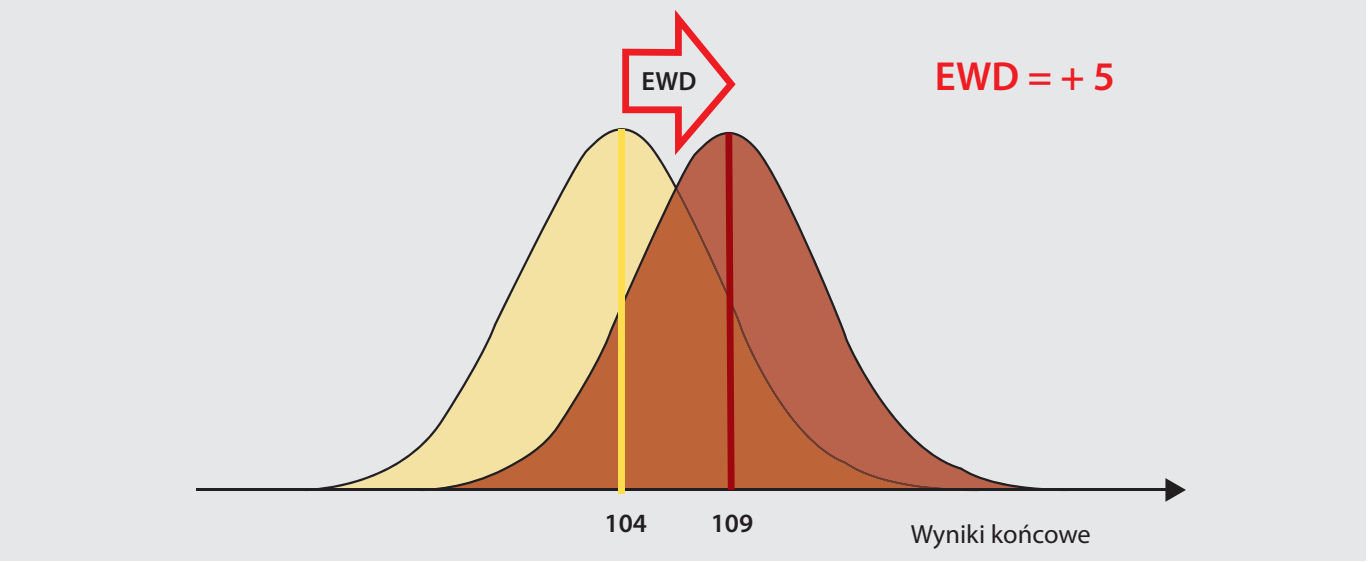
(2) liczymy dla szkoły/grupy średnią arytmetyczną wyników przewidywanych;

(3) liczymy dla szkoły/grupy średnią arytmetyczną wyników faktycznie uzyskanych na sprawdzianie;

(4) w ostatnim kroku odejmujemy od średniej wyników faktycznie uzyskanych średnią wyników przewidywanych; różnica jest wartością wskaźnika EWD.

Proces ten przedstawia rysunek na następnej stronie.

Rysunek 4.19. Ilustracja sposobu obliczania wskaźników EWD dla grupy uczniów. Po lewej rozkład wyników przewidywanych na podstawie modelu EWD, po prawej rozkład wyników faktycznie uzyskanych



Powyższy rysunek przedstawia sytuację, w której wskaźnik EWD jest dodatni, czyli efektywność nauczania w danej szkole/grupie uczniów jest powyżej przeciętnej efektywności w grupie odniesienia. Wartość liczbowa +5 oznacza, że w stosunku do rozkładu przewidywanych wyników udało się dzięki efektywnemu nauczaniu uzyskać średni wynik na sprawdzianie o 5 punktów wyższy (na skali indywidualnych wyników o odchyleniu standardowym 15 punktów, czyli przesunięcie jest znaczące).

Powyższą procedurę uzupełnia wyznaczenie dla obliczonej wartości przedziału ufności, czyli uwzględnienie statystycznej niepewności oszacowania wskaźnika EWD. Tak jak inne wskaźniki EWD, tak też miary EWD dla II etapu edukacyjnego prezentowane są zawsze z przedziałem ufności. Przypomnijmy, że szerokość przedziału ufności zależy od dwóch zasadniczych czynników: (1) liczby uczniów, na podstawie wyników których obliczamy wartość wskaźnika. Wraz ze wzrostem liczby uczniów wzrasta nasza pewność, co do tego jaka jest rzeczywista wartość wskaźnika, a więc zmniejsza się szerokość przedziału ufności; (2) zróżnicowania reszt wyliczanych na podstawie modelu EWD. Mniejsze zróżnicowanie reszt przekłada się na większą pewność, co do tego, jaka jest efektywność nauczania. W efekcie przedział ufności jest węższy.

Z zaznaczeniem przedziału ufności prezentowane są wskaźniki EWD dla szkół podstawowych w specjalnie przygotowanym do tego programie Kalkulator EWD SP.

Trafność wskaźników EWD dla II etapu edukacyjnego

Ze względu na fakt, że wskaźniki EWD dla II etapu edukacyjnego w 2014 roku zostały obliczone po raz pierwszy, mamy niewiele informacji, które mogłyby służyć jako źródło wiedzy o tym, na ile są one trafne. Porównywać możemy jedynie ich własności z charakterystykami wskaźników obliczanych od dłuższego czasu. W efekcie takich porównań odkryto, że zróżnicowanie wyników testu OBUT, które można przypisać podziałowi uczniów na szkoły (wariancja międzyszkolna), jest większe, niż można by się spodziewać, biorąc pod uwagę wyniki innych badań umiejętności prowadzonych wśród dzieci z klas trzecich (por. tabela 4.3.). Gdyby wszystkie szkoły uczyły tak samo (i dawały takie same warunki do rozwiązania testu) różnice w wynikach uczniów byłyby efektem ich indywidualnych cech. W takiej sytuacji nie byłoby wariancji międzyszkolnej (równa byłaby 0%). Jeśli wyniki w każdej szkole byłyby inne, lecz nie różniłyby się wewnątrz szkoły, to wskaźnik ten byłby równy 100%. W przypadku badania OBUT, biorąc pod uwagę wyniki wszystkich piszących, wariancja międzyszkolna równa jest niemal 18%, jeszcze większa jest wariancja w próbie: niemal 22%. Wartości te są znacznie wyższe niż uzyskane w innych badaniach, w których wariancja międzyszkolna osiągnięć szkolnych na koniec I etapu edukacyjnego wahała się w przedziale od 8,6% do 13,7% (Dolata i in., 2014; Jasińska,

Modzelewski, 2013). Może to przemawiać za hipotezą, że przynajmniej w części szkół uzyskane wyniki nie są precyzyjnym obrazem umiejętności uczniów.

Tabela 4.3. Wariancja międzyszkolna w badaniach umiejętności uczniów na koniec I etapu edukacyjnego

Dane		Odsetek wariancji międzyszkolnej
OBUT 2011	populacja	17,7%
	próba	21,9%
Badanie podłużne EWD (2012), TOS3	czytanie	8,6%[a]
	świadomość językowa	11,8%[a]
	matematyka	10,5%[a]
PIRLS 2011	czytanie	12,2%[b]
TIMSS 2011	matematyka	13,7%[b]

[a] Dolata i in., 2014;
[b] Jasińska i Modzelewski, 2013.

Biorąc pod uwagę fakt, że podwyższona wariancja międzyszkolna wyników OBUT może wynikać z uchybień w zakresie standaryzacji testowania w wykorzystaniu wskaźników EWD SP w ewaluacji wewnątrzszkolnej, należy zwrócić uwagę, iż niskie wartości EWD przy wysokich wartościach OBUT mogą wymagać pogłębionej refleksji nie tylko nad efektywnością nauczania w klasach IV – VI, ale również nad sposobem przeprowadzania i oceniania testu OBUT.

Możliwości wykorzystania metody EWD przez szkoły podstawowe

Analizy EWD dla II etapu edukacyjnego są możliwe dzięki specjalnie w tym celu przygotowanej aplikacji komputerowej Kalkulator EWD SP. Można ją bezpłatnie pobrać ze strony projektu www.ewd.edu.pl/pobierz-2/.

Aby przeprowadzić analizę dla szkoły należy, w odpowiedniej formie przygotować i wczytać do programu dane. Przykład kilku wierszy danych do analizy przedstawia tabela 4.4. Użytkownik może dodać swoje własne kolumny, np. zawierające informację o dojazdach ucznia, o tym, który nauczyciel uczył go poszczególnych przedmiotów itp. Informacje dodatkowe mogą okazać się kluczowe dla wykorzystania analiz EWD w procesie ewaluacji wewnątrzszkolnej.

Tabela 4.4. Przykład zestawienia informacji do wprowadzenia do Kalkulatora EWD SP

kod_ucznia	SPR	OBUT_pl	OBUT_mat	Płeć	Dysleksja
A1	22	20	11	M	N
A2	13	17	9	K	N
A3	35	24	16	K	N
A4	22	10	15	M	N
A5	20	20	9	K	N
A6	10	16	9	K	N
A7	21	22	14	K	N

Przejście pierwszych kroków w obsłudze Kalkulatora EWD SP i w interpretacji wyników ułatwiają krótkie filmiki instruktażowe znajdujące się na stronie http://ewd.edu.pl/lista-filmow/ oraz instrukcje „Szybki start" ze strony http://ewd.edu.pl/szybki-start-2/ i „Analizy" (http://ewd.edu.pl/analizy-2/). W materiałach tych przedstawiony został sposób importu danych z plików oraz wskazówki jak samodzielnie przeprowadzić pierwsze analizy.

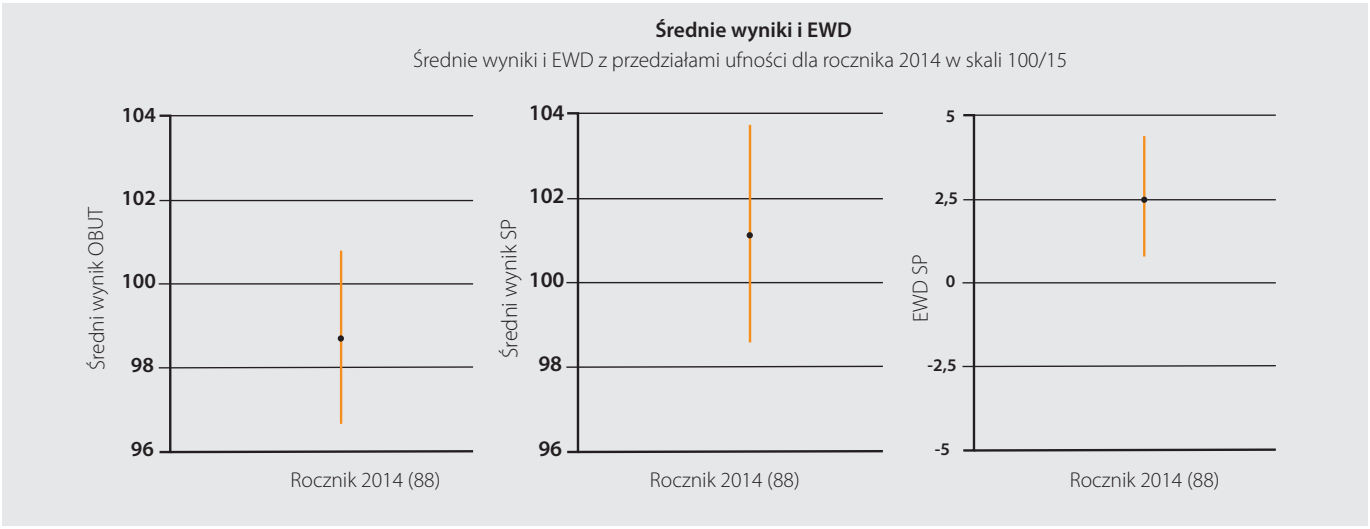
Analiza danych za pomocą kalkulatora służyć powinna przede wszystkim lepszemu zrozumieniu procesów zachodzących w szkole, zwłaszcza w zakresie nauczania przedmiotów objętych sprawdzianem. Korzystanie z Kalkulatora ułatwia wypełnianie wymagań, jakie państwo stawia wobec szkół, zwłaszcza wymagania 11. „Szkoła lub placówka, organizując procesy edukacyjne, uwzględnia wnioski z analizy wyników sprawdzianu (…) oraz innych badań zewnętrznych i wewnętrznych." Badanie wskaźników EWD jest zatem istotnym elementem procesu autoewaluacji szkoły.

Pracę z Kalkulatorem podzielić można na trzy zasadnicze części. Pierwsza to zapoznanie się z ogólnymi wynikami dla szkoły. Drugi to poszukiwanie takich grup uczniów, dla których obserwujemy istotne różnice w zakresie EWD. Trzeci to stawianie i weryfikowanie hipotez na temat takiego stanu rzeczy. Kalkulator, dzięki zestawowi kilku predefiniowanych wykresów, pozwala na graficzną prezentację analizowanych danych. Możliwe do przygotowania typy wykresów to: wyniki przewidywane, wskaźniki EWD, średnie wyniki, rozkład wyników, uprzednie osiągnięcia, rozkład reszt, wykres rozrzutu, krzywe przewidywanego wyniku. Kalkulator umożliwia również zestawianie analizowanych wskaźników w formie tabelarycznej. Wykresy i tabelę można zestawiać ze sobą w jednym widoku, co znacznie ułatwia interpretację uzyskiwanych wyników. Możne je też eksportować do dowolnego edytora tekstu.

Przykładowa analiza danych z wykorzystaniem Kalkulatora EWD SP

Analizę wyników przykładowej szkoły rozpoczniemy od zapoznania się z graficzną prezentacją średnich wyników uczniów w teście OBUT, na sprawdzianie oraz ze wskaźnikiem EWD dla szkoły. Zestawienie takich wykresów przedstawia poniższy rysunek.

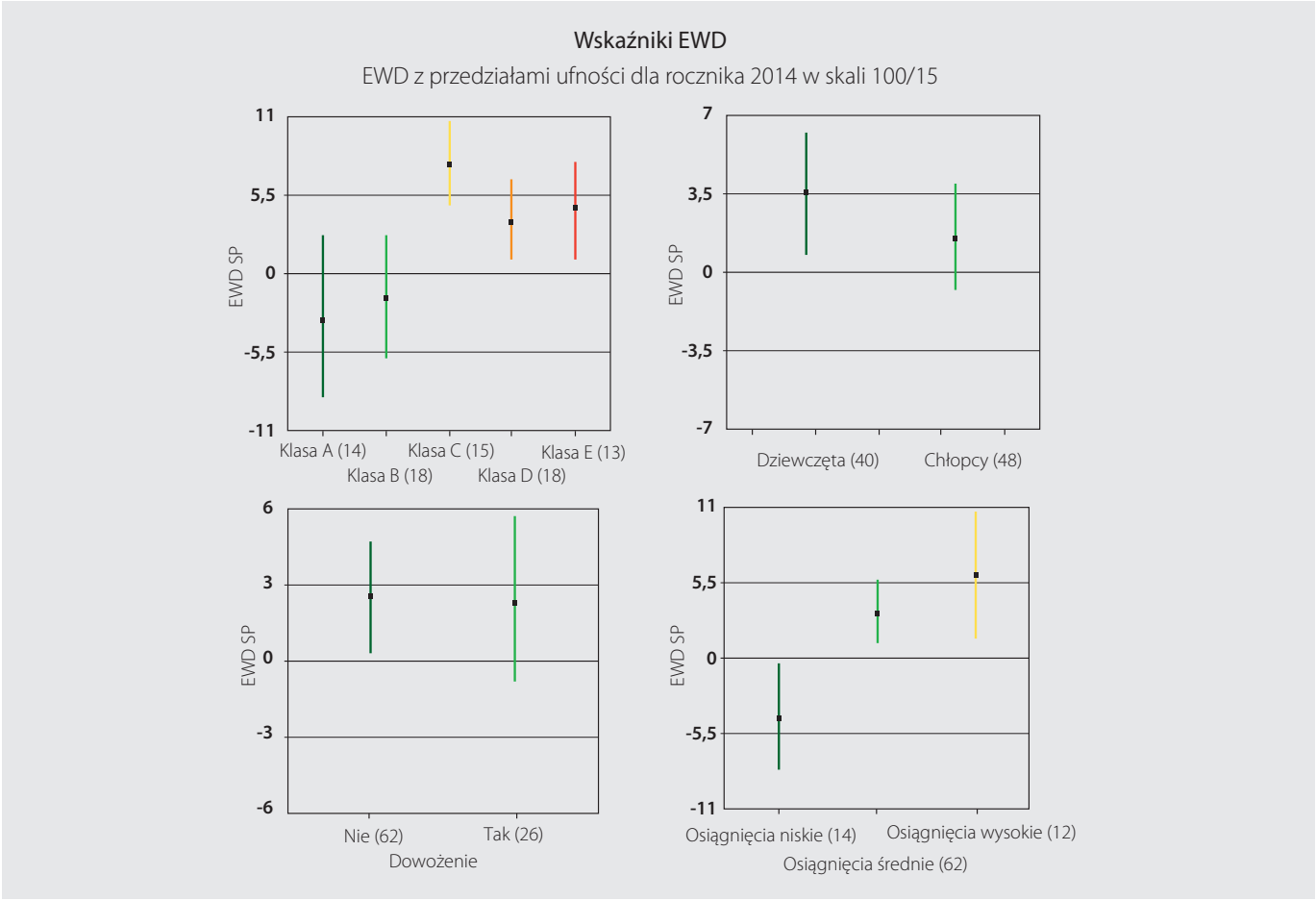
Rysunek 4.20. Średnie wyniki w szkole testu OBUT 2011 i sprawdzianu 2014 oraz wskaźniki EWD – przykład



W analizowanej szkole wynik OBUT jest niższy od średniej. Jednak przedział ufności – przedstawiony na wykresie jako pomarańczowy odcinek – przecina linię o wartości 100, co nakazuje zinterpretować wynik szkoły jako nierozróżnialny statystycznie od średniej w grupie odniesienia – w uproszczeniu można przyjąć, że w kraju. Na pisanym trzy lata później sprawdzianie w klasie VI średni wynik nieznacznie przekracza przeciętną dla kraju, choć i w tym wypadku przedział ufności nakazuje uznanie, że jest on nierozróżnialny statystycznie od średniej w kraju. Wskaźnik EWD dla tej placówki wynosi

2,5 i cały przedział ufności znajduje się ponad osią X, która pokazuje średnią wartość EWD w kraju. Można zatem powiedzieć, że efektywność nauczania w tej szkole jest ponadprzeciętna.
Analizy w Kalkulatorze pozwalają określić, czy w różnych grupach uczniów wskaźniki EWD różnią się. W tym celu można obliczyć wskaźniki EWD osobno dla każdego z oddziałów klasowych. Takie porównania można również wykonywać w podziale na płeć, czy poziom uprzednich osiągnięć. Zestawienie takich wykresów przedstawia poniższy rysunek.

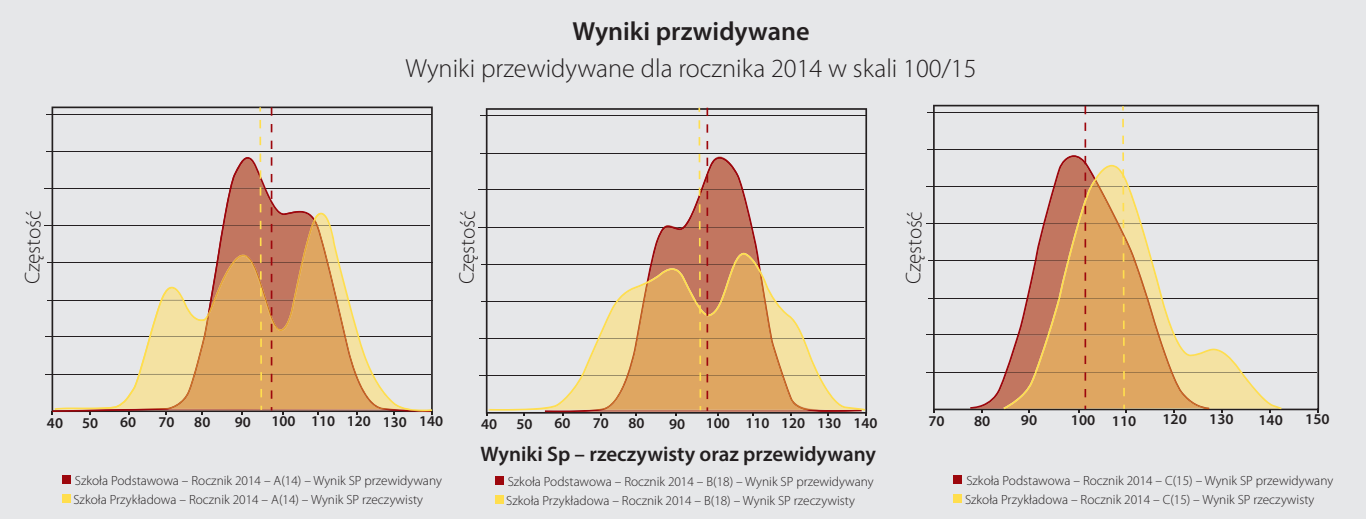
Rysunek 4.21. Wskaźnik EWD szkoły w podziale na różne grupy uczniowskie



Wykresy sporządzone dla analizowanej szkoły sugerują, że nie ze wszystkimi grupami uczniów pracowano równie efektywnie. Efektywność nauczania w obszarach objętych sprawdzianem w klasach A i B była statystycznie istotnie niższa niż w oddziale C. W grupie chłopców i dziewcząt EWD jest porównywalna, podobnie jak wśród uczniów dowożonych i niedowożonych do szkoły. To, co szczególnie zwraca uwagę, to różne wartości EWD dla grup wyróżnionych ze względu na wyniki w testach OBUT, czyli poziom uprzednich osiągnięć. Uzyskane wyniki wskazują, że efektywność nauczania w grupie uczniów o wysokich uprzednich osiągnięciach jest o ok. 10 punktów wyższa niż w grupie uczniów o niskich wynikach na teście OBUT. To bardzo duża różnica: siłę efektu możemy opisać jako 2/3 odchylenia standardowego na skali wyników indywidualnych sprawdzianu. Trzeba w tym miejscu przypomnieć, że model EWD jest tak skonstruowany, że w całej grupie odniesienia wyniki testu OBUT nie są skorelowane z wartością wskaźników efektywności.
Zaobserwowane zróżnicowanie EWD dla oddziałów i dla grup uczniów wyróżnionych ze względu na uprzednie osiągnięcia powinny w analizowanej szkole być impulsem uruchamiającym proces autoewaluacyjny. Jakie są przyczyny zaobserwowanych różnic? By ich dociec zwykle trzeba wykroczyć poza dane z testów osiągnięć szkolnych, jednak Kalkulator EWD też może być do tego pomocny.

Problemy z efektywnym nauczaniem w klasach A i B można zobrazować za pomocą wykresów wyników przewidywanych i faktycznie uzyskanych. Rysunek 4.22. przedstawia odpowiednie wykresy dla oddziałów A, B i C. W dwóch pierwszych widać wyraźnie, że wyniki faktycznie uzyskane są bardziej zróżnicowane niż wyniki przewidywane. Oznacza to, że istniały duże różnice w przyrostach dla poszczególnych uczniów, co sugeruje, że nie ze wszystkimi uczniami pracowano równie efektywnie. Dla porównania w klasie C osiągniecia uczniów były bardziej zbliżone do siebie.

Rysunek 4.22. Przewidywane i faktycznie uzyskane wyniki sprawdzianu w wybranych oddziałach



Z przedstawionej powyżej przykładowej analizy wyników OBUT i sprawdzianu w klasie VI wynika, że choć efektywność nauczania w tej szkole jest ponadprzeciętna, to są jednak obszary, w których poprawa mogłaby się przełożyć na jeszcze lepsze wyniki uczniów w szkole. Wykryte w tej szkole problemy to niska efektywność nauczania w oddziałach A i B oraz w grupie uczniów o niskich wynikach po I etapie edukacyjnym. Analiza wskaźników EWD powinna być punktem wyjścia do autoewaluacji szkoły. Prowadzić powinna do stawiania pytań i poszukiwania przyczyn zaobserwowanego stanu rzeczy oraz ich zweryfikowania na podstawie danych ilościowych (np. ocen szkolnych, danych o frekwencji), czy jakościowych (obserwacji lekcji, danych z wcześniejszych ewaluacji wewnątrzszkolnych). W takim rozumieniu korzystanie z Kalkulatora EWD SP jest sposobem na wzmocnienie ewaluacji wewnętrznej w szkole.

Ograniczenia i rozwój metody EWD na II etapie edukacyjnym

W przypadku wskaźników EWD dla II etapu edukacyjnego podstawowe ograniczenie rozwoju metody wynika z braku stabilnego, powszechnego pomiaru osiągnięć szkolnych po III klasie. Jeżeli szkoła nie bierze udziału w badaniu OBUT, nie może użyć metody EWD do analizy wyników sprawdzianu w klasie VI. Oznacza to znaczące ograniczenie przydatności wyników sprawdzianu z punktu widzenia ewaluacyjnej funkcji systemu egzaminów krajowych. Kolejnym problemem jest brak spójnej z systemem egzaminacyjnym bazy wyników badania OBUT. Sprawne łączenie wyników OBUT i sprawdzianu w klasie VI pozwoliłoby na obliczanie wskaźników EWD w oparciu o wyniki całej populacji uczniów biorących udział w pomiarze umiejętności na zakończenie I etapu edukacyjnego. Trzecie ograniczenie jest związane z niedoskonałą standaryzacją testów stosowanych w badaniu OBUT. Obecnie w sytuacji, gdy obserwowana jest niska wartość wskaźnika EWD przy równoczesnej wysokiej średniej testu OBUT, należy zastanawiać się nie tylko nad efektywnością nauczania w klasach IV –VI, ale również nad sposobem przeprowadzania i oceniania testu OBUT. Poprawa jakości testowania powinna przełożyć się na podwyższenie rzetelności i trafności wskaźników EWD SP. Jednak należy pamiętać, że testy tworzone na potrzeby OBUT mają też do spełnienia inne funkcje niż

tylko pomiar poziomu początkowego w modelach EWD. Z pewnością w kolejnych edycjach tego badania trzeba staranniej przemyśleć jego funkcje i wynikające z tego pożądane własności testów. Wydaje się, że pomimo opisanych ograniczeń wskaźniki EWD mogą być dla szkół podstawowych ważnym źródłem informacji wykorzystywanym w procesie ewaluacji wewnętrznej. Dają unikalną możliwość oparcia refleksji nad działaniami prowadzonymi w szkole na obiektywnych danych.

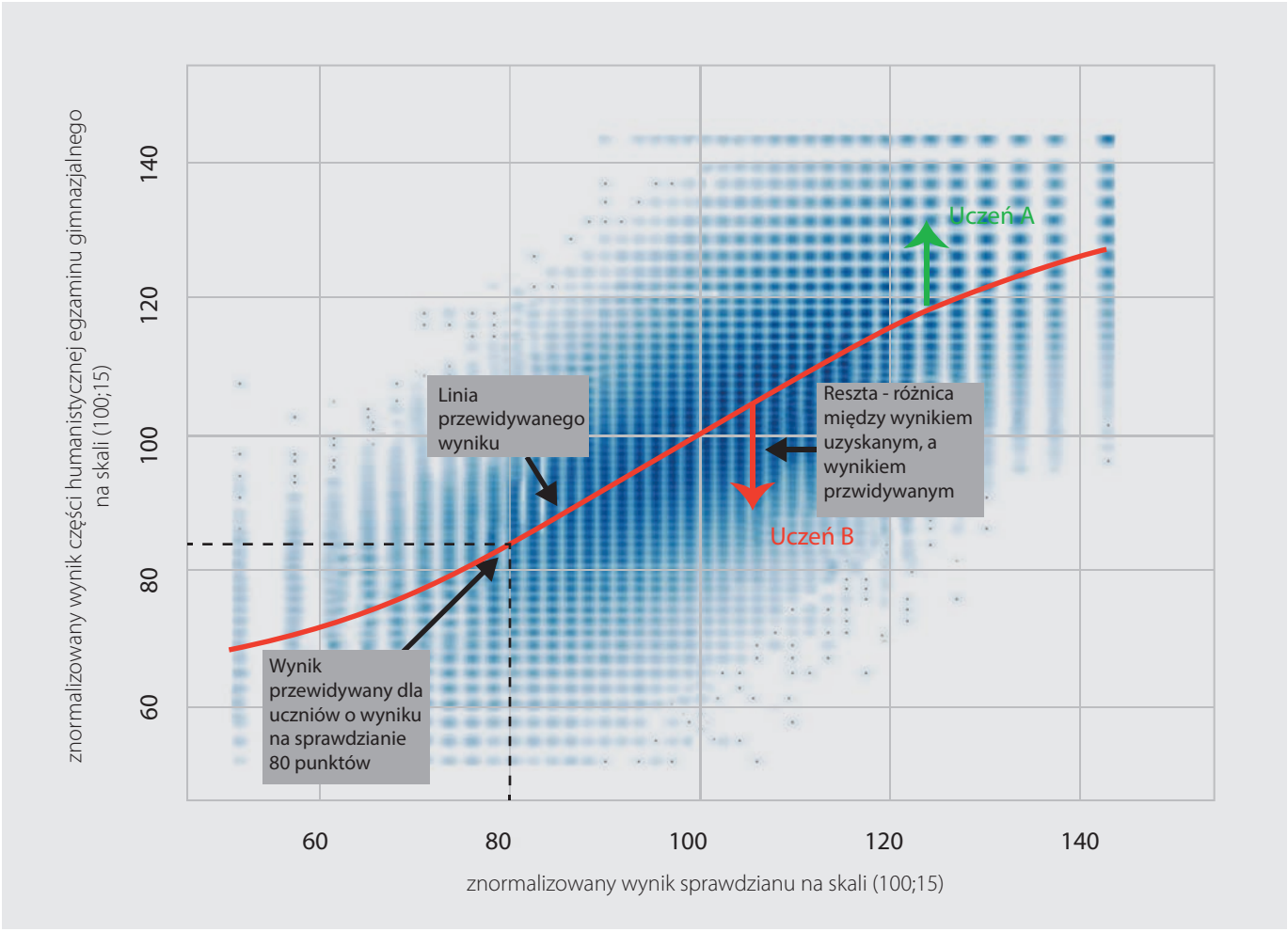
4.4. Wykorzystanie metody edukacyjnej wartości dodanej w gimnazjach

Prace nad modelami edukacyjnej wartości dodanej dla gimnazjów mają najdłuższą tradycję. Opracowaniu metodologii szacowania wskaźników EWD towarzyszyły prace nad narzędziami analizy i materiałami szkoleniowymi dla nauczycieli i dyrektorów gimnazjów oraz innych użytkowników metody. Obecnie (wiosna 2015 roku) dostępne są trzyletnie wskaźniki EWD dla ponad 6 tysięcy gimnazjów, w siedmiu trzyletnich okresach, od okresu 2006–2008 rozpoczynając. Dostępny jest również Kalkulator EWD 100, w którym można prowadzić analizy na danych egzaminacyjnych od 2009 roku.

Jednoroczne i trzyletnie modele edukacyjnej wartości dodanej dla gimnazjów

Jednoroczne modele EWD dla gimnazjów wykorzystywane są w Kalkulatorze EWD 100. Sposób obliczania wskaźników EWD przez Kalkulator EWD jest analogiczny do modelu omówionego w poprzednim podrozdziale. Statystyczną logikę metody pokazuje poniższy wykres.

Rysunek 4.23. Linia przewidywanego wyniku na egzaminie gimnazjalnym w zależności od wyniku na sprawdzianie w klasie VI. Analiza dla przykładowego rocznika



Jeśli w szkole przeważają uczniowie typu A , to efektywność nauczania jest ponadprzeciętna. Jeśli w szkole przeważają uczniowie typu B, czyli uczniowie, którzy uzyskali wyniki poniżej przewidywania, szkoła naucza z efektywnością poniżej przeciętnej, a wskaźnik EWD szkoły będzie ujemny. Wskaźniki EWD dla gimnazjów, tak jak dla szkół podstawowych, są miarami względnymi, a grupą odniesienia są gimnazja o podobnym składzie uczniowskim ze względu na uprzednie osiągnięcia.

Metodę szacowania trochę komplikuje uwzględnienie w modelu zmiennych kontrolnych. Zmienne kontrolne to te czynniki, które mogą mieć wpływ na postęp uczniów, a nie są zależne od szkoły. Jeśli chcemy, aby wskaźniki EWD pokazywały efektywność pracy szkół, takie czynniki powinny być dodatkowo kontrolowane tak, aby różnice pomiędzy uczniami uczącymi się w różnych szkołach nie miały wpływu na wartości wyliczanych dla tych szkół wskaźników EWD. Generalnie przyjmuje się, że uwzględnienie w modelach zmiennych kontrolnych pozytywnie wpływa na własności wyliczanych wskaźników EWD. Można tu jednak sformułować dwa zastrzeżenia. Po pierwsze, zakres kontrolowanych zmiennych powinien być dostosowany do celu, jakiemu mają służyć wskaźniki. Po drugie, w sytuacji nielosowego przydziału uczniów do szkół – co ma miejsce właściwie w każdym systemie szkolnym, choć z różnym nasileniem – część wpływu wynikającego z działań szkół może być w procesie estymacji modelu przypisywana oddziaływaniu zmiennych kontrolnych. Poszerzanie zakresu czynników kontrolowanych w modelu z jednej strony chroni więc przed „niesprawiedliwym" ocenianiem szkół, jednak z drugiej może utrudniać rozróżnianie od siebie szkół pracujących efektywnie i nieefektywnie (Ballou, Sanders i Wright, 2004; McCafrey i in., 2003; McCafrey i in., 2004; OECD, 2008; Raudenbush i Willms, 1995).

W gimnazjalnych modelach edukacyjnej wartości dodanej uwzględnia się dość ubogi zestaw zmiennych kontrolnych. Podstawowym problemem związanym z uwzględnianiem w modelu zmiennych kontrolnych jest słabość polskiego sytemu informacji oświatowej. W praktyce niemożliwe jest dołączenie do wyników egzaminacyjnych żadnych dodatkowych informacji o uczniach ponad te, które przechowywane są w bazach danych okręgowych komisji egzaminacyjnych. Zakres zmiennych uwzględnianych w modelach obejmuje płeć i informacje o dysleksji (podczas sprawdzianu, podczas egzaminu gimnazjalnego oraz interakcję tych dwóch zmiennych).

Włączenie do modeli płci oznacza, że jeśli w skali ogólnokrajowej występują różnice w średnich postępach dziewcząt i chłopców, to nie będą one mieć wpływu na wskaźniki EWD szkół – punktem odniesienia dla dziewcząt będą inne dziewczęta, a dla chłopców inni chłopcy. Analogiczna sytuacja występuje w przypadku informacji o korzystaniu z dostosowań dla dyslektyków, z tym że inny punkt odniesienia jest tu ustalany w ramach każdej z czterech grup: uczniów, którzy nie korzystali z takiego dostosowania, uczniów, którzy korzystali z dostosowania na sprawdzianie, uczniów, którzy korzystali z dostosowania tylko na egzaminie gimnazjalnym i wreszcie uczniów, którzy korzystali z dostosowań podczas obu egzaminów. Istnieją powody, by przypuszczać, że posiadanie opinii poradni o dysleksji nie jest zbyt dobrym wskaźnikiem występowania specyficznych trudności w nauce czytania i pisania (Dolata, 2007). Jest za to jasne, że wyniki egzaminów takich osób nie powinny być bezpośrednio porównywane z wynikami innych zdających, gdyż uczniowie korzystający z dostosowań dla dyslektyków mają wydłużony czas pisania egzaminu i stosuje się wobec nich zmodyfikowane kryteria oceny w zakresie wybranych typów błędów.

Na uwagę zasługuje również grupa laureatów konkursów przedmiotowych. Nie podchodzą oni do egzaminu (w przypadku egzaminu gimnazjalnego – do jednej z jego części, odpowiadającej tematyce konkursu, którego są laureatami), lecz mają przypisywane maksymalne możliwe do uzyskania wyniki. Jednocześnie laureaci stanowią grupę niezwykle zróżnicowaną wewnętrznie. Liczba konkursów dających możliwość zwolnienia z egzaminu jest znaczna, a do tego różna na terenie różnych województw, ponieważ organizatorami konkursów przedmiotowych są kuratoria oświaty. Brak też standaryzacji co do trudności różnych konkursów. Sprawia to, że sensowność porównywania laureatów do innych laureatów jest mocno problematyczna. W związku z tym zdecydowano się nie uwzględniać w modelach informacji o byciu laureatem, uczniowie tacy traktowani są tak samo jak inni, którzy otrzymali z egzaminu maksymalną liczbę punktów. Z punktu widzenia EWD najlepszym

rozwiązaniem problemu związanego z laureatami byłoby, gdyby pisali oni egzaminy tak, jak wszyscy uczniowie, a wyniki konkursów ujawniane były dopiero po napisaniu egzaminów. W ten sposób wyniki konkursów mogłyby być uwzględniane w rekrutacji, jednak w danych egzaminacyjnych dysponowalibyśmy również dobrym oszacowaniem bardziej ogólnych umiejętności laureatów, porównywalnym z wynikami innych uczniów.

Poza modelami statystycznymi pozwalającymi obliczać wskaźniki jednoroczne tworzone są dla gimnazjów modele wieloletnie, które obejmują wyniki trzech kolejnych roczników absolwentów gimnazjów. Wskaźniki jednoroczne i trzyletnie różnią się metodami statystycznymi, ale główną różnicę stanowi ich przeznaczenie i sposób udostępniania odbiorcom.

Tabela 4.5. Porównanie cech jednorocznych i trzyletnich wskaźników EWD.

Wskaźniki jednoroczne	Wskaźniki trzyletnie
obliczane na podstawie wyników jednego rocznika absolwentów	obliczane na podstawie wyników trzech kolejnych roczników absolwentów
dla bardzo małych szkół wskaźniki są niewiarygodne, ze względu na duży wpływ pojedynczych uczniów	mogą zostać opublikowane także dla małych szkół (więcej uczniów)
mogą podlegać dużym wahaniom z roku na rok	mało podatne na krótkookresowe wahania, ale w większym stopniu oparte na danych historycznych, nieodzwierciedlających aktualnej efektywności szkoły
pozwalają przyjrzeć się szczegółom pracy szkoły	pozwalają na bardziej ogólną ocenę pracy szkoły
dostępne tylko dla dyrektorów szkół (lub za ich pośrednictwem) – wymagają dostępu do połączonych danych sprawdzian–egzamin gimnazjalny	powszechnie dostępne
możliwość prowadzenia wielu rodzajów analiz w Kalkulatorze EWD	tylko ściśle określony zestaw wskaźników udostępniany na stronie internetowej
analizy głównie w ramach jednej szkoły	łatwe porównanie między szkołami

Jednoroczne gimnazjalne wskaźniki EWD można było wyznaczać dla szkół po raz pierwszy w 2006 r. Początkowo ich wyliczanie umożliwiały odpowiednio przygotowane arkusze kalkulacyjne, udostępniane odbiorcom pod nazwą Kalkulator EWD. Od 2009 r. udostępniana była szkołom samodzielna aplikacja służąca do wyliczania wskaźników Kalkulator EWD Plus, zastąpiona w 2012 r. przez Kalkulator EWD 100. Do 2011 r. dla każdego roku można było wyznaczać dwa wskaźniki: humanistyczny i matematyczno-przyrodniczy. W 2012 r., w związku ze zmianą formuły egzaminu gimnazjalnego, zakres wskaźników rozszerzono o cztery nowe wskaźniki, które jako miary umiejętności „na wyjściu" wykorzystują wyniki poszczególnych testów, z których składa się nowy egzamin gimnazjalny. Są to wskaźniki z zakresu: języka polskiego, historii i wiedzy o społeczeństwie, matematyki oraz przedmiotów przyrodniczych. Wszystkie jako miarę umiejętności „na wejściu" wykorzystują wyniki sprawdzianu (Pokropek i Żółtak, 2012b). Zmiany wprowadzone w 2012 r., oprócz rozszerzenia liczby wyliczanych wskaźników, objęły zastąpienie w regresji „surowych" wyników egzaminacyjnych wynikami wyrażonymi na skali standardowej oraz uwzględnienie uczniów, których tok kształcenia w gimnazjum był dłuższy o jeden rok (Pokropek i Żółtak, 2012b). Wykorzystanie wyników wyrażonych na standardowej skali 100;15 miało na celu zarówno polepszenie własności tak przekształconych zmiennych w modelach regresji, jak i ułatwienie interpretacji wyników, między innymi poprzez uzyskanie względnej porównywalności między latami. Wadą wyników surowych polskich egzaminów

jest bowiem zróżnicowany pomiędzy latami poziom trudności i rozkład wyników – normalizacja pozwala te problemy w znacznej mierze usunąć, co sprawia, że wyniki z różnych lat mogą być interpretowane tak samo.

Wskaźniki jednoroczne EWD przeznaczone są przede wszystkim na potrzeby ewaluacji wewnątrzszkolnej. Odbiorcom nie są przy tym udostępniane gotowe wskaźniki, wyliczone dla poszczególnych szkół, lecz aplikacja pozwalająca wyliczyć wartość EWD dla dowolnej grupy, określonej przez użytkownika na podstawie danych, wczytanych przez niego do tej aplikacji. Pozwala to na prowadzenie bardziej szczegółowych analiz nad zróżnicowaniem efektywności nauczania, jednak stawia użytkownikom dużo wyższe wymagania zarówno w zakresie technicznych umiejętności posługiwania się dostarczanym im narzędziem, jak i w zakresie interpretacji uzyskiwanych wyników. Grupa osób mogących korzystać ze wskaźników jednorocznych jest przy tym ograniczona do tych, którzy dysponują połączonymi wynikami sprawdzian–egzamin gimnazjalny dla analizowanej grupy uczniów, w praktyce głównie nauczycieli i dyrektorów.

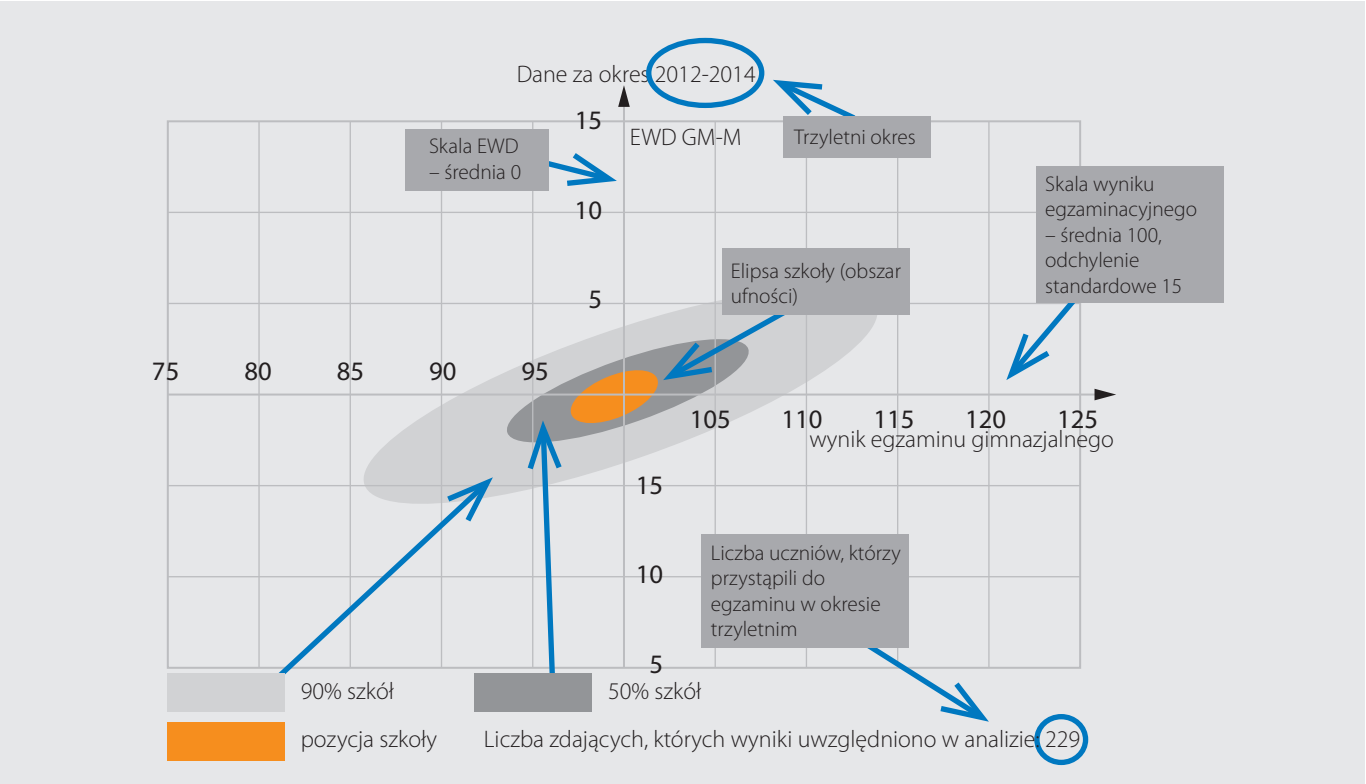
Gimnazjalne wskaźniki trzyletnie opublikowane zostały po raz pierwszy w 2009 r. Analogicznie jak w przypadku wskaźników jednorocznych były to dwa wskaźniki: humanistyczny i matematyczno-przyrodniczy. Do rozszerzenia zakresu wskaźników o cztery kolejne, odnoszące się do nowej struktury egzaminu gimnazjalnego, doszło jednak dopiero w 2014 r., gdy wszystkie trzy sesje egzaminacyjne objęte wskaźnikiem trzyletnim (2012–2014) były przeprowadzone według nowej formuły. Wskaźniki trzyletnie publikowane są w ogólnodostępnym serwisie internetowym (http://ewd.edu.pl/wskazniki/gimnazjum), z myślą o bardzo szerokiej grupie odbiorców, od nauczycieli i dyrektorów, poprzez organy prowadzące szkoły i nadzór pedagogiczny. W związku z tym zdecydowano się na wyliczanie wskaźników na podstawie wyników z trzech kolejnych sesji egzaminu gimnazjalnego tak, by były one bardziej stabilne, zdawały sprawozdanie z przeciętnej efektywności pracy szkoły w dłuższym okresie i nie prowokowały do wyciągania zbyt pochopnych wniosków na podstawie różnic, jakie mogą występować pomiędzy pojedynczymi latami.

Trzyletnie wskaźniki egzaminacyjne dla gimnazjów są łączną informacją o wyniku egzaminacyjnym i edukacyjnej wartości dodanej (EWD) i są prezentowane w formie graficznej. Przy konstrukcji trzyletnich wskaźników EWD zadbano o poprawność statystyczną prezentacji wyniku (Pokropek i Żółtak, 2012a). Do 2011 r. w modelach wykorzystywane były wyniki sprawdzianu i egzaminu gimnazjalnego w prosty sposób przekładane na skalę standardową (tzw. normalizacja ekwikwantylowa), uwzględniano przy tym wyłącznie wyniki uczniów, którzy uczyli się w gimnazjum przez trzy lata (Pokropek i Żółtak, 2012a). Począwszy od wskaźników dla okresu 2012–2014 wyniki egzaminów skalowane są w bardziej zaawansowany sposób (metodologia IRT przy pomocy modelu 2PL/graded response), a w modelach regresji uwzględniono też uczniów, których tok kształcenia był dłuższy o jeden rok. Więcej informacji na temat sposobu modelowania statystycznego trzyletnich gimnazjalnych wskaźników EWD można znaleźć w dokumentacji technicznej, dostępnej na stronie internetowej EWD (Pokropek i Żółtak, 2012b).

Trzyletni wskaźnik egzaminacyjny uwzględnia niepewność pomiarową (95% przedziały ufności) zarówno dla wyniku egzaminacyjnego, jak i dla edukacyjnej wartości dodanej, a jego reprezentacją jest elipsa (obszar ufności). Wielkość obszaru ufności zależy od liczby zdających. Pozycja elipsy jednocześnie informuje o średnim wyniku egzaminacyjnym oraz efektywności nauczania w zakresie sprawdzanym przez egzamin zewnętrzny.

Na poniższym rysunku pokazano najważniejsze elementy graficznej prezentacji trzyletnich wskaźników egzaminacyjnych. Dwie środkowe (szare) elipsy, zwane warstwicami, wskazują obszary, w których koncentruje się odpowiednio około 50% i około 90% wyników szkół. Innymi słowy, warstwice określają obszar najbardziej typowych wyników.

Rysunek 4.24. Podstawowe elementy wykresu trzyletnich wyników egzaminacyjnych/wskaźników EWD



Trafność wskaźników EWD dla gimnazjów

By wskaźniki EWD można było odpowiedzialnie polecać szkołom, musimy mieć pewność, że rzeczywiście odzwierciedlają one efektywność nauczania. Wyniki wstępnych badań, prowadzonych w latach 2005–2007, dawały powody do optymizmu, były jednak zbyt skromne pod względem wielkości próby i zakresu zgromadzonych danych, by wyciągać ostateczne wnioski. Dlatego w latach 2009–2012 wśród gimnazjalistów przeprowadzono szeroko zakrojone, podłużne badanie, którego celem była odpowiedź na pytania o niektóre aspekty trafności wskaźników EWD. Uczestniczyli w nim uczniowie, rodzice, nauczyciele i dyrektorzy szkół gimnazjalnych. Dzięki objęciu badaniem próby dobrze reprezentującej populację publicznych gimnazjów, możliwe było wnioskowanie o wszystkich tego typu szkołach w Polsce.

Ramka 4.2. Opis podłużnego badania uwarunkowań wyników nauczania w gimnazjach

Badanie podłużne w gimnazjach przeprowadzono w latach 2009–2012 i objęto nim reprezentatywną, ogólnopolską losową próbę 292 oddziałów klas pierwszych ze 150 szkół gimnazjalnych (ok. 5 tys. uczniów). Zostało ono zrealizowane w ramach projektu systemowego *Badania dotyczące rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej* (EWD), Śledzono w nim losy uczniów od pierwszej do trzeciej klasy gimnazjum, włączając do badania także ich rodziców oraz nauczycieli i dyrektorów szkół.

Głównym celem badania była weryfikacja, czy obliczane obecnie wskaźniki EWD dla gimnazjów trafnie oddają efektywność nauczania szkół. Zebrane dane dostarczyły również cennej wiedzy dotyczącej indywidualnych, rodzinnych i szkolnych czynników odpowiedzialnych za osiągnięcia szkolne uczniów. Dodatkowo w badaniu monitorowano opinie kadry pedagogicznej na temat egzaminów zewnętrznych.

Podczas pierwszego etapu badania, który odbył się, gdy uczniowie uczęszczali do pierwszych klas, zebrano m.in. dane opisujące uczniów (płeć, data urodzenia, poziom inteligencji,

wcześniejsze doświadczenia edukacyjne, aspiracje) i ich rodziny (różne miary statusu społeczno-ekonomicznego, aspiracje edukacyjne względem dzieci, struktura rodziny). Dyrektorów i nauczycieli pytano o opinie na temat egzaminów zewnętrznych. W kolejnym etapie badania, który odbył się na początku III klasy, sprawdzono osiągnięcia szkolne uczniów w zakresie czytania oraz matematyki. W etapie trzecim ponownie dokonano pomiaru inteligencji uczniów, sprawdzono ich motywację do nauki, integrację szkolną oraz zapytano o przygotowania do egzaminu gimnazjalnego i korzystanie z dodatkowej pomocy w nauce. Od dyrektorów i nauczycieli pozyskano cenne informacje nt. funkcjonowania szkoły (przywództwa, kultury organizacyjnej) oraz stosunku do egzaminów zewnętrznych i EWD. Więcej informacji na temat badania znajduje się na stronie: www.ewd.edu.pl

Istnieje wiele sposobów weryfikacji trafności wskaźników EWD. W badaniu skoncentrowano się na trzech z nich. Po pierwsze sprawdzono, czy dane wykorzystywane do ich wyliczenia, tj. wyniki testów egzaminacyjnych, są odpowiedniej jakości. Ponadto zbadano, czy modele EWD skutecznie kontrolują czynniki niezależne od szkoły, rzutujące na postępy uczniów. Tylko wtedy bowiem będziemy mogli zasadnie przypisać wartość wskaźnika EWD działaniom szkoły. Rozpatrywano czynniki indywidualne (status społeczno-ekonomiczny i inteligencja uczniów, korzystanie z korepetycji) oraz lokalne (społeczno-gospodarcze charakterystyki gmin). Po trzecie zbadano, czy szkoły o wysokim EWD, to szkoły, w których zgodnie z naszą normatywną wiedzą o funkcjonowaniu szkoły dobrze się dzieje. Dlatego zweryfikowano m.in. to, czy w szkołach o wysokim EWD uczniowie szybciej rozwijają się poznawczo. Poniżej przedstawiono wybrane wyniki analiz. Pełny opis wyników badania trafności wskaźników EWD można znaleźć w książce *Trafność metody edukacyjnej wartości dodanej dla gimnazjów* (Dolata i in.,2013).

Jeśli pomiar osiągnięć szkolnych nie jest wystarczająco dobry, nawet najbardziej wyrafinowane metody statystyczne nie dostarczą wartościowych miar. W badaniu zweryfikowaliśmy więc, czy wykorzystane egzaminy spełniają cztery podstawowe kryteria, konieczne, by można było je wykorzystać do obliczania EWD. Były to: rzetelność, odzwierciedlanie realizacji szeroko rozumianych celów kształcenia, pomiar na „wejściu" i „wyjściu" podobnych umiejętności oraz dobrze zdefiniowana jednostka pomiaru.

Przeprowadzone analizy pokazały, że rzetelność testów egzaminacyjnych jest wystarczająca dla tworzenia miar charakteryzujących szkoły. W przypadku wykorzystanego w badaniu sprawdzianu w 2009 roku wyniosła ona 0,83. W przypadku egzaminu gimnazjalnego w 2012 roku wahała się od 0,78 do 0,91 w zależności od części egzaminu. Wartości te są typowe dla wszystkich edycji zarówno sprawdzianu, jak i egzaminu gimnazjalnego.

Drugą cechą, którą powinny charakteryzować się egzaminy wykorzystywane do szacowania EWD, jest dobre odzwierciedlanie przez ich wyniki stopnia, w jakim szkoła zrealizowała najważniejsze cele kształcenia. Krytycy testowania wskazują, że egzaminy pozwalają mierzyć jedynie skuteczność rozwiązania danego zestawu zadań i że można maksymalizować wynik egzaminu przez różne praktyki nauczania opatrywane zbiorczą nazwą „nauczanie pod testy". Praktyki te, choć pozwalałyby osiągać dobre wyniki egzaminacyjne miałyby wątpliwą wartość edukacyjną, bo budowałyby niekorzystne wzorce motywacji, zawężałyby program kształcenia do tego, co sprawdzane na egzaminach, czy też ograniczałyby wpływ potencjału intelektualnego dziecka na szkolne osiągnięcia.

By zweryfikować te zarzuty, sprawdzono, czy uczniowie, którzy dobrze wypadają na egzaminie gimnazjalnym, wykazują także inne pożądane cechy, świadczące o tym, że wysoki wynik egzaminacyjny nie jest okupiony licznymi efektami ubocznymi. Wyniki analiz pokazują, że uczniowie ci otrzymują wyższe oceny w szkole, wykazują się korzystnym wzorcem motywacji (wysoka, wewnętrzna), mają pozytywne nastawienie do nauki i szkoły oraz wierzą we własne możliwości uczenia się. Zaobserwowano również negatywny związek wyników egzaminów z bezradnością intelektualną, czyli

stanem, który powstaje w sytuacji, w której uczeń mimo długotrwałego wysiłku umysłowego nie jest w stanie zrozumieć danego materiału. Ujawnił się również pozytywny, dość silny związek wyników egzaminacyjnych z inteligencją uczniów. Podsumowanie uzyskanych wyników przedstawiono w poniższej tabeli.

Tabela 4.6. Związki między wynikami egzaminów a zewnętrznymi kryteriami

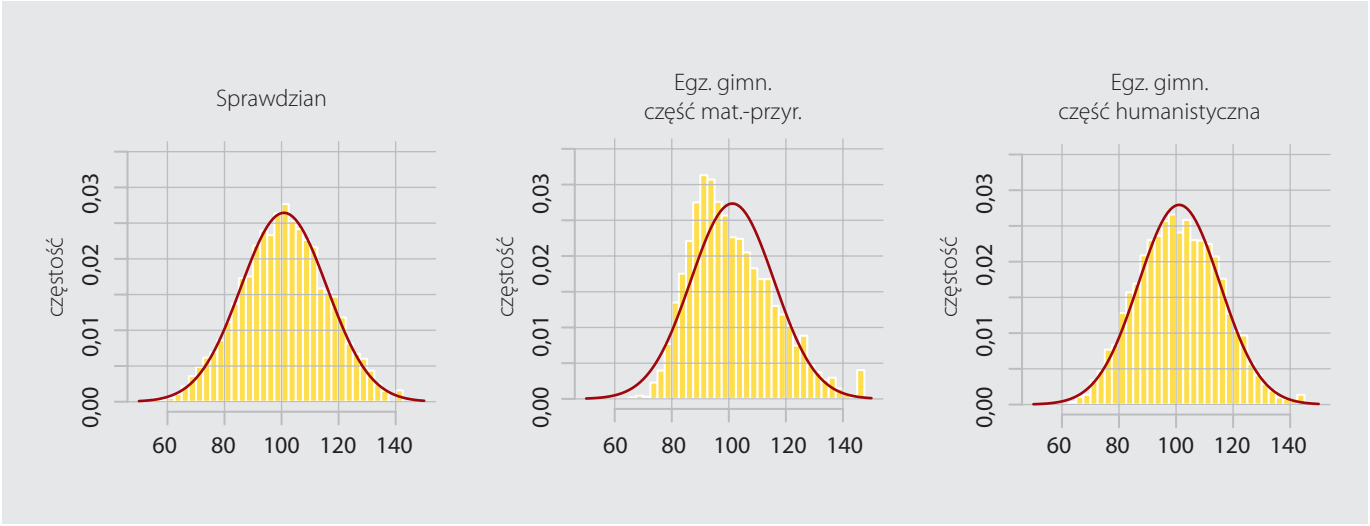
oceny szkolne	+
motywacja niska	x
motywacja wysoka	+
motywacja zewnętrzna	x
motywacja wewnętrzna	+
pozytywne nastawienie do nauki i szkoły	+
wiara we własne możliwości uczenia się	+
bezradność intelektualna	-
inteligencja	+

+ związek pozytywny; - związek negatywny; x brak związku

Wykorzystywane w modelowaniu EWD testy powinny mierzyć te same lub zbliżone umiejętności. Im będą one sobie bliższe, tym lepiej miara EWD będzie odzwierciedlała rzeczywisty wkład szkoły w osiągnięcia uczniów. Obliczając wskaźniki dla gimnazjów, wykorzystuje się wyniki sprawdzianu oraz egzaminu gimnazjalnego. Mimo iż sprawdzian jest testem wiedzy ogólnej, korelacje między nim a poszczególnymi częściami egzaminu gimnazjalnego są dość wysokie i wynoszą od ok. 0,71 do 0,78. Oznacza to, że wyniki na sprawdzianie w wystarczającym stopniu pozwalają przewidzieć wyniki na egzaminie gimnazjalnym. Innymi słowy, mogą być wartościową podstawą wyliczania wskaźników EWD.

Innym ważnym kryterium, które powinny spełniać testy egzaminacyjne wykorzystywane do obliczania EWD, jest dobrze zdefiniowana jednostka pomiaru skali, na której prezentowane są wyniki egzaminacyjne. W szczególności wzrost o jedną jednostkę na skali (np. o jeden punkt) powinien oznaczać równoważny przyrost wiadomości i umiejętności w każdym zakresie skali, tj. zarówno wśród uczniów o niskich, jak i przeciętnych oraz wysokich wynikach. W wypadku skal zdefiniowanych na podstawie odchylenia standardowego – a takie stosujemy w metodzie EWD – oznacza to, że rozkłady wyników powinny mieć kształt rozkładu normalnego. Dlatego do wyliczania wskaźników EWD nie wykorzystuje się wyników surowych. Wyniki egzaminów najpierw się skaluje metodami IRT i przekształca na skalę o średniej 100 i odchyleniu standardowym 15. Na poniższym rysunku pokazano rozkłady wyników wyskalowanych. Widoczne jest, że ich kształty są bliskie rozkładom normalnym (dla czterech testów z egzaminu gimnazjalnego mamy podobny obraz). Oznacza to, że możemy przyjąć, że wyniki testów wykorzystywane do wyliczania wskaźników EWD są przedstawione na skali o dobrze zdefiniowanej jednostce pomiaru.

Rysunek 4.25. Rozkłady wyskalowanych wyników testów egzaminacyjnych: sprawdzian 2009, egzamin gimnazjalny 2012



Podsumowując, przeprowadzone analizy pokazały, że wyniki egzaminów zewnętrznych wykorzystywane do konstrukcji gimnazjalnych wskaźników EWD mają wystarczająco dobre właściwości, by mogły stanowić podstawę wyliczania trafnej miary EWD.

Kolejnym problemem rzutującym na trafność wskaźników EWD jest wpływ na wyniki egzaminacyjne czynników pozaszkolnych, w szczególności pochodzenia społecznego, i inteligencji uczniów. Brak uwzględnienia przy wyliczaniu wskaźników EWD informacji o statusie społecznym rodziny i inteligencji ucznia jest często podnoszonym argumentem krytycznym. Wskazuje się, że status rodziny i poziom zdolności ucznia mogą wpływać na postępy w nauce, a tym samym wyznaczać szansę szkół na wysokie EWD. Jeżeli szkoły różnią się między sobą ze względu na te czynniki, to placówkom rekrutującym uczniów bardziej uzdolnionych lub z rodzin o wyższym statusie może być łatwiej uzyskać wysokie EWD niezależnie od efektywności nauczania.

Analizy pokazują, że gimnazja rzeczywiście znacząco różnią się statusem społecznym rodziców uczniów, którzy do nich trafiają. Największe różnice stwierdzono dla poziomu wykształcenia rodziców i indeksu statusu społeczno-ekonomicznego (ISEI). Około 20% zmienności tych wskaźników to zróżnicowanie międzyszkolne. Również w zakresie poziomu inteligencji uczniów zanotowano znaczne, choć mniejsze niż w wypadku wskaźników statusowych, różnice między gimnazjami. Około 10% zmienności wyników testu inteligencji to zróżnicowanie międzyszkolne.

Status społeczny i inteligencja uczniów mają znaczenie dla osiągnięć szkolnych uczniów. Spośród zbadanych aspektów statusu społecznego ucznia wyniki egzaminu gimnazjalnego najsilniej wyznacza wykształcenie rodziców. W wypadku części humanistycznej przeciętny dystans między dziećmi rodziców o wykształceniu zasadniczym zawodowym i wyższym wynosi prawie 15 punktów, czyli jedno odchylenie standardowe. Dla części matematyczno-przyrodniczej znaczenie wykształcenia rodziców jest troszkę mniejsze. Natomiast siła zależności między wynikami egzaminu gimnazjalnego a inteligencją jest prawie dwukrotnie większa niż obserwowana dla wykształcenia rodziców. Wzrost inteligencji o jedno odchylenie standardowe oznacza wzrost wyników o ok. 8–9 pkt (w zależności od egzaminu).

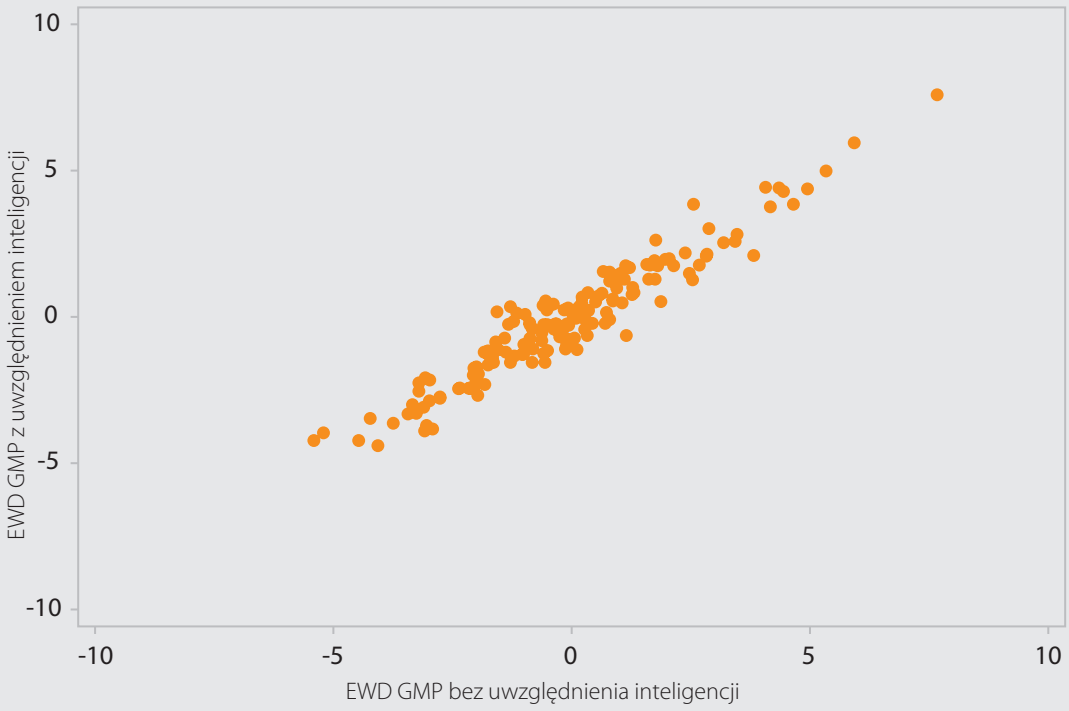
Jednakże uwzględnienie w analizie uprzednich osiągnięć, czyli wykorzystanie modelu EWD, zmienia siłę związku między czynnikami statusowymi, inteligencją a wynikami egzaminu gimnazjalnego. Siła związku wykształcenia rodziców z przyrostem umiejętności staje się ponad dwukrotnie niższa dla obszaru humanistycznego i ponad czterokrotnie niższa dla obszaru matematyczno-przyrodniczego (w porównaniu z siłą związku z wynikami egzaminu). Siła zależności między inteligencją a przyrostami umiejętności dla obszaru humanistycznego jest czterokrotnie słabsza, natomiast dla części matematyczno-przyrodniczej – dwukrotnie. Efekt ten pozostaje jednak znaczący: 4 punkty na jedno odchylenie standardowe inteligencji. Sugeruje to, że wskaźniki EWD nie do końca są niezależne od

tych cech. Dla badania trafności metody EWD ważniejsze jednak od wyników analiz na poziomie ucznia jest zróżnicowanie międzyszkolne. Okazuje się, że gdy kontrolujemy uprzednie osiągnięcia szkolne, zmienne statusowe i inteligencja mają znaczenie dla przyrostu umiejętności w analizach wewnątrz szkół, w znacznie mniejszym zaś stopniu rzutują na międzyszkolną zmienność wyników egzaminacyjnych.

Oszacowane z wykorzystaniem informacji o wykształceniu rodziców uczniów wskaźniki EWD dla szkół są niemal identyczne, jak te wyliczane na podstawie stosowanych obecnie modeli (korelacja na poziomie 0,98–0,99). Oznacza to, że brak w modelu EWD zmiennych statusowych w bardzo niewielkim stopniu obciąża oszacowania wskaźników, a tym samym ma marginalne znaczenie dla trafności metody. Podobny rezultat uzyskano dla inteligencji w obszarze humanistycznym. Trochę niższą korelację uzyskano natomiast dla obszaru matematyczno-przyrodniczego: 0,96. Oznacza to, że w skrajnych sytuacjach notujemy w tym wypadku pewne rozbieżności. Skalę problemu ilustruje poniższy wykres rozrzutu.

Podsumowując, brak uwzględnienia w obliczaniu wskaźników EWD informacji o statusie społecznym rodziny i inteligencji ucznia nie zagraża znacząco trafności metody EWD. Jedynie w wypadku egzaminu w części matematyczno-przyrodniczej brak w modelu EWD informacji o inteligencji ucznia może nieznacznie obciążać szacunki.

Rysunek 4.26. Zależność między wskaźnikami EWD dla szkół wyliczonymi bez uwzględnienia informacji o inteligencji uczniów oraz z uwzględnieniem tej informacji



Gdy myślimy o EWD jako mierze efektywności nauczania w szkole, w sposób naturalny pojawia się pytanie, czy jej wartość zależy od tego, czy uczniowie korzystają z korepetycji. Korepetycje bowiem służą lepszemu opanowaniu materiału, jednak nie stanowią wkładu szkoły w osiągnięcia ucznia. Jednak by korepetycje miały znaczenie dla EWD, musi z nich korzystać znacząca liczba dzieci. Przeprowadzone badanie pokazało, że jest to prawdą w przypadku przedmiotów matematyczno-przyrodniczych – korzystanie z korepetycji z tych przedmiotów choć raz w ciągu 3 lat nauki w gimnazjum

zadeklarowało 22% gimnazjalistów. Co istotne, odsetek uczniów korzystających z korepetycji z przedmiotów matematyczno-przyrodniczych różnił się znacząco między szkołami. W niektórych placówkach uczniowie w ogóle nie korzystali korepetycji, w innych robiło to ponad 50% z nich. Jeżeli EWD zależy od korepetycji, pierwsze z nich mogą znajdować się w niekorzystnej sytuacji, gdyż osiągnięcia szkolne ich uczniów nie otrzymują dodatkowego „zastrzyku" w postaci płatnych zajęć. Korzystanie z korepetycji z przedmiotów humanistycznych zadeklarowało niespełna 5% gimnazjalistów, skala zjawiska jest więc bardzo mała.

Przyjrzyjmy się zatem, jak korzystanie z korepetycji z przedmiotów matematyczno-przyrodniczych powiązane jest z EWD. Pamiętajmy jednak, że EWD jest cechą szkoły. Konieczna jest zatem weryfikacja zależności między odsetkiem uczniów korzystających z korepetycji z przedmiotów matematyczno-przyrodniczych a matematyczno-przyrodniczym EWD szkoły. Korelacja ta jest słaba, dodatnia i wynosi 0,19. Wiadomo jednak, że z korepetycji częściej korzystają uczniowie z rodzin o wyższym statusie społecznym. W związku z tym odsetek korzystających z korepetycji może nieść informację o statusie społecznym uczniów. Gdy uwzględnimy to w analizie, korelacja między odsetkiem uczniów pobierających korepetycje z przedmiotów matematyczno-przyrodniczych a EWD szkoły spada do zera.

W wielu badaniach wskazuje się, że czynniki lokalne wpływają na poziom umiejętności uczniów. W projekcie EWD zbadaliśmy, czy i jak poszczególne cechy otoczenia szkoły powiązane są z wynikami egzaminacyjnymi i czy wskaźniki EWD są wrażliwe na wpływ tych czynników. W analizach uwzględniono takie czynniki charakteryzujące otoczenie społeczno-gospodarcze szkoły jak: poziom bezrobocia, wielkość gminy, nowoczesność gospodarki (ilość firm high-tech z siedzibą w danej gminie w stosunku do liczby mieszkańców), powszechność opieki przedszkolnej (odsetek dzieci w wieku 3–5 lat z terenu gminy uczęszczających do przedszkola), wydatki na oświatę, zasoby kulturowe (liczba bibliotek publicznych na tysiąc mieszkańców gminy) oraz zamożność (dochody mieszkańców gminy oraz zakres i skala pomocy społecznej udzielanej mieszkańcom). Informacje te pozyskano z Banku Danych Lokalnych Głównego Urzędu Statystycznego. Dotyczyły one gmin, w których zlokalizowane są szkoły, i dotyczyły roku 2010, czyli środkowego roku dla cyklu kształcenia badanej grupy gimnazjalistów (początek nauki w gimnazjum: 2009, koniec: 2012)

Spośród analizowanych charakterystyk gmin sześć istotnie statystycznie wiąże się wynikami nauczania: wielkość gminy, gospodarka nowoczesnych technologii, powszechność opieki przedszkolnej, poziom bezrobocia, wydatki na oświatę oraz zasoby kulturowe. Trzy pierwsze okazały się skorelowane zarówno z umiejętnościami humanistycznymi, jak i matematyczno-przyrodniczymi. Wysokość gminnych wydatków na oświatę i poziom bezrobocia korelowały jedynie z poziomem umiejętności humanistycznych, natomiast zasoby kulturowe – z osiągnięciami matematyczno-przyrodniczymi. Wyniki analiz pokazały, że im bardziej nowoczesna jest lokalna gospodarka oraz im więcej dzieci objętych jest opieką przedszkolną, tym wyższe są osiągnięcia uczniów. Z kolei wyższy poziom bezrobocia związany był z obniżeniem tych osiągnięć. Znaczenie pozostałych czynników (wielkości gminy, jej wydatki na oświatę i zasoby kulturowe) okazało się statystycznie istotne, ale w praktyce nieznaczące (minimalna siła efektu).

Czy czynniki te powiązane są z efektywnością pracy szkoły? Okazało się, że ze wskaźnikami EWD, zarówno w zakresie przedmiotów humanistycznych, jak i matematyczno-przyrodniczych, istotnie powiązany jest jedynie poziom bezrobocia w gminie. Im większe bezrobocie w gminie, tym trudniej szkole uzyskać wysoką efektywność nauczania. Zależność to jest jednak w porównaniu z zależnością z wynikami egzaminacyjnymi znacznie słabsza. Oznacza to, że modele statystyczne EWD dla gimnazjów pozwalają częściowo kontrolować ten efekt. Dodatkowo dla efektywności pracy szkoły w zakresie umiejętności matematyczno-przyrodniczych znaczenie ma także wielkość gminy oraz usytuowanie w gminie przedsiębiorstw zawansowanych technologii. Oba te czynniki również znacznie słabiej oddziałują na efektywność pracy szkoły niż na same osiągnięcia uczniów. Podsumowanie wyników analiz przedstawiono w poniższej tabeli.

Tabela 4.7. Powiązanie czynników społeczno-gospodarczych z osiągnięciami uczniów i efektywnością nauczania (EWD) w zakresie przedmiotów humanistycznych oraz matematyczno-przyrodniczych

Czynnik	Humanistyczne		Mat-przyr.	
	osiągnięcia	EWD	osiągnięcia	EWD
poziom bezrobocia	-	-	x	-
wielkość gminy	-	x	-	-
przedsiębiorstwa zaawansowanych technologii	+	x	+	+
powszechność opieki przedszkolnej	+	x	+	x
wydatki na oświatę	-	x	x	x
zasoby kulturowe	x	x	-	x
zamożność gminy	x	x	x	x

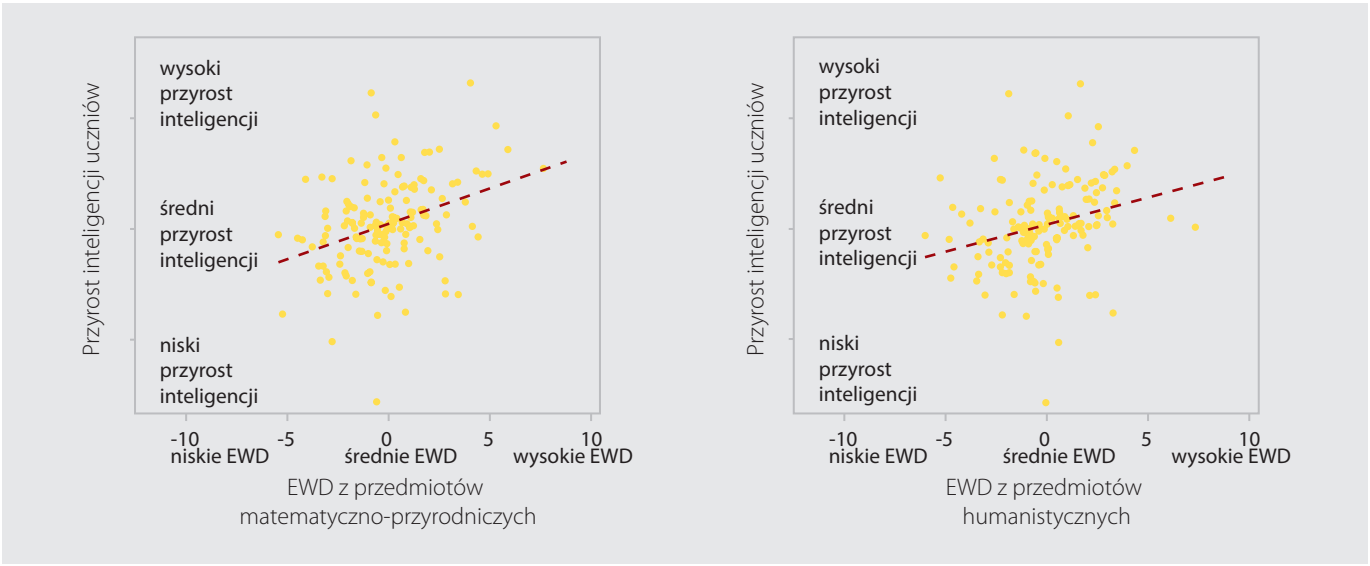
+ związek pozytywny, - związek negatywny, x brak związku

Okazuje się zatem, że choć niektóre cechy społeczno-gospodarczego otoczenia gimnazjum mają znaczenie dla wyników uzyskiwanych przez uczniów na egzaminach, to zdecydowanie mniej czynników tego typu związanych jest z efektywnością pracy szkoły, mierzoną wskaźnikami EWD. Można więc powiedzieć, że miary EWD są zdecydowanie mniej wrażliwe na wpływ środowiska społeczno-gospodarczego, w którym działa szkoła.

Krytycy EWD zarzucają niekiedy, iż wysoka wartość wskaźnika jest dowodem testomanii panującej w szkole i świadczy o silnym nacisku na przygotowanie uczniów do egzaminów, kosztem ich ogólnego rozwoju intelektualnego. Gdyby teza ta była prawdziwa, w szkołach o wysokim EWD obserwowalibyśmy wolniejszy rozwój poznawczy uczniów. Czy rzeczywiście ma to miejsce? W celu weryfikacji tego zarzutu sprawdziliśmy zależność między EWD szkoły a przyrostem inteligencji uczniów w toku nauki w gimnazjum. By możliwe było obliczenie średnich przyrostów inteligencji uczniów w każdej badanej szkole, mierzono ją dwukrotnie – w pierwszej i trzeciej klasie.

Przeprowadzone analizy pokazały, że szkoły rzeczywiście różnią się między sobą średnią wielkością przyrostów inteligencji swoich uczniów (podział na szkoły wyjaśnia ok. 15% tego zróżnicowania). Jednak zależność okazała się odwrotna od przewidywanej przez krytyków EWD. Zaobserwowano statystycznie istotną pozytywną korelację między EWD szkoły a wielkością średnich przyrostów inteligencji jej uczniów. Im wyższa wartość EWD szkoły, tym większe przyrosty inteligencji. Związek okazał się silniejszy w przypadku EWD w obszarze matematyczno-przyrodniczym (korelacja wynosi 0,32) niż humanistycznym (korelacja równa 0,24). Zależność tę zobrazowano na poniższym wykresie. Kropki reprezentują badane szkoły, niebieską linią pokazano najlepsze przybliżenie dla zaobserwowanego związku.

Rysunek 4.27. Związek EWD z przyrostami inteligencji w szkołach



Wniosek płynący z analiz zaprzecza więc wysuwanemu pod adresem EWD zarzutowi. W szkołach o wysokiej wartości wskaźników uczniowie robią nie tylko większy niż przeciętnie postęp w zakresie przyswajania wiadomości i umiejętności szkolnych, ale także szybciej rozwijają się poznawczo. Wynik ten stanowi ważki argument na rzecz tezy o trafności wskaźników.

Przeprowadzone analizy dostarczają wiele argumentów na rzecz trafności wskaźników EWD, ale wskazują też na słabsze ich strony:

- Egzaminy zewnętrzne wykorzystane do konstrukcji gimnazjalnych wskaźników EWD mają wystarczająco dobre właściwości, by mogły stanowić podstawę wyliczania trafnej miary EWD.
- Brak uwzględnienia w obliczaniu wskaźników EWD dla szkół informacji o statusie społecznym rodziny i inteligencji ucznia nie zagraża znacząco trafności metody EWD. Jednak w wypadku egzaminu w części matematyczno-przyrodniczej brak w modelu EWD informacji o inteligencji ucznia może nieznacznie obciążać szacunki.
- Gimnazja, w których uczniowie uczęszczają na korepetycje, nie mają wyższego EWD. Korepetycje nie mają znaczenia dla EWD jako miary na poziomie szkoły.
- Choć wiele cech społeczno-gospodarczego otoczenia gimnazjum ma znaczenie dla wyników uzyskiwanych przez uczniów na egzaminach, to zdecydowanie mniej czynników tego typu związanych jest z efektywnością pracy szkoły mierzoną wskaźnikami EWD.
- Powiązanie wskaźników EWD z przyrostami inteligencji dowodzi, że miara EWD nie wskazuje jedynie na to, na ile dobrze szkoła radzi sobie z przygotowaniem uczniów do egzaminów zewnętrznych. W szkołach o wysokim EWD szybciej przebiega ogólny rozwój poznawczy uczniów.

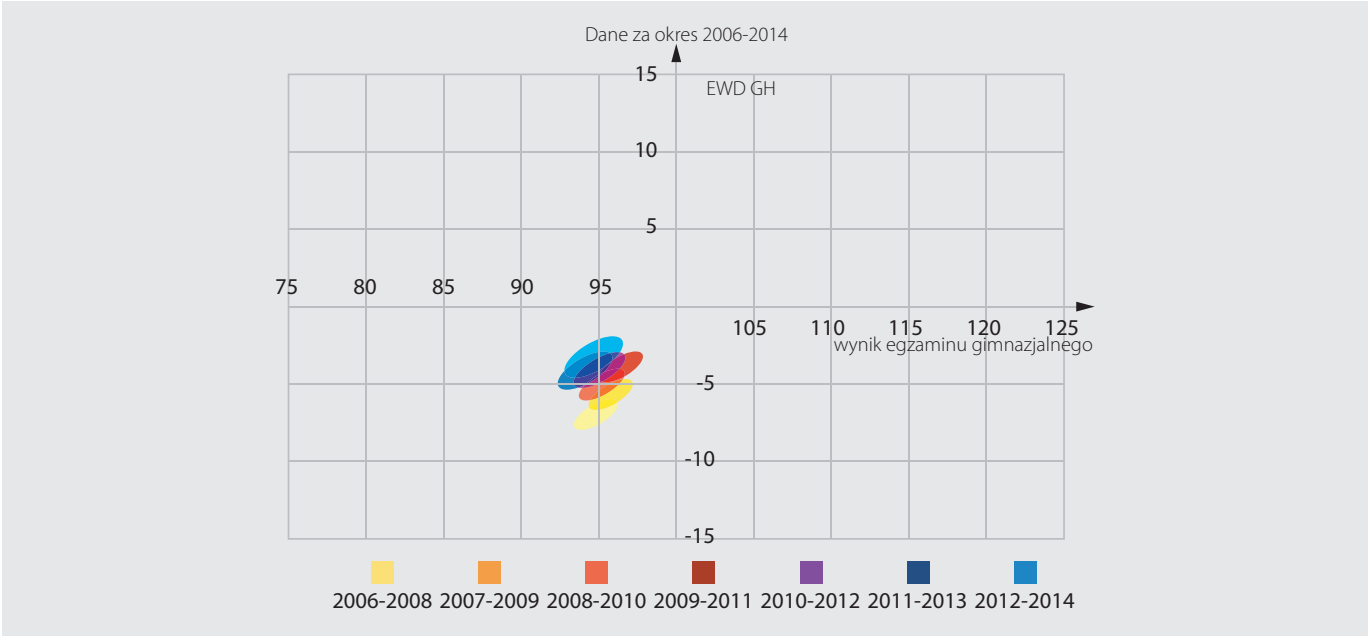
Możliwości wykorzystania przez gimnazja trzyletnich wskaźników EWD

Trzyletnie wskaźniki egzaminacyjne dają ogólny obraz wyników i efektywności nauczania w szkole. Ich przydatność dla szkół jest związana przede wszystkim z możliwością śledzenia zmian w czasie i porównywania swojej szkoły z innymi. Śledzenie zmian w czasie pozwala z większą pewnością wnioskować o zależnościach przyczynowo-skutkowych: o ile w otoczeniu szkoły nie zaszły jakieś znaczące zmiany, to zmiany w czasie wskaźników EWD można przypisać procesom zachodzącym w szkole. Możliwości wykorzystania przez szkoły trzyletnich miar EWD zilustrujemy kilkoma przykładami.

Gimnazjum A osiąga w analizowanym okresie 2006–2014 średnie wyniki egzaminacyjne niższe od średniej w kraju o około 1/3 odchylenia standardowego. Natomiast w zakresie EWD obserwujemy dość systematyczny trend wzrostowy. Oznacza to, że szkoła rekrutuje do klas pierwszych uczniów o coraz niższych uprzednich osiągnięciach, ale pracując z nimi coraz bardziej efektywnie, i uzyskuje

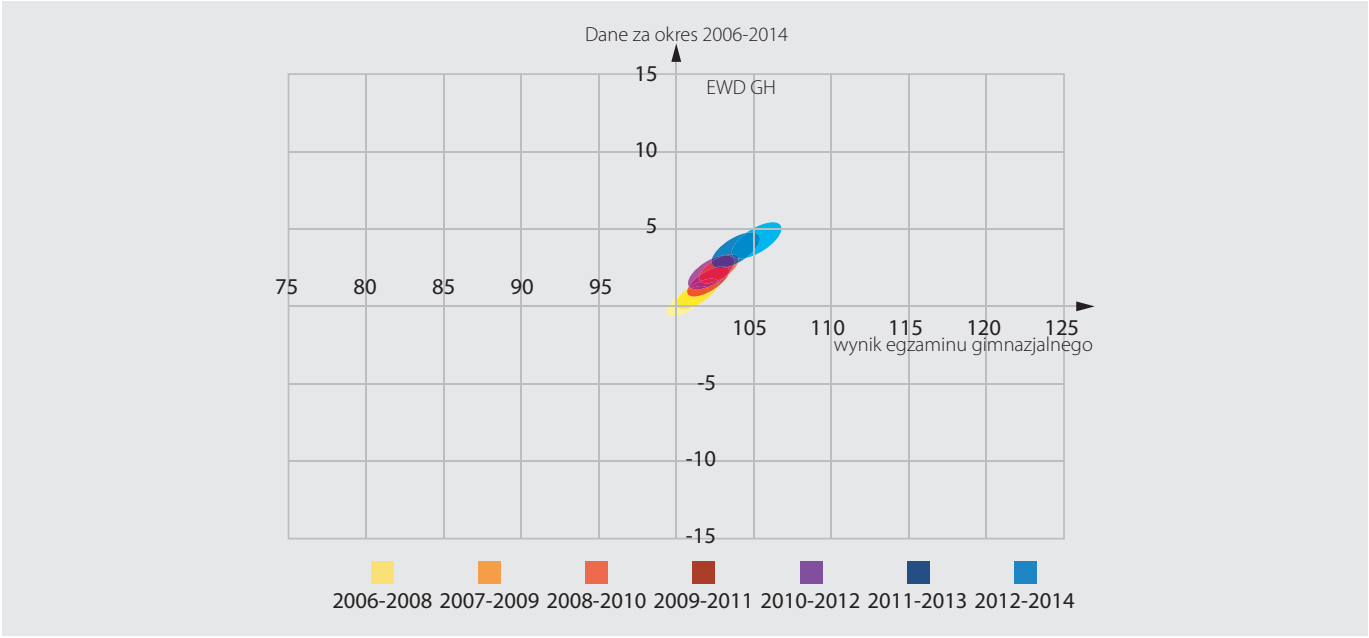
stabilne w czasie wyniki końcowe. Utrzymanie wyniku egzaminacyjnego na stałym poziomie możliwe jest tylko dzięki zwiększaniu efektywności nauczania, która wyraża się w rosnącym wskaźniku EWD. Ponadto rosnąca wielkość elipsy pokazuje, że szkoła rekrutuje do klas pierwszych coraz mniej uczniów: w latach 2006–2008 do egzaminu gimnazjalnego przystąpiło 536 uczniów, a w ostatnim okresie trzyletnim 2012–2014 już tylko 397 uczniów.

Rysunek 4.28. Zmiany trzyletnich egzaminacyjnych wskaźników dla gimnazjum A



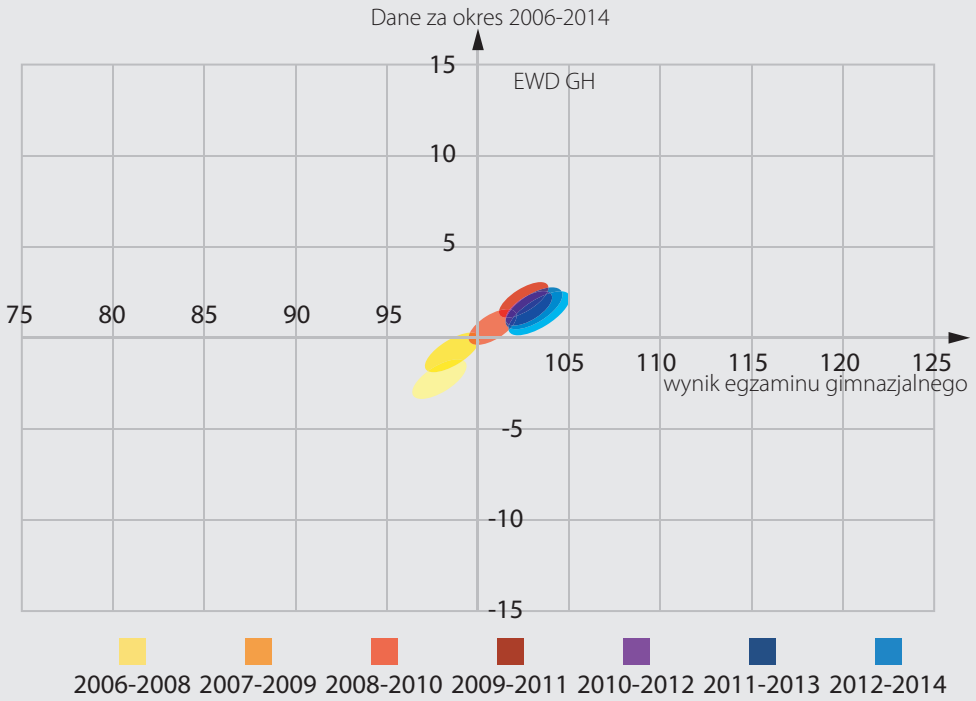
Gimnazjum B osiąga coraz wyższe średnie wyniki egzaminacyjne, które w ostatnim okresie trzyletnim (2012–2014) były wyższe od średnich wyników w kraju o 1/3 odchylenia standardowego. Jest to efektem stale rosnącej efektywności nauczania przedmiotów humanistycznych mierzonej wskaźnikiem EWD.

Rysunek 4.29. Zmiany trzyletnich wskaźników egzaminacyjnych dla gimnazjum B



Gimnazjum C osiąga ponadprzeciętne wyniki egzaminacyjne w części humanistycznej egzaminu gimnazjalnego, przy dodatnim wskaźniku EWD. W latach 2006–2011 obserwujemy wyraźny trend wzrostowy, od 2011 roku wyniki i efektywność stabilizują się na poziomie trochę przekraczającym średnią krajową.

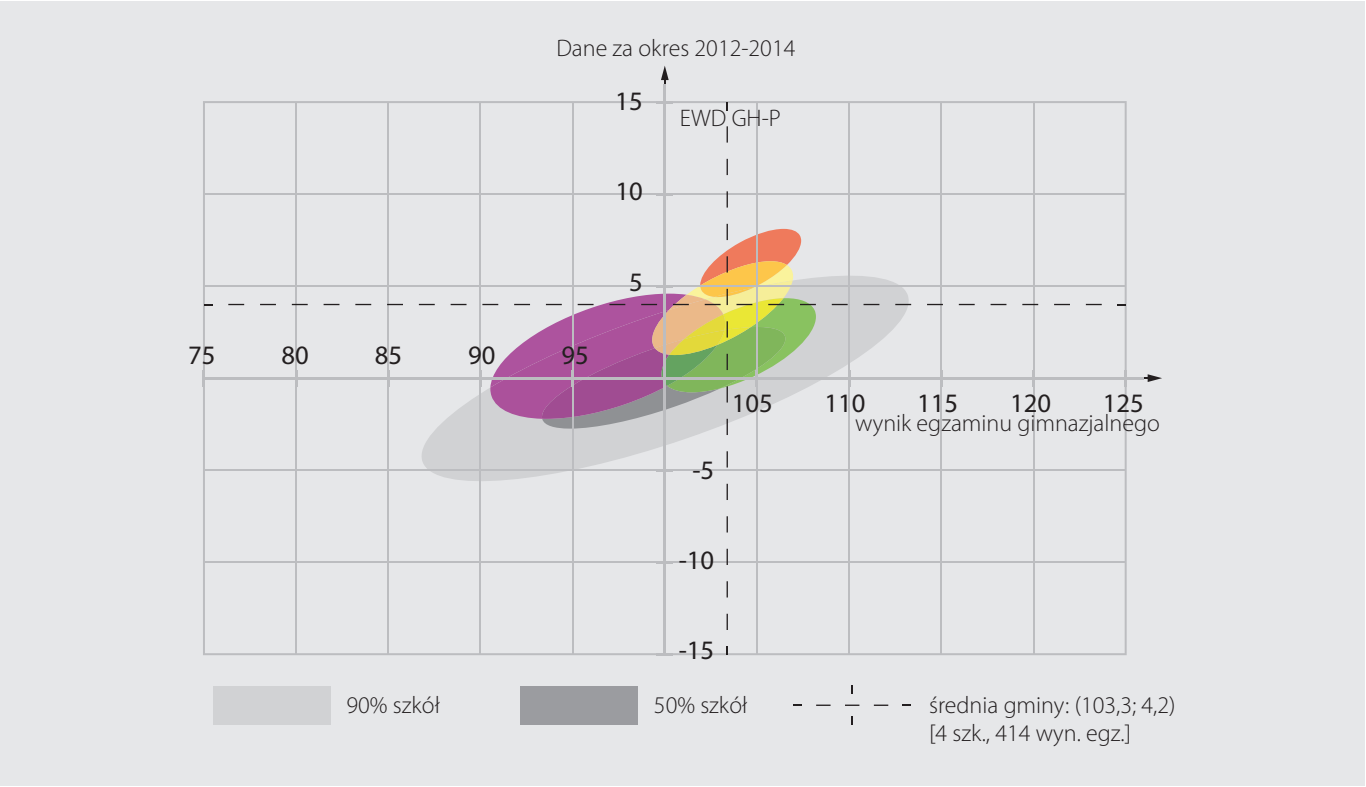
Rysunek 4.30. Zmiany trzyletnich wskaźników egzaminacyjnych dla gimnazjum C



Wskaźniki EWD są miarami względnymi. Ważne jest, aby móc określić pozycję szkoły nie tylko względem krajowego, ale także względem lokalnych układów odniesienia, takich jak gmina, powiat czy województwo. Dzięki możliwościom strony internetowej, na której prezentowane są trzyletnie wskaźniki EWD, można interaktywnie wybrać potrzebny w danej analizie punkt widzenia.

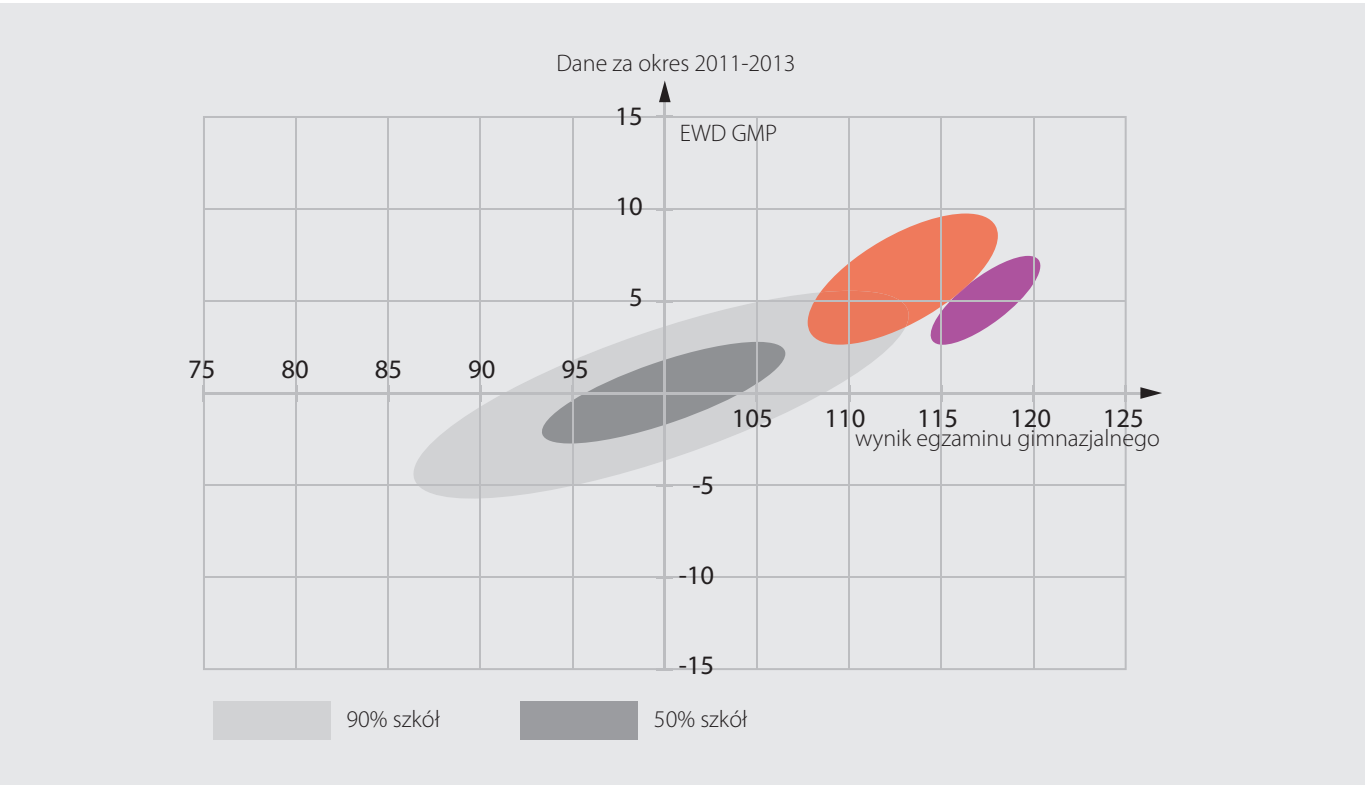
Gmina Z. Lokalny układ odniesienia dla gminy Z dla języka polskiego (zaznaczony na poniższym rysunku przerywaną linią) jest wyraźnie przesunięty w kierunku wyższych wyników i ponadprzeciętnego EWD. W tej gminie większość uczniów uzyskuje wyniki powyżej średniej krajowej, jednak różnie się to rozkłada między szkołami. Gimnazjum, którego pozycję w układzie współrzędnych opisuje fioletowa elipsa, osiąga na tle kraju przeciętne wyniki i EWD, natomiast w gminnym układzie odniesienia jego wyniki i efektywność nauczania języka polskiego mierzona EWD są statystycznie niższe od gimnazjum, którego pozycje określa pomarańczowa elipsa.

Rysunek 4.31. Lokalny układ odniesienia dla gminy Z



Gimnazja żeńskie. Nie zawsze lokalny punkt odniesienia jest najlepszy do porównań. Jeśli znamy szkoły pracujące w podobnych warunkach i chcemy się z nimi właśnie porównywać, warto zobaczyć swoją szkołę na tle tych właśnie placówek.

Rysunek 4.32. Porównanie dwóch żeńskich gimnazjów.



Na powyższym wykresie pokazano dwie żeńskie szkoły o podobnej efektywności nauczania i porównywalnych wynikach końcowych.

Możliwości wykorzystania przez gimnazja jednorocznych wskaźników EWD

Trzyletnie wskaźniki EWD wyliczane są dla całych szkół. Ten poziom analizy jest niewystarczający do zrozumienia procesów zachodzących w szkole. Do tego celu potrzebne jest narzędzie, które wykaże zmiany w krótszych niż trzyletnie okresach czasu, pozwoli wyliczyć wskaźniki EWD dla mniejszych i specyficznych dla szkoły grup uczniów, umożliwi porównanie między latami nie tylko efektów pracy szkoły jako całości, ale również efektywności nauczania w wybranych grupach uczniowskich. W dużych szkołach, o zwykle dużym zróżnicowaniu między oddziałami klasowymi (m.in. pod względem składu uczniowskiego, uczących nauczycieli lub profilu klas – klasy integracyjne, klasy sportowe) do opisu pracy szkoły lepiej nadają się jednoroczne wskaźniki.

W analizach wewnątrzszkolnych standardowo oblicza się jednoroczne wskaźniki dla oddziałów, analizuje się efekt płci, efektywność nauczania w grupach o różnych uprzednich osiągnięciach. Jeśli dysponujemy dodatkowymi informacjami o uczniach, to można je wykorzystać jako kryteria podziału na specyficzne grupy uczniowskie. Można np. analizować wskaźniki EWD w grupie uczniów korzystających z dodatkowych zajęć, uczniów dojeżdżających do szkoły, uczniów-absolwentów różnych szkół podstawowych. Ten katalog można dowolnie poszerzać i będzie on inny dla każdej szkoły. Nie chodzi bowiem jedynie o samo wyznaczenie wskaźnika EWD – ważniejsze jest wyodrębnienie grup uczniowskich o istotnie różnej efektywności nauczania mierzonej EWD. Próba zrozumienia, dlaczego te grupy różnią się efektywnością nauczania, powinna poszerzyć wiedzę o czynnikach efektywności specyficznych dla danej szkoły.

Analizy z wykorzystaniem jednorocznych wskaźników EWD są możliwe w Kalkulatorze EWD 100. Do kalkulatora zostały zaimplementowane modele jednorocznej EWD uwzględniające uczniów drugorocznych. W Kalkulatorze EWD 100 można od rocznika 2012 wyliczyć wskaźniki odpowiadające czterem arkuszom gimnazjalnym: GH-P, GH-H, GM-M, GM-P oraz syntetyczne wskaźniki dla całego obszaru humanistycznego (GH) i całego obszaru matematyczno-przyrodniczego (GMP). Dla wcześniejszych lat Kalkulator EWD oblicza wskaźniki dla części humanistycznej (GH) oraz matematyczno-przyrodniczej (GMP). Syntetyczne wskaźniki GH i GMP pozwalają porównywać wyniki od 2012 roku z wcześniejszymi rocznikami.

W Kalkulatorze EWD można w prosty sposób wykonywać różne analizy statystyczne wyników egzaminacyjnych. Można zrobić:

- **wykres rozrzutu** – pozwala na jednym wykresie zobaczyć wynik ucznia na egzaminie z poprzedniego etapu kształcenia i na egzaminie końcowym oraz krzywą przewidywanego wyniku;
- **wykres prezentujący wskaźniki EWD wraz z przedziałami ufności –** pozwala on porównać ze względu na EWD różne grupy uczniów i śledzić zmiany EWD w czasie;
- **wykres prezentujący średnie wyników egzaminacyjnych wraz z przedziałami ufności¸ –** pozwala on porównać różne grupy uczniów i śledzić zmiany w czasie;
- **wykres procentowy skumulowany dla uprzednich osiągnięć szkolnych** – pozwala porównywać między sobą grupy uczniów ze względu na uprzednie osiągnięcia;
- **rozkład wyników przewidywanych i rzeczywiście uzyskanych –** wykres jest jednym ze sposobów graficznej prezentacji wskaźników EWD;
- **rozkłady reszt** – pozwala porównać zróżnicowanie reszt w szkole z rozkładem krajowym.

Uzupełnieniem powyższych narzędzi graficznych jest możliwość sporządzania tabel, które mogą zawierać wiele wskaźników liczbowych, dotyczących zarówno EWD, jak i wyników egzaminacyjnych. Kalkulator EWD 100 można pobrać bezpłatnie ze strony http://ewd.edu.pl/kewd100/pobierz

Analizę wyników egzaminacyjnych wzbogaconą o wskaźniki EWD można wykorzystać w szkole zarówno na etapie opisu (jak jest?), jak i na etapie identyfikowania obszarów do zmiany (jak powinno być?) i/lub określenia przedmiotu ewaluacji wewnętrznej. W szkole podejmuje się różnorodne działania, które mają sprzyjać poprawie efektywności nauczania – naturalnym sposobem weryfikacji

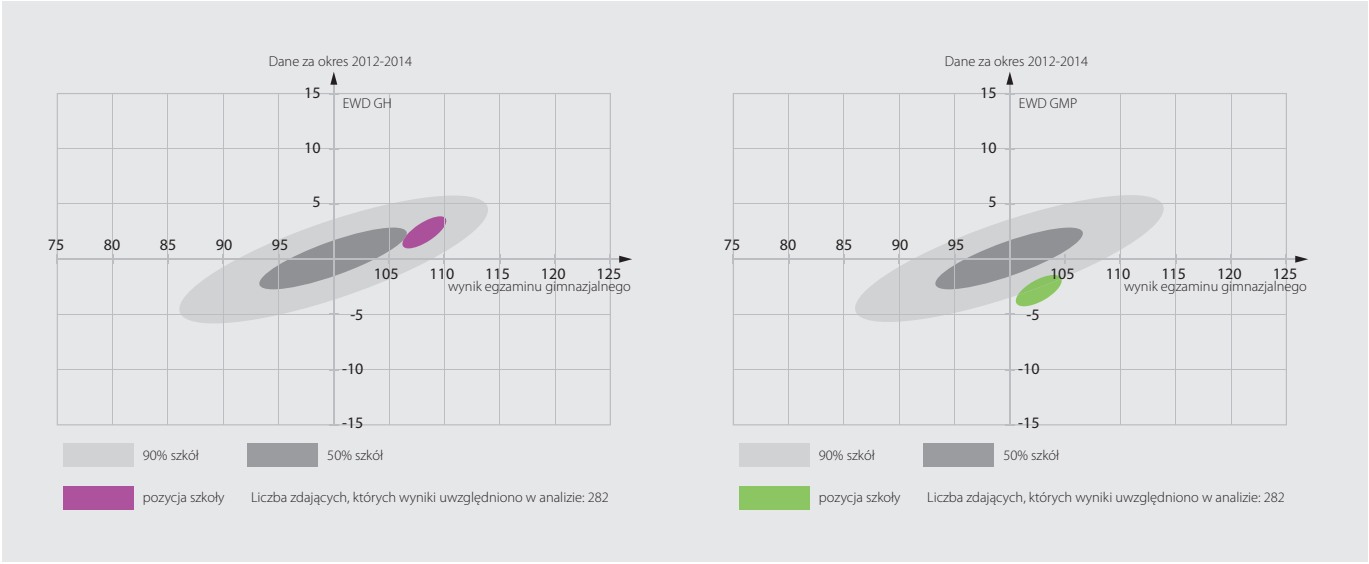
skuteczności tych działań jest analiza wskaźników EWD. W tym przypadku należy pamiętać o tym, że efekty podejmowanych działań rzadko kiedy będą widoczne w krótkim okresie czasu.

Można dokonywać zarówno analiz w czasie, jak i analiz przekrojowych w podziale na różne grupy uczniowskie. W analizach poszukujemy takich grup, które statystycznie różnią się wskaźnikami EWD, bo pozwoli to poszukiwać czynników wpływających na efektywność nauczania. Jeśli kryterium podziału jest czas, stawiamy pytanie: Co spowodowało wzrost lub spadek wskaźników EWD w czasie? Jeśli kryterium podziału jest grupa uczniowska, to stawiamy pytanie: Dlaczego dany podział generuje różnicę we wskaźnikach EWD?

Prześledźmy możliwości tkwiące w analizach EWD na przykładzie dwóch gimnazjów: Gimnazjum SZ oraz Gimnazjum P. Więcej przykładów zastosowania wskaźników EWD do analiz wewnątrzszkolnych można znaleźć w materiałach pomocniczych przygotowywanych dla dyrektorów i nauczycieli gimnazjów (Stożek, 2008, 2010).

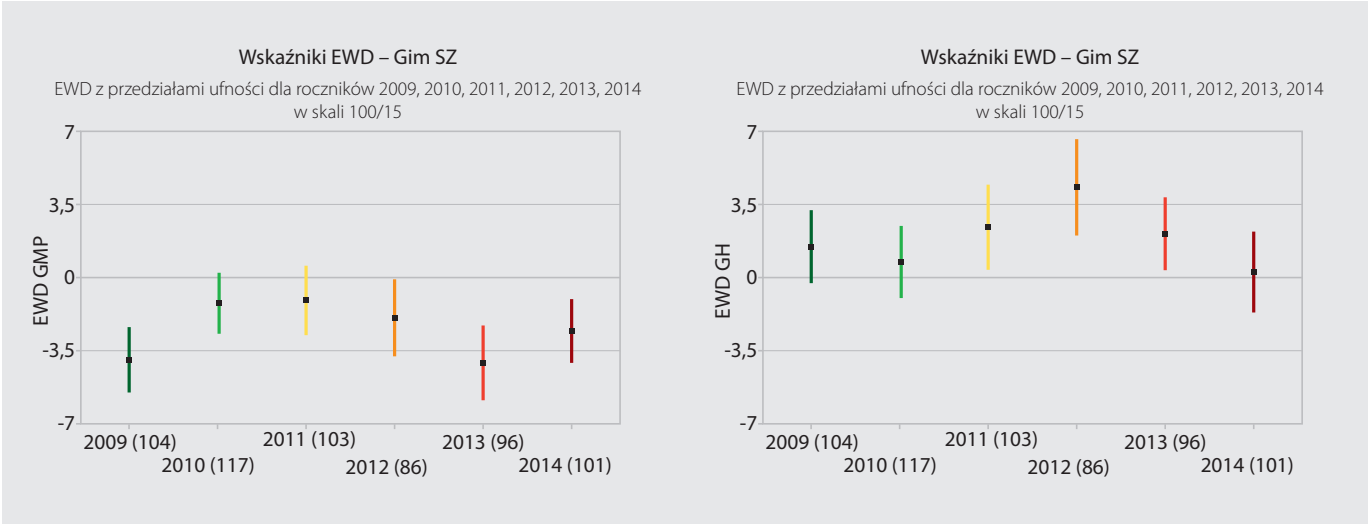
Gimnazjum SZ. Uczniowie gimnazjum SZ uzyskują na egzaminach gimnazjalnych ponadprzeciętne wyniki: w trzyletnim okresie 2012–2014 w części humanistycznej są one o ok. 1/2 odchylenia standardowego wyższe od średniej krajowej, natomiast w części matematyczno-przyrodniczej przekraczają średnią o ok. 1/5 odchylenia standardowego. Jak pokazują poniższe wykresy EWD w obszarze humanistycznym jest dodatnie, ale w zakresie przedmiotów matematyczno-przyrodniczych jest ujemne. Znacząca różnica we wskaźnikach EWD wskazuje na niską efektywność nauczania w zakresie przedmiotów matematyczno-przyrodniczych i powinna być powodem do niepokoju.

Rysunek 4.33. Trzyletnie wskaźniki egzaminacyjne dla gimnazjum SZ za lata 2012–2014



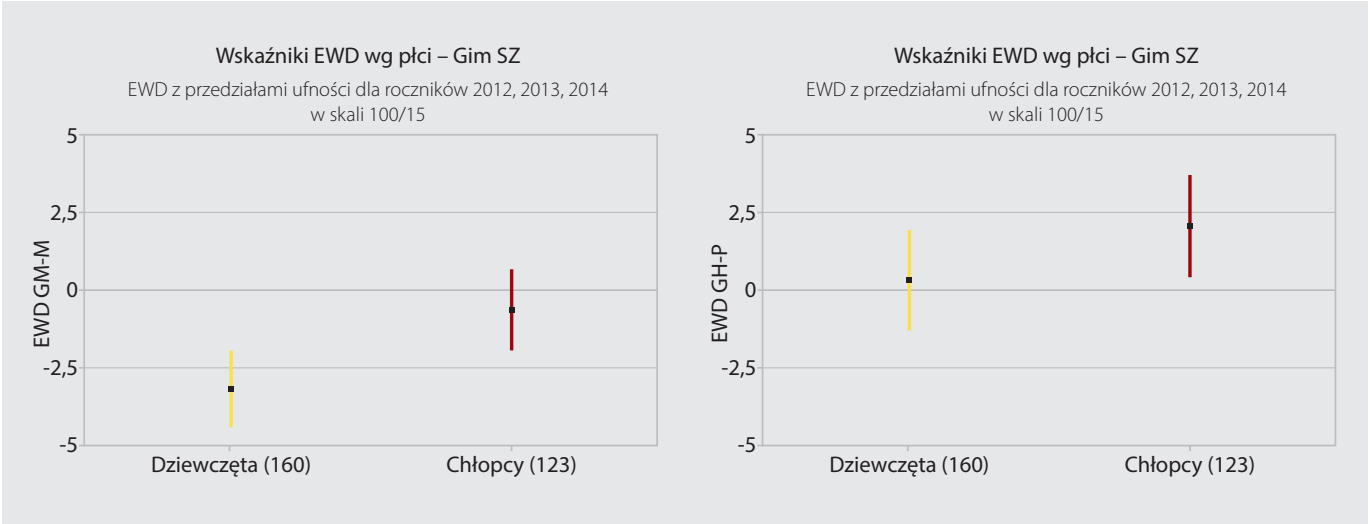
Bardziej szczegółowy obraz można uzyskać dzięki analizom wskaźników jednorocznych wykonanym za pomocą Kalkulatora EWD. Znaczącą (i utrzymująca się w czasie) różnicę w efektywności nauczania przedmiotów matematyczno-przyrodniczych i przedmiotów humanistycznych potwierdza analiza jednorocznych wskaźników EWD. Warto zatem przemyśleć w szkole czynniki, które mogą mieć na to wpływ.

Rysunek 4.34. Jednoroczne wskaźniki EWD za lata 2009–2014 dla gimnazjum SZ



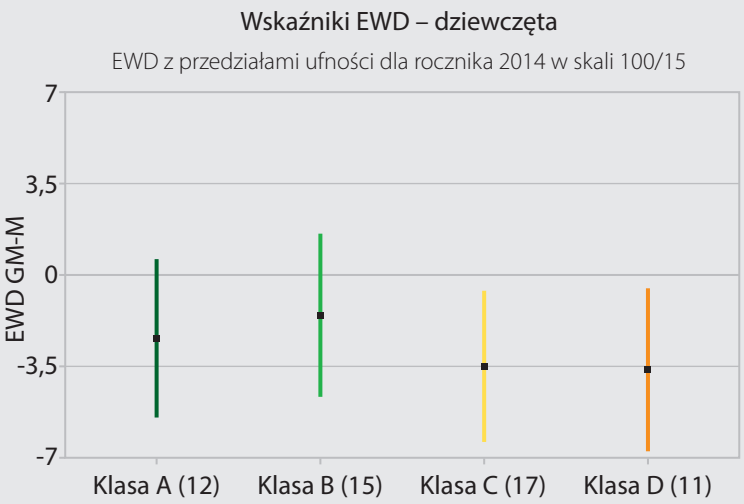
Kalkulator EWD zasadniczo przeznaczony jest do analizowania wskaźników jednorocznych, ale można obliczać wskaźniki z kilku lat. Tak zbudowany wskaźnik EWD dla matematyki za lata 2012–2014 pokazuje znaczącą statystycznie różnicę między efektywnością nauczania dziewcząt i chłopców, natomiast w przypadku wskaźnika EWD dla języka polskiego różnica ta nie jest istotna statystycznie. Efektywność nauczania matematyki w przypadku dziewcząt jest niska, znacząco poniżej średniej krajowej. Dla chłopców natomiast efektywność nauczania matematyki jest bliska przeciętnej.

Rysunek 4.35. Wskaźniki EWD dla dziewcząt (K) i chłopców (M) w gimnazjum SZ za lata 2012–2014



Wróćmy do danych tylko z 2014 roku. Poniższy wykres pokazuje EWD matematyczne dla dziewcząt, ale w rozbiciu na oddziały.

Rysunek 4.36. Jednoroczne wskaźniki EWD dla dziewcząt w podziale na oddziały klasowe w gimnazjum SZ w 2014 roku.

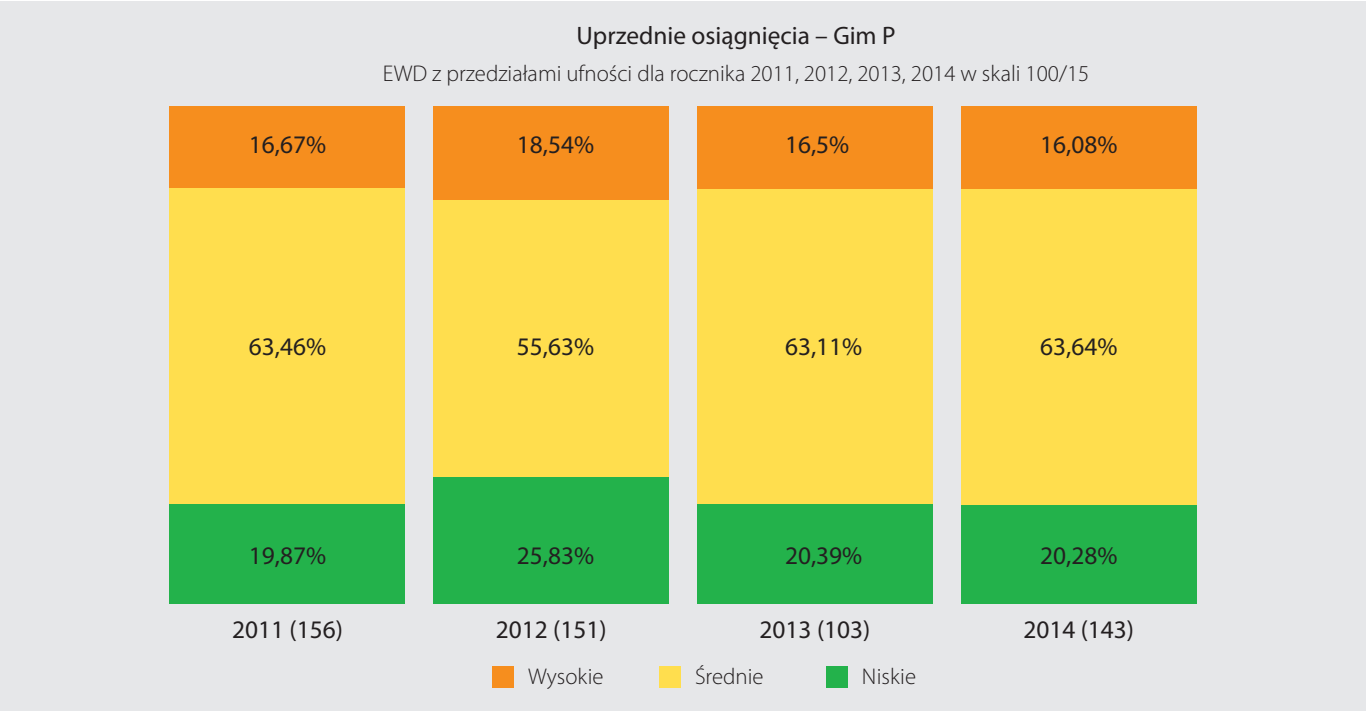


W roku 2014 w żadnym z oddziałów klasowych nie udało się osiągnąć z dziewczętami ponadprzeciętnej efektywności nauczania z matematyki. Problem niskiej efektywności nauczania matematyki w grupie dziewcząt jest trwałą charakterystyką tej szkoły i warto go uczynić przedmiotem ewaluacji wewnętrznej w szkole.

Gimnazjum P. W Gimnazjum P przed czterema laty wdrożono nowy model nauczania oparty na nauczaniu przez doświadczanie. Metodę zaczęli stosować przede wszystkim nauczyciele przedmiotów matematyczno-przyrodniczych. Metoda wpływa na zwiększone zaangażowanie uczniów w proces uczenia się, wzmacnia wiarę we własne możliwości uczniów, sprzyja samodzielnej pracy. W 2014 roku do egzaminu gimnazjalnego przystąpili uczniowie, którzy przez 3 lata nauki w gimnazjum byli nauczani z zastosowaniem tego modelu nauczania. Sprawdźmy zatem, czy wskaźniki EWD pokażą wzrost efektywności nauczania w zakresie przedmiotów matematyczno-przyrodniczych w tej szkole.

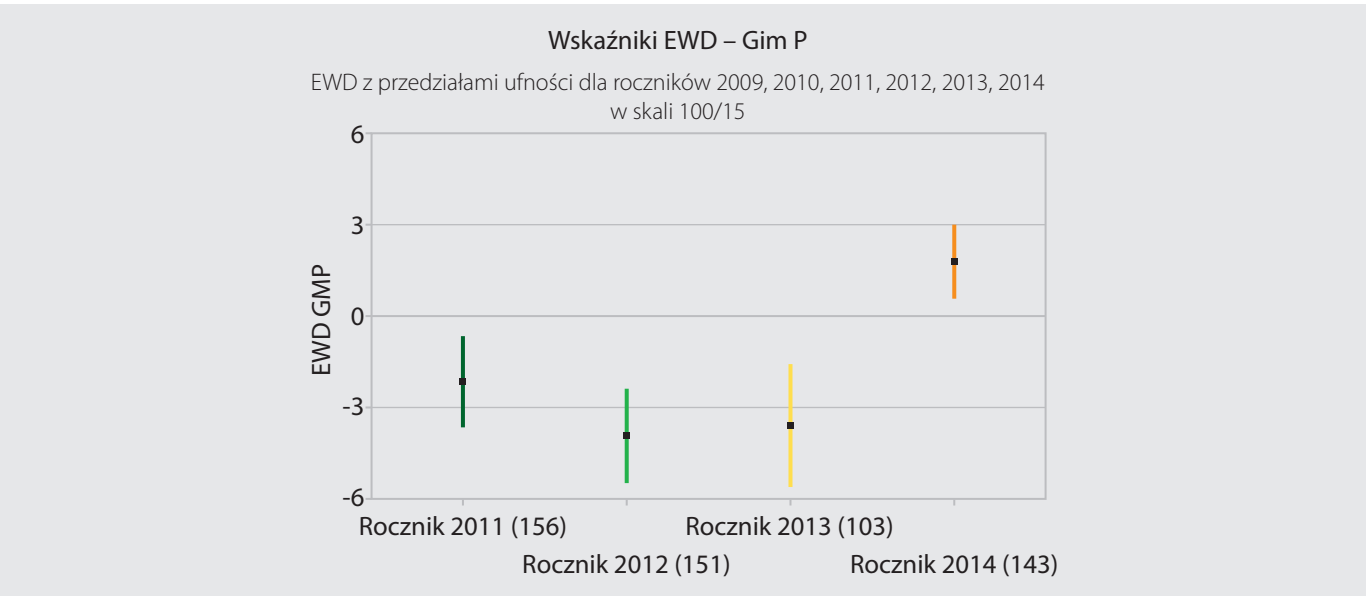
Szkoła przez lata uzyskiwała niskie wyniki egzaminacyjne przy ujemnym wskaźniku EWD. Rekrutowała uczniów o przeciętnych uprzednich osiągnięciach, co pokazano na poniższym wykresie.

Rysunek 4.37. Struktura uczniów gimnazjum P wg wyników uzyskanych na sprawdzianie po szóstej klasie. W opisach podano rok przystąpienia do egzaminu gimnazjalnego, a w nawiasach liczbę uczniów. Niskie wyniki odpowiadają wynikom na sprawdzianie z 1., 2., 3. przedziału staninowego, wyniki wysokie – wynikom na sprawdzianie z 7., 8., 9 przedziału staninowego



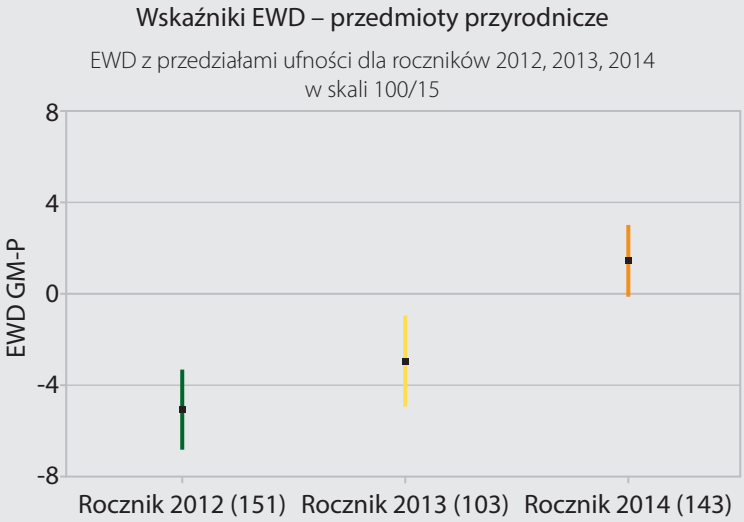
Roczniki, które w latach 2011–2014 przystępowały do egzaminu gimnazjalnego, były do siebie podobne ze względu na uprzednie osiągnięcia szkolne. Około 20% uczniów to uczniowie o niskich wynikach na sprawdzianie szóstoklasisty, a 16–18% to uczniowie o wysokich wynikach na sprawdzianie.

Rysunek 4.38. Efektywność nauczania w zakresie przedmiotów matematyczno-przyrodniczych w latach 2011–2014.



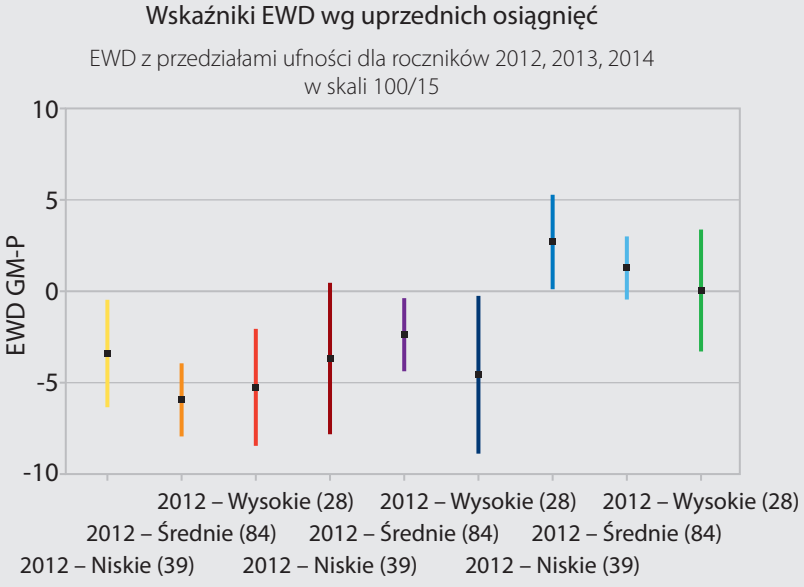
W latach 2011–2013 EWD była ujemna, w roku 2014 ponadprzeciętna.
W dalszych analizach ograniczymy się do roczników 2012, 2013 i 2014, co pozwoli nam rozpatrywać wskaźniki EWD tylko dla przedmiotów przyrodniczych.

Rysunek 4.39. Efektywność nauczania w zakresie przedmiotów przyrodniczych w latach 2012–2014

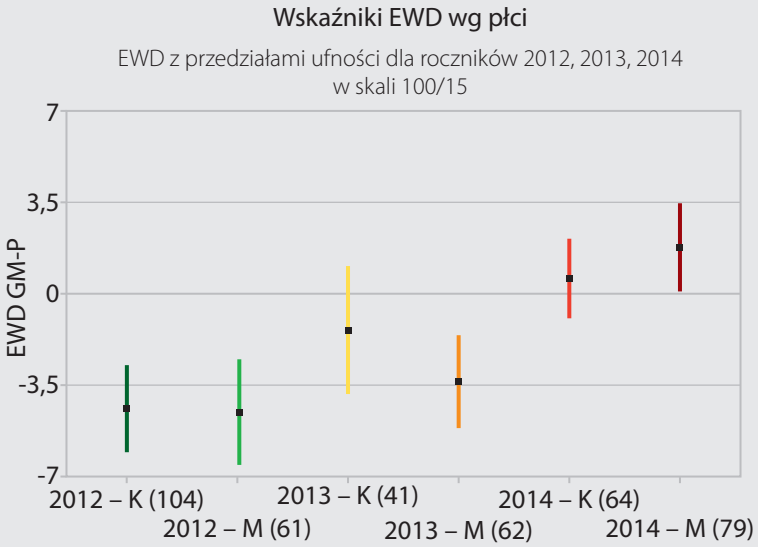


We wskaźnikach EWD z przedmiotów przyrodniczych obserwujemy wzrost od niższej niż przeciętna efektywności nauczania w 2012 roku do ponadprzeciętnej efektywności w roku 2014. Sprawdźmy, w jakiej grupie uczniów uzyskano największy względny przyrost umiejętności. Czy wyższe EWD szkoły uzyskano dzięki lepszej pracy z uczniami słabszymi, a może ponadprzeciętne EWD uzyskali tylko uczniowie o wysokich wynikach na sprawdzianie? A może wdrożona w szkole metoda skuteczniej wpływa na postępy chłopców niż dziewczynek?
Odpowiedzi na te pytania dostarcza analiza wskaźników EWD wg uprzednich osiągnięć szkolnych oraz wg płci.

Rysunek 4.40. Efektywność nauczania w zakresie przedmiotów przyrodniczych w latach 2012–2014 w podziale na uprzednia osiągnięcia szkolne



Rysunek 4.41. Efektywność nauczania w zakresie przedmiotów przyrodniczych w latach 2012-2014 w podziale na płeć (K – dziewczęta, M – chłopcy)



W 2014 roku uzyskano porównywalną, wysoką efektywność nauczania we wszystkich analizowanych grupach uczniowskich.

Te wyniki przemawiają za tym, że udało się poprzez zastosowanie nowej metody dydaktycznej podnieść efektywność nauczania przedmiotów przyrodniczych. Należy spodziewać się, że w kolejnych latach wskaźniki EWD szkoły utrzymają się na ponadprzeciętnym poziomie.

Ograniczenia i rozwój metody EWD dla gimnazjów

Nowa formuła egzaminu gimnazjalnego w 2012 roku spowodowała, że wzrosła liczba wskaźników możliwych do wyliczenia dla każdego gimnazjum – nie tylko zbiorcze wskaźniki humanistyczne i matematyczno-przyrodnicze, ale również wskaźniki dla języka polskiego, historii i WOS-u, matematyki oraz przedmiotów przyrodniczych. Podobnie nowa formuła sprawdzianu w 2015 roku da impuls do prac nad szacowaniem nowych wskaźników EWD gimnazjalnych w 2018 roku, między innymi dla języków obcych. Odejście od ponadprzedmiotowego charakteru sprawdzianu zmusi do rewizji modeli szacowania EWD. Może to na przykład oznaczać, że inaczej będą liczone gimnazjalne wskaźniki EWD dla matematyki i polskiego (jako miary uprzednich osiągnięć będzie można używać wyników z wyodrębnionych na sprawdzianie testów przedmiotowych). Konstrukcji nowych wskaźników powinny towarzyszyć badania nad porównywalnością tych wskaźników z miarami liczonymi przed 2018 rokiem.

Ważnym problemem, który warto w tym miejscu zasygnalizować, jest sposób traktowania laureatów konkursów przedmiotowych. Wskaźniki EWD mają ograniczoną wartość dla szkół skupiających dużą liczbę laureatów konkursów przedmiotowych. Stanowią oni specyficzną grupę uczniów. Nie przystępują do egzaminu (w przypadku egzaminu gimnazjalnego – do jednej z jego części, odpowiadającej tematyce konkursu, którego są laureatami), lecz mają przypisywaną maksymalną liczbę punktów. Jednocześnie laureaci stanowią grupę niezwykle zróżnicowaną wewnętrznie. Liczba konkursów dających możliwość zwolnienia z egzaminu jest znaczna, a do tego różna w zależności od województwa. Ponadto, konkursy opracowywane lokalnie nie są porównywalne co do stopnia trudności. Sprawia to, że nie ma sensu w modelu EWD traktowanie bycia laureatem jako ziemnej kontrolnej. Z punktu widzenia wskaźników EWD najlepszym rozwiązaniem problemu związanego z laureatami byłoby, gdyby pisali oni egzaminy tak jak wszyscy uczniowie, a wyniki konkursów mogłyby być uwzględniane w rekrutacji do szkół ponadgimnazjalnych w inny sposób.

W trzyletnich wskaźnikach EWD do skalowania wyników wykorzystywana jest informacja o rozwiązaniu każdego zadania w teście, natomiast w kalkulatorze EWD 100 wykorzystuje się, głównie ze względu na prostotę obsługi aplikacji, skalowanie wyników na podstawie sumy punktów uzyskanych przez ucznia za rozwiązane zadań z arkusza egzaminacyjnego. Ujednolicenie podejścia byłoby możliwe, gdyby Kalkulator EWD 100 mógł pobierać dane z centralnej bazy połączonych i wyskalowanych wyników, co wymaga jednak rozwiązań dotyczących integracji egzaminacyjnych baz danych i opracowania systemu autoryzowanego dostępu do tych danych.

4.5. Wykorzystanie metody edukacyjnej wartości dodanej w liceach ogólnokształcących i technikach

W wypadku liceów ogólnokształcących i techników najlepszą, dostępną miarą zasobów „na wejściu" są wyniki uczniów na egzaminie gimnazjalnym. W maturalnych modelach EWD jako zmienne kontrolne wykorzystywane są informacje o płci i dysleksji ucznia oraz o tym, czy uczeń miał na etapie ponadgimnazjalnym wydłużony tok nauczania. Wskaźniki EWD są więc liczone w taki sposób, że szansa na ich wysoką wartość nie zależy od tego, z uczniami o jakich cechach (przede wszystkim wynikach egzaminu gimnazjalnego) pracowała szkoła. Szkoła, która świetnie pracowała z uczniami o przeciętnych wynikach na egzaminie gimnazjalnym, będzie charakteryzować się dodatnią wartość EWD, mimo że jej średnie wyniki matury będą prawdopodobnie niższe niż szkoły, która słabo pracowała z uczniami, którzy osiągnęli wysokie wyniki po gimnazjum.

Modele EWD dla liceów ogólnokształcących i techników

Wskaźniki EWD na poziomie ponadgimnazjalnym uwzględniają wyłącznie osiągnięcia w zakresie kształcenia ogólnego. Wskaźniki EWD wyliczane są oddzielnie dla liceów ogólnokształcących i dla techników. Oznacza to, że licea ogólnokształcące porównywane są do innych liceów ogólnokształcących, a technika do innych techników. W związku z tym nie ma możliwości dokonania bezpośredniego porównania dwóch szkół, z których jedna to liceum ogólnokształcące, a druga to technikum. Rozwiązanie takie przyjęto ze względu na istotne różnice pomiędzy tymi dwoma typami szkół. Licea ogólnokształcące to szkoły o profilu ogólnym, ukierunkowanym na przygotowanie do podjęcia studiów. Technika z kolei łączą kształcenie ogólne z kształceniem zawodowym. Duża cześć uczniów kończących te szkoły w ogóle nie decyduje się na przystąpienie do matury, a przystępujący często ograniczają się do zdawania jedynie obowiązkowego zestawu przedmiotów. Dodatkowo szkoły te różnią się długością cyklu kształcenia (trzyletni w LO i czteroletni w technikach). Są więc istotne powody, by twierdzić, że rozwój umiejętności ogólnych, mierzonych na maturze, przebiega odmiennie w obu tych typach szkół. Co więcej, potwierdzają to wyraźne rozbieżności w wynikach matur – zdecydowanie niższe wśród uczniów techników. W związku z tym istotne wydaje się uwzględnienie w modelowaniu EWD zróżnicowania zdających, związanego z typem szkoły, do której uczęszczali.

Jednoroczne i trzyletnie wskaźniki EWD dla liceów ogólnokształcących i techników

Wskaźniki EWD dla szkół maturalnych, podobnie jak dla gimnazjów, wyliczane są w dwóch podstawowych wariantach: jako tak zwane wskaźniki jednoroczne i trzyletnie. Pierwsze z nich uwzględniają wyniki jednego rocznika absolwentów, podczas gdy przy wyliczaniu tych drugich brane są pod uwagę osiągnięcia trzech kolejnych roczników absolwentów. Oba typy wskaźników różnią się metodami statystycznymi wykorzystywanymi w modelowaniu EWD, ale za ważniejszą różnicę należy uznać ich przeznaczenie i sposób udostępniania odbiorcom.

Wskaźniki trzyletnie dla liceów i techników, tak jak gimnazjalne, publikowane są w ogólnodostępnym serwisie internetowym, z myślą o bardzo szerokiej grupie odbiorców, od nauczycieli i dyrektorów, poprzez organy zarządzające i nadzór pedagogiczny, po rodziców i uczniów. Wskaźniki jednoroczne przeznaczone są przede wszystkim na potrzeby ewaluacji wewnątrzszkolnej. Grupa

osób mogących korzystać ze wskaźników jednorocznych jest ograniczona do tych, którzy mogą otrzymać z okręgowych komisji egzaminacyjnych dane z wynikami uczniów – w praktyce głównie nauczycieli i dyrektorów szkół.
Ze względu na zakres treści nauczania objętych danym wskaźnikiem wyróżniamy cztery różne typy wskaźników trzyletnich dla szkół maturalnych:

- w zakresie języka polskiego;
- w zakresie matematyki;
- w zakresie przedmiotów humanistycznych – obejmujące wyniki z języka polskiego, historii i wiedzy o społeczeństwie;
- w zakresie przedmiotów matematyczno-przyrodniczych – obejmujące wyniki z matematyki, biologii, chemii, fizyki, geografii i informatyki.

Syntetyczne miary osiągnięć maturzystów tworzone są z uwzględnieniem wyników z wszystkich przedmiotów, które zdawali poszczególni uczniowie (czy to na poziomie podstawowym, czy rozszerzonym), a które przypisane są do danego wskaźnika. Stworzenie takich złożonych miar osiągnięć jest możliwe dzięki zastosowaniu odpowiednich metod statystycznych, a konkretnie teorii odpowiedzi na zadanie testowe (IRT). W dalszej części tekstu opisane zostały wybrane problemy, z którymi można sobie dzięki tej metodzie poradzić. W przypadku wskaźników EWD w zakresie języka polskiego i przedmiotów humanistycznych jako miara osiągnięć „na wejściu" wykorzystywane są wyniki części humanistycznej egzaminu gimnazjalnego. W przypadku wskaźników EWD w zakresie matematyki i przedmiotów matematyczno-przyrodniczych są to z kolei wyniki części matematyczno-przyrodniczej egzaminu gimnazjalnego.
Wskaźniki jednoroczne EWD dla LO i techników, pozwalające przy użyciu Kalkulatora EWD na prowadzenie analiz wewnątrzszkolnych w zakresie nauczania matematyki, zostały udostępnione po raz pierwszy dopiero w 2013 r., przy czym wyliczono i udostępniono poprzez Kalkulator EWD modele również dla lat wcześniejszych, do roku 2010 włącznie. Pierwotnie zostały one wyliczone w oparciu wyłącznie o wyniki matury z matematyki na poziomie podstawowym (jako wskaźnik osiągnięć na wejściu wykorzystano wyniki części matematyczno-przyrodniczej egzaminu gimnazjalnego), co pozwoliło wykorzystać analogiczną metodologię jak w przypadku jednorocznych wskaźników gimnazjalnych. Wskaźniki jednoroczne uwzględniające wyłącznie poziom podstawowy matury z matematyki miały jednak istotną wadę. Ze względu na występujący w teście maturalnym z matematyki na poziomie podstawowym efekt sufitowy model statystyczny zaniżał EWD liceów, których uczniowie mieli bardzo wysokie umiejętności matematyczne. Aby przezwyciężyć ten problem, w roku 2014 zastosowano skalowanie wyników maturalnych modelem Rascha. Pozwoliło to uwzględnić wyniki matury z matematyki zarówno na poziomie podstawowym, jak i na poziomie rozszerzonym, jednocześnie pozostawiając możliwość posługiwania się w Kalkulatorze EWD wyłącznie informacją o sumie punktów zdobytych w każdej z tych części egzaminu. Obecnie prowadzone są prace nad zaadaptowaniem tej metodologii do wyliczenia jednorocznych wskaźników EWD w zakresie języka polskiego, wymaga to jednak rozwiązania dodatkowych trudności, związanych z wyborem tematu wypracowania.

Skalowanie wyników matury

Polska matura jest egzaminem o bardzo złożonej strukturze. W latach 2010–2014 składała się z dwudziestu różnych przedmiotów, w tym sześciu języków obcych, nie wliczając egzaminów z języków mniejszości narodowych i etnicznych. Dodatkowo dla zdecydowanej większości przedmiotów układane są w każdym roku arkusze na dwóch różnych poziomach trudności: podstawowym i zaawansowanym.
Liczba piszących poszczególne części egzaminu maturalnego jest bardzo zróżnicowana. Wszyscy zdający maturę – obecnie około 300 tys. uczniów szkół ponadgimnazjalnych rocznie – podchodzą do przedmiotów obowiązkowych: języka polskiego i matematyki na poziomie podstawowym.

Spośród języków obcych, z których obowiązkowo trzeba wybrać jeden na poziomie podstawowym, zwykle około 85% zdających wybiera język angielski. Najpopularniejsze przedmioty zdawane jako dodatkowe to w ostatnich latach geografia i biologia na obu poziomach oraz chemia na poziomie rozszerzonym i WOS na poziomie podstawowym – wybiera je od około 20 tys. do około 45 tys. zdających. Są też przedmioty zdawane rokrocznie przez nie więcej niż kilkaset osób, jak filozofia, historia muzyki, wiedza o tańcu czy język łaciński i kultura antyczna.

Takie „rozdrobnienie" egzaminu maturalnego stanowi istotne wyzwanie w sytuacji, gdy chcemy wykorzystać jego wyniki do analizy np. efektywności nauczania w szkole. Ze względu na to, że w bardzo wielu szkołach poszczególne przedmioty (a dokładniej poszczególne przedmioty na poszczególnych poziomach) zdawane są przez bardzo niewielu uczniów (wyjąwszy oczywiście przedmioty obowiązkowe), wydaje się, że wykorzystanie wyników z tych przedmiotów do konstrukcji oddzielnych wskaźników efektywności pracy szkół w zakresie pojedynczych przedmiotów byłoby rozwiązaniem mało użytecznym, bowiem wskaźniki uzyskiwane na poziomie szkół byłyby w większości wypadków obarczone ogromną niepewnością statystyczną. W związku z tym dużo odpowiedniejsze do wykorzystania w tym kontekście wydają się bardziej ogólne miary umiejętności zdających egzamin maturalny, które byłyby w stanie uwzględnić jednocześnie wyniki różnych części tego egzaminu.

Takie złożone miary umiejętności możemy określić mianem kompozycyjnych, jako że łączą one ze sobą wyniki wielu bardziej szczegółowych pomiarów w ogólniejsze, bardziej syntetyczne wskaźniki. Wykorzystanie metody IRT pozwala skonstruować takie miary osiągnięć nawet na podstawie egzaminu o bardzo złożonej strukturze. Można powiedzieć, że w efekcie zastosowania metody IRT otrzymujemy pewien rodzaj łączenia ze sobą wyników różnych części egzaminu, a więc sprowadzania ich wyników do wspólnej skali. Oczywiście, aby było to możliwe, konieczne jest spełnienie warunku zdawania przez te same osoby różnych części egzaminu. W kontekście polskiej matury zasadnicze znaczenie dla możliwości konstruowania takich miar ma oczywiście występowanie dwóch przedmiotów obowiązkowo pisanych przez wszystkich: języka polskiego i matematyki na poziomie podstawowym.

Jak wspomniano już wcześniej, na potrzeby wyliczania wskaźników EWD konstruowane są na podstawie wyników matury cztery różne miary osiągnięć:

- w zakresie języka polskiego – na poziomie podstawowym i rozszerzonym;
- w zakresie matematyki – na poziomie podstawowym i rozszerzonym;
- w zakresie przedmiotów humanistycznych – na podstawie wyników z języka polskiego, historii i wiedzy o społeczeństwie na poziomie podstawowym i rozszerzonym;
- w zakresie przedmiotów matematyczno-przyrodniczych – na podstawie wyników z matematyki, biologii, chemii, fizyki, geografii i informatyki na poziomie podstawowym i rozszerzonym.

Dwie pierwsze są dużo prostsze – uwzględniają wyłącznie oba poziomy trudności jednego przedmiotu. Dwie kolejne, a zwłaszcza miara matematyczno-przyrodnicza, są bardzo skomplikowanymi miarami kompozycyjnymi. Każda z tych miar prezentowana jest przy pomocy skali standardowej o średniej 100 i odchyleniu standardowym 15, z tym że przeliczenie parametrów skali dokonywane jest oddzielnie dla uczniów techników i oddzielnie dla uczniów liceów ogólnokształcących. Tak więc w przypadku techników grupę odniesienia stanowią wszyscy uczniowie techników, którzy zdają maturę po raz pierwszy, a w przypadku liceów, wszyscy uczniowie liceów, którzy zdają maturę po raz pierwszy.

Poniżej omówiony zostanie tylko jeden szczególny problem, z jakim musimy się zmierzyć przy konstrukcji kompozycyjnych miar osiągnięć na podstawie wyników matury, a mianowicie kwestia wyboru przez uczniów zestawu zdawanych przedmiotów[50].

[50] Czytelników zainteresowanych bardziej szczegółowymi analizami odsyłamy do publikacji Tomasza Żółtaka (Żółtak, 2013: s. 13) i książki opisującej problemy modelowania cech ukrytych (Pokropek, 2015).

Problem autoselekcji przy wyborze przedmiotów zdawanych na maturze

Jak wspomniano, specyfika egzaminu maturalnego polega na tym, że uczniowie sami decydują, jakie przedmioty i na jakim poziomie będą zdawać. Oczywiście wyborów tych nie dokonują losowo, lecz w sposób strategiczny, związany z preferowanymi przedmiotami (w dużej mierze zależy to od wymagań rekrutacyjnych na kierunkach studiów, na które chce się dostać maturzysta) i poziomem swoich umiejętności (należy się spodziewać, że przedmioty na poziomie rozszerzonym wybierają uczniowie o wyższym poziomie umiejętności). Aby uzyskać dobre oszacowania parametrów modelu skalowania, a potem także dobre oszacowania poziomu umiejętności zdających, konieczne jest więc uwzględnienie w jakiś sposób w modelu występowania takich procesów wyborów.

W tym celu zdecydowano się na włączenie do modelu dodatkowych zmiennych (po jednej dla każdej części egzaminu) traktowanych tak jak zadania, a opisujących, czy uczeń zdawał daną część egzaminu, czy też nie (wzorowano się na rozwiązaniu zaproponowanym w publikacji Korobko i współpracowników – 2008). Na podstawie charakterystyk takich „zadań", które określamy mianem parametrów selekcji, możemy potem określić, w jakim stopniu wybór poszczególnych części egzaminu powiązany jest z poziomem mierzonych umiejętności.

Problemem bardzo zbliżonym do wyboru zdawanych przedmiotów jest wybór tematu wypracowania z języka polskiego, a także wypowiedzi pisemnych wchodzących w skład arkuszy z WOS-u i historii na poziomie rozszerzonym. Może on zostać rozwiązany w analogiczny sposób, tj. poprzez potraktowanie tych samych kryteriów oceny wypracowania jako różnych zadań – w zależności od wybranego tematu – i dodanie do modelu dodatkowego „zadania" opisującego dokonany wybór tematu (Pokropek, 2011).

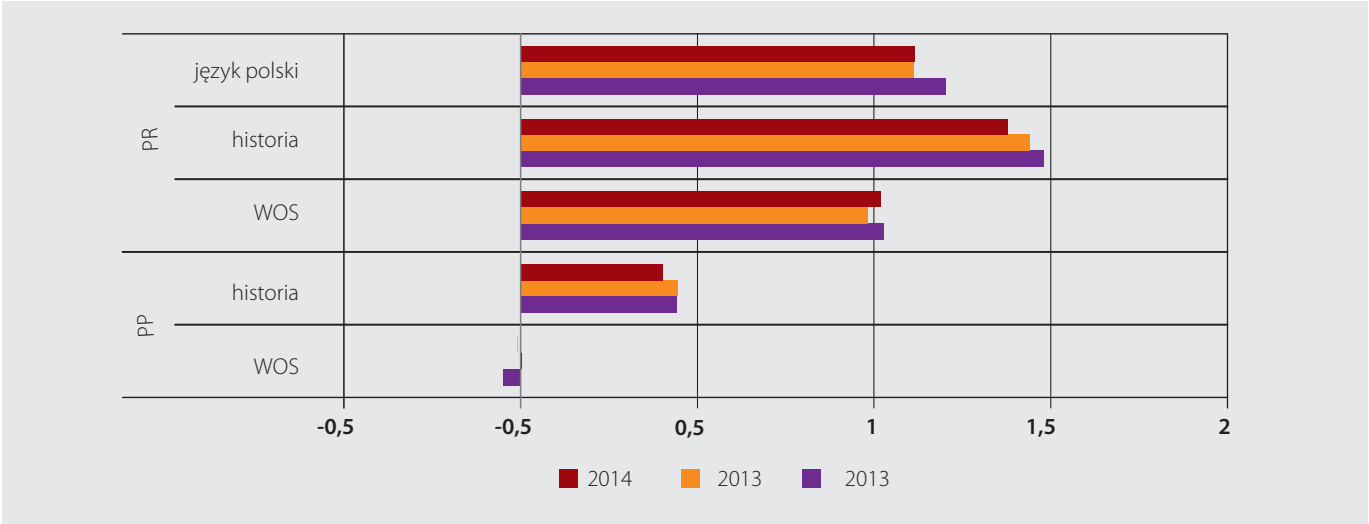
W oparciu o wyliczenia dokonane w procesie konstruowania naszych kompozycyjnych miar osiągnięć możemy przyjrzeć się „elitarności" poszczególnych przedmiotów maturalnych, to jest temu, jak silne są zależności pomiędzy poziomem umiejętności a skłonnością do zdawania danego przedmiotu na danym poziomie. Jak opisano wyżej, mówią nam o tym tzw. parametry selekcji. Ich wartości z poszczególnych modeli skalowania z lat 2012–2014 przedstawione zostały na poniższych rysunkach. Wartości parametrów większe od zera wskazują, że wybór danego przedmiotu (na danym poziomie) jest statystycznie pozytywnie powiązany z poziomem osiągnięć. Inaczej mówiąc, dany przedmiot wybierali raczej zdający o wysokich umiejętnościach w danym obszarze nauczania. Im wartość parametru większa, tym częściej dany przedmiot (na danym poziomie) wybierają uczniowie o wysokich umiejętnościach, a tym rzadziej uczniowie o niskich umiejętnościach. Ujemne wartości parametrów selekcji wskazują z kolei, że dany przedmiot, na danym poziomie wybierali raczej zdający o poziomie umiejętności niższym od średniej wśród wszystkich zdających.

Biorąc pod uwagę konstrukcję egzaminu maturalnego, oczekiwalibyśmy dodatnich wartości współczynników selekcji związanych z wyborem przedmiotów na poziomie rozszerzonym. Oczekiwalibyśmy również, że w ramach każdego przedmiotu wartość parametru selekcji będzie większa dla poziomu rozszerzonego niż dla poziomu podstawowego. Okazuje się, że oczekiwania te znajdują potwierdzenie w danych. Wartości parametrów selekcji dla poziomu rozszerzonego wyniosły od 0,63 (geografia w 2014 r.) do 2,2 (matematyka w 2012 r.) dla przedmiotów matematyczno-przyrodniczych i od 0,98 (WOS w 2013 r.) do 1,48 (historia w 2012 r.) dla przedmiotów humanistycznych. W przypadku każdego przedmiotu wartości parametrów selekcji dla poziomu rozszerzonego są też wyraźnie wyższe niż dla poziomu podstawowego.

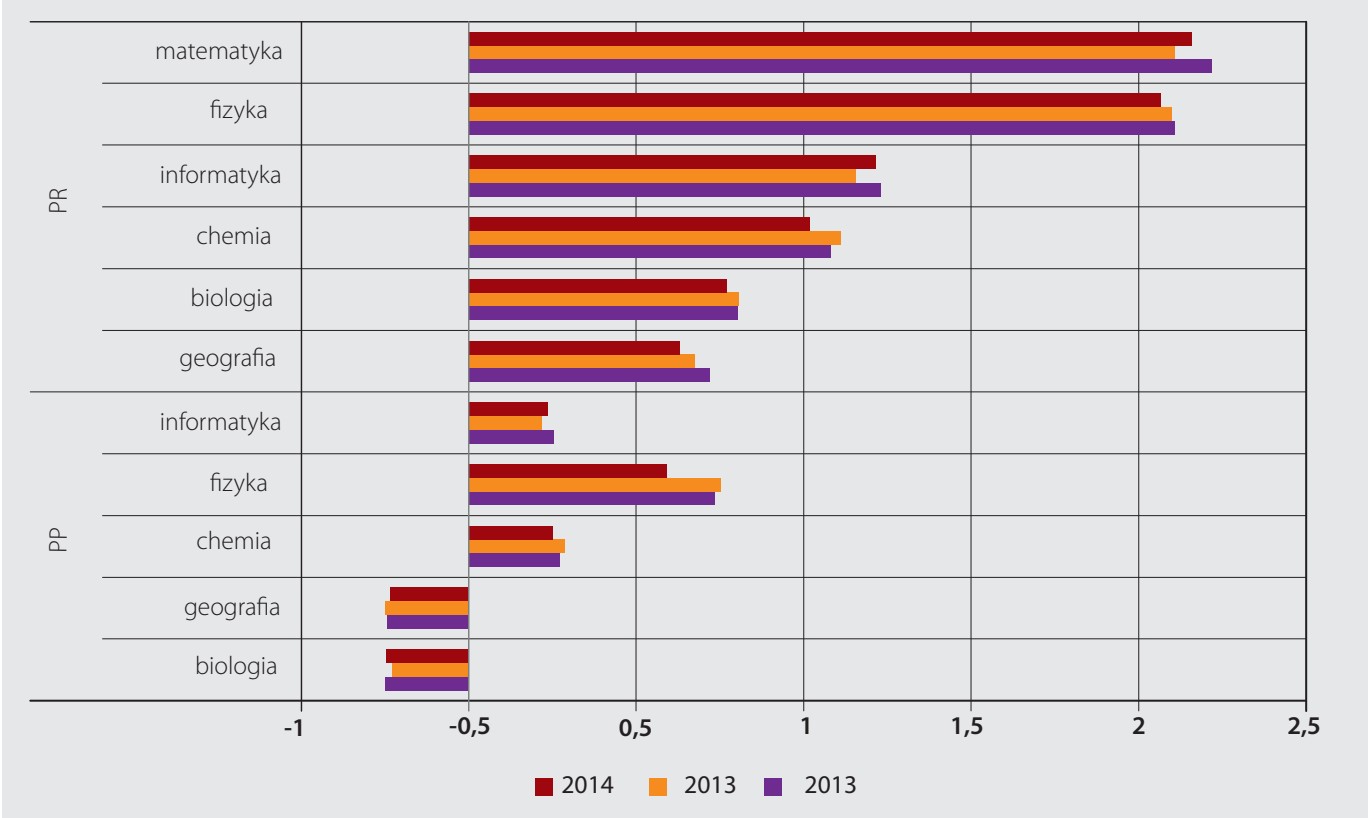
Zdecydowanie najbardziej poziomem umiejętności wyróżniają się zdający rozszerzoną matematykę i fizykę. Mniej „elitarne" okazały się rozszerzona biologia i geografia. Pozostałe przedmioty przyrodnicze oraz przedmioty humanistyczne na poziomie rozszerzonym plasowały się pomiędzy tymi dwoma grupami przedmiotów. Jeśli chodzi o wybór przedmiotów na poziomie podstawowym, to możemy wyróżnić takie, które wybierane są przez raczej słabszych uczniów (biologia i geografia – wartości parametrów selekcji około –0,2), takie, których wybór jest właściwie bez związku z poziomem umiejętności (WOS), oraz takie, które wybierają raczej lepsi zdający (chemia, informatyka,

historia, fizyka). Warto zauważyć, że wartości parametrów selekcji dla fizyki na poziomie podstawowym były porównywalnie z tymi dla rozszerzonej biologii czy geografii.
Warta odnotowania jest także duża stabilność wartości parametrów selekcji pomiędzy latami. Wskazuje to, że strategie wyboru przedmiotów przez zdających są – przynajmniej w zakresie, w jakim rolę odgrywa tu poziom umiejętności – dosyć stałe, pomimo sporej zmienności w wymaganiach rekrutacyjnych uczelni oraz czasami zróżnicowanego poziomu trudności danej części egzaminu między latami.

Rysunek 4.42. Wartości parametrów selekcji pokazujących siłę związku pomiędzy poziomem umiejętności humanistycznych a skłonnością do zdawania poszczególnych przedmiotów na poszczególnych poziomach



Rysunek 4.43. Wartości parametrów selekcji pokazujących siłę związku pomiędzy poziomem umiejętności matematyczno-przyrodniczych a skłonnością do zdawania poszczególnych przedmiotów na poszczególnych poziomach



Trafność wskaźników EWD dla liceów ogólnokształcących i techników

Analizowaniu trafności wskaźników EWD dla liceów ogólnokształcących i techników poświęcone zostało badanie *Ścieżki rozwoju edukacyjnego młodzieży – szkoły gimnazjalne*, zrealizowane w latach 2009–2013 przez IFiS PAN (Karwowski, 2013). W jego ramach analizowano między innymi kwestię, na ile nieuwzględnienie w polskich modelach EWD dodatkowych zmiennych kontrolnych, przede wszystkim opisujących status społeczno-ekonomiczny rodziny ucznia, może prowadzić do zniekształcania informacji na temat efektywności pracy szkoły. Badano też wpływ na osiągnięcia inteligencji ogólnej oraz, choć w bardzo ograniczonym zakresie, udziału w korepetycjach i zajęciach pozalekcyjnych.

Ramka 4.3. Opis podłużnego badania uwarunkowań wyników nauczania w szkołach podstawowych

Badanie *Ścieżki rozwoju edukacyjnego młodzieży – szkoły gimnazjalne* prowadzone było w latach 2009–2013 i objęło reprezentatywną, ogólnopolską próbę losową łącznie 200 oddziałów klas pierwszych ze 100 liceów ogólnokształcących (ok. 2,4 tys. uczniów), 60 techników i liceów profilowanych (ok. 1,2 tys. uczniów) oraz 40 zasadniczych szkół zawodowych (ok. 700 uczniów). Badanie realizowane było w ramach projektu *Badania dotyczące rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej* (EWD), a jego realizację w wyniku otwartego przetargu zlecono Instytutowi Filozofii i Socjologii Polskiej Akademii Nauk. Badanie objęło pełen cykl kształcenia na etapie ponadgimnazjalnym (śledzone są losy tych samych uczniów). Uczestniczyli w nim uczniowie, ich rodzice, nauczyciele oraz dyrektorzy szkół.

Badanie miało na celu lepsze poznanie indywidualnych, rodzinnych i szkolnych czynników odpowiedzialnych za osiągnięcia szkolne uczniów. Było to punktem wyjścia do oceny, na ile stosowane w Polsce modele EWD dla szkół ponadgimnazjalnych pozwalają w odpowiedni sposób kontrolować czynniki niezależne od działań szkoły. Analizowano także, czym charakteryzują się szkoły osiągające wysokie, a czym osiągające niskie wartości wskaźników EWD.

Podczas pierwszego etapu badania, który odbył się, kiedy badani uczniowie uczęszczali do pierwszych klas, zebrano dane z wykorzystaniem narzędzi z badania PISA 2009. Dotyczyło to zarówno pomiaru umiejętności w zakresie czytania i interpretacji, matematyki i rozumowania w naukach przyrodniczych, jak i kwestionariuszy dla uczniów, nauczycieli i dyrektorów szkół. Podczas kolejnego etapu badania, zrealizowanego na początku drugiej klasy, wykorzystano trzy testy psychologiczne: test matryc Ravena (inteligencja), test kompetencji społecznych i test nadziei na sukces. W rok po pierwszym etapie badania zrealizowano etap trzeci, w którym powtórnie zmierzono umiejętności uczniów z wykorzystaniem testów PISA oraz zastosowano testy psychologiczne do oceny samooceny i lęku. Zebrano też wtedy informacje o funkcjonowaniu uczniów w środowisku rówieśniczym (Karwowski, 2013). Więcej informacji na temat badania można znaleźć na stronie: http://www.ifispan.waw.pl/index.php?lang=pl&m=page&pg_id=215

Wyniki analiz wskazują, że na etapie edukacji ponadgimnazjalnej wykorzystanie w modelach EWD dla liceów ogólnokształcących i techników informacji o wynikach uczniów na egzaminie gimnazjalnym pozwala niemal całkowicie kontrolować wpływ czynników statusowych. W dużym stopniu, choć nie tak skutecznie jak w wypadku SES, pozwala też kontrolować wpływ inteligencji, rozumianej jako potencjał poznawczy wynikający z czynników biologicznych. Pewną poprawę w tym zakresie można by zapewne uzyskać, zwiększając rzetelność egzaminu gimnazjalnego.

Analizowano również, jakie cechy szkoły stanowią korelaty wysokich wartości wskaźników EWD. Zarówno w LO, jak i w technikach wysokim wartościom EWD sprzyjało efektywne wykorzystanie czasu na lekcji (czas poświęcony na naukę w stosunku do czasu poświęcanego na utrzymanie porządku w klasie i na czynności administracyjne) oraz proszkolny klimat panujący wśród uczniów (zarówno gdy był on oceniany przez nauczycieli, jak i gdy uczniowie oceniali nastawienie swoich koleżanek i kolegów). Co ciekawe, częste stosowanie w nauczaniu metod podających (mierzone poprzez deklaracje nauczycieli nt. stosowanych technik nauczania) negatywnie korelowało z EWD w liceach ogólnokształcących, ale w technikach albo było to bez znaczenia, albo, w przypadku języka polskiego, sprzyjało wyższym wartościom EWD. W liceach ogólnokształcących wysokie EWD szkoły było powiązane z pozytywnymi postawami wobec czytania (deklaracje, że czyta się często i chętnie), nasileniem wśród uczniów danej szkoły relacji koleżeńskich i relacji przyjaźni oraz bogatą ofertą zajęć dodatkowych organizowanych przez szkołę. Wyniki te są ogólnie rzecz biorąc zgodne z oczekiwaniami co do tego, jakimi cechami powinny charakteryzować się „dobre" szkoły i choć nie może to być decydującym argumentem, jednak przemawia na korzyść twierdzenia o trafności wskaźników EWD.

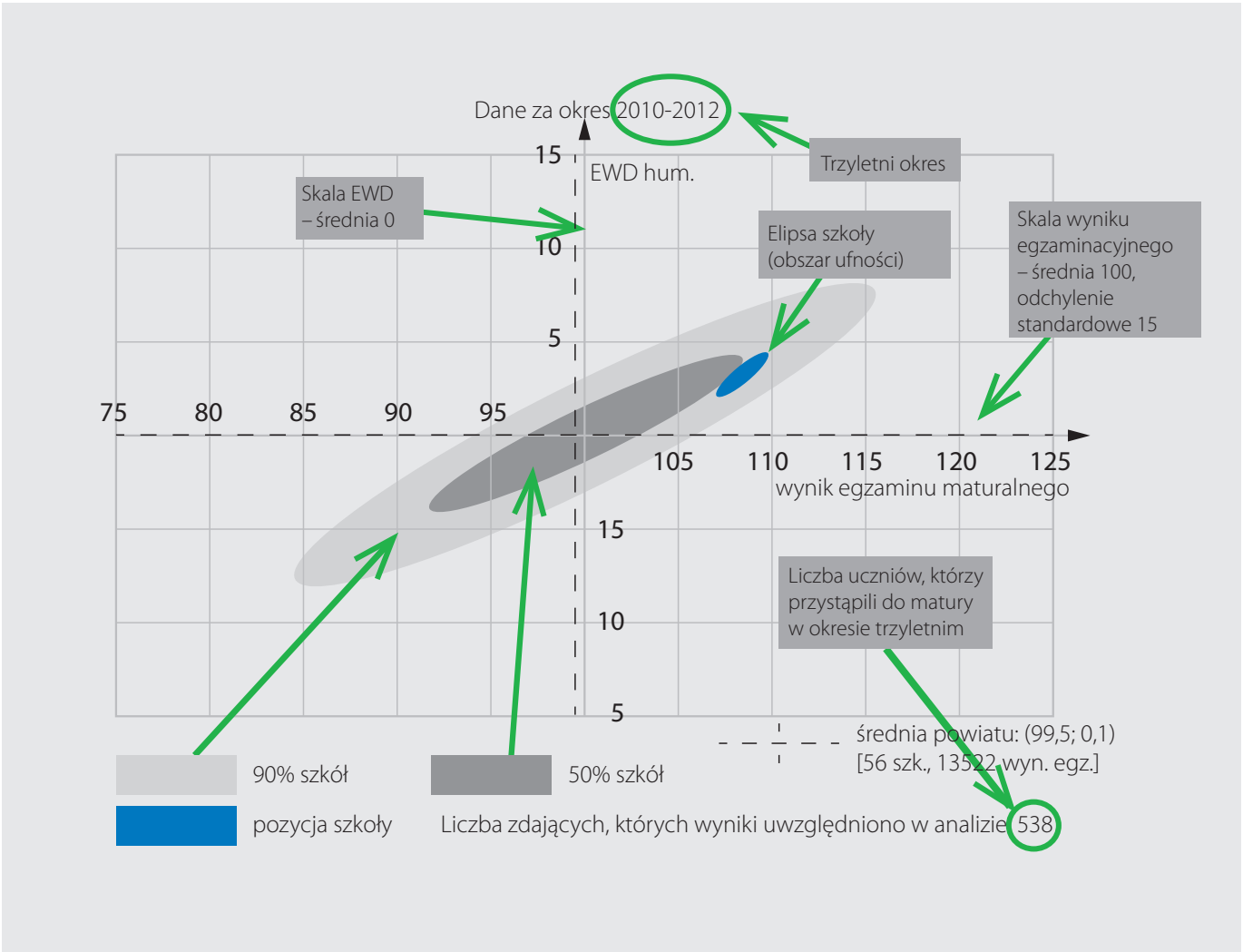
Sposób prezentacji wskaźników EWD dla liceów ogólnokształcących i techników

Dla prezentacji trzyletnich wskaźników maturalnych EWD, podobnie jak dla gimnazjów, przyjęto formę graficzną: na jednym wykresie pokazano jednocześnie wyniki egzaminacyjne oraz EWD. Szkołę reprezentuje elipsa (obszar ufności), wielkość której zależy od liczby uczniów w szkole. Elipsę szkoły należy traktować jako obszar ufności, czyli obszar, w którym z 95% prawdopodobieństwem znajduje się prawdziwy wynik szkoły. Im więcej uczniów przystąpiło w szkole do egzaminu maturalnego, tym dokładniej szacowane są wyniki i EWD, a więc, tym mniejsza jest elipsa.

Na podstawie wykresu możemy też wyznaczyć jednowymiarowe przedziały ufności, oddzielnie dla średnich wyników szkoły i oddzielnie dla EWD. Uzyskamy to, rzutując elipsę odpowiednio na oś wyników matury lub na oś EWD. Należy jednak pamiętać, że po dokonaniu takich rzutowań tracimy informację o współzależności między tymi dwoma cechami szkoły, o której informuje nas kształt całej elipsy.

Na rysunku 4.44. pokazano wszystkie najważniejsze elementy graficznej prezentacji trzyletnich wskaźników maturalnych. Dwie środkowe (szare) elipsy, wskazują obszary, w których koncentruje się odpowiednio 50% i 90% średnich wyników szkół. Innymi słowy, szare elipsy określają obszary najbardziej typowych wyników.

Rysunek 4.44. Podstawowe elementy graficznej prezentacji trzyletnich wskaźników maturalnych



Uzupełnieniem graficznej prezentacji wyniku egzaminacyjnego i EWD są tabele, w których podano informację o liczbie osób, które przystąpiły w szkole do egzaminu maturalnego w analizowanym okresie, wraz ze strukturą wyboru przedmiotów i poziomów egzaminów.

Kalkulator EWD 100 daje możliwość przeprowadzenia wielu różnych typów analiz wewnątrzszkolnych, odnoszących się zarówno do wyników matury z matematyki, EWD, jak i do wyników egzaminu gimnazjalnego. Funkcjonalności Kalkulatora zostaną przedstawione w części opisującej praktyczne wykorzystanie wskaźników EWD. Warto jedynie zaznaczyć, że informacje o EWD i/lub średnich wynikach matury prezentowane są w nim wyłącznie w formie jednowymiarowych przedziałów ufności (reprezentowanych graficznie w postaci odcinków). Nie ma możliwości wyrysowania w Kalkulatorze EWD obszaru ufności (elipsy), który pokazywałby łączną informację zarówno o wynikach matury, jak i o EWD.

Zarówno w przypadku wskaźników jednorocznych, jak i trzyletnich należy pamiętać, że wskaźniki EWD mają charakter względny. W skali kraju wskaźnik EWD ma z definicji wartość równą zeru. Wskaźnik EWD dla liceum ogólnokształcącego mówi o tym, na ile wysokie/niskie wyniki matury uzyskali jego absolwenci w porównaniu do uczniów o analogicznych wynikach na egzaminie gimnazjalnym. Podobnie wskaźnik EWD dla technikum pokazuje, na ile wysokie/niskie wyniki matury uzyskali jego absolwenci w porównaniu do uczniów techników w Polsce przy statystycznej kontroli wyników na egzaminie gimnazjalnym. Wartość dodatnia EWD wskazuje na ponadprzeciętną w skali kraju efektywność nauczania, wartość ujemna na niższą niż przeciętna efektywność.

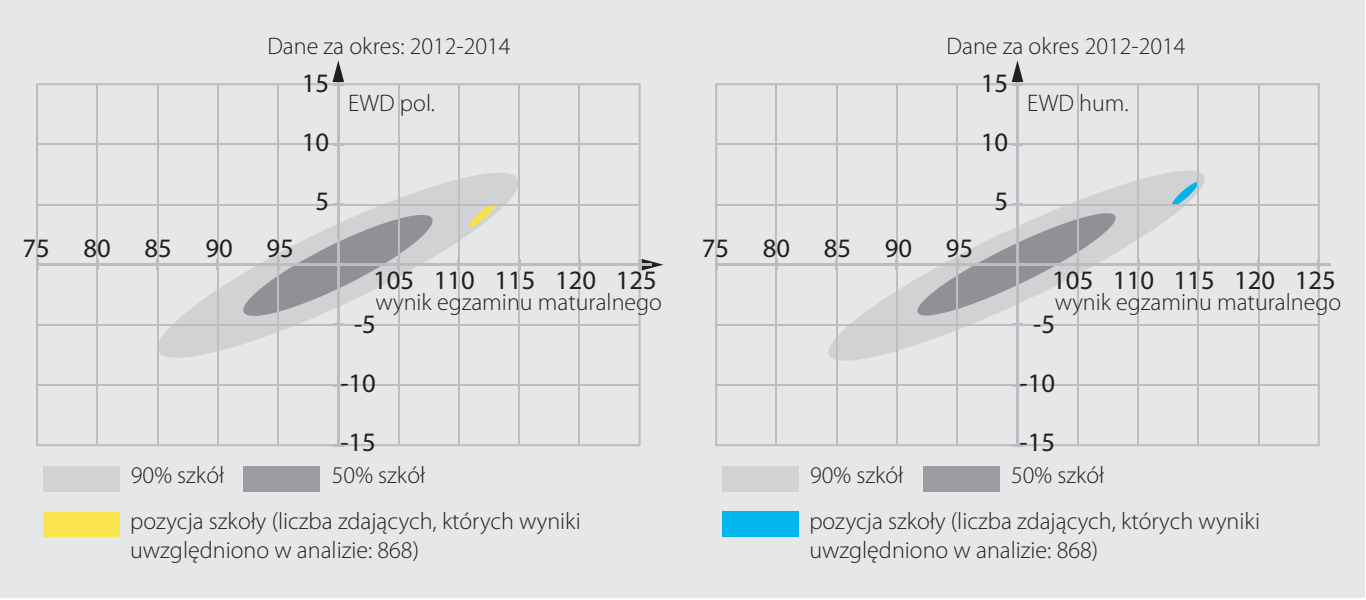
Możliwości wykorzystania metody EWD przez licea ogólnokształcące i technika

W tym rozdziale przedstawiono dwa przykłady wykorzystania metody EWD do analizy wyników egzaminacyjnych w liceach i technikach. Zaprezentowane wykresy pokazują potencjał analityczny trzyletnich wskaźników EWD i Kalkulatora EWD 100.

Liceum Ogólnokształcące A. To bardzo duża wielkomiejska szkoła, ze stuletnią tradycją. Po jej skończeniu większość młodzieży dostaje się na wybrany kierunek studiów.

Zobaczmy, jaki obraz szkoły rysuje się na podstawie trzyletnich wskaźników EWD za okres 2012–2014.

Rysunek 4.45. Liceum A. Przedmioty humanistyczne

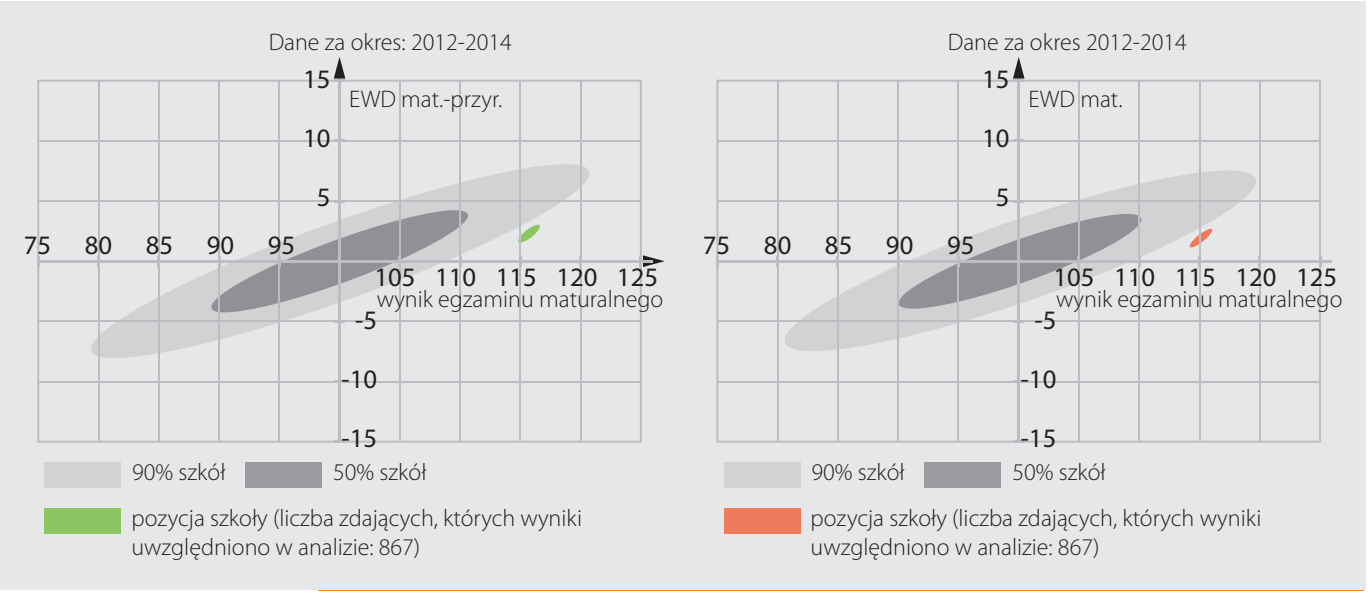


Przedmiot	Łącznie	Poziom rozszerzony
j. polski	868	272
historia	133	133
WOS	168	167

W liceum prawie 1/3 uczniów zdaje język polski na poziomie rozszerzonym, jednak można zauważyć, że wyższą efektywność i wyniki egzaminacyjne osiągają zdający, gdy wskaźniki liczone są łącznie dla wszystkich przedmiotów humanistycznych, niż dla samego języka polskiego. Elipsa wyników szkoły (EWD oraz wynik egzaminacyjny) znajduje się wtedy nad przerywaną zieloną kreską, natomiast dla samego języka polskiego efektywność liczona wskaźnikiem EWD jest niższa (pod przerywaną kreską). Analogiczna zależność występuje w odniesieniu do wyników matury, tj. elipsa opisująca wyniki w zakresie samego języka polskiego przesunięta jest w lewo względem tej, która uwzględnia wszystkie przedmioty humanistyczne. Wyniki z historii i wiedzy o społeczeństwie przyczyniają się do osiągania przez szkołę wyższych wyników egzaminacyjnych i wyższej efektywności.

W części matematyczno-przyrodniczej nie występują różnice w wartościach wskaźników z przedmiotów matematyczno-przyrodniczych i dla samej matematyki.

Rysunek 4.46. Liceum A. Przedmioty matematyczno-przyrodnicze



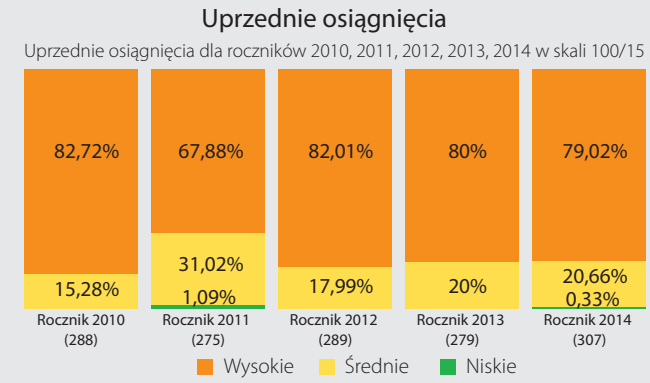
Przedmiot	Łącznie	Poziom rozszerzony
matematyka	868	438
biologia	203	199
chemia	229	226
fizyka	38	33
geografia	135	133
informatyka	20	19

Wyniki z przedmiotów matematyczno-przyrodniczych są o jedno odchylenie standardowe wyższe od średnich wyników dla szkół licealnych, efektywność pracy szkoły jest ponadprzeciętna, ale nie jest bardzo wysoka.

Dzięki Kalkulatorowi EWD 100 można dokładniej przyjrzeć się efektywności nauczania w liceum A. Analizy wyników szkoły zostały wykonane dla danych egzaminacyjnych z matematyki za lata 2010–2014. Kalkulator EWD 100 dla szkół maturalnych umożliwia wykonywanie obliczeń statystycznych dla połączonych wyników egzaminu gimnazjalnego z części matematyczno-przyrodniczej i **matematyki** zdawanej na maturze (poziom podstawowy i rozszerzony).

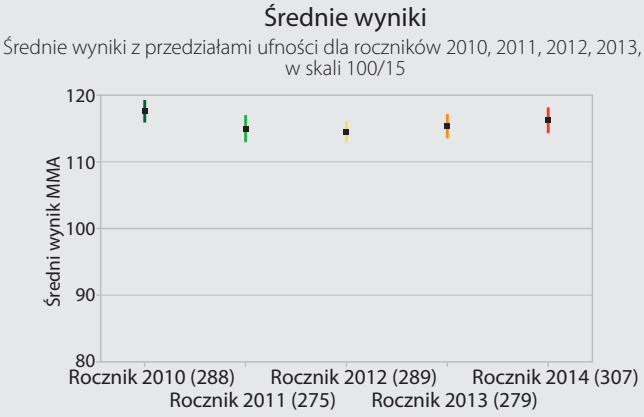
W LO A uczy się młodzież o bardzo wysokich osiągnięciach na egzaminie gimnazjalnym. Pokazuje nam to wykres *Uprzednie osiągnięcia,* jedna z analiz Kalkulatora EWD 100.

Rysunek 4.47. Liceum A. Analizy wykonane za pomocą Kalkulatora EWD 100



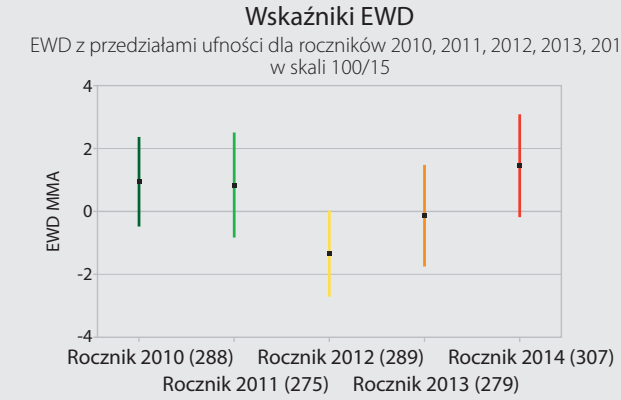
Wyniki z egzaminu gimnazjalnego młodzieży zdającej maturę w latach 2010–2014

Około 4/5 populacji uczęszczającej do LO A przychodzi z wynikami wysokimi po egzaminie gimnazjalnym w części matematyczno-przyrodniczej, sporadycznie zdarzają się uczniowie z wynikami niskimi.



Wyniki oraz efektywność pracy szkoły w zakresie nauczania matematyki w latach 2010–2014

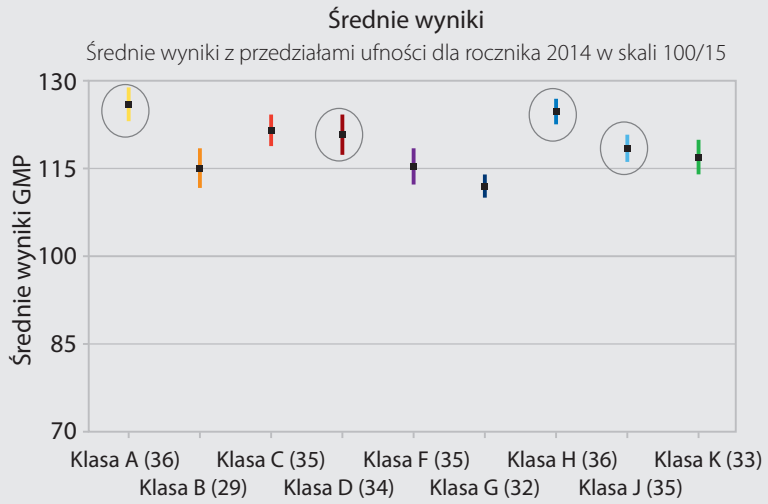
Szkoła osiąga co roku bardzo wysokie wyniki egzaminacyjne–powyżej jednego odchylenia standardowego wyżej, niż średni wynik szkół licealnych w Polsce.
Wyniki szkoły z matematyki są porównywalne w latach 2010–2014.



Wartości EWD z matematyki są bliskie zeru, w latach 2012–2014 widać trend wzrostowy.

Wyniki analizy wskazują, że wysokie wyniki egzaminu maturalnego z matematyki są głównie efektem tego, że do liceum A przychodzi młodzież z bardzo wysokimi wynikami z matematyki na egzaminie gimnazjalnym, natomiast wpływ edukacyjny szkoły jest przeciętny (od roku 2012 zarysowuje się trend wzrostowy EWD).
Przejdźmy do pogłębionych analiz wewnątrzszkolnych. Zacznijmy od analizy uprzednich osiągnięć uczniów w poszczególnych oddziałach liceum A.

Rysunek 4.48. Średni wynik uczniów na egzaminie gimnazjalnym z podziałem na oddziały



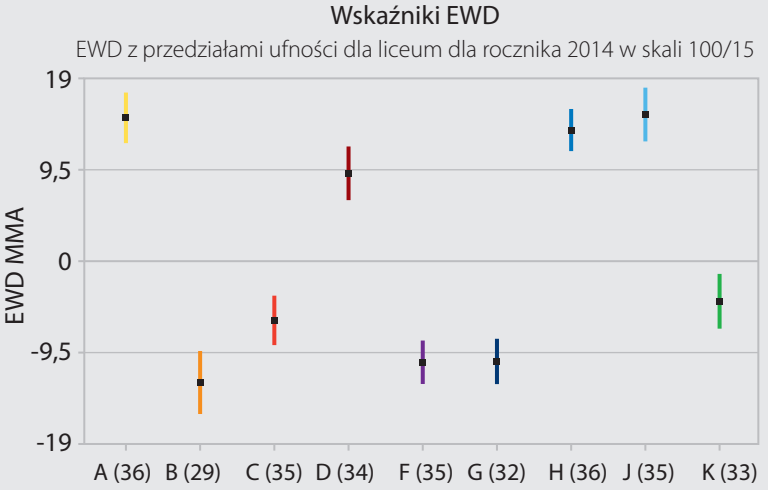
Średnie wyniki z egzaminu gimnazjalnego młodzieży zdającej maturę w roku 2014

Wyniki uczniów z części matematyczno-przyrodniczej egzaminu gimnazjalnego w roku 2011 w oddziałach dla rocznika, który zdawał maturę w roku 2014.

Widać znaczne zróżnicowanie średnich wyników w oddziałach.

Oddziały zaznaczone kółkiem (A, D, H i J) realizowały program z dodatkową liczbą godzin z matematyki. Do klas A i H została wybrana młodzież o bardzo wysokich wynikach z egzaminu gimnazjalnego (część matematyczno-przyrodnicza): 1,5 odchylenia standardowego powyżej średniej w Polsce. Przyjrzyjmy się zatem EWD w poszczególnych oddziałach

Rysunek 4.49. EWD w poszczególnych oddziałach



Efektywność pracy szkoły z podziałem na klasy (egzamin maturalny z matematyki 2014 r.)

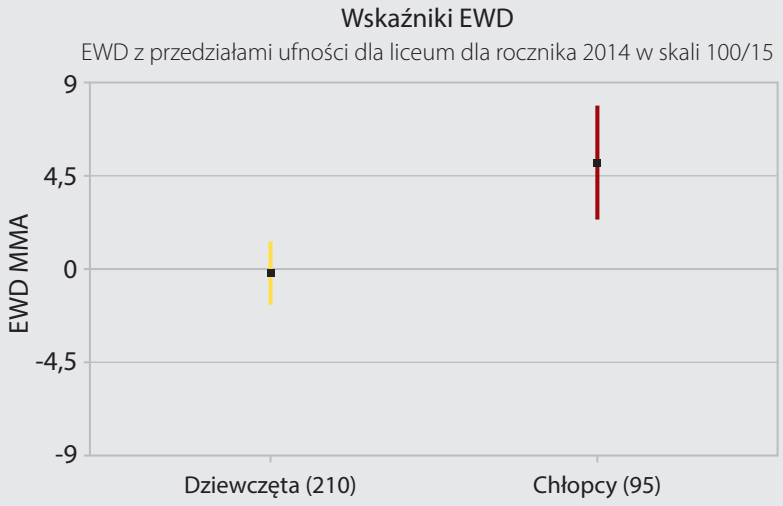
Klasy osiągajace wysoką efektywność z matematyki, to klasy o zwiększonej liczbie godzin tego przedmiotu. W klasach, w których nie poszerzono liczby godzin matematyki, efektywność nauczania tego przedmiotu była niska.

Ponadprzeciętną efektywność nauczania matematyki osiągnięto w oddziałach, w których matematyka była realizowana w zwiększonej liczbie godzin. W pozostałych klasach, ze standardową liczbą godzin, efektywność mierzona wskaźnikiem EWD była poniżej przeciętnej.

W następnym etapie analiz sprawdźmy, czy w liceum A obserwujemy zróżnicowanie efektywności nauczania matematyki ze względu na płeć uczniów.

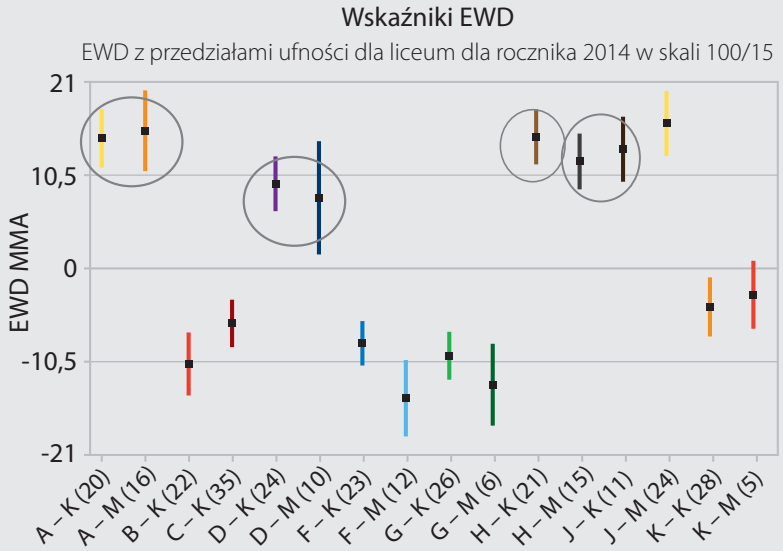
Rysunek 4.50. Płeć a efektywność nauczania matematyki

Efektywność pracy szkoły z podziałem na klasy (egzamin maturalny z matematyki 2014 r.)



Szkoła

Wskaźnik EWD jest wyższy dla chłopców niż dla dziewcząt. Z danych wynika, że w szkole maturę rozszerzoną z matematyki zdaje dwukrotnie więcej dziewcząt niż chłopców.

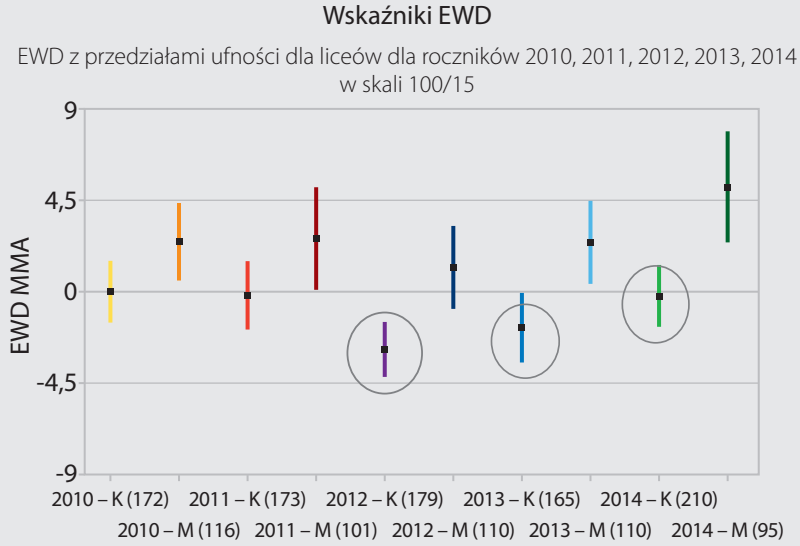


Klasy

W klasach o profilu matematycznym, (zaznaczone kółkiem), chłopcy i dziewczęt uczą się z porównywalną efektywnością, w klasach o innych profilach efektywność nauczania matematyki zarówno dla chłopców, jak i dla dziewcząt jest niska.

Wskaźnik EWD jest wyższy dla chłopców niż dla dziewcząt a z danych wynika, że w szkole maturę rozszerzoną z matematyki zdaje dwukrotnie więcej dziewcząt niż chłopców. Jednak analiza, w której uwzględniono zarówno podział na oddziały, jak i płeć maturzystów, pokazuje, że efektywność nauczania matematyki w szkole zdecydowanie silniej zależy od profilu klasy, do której uczęszcza maturzysta, niż od płci ucznia. Uczniowie wybierający profile niematematyczne osiągają bardzo niski wskaźnik EWD z matematyki. Analogiczną zależność można obserwować od roku 2010, równocześnie do liceum uczęszcza coraz więcej dziewcząt oraz maleje liczba chłopców.

Rysunek 4.51. Płeć a efektywność nauczania matematyki w latach 2010 – 2014



Zależność wskaźników egzaminacyjnych od płci zdających

Zdecydowanie wyższą efektywność pracy (EWD) z matematyki mają absolwenci -chłopcy. Wyraźnie widać ten efekt w ostatnich trzech latach.

Od roku 2010 dziewczęta nie miały ani razu wskaźnika EWD powyżej zera, a w dwóch rocznikach (2012, 2013) osiągały poziom poniżej przeciętnej.

Uzyskane w liceum A wyniki analiz można podsumować w następujący sposób:

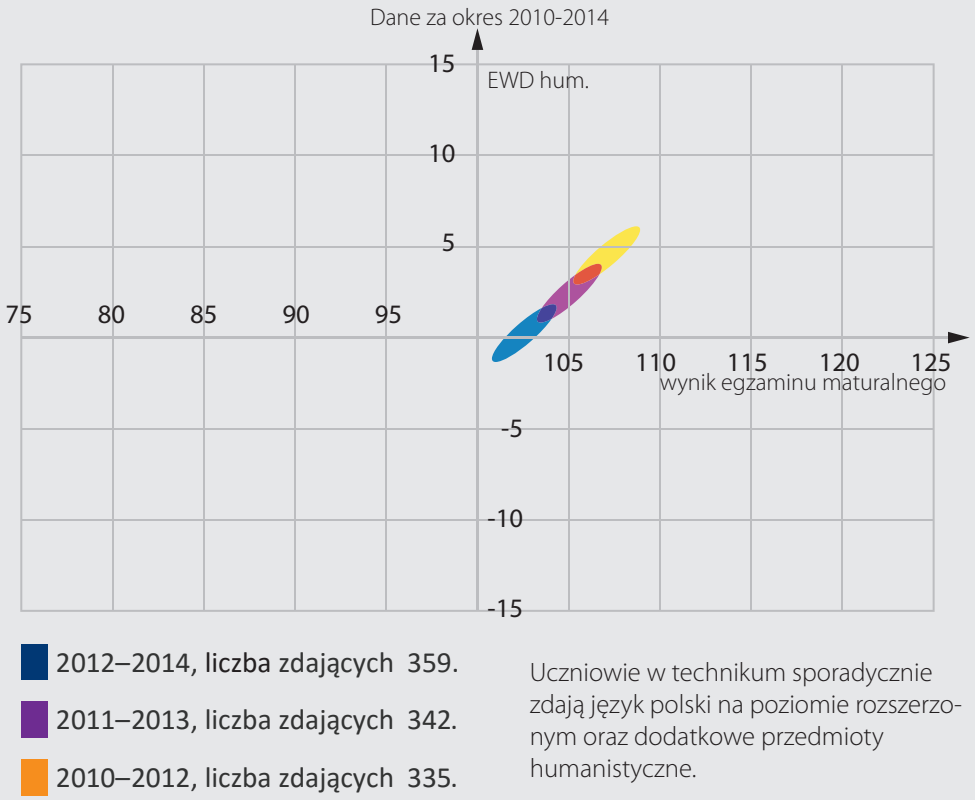
- Wyższe wyniki egzaminacyjne i efektywność (EWD) z historii i WOS-u na maturze niż z języka polskiego.
- Wysokie wartości wskaźników EWD z matematyki w klasach z dodatkowymi godzinami nauczania przedmiotu oraz znacząco niższa efektywność nauczania matematyki w oddziałach o profilach innych niż matematyczne.

Uzyskane wyniki to oczywiście początek, a nie koniec działań ewaluacyjnych.

Technikum LR. Szkoła znajduje się w mieście wojewódzkim. Była to szkoła o profilu energetyczno-elektrycznym, która w związku ze zmianami na rynku pracy modyfikuje swój profil na informatyczno-elektroniczny. Szkoła przygotowuje do pracy w zawodach technik elektryk, technik elektronik oraz technik informatyk.

Analizę zacznijmy od przedmiotów humanistycznych.

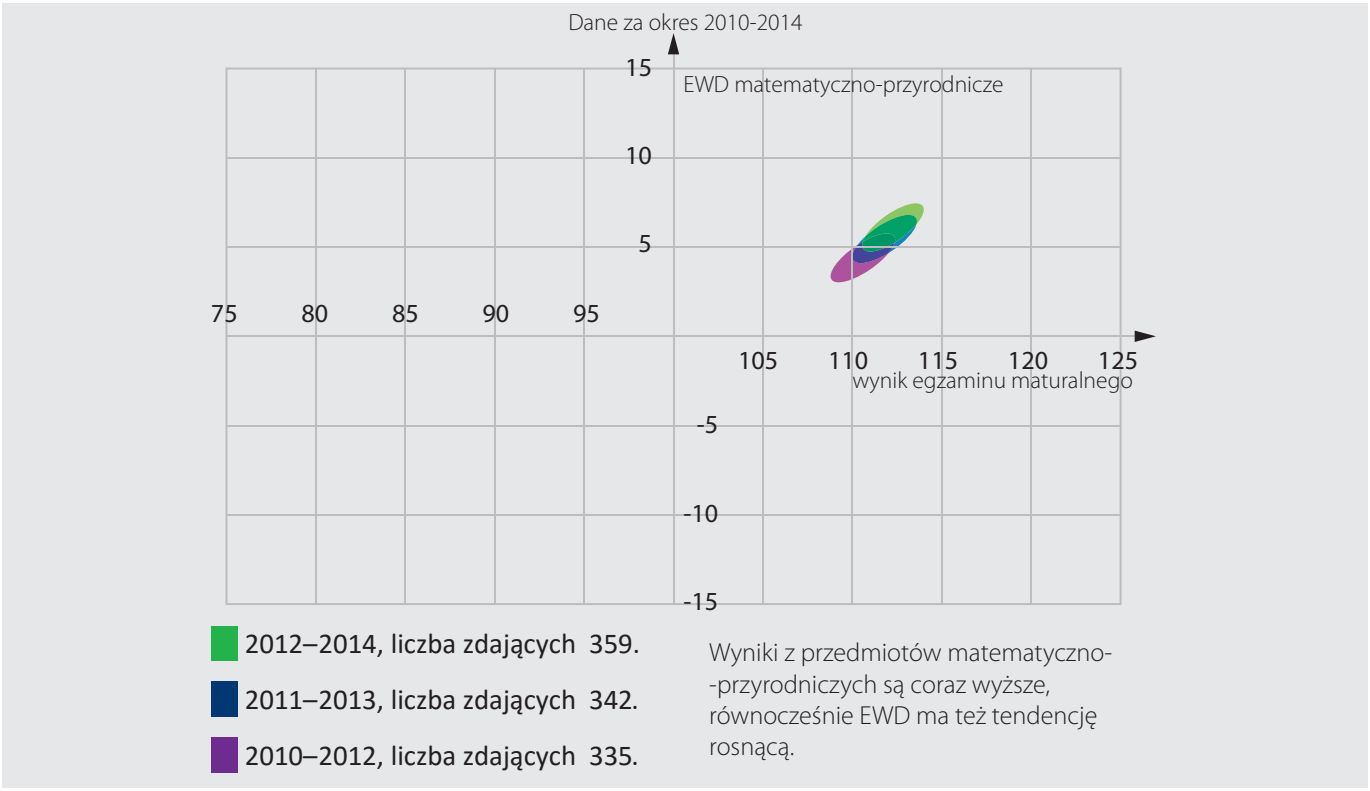
Rysunek 4.52. Technikum LR. EWD w obszarze humanistycznym



Liczba osób zdających poszczególne przedmioty w latach 2012–2014		
przedmiot	**łącznie**	**poziom rozszerzony**
j. polski	359	5
historia	2	0
WOS	2	0

W zakresie przedmiotów humanistycznych w technikum LR obserwujemy spadek efektywności nauczania i maturzyści osiągają coraz niższe wyniki egzaminacyjne.

Rysunek 4.53. Technikum LR. EWD w obszarze matematyczno-przyrodniczym



Z przedmiotów matematyczno-przyrodniczych technikum LR osiąga wysokie wyniki egzaminacyjne oraz pracuje z wysoką efektywnością. Dla okresu 2012–2014 przedział ufności dla technikum MP w obszarze osiągnięć matematyczno-przyrodniczych rozciąga się od około 112 do 114 punktów, czyli o mniej więcej 2/3 odchylenia standardowego powyżej średniej krajowej dla techników. Wskaźnik EWD również jest wysoki i mieści się w przedziale (5; 8). Z danych uzyskanych od szkoły wiemy, że prawie wszyscy uczniowie podchodzą do egzaminów maturalnych i zawodowych i zdecydowana większość z sukcesem.

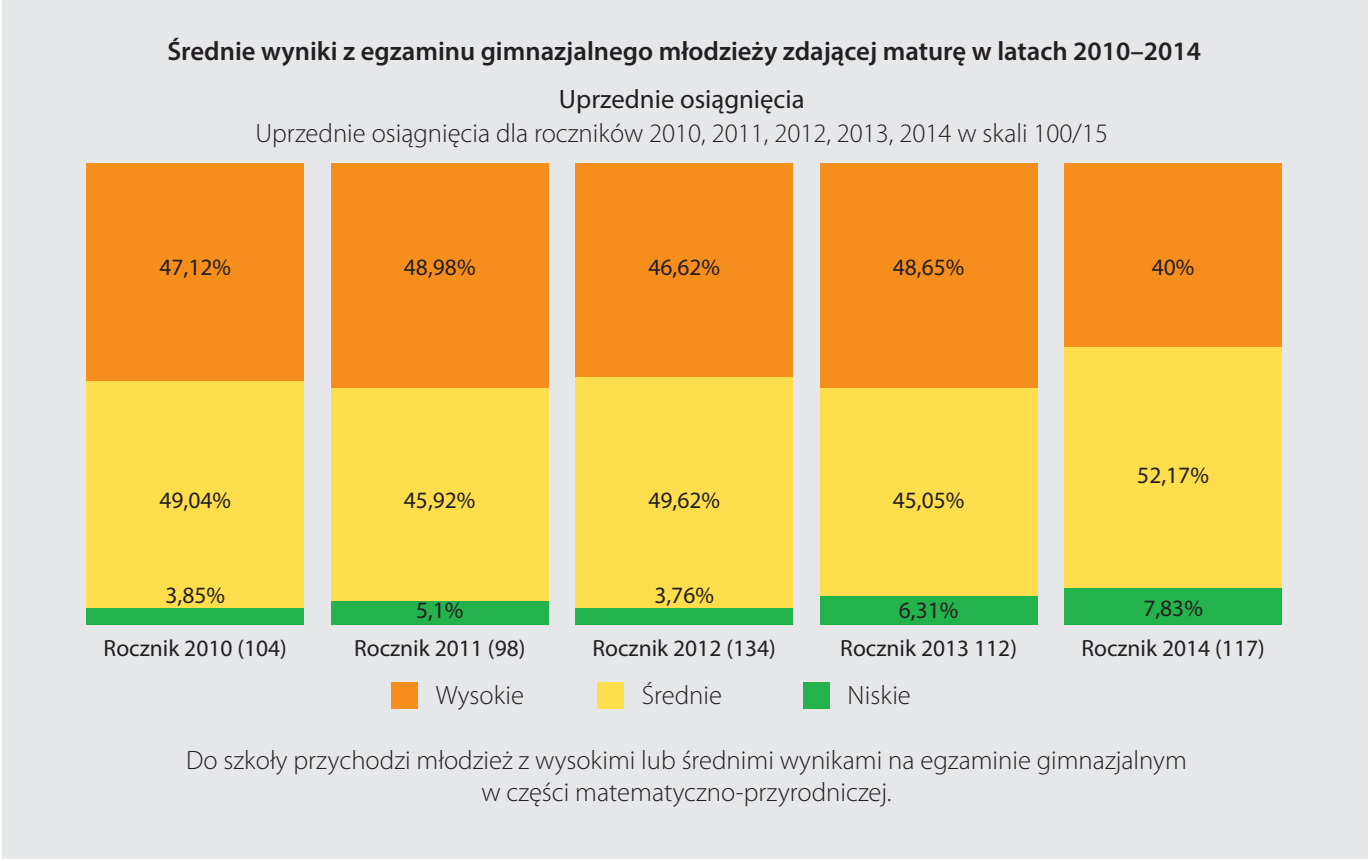
Jak widać w poniższej tabeli, przedmiotem najczęściej zdawanym w technikum LR jako przedmiot dodatkowy jest **informatyka.** W latach 2010–2012 zdawały go 44 osoby, a w okresie 2012–2014 już 82 osoby. Jest to zgodne z obecnym profilem technikum (informatyczno-elektroniczne). W tej szkole matematyka na poziomie rozszerzonym jest wybierana każdego roku przez ponad jedną czwartą maturzystów. Maturzyści z klasy elektronicznej stosunkowo często, jak na technika, zdają na maturze fizykę jako przedmiot dodatkowy.

Tabela 4.8. Liczba osób zdających przedmioty matematyczno-przyrodnicze w technikum LR:

	2010–2012		2011–2013		2012–2014	
przedmiot	**łącznie**	**poziom rozszerzony**	**łącznie**	**poziom rozszerzony**	**łącznie**	**poziom rozszerzony**
matematyka	335	88	342	92	359	93
biologia	0	0	0	0	0	0
chemia	0	0	1	0	1	0
fizyka	18	8	23	9	21	7
geografia	35	8	24	7	17	3
informatyka	44	31	65	53	82	80

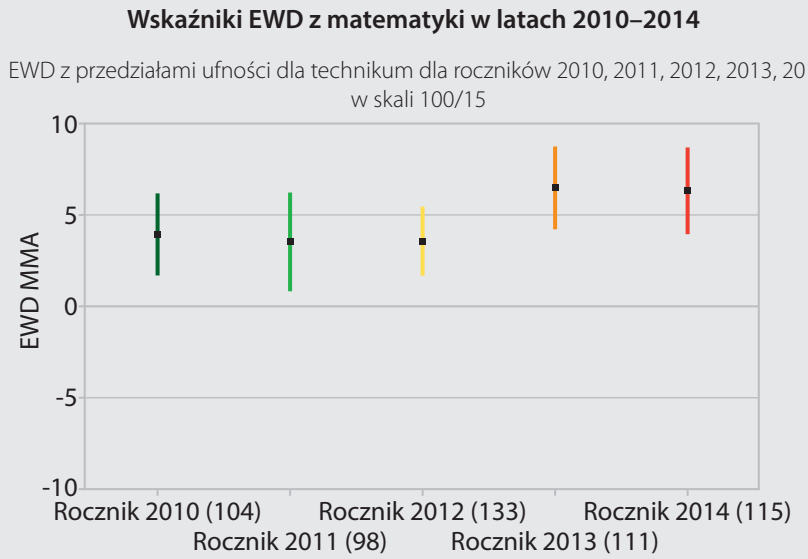
Dokładniejszą informację o efektywności nauczania w technikum LR dostarczą nam analizy wykonane przy użyciu programu Kalkulator EWD 100. Do analiz wykorzystano dane egzaminacyjne z lat 2010–2014.
Zacznijmy od analizy naboru do szkoły.

Rysunek 4.54. Technikum LR. Wyniki egzaminu gimnazjalnego w części matematyczno-przyrodniczej roczników przystępujących do matury w latach 2010–2014



Technikum LR jest szkołą o stabilnym naborze uczniów, mniej więcej połowa uczniów przychodzi z wynikami średnimi z egzaminu gimnazjalnego, druga połowa ma wyniki wysokie, nieliczni to absolwenci gimnazjów z niskim wynikiem z przedmiotów matematyczno-przyrodniczych.
Wykorzystując Kalkulator EWD 100, prześledźmy bardziej wnikliwie efektywność nauczania matematyki.

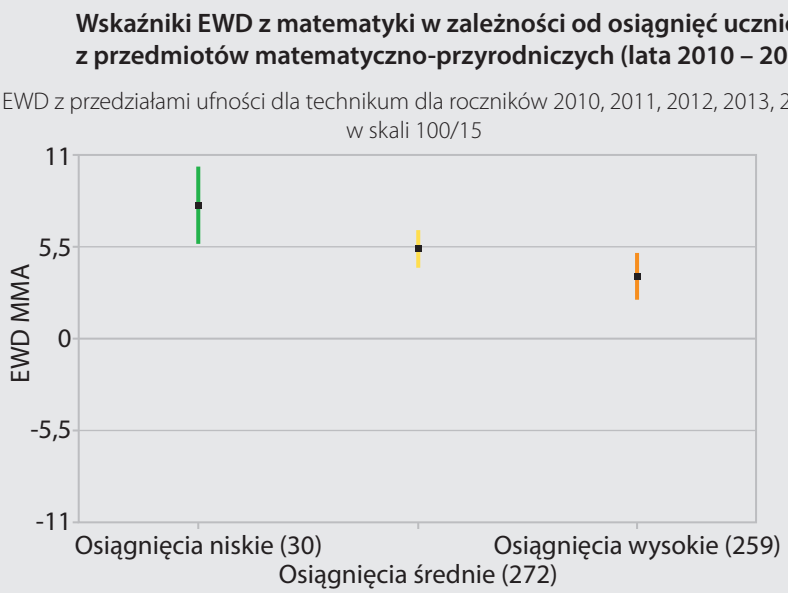
Rysunek 4.55. Technikum LR. Jednoroczne wskaźniki EWD z matematyki w latach 2010–2014



EWD z matematyki w latach 2010–2014 jest ponadprzeciętne, w dwóch ostatnich latach można powiedzieć, że jest nawet bardzo wysokie.

Wskaźnik EWD z matematyki w analizowanym okresie waha się od plus 3 do plus 7 punktów. Oznacza to ponadprzeciętną efektywność nauczania matematyki na tle wszystkich techników w kraju. Przyjrzyjmy się efektywności nauczania matematyki w podziale na trzy poziomy osiągnięć na egzaminie gimnazjalnym. Wyniki odpowiedniej analizy wykonanej w Kalkulatorze EWD pokazuje poniższy rysunek.

Rysunek 4.56. Technikum LR. Łączny wskaźnik EWD z matematyki w latach 2010–2014 dla uczniów o trzech poziomach uprzednich osiągnięć



Dla całego analizowanego okresu można zauważyć interesujące różnice EWD w grupach uczniów wyróżnionych ze względu na wyniki na egzaminie gimnazjalnym.

W latach 2010–2014 wtechnikum LR zdawało maturę 30 absolwentów z niskimi wynikami na egzaminie gimnazjalnym i z nimi szkoła pracowała najefektywniej.
Zajmijmy się teraz analizami różnic między oddziałami. Jak przydziela się uczniów do oddziałów ze względu na uprzednie osiągnięcia?

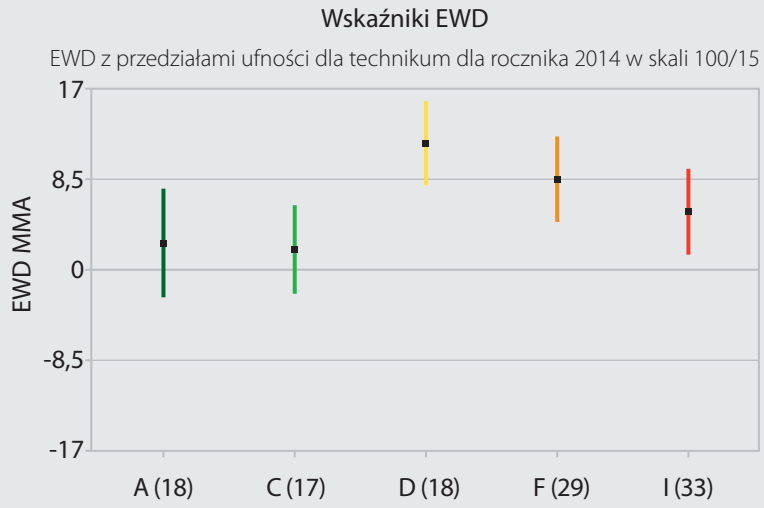
Rysunek 4.57. Technikum LR. Uprzednie osiągnięcia maturzystów w podziale na oddziały. Dwa sposoby analizy: rozkład procentowy trzech grup uczniów wyróżnionych ze względu na wyniki na egzaminie gimnazjalnym i średnie wyników egzaminu gimnazjalnego wraz z przedziałami ufności

Wykres *Uprzednie osiągnięcie* dostarcza informacji o strategii przyjmowania młodzieży do klas I w technikum LR. Najwięcej uczniów o niskich wynikach przyjęto do klasy A (elektryk), w klasie tej rozkład uczniów jest porównywalny z rozkładem ogólnopolskim, w pozostałych klasach osiągnięcia uczniów z egzaminu gimnazjalnego są wyższe: w klasach D, F, I wyniki egzaminacyjne są wysokie i średnie. Porównywanie średnich wyników po egzaminie gimnazjalnym (matematyczno-przyrodniczym) pokazuje, że statystycznie różnią się one tylko pomiędzy klasami A oraz I, ponieważ pozostałe przedziały ufności dla klas nie są rozłączne.

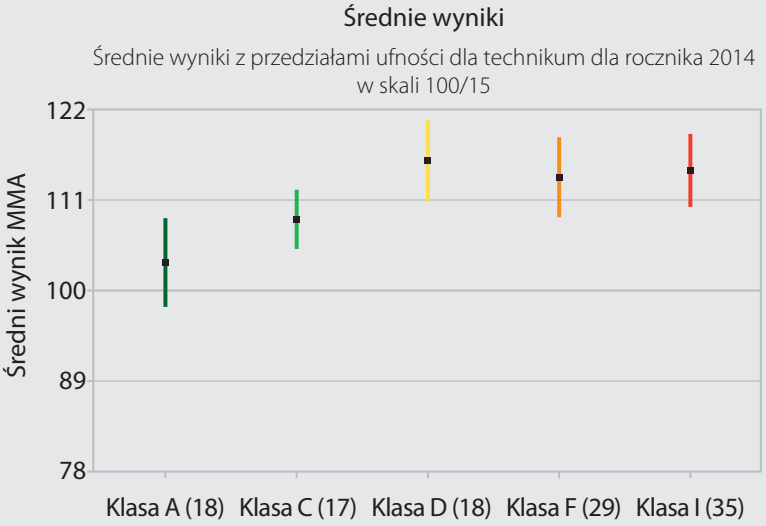
Na koniec prezentacji przykładowych analiz EWD dla techników przyjrzyjmy się międzyoddziałowemu zróżnicowaniu EWD i wyników matury z matematyki.

Rysunek 4.58. Technikum LR. Wskaźniki EWD i średnie wyniki z matury z matematyki w podziale na oddziały

Wskaźniki EWD i średnie wyniki oddziałów, matura z matematyki w 2014 r.



Wskaźniki EWD są znacząco zróżnicowane. Istotne statystycznie różnice występują między klasami A i C oraz klasą D.



Istotne statystycznie różnice w wynikach egzaminu maturalnego z matematyki występują pomiędzy klasami A oraz D, F i I.

Uzyskane w technikum LR wyniki analiz można podsumować w następujący sposób:

- Coraz niższe wyniki egzaminacyjne i EWD z przedmiotów humanistycznych.
- Stabilne, ponadprzeciętne wartości wskaźnika EWD w obszarze matematyczno-przyrodniczym.
- Znaczące różnice w efektywności nauczania matematyki w dwóch klasach o profilu elektronicznym (klasy C i D), przy porównywalnych osiągnięciach młodzieży z egzaminu gimnazjalnego.

Ograniczenia i rozwój metody EWD dla liceów ogólnokształcących i techników

Wskaźniki EWD dla liceów ogólnokształcących i techników wyliczane są obecnie przy użyciu bardzo złożonego instrumentarium statystycznego, przede wszystkim w zakresie skalowania wyników matury. Choć generalnie można uznać je za trafnie obrazujące efektywność pracy szkół w zakresie kształcenia ogólnego, to jednak mają one pewne istotne ograniczenia. Niestety przezwyciężenie większości z nich wydaje się obecnie bardzo trudne, a niektórych po prostu niemożliwe.

Pierwsze istotne ograniczenie odnosi się specyficznie do techników i polega na nieuwzględnieniu efektów pracy szkoły w zakresie kształcenia zawodowego, co znacznie zmniejsza zainteresowanie

wskaźnikami EWD ze strony szkół tego rodzaju. Niestety kwestia ta wydaje się obecnie niemożliwa do rozwiązania.

Drugi problem to ograniczenie zestawu wskaźników jednorocznych, umożliwiających prowadzenie analiz wewnątrzszkolnych, jedynie do wskaźnika EWD w zakresie matematyki. Stworzenie wskaźników jednorocznych obejmujących kilka przedmiotów napotyka na problem przeliczania wyników poszczególnych części egzaminu na wartości ogólnych, kompozycyjnych miar osiągnięć. Schematy służące do takiego przeliczania byłyby bowiem bardzo złożone i trudne do zaimplementowania w Kalkulatorze EWD. Niemniej wielu użytkowników Kalkulatora EWD zgłasza zapotrzebowanie na przygotowanie wskaźników w zakresie języka polskiego. Wydaje się to postulat możliwy do spełnienia, przy wykorzystaniu podobnych narzędzi jak wykorzystywane obecnie do wyliczania wskaźników w zakresie matematyki, choć dodatkową komplikację stanowi fakt, że przy wyliczaniu miar osiągnięć uczniów należałoby również uwzględnić wybór tematu wypracowania. Niestety, ze względu na niepożądane cechy psychometryczne arkuszy maturalnych w zakresie języka polskiego, przede wszystkim ich niezbyt wysoką rzetelność i, co za tym idzie, niezbyt silny związek z wynikami egzaminu gimnazjalnego, wskaźniki te mogą cechować się znaczną niedokładnością. Działania zmierzające do ich przygotowania są obecnie w toku, a o ich efektach będzie można się przekonać najprawdopodobniej jesienią tego roku.

Trzeci problem, o którym warto wspomnieć, to fakt, że w modelowaniu EWD możemy uwzględnić wyniki tylko tych absolwentów szkół maturalnych, którzy przystępują do egzaminu dojrzałości. O ile w liceach ogólnokształcących dla młodzieży jest to problem marginalny, to w pozostałych typach szkół jest on znaczący. Niestety brak w bazach danych egzaminacyjnych informacji o absolwentach, którzy nie podchodzą do matury, uniemożliwia jakiekolwiek analizy tego zagadnienia.

Trudno powiedzieć, jaki wpływ na własności modeli i wskaźników EWD będą mieć zmiany w formule matury 2015. Nie wprowadzają one co prawda żadnych elementów, które wymagałyby dodatkowego komplikowania modeli, jednak jak na razie otwarte pozostaje pytanie o własności psychometryczne egzaminu w nowej formule. Warto też zauważyć, że w związku ze zmianami w egzaminie gimnazjalnym, jakie zaszły w 2012 r., konieczne będzie przeanalizowanie kwestii, czy w przypadku wskaźników w zakresie języka polskiego i matematyki nie byłoby lepiej zawęzić miary osiągnięć „na wejściu" tylko do wyników testu odpowiednio z języka polskiego i matematyki.

Bibliografia

Ballou, D., Sanders, W. i Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics* 29, 37–65.

Braun, H. i Wainer, H. (2007). Value-added modeling. w Rao, C.R. i Sinharay, S. (red.), *Handbook of Statistics* 26: Psychometrics, s. 475–501. Amsterdam, Elsevier.

Conijn, J. M., Emons, W. H. M., i Sijtsma, K. (2014). Statistic lz -Based Person-Fit Methods for Noncognitive Multiscale Measures. *Applied Psychological Measurement*, 38(2), 122–136.

Dolata, R. (red.) (2007). *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania.* Warszawa: Centralna Komisja Egzaminacyjna.

Dolata, R. (2006a). *Edukacyjna wartość dodana w komunikowaniu wyników egzaminów zewnętrznych.* Egzamin. Biuletyn Badawczy CKE, 8, s. 9-20.

Dolata, R. (2006b). *Efektywność nauczania w gimnazjach miasta X. Analiza edukacyjnej wartości dodanej.* Egzamin. Biuletyn Badawczy CKE, 8, s. 28-37.

Dolata, R. (2008). *Szkoła-segregacje-nierówności*. Warszawa, Wydawnictwa Uniwersytetu Warszawskiego.

Dolata, R., Hawrot, A., Hummeny, G., Jasińska, A., Koniewski M., Majkut P. i Żółtak T. (2013). *Trafność metody edukacyjnej wartości dodanej dla gimnazjów*. Warszawa, Instytut Badań Edukacyjnych.

Dolata, R., Jasińska, A. i Modzelewski, M. (2012). Wykorzystanie krajowych egzaminów jako instrumentu polityki oświatowej na przykładzie różnicowania się gimnazjów w dużych miastach. *Polityka Społeczna*, nr tematyczny 1, s. 41-47.

Dolata, R. i Pokropek, A. (2012). Czy warto urodzić się w styczniu? Wiek biologiczny a wyniki egzaminacyjne. w: B. Niemierko i M. K. Szmigel (red.), *Regionalne i lokalne diagnozy edukacyjne: XVIII Krajowa Konferencja Diagnostyki Edukacyjnej, Wrocław, 21–23 września 2012 r.* Kraków: Grupa Tomami.

Dolata, R., Hawrot, A., Humenny, G., Jasińska-Maciążek, A., Koniewski, M. i Majkut, P. (2014). *Kontekstowy model oceny efektywności nauczania po pierwszym etapie edukacyjnym*. Warszawa: Instytut Badań Edukacyjnych.

Hanushek, E. A. (2003). The Failure of Input-based Schooling Policies. *Economic Journal* 113, February, s. F64-F98.

Jacob, B. A. i Levitt, S. D. (2003a). Catching cheating teachers: The results of an unusual experiment in implementing theory. *National Bureau of Economic Research*.

Jacob, B. A. i Levitt, S. D. (2003b). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, 118(3), 843–877.

Jasińska, A. i Modzelewski, M. (2014). Testy osiągnięć szkolnych TOS 3 jako przykład narzędzia skonstruowanego z wykorzystaniem modelu Rascha. *Edukacja*, 2(127), 85–107.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4), 277–298.

Karwowski, M. (Red.). (2013). *Ścieżki rozwoju edukacyjnego młodzieży – szkoły pogimnazjalne*. Warszawa: Wydawnictwo Instytutu Filozofii i Socjologii PAN.

Korobko, O. B., Glas, C. A. W., Bosker, R. J. i Luyten, J. W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, 45(2), 139–157.

Markman, J. M., Hanushek, E. A., Kain, J. F. i Rivkin S. G., (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, vol. 18(5), s. 527-544.

McCafrey, D. F., Lockwood, J. R., Koretz, D. M. i Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, Kalifornia: RAND.

McCafrey, D. F., Lockwood, J. R., Koretz, D. M., Louis, T. A. i Hamilton, L. S. (2004). Models for value-added modeling of teacher efects. *Journal of Educational and Behavioral Statistics*, 29, 67–101.

Meyer, R. (1997). Value-Added Indicators of School Performance: A Primer. *Economics of Education Review*, Vol. 16, No.3, s. 283-301.

OECD. (2008). *Measuring improvements in learning outcomes: best practices to assess the value- added of schools.* Paryż: OECD.

Pokropek, A. (2011). Matura z języka polskiego. Wybrane problemy psychometryczne. W: B. Niemierko i M. K. Szmigel (red.), *Ewaluacja w edukacji: koncepcje, metody, perspektywy.* Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Pokropek, A. (red.). *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania*. Warszawa: Instytut Badań Edukacyjnych.

Pokropek, A. i Żółtak, T. (2012a). *Trzyletni wskaźnik egzaminacyjny.* Dokumentacja techniczna. Wersja 2.0. http://2013.ewd.edu.pl/downloads/Dokumentacja%20techniczna%20v2.0.pdf

Pokropek, A. i Żółtak, T. (2012b). Nowe modele jednorocznej EWD. W: K. Szmigiel i B. Niemierko (red.), *Regionalne i lokalne diagnozy edukacyjne: XVIII Krajowa Konferencja Diagnostyki Edukacyjnej, Wrocław, 21–23 września 2012 r.* Kraków: Grupa Tomami.

Raudenbush, S. W. i Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20, 307–335.

Rubin, D. B., Stuart E. A. i Zanutto, E. L, (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics* 29:103–116.

Schagen, I. i Hutchinson, D. (2003). Adding value in educational research – the marriage of data and analitical power. *British Educational Research Journal*, vol. 29, no. 5.

Stożek, E. (2008). *Analiza wyników egzaminów zewnętrznych z wykorzystaniem metody EWD w ewaluacji wewnątrzszkolnej.* http://2013.ewd.edu.pl/materialy-szkoleniowe/broszura_ewd.pdf

Stożek, E. (2009). EWD w ręku dyrektora szkoły. *Dyrektor Szkoły* 12(192)/2009, s.I-VIII.

Stożek, E. (2010). *Analiza wyników egzaminów zewnętrznych z wykorzystaniem metody EWD w ewaluacji wewnątrzszkolnej. Materiał pomocniczy dla dyrektorów i nauczycieli gimnazjów* http://2013.ewd.edu.pl/materialy-szkoleniowe/broszura_2010.pdf

Żółtak, T. (2013a). *Statystyczne modelowanie wskaźników edukacyjnej wartości dodanej – podsumowanie polskich doświadczeń.* Analizy IBE/02/2013.

Żółtak, T. (2013). Gimnazjalne wskaźniki EWD. w: R. Dolata (red.), *Trafność metody edukacyjnej wartości dodanej. Raport podsumowujący wyniki badania podłużnego w gimnazjach*. Warszawa, Instytut Badań Edukacyjnych w Warszawie, s.90-102

5. Wyniki egzaminów zewnętrznych w pracy szkoły

Katarzyna Matuszczak, Olga Wasilewska

Wstęp

Szkoły w Polsce od wielu już lat mają dostęp do wyników egzaminacyjnych, które mogą wykorzystywać w bardzo różnych celach. Możliwości prowadzenia analiz wzbogaciło pojawienie się wskaźników edukacyjnej wartości dodanej i Kalkulatora EWD, który jest dostępny dla gimnazjów, liceów ogólnokształcących i techników. Od niedawna możliwe jest też wyliczenie wskaźników EWD dla klas IV-VI szkół podstawowych, pojawiają się też nowe narzędzia, takie jak np. porównywalne wyniki egzaminacyjne. W jakim stopniu i w jaki sposób szkoły korzystają z możliwości, które daje im system egzaminacyjny? Na ile korzystanie z danych egzaminacyjnych jest efektem presji zewnętrznej, a na ile wynika z przekonania, że są w one w stanie pomóc w doskonaleniu nauczania na poziomie szkoły? Jakie obowiązki szkół w zakresie wykorzystywania wyników egzaminacyjnych wynikają z regulacji prawnych – związanych z nadzorem pedagogicznym? Jakie problemy napotykają szkoły w wykorzystaniu danych egzaminacyjnych? Jak organizują proces przeprowadzania analiz? A wreszcie, czy wnioski z analiz przekładają się na działania służące rozwojowi szkół? Postaramy się przybliżyć odpowiedzi na te pytania, korzystając z wyników badań oraz danych gromadzonych przez kuratoria oświaty w trakcie ewaluacji zewnętrznej.

W rozdziale tym spojrzymy zatem na kwestię egzaminów zewnętrznych z perspektywy szkoły. W pierwszej części rozdziału opiszemy, co o analizach danych egzaminacyjnych mówią regulacje prawne i zalecenia dotyczące ewaluacji zewnętrznej i wewnętrznej. W szczególny sposób widoczne jest w tym kontekście – charakterystyczne także dla innych systemów edukacyjnych – napięcie między wykorzystaniem danych na potrzeby rozliczalności: odpowiedzialności placówki za efekty jej działań, a potrzebami samodoskonalenia i rozwoju szkoły (por. m.in. Earl i Fullan, 2003; Faubert, 2009, OECD, 2013). Możemy mówić bowiem o różnych, w pewnym stopniu przeciwstawnych celach analiz danych. Te same mechanizmy mają za zadanie z jednej strony dostarczyć środowisku (w tym np. rodzicom, władzom samorządowym) informacji o jakości prowadzonych działań w kontekście kontroli wypełniania określonych standardów i procedur, a z drugiej pełnić funkcję prorozwojową – wspierać szkoły we wprowadzaniu usprawnień dydaktycznych i organizacyjnych. Szkoły natomiast mają sprostać tym obydwu zadaniom jednocześnie. Temu, jak sobie radzą z tym wyzwaniem, przyglądamy się w kolejnych częściach tego rozdziału. Cele i zakres analiz wyników egzaminacyjnych prowadzonych w szkołach omawiane są w części drugiej. Natomiast w części trzeciej bliżej przyglądamy się różnym podejściom szkół do prowadzenia analiz danych egzaminacyjnych. Przybliżamy też praktyczne problemy, jakie napotykają placówki, prowadząc analizy i starając się je wykorzystywać w różnych procesach. Na wykorzystanie danych przez kadrę szkół mają wpływ różne czynniki, o których piszemy w części czwartej rozdziału. Związane są one z charakterem samych danych, organizacją pracy szkoły, kompetencjami kadry szkół oraz ich przekonaniami i postawami. Nie bez znaczenia jest również dostęp placówki do wspomagania, jak i oczekiwania otoczenia szkoły związane z wynikami egzaminacyjnymi.

5.1. Uwarunkowania prawne – wyniki egzaminacyjne w ramach systemu nadzoru pedagogicznego

Jednym z głównych wyzwań związanych z ewaluacją i ocenianiem pracy szkoły, przed jakim stają systemy edukacyjne w różnych krajach, jest ich spójność i odpowiednie powiązanie różnych jego elementów (OECD, 2013). Warto więc w pierwszej kolejności przyjrzeć się miejscu danych egzaminacyjnych w systemie nadzoru pedagogicznego. Skupimy się przy tym na kluczowej – z perspektywy oceny pracy szkoły – części nadzoru pedagogicznego, tj. ewaluacji. W polskim systemie edukacji dzieli się ją na ewaluację zewnętrzną i wewnętrzną. Obie formy ewaluacji mają wpływ na prowadzenie analiz danych egzaminacyjnych w szkołach.

5.1.1. Dane egzaminacyjne w wymaganiach wobec szkół

Ramy odniesienia dla prowadzenia ewaluacji zewnętrznej w latach szkolnych 2009/2011 – 2014/2015 stanowiły określone w rozporządzeniu w sprawie nadzoru pedagogicznego[51] wymagania wobec szkół i placówek, obejmujące – zgodnie z założeniami – priorytetowe i strategiczne z perspektywy państwa kwestie. Do końca roku szkolnego 2012/2013 w ramach ewaluacji zewnętrznych badano spełnianie 17 wymagań uporządkowanych tematyczne w 4 obszarach. W 2013 roku zmodyfikowano brzmienie wymagań, zmniejszono ich liczbę do 12 oraz zrezygnowano z przypisywania wymagań do obszarów. Ewaluacje zewnętrzne prowadzone są przez specjalnie przeszkolonych wizytatorów ds. ewaluacji, którzy zbierają i analizują informacje dotyczące funkcjonowania szkoły w odniesieniu do poszczególnych wymagań, a także wskazują, czy szkoła je spełnia. Zgodnie z przepisami obowiązującymi do końca roku szkolnego 2014/2015, wizytatorzy określali poziom spełnienia wymagań – w skali od A do E (A – bardzo wysoki, a E – niski), przy czym szczegółowo określone były charakterystyki wymagań na dwóch poziomach: B (wysokim) i D (podstawowym). W lutym 2015 r. dodano w ustawie o systemie oświaty art. 21a, w którym znalazły się zapisy dotyczące wymagań wobec szkół i placówek a minister właściwy do spraw oświaty i wychowania został zobowiązany do wydania odrębnego rozporządzenia w tym zakresie. 1 września 2015 r. weszły w życie rozporządzenia Ministra Edukacji Narodowej: z dnia 6 sierpnia 2015 r. w sprawie wymagań wobec szkół i placówek oraz z dnia 27 sierpnia 2015 r. w sprawie nadzoru pedagogicznego. W konsekwencji wprowadzonych zmian obecnie wizytatorzy przeprowadzający ewaluacje zewnętrzne ustalają, czy szkoła spełnia badane wymagania bez określania poziomu spełniania tych wymagań.

Można wskazać dwa poziomy powiązań pomiędzy wynikami egzaminacyjnymi, a ewaluacją zewnętrzną i wymaganiami wobec szkół. Po pierwsze na podstawie wymagań sprawdzanych w trakcie ewaluacji szkoły są zobowiązane do analizowania wyników egzaminów zewnętrznych, wyciągania wniosków, oraz podejmowania na tej podstawie działań służących podnoszeniu jakości procesów edukacyjnych. Ponadto wyniki egzaminacyjne uzyskiwane przez szkołę stanowią także punkt odniesienia, źródło informacji wykorzystywane w trakcie ewaluacji przez wizytatorów do oceny efektywności działań prowadzonych przez szkołę. Badanie w procesie ewaluacji spełniania przez szkołę wymagań dotyczących efektów kształcenia związane jest z pozyskiwaniem danych na ten temat z dostępnych wizytatorom źródeł, a w szczególności związane jest z wykorzystaniem wyników egzaminacyjnych danej szkoły i wyników EWD.

Analizując pierwszy poziom powiązań należy przyjrzeć się w pierwszej kolejności obowiązującemu od początku roku szkolnego 2013/2014 Wymaganiu 11. „Szkoła lub placówka, organizując procesy edukacyjne, uwzględnia wnioski z analizy wyników sprawdzianu, egzaminu gimnazjalnego, egzaminu maturalnego, egzaminu potwierdzającego kwalifikacje zawodowe i egzaminu potwierdzającego kwalifikacje w zawodzie oraz innych badań zewnętrznych i wewnętrznych" (por. tabela 5.1).

---

[51] *Rozporządzenie Ministra Edukacji Narodowej z dnia 7 października 2009 r. w sprawie nadzoru pedagogicznego (Dz.U. Nr 168, poz. 1324). Rozporządzenie Ministra Edukacji Narodowej z dnia 10 maja 2013 r. zmieniające rozporządzenie w sprawie nadzoru pedagogicznego (Dz.U. poz. 560).*

Tabela 5.1. Wymagania dotyczące analizy wyników egzaminacyjnych obowiązujące w szkołach

WYMAGANIA DOTYCZĄCE ANALIZY WYNIKÓW EGZAMINACYJNYCH OBOWIĄZUJĄCE W SZKOŁACH	
Od roku szkolnego 2013/2014 po modyfikacjach z 2015r. :	
Wymaganie: „Szkoła lub placówka, organizując procesy edukacyjne, uwzględnia wnioski z analizy wyników sprawdzianu, egzaminu gimnazjalnego, egzaminu maturalnego, egzaminu potwierdzającego kwalifikacje zawodowe i egzaminu potwierdzającego kwalifikacje w zawodzie oraz innych badań zewnętrznych i wewnętrznych".	
Charakterystyka wymagania na poziomie podstawowym	Charakterystyka wymagania na poziomie wysokim
W szkole lub placówce analizuje się wyniki sprawdzianu i egzaminów oraz wyniki ewaluacji zewnętrznej i wewnętrznej. Analizy prowadzą do formułowania wniosków i rekomendacji, na podstawie których nauczyciele planują i podejmują działania służące podnoszeniu jakości procesów edukacyjnych.[A] Działania prowadzone przez szkołę lub placówkę są monitorowane i analizowane, a w razie potrzeb – modyfikowane.	W szkole lub placówce wykorzystuje się wyniki badań zewnętrznych innych niż wyniki sprawdzianu i egzaminu[B] i prowadzi badania wewnętrzne, odpowiednio do potrzeb szkoły lub placówki, w tym badania osiągnięć uczniów i losów absolwentów.
Od roku szkolnego 2009/2010 do roku szkolnego 2012/2013:	
Wymaganie: „Analizuje się wyniki sprawdzianu/ egzaminu gimnazjalnego/ egzaminu maturalnego i egzaminu potwierdzającego kwalifikacje zawodowe".	
Charakterystyka wymagania na poziomie podstawowym	Charakterystyka wymagania na poziomie wysokim.
Wyniki sprawdzianu i egzaminów są analizowane w celu poprawy jakości pracy szkoły lub placówki. W szkole lub placówce są wdrażane wnioski z analizy wyników sprawdzianu i egzaminów.	Do analizy wyników sprawdzianu i egzaminów wykorzystuje się różnorodne metody analizy wyników. Wdrażane w szkole lub placówce wnioski z analizy wyników sprawdzianu i egzaminów przyczyniają się do wzrostu efektów kształcenia.

Źródło: Rozporządzenie Ministra Edukacji Narodowej z dnia 7 października 2009 r. w sprawie nadzoru pedagogicznego. Rozporządzenie Ministra Edukacji Narodowej z dnia 10 maja 2013 r. zmieniające rozporządzenie w sprawie nadzoru pedagogicznego. Rozporządzenie Ministra Edukacji Narodowej z dnia 6 sierpnia 2015 r. w sprawie wymagań wobec szkół i placówek.

Istotą tego wymagania nie jest sama konieczność analizowania przez szkołę dostępnych danych (wyników egzaminów, a także rezultatów ewaluacji wewnętrznej i zewnętrznej), ale wyciąganie wniosków z tych analiz i uwzględnienie ich w prowadzonych działaniach, które to następnie mają także być przedmiotem dalszej refleksji (monitorowanie, modyfikowanie w sytuacjach, gdy jest taka potrzeba). Szkoła realizuje działania na poziomie wysokim, jeżeli wykorzystuje dodatkowo wyniki badań zewnętrznych i prowadzi odpowiednio dopasowane do potrzeb szkoły badania własne, w tym badania osiągnięć uczniów i losów absolwentów (w tym przypadku nie chodzi jednak o ewaluację wewnętrzną, której realizacja jest uwzględniana w ramach wymagania 12 „Zarządzanie szkołą lub placówką służy jej rozwojowi"). Zarówno w przypadku wykorzystywanych badań zewnętrznych, jak i badań własnych, kluczowe znowu jest ich wykorzystanie przez szkołę.

[A] Sformułowanie „służące podnoszeniu jakości procesów edukacyjnych" dodane w rozporządzeniu z sierpnia 2015 roku.
[B] Sformułowanie „innych niż wyniki sprawdzianu i egzaminu" dodane w rozporządzeniu z sierpnia 2015 roku.

Sam sposób prowadzenia analiz, stosowane metody, ich poprawność czy wartościowość wyciąganych wniosków nie jest przedmiotem oceny przy ewaluacji zewnętrznej. Niemniej jednak, zgodnie ze wskazówkami dla wizytatorów, zauważone w szkolnych analizach nieprawidłowości nie powinny być ignorowane – wizytatorzy nie mogą też „akceptować błędów popełnionych przez respondentów i na tej podstawie formułować wnioski, które są nieuprawnione" (Gocłowska, 2013, s. 50). We wcześniejszym brzmieniu wymagania, obowiązującego w latach 2009–2013 („Analizuje się wyniki sprawdzianu/ egzaminu gimnazjalnego/ egzaminu maturalnego i egzaminu potwierdzającego kwalifikacje zawodowe"), kwestia organizacji i sposobu prowadzenia analiz przez szkołę była uwzględniana, zresztą sama nazwa wymagania odnosiła się do „analizy" egzaminu, a nie „uwzględniania" wniosków z analiz. Wizytatorzy przyglądali się sposobom prowadzenia analiz i stosowanym metodom, chociaż wdrażanie wniosków było również kluczowe i wymagane też już na poziomie podstawowym. W porównaniu z wcześniejszą wersją wymagania, w 2013 r. rozszerzono także zakres informacji na temat pracy szkoły, które powinny być spożytkowane przez placówki – nie tylko wyniki egzaminów zewnętrznych, ale też wyniki ewaluacji zewnętrznej i wewnętrznej, badań prowadzonych przez szkołę i innych badań zewnętrznych.

Kwestia analiz wyników egzaminacyjnych i wykorzystania ich wyników przez szkoły pojawia się także pośrednio, w ramach innych wymagań wobec szkół. Dotyczy to zwłaszcza wymagania 3. „Uczniowie nabywają wiadomości i umiejętności określone w podstawie programowej" (por. tabela 5.2.).

Tabela 5.2. Wymaganie 3. Uczniowie nabywają wiadomości i umiejętności określone w podstawie programowej

Wymaganie 3. Uczniowie nabywają wiadomości i umiejętności określone w podstawie programowej	
(obowiązujące od roku szkolnego 2013/2014 po modyfikacjach z sierpnia 2015r.)	
Charakterystyka wymagania na poziomie podstawowym	**Charakterystyka wymagania na poziomie wysokim**
a) W szkole lub placówce realizuje się podstawę programową z uwzględnieniem osiągnięć uczniów z poprzedniego etapu edukacyjnego. Uczniowie nabywają wiadomości i umiejętności określone w podstawie programowej i wykorzystują je podczas wykonywania zadań i rozwiązywania problemów[c].	d) Wdrażane wnioski z monitorowania i analizowania osiągnięć uczniów przyczyniają się do wzrostu efektów uczenia się i osiągania różnorodnych sukcesów edukacyjnych uczniów.
b) Podstawa programowa jest realizowana z wykorzystaniem zalecanych warunków i sposobów jej realizacji.	e) Wyniki analizy osiągnięć uczniów, w tym uczniów, którzy ukończyli dany etap edukacyjny, potwierdzają skuteczność podejmowanych działań dydaktyczno-wychowawczych.
c) W szkole lub placówce monitoruje się i analizuje osiągnięcia każdego ucznia, z uwzględnieniem jego możliwości rozwojowych, formułuje się i wdraża wnioski z tych analiz.	

Źródło: Rozporządzenie Ministra Edukacji Narodowej z dnia 10 maja 2013 r. zmieniające rozporządzenie w sprawie nadzoru pedagogicznego. Rozporządzenie Ministra Edukacji Narodowej z dnia 6 sierpnia 2015 r. w sprawie wymagań wobec szkół i placówek.

Wspomniana w powyższej tabeli w pkt. a) diagnoza osiągnięć uczniów z poprzedniego etapu edukacyjnego, może wykorzystywać dane egzaminacyjne właśnie z poprzedniego etapu (w przypadku

[c] Sformułowanie „uczniowie nabywają wiadomości i umiejętności określone w podstawie programowej i wykorzystują je podczas wykonywania zadań i rozwiązywania problemów" dodane w rozporządzeniu z sierpnia 2015 roku, wtedy też usunięty został dodatkowy zapis na poziomie wysokim: „uczniowie odnoszą sukces na wyższym etapie kształcenia lub na rynku pracy".

gimnazjów i szkół ponadgimnazjalnych): „Podczas analizy danych postaraj się zwrócić uwagę, czy nauczyciele uwzględniają wcześniejsze osiągnięcia uczniów, czy rozpoznają je, na przykład poprzez diagnozy wstępne, analizę dokumentów (świadectw, zaświadczeń z egzaminów zewnętrznych, opinii i orzeczeń wydanych przez poradnie psychologiczno-pedagogiczne), pozyskiwanie informacji od uczniów i ich rodziców" (Gocłowska, 2013, s. 20). Również przy tym wymaganiu nacisk położono na celowość prowadzonych diagnoz i wykorzystanie wynikającej z nich wiedzy przy realizacji podstawy programowej. Wiele szkół odwołuje się także do analiz wyników egzaminacyjnych w kontekście punktu c), często jako uzupełnienie innych źródeł danych.

Warto także zwrócić uwagę na wymaganie 12: „Zarządzanie szkołą lub placówka służy jej rozwojowi". Wśród różnych zagadnień oceniany jest również sposób realizacji ewaluacji wewnętrznej i sposób wykorzystania wniosków z nadzoru pedagogicznego przez szkołę. Analizowane jest między innymi to, czy ewaluacja wewnętrzna jest przeprowadzana wspólnie z nauczycielami i czy w procesie zarządzania, na podstawie wniosków wynikających z nadzoru pedagogicznego, podejmuje się działania służące rozwojowi szkoły. Biorąc pod uwagę, że dane egzaminacyjne mogą być wykorzystywane w ewaluacji wewnętrznej, to także i to wymaganie – pośrednio – wiąże się z wynikami egzaminacyjnymi. Podobnie jak przy wcześniej wspomnianych wymaganiach, także tu przedmiotem analiz wizytatorów jest przede wszystkim wykorzystanie wniosków z nadzoru. Jeśli zaś chodzi o sposób realizacji ewaluacji wewnętrznej, to nacisk kładziony jest na zaangażowanie nauczycieli w jej prowadzenie.

Kolejny poziom powiązań pomiędzy wynikami egzaminacyjnymi a ewaluacją zewnętrzną to wykorzystanie wyników egzaminacyjnych przez wizytatorów do oceny pracy szkoły. W przypadku wspomnianego wymagania 3. „Uczniowie nabywają wiadomości i umiejętności określone w podstawie programowej" (por. tabela 5.2) szkoły realizujące działania na poziomie wysokim to szkoły, w których: „wdrażane wnioski z monitorowania i analizowania osiągnięć uczniów przyczyniają się do wzrostu efektów uczenia się i osiągania różnorodnych sukcesów edukacyjnych uczniów. Wyniki analizy osiągnięć uczniów, w tym uczniów, którzy ukończyli dany etap edukacyjny, potwierdzają skuteczność podejmowanych działań dydaktyczno-wychowawczych". Co to w praktyce oznacza? Stawiane przez wizytatorów pytania badawcze dotyczące tego obszaru powinny koncentrować się na działaniach szkoły, podjętych na podstawie monitorowania osiągnięć uczniów, które przyczyniły się do wzrostu wyników egzaminacyjnych lub innych osiągnięć edukacyjnych. Przedmiotem analiz ma więc być skuteczność prowadzonych działań, a jednym z jej wyznaczników powinny być wyniki egzaminacyjne analizowane przez wizytatora w ramach analizy danych zastanych.

Pytania do analizy danych zastanych, na które wizytator ma znaleźć odpowiedź to:

– Jakie są wyniki egzaminów lub sprawdzianów w ostatnich trzech latach? Jakie są zmiany (proszę przedstawić uogólniony opis trendów, np. zmiany staninów, w zależności od typu szkoły trzyletniego wskaźnika EWD)?

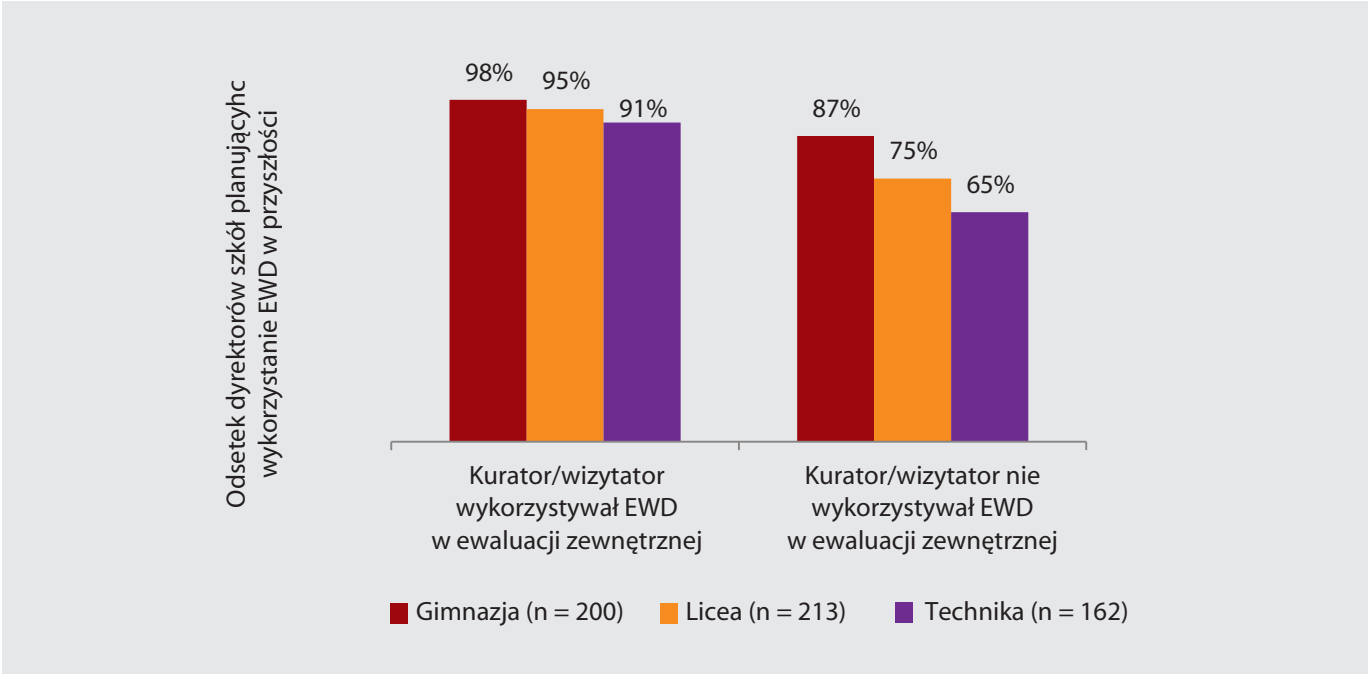
– Czy w świetle tych danych jest widoczny wzrost efektów kształcenia? W jakich obszarach można to zauważyć? (źródło: konceptualizacja wymagania)

Poza powyższym wymaganiem wyniki egzaminacyjne uczniów ze szkoły poddawanej ewaluacji zewnętrznej formalnie nie mają stanowić punktu odniesienia dla ogólnej oceny szkoły.

System nadzoru niejako ukierunkowuje, czy wręcz wymusza, określone działania kadry szkoły w zakresie wykorzystania wyników egzaminacyjnych. Wizytatorzy poprzez swoje działania pośrednio upowszechniają wśród szkół przekonanie, że powinni analizować wyniki egzaminacyjne. Dobrym przykładem są wskaźniki EWD, o których nie wspomina się w samych wymaganiach, ale występują w dodatkowych materiałach dla wizytatorów. Z badań wynika, że w około połowie badanych gimnazjów i szkół ponadgimnazjalnych, w których odbyła się ewaluacja zewnętrzna, dyrektorzy szkół zadeklarowali, że wizytator wykorzystywał u nich wskaźniki EWD do oceny pracy szkoły (Matuszczak, Zielonka i Bąbiak, 2014). Jednocześnie okazuje się, że te szkoły, w których wizytator – w opinii

dyrektora – stosował EWD podczas ewaluacji zewnętrznej, częściej deklarują swoje zainteresowanie wykorzystaniem wskaźników w przyszłości (por. rysunek 5.1).

Rysunek 5.1. Deklaracja wykorzystania wskaźników EWD w przyszłości według wykorzystywania wskaźników EWD przez kuratora oraz w podziale na typ szkoły



Źródło: Opracowanie własne na podstawie badania IBE prowadzonego techniką CATI wśród dyrektorów szkół w 2013 roku. Podstawa procentowania: dyrektorzy, których placówki były objęte ewaluacją zewnętrzną i posiadali wiedzę o wykorzystywaniu EWD przez kuratora/wizytatora.

Przedstawiony opis powiązań pomiędzy wynikami egzaminacyjnymi a ewaluacją zewnętrzną bazuje na założeniach dotyczących sposobu prowadzenia ewaluacji. Analizy raportów z ewaluacji zewnętrznych z pierwszych lat funkcjonowania systemu pokazywały, że praktyka różni się od założeń. Wskazuje się między innymi na problem nieprzystających uzasadnień i wniosków do przyznawanych przez wizytatorów poziomów spełniania przez szkoły wymagań, czy też błędy pojawiające się w raportach, związane na przykład z nieprawidłowym użyciem terminologii, czy przytaczaniem przez wizytatorów nieuprawnionych i nieprawdziwych wniosków z analiz egzaminacyjnych prowadzonych przez szkoły (Stożek, 2010; Skórska, Koniewski i Majkut, 2012). W *Poradniku wizytatora* autorzy podkreślają między innymi problem braku precyzyjnych uzasadnień ocen i niejasnego zastosowania kryteriów, jak również występowanie w przygotowywanych przez wizytatorów raportach z ewaluacji nieprecyzyjnych i błędnych sformułowań związanych z analizą wyników egzaminacyjnych. (Gocłowska 2013).

5.1.2. Dane egzaminacyjne w ewaluacji wewnętrznej

Ewaluacja wewnętrzna to badanie przeprowadzane w szkole lub placówce pod nadzorem jej dyrektora. Zgodnie z założeniami i zapisami rozporządzenia o nadzorze wybór przedmiotu ewaluacji wewnętrznej, sposobu jej organizacji, metod wykorzystywanych w trakcie jej prowadzenia pozostaje w gestii dyrektora – zakres i sposób prowadzenia ewaluacji wewnętrznej to autonomiczna decyzja szkoły. Kluczowe z tej perspektywy jest to, aby wyniki ewaluacji wewnętrznej były wykorzystywane do doskonalenia jakości pracy szkoły, co sprawdzane jest m.in. w ramach ewaluacji zewnętrznej.
Formalnie nie ma więc żadnych wytycznych dotyczących sposobów wykorzystania wyników egzaminacyjnych w ramach ewaluacji wewnętrznej. Ale jednym z popularniejszych wśród szkół obszarów

ewaluacji są badania efektów nauczania, które są często rozumiane właśnie jako analizy wyników osiąganych przez uczniów w egzaminach zewnętrznych. W praktyce, w wielu szkołach analizy danych egzaminacyjnych są więc włączane do ewaluacji wewnętrznej (wpisywanej w szkolny plan nadzoru pedagogicznego). W niektórych szkołach ewaluacje dotyczące tej kwestii prowadzone są co roku, na przykład wraz z ewaluacją w innym obszarze. Taki wybór jest zazwyczaj uzasadniany tym, że jest to z perspektywy pracy szkoły kluczowa kwestia. W innych placówkach analizy takie prowadzone są niezależnie od ewaluacji i postrzegane są raczej jako odrębny proces (Wasilewska, Rybińska i Muzyk, 2014, Kędracka, Matuszczak, Rappe i Stożek ,2013). Być może wiąże się to z popularnością wśród szkół badań ankietowych jako podstawowego narzędzia wykorzystywanego w ewaluacji wewnętrznej, a także wzorowania ewaluacji wewnętrznej na zewnętrznej i traktowania badań ankietowych jako niejako „wymaganego" sposobu prowadzenia ewaluacji w szkole (por. Elsner i Bednarek , 2012, Wasilewska i in., 2014). Dość często spotkać się można z sytuacją, że w szkołach tworzone są z jednej strony dokumenty zwane raportami z ewaluacji wewnętrznej, a z drugiej dokumenty z analiz wyników egzaminacyjnych – przy czym oba uwzględniają analizy wyników egzaminacyjnych.

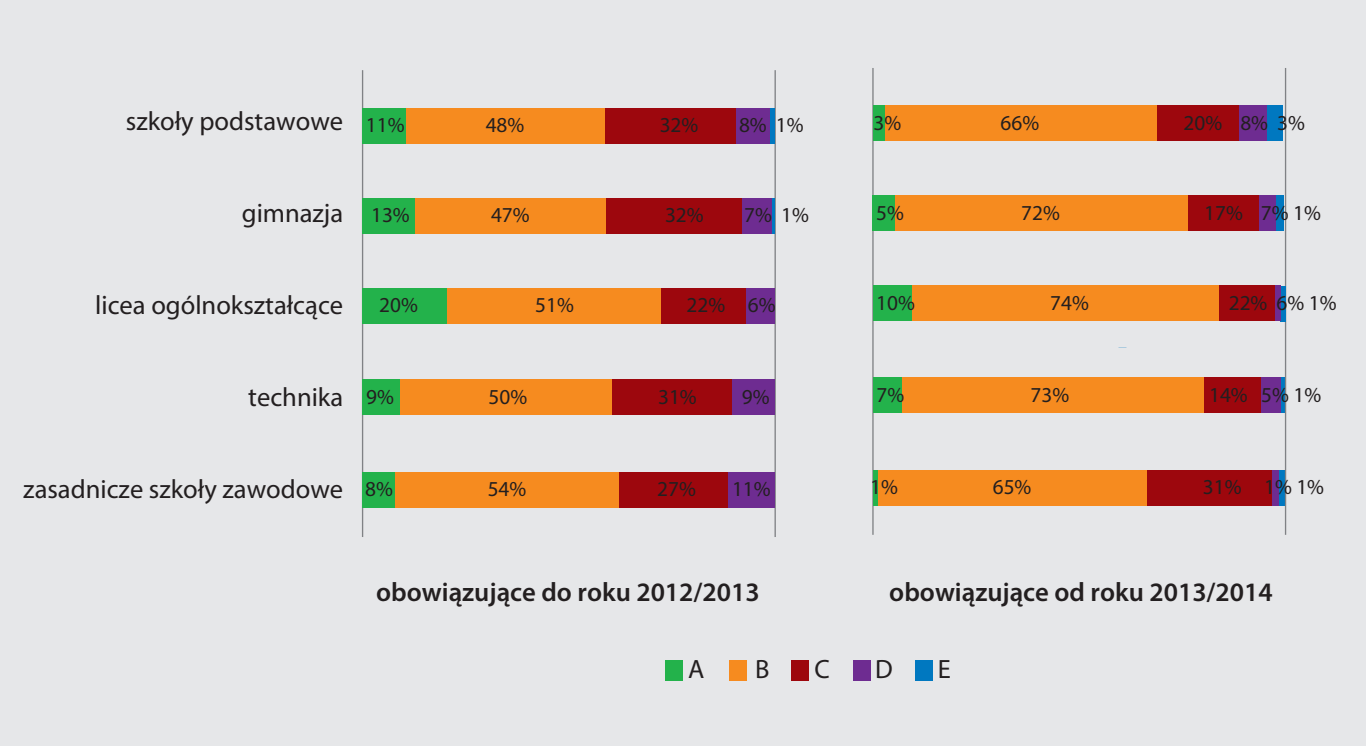
5.2. Cele i zakres wykorzystania danych egzaminacyjnych w szkołach

Biorąc pod uwagę przedstawione wymagania państwa względem szkół, można stwierdzić, że w ramach polskiego systemu edukacji kluczowym pytaniem staje się nie to, czy szkoły analizują wyniki egzaminacyjne, ale w jakim celu to robią. Jak wspomniano we wstępie rozdziału, dane egzaminacyjne mogą być analizowane w szkołach w różnych celach – zarówno rozwojowych, w przypadku których kluczowe jest znaczenie motywacji wewnętrznej użytkowników danych, jak i związanych z wymogami zewnętrznymi i koniecznością rozliczenia pracy szkoły przed instytucjami zewnętrznymi. W tym drugim przypadku można mówić raczej o motywacji zewnętrznej (Marciniak i Ronka, 2012).

5.2.1. Analiza i wykorzystanie wyników egzaminacyjnych według danych nadzoru pedagogicznego

Z analiz danych z nadzoru z lat 2009-2014 wynika, że wizytatorzy wysoko oceniają działania szkół w zakresie korzystania z wyników egzaminacyjnych. W przypadku wymagań związanych z wykorzystaniem wyników egzaminacyjnych odsetek przyznanych ocen na poziomie podstawowym i niskim (D i E) w różnych typach szkół nie przekracza 11%, a w ocenach dominuje poziom wysoki (por. rysunek 5.2). Ze względu na zmianę brzmienia wymagań warto oddzielnie przyjrzeć się wynikom ewaluacji zewnętrznych prowadzonych w pierwszych czterech latach szkolnych obowiązywania nowego systemu nadzoru oraz realizowanych od początku roku szkolnego 2013/2014. Po zmianie sformułowania wymagania dotyczącego analiz wyników egzaminacyjnych, znacznie więcej szkół wizytatorzy ocenili na poziomie wysokim (B) niż w okresie sprzed modyfikacji – w zależności od typu szkoły w pierwszym okresie było to od 47% do 54%, w drugim zaś od 65% do 74%. Wzrost ten wiąże się zarazem z mniejszym udziałem ocen najwyższych (A – poziom bardzo wysoki) i średnich (C – poziom średni). W przypadku obu wersji wymagań (sprzed i po zmianie), trochę lepiej wypadają na tle innych szkół licea ogólnokształcące.

Rysunek 5.2. Poziomy spełniania wymagań związanych z wykorzystaniem wyników egzaminacyjnych a) wymagania obowiązującego do roku szkolnego 2012/2013 b) obowiązującego od roku szkolnego 2013/2014



Źródło: Opracowanie własne na podstawie danych z platformy SEO: a) za okres 1.09.2009–31.08.2013 b) za okres 1.09.2013–31.12.2014, [pobrane 18 lutego 2015] N = 4847.

Bardziej szczegółowe informacje na temat ocen przyznawanych przez wizytatorów można uzyskać, analizując poszczególne obszary badane przez nich w ramach wymagań. Przedstawiono je w tabelach 5.3 i 5.4. W przypadku wymagania obowiązującego do roku szkolnego 2012/2013 widać wyraźnie, że prawie wszystkie szkoły uznano za spełniające wymaganie na poziomie podstawowym (w szkołach tych, zdaniem wizytatorów, prowadzone są analizy egzaminów zewnętrznych, które realizowane są w celu poprawy jakości pracy szkoły, a wnioski z tych analiz są wdrażane). Niewielkiej części szkół (od 10% w przypadku szkół podstawowych, do 16% w odniesieniu do szkół zawodowych) uzyskanie wysokiego poziomu spełniania wymagania uniemożliwił brak stosowania przez nie jakościowych i ilościowych metod analiz. Z danych nadzoru wynika również, że w ¼ szkół wdrażane przez nie wnioski z analiz wyników egzaminacyjnych nie przyczyniały się do wzrostu efektów kształcenia. Być może niższe wyniki szkół pod względem tego kryterium związane są między innymi z tym, że jest ono trudne do weryfikacji i oceny. Nie widać znaczących różnic w ocenach ze względu na typ szkoły. Jedynie w odniesieniu do stosowanych metod analiz stosunkowo słabiej wypadły szkoły zasadnicze zawodowe. Nieco lepiej oceniono z kolei efektywność działań wprowadzonych na skutek analiz w liceach ogólnokształcących.

Tabela 5.3. Odsetek szkół spełniających poszczególne obszary wymagania: *Analizuje się wyniki sprawdzianu/ egzaminu gimnazjalnego/ egzaminu maturalnego i egzaminu potwierdzającego kwalifikacje zawodowe* (obowiązującego do roku szkolnego 2012/2013)

Kryterium/typ placówki	W szkole przeprowadzana jest analiza wyników egzaminów zewnętrznych (D)	Analiza jest prowadzona w celu poprawy jakości pracy szkoły (D)	Wnioski z analizy są wdrażane (D)	W szkole stosuje się jakościowe i ilościowe metody analizy (B)	Wdrażane wnioski przyczyniają się do wzrostu efektów kształcenia (B)
szkoły podstawowe (N=2194)	99,8%	99,7%	99,4%	89,7%	72,0%
gimnazja (N=1516)	99,6%	99,6%	99,3%	89,0%	73,0%
licea ogólnokształcące (N=537)	99,8%	99,8%	99,4%	90,1%	82,5%
technika (N=385)	99,5%	99,7%	99,5%	86,7%	74,6%
zasadnicze szkoły zawodowe (N=228)	99,6%	99,6%	100,0%	84,2%	76,2%
Ogółem (N=4860)	99,7%	99,7%	99,4%	89,0%	73,9%

Źródło: opracowanie własne na podstawie danych z platformy SEO za okres 1.09.2009-31.08.2013. N=4860 [pobrane 18 lutego 2015].

Dane z ewaluacji zewnętrznych prowadzonych od roku 2013/2014 pokazują, że ocena podejścia szkół do wyników egzaminacyjnych dokonywana przez wizytatorów była podobna. Z ich analiz wynika, że prawie wszystkie szkoły spełniają poszczególne obszary na poziomie podstawowym, a więc zdaniem wizytatorów szkoły powszechnie analizują wyniki egzaminów, a także wnioski z ewaluacji zewnętrznej i wewnętrznej. W ocenie wizytatorów szkoły na podstawie tych analiz podejmują działania, które potem są przez nie monitorowane i w razie potrzeb modyfikowane (por. tabela 5.4). Nieco gorzej wizytatorzy oceniają wykorzystanie przez szkoły badań zewnętrznych (zwłaszcza w zasadniczych szkołach zawodowych) oraz prowadzenie przez nie własnych badań.

Tabela 5.4. Odsetek szkół spełniających poszczególne obszary wymagania: *Szkoła lub placówka, organizując procesy edukacyjne, uwzględnia wnioski z analizy wyników sprawdzianu, egzaminu gimnazjalnego, egzaminu maturalnego, egzaminu potwierdzającego kwalifikacje zawodowe i egzaminu potwierdzającego kwalifikacje w zawodzie oraz innych badań zewnętrznych i wewnętrznych* (obowiązującego od roku szkolnego 2013/2014)

Kryterium/typ placówki	W szkole lub placówce analizuje się wyniki sprawdzianu i egzaminów oraz wyniki ewaluacji zewnętrznej i wewnętrznej. Analizy prowadzą do formułowania wniosków i rekomendacji, na podstawie których planuje się i podejmuje działania (D)	Działania prowadzone przez szkołę lub placówkę są monitorowane i analizowane, a w razie potrzeby modyfikowane (D)	W szkole lub placówce wykorzystuje się wyniki badań zewnętrznych (B)	W szkole lub placówce prowadzi się badania odpowiednio do potrzeb szkoły lub placówki, w tym osiągnięć uczniów i losów absolwentów (B)
szkoły podstawowe (N = 1074)	97,4%	97,2%	80,1%	82,0%
gimnazja (N = 542)	98,2%	98,2%	83,5%	90,0%
licea ogólnokształcące (N = 144)	99,3%	99,3%	86,8%	95,8%
technika (N = 127)	99,2%	99,2%	83,5%	92,9%
zasadnicze szkoły zawodowe (N = 78)	98,7%	98,7%	67,9%	96,2%
Ogółem (N = 1965)	97,9%	97,8%	81,3%	86,5%

Źródło: Opracowanie własne na podstawie danych z platformy SEO za okres 1.09.2013–31.12.2014 [pobrane 18 lutego 2015]. N = 1965

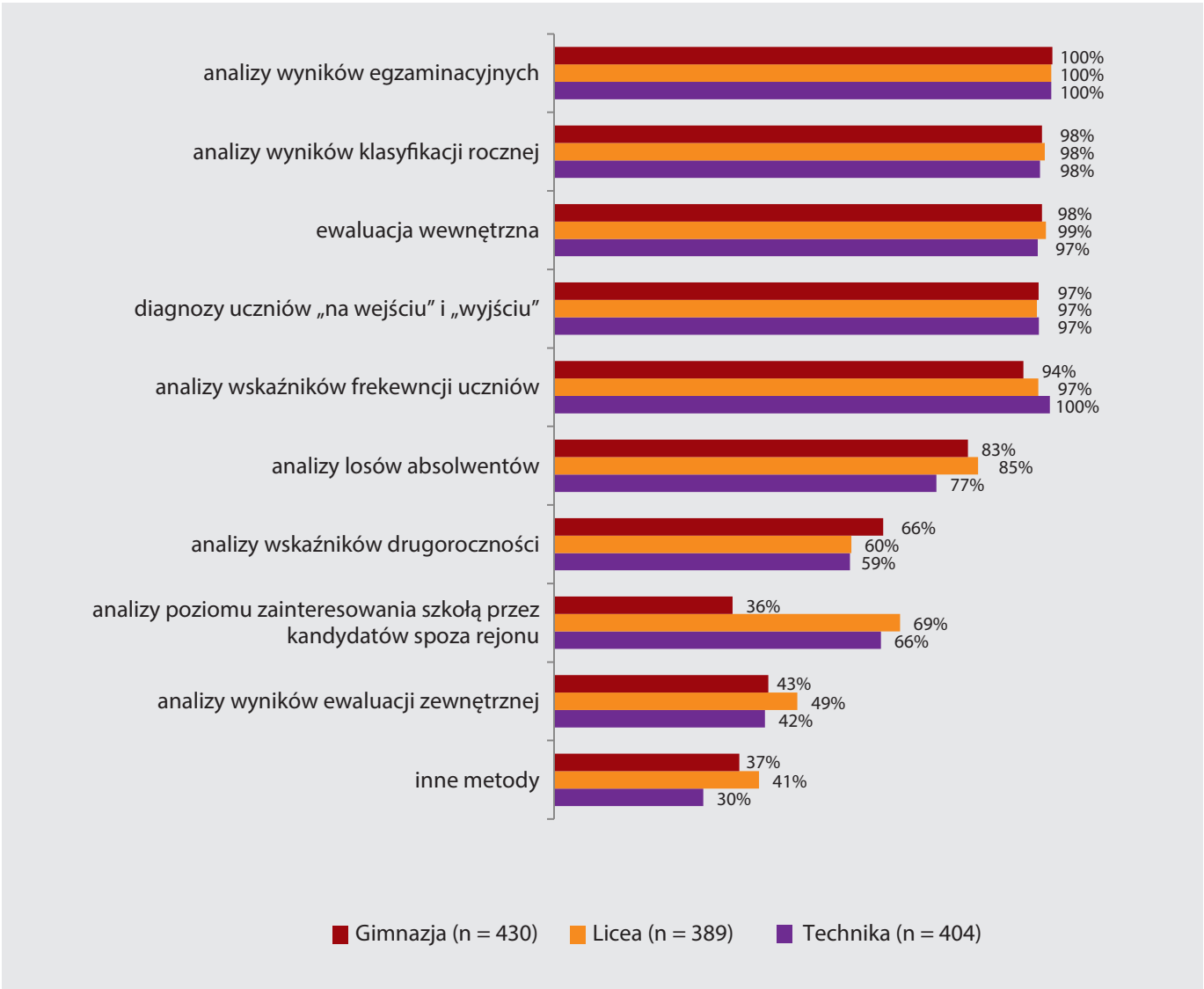
Podsumowując, obraz wyłaniający się z analizy ocen przyznawanych przez wizytatorów jest bardzo pozytywny. Prawie wszystkie szkoły poddane ewaluacji zewnętrznej w odniesieniu do tego wymagania, w obu analizowanych okresach w ocenie wizytatorów, nie tylko analizują wyniki egzaminacyjne, ale także wykorzystują wnioski z nich wynikające w pracy szkoły. Dodatkowo, zdaniem wizytatorów, większość szkół spełnia obszary na poziomie wysokim. Przy ocenie tych wyników trzeba jednak pamiętać o ograniczeniach związanych z charakterem omawianych danych. Zebrane zostały one w ramach ewaluacji zewnętrznych postrzeganych przez szkoły w kategoriach kontroli, przed którą trzeba pokazać się z jak najlepszej strony. Ponadto, jak już wspomniano, także praktyczna realizacja przez wizytatorów ewaluacji zewnętrznych odbiega niekiedy od założeń.

5.2.2. Wykorzystanie danych egzaminacyjnych i innych źródeł danych w świetle badań

Wyniki badań dostarczają także informacji na temat deklaracji szkół odnośnie do celów i zakresu wykorzystania przez nie danych egzaminacyjnych. Wyniki egzaminacyjne są tylko jednym z wielu rodzajów źródeł informacji, które mogą wykorzystywać szkoły. Jakie jest zatem miejsce wyników egzaminacyjnych wśród tych danych? W 2013 roku zapytano dyrektorów gimnazjów, liceów

i techników o to, jakie źródła informacji wykorzystują do oceny efektywności nauczania (Matuszczak i in., 2014). Okazuje się, że prawie 100% z nich deklaruje, że przeprowadza w tym celu analizy wyników egzaminacyjnych (por. rysunek 5.3). Równie często wykorzystywane są analizy wyników klasyfikacji rocznej, ewaluacji wewnętrznej, wyników badań diagnostycznych uczniów „na wejściu" i „wyjściu" danego etapu edukacyjnego oraz analiz frekwencji uczniów. Sporo, bo aż 77–88% dyrektorów deklaruje, że przeprowadza analizy losów absolwentów. Na dalszych miejscach znalazły się wskaźniki drugoroczności oraz poziomu zainteresowania szkołą przez uczniów. Niższy odsetek wykorzystania wyników ewaluacji zewnętrznej (42–49%) wynika stąd, że miała ona miejsce jeszcze nie we wszystkich szkołach. Inne, zgodnie z deklaracjami, stosowane w gimnazjach, liceach i technikach metody oceny efektywności nauczania, to najczęściej wewnętrzne badania oraz analizy postępu wiedzy i osiągnięć, wykorzystywanie wskaźników związanych z udziałem w olimpiadach i konkursach, wyników sprawdzianów wewnętrznych, ocen semestralnych, średnich uzyskiwanych przez uczniów ocen.

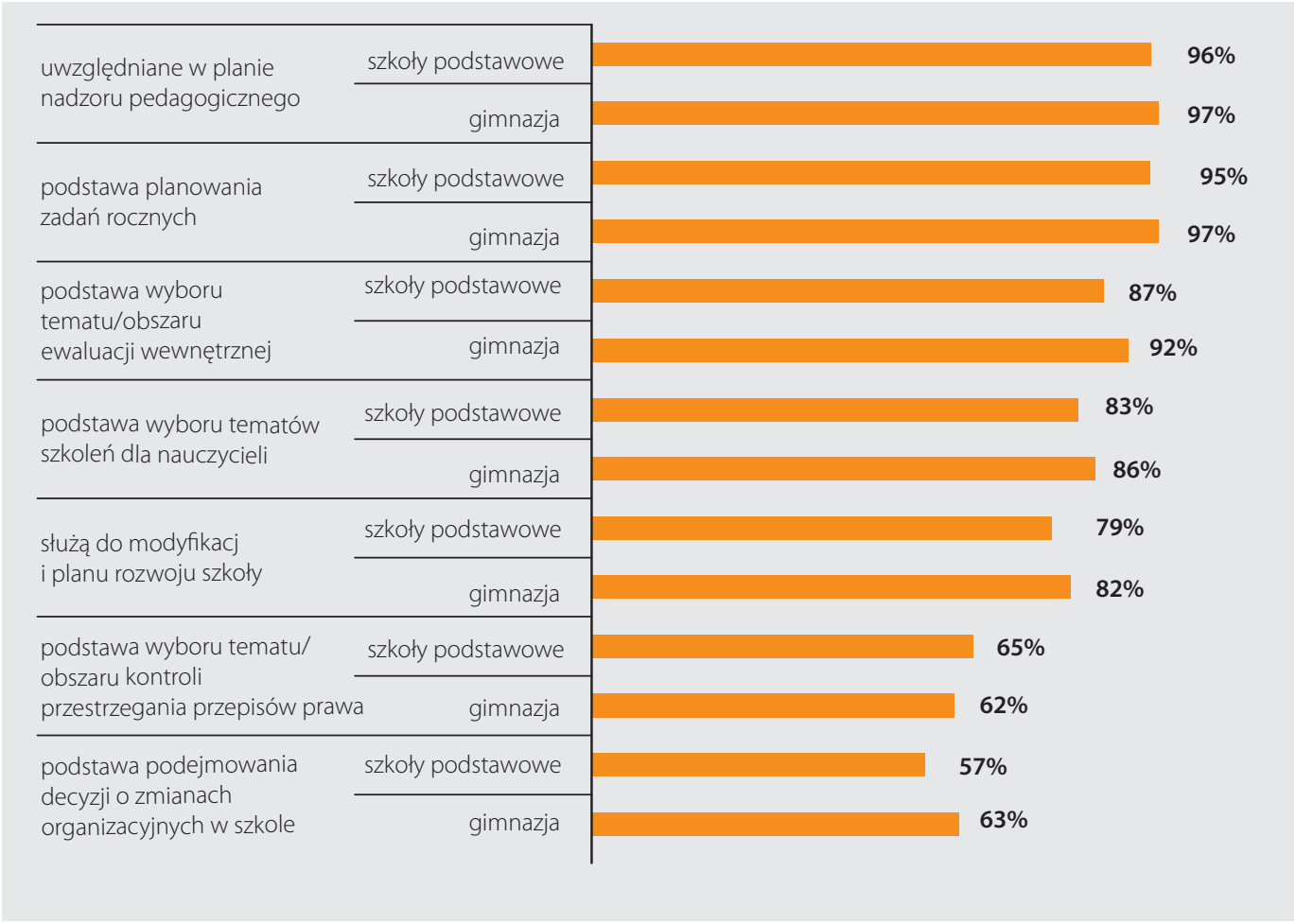
Rysunek 5.3. Metody wykorzystywane do oceny efektywności nauczania stosowane dotychczas w szkole według typu szkoły



Źródło: Opracowanie własne na podstawie badania IBE prowadzonego techniką CATI wśród dyrektorów szkół w 2013 roku. Podstawa procentowania: wszyscy zbadani.

Do czego zatem wyniki egzaminacyjne są, zdaniem kadry szkół, przede wszystkim wykorzystywane? Prawie wszyscy dyrektorzy szkół podstawowych (96%) i gimnazjów (97%), wskazali w 2012 roku w badaniach podłużnych IBE, że uwzględniają analizy wyników egzaminacyjnych w planie nadzoru pedagogicznego, oraz że wyniki są dla nich podstawą planowania zadań rocznych. Jednocześnie większość dyrektorów stwierdza, że analizy wyników egzaminacyjnych stanowią podstawę wyboru obszaru ewaluacji wewnętrznej (87%–SP, 92%–G), podstawę wyboru tematów szkoleń dla nauczycieli (83%–SP, 86%–G) i modyfikacji planu rozwoju szkoły.

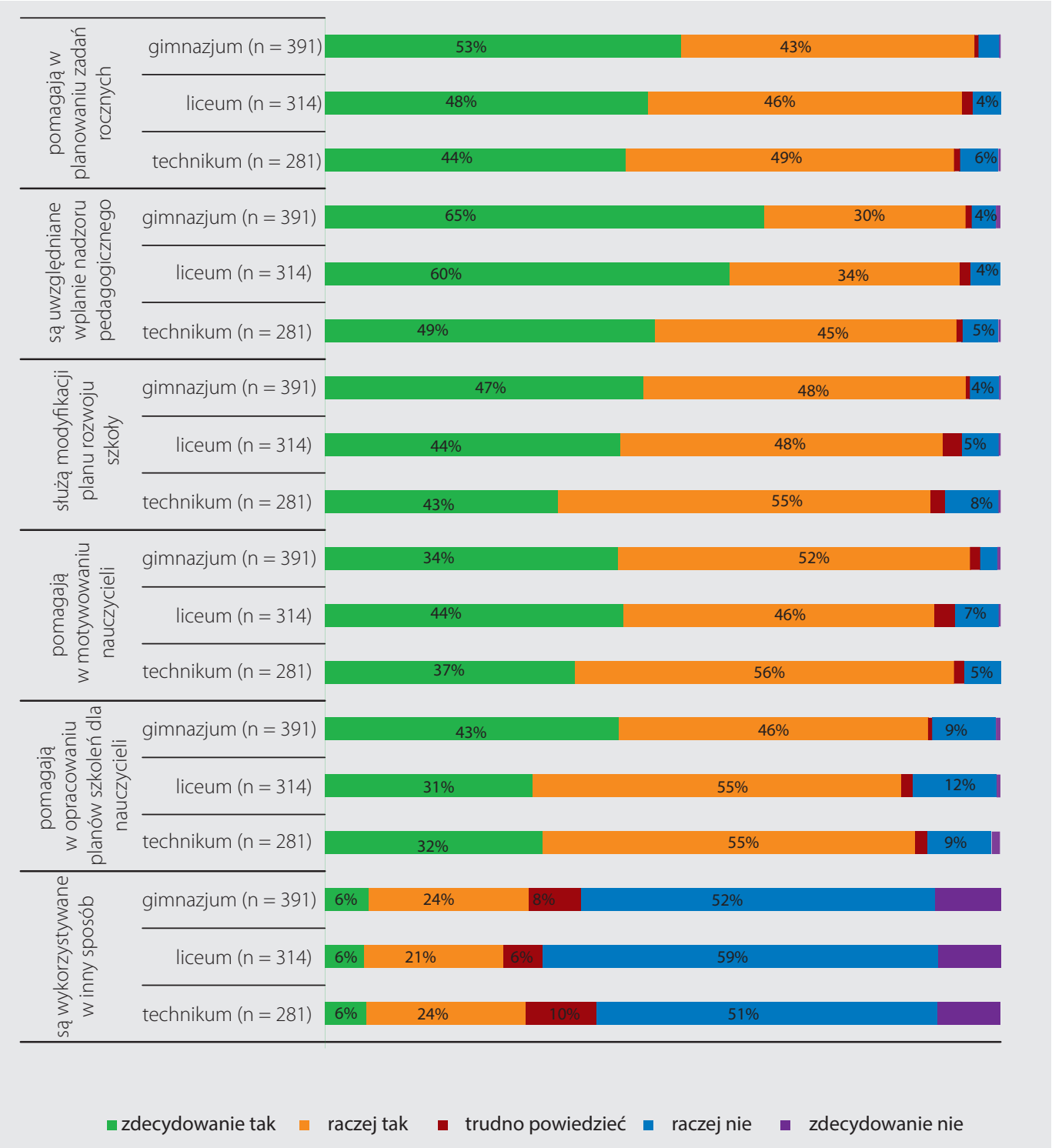
Rysunek 5.4. Deklaracja dyrektorów wykorzystywania analiz wyników egzaminacyjnych do zarządzania szkołą; 2012 r.



Opracowanie własne na podstawie danych z badań podłużnych IBE, 2012 rok (SP N = 180, gimnazja N = 150); pytania: „W jaki sposób wykorzystywane są wnioski z analizy wyników egzaminacyjnych do zarządzania Pani/Pana szkołą?" oraz „Na podstawie analizy wyników egzaminacyjnych podejmowane są decyzje…".

Podobne opinie wyrażali dyrektorzy korzystający z EWD. Wskaźniki EWD są przydatne zarówno do planowania zadań rocznych, modyfikacji planu rozwoju szkoły, przygotowania planu nadzoru pedagogicznego, jak i motywowania nauczycieli oraz opracowaniu planów szkoleń dla nauczycieli. W prawie każdym aspekcie (z wyjątkiem motywowania nauczycieli) najwyżej przydatność wskaźników EWD oceniają dyrektorzy gimnazjów, którzy jednocześnie najdłużej spośród badanych typów szkół dysponują ogólnodostępnymi wskaźnikami EWD oraz Kalkulatorem do analiz wewnątrzszkolnych.

Rysunek 5.5. Opinie dyrektorów o przydatności EWD według typu szkoły

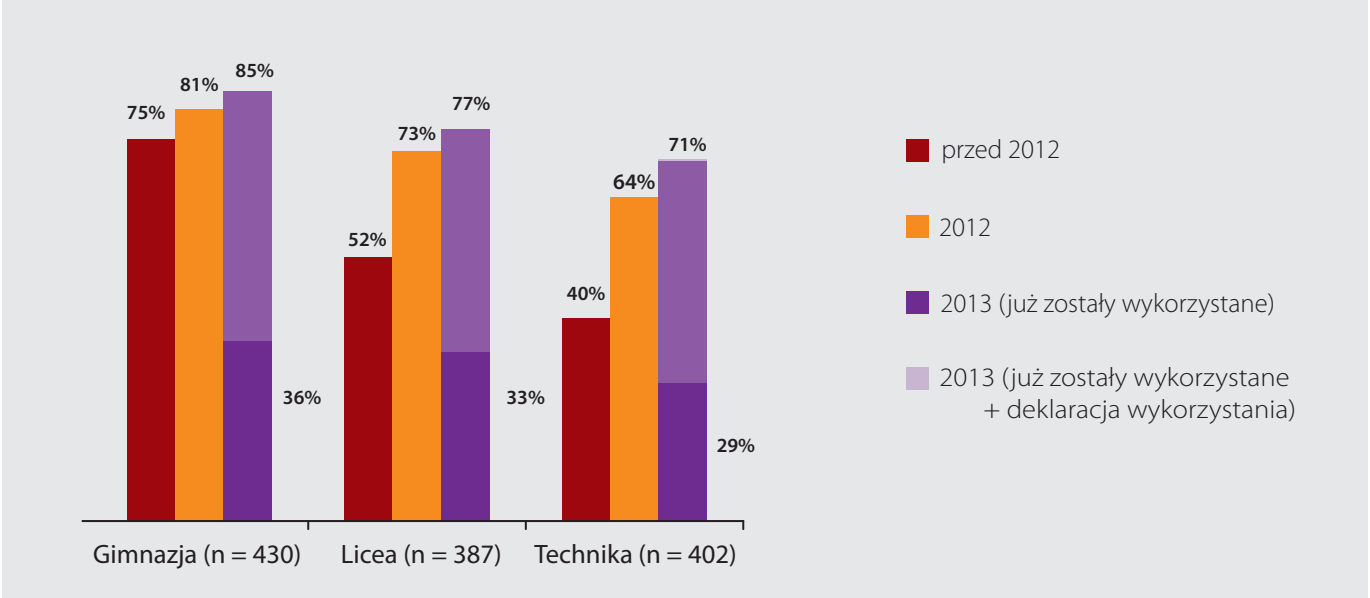


Źródło: Opracowanie własne na podstawie badania IBE prowadzonego techniką CATI wśród dyrektorów szkół w 2013 roku. Podstawa procentowania: dyrektorzy szkół, w których wykorzystuje się wskaźniki EWD do oceny efektywności nauczania.

Wyniki EWD są też – jak deklarują dyrektorzy - w większym stopniu wykorzystywane wewnątrz szkoły niż w kontaktach z jej otoczeniem. Dyrektorzy gimnazjów, liceów i techników najczęściej wskazują, że korzystają z analiz EWD w dyskusjach na radzie pedagogicznej (od 97% do 99%). Rzadziej natomiast w komunikacji z organem prowadzącym (70–78%), a najrzadziej z kuratorium oświaty (43–52%).

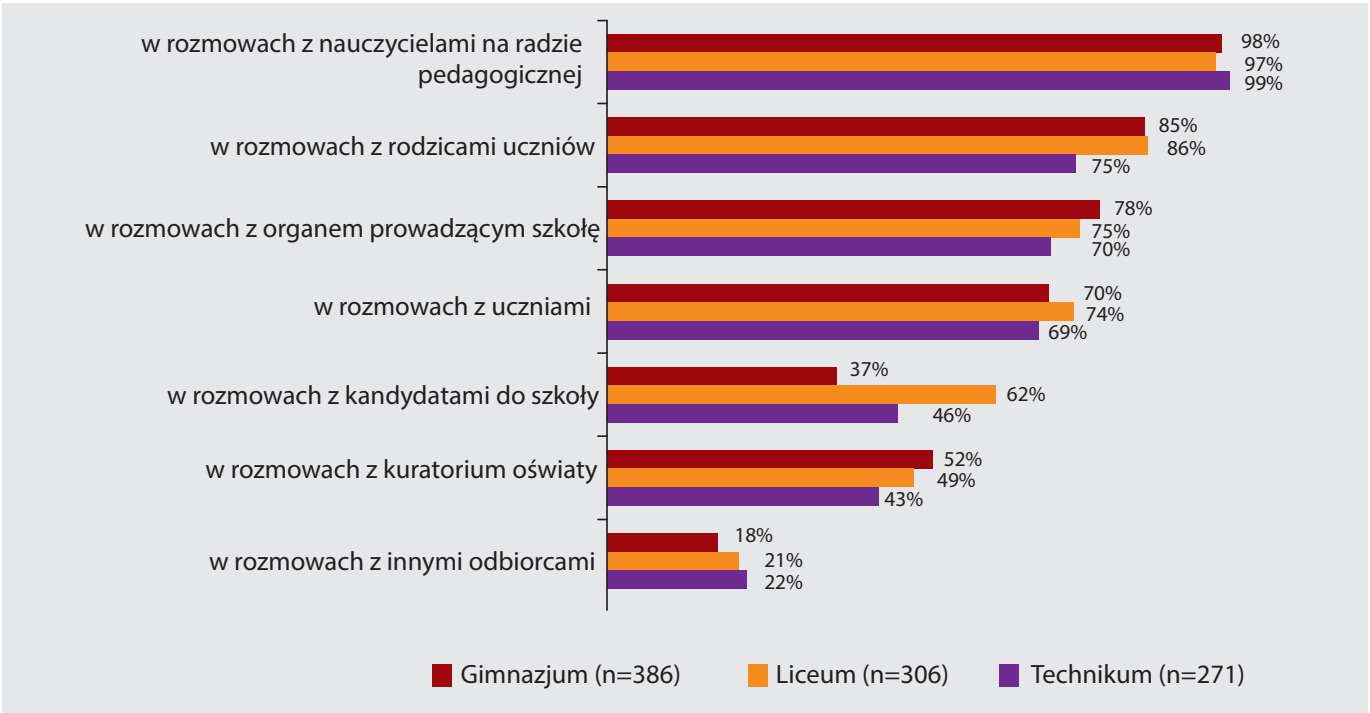
Wzrasta też odsetek szkół, w których deklaruje się korzystanie z EWD (por. Rysunek 5.6). Widać tu wyraźną przewagę gimnazjów, przy czym nie bez znaczenia jest to, że szkoły te mają dostęp do wskaźników EWD już od 2009 r., a licea i technika od 2012 r.

Rysunek 5.6. Deklaracja dyrektorów dotycząca wykorzystywania wskaźników EWD do oceny efektywności nauczania w różnych latach, według rodzaju szkoły



Źródło: Opracowanie własne na podstawie badania IBE prowadzonego techniką CATI wśród dyrektorów szkół w 2013 roku. Podstawa procentowania: dyrektorzy, z wyjątkiem tych, którzy wskazali, iż nie orientują się, czym jest EWD.

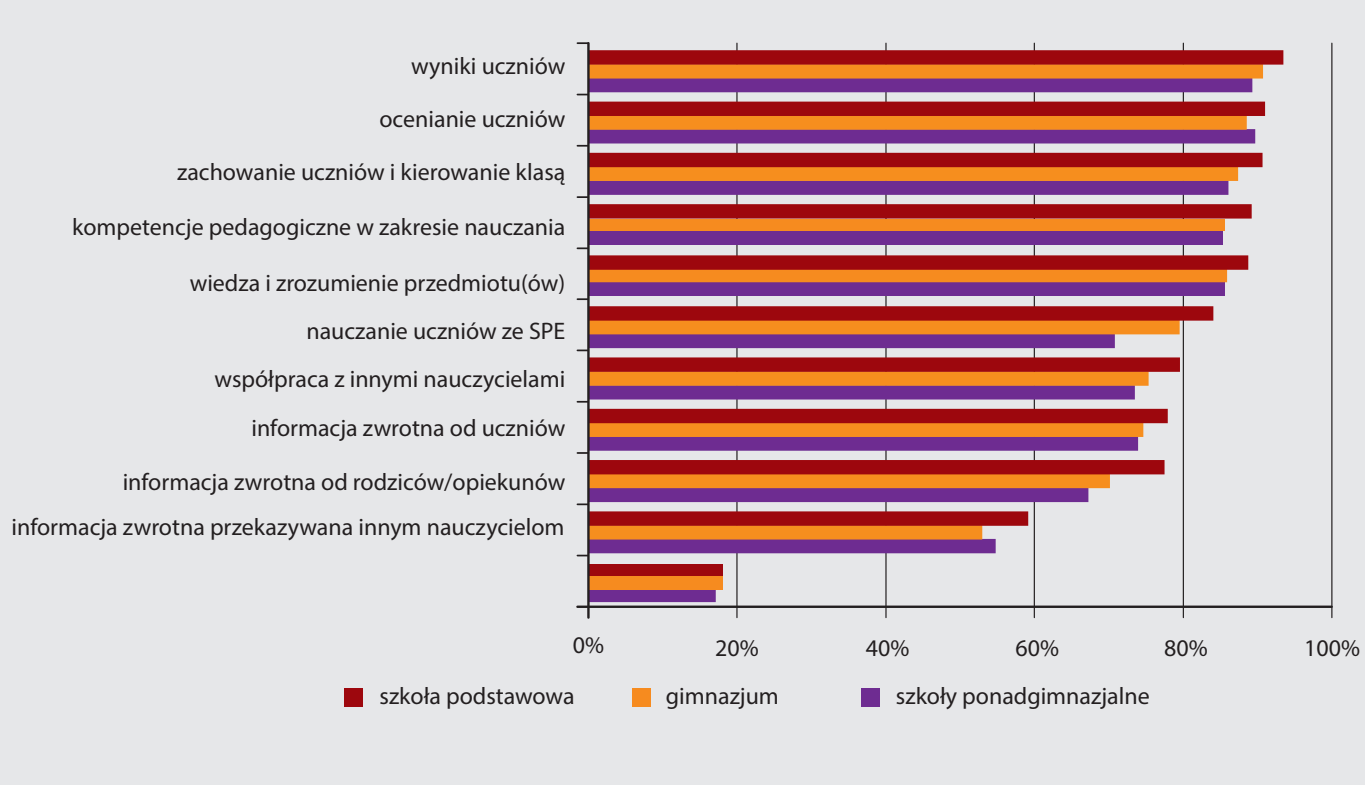
Rysunek 5.7. Deklaracje dyrektorów dotycząca powoływania się na wyniki analiz EWD, 2013 r.



Źródło: na podstawie badania IBE prowadzonego techniką CATI wśród dyrektorów szkół w 2013 roku. Podstawa procentowania: dyrektorzy szkół, w których wykorzystuje się wskaźniki EWD do oceny efektywności nauczania.

Szczególnie istotną kwestią jest potencjalne zastosowanie analiz wyników egzaminacyjnych do rozwoju praktyk dydaktycznych nauczycieli. Wyniki egzaminacyjne mogą być wykorzystywane jako element przekazywanej nauczycielom informacji zwrotnej dotyczącej ich pracy. Relacje nauczycieli uczestniczących w polskiej edycji badania TALIS 2013 pokazują, że w Polsce wśród różnych zagadnień, na które kładziono nacisk w ramach przekazywanej nauczycielom informacji dominują właśnie wyniki uczniów (rysunek 5.8). Jest to charakterystyczne dla szkół ze wszystkich etapów edukacyjnych – 89% nauczycieli szkół ponadgimnazjalnych, 91% nauczycieli gimnazjów i aż 94% nauczycieli szkół podstawowych wskazywało, że wyniki uczniów zostały uwzględnione jako bardzo ważne, bądź umiarkowanie ważne w ramach otrzymywanej przez nich informacji zwrotnej. Warto dodać, że w perspektywie ostatnich kilku lat coraz więcej nauczycieli we wszystkich krajach biorących udział w badaniu TALIS deklaruje, że w uzyskiwanej przez nich informacji zwrotnej kwestia wyników uczniów była ważnym zagadnieniem, przy czym Polska znajduje się w tym przypadku powyżej średniej (OECD, 2014).

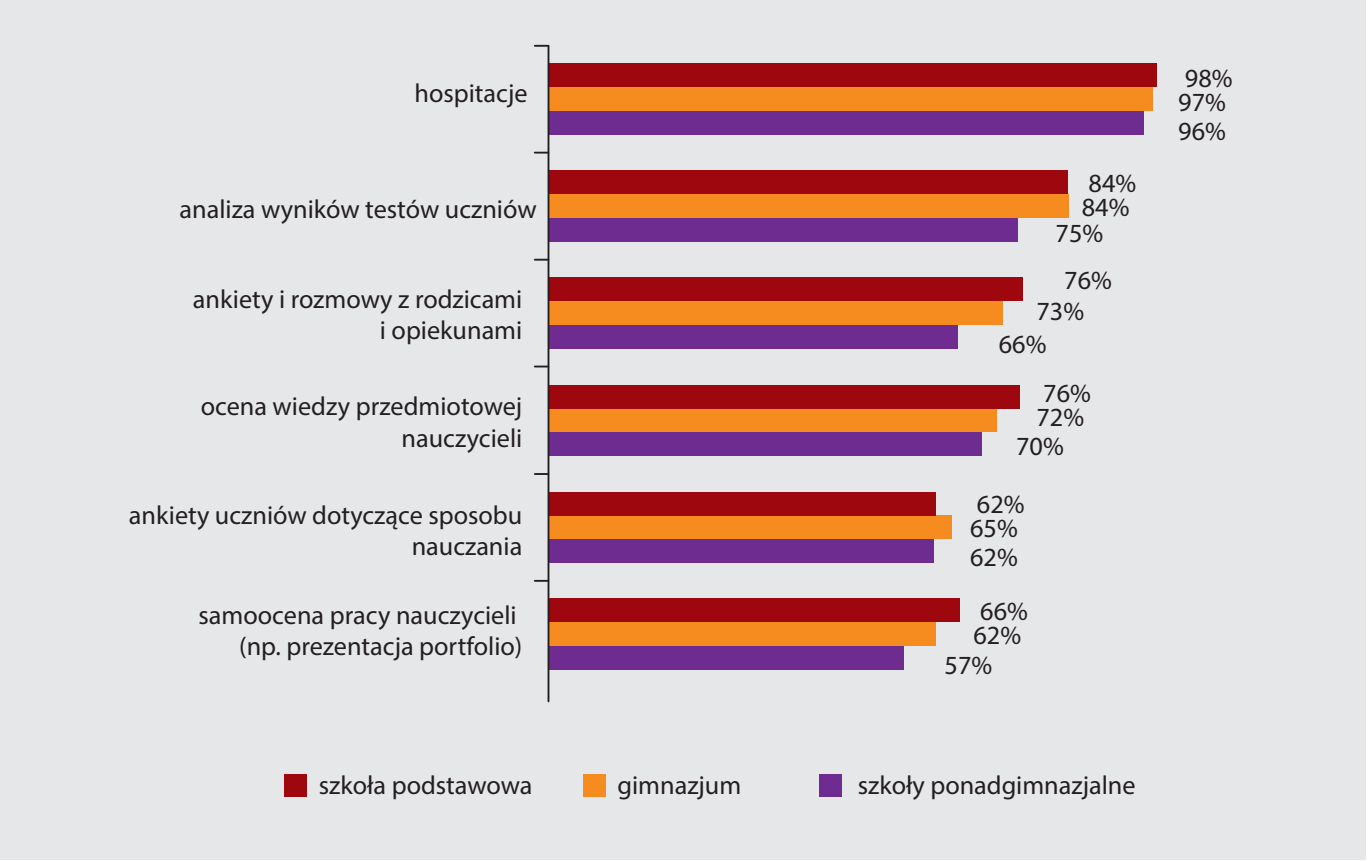
Rysunek 5.8. Odsetek nauczycieli polskich szkół deklarujących, że w ramach otrzymywanej informacji zwrotnej poszczególne obszary uwzględnione zostały jako bardzo ważne, bądź umiarkowanie ważne



Źródło: Dane OECD, TALIS 2013.

W badaniu TALIS analizowano także różne metody przekazywania informacji zwrotnej, a właściwie swego rodzaju źródła stanowiące podstawę tej informacji. Analiza wyników testów uczniów była drugą – po informacji uzyskanej w efekcie hospitacji – najczęściej wskazywaną „metodą" przekazywania informacji zwrotnej spośród sześciu uwzględnionych w ramach badania. W zależności od poziomu edukacyjnego wskazywało na nią, zależnie od rodzaju szkoły, 75–84% nauczycieli. Nie wiadomo jednak, co kryje się pod sformułowaniem „analiza wyników testów".

Rysunek 5.9. Odsetek nauczycieli polskich szkół wskazujących otrzymywanie informacji zwrotnej z zastosowaniem poszczególnych stosowanych w szkole metod (od wszystkich podmiotów łącznie)



Źródło: Dane OECD, TALIS 2013.

Dodajmy, że pod względem znaczenia wyników testów polscy nauczyciele gimnazjów plasują się zdecydowanie powyżej średniej krajów biorących udział w badaniu TALIS – aż 84% deklaruje uzyskiwanie informacji zwrotnej po analizach wyników egzaminacyjnych, przy czym najczęściej taką informację przekazują sami dyrektorzy (56%). Średnia dla państw, które brały udział w badaniu jest w obu przypadkach zdecydowanie niższa (odpowiednio 64% i 24%) (OECD, 2014).

5.3. Praktyka wykorzystania danych egzaminacyjnych w szkołach

Z przedstawionych we wcześniejszych częściach rozdziału danych wynikałoby, że w szkołach powszechnie przeprowadza się analizy wyników egzaminów zewnętrznych i wykorzystuje ich wyniki w bardzo zróżnicowany sposób. Przytoczone badania międzynarodowe dodatkowo pokazują, że Polska pod tym względem znajduje się zazwyczaj w czołówce – wśród państw, w których szczególnie często wskazuje się na uwzględnianie wyników uczniów, w tym wyników egzaminacyjnych w pracy szkoły. Możemy stwierdzić, że wyniki egzaminacyjne w przypadku kadry szkół postrzegane są jako coś ważnego, co powinno być uwzględniane w pracy szkoły. Tak wysoki poziom deklaracji wiąże się między innymi z dużą świadomością badanych dotyczącego obowiązku przeprowadzania analiz wyników egzaminacyjnych, wynikającego z regulacji prawnych. Nie bez znaczenia wydają się również –omawiane w pierwszym rozdziale raportu – oczekiwania względem szkół ze strony instytucji systemu oświaty, ale też presja otoczenia – zainteresowanie mediów, rodziców wynikami egzaminacyjnymi.

Jednocześnie trzeba zauważyć, że pojęcia stosowane w badaniach ankietowych na temat wykorzystania wyników egzaminacyjnych mogą być różnie rozumiane przez dyrektorów szkół i nauczycieli.

Kiedy kończy się refleksja nad wykorzystaniem wyników egzaminacyjnych, a zaczyna ich wykorzystanie? Czy zapoznanie się z wynikami EWD przedstawionymi na elipsie bądź średnimi wynikami z egzaminu to już analiza wyników? W świetle przedstawionych badań nie mamy więc pełnego obrazu dotyczącego tego, jak w praktyce wyglądają interpretacje wyników analiz oraz jak faktycznie wnioski z tych analiz wykorzystywane są do podejmowania decyzji.

Kwestia rozbieżności pomiędzy deklaracjami dotyczącymi wykorzystania analiz wyników egzaminacyjnych przez kadrę szkół a faktycznym ich zastosowaniem identyfikowana była m.in. przez autorów badań dotyczących korzystania z danych przez szkoły, prowadzonych w 5 krajach Unii Europejskiej (oprócz Polski, także w Holandii, na Litwie, w Niemczech i Wielkiej Brytanii – „Data use project"). Badani we wszystkich krajach deklarowali wykorzystanie danych m.in. w celu kształtowania programu szkoły, planowania zmian, doskonalenia pracy nauczycieli. Wykorzystanie to było jednak bardzo powierzchowne. Tylko w Wielkiej Brytanii nauczyciele byli w stanie wskazać konkretne przykłady wykorzystania danych (Schildkamp, Karbautzki i Vanhoof, 2014).

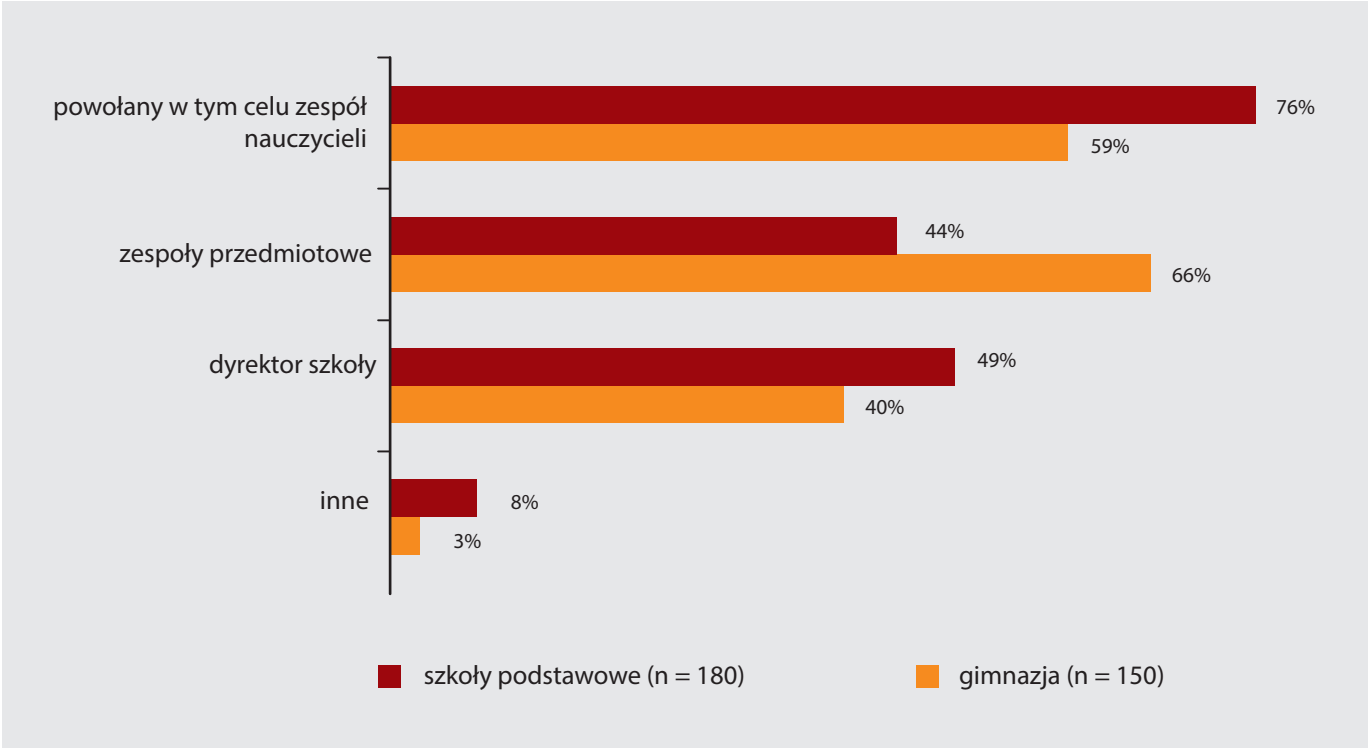
Co zatem w praktyce szkoły robią z wynikami egzaminów zewnętrznych? Czy jedynie je gromadzą, czy również analizują? A jeśli tak, to co analizują – średnie wyniki, wskaźniki EWD, dodatkowe dane kontekstowe mówiące o uwarunkowaniach szkolnych i pozaszkolnych? Na jakie pytania dyrektorzy i nauczyciele poszukują odpowiedzi w swoich analizach? Czy są to pytania o efekty, czy również o efektywność nauczania? Czy formułują hipotezy odnośnie do możliwych przyczyn zidentyfikowanych problemów? Wreszcie, czy rezultatem analiz są propozycje usprawnień? Jeśli tak, to jakiego typu są to działania? Na powyższe pytania danych dostarczają badania jakościowe przeprowadzone w IBE w latach 2013–2014 w 58 gimnazjach, liceach i technikach, obserwacje uczestniczące na warsztatach rad pedagogicznych na temat EWD oraz przegląd raportów z analiz wyników egzaminacyjnych zebranych w 2014 roku w ok. 50 szkołach podstawowych. Do analiz wykorzystano także materiał badawczy zebrany przez wizytatorów w ramach ewaluacji zewnętrznych, jak również wybrane badania ilościowe prowadzone przez IBE.

5.3.1. Organizacja analiz wyników egzaminacyjnych

Kto analizuje wyniki egzaminacyjne?

Dyrektorzy i nauczyciele w różny sposób organizują w szkołach proces przeprowadzania analiz wyników egzaminacyjnych. Prowadzone są one przede wszystkim zespołowo – w ramach specjalnie powołanych zespołów zadaniowych lub w ramach zespołów przedmiotowych. Rzadziej analizy prowadzone są bezpośrednio przez samego dyrektora.

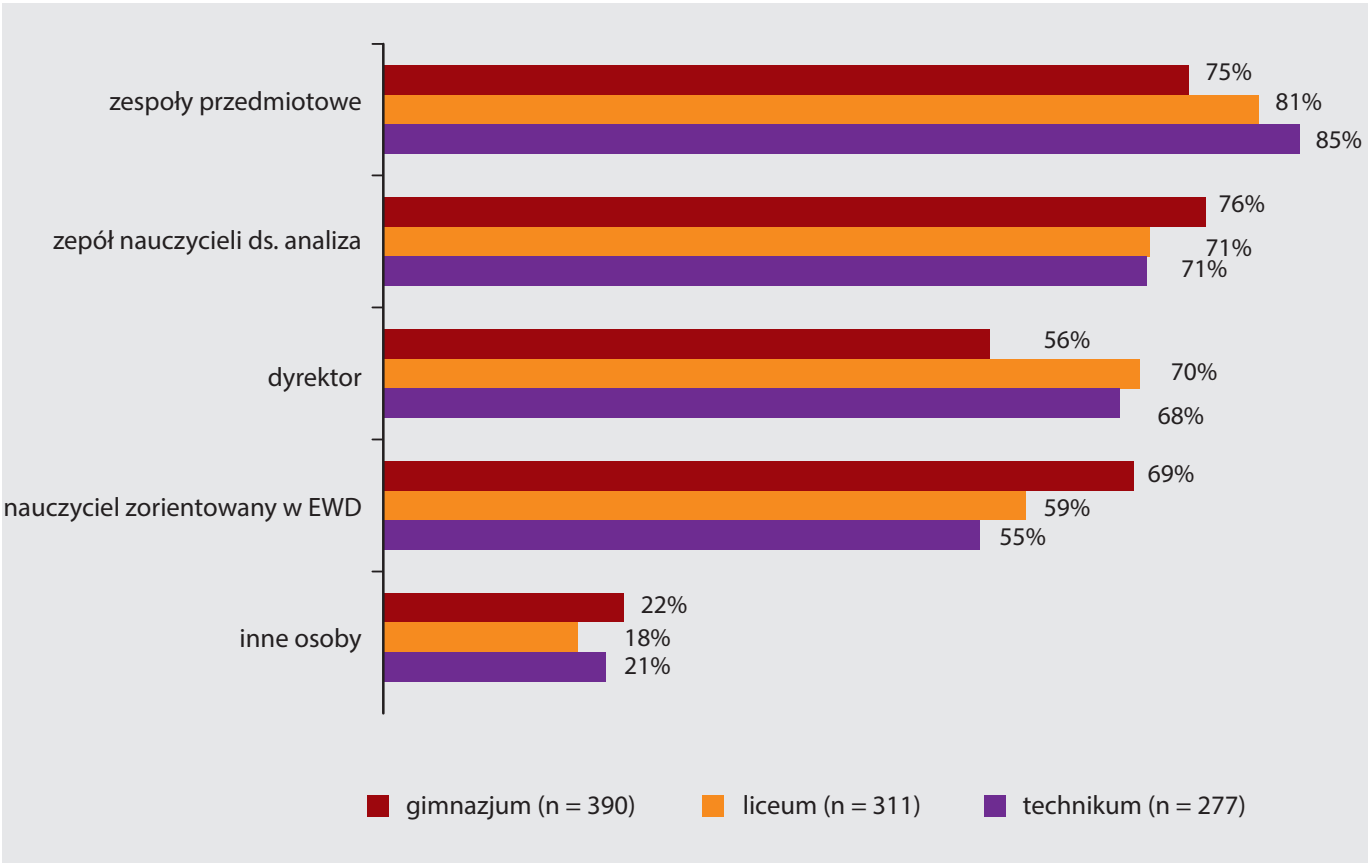
Rysunek 5.10. Deklaracje dyrektorów dotyczące tego, kto analizuje wyniki egzaminacyjne, wg typu szkoły (2012 r.)



Źródło: na podstawie: Badania podłużne w szkołach podstawowych i gimnazjach (III etap – 2012); możliwość kilku odpowiedzi.

W podobny sposób organizowane są analizy wskaźników EWD. Z badania przeprowadzonego w 2013 roku wynika, że w liceach, technikach i gimnazjach analizy najczęściej przeprowadzane są w ramach poszczególnych zespołów przedmiotowych (75–85%) lub specjalnie w tym celu powołanych zespołów ds. analiz (71–76%), w dalszej kolejności jest to zadanie dyrektora (56–70%) (por. rysunek 5.11). Różnica w tym przypadku polega na dosyć częstym prowadzeniu analiz przez 1 nauczyciela – osobę szczególnie dobrze zorientowaną w zagadnieniach EWD. Co ciekawe, w zespołach ds. ewaluacji analizy wskaźników EWD przeprowadzane są niezwykle rzadko (5%, w ramach kategorii inne). Zjawisko to może być kolejnym potwierdzeniem tego, że w świadomości dyrektorów szkół i praktyce szkolnej analizy przeprowadzane w ramach ewaluacji wewnętrznej wydają się czymś odrębnym od analiz wyników egzaminacyjnych.

Rysunek 5.11. Deklaracje dyrektorów odnośnie tego, kto analizuje wskaźniki EWD, wg typu szkoły (2013 r.)



Źródło: badanie CATI przeprowadzone przez IBE wśród dyrektorów w 2013 r., podstawa procentowania: dyrektorzy szkół, w których wykorzystuje się wskaźniki EWD do oceny efektywności nauczania.

W ramach wymagań państwa wobec szkół, obowiązujących do roku szkolnego 2012/2013, jednym z obszarów podlegających ewaluacji zewnętrznej był sposób prowadzenia analizy wyników egzaminacyjnych. Z zebranego przez wizytatorów materiału wynika, że wstępne analizy wyników prowadzone są najczęściej w ramach zespołów zadaniowych lub przedmiotowych. W dużej części szkół wyniki analiz przedstawiane są i omawiane na spotkaniu rady pedagogicznej, w trakcie której formułuje się i zatwierdza wnioski (Ligęza, 2013; Ligęza i Franczak, BDW; Milecka, 2014a). W części szkół, niezależnie od analizy zespołowej, każdy z nauczycieli prowadzi własne analizy wyników, rzadko analizy prowadzone są tylko przez jedną osobę. Zespołowe analizy danych są bardziej powszechne w szkołach podstawowych i gimnazjach, w których to sprawdzian i egzamin mają charakter ponadprzedmiotowy. W przypadku szkół ponadgimnazjalnych – zwłaszcza liceów ogólnokształcących – odwrotnie – większy jest udział analiz prowadzonych indywidualnie przez nauczycieli (Milecka, 2014a).

Jakie analizy są przeprowadzane w szkołach

W trakcie ewaluacji zewnętrznych we wspomnianym okresie przyglądano się bardziej szczegółowo stosowanym metodom analiz wyników egzaminacyjnych w podziale na analizy ilościowe i jakościowe, przy czym te pierwsze zdecydowanie dominowały wśród analiz szkolnych. Z danych zebranych przez wizytatorów wynika, że w ramach analiz ilościowych najpowszechniej stosowane w szkołach metody to (Ligęza, 2013; Ligęza i Franczak BDW; Milecka, 2014a):

- przyglądanie się średniej szkoły, trochę rzadziej średnim poszczególnych oddziałów klasowych oraz poszczególnych uczniów,
- przyglądane się położeniu średniej szkoły na skali staninowej;

- przyglądanie się rozkładowi wyników uczniów przede wszystkim na skali staninowej, też określanie liczebności uczniów w określonych staninach/centylach, określanie liczebności uczniów z najniższymi najwyższymi wynikami;
- analizowanie łatwości, przede wszystkim w odniesieniu do zadania, w dalszej kolejności konkretnej umiejętności i standardu;
- w przypadku szkół ponadgimnazjalnych dodatkowo: analiza zdawalności egzaminu, a także liczby osób przystępujących do egzaminu z różnych przedmiotów, na poszczególnych poziomach.

Wiele szkół analizuje wyniki egzaminacyjne w celu porównania się z innymi placówkami. Odnoszą się zatem do wyników gminy, powiatu, województwa bądź kraju. Zdecydowanie rzadziej natomiast porównują się do siebie samych, analizując swoje własne wyniki na przestrzeni lat (Ligęza, 2013; Ligęza i Franczak BDW; Milecka, 2014a). Jest to stała tendencja wskazywana we wszystkich analizach danych z nadzoru pedagogicznego w tym okresie.

Określenie, które sposoby analiz wyników egzaminacyjnych można uznać za jakościowe, a które nie, przysparzało trudności zarówno wizytatorom, jak i dyrektorom oraz nauczycielom. W danych zgromadzonych przez wizytatorów (Ligęza, 2013; Ligęza i Franczak BDW; Milecka, 2014a), wśród wymienianych analiz jakościowych szkół, najczęściej pojawiały się zestawienia łatwości zadania z jego typem, identyfikowanie zadań i umiejętności sprawiających największe problemy, a także analiza kontekstowa. Wizytatorzy wskazywali także, że w szkołach przeprowadzane są porównania wyników egzaminacyjnych z klasyfikacją końcoworoczną, ocenami z poszczególnych przedmiotów i wynikami próbnych egzaminów. Warto podkreślić, że wśród branych przez szkoły pod uwagę czynników kontekstowych najczęściej nacisk kładzie się na czynniki uczniowskie (np. opinie poradni pedagogicznych, frekwencję, różnego rodzaju dysfunkcje), w dalszej kolejności na czynniki środowiskowe (sytuacja rodzinna, miejsce zamieszkania, wykształcenie i status rodziców i inne). Bardzo rzadko natomiast pojawiają się uwarunkowania związane z pracą nauczycieli, czy organizacją pracy szkoły (stosowane metody nauczania, realizowane programy, sposoby organizacji zajęć lekcyjnych i pozalekcyjnych, absencja nauczycieli i inne).

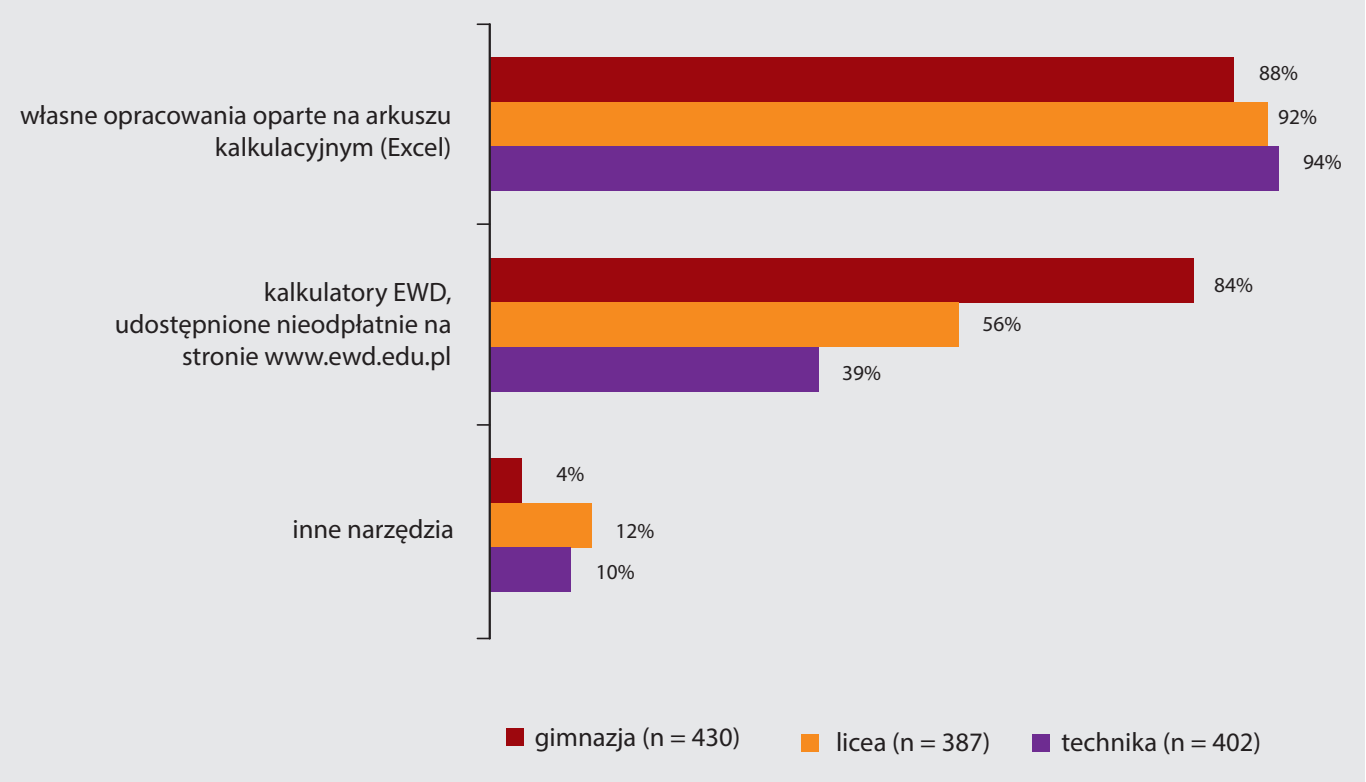
Dane zebrane przez wizytatorów pokazują także, że analiza wyników egzaminacyjnych nie jest dla szkół zadaniem łatwym, identyfikowane są różne trudności i błędy. Wydaje się zatem, że nadal w pewnym stopniu aktualna jest diagnoza z początków funkcjonowania systemu nadzoru: „W gimnazjum analizowane są wyniki egzaminów zewnętrznych. Powyższe zdanie można przeczytać praktycznie we wszystkich raportach ewaluacyjnych, bo wszystkie szkoły analizy wykonują. Jednak z lektury raportów ewaluacyjnych wyłania się smutny obraz tego, w jak nieporadny, powierzchowny sposób szkoły pracują z danymi egzaminacyjnymi. Wizytatorzy-ewaluatorzy poddają się tej nieporadności, powtarzają opinie dyrektorów i nauczycieli sformułowane w ankietach oraz wywiadach i nie próbują krytycznie dociekać, jaki faktyczny użytek z wyników egzaminacyjnych robią szkoły" (Stożek, 2010).

Jednym z typowych problemów zaobserwowanych podczas ewaluacji zewnetrznych jest porównywanie średniego wyniku szkoły pomiędzy latami (Ligęza, 2013; Ligęza i Franczak, BDW; Milecka, 2014a). Niemniej jednak z analizy danych nadzoru pedagogicznego wynika również, że zauważalny jest wzrost popularności miar względnych (skali staninowej w przypadku szkół podstawowych i gimnazjów, rozkładów centylowych w przypadku szkół ponadgimnazjalnych) (Milecka, 2014a). Materiały te potwierdzają również, wskazywany we wcześniejszych częściach raportu, obserwowany wzrost wykorzystania przez szkoły wskaźników EWD. Jednocześnie podkreślane jest, że w przypadku szkół deklarujących wykorzystanie EWD raczej dominuje analiza wskaźników rocznych bądź trzyletnich, rzadziej stosowane są pogłębione analizy z użyciem Kalkulatora EWD, zaś na przykład bardziej zaawansowane analizy dla grup uczniów wyodrębnionych ze względu na różne cechy prowadzone są w jednostkowych przypadkach (por. Milecka, 2014a).

Stosowanie przez kadrę szkoły Kalkulatora EWD do analiz wewnątrzszkolnych może być jednym z wskaźników zaawansowania placówki w pogłębionych analizach. Wyniki badań ilościowych IBE

prowadzonych w 2013 roku wśród dyrektorów szkół wskazują, że analizy wyników egzaminów zewnętrznych najczęściej mają postać własnych opracowań przygotowywanych z wykorzystaniem arkusza kalkulacyjnego (88–94%) (por. rysunek 5.12). Kalkulatory EWD wykorzystywane są w tym celu najczęściej w gimnazjach (84%). Szkoły ponadgimnazjalne deklarowały stosowanie Kalkulatora nieco rzadziej, co wynika w dużym stopniu z późniejszego terminu udostępnienia im modułu maturalnego Kalkulatora (od lutego 2013). W technikach 39% dyrektorów deklarowało wykorzystywanie Kalkulatorów EWD, a w liceach 56%.

Rysunek 5.12. Deklaracje dyrektorów dotyczące stosowania narzędzi do analizy wyników egzaminów zewnętrznych, wg typu szkoły



Źródło: Na podstawie badania CATI przeprowadzonego przez IBE wśród dyrektorów w 2013 r., podstawa procentowania: dyrektorzy, z wyjątkiem tych, którzy wskazali, iż nie orientują się, czym jest edukacyjna wartość dodana.

Z kolei w szkołach podstawowych pod koniec 2014 roku jedynie 16% dyrektorów przeprowadziło analizy EWD z użyciem Kalkulatora. Należy pamiętać jednak, że szkoły podstawowe dysponują Kalkulatorem dopiero od sierpnia 2014 roku, a więc użytkownicy, którzy dysponowali danymi do Kalkulatora (tj. wynikami OBUT 2011 i sprawdzianu 2014), nie mieli jeszcze zbyt wiele czasu, aby poznać to narzędzie.

5.3.2. Formułowanie i wdrażanie rekomendacji

Zdefiniowanie, zaplanowanie i wdrożenie działań mających na celu poprawę wyników egzaminacyjnych oraz efektywności nauczania szkoły nie jest zadaniem łatwym. Porównania międzynarodowe pokazują, że jednym z głównych problemów i wyzwań w wielu krajach jest przełożenie wyników analiz na praktykę w klasie (OECD, 2013). Po pierwsze, trudności sprawia szkołom zidentyfikowanie możliwych przyczyn, z których może wynikać niska efektywność nauczania – hipotez, które mogłyby być sprawdzane z wykorzystaniem posiadanych przez szkołę danych. Trudna jest też ich weryfikacja. Problem sprawia także etap formułowania konkretnych rozwiązań. Raporty z analiz wyników

egzaminacyjnych szkół podstawowych ograniczają się zazwyczaj się do prezentacji samych wykresów i zestawień z analiz wyników egzaminacyjnych lub EWD. W większości przypadków opracowania te nie zawierają propozycji usprawnień lub opisują je w sposób ogólnikowy (np. „prowadzenie zajęć wyrównawczych"). Pytani o to dyrektorzy tłumaczą, że raporty stanowią jedynie punkt wyjścia do dyskusji z nauczycielami, w których wypracowywane są propozycje zmian. Jednak rozmowy z dyrektorami i nauczycielami szkół gimnazjalnych, liceów i techników (Kędracka i in., 2013) pokazują, że powiązanie analizy danego problemu z planowanymi usprawnieniami pracy szkół jest dla wielu z nich problemem. Innym zaobserwowanym zjawiskiem jest wiązanie już wcześniej zapoczątkowanych działań z analizami wyników egzaminacyjnych, co wynika z chęci udokumentowania, że szkoła spełnia oczekiwania systemu i wykorzystuje w praktyce wyniki analiz.

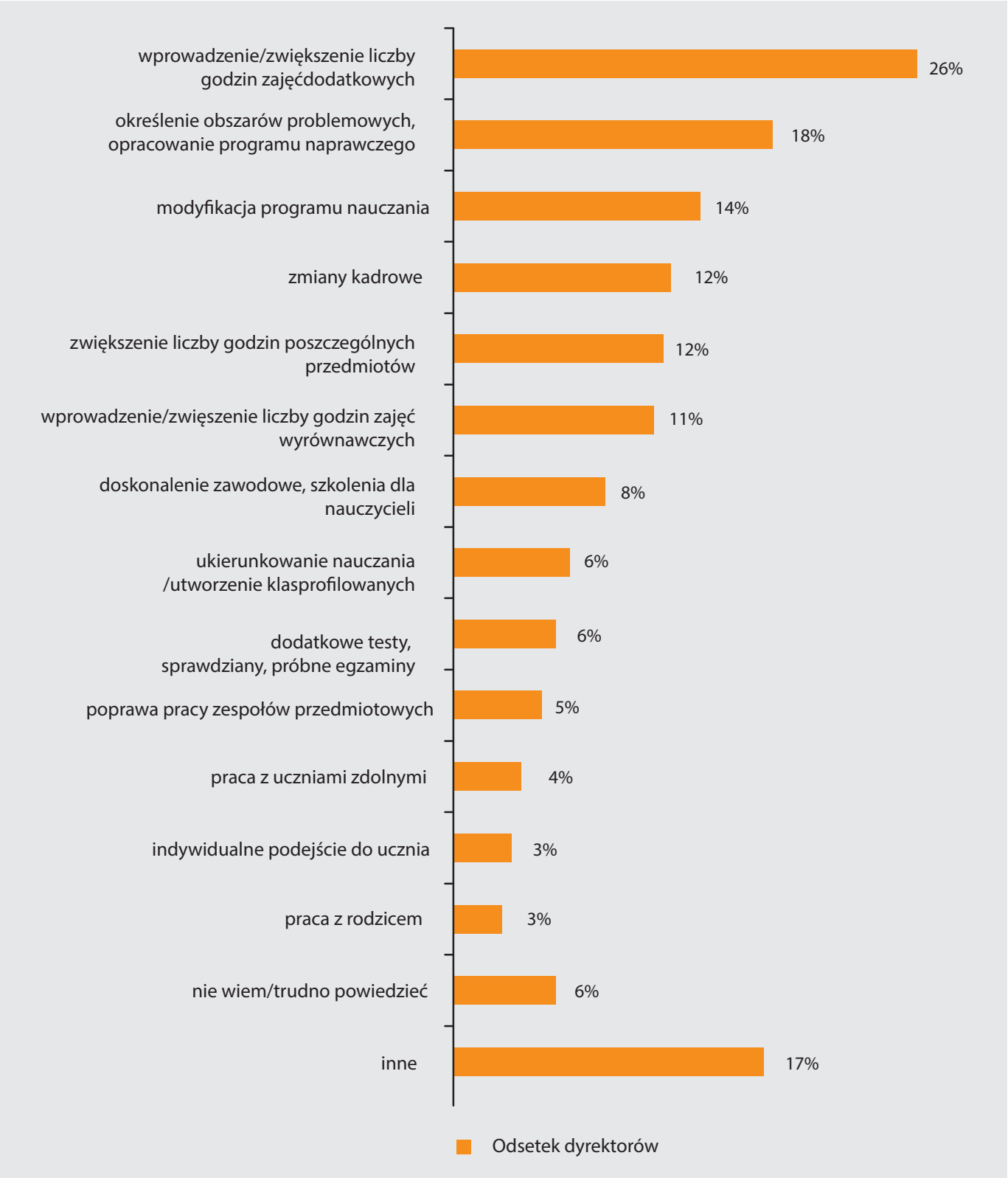
W przygotowywanych w szkołach opracowaniach wyników analiz spotkać też można rekomendacje nieadekwatne do postawionego problemu dydaktycznego, wychowawczego czy organizacyjnego. Zdarzają się też dokumenty, w których analizy są adekwatne, ale rekomendacje nierealistyczne, np. zbyt szybko oczekuje się mierzalnych efektów lub stawia się zbyt ambitne cele, nie dostrzegając znaczenia drobnych modyfikacji w praktykach dydaktyczno-wychowawczych.

O zakresie i sposobach wykorzystania wniosków z analiz danych egzaminacyjnych wiele się można dowiedzieć również z omawianych wcześniej materiałów zebranych przez wizytatorów (Ligęza, 2013; Ligęza i Franczak, BDW; Milecka, 2014a, 2014 b). Wynika z nich, że wśród działań podejmowanych w szkołach po analizach wyników egzaminacyjnych dominuje zwiększenie nacisku na te umiejętności, które w analizach uznano za słabiej opanowane, bądź też zwiększanie czasu przeznaczanego na nauczanie: organizacja dodatkowych zajęć, na których są one kształtowane, zwiększenie liczby godzin z przedmiotów egzaminacyjnych, zmiany w ofercie zajęć pozalekcyjnych/fakultatywnych, w ramach których nacisk kładzie się na zajęcia przygotowujące do egzaminu. Jak pokazano w rozdziale 1, tego rodzaju praktyki mogą przynosić zarówno pozytywne skutki, jak i negatywne, związane przede wszystkim z odwracaniem uwagi od tych ważnych umiejętności, które nie są mierzone na egzaminie. Inne często deklarowane działania wiążą się bardziej z kształtowaniem umiejętności służących poprawie wyniku egzaminacyjnego niż kształtowaniu umiejętności zapisanych w podstawie, np. doskonalenie umiejętności pracy z arkuszem egzaminacyjnym czy ćwiczenie typowych zadań egzaminacyjnych. Mniej powszechne, choć wskazywane w części szkół (zwłaszcza w ostatnich latach), są działania nastawione na zmiany metod pracy z uczniem, zwiększanie motywacji uczniów czy poprawę jakości współpracy z rodzicami (np. większa indywidualizacja nauczania, metody aktywizujące, realizacja projektów zewnętrznych). Kładziono także większy nacisk na monitorowanie opanowania standardów egzaminacyjnych, zwiększenie liczby testów kompetencji i egzaminów próbnych. Co ciekawe, szkoły we wnioskach z analiz danych egzaminacyjnych bardzo rzadko wskazują na działania związane bezpośrednio z doskonaleniem nauczycieli (Ligęza, 2013; Ligęza i Franczak, BDW; Milecka, 2014a, 2014 b).

Dyrektorów gimnazjów, liceów i techników pytano także o to, w jakiej mierze przeprowadzone w szkole analizy EWD pomogły w podjęciu decyzji, które przyczyniły się do poprawy efektywności nauczania w szkole (Matuszczak i in. 2014). Większość dyrektorów przyznaje, że analizy EWD pomogły w dużym (59–69%) lub bardzo dużym stopniu (9–10%). Pozostali w większości ocenili analizy EWD jako pomocne w niewielkim stopniu (18–26%).

Najczęściej deklarowano właśnie wprowadzenie lub zwiększenie liczby godzin określonych zajęć: dodatkowych lub wyrównawczych, czy zajęć z poszczególnych przedmiotów. Dosyć często wskazywano na opracowanie programów naprawczych i modyfikacje programu nauczania.

Rysunek 5.13. Decyzje przy których pomocne było wykorzystanie wskaźników EWD



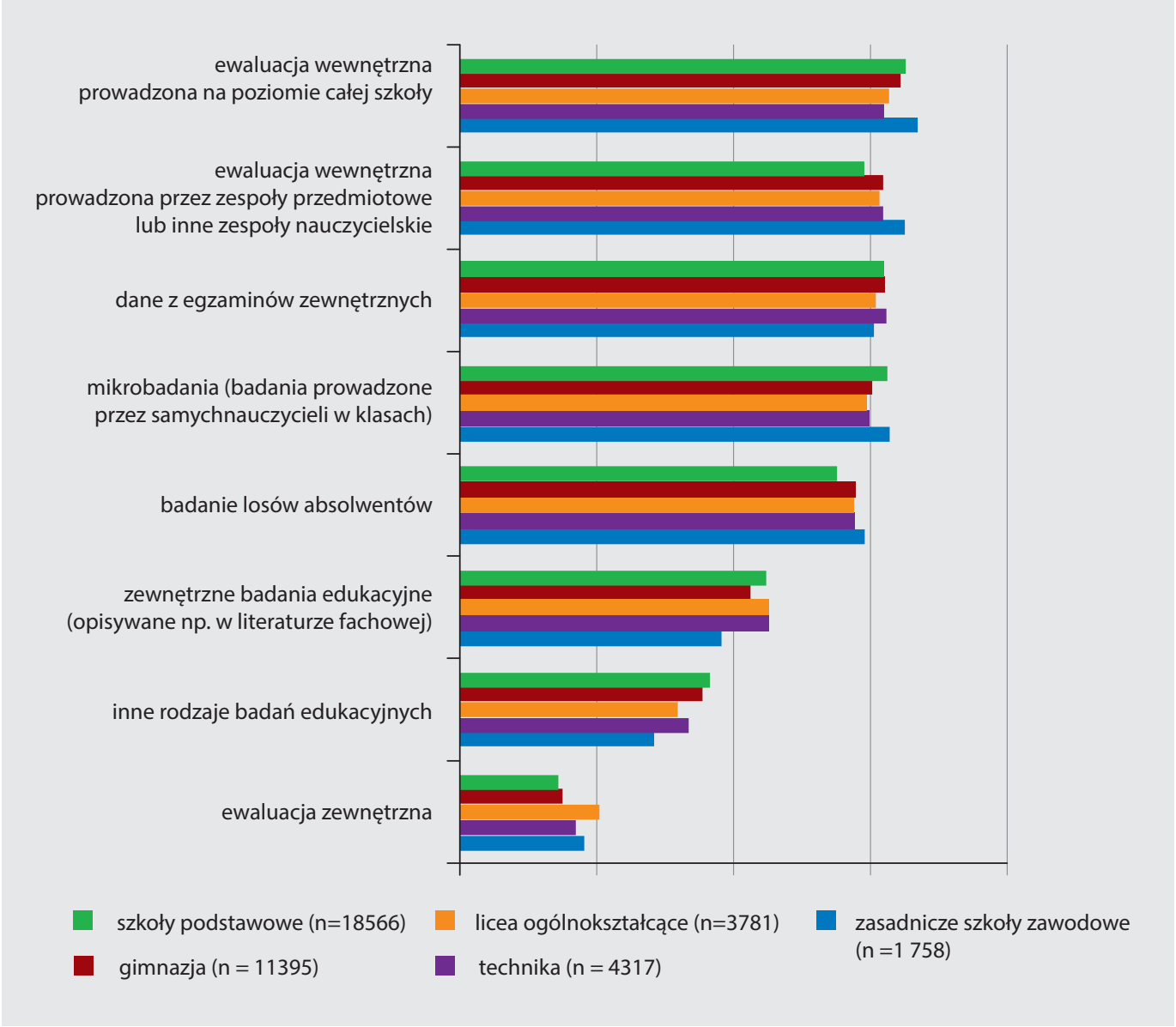
Źródło: Na podstawie badania CATI przeprowadzonego przez IBE wśród dyrektorów w 2013 r., podstawa procentowania: dyrektorzy, dla których EWD pomogło w podjęciu decyzji w dużym lub bardzo dużym stopniu (N = 776).

Z danych nadzoru pedagogicznego wynika, że o ile analizy często prowadzi się w zespołach, to już wdrożenie wniosków zazwyczaj odbywa się na poziomie działań planowanych i podejmowanych przez poszczególnych nauczycieli. Wielu badanych stwierdza, że indywidualnie planują pracę

z uczniem, nieliczni wskazują, że w szkołach w konsekwencji analiz wyników egzaminacyjnych realizowane są wspólne działania (Ligęza, 2013). Należy przy tym dodać, że w przypadku analiz dotyczących wymagania obowiązującego od roku szkolnego 2013/2014 zauważyć można rozdźwięk pomiędzy wypowiedziami dyrektorów szkół i nauczycieli. Dyrektorzy deklarują, że uwzględniają wnioski z prowadzonych analiz w organizacji procesów edukacyjnych i wykorzystują je do podnoszenia jakości pracy szkoły. Wskazują średnio co najmniej kilka zmian wprowadzonych na poziomie szkolnym (Milecka, 2014b).

Z kolei z zebranych ankiet wynika, że odsetek nauczycieli deklarujących bezpośrednie wykorzystanie analiz wyników egzaminacyjnych w swojej pracy nie przekracza kilkunastu procent. Na podobnym poziomie kształtuje się odsetek nauczycieli deklarujących korzystanie z wyników ewaluacji zewnętrznej i wewnętrznej czy twierdzących, że prowadzą własne badania w klasie (tzw. mikrobadania). Co ciekawe, pomimo zróżnicowanego zakresu informacji dostępnej dla nauczycieli poszczególnych rodzajów szkół odsetek ten nie różni się między nauczycielami uczącymi w szkołach podstawowych, gimnazjach i szkołach ponadgimnazjalnych.

Rysunek 5.14. Wykorzystanie przez nauczycieli poszczególnych rodzajów badań w pracy, wg typu szkoły



Źródło: Opracowanie własne na podstawie danych z platformy SEO za okres 1.09.2013–31.12.2014 [pobrane 18 lutego 2015]. N = 39817

Problem indywidualnego wdrażania wniosków z analiz i niespójności pomiędzy wypowiedziami dyrektorów szkół i nauczycieli dotyka szerszej kwestii, którą jest przełożenie wyników analiz na nauczanie na poziomie klasy. Szkoły mają trudności z powiązaniem analiz efektów/efektywności kształcenia z własną pracą. Stanowi to problem zarówno na etapie diagnozowania przyczyn niepowodzeń/sukcesów szkoły, wyznaczania kierunków dalszego rozwoju, jak i podejmowania konkretnych działań. Na przykład, identyfikując problemy, szkoły głównie wskazują na przyczyny zewnętrzne, niezależne od szkoły (Kędracka i in., 2013). Wiąże się to bardziej z ogólnym wyzwaniem strategicznego, planowego podejścia do określania celów i sposobów wprowadzania zmian w szkole. Stąd wprowadzane działania mają często charakter punktowy, dominują znane szkołom schematy i metody postępowania (Wasilewska i in., 2014). Kluczowe znaczenie ma tu rola dyrektora, o czym szerzej piszemy w części 4 rozdziału.

5.3.3. Zróżnicowane zaawansowanie użytkowników analiz

Badania IBE pokazują, że większość szkół dopiero uczy się, jak i w jakim celu analizować, interpretować i wykorzystywać dane z systemu egzaminacyjnego, tak aby przeprowadzone analizy rzeczywiście okazały się przydatne w rozwoju ich placówek (Stożek, Kędracka i Rappe, 2015). Szkoły gromadzą różne dane, w tym wyniki egzaminacyjne, lecz ich nie przetwarzają w świadomy i celowy sposób, tak aby odpowiedzieć sobie na postawione przez siebie pytania (związane z np. dydaktyką). Raczej nie sięgają również po wskaźniki edukacyjnej wartości dodanej.

Główną motywacją do gromadzenia danych w takich szkołach jest świadomość zobowiązań regulacji prawnych (motywacja zewnętrzna). Umiejętności analityczne osób zajmujących się analizami są niskie. W tego typu szkołach częściej spotkać można negatywne postawy względem samego systemu egzaminów zewnętrznych, w tym jakości egzaminów.

W nieco bardziej zaawansowanych szkołach dane analizowane są w bardziej świadomy sposób, ale wciąż głównie w celach zaspokojenia potrzeb informacyjnych podmiotów nadzorujących (organów prowadzących lub kuratoriów oświaty). Analizy ograniczają się zwykle do zdawalności egzaminów, średnich wyników, trudności poszczególnych zadań. Rzadko występują pogłębione, bardziej jakościowe analizy, które mogłyby wskazywać na próby poszukiwania problemów. W analizach nie uwzględnia się dodatkowych danych kontekstowych, które pozwoliłyby bardziej zrozumieć uwarunkowania osiągnięć poszczególnych uczniów i oddziałów klasowych. Szkoły szukają raczej odpowiedzi na pytanie o skuteczność, a nie efektywność prowadzonych działań dydaktycznych (wykorzystanie potencjału uczniów), o której mówią np. wskaźniki EWD. W szkołach tego typu przywiązuje się również bardzo dużą wagę do dokumentowania wyników analiz, głównie w celu przedstawienia ich organom nadzorującym.

Szkoły jeszcze bardziej zaawansowane w pracy z danymi, w których pojawia się już głębsze myślenie ewaluacyjne o pracy szkoły, co prawda przeprowadzają już analizy, kierując się potrzebami szkoły (w raportach pojawiają się np. hipotezy badawcze, problemy do rozwiązania), lecz użytkownicy ci wciąż napotykają na problemy z interpretacją wyników analiz i przełożeniem wniosków z nich płynących na propozycje usprawnień dydaktycznych i organizacyjnych. W takich szkołach rodzą się już pytania: „czy my robimy dobrze? co dany wynik oznacza – że jest dobrze, czy raczej wciąż źle?" Ale widoczne są trudności z formułowaniem hipotez i ich weryfikacji.

W niektórych placówkach myślenie ewaluacyjne o szkole jest już sprzężone z codzienną pracą, a umiejętności analityczne osób pracujących z danymi raczej wysokie. Na podstawie wniosków z analiz formułowane są więc hipotezy dotyczące przyczyn danego stanu rzeczy oraz rozwiązania organizacyjno-dydaktyczne. Użytkownicy wskazują na jego przyczyny leżące po stronie szkoły, takie jak np. duża rotacja nauczycieli czy brak współpracy międzyprzedmiotowej. Pojawiają się też propozycje rozwiązania zidentyfikowanych problemów, jednak często nie są one wprowadzane w życie, a więc nie jest sprawdzane w szkole, czy dane podejście ma szansę rzeczywiście przyczynić się do

rozwiązania danych problemów. Widoczne jest także zbyt szerokie definiowanie problemów i celów, czy też proponowanie nieadekwatnych do postawionego problemu działań.
Są też placówki (często wśród gimnazjów, które najdłużej spośród wszystkich typów szkół mają dostęp do wskaźników EWD) o wysokiej kulturze pracy z danymi. W placówkach tych proces analizowania wyników egzaminacyjnych oraz wielu innych danych jest stałym elementem funkcjonowania szkoły. Analizy przeprowadzane są co roku, wokół analiz prowadzona jest dyskusja nauczycieli i dyrekcji, umacniana jest współpraca nauczycieli wymieniających się doświadczeniami dydaktycznymi, testowane są różne usprawnienia służące poprawie efektywności nauczania, a na tej podstawie wyciągane wnioski do kolejnych usprawnień – szkoła staje się organizacją uczącą się od siebie samej, na podstawie swoich własnych doświadczeń. Ale nawet w szkołach bardziej zaawansowanych w analizie i interpretacji wyników, zarówno dyrektorzy jak i nauczyciele mają często trudności z wskazaniem efektów podjętych po analizie wyników egzaminacyjnych działań. Związane jest to często z długim okresem czasu, jaki musi upłynąć aby rzeczywiście można było zaobserwować pozytywne oddziaływanie na uczniów.

5.4. Co sprzyja, wykorzystywaniu zaawansowanych analiz, a co je utrudnia?

Dlaczego zatem jest tak, że niektóre szkoły wyciągają wnioski z analiz wyników egzaminacyjnych i przekładają je na praktykę szkolną, a inne nie? Dlaczego niektóre placówki charakteryzują się wysoką kulturą pracy z danymi, a inne dużo niższą? Z badań międzynarodowych (Breiter i Karbautzki, 2011) wynika, że zależy to od wielu czynników. Znaczenie ma dostępność danych i ich jakość, ale też kompetencje analityczne nauczycieli i dyrektorów i ich przekonania o użyteczności danych. Zwraca się też uwagę na znaczenie współpracy nauczycieli w szkole, jakości przywództwa czy wizji pracy szkoły i obowiązujących w niej wartości. Na wykorzystanie danych ma też wpływ dostęp do oferty wspomagania pracy szkoły i jakość dostępnego doskonalenia. Potwierdzają to wyniki badań prowadzonych w Polsce.
Z badań podłużnych IBE wiemy, że podejście szkół podstawowych i gimnazjów do analizy/wykorzystania wyników egzaminacyjnych związane jest z motywacją zewnętrzną kadry szkół (oczekiwaniami środowiska), motywacją wewnętrzną (przekonaniami użytkowników o przydatności wyników egzaminacyjnych), kompetencjami osób zajmujących się analizami, a także z otwartością kadry szkoły na współpracę. Wnioski te znajdują potwierdzenie również w badaniach jakościowych dotyczących wykorzystania edukacyjnej wartości dodanej (Matuszczak i in. 2014; Stożek i in., 2015). Okazuje się zatem, że wykorzystanie metody EWD przez szkoły warunkują wiedza dyrektora i nauczycieli (znajomość metody i jej możliwości), umiejętności analityczne oraz postawy (np. stopień akceptacji, zaufanie do metody, otwartość na uczenie się).

5.4.1. Jakość i dostępność danych

Warunkiem korzystania przez szkoły z analiz wyników egzaminacyjnych jest dobra jakość i dostępność danych, którymi dysponują placówki. W drugim rozdziale opisaliśmy kiedy i w jakiej formie szkoły otrzymują co roku dane egzaminacyjne, a w kolejnych opisaliśmy sposób informowania o EWD i porównywalnych wynikach egzaminacyjnych. Warto jednak tutaj przypomnieć, że o ile surowe wyniki egzaminacyjne otrzymują od momentu stworzenia systemu wszystkie typy szkół, to wskaźniki EWD są udostępniane na stronach internetowych jedynie dla gimnazjów, liceów i techników. Natomiast szkoły podstawowe dopiero od 2014 roku mają możliwość wyliczania na swoje potrzeby edukacyjnej wartości dodanej, bazując na wynikach „badań diagnostycznych OBUT" (dane na wejściu) oraz „sprawdzianu szóstoklasisty" (dane na wyjściu). Wskaźniki EWD dla szkół podstawowych nie są ogólnodostępne w sieci – danymi służącymi do ich wyliczenia dysponuje wyłącznie

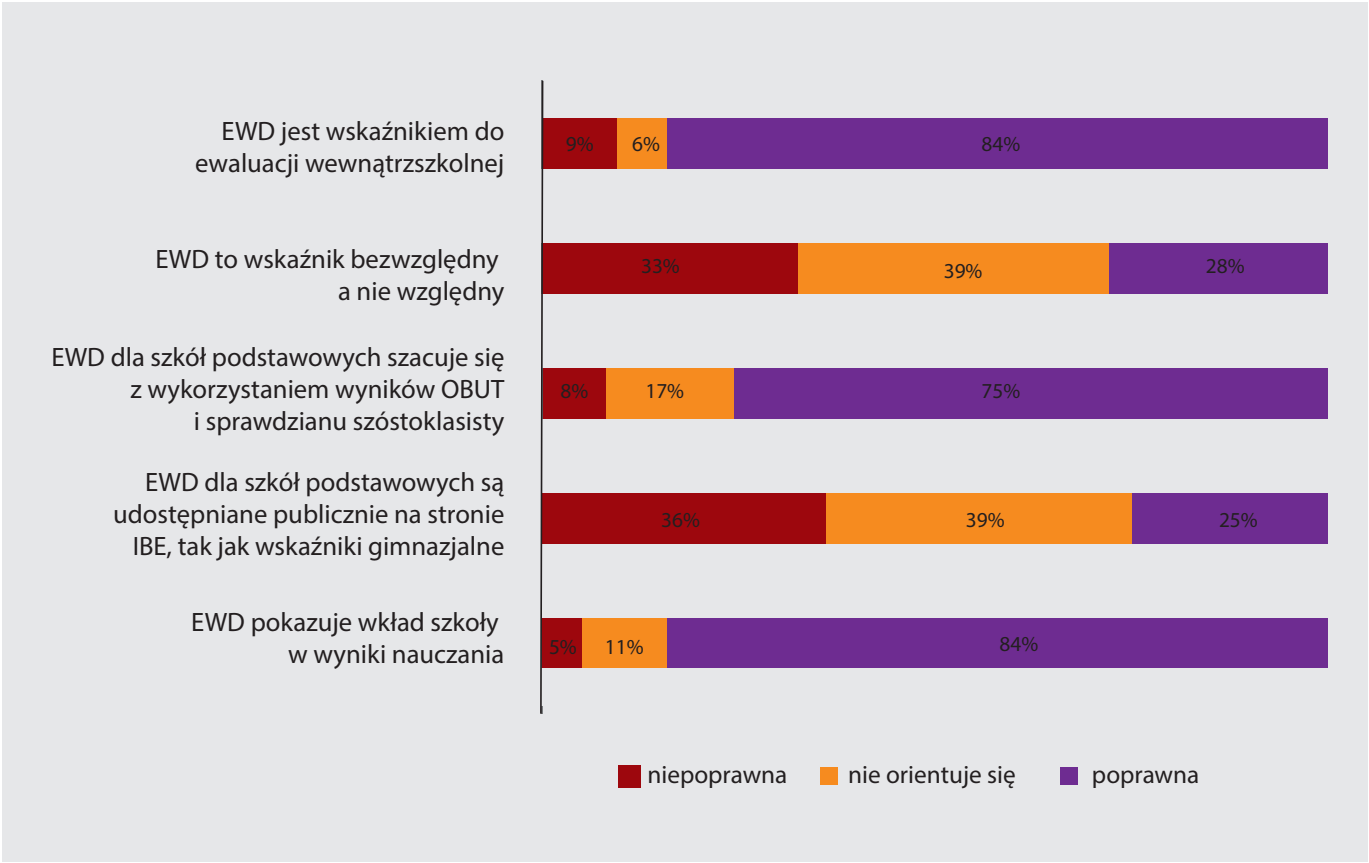
dyrektor szkoły. Zaawansowanie użytkowników w stosowaniu wskaźników EWD związane jest zatem między innymi z harmonogramem udostępniania wskaźników EWD dla różnych typów szkół.

5.4.2. Cechy użytkowników

Wiedza i kompetencje

O wiedzy i umiejętnościach analitycznych kadry szkoły możemy wnioskować jedynie na podstawie deklaracji dyrektorów szkół zebranych w ramach badań dotyczących edukacyjnej wartości dodanej. Z sondażu przeprowadzonego pod koniec 2014 roku wynika, że świadomość dyrektorów szkół podstawowych istnienia edukacyjnej wartości dodanej jest już bardzo wysoka (Fila, Matuszczak i Zielonka, 2015). Prawie wszyscy (99%) zadeklarowali, iż słyszeli o takich wskaźnikach, a tylko 3% dyrektorów, którzy uczestniczyli w edycji OBUT z 2010/2011 roku, nigdy nie słyszała o Kalkulatorze EWD (por. rysunek 5.15). Zdecydowana większość dyrektorów, którzy słyszeli wcześniej o EWD, ma świadomość, że jest to wskaźnik mogący służyć do ewaluacji wewnątrzszkolnej i że pokazuje wkład szkoły w wyniki nauczania. Dyrektorzy wiedzą również, z wykorzystaniem jakich danych wskaźniki są wyliczane (75%). Mają jednak trudność z określeniem, czy wskaźniki dla szkół podstawowych są ogólnodostępne w sieci – tylko 25% udzieliło poprawnej odpowiedzi.

Rysunek 5.15. Odpowiedzi dyrektorów SP na pytanie o wiedzę o wskaźnikach EWD, według udziału w edycji 2010/2011 OBUT

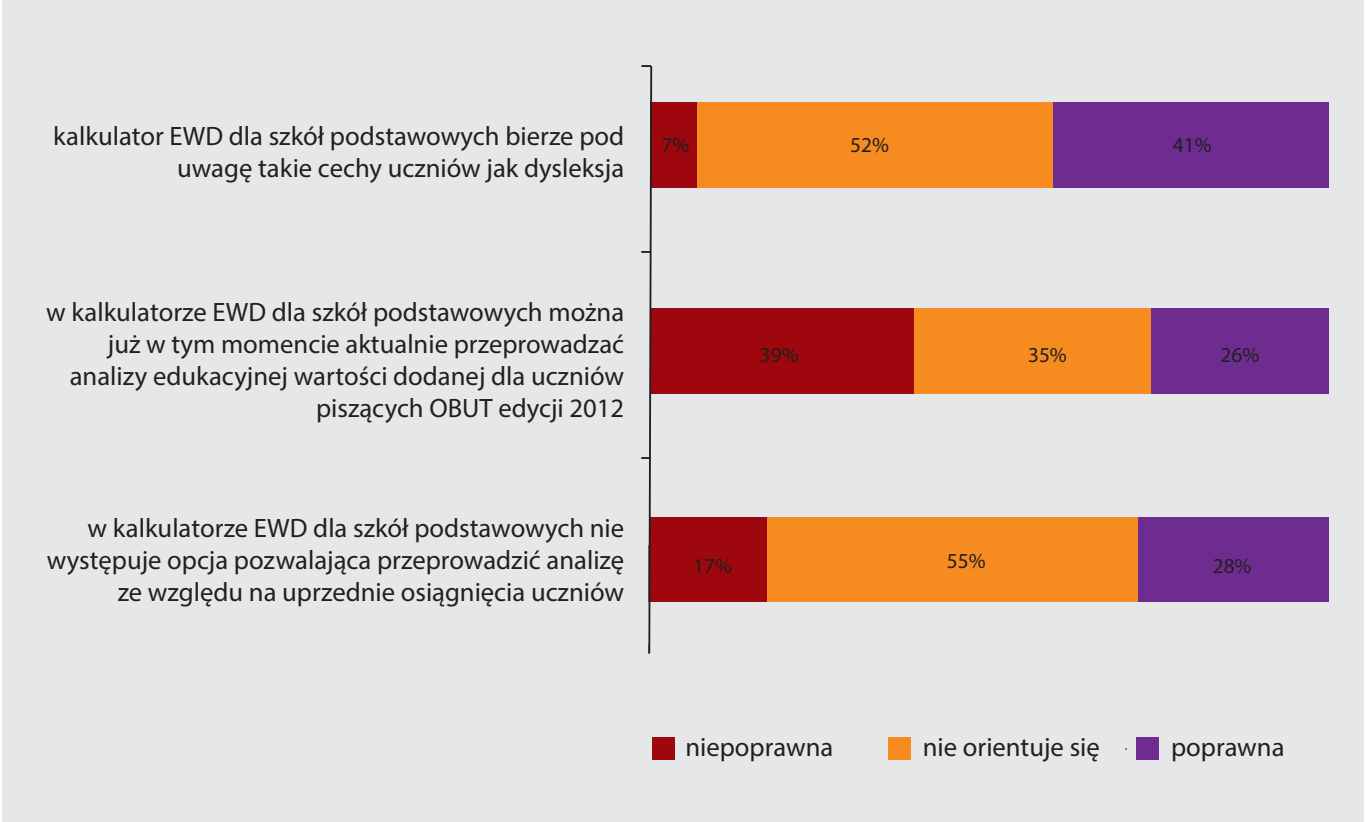


Źródło: Opracowanie własne na podstawie badania ankietowego IBE wśród dyrektorów przeprowadzonego w 2014 r.. Podstawa procentowania: dyrektorzy, których słyszeli wcześniej o EWD (n = 390).

Jednocześnie dyrektorzy, którzy deklarują, że korzystali już z Kalkulatora EWD, mają pewne istotne braki w wiedzy na temat funkcji tego narzędzia (por. rysunek 5.16). Większość badanych w tej grupie (52%) nie wiedziała, czy Kalkulator EWD bierze pod uwagę dysleksję uczniów (prawidłowa odpowiedź: tak). Podobnie, większość (55%) nie wiedziała, czy w Kalkulatorze EWD występuje opcja

pozwalająca przeprowadzić analizę ze względu na uprzednie osiągnięcia uczniów (prawidłowa odpowiedź: tak). Jedynie 26% dyrektorów szkół podstawowych prawidłowo odpowiedziało na pytanie, czy w momencie badania można było analizować EWD dla uczniów piszących OBUT w edycji 2012. Oznacza to, że tylko te osoby stosunkowo dobrze orientowały się w EWD dla szkół podstawowych.

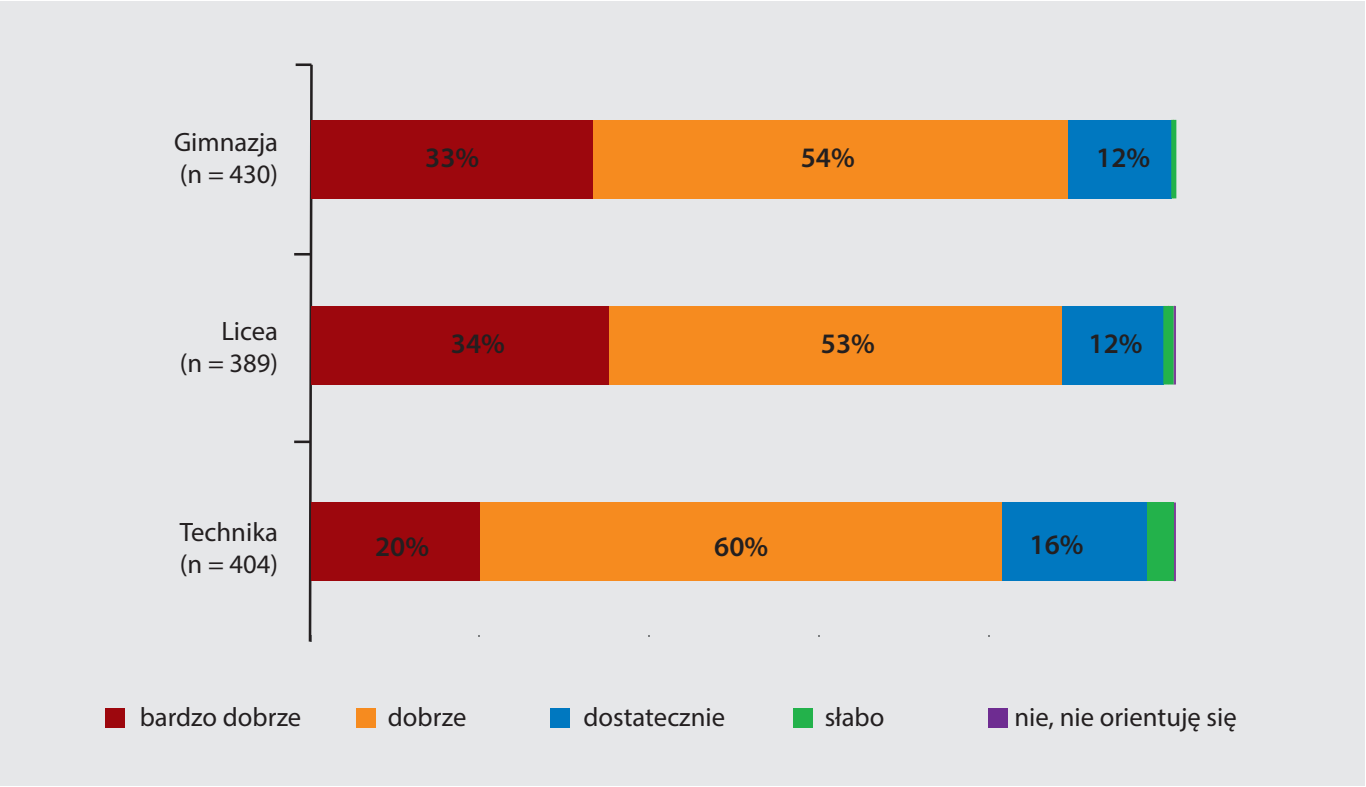
Rysunek 5.16. Odpowiedzi dyrektorów na pytanie o wiedzę o Kalkulatorze EWD dla szkół podstawowych



Źródło: Opracowanie własne na podstawie badania ankietowego IBE. Podstawa procentowania: dyrektorzy, których zapoznali się z kalkulatorem EWD dla szkół podstawowych (n = 126).

Jeśli chodzi o dyrektorów gimnazjów, liceów i techników, to większość ocenia swoją wiedzę w zakresie EWD co najmniej wysoko (por. rysunek 5.17). Większość (87%) dyrektorów liceów i gimnazjów uważa, że dobrze lub bardzo dobrze orientuje się w EWD. Nieco niższe oceny obserwujemy u dyrektorów techników. Wynika to w dużej mierze stąd, że technika od stosunkowo niedawna mają do dyspozycji EWD oraz z faktu, że często dla techników ważniejszym źródłem informacji o pracy szkoły są wyniki egzaminów zawodowych.

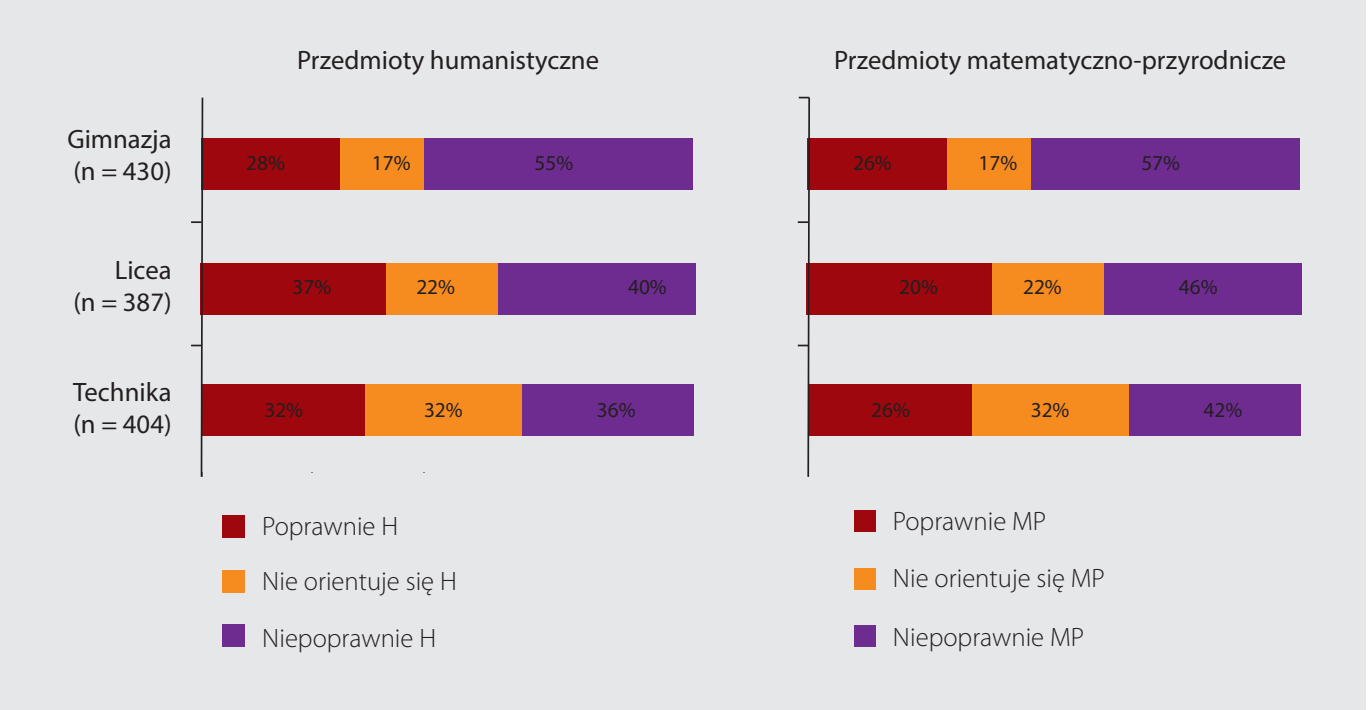
Rysunek 5.17. Samoocena wiedzy dyrektorów gimnazjów, liceów i techników na temat EWD według typu szkoły



Źródło: Opracowanie własne na podstawie badania IBE prowadzonego techniką CATI wśród dyrektorów w 2013 r.. podstawa procentowania: wszyscy zbadani.

Jednocześnie jednak badani niezbyt orientują się w zakresie wartości trzyletnich wskaźników osiągniętych przez ich szkoły (por. rysunek 5.18). Z przedmiotów humanistycznych jedynie 37% dyrektorów liceów, 32% techników i 28% gimnazjów potrafiło w 2013 roku podać rzeczywistą wartość EWD. Natomiast z matematyczno - przyrodniczych odpowiednio – 32% dyrektorów liceów oraz 26% gimnazjów i techników. Natomiast z matematyczno-przyrodniczych odpowiednio – 32% liceów oraz 26% gimnazjów i techników. Należy jednak zauważyć, że odpowiedź na pytanie o wartości wskaźników mogła przysparzać niektórym dyrektorom trudności w przypadku, gdy uzyskany przez szkołę wskaźnik EWD był istotnie statystycznie różny od zera.

Rysunek 5.18. Poprawność wskazania wartości trzyletnich EWD z przedmiotów humanistycznych i matematyczno-przyrodniczych uzyskiwanych przez szkołę, którą kieruje badany dyrektor, według typu szkoły



Źródło: Opracowanie własne na podstawie badania IBE prowadzonego techniką CATI wśród dyrektorów szkół w 2013 roku. Podstawa procentowania: zbadani, którzy wskazali, iż orientują się, czym jest EWD.

Podsumowując, deklaracje dyrektorów pokazują, że ich wiedza na temat edukacyjnej wartości dodanej może nie być wystarczająca, aby mogli w pełni wykorzystać potencjał tego narzędzia i zarazem upowszechniać jego wykorzystanie wśród nauczycieli. Hipotezę tę wydają się również potwierdzać badania jakościowe IBE wśród nauczycieli i dyrektorów. Należy jednak zauważyć, że w badanych jakościowo liceach, technikach i gimnazjach to dyrektorzy prezentowali wyższy poziom wiedzy o EWD niż nauczyciele. Okazuje się, że to kadra kierownicza głównie pozyskiwała informacje o wskaźnikach EWD na konferencjach dla dyrektorów i naradach szkoleniowych. Niepełna wiedza i brak umiejętności nauczycieli w zakresie analiz EWD może stanowić zatem istotne utrudnienie na drodze do pogłębionych analiz wyników egzaminacyjnych.

5.4.3. Postawy i przekonania

Względem systemu egzaminów zewnętrznych

Kolejnym czynnikiem są postawy i przekonania dyrektorów szkół i nauczycieli względem samego systemu egzaminów. Badano je w 2009 r., zadając im pytania otwarte oraz prosząc o ustosunkowanie się do listy stwierdzeń (Rappe, 2013). Dyrektorzy i nauczyciele dostrzegali wówczas zarówno dobre, jak i złe strony egzaminów. W wypowiedziach nauczycieli podstawową wskazywaną korzyścią egzaminów zewnętrznych jest ich pozytywny wpływ na motywację uczniów i nauczycieli, doskonalenie metod pracy i zwiększanie odpowiedzialności nauczycieli za efekty kształcenia. Rzadziej wymieniane przez nauczycieli korzyści wiążą się ze sprawiedliwością wyników czy zwiększeniem ich obiektywności.

W przypadku słabych stron systemu egzaminacyjnego wskazywano na zjawisko uczenia w szkole „pod egzamin", „pod klucz", a także – związanego z tym - ograniczenia we wyrażaniu poglądów i braku możliwości wyrażania sądów. Egzaminom zarzucano premiowanie schematyczności, „zabijanie twórczości i kreatywności". Spory odsetek dyrektorów i nauczycieli widział również w egzaminach zagrożenie niemiarodajnego i niesprawiedliwego oceniania szkół wyłącznie przez pryzmat

wskaźników egzaminacyjnych. Podobne opinie rzadziej – ale na tle innych zagadnień także stosunkowo często – pojawiały się wśród wypowiedzi dyrektorów i nauczycieli gimnazjów. Część dyrektorów szkół podkreśla negatywny wpływ egzaminów w postaci tworzenia rankingów szkół, rywalizacji między szkołami i selekcji szkół. Nauczycieli niepokoiło również ocenianie szkół/nauczycieli przez pryzmat wyników egzaminacyjnych. Większość nauczycieli nie uważa, aby wynik sprawdzianu „zależał przede wszystkim od pracy nauczyciela". Nie zgadzają się również ze stwierdzeniem, że wyniki sprawdzianów/egzaminów „pozwalają organom prowadzącym racjonalnie oceniać szkoły" czy oceniać „skuteczność nauczania w szkole". Ponad połowa nauczycieli nie zgadza się również z opinią, że „zadania na sprawdzianie dobrze sprawdzały to, co w edukacji jest najważniejsze".

Tabela 5.5. Opinie nauczycieli szkół podstawowych i gimnazjów o systemie egzaminów zewnętrznych (%)

Stwierdzenie	Nauczyciele SP			Nauczyciele gimnazjów		
	tak	nie	trudno powiedzieć	tak	nie	trudno powiedzieć
1. Wyniki sprawdzianu/egzaminu pozwalają na ocenę skuteczności nauczania w szkole.	35	56	10	24	63	12
2. Sprawdzian/egzamin motywuje uczniów do systematycznej nauki.	30	54	16	36	51	13
3. Wynik sprawdzianu/egzaminu zależy przede wszystkim od pracy nauczyciela.	10	86	4	9	86	5
4. Sprawdzian/egzamin mobilizuje szkoły do ciągłego wysiłku i doskonalenia metod nauczania.	69	21	10	77	16	7
5. Sprawdzian/egzamin zwiększa społeczne zainteresowanie oświatą.	45	30	25	44	27	29
6. Zadania na sprawdzianie/egzaminie dobrze sprawdzają to, co w edukacji jest najważniejsze.	25	51	24	25	54	21
7. Sprawdzian/egzamin przygotowuje uczniów do radzenia sobie w trudnych sytuacjach.	34	49	17	26	55	19
8. Informacje o wynikach szkoły na egzaminie pozwalają lepiej planować działania dydaktyczne.	62	23	15	64	19	17
9. Wynik sprawdzianu/egzaminu jest dobrym kryterium rekrutacji uczniów do szkoły gimnazjalnej/ponadgimnazjalnej.	25	56	19	42	42	17
10.Wyniki sprawdzianów/egzaminów pozwalają organom prowadzącym racjonalnie oceniać szkoły.	15	70	15	25	64	11
11. Jednolity w całym kraju sprawdzian/egzamin pozwala porównywać osiągnięcia szkolne uczniów.	66	22	12	79	16	5
12. Sprawdzian/egzamin zwiększa zaangażowanie nauczycieli w pracę.	52	32	16	57	31	12
13. Szkoły zamieniają ostatnie miesiące nauki w ciągłe przerabianie zadań testowych i niekończący się próbny sprawdzian.	53	36	10	43	49	8

Stwierdzenie	Nauczyciele SP			Nauczyciele gimnazjów		
	tak	nie	trudno powiedzieć	tak	nie	trudno powiedzieć
14. Ocena pracy szkoły i nauczycieli na podstawie wyników sprawdzianu/egzaminu jest niesprawiedliwa.	80	13	7	78	12	10
15. Szkoły pod presją sprawdzianów/egzaminów coraz mniej chętnie przyjmują słabszych uczniów.	38	36	25	36	42	22
16. Nauczyciele pomijają ważne treści programowe nieobecne na sprawdzianie/egzaminie.	21	61	18	21	63	15
17. Szkoły skupiające najbardziej uzdolnionych uczniów często spoczywają na laurach.	25	43	32	35	39	26
18. Nauczyciele pracujący ze słabszymi uczniami odczuwają frustrację, wynikającą z niskich wyników swoich uczniów na sprawdzianie.	81	11	8	74	17	10
19. Sprawdzian/egzamin zmniejsza szanse uczniów wywodzących się z biednych rodzin na dostanie się do dobrej szkoły gimnazjalnej.	35	47	17	22	68	10
20. Sprawdzian/egzamin, zamiast skłaniać szkoły do lepszej pracy, napędza rynek korepetycji i różnego typu kursów.	51	28	21	35	50	15
21. Wynik sprawdzianu/egzaminu jest w dużej mierze przypadkowy.	37	46	17	32	53	15
22. Sprawdzian/egzamin osłabia bieżącą motywację uczniów do nauki, skłaniając do uczenia się na ostatnią chwilę.	25	55	20	34	54	12
23. Wynik sprawdzianu w zbyt dużym stopniu decyduje o dalszych losach szkolnych ucznia.	45	39	17	51	37	12
24. Sprawdzian/egzamin osłabia autorytet nauczyciela.	21	56	23	14	70	16

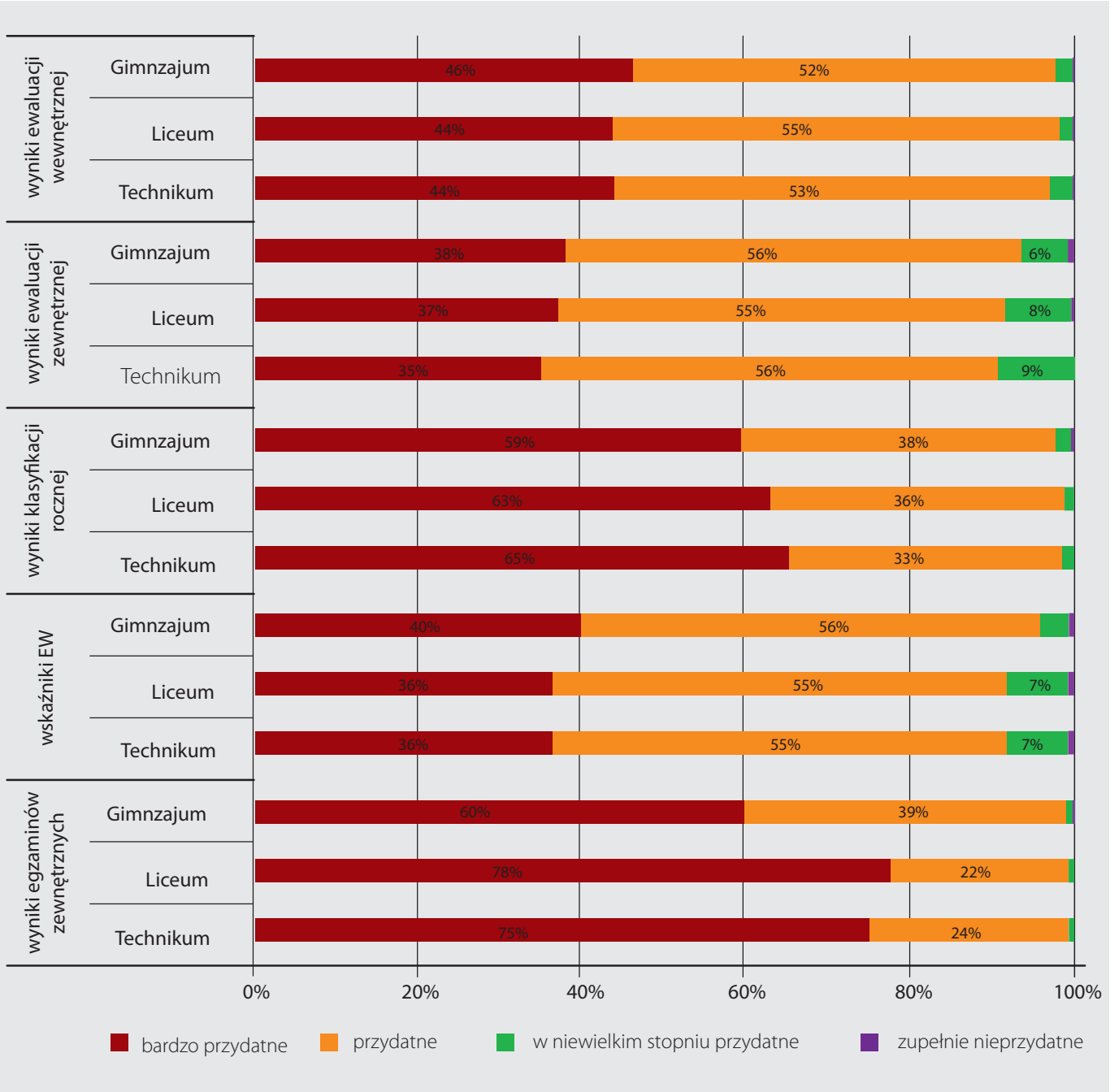
Źródło: Opracowanie własne na podstawie wyników badań przedstawionych w: Peter, J. i Rappe A. (2010) Opinie nauczycieli i egzaminatorów o systemie egzaminów zewnętrznych, Toruń: KDE. N = 2432 nauczycieli szkól podstawowych i 2839 nauczycieli gimnazjalnych.

Warto przy tym zauważyć, że ci nauczyciele, którzy są jednocześnie egzaminatorami, prezentują postawy mniej krytyczne względem systemu niż pozostali nauczyciele (Rappe i Peter, 2010). Zauważają oni więcej zalet oraz mniej wad egzaminów zewnętrznych (choć ich opinia zależy również od rodzaju egzaminu).

Względem EWD i efektywności kształcenia

Dyrektorzy gimnazjów, liceów i techników deklarują (por. rysunek 5.19), że wskaźniki EWD są przydatnym źródłem informacji w zarządzaniu szkołą. Niewielka część dyrektorów jest zdania, że wskaźniki nie są przydatne lub że są przydatne tylko w niewielkim stopniu.

Rysunek 5.19. Ocena przydatności różnych informacji w zarządzaniu szkołą, wg typu szkoły



Źródło: Opracowanie własne na podstawie badania IBE prowadzonego techniką CATI wśród dyrektorów szkół w 2013 roku. gimnazja N=430, licea N = 389, technika N = 404.

Bardziej zniuansowany obraz postaw dyrektorów i nauczycieli pokazują badania jakościowe. Początkujący użytkownicy metody posiadają jedynie ogólne informacje na temat EWD, zetknęli się ze sposobem prezentacji wyników w formie elips na wykresach, potrafią odnaleźć wskaźniki 3-letnie własnej placówki. Wśród nich zaobserwowano nauczycieli zarówno otwartych na uczenie się, jak i prezentujących postawę „zamkniętą". Niektórzy użytkownicy są przekonani, że efektywności nauczania nie da się zmierzyć żadnymi wskaźnikami. Użytkownicy bardziej otwarci dostrzegają wartość informacyjną tego wskaźnika i planują wdrożenie EWD do swojego warsztatu pracy nauczyciela, natomiast zamknięci – wręcz przeciwnie.

Niektórzy użytkownicy, szczególnie w szkołach o niskim – w ich poczuciu – EWD, uważają, że wskaźnik ten nie odzwierciedla wysiłku, jaki wkładają w podniesienie efektywności nauczania. Tłumaczą, że są w związku z tym zdemotywowani, aby poszerzać swoją wiedzę w zakresie stosowania EWD.
Część kadry szkół, szczególnie tych charakteryzujących się wysokimi – w ich odczuciu – wskaźnikami EWD, uważa, że analizy EWD są mało użyteczne, bo potwierdzają tylko to, co było już wiadomo wcześniej.
Niektórzy nowi użytkownicy EWD, zwłaszcza ze szkół podstawowych (Pfeiffer, 2015) uważają, że zarówno gromadzenie danych, jak i analizy są zbyt czasochłonne. Pojawiają się też obawy dotyczące błędnych sformułowań wniosków z analiz EWD, wynikających z braku kompetencji prowadzących analizy (Kowalewska, 2014)
Warto również wspomnieć, że niektórzy nauczyciele i dyrektorzy szkół podstawowych podchodzą do analiz EWD z pewną rezerwą, z uwagi na swoje wątpliwości odnośnie do jakości danych OBUT, które są wykorzystywane do wyliczenia EWD dla szkół podstawowych (Pfeiffer, 2015). Uważają oni, że sposób oceny arkuszy diagnostycznych OBUT przez wychowawców klas które piszą OBUT, a nie egzaminatorów zewnętrznych, nie gwarantuje rzetelności tych danych. Jednocześnie jednak, w tych szkołach, w których arkusze OBUT sprawdzane były wspólnie przez wychowańców klas III, przez nauczycieli klas IV lub wychowawcę innej klasy niż oceniana, opinie o niskiej rzetelności wyników OBUT, a tym samym EWD, występują rzadziej.
We wszystkich typach szkół, dla których liczone są wskaźniki EWD, można odnotować przypadki zgłaszania obaw grona pedagogicznego dotyczących tego, w jaki sposób informacje, których dostarczają analizy EWD, zostaną wykorzystane przez otoczenie szkoły, w tym szczególnie przez organy prowadzące, ale także nadzór pedagogiczny, rzadziej – rodziców uczniów. Nauczyciele obawiają się, że brak wiedzy pracowników organów prowadzących powodować może mylne interpretacje wskaźników EWD i podejmowanie przez nich nieadekwatnych, krzywdzących szkoły i nauczycieli decyzji: „Uczestnicy [badania uczestniczącego w SP] wyrażali wiele obaw dotyczących poprawnego wykorzystania wskaźnika przez organy prowadzące i nadzorujące oraz innych odbiorców informacji o wynikach egzaminacyjnych. Dopytywali o pojawiające się w prasie artykuły na temat EWD w kontekście wykorzystania wskaźnika do typowania szkół do likwidacji. Atmosfera wokół nowej miary jest dość „nerwowa" (Pfeiffer, 2015, s. 22).
Na podejście do analiz wyników egzaminacyjnych EWD i zaufanie do tej metody mają również wpływ przekonania kadry szkoły odnośnie do czynników, które wpływają na efektywność nauczania i tego, czy czynniki te brane są pod uwagę w oszacowaniach EWD. Niektórzy uważają, że szkoła ma istotny wpływ na ucznia, inni natomiast, że jednak niewielki. Dla tej ostatniej grupy osiągnięcia uczniów zależą w głównej mierze od czynników środowiskowych i indywidualnych cech ucznia, będących poza oddziaływaniem szkoły. Ta postawa jest spójna również z ich przekonaniem, że na drodze do wysokiej efektywności nauczania stoją ciągłe zmiany w podstawie programowej. Tak o zaobserwowanych postawach tego typu w niektórych szkołach podstawowych pisze Pfeiffer (2015, s. 28): „Nauczyciele, szukając wyjaśnienia poziomu efektywności nauczania w szkole, odnoszą się głównie do czynników rodzinnych, środowiskowych oraz indywidualnych (brak współpracy z rodzicami, brak zaangażowania rodziców w kreowanie sukcesu edukacyjnego dziecka, brak motywacji uczniów, dysfunkcje). Widoczny jest opór przed szukaniem powiązań efektywności z czynnikami pedagogicznymi, szkolnymi."
Inną postawę względem EWD i mierzenia efektywności nauczania prezentują ci spośród grona pedagogicznego, którzy – chociaż świadomi czynników środowiskowych - jednocześnie są przekonani, że szkoła może mieć istotny wpływ na uczniów. Łączą oni efektywność nauczania z takimi działaniami szkoły jak: indywidualizacja nauczania, współpraca z rodzicami, liczba i jakość zajęć wyrównawczych oraz zajęć ukierunkowanych na wspieranie ucznia zdolnego, wykorzystaniem wniosków z analiz wyników egzaminacyjnych, diagnozy zainteresowań uczniów, koleżeńskie obserwacje lekcji, a w szkołach podstawowych – szczególnie ze współpraca nauczycieli klas III i IV.

Przekonanie, że analizy nie stanowią elementu pracy nauczyciela

Pomimo tego, że od kilku lat szkoły są zobowiązane przeprowadzać analizy wyników egzaminacyjnych, to jednak wciąż część nauczycieli nie postrzega tego typu działań jako element swojego warsztatu dydaktycznego. Zapoznając się z takimi narzędziami jak Kalkulator EWD, obawiają się czasochłonności analiz, nie są skłonni się ich uczyć, gdyż nie łączą wyników analiz ze swoją pracą dydaktyczną. Badania pokazują, że procesy mające pomóc ocenić jakość pracy szkoły i nauczycieli, jak np. ewaluacja wewnętrzna czy właśnie analizy wyników egzaminacyjnych postrzegane są przez niektórych nauczycieli jako kwestie niejako dodatkowe, rozgraniczane od pracy dydaktyczno-wychowawczej, a nie jak ich komplementarna część.

Obawy o pogorszenie relacji koleżeńskich wśród kadry szkoły

Okazuje się również, że niektórzy nauczyciele podchodzą z rezerwą do analiz wyników egzaminacyjnych dlatego, że obawiają się, że przyczynią się one do zwiększenia rywalizacji pomiędzy nauczycielami i pogorszą atmosferę panującą w szkole. Obawiając się o wzajemne obwinianie za niezadowalające wyniki szkoły, użytkownicy ci wolą raczej unikać tego tematu. Lęki te ilustruje następujący cytat dotyczący EWD: „Nauczyciele pozytywnie odnieśli się do EWD na poziomie szkoły, obawiali się, że analizy EWD na poziomie klas spowodują pogorszenie relacji koleżeńskich między nauczycielami, rywalizację, lęk przed rozliczalnością i odpowiedzialnością." (Pfeiffer, 2015, s. 24)

Motywacja wewnętrzna do rozwoju

Kluczowym czynnikiem sprzyjającym wykorzystaniu wyników egzaminacyjnych jest motywacja wewnętrzna użytkowników do rozwoju, ich ciekawość poznawcza. Potwierdzają to wypowiedzi dyrektorów i nauczycieli: „Oczekiwaliśmy tego. Także skłoniła nas własna ciekawość i chęć zobaczenia"[nauczyciel liceum, id 16, 2014] oraz „(…) przyczyną zastosowania EWD w szkole była chęć posiadania informacji, czy idziemy w dobrym kierunku, w dobrym tempie…" [dyrektor, id 16, 2014]
Aby możliwe było sformułowanie hipotez dotyczących możliwych przyczyn niskich wyników egzaminacyjnych, czy też niskiej efektywności nauczania, istotna jest również otwartość użytkowników na krytykę, szukanie przyczyn niepowodzeń również we własnym postępowaniu, a nie tylko w czynnikach zewnętrznych. Podczas obserwacji przeprowadzania analiz EWD w szkołach podstawowych zauważono, że u wielu użytkowników EWD występuje „zbyt silna potrzeba obarczania innych i samousprawiedliwiania się. Jeżeli jest sukces, każdy z obecnych ma w tym swoją część udziału; jeśli porażka – nie ma winnych wśród nas." (Pfeiffer, 2015, s. 34)
Decydujące znaczenie ma oczywiście także poczucie celowości prowadzenia analiz, wynikające z wewnętrznych przesłanek, a także pewnego rodzaju odwaga konieczna do konfrontacji swojej pracy z jej efektami. W tym kontekście podkreślana bywa przez użytkowników chęć uzyskania „obiektywnej" oceny efektów swojej pracy, z niejako zewnętrznego, wiarygodnego źródła.

5.4.4. Organizacja pracy szkoły

Każda szkoła charakteryzuje się określoną kulturą organizacyjną, a jednym z jej elementów jest kultura pracy z danymi. Zależy ona między innymi od postawy i działań dyrekcji szkoły oraz współpracy grona pedagogicznego, ale także od wizji pracy placówki oraz dostępu kadry szkoły do wsparcia w przeprowadzaniu analiz danych, w tym wyników egzaminacyjnych.

Przywództwo

Wykorzystanie danych egzaminacyjnych przez dyrektora szkoły może sprzyjać sprawniejszemu podejmowaniu przez niego decyzji. Ponadto, aktywna postawa dyrektora - jego kompetencje analityczne oraz przekonanie o tym, że wyniki egzaminów zewnętrznych niosą informacje, które mogą być użyteczne dla nauczycieli w ich pracy dydaktycznej i rozwoju zawodowym – jest kluczowa dla rozwoju wysokiej kultury pracy z danymi. Z badań międzynarodowych wynika, że to w jaki sposób

dane są analizowane, interpretowane i wykorzystane w szkole zależy przede wszystkim od dyrektora placówki (Vanhoof, Vanlommel, Thijs, Vanderlocht, 2013) oraz współpracy grona pedagogicznego. Potwierdzają to również Fischer i Taylor (2012), według, których doskonalenie pracy szkoły wymaga odpowiedniego wykorzystania danych, współpracy między nauczycielami i skutecznego przywództwa. Wnioski te znajdują również poparcie w badaniach IBE.
Na przykład, w badanych gimnazjach oraz szkołach ponadgimnazjalnych, tam gdzie dyrektor był przekonany o użyteczności danych egzaminacyjnych oraz EWD, inicjował gromadzenie ich i udostępnianie wszystkim nauczycielom, regularne przeprowadzanie analiz w zespołach przedmiotowych (lub zespołach zadaniowych), w takich placówkach poziom analizy danych ma szansę przełożyć się na rozwój nauczycieli i całej szkoły.

Współpraca nauczycieli

Jednak bez współpracy nauczycieli, wspólnej akceptacji wniosków z analiz wyników egzaminów zewnętrznych, działania naprawcze nie mogą przynieść szkole oczekiwanych korzyści. Wspólne analizy danych egzaminacyjnych powinny prowadzić do refleksji nad ich znaczeniem w całym gronie pedagogicznym, do współpracy między-przedmiotowej: *Coś trzeba zmienić. Każdy tam w obrębie własnego przedmiotu (…) problem jest i jak gdyby każdy pracuje nad własnym przedmiotem i efekt jest mniejszy lub większy, ale jest, ale nie pojawia się głębsza współpraca między przedmiotami."* [nauczyciel, id 46, 2014]

Czas na dialog wokół wyników egzaminacyjnych

Wykorzystaniu wyników egzaminacyjnych powinna sprzyjać odpowiednia organizacja systemu analiz, pewna instytucjonalizacja procesu analiz w ramach szkoły, tj. istnienie jasnych zasad gromadzenia danych, wyznaczenie osób odpowiedzialnych za przeprowadzanie analiz, a także przeznaczenie w ramach organizacji pracy szkoły odpowiedniego czasu na regularną pracę nauczycieli z danymi. W wielu badanych szkołach powszechnie wskazywanym przez nauczycieli problemem z analizami wyników egzaminacyjnych jest kwestia braku czasu (P01, Kowalewska, 2015), w tym konieczność ich realizacji kosztem innych – z perspektywy dyrektorów i nauczycieli – bardziej istotnych zadań związanych z dydaktyką. Podobne problemy zidentyfikowano np. w odniesieniu do prowadzonej w szkołach ewaluacji wewnętrznej (Wasilewska i in, 2014).

5.4.5. Dostęp do wspomagania placówki w analizach danych

Aby możliwe było podnoszenie kompetencji użytkowników w zakresie przeprowadzania analiz wyników egzaminacyjnych istotne jest zapewnienie placówkom dostępu do odpowiedniej jakości wsparcia. Tego typu pomoc, na stosunkowo niewielką - w stosunku do ogólnopolskich potrzeb - skalę, była zapewniona szkołom w projekcie systemowym EWD, który jednak miał charakter głównie badawczy, a nie upowszechniający. Podczas kilku lat trwania tego projektu wypracowano jednak pokaźny pakiet materiałów dydaktycznych na szkolenia użytkowników analiz danych. Wyszkolono również grono ekspertów EWD w każdym województwie, którzy wypracowują dobre praktyki szkoleniowe pracy z danymi w ramach całych rad pedagogicznych. Bardzo ważne jest aby te rozwiązania zostały włączone do systemu, aby ze gromadzonej wiedzy i doświadczenia mogły korzystać wszystkie szkoły w Polsce.

5.4.6. Oczekiwania otoczenia szkoły

Postawy i wymagania podmiotów zewnętrznych – głównie organów nadzoru pedagogicznego i organów prowadzących - oddziałują na to, czy kadra szkoły będzie patrzeć na wyniki egzaminów jako źródło informacji kluczowe dla kształtowania swojej pracy i rozwoju szkoły, czy też będzie

koncentrować się na tym, jak dane zostaną odebrane na zewnątrz, jakie konsekwencje dla szkoły mogą przynieść.

Nie bez znaczenia są także oczekiwania rodziców uczniów, zwłaszcza w kontekście ich rosnących aspiracji edukacyjnych względem dzieci. Wyniki przez nie osiągane osiągane są jednym z istotnych czynników branych przez rodziców pod uwagę przy ocenie pracy szkoły, jednak – jak pokazują badania - nie zawsze najważniejszym. Na zadowolenie rodziców ze szkoły znaczący wpływ ma na przykład opinia na temat ogólnej atmosfery panującej w szkole (Badania rodziców, 2009).

Nadzór pedagogiczny

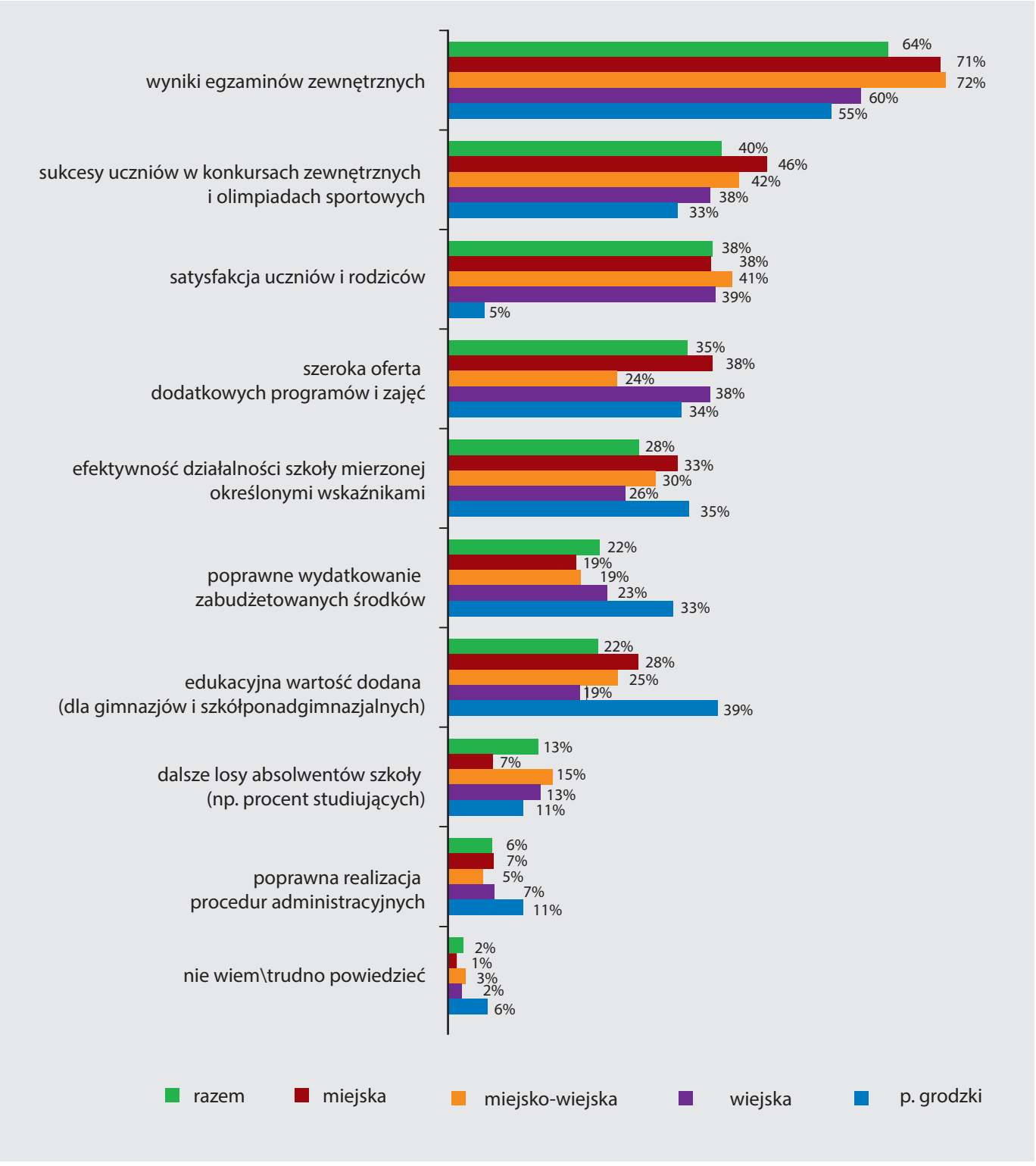
Wykorzystanie wniosków z analizy wyników egzaminacyjnych stanowi jedno z wymagań państwa wobec szkół, sprawdzane w ramach ewaluacji zewnętrznej. Wiemy też, że wyniki egzaminacyjne są wykorzystywane przez wizytatorów do oceny spełniania przez szkoły innych wymagań. Z jednej strony zatem już samo doświadczenie ewaluacji zewnętrznej, kontakt z wizytatorami, pracownikami kuratorium może być bodźcem wpływającym na podejście szkół do analiz egzaminacyjnych. Z drugiej strony, kuratoria oświaty, mogą również podejmować inne działania, które mogą sprzyjać wykorzystaniu przez szkoły wyników egzaminacyjnych i EWD, np. poprzez organizowanie konkursów na granty szkoleniowe. Pytanie brzmi, czy będzie to bodziec pozytywny, tj. inspirujący szkołę do bardziej celowej, wynikającej z jej potrzeby, pracy z danymi egzaminacyjnymi, czy też czy będzie prowadził raczej do pozorowanych analiz wyników egzaminacyjnych. W tym kontekście pojawiają się dwa kluczowe, zidentyfikowane podczas badań, problemy. Otóż niektórzy pracownicy nadzoru są zdania, że rozmowa o sposobach prowadzenia analiz wyników egzaminacyjnych, czy możliwych korzyści z nich wynikających nie należy do ich zadań jako wizytatorów ds. ewaluacji: *„W związku ze specyfiką naszej pracy, z jednej strony nadzór pedagogiczny ten nowy, w którym oddzielono wspomaganie od jednak nadzoru czystego w sensie ewaluacja a kontrola, my nie możemy wpadać w coś takiego jak szkolenie rad pedagogicznych…"* [wizytator, id 07, 2013] Trzeba również mieć na uwadze problem bardzo niepełnej, bądź powierzchownej wiedzy części wizytatorów na temat analiz wyników egzaminacyjnych oraz EWD.

Organy prowadzące

Jednostki samorządu terytorialnego, pełniące funkcje organów prowadzącymi szkoły zobowiązane są do corocznego składania swoim organom stanowiącym informacji o stanie realizacji zadań oświatowych. Z zapisu art. 5a ust. 4 ustawy o systemie oświaty wynika, że w ramach takiej informacji powinny zostać przedstawione także wyniki egzaminów zewnętrznych. Samorząd ma dużą swobodę w tym w jaki sposób i w jakich celach wykorzystuje dane egzaminacyjne do zarządzania oświatą na swoim terenie. Potencjalne ich wykorzystanie zależy zatem od indywidualnego podejścia danego organu prowadzącego. Należy mieć świadomość występujących wśród niektórych samorządowców opinii, że kwestia monitorowania efektów pracy szkoły i prowadzenie w tym celu analiz wyników egzaminacyjnych nie leży w zakresie kompetencji samorządu: *„To nie jest zadanie kompetencyjnie przypisane organowi prowadzącemu. To jest otoczenie kompetencji organu prowadzącego. Co oznacza, że jeżeli będą tam ciekawi ludzie, to się tym zajmą, a jak nie będzie ciekawych ludzi… Nikt przeważnie nie chce brać sobie więcej roboty niż podejmowanie to, co musi."* [pracownik organu prowadzącego, id 30, 2013]

Jednocześnie jednak wśród wielu przedstawicieli samorządów zaobserwowano duże zainteresowanie wynikami egzaminacyjnymi jako wyznacznikami jakości pracy szkoły. Z ogólnopolskiego badania przeprowadzonego w 2012 roku na zlecenie ORE na próbie 320 gmin (UW/Millward Brown/ORE, 2012), wynika, że wyniki egzaminów zewnętrznych są dla organów prowadzących właśnie najważniejszym czynnikiem określającym jakość pracy szkoły (dla 64% badanych), w dalszej kolejności pojawiły się sukcesy uczniów w konkursach i olimpiadach (40%), a dopiero na trzecim miejscu satysfakcja uczniów i ich rodziców (38%). Edukacyjna wartość dodana pojawiała się jako jeden z 3 najważniejszych czynników w wypowiedziach 22% respondentów.

Rysunek 5.20. Najważniejsze aspekty jakości pracy szkoły według typów gmin



Źródło: opracowanie własne na podstawie danych z Zarządzanie oświatą w gminach. Raport z badania ankietowego, Uniwersytet Warszawski/ORE/ Millward Brown, Warszawa 2012; N=320.

Należy podkreślić jednak, że korzystanie z wyników egzaminacyjnych przez pracowników samorządów często wiąże się z chęcią tworzenia przez nich porównań pomiędzy szkołami. Na przykład, niektórzy pracownicy samorządów, w celu przygotowania rankingów szkół na podstawie EWD zaczęli wręcz stosować „quasi-metody" pomiaru na elipsach wartości punktowych: „*Respondenci stosują*

„chałupnicze" metody analizy EWD np. wymierzając linijką środek elipsy i wykorzystując tę wiedzę do porównań wskaźników EWD między szkołami." (Raport G_01, IBE, 2013, s.17)

Wskazać można jednak również przykłady samorządów, które starają się wykorzystywać wnioski z analiz wyników egzaminacyjnych do podejmowania konkretnych decyzji mających wpływ na funkcjonowanie szkół. Niektóre wydają się wykorzystywać wnioski z analiz wyników egzaminów i EWD aby wspomagać szkoły (Matuszczak i in. 2013), np. poprzez przyznawanie nagród finansowych dyrektorom uzyskującym wysokie wskaźniki efektywności kształcenia w swoich placówkach, planowanie wydatków na doszkalanie nauczycieli, czy też dodatkowe godziny zajęć lekcyjnych w szkołach, które nie uzyskują zadowalających wskaźników EWD. Z kolei innym samorządom wyniki egzaminacyjne i EWD pomagają np. w podejmowaniu decyzji dotyczących restrukturyzacji sieci placówek. Brakuje jednak danych, aby ocenić, jaka jest skala powyższych podejść organów prowadzących. Natomiast ważne wydaje się nie tylko to, że taki sposób postępowania ma miejsce, ale również sama świadomość szkół, że tak może być. Badania pokazują, że takie obawy wśród szkół są dosyć powszechne.

5.5. Podsumowanie

W wielu krajach i różnych systemach edukacyjnych duża ilość dostępnych danych niekoniecznie przekłada się na odpowiednie ich wykorzystywanie i wartościowe tego efekty (Vanhoof, Schildkamp, 2014). Przed takim wyzwaniem znajdują się także polskie szkoły mające dostęp między innymi do bardzo bogatego źródła informacji jakim są wyniki egzaminacyjne. Pojawia się więc pytanie, jak analizować te dane i w jakim celu je wykorzystywać. Na poziomie deklaracji analizy wyników egzaminacyjnych są powszechnie prowadzone, a ich wyniki uwzględniane w szeregu decyzji podejmowanych przez szkoły. Podobne zresztą są wnioski z analizy poziomu spełniania wymagania związanego z analizą wyników egzaminacyjnych w ramach nadzoru pedagogicznego. Dodatkowo badania międzynarodowe pokazują, że dyrektorzy i nauczyciele z polskich szkół znajdują się zazwyczaj wśród państw, w których najczęściej deklarowane jest wykorzystywania wyników egzaminacyjnych w różnych aspektach. Zaś deklarowane wykorzystywanie wyników egzaminacyjnych w większości państw w ostatnich latach bardzo rośnie. Wyniki badań pokazują, że analiza danych egzaminacyjnych i same wyniki egzaminacyjne postrzegane są przez kadrę szkół jako bardzo istotne źródło informacji, które powinno być wykorzystywane do podejmowania decyzji. Taki sposób myślenia może być jednym z wyznaczników rosnących oczekiwań względem szkół, a w szczególności oczekiwań odnoszących się do osiągania wysokich wyników uczniów.

Z pogłębionych analiz z kolei wynika, że szkoły mają liczne problemy z prowadzonymi analizami, a większość z nich dopiero się uczy jak i w jakim celu analizować i wykorzystywać dane z systemu egzaminacyjnego. Zebrany materiał pokazuje, że motywacja do gromadzenia i przetwarzania danych w szkołach ma zazwyczaj charakter zewnętrzny – związany z regulacjami, które obligują szkoły do wykorzystywania wyników egzaminacyjnych, a nie z poszukiwaniem odpowiedzi na pojawiające się wyzywania, chęcią wprowadzania zmian w szkole. Cele „prorozwojowe" znajdują się więc na dalszym planie, a wyznacznikiem tego może być na przykład fakt, że zdecydowanie bardziej popularne jest odnoszenie wyników szkoły do wyników innych placówek, z gminy, czy regionu niż porównywanie wyników szkoły miedzy latami, czy rzadkie uwzględnianie wśród czynników kontekstowych uwarunkowań związanych z pracą nauczycieli, czy organizacją pracy szkoły. Oczywiście badania zidentyfikowały także przypadki szkół bardziej zaawansowanych w analizie i interpretacji wyników, jednakże nawet w tych placówkach dyrektorzy i nauczyciele mają często trudności z wskazaniem efektów podjętych po analizie wyników egzaminacyjnych działań.

Z perspektywy analiz wyników egzaminacyjnych szczególnie problematyczne jest odpowiednie formułowanie wniosków i rekomendacji na podstawie prowadzonych analiz i potem dalsze ich wdrażanie przez poszczególnych nauczycieli. Podobnie jak w wielu krajach głównym wyzwaniem

okazuje się więc wdrożenie wyników analiz do szkolnej praktyki w klasie. W tej chwili wśród działań podejmowanych po analizach wyników egzaminacyjnych dominuje raczej intensyfikacja pracy nad konkretnymi umiejętnościami/przedmiotami, przede wszystkim poprzez zwiększenie liczby konkretnych ćwiczeń i liczby zajęć na których pracuje się nad umiejętnościami uznanymi za słabiej opanowane, a nie bardziej ogólne doskonalenie sposobów nauczania, czy modyfikacje w warsztacie pracy. Potwierdza to częściowo obserwacje z innych krajów opisane w rozdziale 1 raportu: obok pozytywnych skutków wykorzystywania danych egzaminacyjnych widoczne są też skutki negatywne. Dlatego tak ważne są działania mające na celu wzmacnianie kompetencji analizy danych i uświadamianie ich zalet, ale i ograniczeń, oraz wzbogacania zakresu danych wykorzystywanych przez szkoły.

Bibliografia

Badania rodziców. Przemoc i inne problemy w polskiej szkole. (2009). Pobrane z: http://www.szkola-bezprzemocy.pl/996,badania-rodzicow-2009

Breiter, A. i Karbautzki, L. (2011). Data Use in Schools – A Cross-Country Study. Institute for Information Managmement, University of Brenen. WAB #1792909. Pobrane z: https://sites.google.com/site/comeniusdatause/documents/dissemination-materials/scientific-articles

Earl, L. i Fullan, M. (2003). Using data in leadership for learning, *Cambridge Journal of Education* Vol. 33, No. 3, 383-394 November. Oxfordshire: Carfax Publishing.

Elsner, D. i Bednarek, K. (2012). *Dwa lata ewaluacji wewnętrznej w opiniach środowisk edukacyjnych. Doniesienie z badań.* Publikacja Programu NPSEO. Pobrane z: http://www.npseo.pl/data/various/files/III_1%20Dr%20D_Elsner_K_Bednarek_nowe.pdf

Faubert, V. (2009). *School Evaluation: Current Practices in OECD Countries and a Literature Review*. OECD Education Working Paper No. 42, EDU/WKP(2009)21. Paris: OECD.

Fila, J., Matuszczak, K. i Zielonka, P. (2015). *Wykorzystanie efektów projektu systemowego OBUT przez szkoły*. Warszawa: Instytut Badań Edukacyjnych (raport niepublikowany).

Fischer, J.M. i Taylor, F. (2012). Wspieranie zespołów nauczycieli w procesie podejmowania decyzji. W: G. Mazurkiewicz (red.), *Jakość edukacji. Różnorodne perspektywy* (s. 235-249). Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Gocłowska, A. (red.). (2013). *Ewaluacja zewnętrzna. Poradnik wizytatora*. Warszawa: Ośrodek Rozwoju Edukacji.

Kędracka, E., Matuszczak, K., Rappe, A. i Stożek, E. (2013). *Badanie na temat wykorzystania edukacyjnej wartości dodanej (EWD) przez szkoły ponadgimnazjalne*. Warszawa: Instytut Badań Edukacyjnych.

Konceptualizacje wymagań. Pobrane 18 lutego 2015 i 10 września 2015 z www.npseo.pl

Koniewski, M. (2015). *Materiały szkoleniowe Wiosennej Szkoły EWD*. Warszawa: Instytut Badań Edukacyjnych.

Kowalewska, G. (2014). *Raport z badania uczestniczącego w szkole podstawowej*. Warszawa: Instytut Badań Edukacyjnych (raport niepublikowany).

Ligęza, A. (2013). Analiza danych dotyczących wymagania „Analizuje się wyniki sprawdzianu, egzaminu gimnazjalnego, egzaminu maturalnego i egzaminu potwierdzającego kwalifikacje zawodowe". W: G. Mazurkiewicz (red.), *Jakość edukacji. Dane i wnioski z ewaluacji zewnętrznych prowadzonych w latach 2010-2011* (s. 19-25). Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Ligęza, A. i Franczak, J. (BDW). *Jak analizuje się wyniki egzaminów zewnętrznych w polskich szkołach. Raport z wyników ewaluacji zewnętrznej*. Pobrane 18 lutego 2015 z:

http://www.npseo.pl/data/various/files/Agata%20Lig%C4%99za%20Justyna%20Franczak-%20analiza%20wynik%C3%B3w%202011_12.pdf.

Marciniak, M. i Ronka, D. (2012). *Development of the Data Use Professional Development Course*, Paper for the International Congress for School Effectiveness and Improvement (ICSEI), 25th ICSEI Congress in January 2012, Malmö, Sweden.

Matuszczak, K., Zielonka, P. i Bąbiak, I. (2014). *Raport z ewaluacji wewnętrznej projektu EWD*. Warszawa: Instytut Badań Edukacyjnych.

Milecka, B. (2014a). Jak analizuje się wyniki egzaminów zewnętrznych w polskich szkołach. Raport z wyników ewaluacji. W: G. Mazurkiewicz i A. Gocłowska (red.), *Jakość edukacji. Dane i wnioski z ewaluacji zewnętrznych prowadzonych w latach 2012-2013* (s. 188-197). Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Milecka, B. (2014b). W jaki sposób i jakie analizy wykorzystuje się w polskich szkołach do doskonalenia procesów edukacyjnych. W: G. Mazurkiewicz i A. Gocłowska (red.), *Jakość edukacji. Dane i wnioski z ewaluacji zewnętrznych prowadzonych w latach 2013-2014* (s. 213-227). Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

OECD (2013). Synergies for Better Learning: An International Perspective on Evaluation and Assessment, OECD Publishing.

OECD (2014). TALIS 2013 Results: An International Perspective on teaching and Learning, OECD Publishing. http://dx.doi.org/10.1787/9789264196261-en

Peter, J. i Rappe, A. (2010). Opinie nauczycieli i egzaminatorów o systemie egzaminów zewnętrznych. W: B. Niemierko i M.K. Szmigel (red.), *Teraźniejszość i przyszłość oceniania szkolnego* (s. 459-466). Kraków: Grupa Tomami.

Pfeiffer, A. (2015). *Raport zbiorczy z badania uczestniczącego w szkołach podstawowych*. Warszawa: Instytut Badań Edukacyjnych.

Rappe, A. (2013). Opinie nauczycieli i dyrektorów o systemie egzaminów zewnętrznych na podstawie badań prowadzonych przez zespół EWD. W: B. Niemierko i M.K. Szmigel (red.), *Polska edukacja w świetle diagnoz prowadzonych z różnych perspektyw badawczych* (s. 254-267). Kraków: Grupa Tomami.

Rozporządzenie Ministra Edukacji Narodowej z dnia 27 sierpnia 2015 r. w sprawie nadzoru pedagogicznego.

Rozporządzenie Ministra Edukacji Narodowej z dnia 6 sierpnia 2015 r. w sprawie wymagań wobec szkół i placówek.

Rozporządzenie Ministra Edukacji Narodowej z dnia 10 maja 2013 r. zmieniające rozporządzenie w sprawie nadzoru pedagogicznego.

Rozporządzenie Ministra Edukacji Narodowej z dnia 7 października 2009 r. w sprawie nadzoru pedagogicznego.

Schildkamp, K., Karbautzki, L. i Vanhoof, J. (2014). Exploring data use practices around Europe: Identifying enablers and barriers, *Studies in Educational Evaluation*, 42(2014) 15-24.

Skórska, P., Koniewski, M. i Majkut, P. (2012). Obiektywność oceny stopnia spełnienia wymagań stawianych przez państwo szkołom w ramach Systemu Ewaluacji Oświaty. W: B. Niemierko i M.K. Szmigel (red.), *Regionalne i lokalne diagnozy edukacyjne* (s. 288-299). Kraków: Grupa Tomami.

Stożek, E. (2010). Dane egzaminacyjne w ewaluacji zewnętrznej, *Dyrektor szkoły* nr 9/2010, s.18-21.

Stożek, E. i Hawrot, A. (2014). Dlaczego szkoły analizują wyniki egzaminacyjne? W: B. Niemierko i M.K. Szmigel (red.). *Diagnozy Edukacyjne. Dorobek i nowe zadania* (s. 230-239). Kraków: Grupa Tomami.

Stożek, E., Kędracka, E. i Rappe, A. (2015). *Opis poziomów wykorzystania wskaźników EWD przez szkoły*, artykuł niepublikowany.

Uniwersytet Warszawski, ORE, Millward Brown. (2012). *Zarządzanie oświatą w gminach. Raport z badania ankietowego*, Warszawa.

Vanhoof, J. i Schildkamp, K. (2014). From 'professional development for data use' to 'data use for professional development, *Studies in Educational Evaluation*, 42(2014) 1-4.

Vanhoof, J., Vanlommel, K., Thijs S. i Vanderlocht, H. (2013). Data use by Flemish school principals: impact of attitude, self-efficacy and external expectations. *Educational Studies* 40 (2014), 48-62.

Wasilewska, O., Rybińska, A. i Muzyk, A. (2014). *Wykorzystanie ewaluacji zewnętrznej i wewnętrznej przez szkoły*. Warszawa: Instytut Badań Edukacyjnych.

Instytut Badań Edukacyjnych

Głównym zadaniem Instytutu jest prowadzenie badań, analiz i prac przydatnych w rozwoju polityki i praktyki edukacyjnej.

Instytut zatrudnia badaczy zajmujących się edukacją – pedagogów, socjologów, psychologów, ekonomistów, politologów i przedstawicieli innych dyscyplin naukowych – wybitnych specjalistów w swoich dziedzinach, o różnorakich doświadczeniach zawodowych, które obejmują, oprócz badań naukowych, także pracę dydaktyczną, doświadczenie w administracji publicznej czy działalność w organizacjach pozarządowych.

Instytut w Polsce uczestniczy w realizacji międzynarodowych projektów badawczych, w tym *PIAAC, PISA, TALIS, ESLC, SHARE, TIMSS i PIRLS* oraz projektów systemowych współfinansowanych przez Unię Europejską ze środków Europejskiego Funduszu Społecznego.

Instytut Badań Edukacyjnych

ul. Górczewska 8, 01-180 Warszawa | tel. +48 22 241 71 00

ibe@ibe.edu.pl | **www.ibe.edu.pl**

Raport współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.