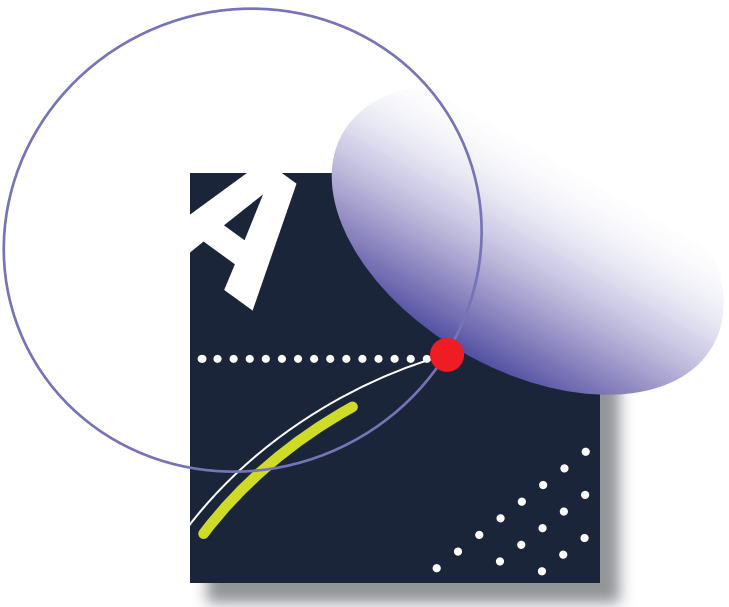
BADANIE EDUKACJI

Roman Dolata, Maciej Jakubowski, Artur Pokropek

Polska oświata w międzynarodowych badaniach umiejętności uczniów PISA OECD

Wyniki, trendy, kontekst i porównywalność



WYDAWNICTWA UNIWERSYTETU WARSZAWSKIEGO
WUW

Polska oświata w międzynarodowych badaniach umiejętności uczniów PISA OECD

Roman Dolata, Maciej Jakubowski, Artur Pokropek

Polska oświata w międzynarodowych badaniach umiejętności uczniów PISA OECD

Wyniki, trendy, kontekst i porównywalność

WUW

Warszawa 2013

Recenzent
dr hab. Roman Konarski, prof. UG

Redaktor prowadzący
Ewa Wyszyńska

Redakcja
Elwira Wyszyńska

Redakcja techniczna
Zofia Kosińska

Korekta
Bożena Gorlewska

Projekt okładki i stron tytułowych
Katarzyna A. Jarnuszkiewicz

Skład i łamanie
Marcin Szcześniak

ISBN 978-83-235-1011-6
ISBN 978-83-235-2023-8 PDF

Publikacja dofinansowana przez Wydział Pedagogiczny UW

Teksty składające się na książkę powstały w ramach projektu badawczego „Analiza porównawcza wyników międzynarodowych badań umiejętności uczniów PISA w oparciu o semiparametryczne metody dopasowania oraz hierarchiczne modele liniowe", NN114 446336, afiliowanego przy CASE – Centrum Analiz Społeczno-Ekonomicznych, Fundacja Naukowa.

Wydawnictwa Uniwersytetu Warszawskiego
00-497 Warszawa, ul. Nowy Świat 4
www.wuw.pl; e-mail: wuw@uw.edu.pl
Dział Handlowy WUW: tel. +48 22 55-31-333; e-mail: dz.handlowy@uw.edu.pl
Księgarnia internetowa: www.wuw.pl/ksiegarnia

Wydanie 1

Spis treści

Wprowadzenie 7

ROZDZIAŁ 1. Wyniki polskich piętnastolatków w perspektywie porównawczej 11

1.1. Problemy skalowania 11
1.2. Korekta wyników uwzględniająca status społeczno-ekonomiczny rodziny ucznia 17
1.3. Wyniki PISA 2009 19
1.4. Zmiany w poziomie umiejętności w latach 2000–2009 32
1.5. Trendy w zróżnicowaniu poziomu umiejętności 44

ROZDZIAŁ 2. Osiągnięcia uczniów w szkole podstawowej i ich dalszy rozwój – między PIRLS a PISA 51

2.1. Badanie umiejętności czytania w PIRLS i PISA 51
2.2. Miary międzynarodowej wartości dodanej na podstawie PIRLS i PISA 55
2.3. Problemy z porównywalnością testów 56
2.4. Szacowanie błędów ekwiwalentności konstruktów oraz modelu skalowania 58
2.5. Łączenie oszacowanych błędów 61
2.6. Metody korekty parametrów 62
2.7. Wyniki analiz rozwoju umiejętności czytania 63
Załącznik 75

ROZDZIAŁ 3. Przyrost poziomu umiejętności mierzonych w PISA w kolejnych latach nauki szkolnej 82

3.1. Wyniki PISA 2000 i 2006 z podziałem na typy szkół 83
3.2. Efektywność nauczania w różnych typach szkół ponadgimnazjalnych – model wartości dodanej 87

ROZDZIAŁ 4. Rodzinne uwarunkowania poziomu umiejętności: pieniądze czy książki? ... 98

4.1. Dane wykorzystane w analizie uwarunkowań rodzinnych ... 100
4.2. Metoda analizy uwarunkowań rodzinnych ... 103
4.3. Uwarunkowania rodzinne na poziomie gimnazjum i szkoły ponadgimnazjalnej ... 105
4.4. Analiza porównawcza uwarunkowań rodzinnych ... 114

ROZDZIAŁ 5. Motywacja do uczenia się a umiejętności z zakresu przedmiotów przyrodniczych ... 120

5.1. Główne podejścia teoretyczne do badania motywacji ... 120
5.2. Motywacja do nauki przedmiotów przyrodniczych a wyniki testu PISA 2006 – analiza porównawcza ... 122

Podsumowanie ... 128

ANEKS. Wyniki egzaminacyjne na skali PISA 2006 ... 134

A.1. Wyrażanie wyników egzaminacyjnych na skalach PISA ... 134
A.2. Egzamin gimnazjalny a test PISA ... 134
A.3. Wiązanie egzaminu gimnazjalnego z PISA ... 138
A.4. Wyniki egzaminacyjne na skali PISA w województwach ... 140
A.5. Wyniki egzaminacyjne na skali PISA w powiatach ... 142

Literatura ... 157

Spis tabel ... 160

Spis wykresów ... 162

Spis rysunków ... 165

Wprowadzenie

Program Międzynarodowej Oceny Umiejętności Uczniów (*Programme for International Student Assessment* – PISA) to międzynarodowe badanie umiejętności i wiadomości uczniów realizowane w Polsce od samego początku jego istnienia. Badanie to zarządzane jest przez Organizację Współpracy Gospodarczej i Rozwoju (OECD), zrzeszającą 34 najwyżej rozwinięte demokratyczne państwa świata, w tym Polskę. PISA realizowana jest jednak w znacznie większej liczbie krajów. Chcą one porównać umiejętności własnych uczniów z uczniami z krajów OECD. Badanie PISA prowadzone jest w każdym państwie na próbie reprezentatywnej dla populacji piętnastolatków, na podstawie tych samych zasad oraz tych samych testów wiadomości i umiejętności uczniów. W 2009 roku testy PISA rozwiązało niemal pół miliona uczniów, których wyniki są reprezentatywne dla populacji niemal 30 milionów piętnastolatków na całym świecie. Łącznie kraje uczestniczące w PISA wytwarzają ponad 80% światowego PKB. Co prawda oficjalne wyniki PISA 2009 nie obejmują całości populacji uczniów Chin oraz Indii, ale kraje te także uczestniczyły w badaniu, przedstawiając wyniki o ograniczonej reprezentatywności[1]. Można jednak uznać, że ostatnia edycja badania PISA z 2009 roku objęła swoim zasięgiem zdecydowaną większość krajów mających wpływ na rozwój światowej gospodarki.

Projekt PISA jest realizowany w cyklu trzyletnim. Pierwsze badanie miało miejsce w 2000 roku. W każdym z cykli oceniane są umiejętności w trzech obszarach: czytanie, matematyka i przedmioty przyrodnicze (angielski termin *science* obejmuje fizykę, chemię, biologię i geografię). W każdym cyklu jeden z obszarów ma rangę obszaru głównego i jest badany

[1] Szanghaj oraz Hongkong uczestniczyły w PISA 2009 oficjalnie. Dziesięć prowincji Chin uczestniczyło w programie pilotażowym, jednak wyniki nie zostały dołączone do raportu. Ukazały się one jedynie w chińskiej prasie. Pokazano, że wszystkie prowincje uzyskały wyniki powyżej średniej OECD, jednak prowincja, w której leży Szanghaj, miała wyniki najwyższe. Wyniki dla Indii zostały opublikowane w ramach prezentacji wyników PISA 2009+ (Walker, 2011).

znacznie obszerniej. W 2000 roku szczegółowej ocenie poddano czytanie, w 2003 matematykę, w 2006 nauki ścisłe, a w 2009 ponownie czytanie. W praktyce testowania oznacza to, że główna dziedzina oceniana jest na podstawie ponad 100 zadań testowych dotyczących wielu aspektów danej umiejętności, pozostałe zaś za pomocą znacznie mniej obszernych testów.

Wyniki badania PISA są podawane na trzech podstawowych skalach: czytania, matematyki i przedmiotów przyrodniczych, ale dla głównej dziedziny także na podskalach. Przykładowo, w 2006 roku, gdy badano przede wszystkim przedmioty przyrodnicze, podano wynik nie tylko dla całego obszaru, lecz także dla trzech podskal: umiejętności rozpoznawania zagadnień naukowych, wyjaśniania zjawisk przyrodniczych w sposób naukowy, a także interpretacji i wykorzystywania wyników i dowodów naukowych. Wyniki z zakresu przedmiotów przyrodniczych przedstawiono w podziale na obszary wiedzy: układy nieożywione, układy ożywione oraz Ziemia i kosmos.

W edycji PISA 2009 w najszerszym zakresie ponownie było badane czytanie, a w edycji PISA 2012 najdokładniejszy pomiar dotyczył matematyki. Oprócz pomiaru wiedzy i umiejętności uczniów badanie PISA gromadzi także bogate informacje dotyczące rodzinnych i szkolnych kontekstów procesów nauczania – uczenia się.

Niemal wszystkie zebrane dane są publikowane w ogólnie dostępnych bazach danych, co umożliwia badaczom z całego świata prowadzenie niezależnych analiz. Bazy te stanowią niezwykle bogaty materiał do badań, niestety, rzadko wykorzystywany w naszym kraju. Niewiele powstało prac badawczych analizujących wyniki i kontekst osiągnięć polskich uczniów. Oczywiście raporty krajowe z każdej edycji badania PISA publikowane przez polski zespół PISA (por. Federowicz, 2007), stanowią dobry materiał opisowy prezentujący uzyskane w Polsce wyniki, ale brak szczegółowych, krytycznych analiz uzyskanych efektów.

Wyniki badania PISA są ciekawe głównie w perspektywie porównawczej, ale fakt, że mamy już za sobą cztery edycje tego badania, sprawia, że równie interesująca jest perspektywa analiz trendów czasowych w obrębie jednego systemu oświaty. Dodatkowo wartość danych PISA wzrasta dzięki zrealizowaniu w Polsce w 2006 roku tzw. opcji narodowej, która rozszerzyła pomiary PISA na I i II klasę szkół ponadgimnazjalnych. W książce zajmujemy się zarówno porównaniami międzynarodowymi, jak i osiągnięciami polskich uczniów na przestrzeni kilku edycji PISA i kilku lat nauki szkolnej. Wnikliwe badanie trendów czasowych zostało zainicjowane przez Bank Światowy, zainteresowany przyczynami znaczącego wzrostu poziomu umiejętności polskich uczniów mierzonych w PISA

w zakresie umiejętności czytania. Wzrost ten został dostrzeżony w wielu krajach i wzbudził chęć poznania przyczyn tak znacznego polepszenia się wyników polskich uczniów.

PISA wypracowała własną koncepcję badania umiejętności i wiadomości uczniów, które są określane wspólnym terminem *literacy* (tłumaczone dość dowolnie jako „rozumowanie" lub też „biegłość"). Koncepcja ta zakłada, że w badaniu mierzyć się będzie przyswojenie wiadomości i opanowanie umiejętności niezbędnych uczniom w życiu dorosłym, na rynku pracy i do tego, aby w pełni swobodnie funkcjonowali w społeczeństwie. Współcześnie we wszystkich międzynarodowych badaniach wiadomości i umiejętności, takich jak PIRLS czy TIMSS, metody pomiaru są podobne. To, co wyróżnia PISA, to właśnie definiowanie mierzonych umiejętności w perspektywie szerszej niż programy szkolne, w pewnym oderwaniu od tego, czego naucza się w szkole (programy szkolne stanowią główną wykładnię pomiaru dla PIRLS oraz TIMSS). Sam sposób definiowania mierzonych umiejętności to długotrwały proces konsultacji międzynarodowych ekspertów dokumentowany w osobnych publikacjach (tzw. *PISA Framework*). Eksperci programu PISA twierdzą, że skonstruowane przez nich podejście daje podstawę do oceny umiejętności przydatnych w dorosłym życiu, u którego progu stoją piętnastolatki.

Dane PISA można analizować w dwóch, równie ważnych dla każdego systemu edukacyjnego wymiarach. Pierwszy z nich dotyczy poziomu wiadomości i umiejętności uczniów. Mierzy się go przede wszystkim średnim wynikiem uczniów w danym kraju, ale i procentem uczniów, którzy osiągnęli wyróżniony poziom umiejętności (na przykład w PISA 2006 zdefiniowano sześć poziomów umiejętności w przedmiotach przyrodniczych: od podstawowego do zaawansowanego). Drugi wymiar dotyczy nierówności edukacyjnych. Tutaj mierzone jest zróżnicowanie wyników, podawane przede wszystkim jako odchylenie standardowe ogółu wyników uczniów, ale i jako procent wariancji wyników, wyjaśniany przez przynależność do szkoły lub też moc zależności między osiągnięciami a statusem społeczno-ekonomicznym rodziny ucznia (im silniejsza, tym większe nierówności ze względu na pochodzenie społeczne). Dalej przedstawiamy te wyniki dla Polski, opierając się na oficjalnych raportach oraz własnych analizach baz danych PISA.

Innym badaniem międzynarodowym obecnym w Polsce i wykorzystywanym w tym opracowaniu jest PIRLS (*Progress in International Reading Literacy Study*). Stawia ono sobie za cel pomiar biegłości w czytaniu wśród dzieci mających za sobą czwarty rok nauki. Pomiary PIRLS odbywają się w cyklu pięcioletnim, w roku 2006 przeprowadzono je w 40 krajach.

Książka ta ma na celu przedstawienie wyników rozszerzonych, krytycznych analiz wybranych danych z międzynarodowych badań umiejętności uczniów, ze szczególnym uwypukleniem wyników polskich. Analizy te opierają się na zaawansowanych metodach statystycznych, jednak ich opis staramy się ograniczyć do niezbędnego minimum, skupiając się na omówieniu głównych rezultatów i ich interpretacji. Szczegółowe informacje dotyczące metodologii analiz są dostępne w artykułach, do których podajemy odnośniki.

Rozdział 1 przedstawia główne wyniki z badania PISA. Nie jest to jednak proste przypomnienie wyników z raportu PISA, prezentowane są w nim oryginalne analizy wykorzystujące od nowa wyskalowane indywidualne wyniki polskich uczniów. Nowe skale, także skorygowane o zmiany w składzie prób uczniów, a nawet zmiany w cechach rodzin uczniów, pokazują bardziej spójny obraz przemian umiejętności piętnastolatków w Polsce. W rozdziale 2 porównania dopełniają analizy wykorzystujące dane PIRLS oraz nowatorska analiza przyrostu umiejętności uczniów w zakresie czytania między końcem klasy III szkoły podstawowej (wyniki badania PIRLS) a końcem gimnazjum (wyniki badania PISA). Rozdział 3 został poświęcony analizie danych dla klas I i II szkół ponadgimnazjalnych w Polsce. Dane te zostały zebrane w ramach tzw. opcji narodowej badania PISA przeprowadzonego w 2006 roku. W rozdziale 4 przedstawiono wyniki analiz wpływu różnych aspektów pochodzenia społeczno-ekonomicznego uczniów na jego wyniki w testach PISA. Rozdział 5 został poświęcony analizie motywacji uczniów, a rozdział 6 podsumowuje główne rezultaty opisane w książce.

Oryginalnym i równocześnie niezwykle ciekawym uzupełnieniem książki jest Aneks. Zawiera on opis transpozycji wyników egzaminu gimnazjalnego na skale pomiarowe PISA 2006. Wyrażenie wyników egzaminu gimnazjalnego na skalach PISA 2006 było możliwe dzięki połączeniu dla próby polskich piętnastolatków informacji o wykonaniu zadań z testów egzaminacyjnych i testów PISA oraz wspólnemu wyskalowaniu tych danych. Przeskalowane wyniki pozwalają porównywać osiągnięcia polskich gimnazjalistów w różnych podgrupach z wynikami w państwach, które uczestniczyły w badaniu PISA 2006. W Aneksie prezentujemy średnie wyniki dla województw i powiatów. Równocześnie sama procedura wspólnego skalowania testów PISA i testów egzaminacyjnych dostarcza niezwykle ciekawych wyników, wskazujących na daleko idące podobieństwo tych narzędzi pomiarowych.

ROZDZIAŁ 1

Wyniki polskich piętnastolatków w perspektywie porównawczej

Badanie PISA obejmuje populacje piętnastolatków we wszystkich 34 krajach członkowskich OECD, a także w kilkudziesięciu krajach partnerskich. Metodologia pomiaru stosowana w tym badaniu jest bardzo bliska metodom innych znanych badań międzynarodowych: TIMSS (*Trends in International Mathematics and Science Study*) – które jest badaniem umiejętności matematycznych i umiejętności z zakresu przedmiotów przyrodniczych wśród czwarto- i ośmioklasistów, w którym niestety Polska do 2011 roku nie uczestniczyła (po raz pierwszy w 2011 roku w tzw. małym TIMSS), czy PIRLS, czyli badaniu umiejętności czytania, w którym Polska uczestniczyła w 2006 i w 2011 roku. Metody stosowane w PISA i innych międzynarodowych badaniach osiągnięć szkolnych zostały rozwinięte w ramach prac nad amerykańskim systemem ogólnonarodowych testów NAEP (*National Assessment of Educational Progress*) i dostosowane do badań międzynarodowych. Obecnie metody te stanowią najbardziej zaawansowane sposoby pomiaru i raportowania umiejętności uczniów. Stanowią one punkt odniesienia dla wszystkich innych badań edukacyjnych (Aitkin i Aitkin, 2011).

1.1. Problemy skalowania

W badaniu PISA zakłada się, że określenie wyniku ucznia nie jest ostatecznym celem. Autorom badania zależy na uzyskaniu jak najbardziej precyzyjnej oceny wyników uczniów w danej populacji (najczęściej w danym kraju) lub też w wybranych kategoriach (np. dziewczynki w Polsce). Aby osiągnąć ten cel, potrzebne są dwa elementy:

1. odpowiednia próba uczniów;
2. odpowiednia próba zadań.

Aby wypowiadać się o populacji danego kraju w sposób wiarygodny, potrzebna jest odpowiednio duża losowa próba uczniów, która zapewni,

iż wnioskowania na jej podstawie nie będą obarczone żadnymi systematycznymi błędami. Jeżeli uczniów (a precyzyjniej szkoły, bo w badaniu PISA najpierw losowane są szkoły, a następnie z każdej szkoły uczniowie) będziemy dobierać losowo, to mamy pewność, iż przy dostatecznie dużej liczbie uczniów nasza próba będzie reprezentatywna. Inaczej mówiąc, próba ta w sposób precyzyjny będzie odzwierciadlać wszystkie charakterystyki danej populacji, np. w próbie odnajdziemy taką samą proporcję uczniów z miast i ze wsi, taką samą proporcję dziewcząt i chłopców czy taki sam procent uczniów słabszych i lepszych jak w całej populacji. Losowanie uczniów nie jest sprawą błahą i wymaga szczegółowej wiedzy z zakresu teorii doboru próby. Badanie PISA charakteryzuje się w tym zakresie bardzo wysokimi standardami. Procedury losowania narzucane krajom biorącym udział w badaniu gwarantują takie losowanie uczniów, jakie pozwala uzyskać zadowalającą reprezentatywność prób krajowych, a tym samym odpowiedni stopień porównywalności między krajami. Osiągane jest to między innymi przez odpowiednio dużą liczebność próby (na ogół około 5000 osób), restrykcyjne wymagania co do realizacji badania (minimum 65% na poziomie szkół), posługiwanie się dodatkowymi zmiennymi w celu optymalizacji doboru oraz pełną transparentność procesu losowania (OECD, 2012).

Ponieważ w programie PISA ocenia się opanowanie przez uczniów szerokiego spektrum wiadomości i umiejętności, potrzebna jest odpowiednia liczba starannie przygotowanych zadań. Badanie osiągnięć uczniów opiera się na założeniu, że wykonanie danego typu zadań jest zależne od poziomu bezpośrednio nieobserwowalnej, złożonej umiejętności (cecha ukryta). Na przykład wykonanie zadań wymagających czytania różnego rodzaju tekstów zależy od wieloaspektowej umiejętności, nazywanej umownie czytaniem. Według założeń kompetencja ta odpowiada za to, jak uczniowie radzą sobie w zadaniach sprawdzających umiejętność czytania w różnych sytuacjach życiowych. Dzięki niej osoba potrafi przeczytać ulotkę leku i zastosować go we właściwy sposób, potrafi przyswoić informacje podawane w gazecie i ustosunkować się do nich itd. Takie rozumienie przedmiotu narzuca sposób jego mierzenia. Pomiar umiejętności musi być wielostronny, uwzględniać wiele możliwych sytuacji, dotyczyć czytania różnego rodzaju tekstów (naukowe, publicystyczne, użytkowe itd.) w różnych kontekstach. Pomiar musi zatem opierać się na dużej liczbie różnorodnych zadań testowych. Im więcej dobrych pytań, tym pomiar jest lepszy.

Oczywiście nie tylko liczba, lecz także jakość zadań przekłada się na wartość wyników pomiaru. Zadania w PISA są wielokrotnie próbnie

testowane. Szczególną uwagę poświęca się też ich tłumaczeniu. Sprawdza się, czy te same zadania mają podobne parametry w różnych krajach. Jeśli mimo długiego procesu przygotowywania zadań (badania ich właściwości w różnych krajach, wielokrotnego sprawdzania tłumaczeń, usuwania wątków odczytywanych różnie w zależności od kultury i historii danego kraju itd.) w badaniu zasadniczym okazuje się, że jakieś zadanie jest w jakimś kraju znacznie trudniejsze lub łatwiejsze w porównaniu do wyników w pozostałych krajach, to jest ono usuwane przed skalowaniem wyników uczniów tego kraju. Przykładowo, jeśli w procesie przygotowania zadań „przepuszczono" zadanie z czytania, przywołujące postać świętego Mikołaja, to zadanie to zapewne zostanie usunięte w krajach arabskich, gdyż najprostsze statystyki pokażą, że nawet najlepiej czytający uczniowie z tego regionu nie są w stanie go rozwiązać, podczas gdy nie sprawia ono trudności uczniom polskim, którym zarówno święty Mikołaj, jak i zima są znacznie bliższe.

Postulat dużej liczby odpowiedniej jakości zadań w zderzeniu z ograniczeniami w zakresie czasu testowania jednego ucznia jest źródłem poważnego utrudnienia. Wiadomo przecież, że uczeń nie może być testowany w nieskończoność. Testowanie w badaniu PISA trwa dwie godziny (z krótką przerwą) i wydaje się, że każda dłuższa próba testowania naraziłaby badanie na poważne problemy: uczniowie w końcu zmęczyliby się, zabrakłoby im motywacji itd. Podobnie kilkudniowe testowanie uczniów, niezależnie od tego, że byłoby bardzo kosztowne, najprawdopodobniej zniechęcałoby uczniów do uczestnictwa w badaniu. Z tych powodów opracowano taki system, że uczniowie rozwiązują różne zestawy zadań testowych, tzw. *booklets* (dosłownie „książeczki" czy „zeszyty"). Każdy uczeń dostaje więc tylko część zadań przygotowanych z danej dziedziny. Na przykład w 2009 roku w PISA wszystkie zadania zostały pogrupowane w 13 zestawów (7 z czytania ze zrozumieniem, 3 z przedmiotów przyrodniczych i 3 z matematyki). Na rozwiązanie każdego zestawu przeznaczono 30 minut. Zestawy ułożono w 13 zeszytów, z których każdy składał się z 4 zestawów zgodnie ze schematem rotacyjnym przedstawionym w tabeli 1.1. Symbole od S1 do S3 oznaczają zestawy przyrodnicze, R1 do R7 zestawy czytania ze zrozumieniem, a M1 do M3 to zestawy matematyczne. Przyporządkowanie zestawów opiera się na tzw. zrównoważonym, niekompletnym schemacie blokowym. Każdy z zestawów występuje po jednym razie na każdej z czterech pozycji w zeszycie. Wykluczona jest również możliwość znalezienia się danej pary zestawów w dwóch różnych zeszytach.

Tabela 1.1. Rozlokowanie zestawów zadań w zeszytach testowych

Zeszyt	Zestawy zadań			
1	M1	R1	R3	M3
2	R1	S1	R4	R7
3	S1	R3	M2	S3
4	R3	R4	S2	R2
5	R4	M2	R5	M1
6	R5	R6	R7	R3
7	R6	M3	S3	R4
8	R2	M1	S1	R6
9	M2	S2	R6	R1
10	S2	R5	M3	S1
11	M3	R7	R2	M2
12	R7	S3	M1	S2
13	S3	R2	R1	R5

Klasyczne oszacowanie umiejętności ucznia przez określanie liczby poprawnych odpowiedzi nie zdaje w takiej sytuacji egzaminu. Stoimy przed dwoma problemami, z którymi nie radzi sobie dobrze tzw. klasyczna teoria testów. Po pierwsze, każdy uczeń rozwiązał jedynie część z całej baterii zadań. Po drugie, wynik każdego ucznia w klasycznej teorii testów jest wyrażony na skali tzw. dyskretnej, czyli wyniki pomiaru są liczbami całkowitymi. W wielu opracowaniach (np. Davier, Gonzalez i Mislevy, 2009) wykazano, iż stosowanie klasycznych metod może powodować błędy w szacowaniu różnic między grupami w średnich wynikach (na przykład chłopców i dziewczynek, uczniów ze wsi i z miasta) czy też w szacowaniu zależności między wynikami a interesującymi nas zmiennymi (np. wynikami a pochodzeniem społecznym ucznia, czyli zmiennymi wykorzystywanymi w niemal każdym badaniu edukacyjnym).

Problemy te rozwiązuje model Rascha z tzw. *plausible values* (dosłownie model „prawdopodobnych wartości"). W modelu tym uczniom przypisywanych jest pięć prawdopodobnych wyników – wartości mierzonej umiejętności – po uwzględnieniu odpowiedzi na wszystkie zadania testowe, a także związków odpowiedzi na zadania testowe z cechami uczniów w całej populacji (Wu, Adams, Wilson i Haldane, 2007). Informacje o różnych dodatkowych charakterystykach uczniów są zbierane w badaniu ankietowym. Ankietę wypełnia uczeń po teście wiadomości i umiejętności.

Ten dość skomplikowany model statystyczny, przypisujący pięć wyników każdemu uczniowi, dobrze odzwierciedla niepewność pomiaru. Wykazano, że posługując się pięcioma *plausible values*, można lepiej odtworzyć rozkład prawdziwych umiejętności w całej populacji. Tak więc dzięki użyciu modelu Rascha z *plausible values* nie tylko średnie wyniki w całej populacji, lecz także wyniki dla podgrup, wariancja tych wyników oraz dalsze analizy odnoszące osiągnięcia uczniów do ich innych cech, programów edukacyjnych itp., ukazują wartości bliskie prawdziwym. Co więcej, model *plausible values* przewiduje wyniki dla uczniów, którzy rozwiązywali różne zadania testowe, dzięki uwzględnieniu cech tych uczniów, ale i temu, że każdy z zeszytów testowych zawiera zadania wspólne z innym zeszytami, a model statystyczny bierze pod uwagę równocześnie wszystkich uczniów i wszystkie zadania testowe[2].

Jak już wspomniano, nie jest wskazane, aby w ten sposób oceniać umiejętności pojedynczych uczniów, jednak metoda ta sprawia, że wartości dla całej populacji są liczone bardziej precyzyjnie, lepiej odzwierciedlają prawdziwe rozkłady umiejętności i wiedzy, a także pozwalają na oszacowanie rzeczywistych relacji między wynikami a cechami uczniów i szkół (Davier, Gonzalez i Mislevy, 2009).

Sposób skalowania wyników PISA jest tak opracowany, aby umożliwić jak najbardziej wiarygodne porównania międzynarodowe. Jednak wyniki badania PISA są także od ponad 10 lat intensywnie wykorzystywane do analiz wewnątrzkrajowych, w tym do analizy zmian w czasie poziomu umiejętności w poszczególnych krajach. Schemat skalowania stosowany w badaniu PISA z perspektywy tak postawionego celu nie jest schematem najlepszym, dostosowanie parametrów do danych ze wszystkich krajów powoduje bowiem, że skalowanie nie jest optymalne, jeśli patrzymy na wyniki tylko jednego kraju na przestrzeni lat.

Aby to wyjaśnić, musimy odwołać się do kilku technicznych szczegółów skalowania PISA, o których do tej pory nie wspomnieliśmy. Wyniki PISA są skalowane w kilku krokach. W pierwszym skaluje się trudność pytań (modelem Rascha) na losowo dobranej próbie 500 uczniów z każdego kraju OECD (OECD, 2009). W ten sposób wkład każdego kraju do oszacowań parametru trudności dla poszczególnych zadań testowych jest taki sam. Tak oszacowane parametry są optymalne z perspektywy międzynarodowej i stanowią najlepsze rozwiązanie dla badań porównawczych.

[2] Dokładniej – model bezpośrednio szacuje wyniki w kilku dziedzinach (por. OECD, 2012, rozdział 9).

Po wstępnym wyskalowaniu uzyskane międzynarodowe parametry aplikuje się do modelu Rascha, w którym skaluje się wyniki już dla wszystkich uczniów z wykorzystaniem opisanych *plausible values*. Aby zapewnić porównywalność skal między latami, wykorzystuje się zestaw powtarzanych w każdym badaniu zadań do oszacowania funkcji łączącej testy między edycjami. Na podstawie tej funkcji przekształca się utworzone skale z różnych edycji, by uzyskać wyniki przedstawiające trendy czasowe. Wszystko w tej procedurze jest podyktowane głównym celem badania: porównywalnością międzynarodową.

Skalowanie pytań dla wszystkich krajów jednocześnie może prowadzić do gorszego dopasowania modelu do danych w poszczególnych krajach. W skrajnych sytuacjach, gdy błędy dopasowania w modelu skalowania skumulują się, może to prowadzić do znacznych nieprawidłowości. Inaczej mówiąc, model w badaniu międzynarodowym optymalizuje dopasowanie parametrów zadań testowych dla wszystkich państw, przez co z perspektywy pojedynczego kraju rozwiązanie międzynarodowe nie musi być najlepsze. Ponadto skalowanie przy zachowaniu wymogu porównywalności międzynarodowej zmusiło twórców badania do pewnych kompromisów: we wstępnym skalowaniu używa się tylko próbki uczniów, co zmniejsza precyzję oszacowania. Aby skalowanie dla próby składającej się z kilkudziesięciu krajów było wykonalne, stosuje się model Rascha, który obok wielu zalet ma zasadniczą wadę: jego dopasowanie do danych opiera się na jednym parametrze (trudności), czyli z definicji model ten jest gorzej dopasowany niż model np. dwuparametryczny (Baker i Kim, 2004). Model dwuparametryczny jest jednak dużo bardziej wymagający obliczeniowo niż model Rascha. Kolejnym kompromisem jest sposób zrównywania wyników między latami. Zrównywanie w PISA opiera się na przekształceniu liniowym danych z kolejnych edycji badania. W tej procedurze najpierw skaluje się wyniki z poszczególnych edycji, a potem je liniowo przekształca tak, aby sprowadzić je do wspólnej skali. Bardziej efektywne zrównywanie, polegające na skalowaniu czterech edycji PISA w jednym modelu, jest praktycznie niewykonalne z powodów wielkości próby (por. OECD, 2009; Gebhardt i Adams, 2007).

Ponieważ w tej książce interesuje nas przede wszystkim Polska, postanowiliśmy podejść do trendów w inny sposób. Zmiana naszego podejścia dotyczy przede wszystkim trzech aspektów skalowania wyników. Po pierwsze, postanowiliśmy wyskalować wyniki tylko dla Polski, aby uzyskać dopasowania najlepsze z punktu widzenia naszego kraju. Po drugie, postanowiliśmy zastosować model dwuparametryczny, czyli taki,

który pozwala na lepsze dopasowanie do danych niż model Rascha (Baker i Kim, 2004). Po trzecie, zrównanie wyników w kolejnych edycjach PISA postanowiliśmy osiągnąć przez jedną łączną kalibrację wszystkich zadań, począwszy od 2000 roku (por. Kolen i Brennan, 2004; Davier i Davier, 2007). Używając tych rozwiązań, uzyskaliśmy skale lepiej dopasowane do polskich danych, ale mogą być one podstawą do analizy trendów wyłącznie dla naszego kraju.

1.2. Korekta wyników uwzględniająca status społeczno-ekonomiczny rodziny ucznia

W rozdziale tym prezentujemy wyniki skorygowane także ze względu na zmiany w czasie charakterystyk społecznych uczniów. Korekta taka ma sens zarówno przy porównaniach wyników polskich uczniów uzyskanych w 2009 roku z innymi krajami, jak i przy analizie trendów. Przy porównaniach z innymi krajami korekta wyników ze względu na status społeczno-ekonomiczny rodzin uczniów pozwala uwzględnić to, że rozkłady zmiennych statusowych w krajach biorących udział w badaniu PISA znacząco się różnią. Wykorzystując wyniki PISA do oceny efektywności krajowych systemów edukacyjnych, warto wziąć pod uwagę cechy statusu społecznego uczniów. W pewnych analizach warto także wyłączyć z porównań imigrantów, ich wyniki zależą bowiem w dużej mierze od tego, w jakim kraju się urodzili i pobierali pierwsze nauki.

Z kolei korekta trendów uwzględniająca zmiany w czasie rozkładu cech statusowych ucznia pomaga w dokonaniu porównań, na które nie będą miały wpływu przemiany społeczne, niezależne od szkół. Przykładowo, jeśli ogólnym trendem w populacji jest zdobywanie coraz wyższego wykształcenia, to po pewnym czasie szkoły będą miały niejako ułatwione zadanie, będą bowiem pracować z uczniami pochodzącymi z coraz lepiej wykształconych rodzin. Jeśli chcemy ocenić efektywność systemów edukacji, to tego rodzaju zmiany kontekstu powinniśmy wziąć pod uwagę.

Korektę wyników o zmienne kontekstowe można dokonać w dość prosty sposób, stosując odpowiednie modele statystyczne. Metody przez nas wykorzystane różnią się nieco zależnie od analizowanych danych, zasada działania jest jednak zawsze ta sama. W każdym przypadku prezentujemy wyniki oryginalne, gdzie populacje mogą dowolnie różnić się między krajami pod względem statusu społeczno-ekonomicznego rodzin uczniów. Prezentujemy też jednak wyniki skorygowane, które dzięki odpowiednim zabiegom statystycznym pokazują, jak układałyby się wyniki