Maciej Jakubowski
Artur Pokropek

BADAJĄC EGZAMINY

Podejście ilościowe w badaniach edukacyjnych

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Centralna Komisja Egzaminacyjna

Maciej Jakubowski
Artur Pokropek

BADAJĄC EGZAMINY

Podejście ilościowe w badaniach edukacyjnych

Centralna Komisja Egzaminacyjna
Warszawa 2009

Maciej Jakubowski, Artur Pokropek
Badając egzaminy. Podejście ilościowe w badaniach edukacyjnych

Redakcja: *Klemens Stróżyński*
Opracowanie graficzne i łamanie: *Anna Michalak*
Korekta: *Magdalena Pokropek*

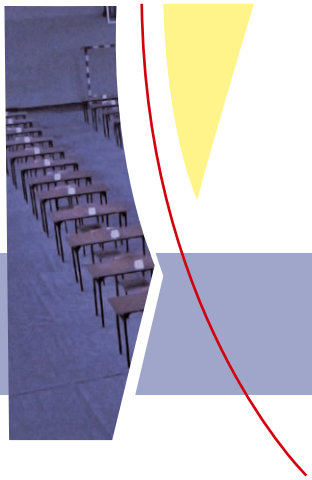
Publikacja współfinansowana przez Unię Europejską
w ramach Europejskiego Funduszu Społecznego.
Publikacja wydana w ramach Programu Operacyjnego *Kapitał Ludzki*,
projekt *Specjalistyczne publikacje dotyczące egzaminów zewnętrznych*.

Wydawca: Centralna Komisja Egzaminacyjna
Warszawa 2009

Druk: Centrum Obsługi Kancelarii Prezesa Rady Ministrów,
Wydział Wydawnictw i Poligrafii, Warszawa

978-83-7400-265-3

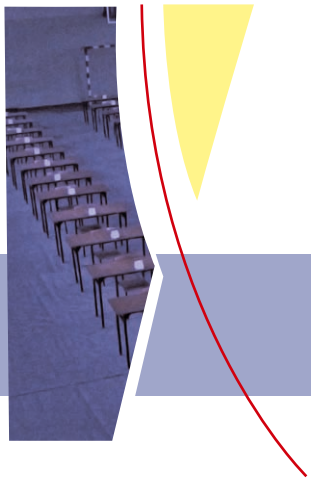
PUBLIKACJA JEST DYSTRYBUOWANA BEZPŁATNIE



Spis treści

Wstęp 5
1. Zagadnienia pomiaru 8
1.1. Problemy związane z pomiarem 9
Trafność 9
Reprezentatywność 10
Rzetelność 11
Zagadnienie skali 14
Podsumowanie – problemy pomiaru 19
1.2. Klasyczna teoria testu 20
Wynik prawdziwy i błąd pomiaru 21
Błędy systematyczne i niesystematyczne (losowe) 25
Rzetelność 26
Klasyczna teoria testu jako pomoc w analizie testów i zadań 28
Alfa Cronbacha – dolna granica rzetelności 29
Łatwość 32
Moc różnicująca 34
Analiza graficzna 38
1.3. Teoria odpowiedzi na zadania testowe (item response theory, IRT) 41
Podstawowe definicje 41
Poziom umiejętności uczniów 1PL (model Rasha) 46
Model dwuparametryczny 49
Dopasowanie krzywej charakterystycznej do modelu 52
Krzywa informacyjna zadania 53
Dopasowanie nieparametryczne 55
Analiza testu jako całości 56
Krzywa charakterystyczna testu 56
Krzywa informacyjna testu 58
Strategia budowania testów 59

Pułapki teorii odpowiedzi na zadanie testowe 62
1.4. Cele a metody pomiaru: od egzaminów zewnętrznych po badania międzynarodowe 65
Zależność między celami a metodami pomiaru umiejętności i wiedzy ucznia 65
Pomiar umiejętności i wiedzy uczniów w badaniach międzynarodowych 68
2. Wnioskowanie o populacji 74
2.1. Próby losowe – kiedy badanie można nazwać reprezentatywnym? 76
Błąd ekologiczny i atomistyczny 82
Paradoks Simpsona 83
Losowość doboru próby i reprezentatywność danych 85
2.2. Próby nielosowe: dobór celowy i dane zastane 89
2.3. Badania reprezentatywne a określanie związków przyczynowo-skutkowych 92
2.4. Eksperymenty 99
2.5. Naturalne eksperymenty 104
2.6. Badania z użyciem danych z całej populacji 105
2.7. Badania na małych grupach 108
3. Ewaluacja metodami edukacyjnej wartości dodanej 110
3.1. Modele EWD – wprowadzenie 113
3.2. Przykłady wewnątrzszkolnych lub wewnątrzgminnych analiz EWD 123
Podsumowanie 127
4. Przegląd wyników badań w Polsce 129
4.1. Badania wyników egzaminów całej populacji uczniów 130
System edukacyjny „z lotu ptaka” 130
Badania efektywności szkół 135
4.2. Badania wyników egzaminów zewnętrznych na próbach reprezentatywnych 142
Badania na mniejszych populacjach 145
4.3. Porównawcze badania międzynarodowe 147
Badania PISA 147
Badanie nad edukacją obywatelską (Civic Education Study) 153
Studium przypadku – eksperyment kwidzyński 161
Badania jakościowe na małych, nielosowych próbach 162
5. Bazy danych 166
5.1. Bazy międzynarodowe 167
Badanie PISA 167
Badanie CivEd 175
Badania PIRLS i TIMSS 175
5.2. Polskie bazy danych 181
Baza Pentor-CKE 181
Baza EWD 184
Baza „Trzecioklasiści” 187
6. Słowniczek 189



Wstęp

Niniejsza publikacja ma dwojaki charakter. Po pierwsze, stanowi przewodnik po stosowanych w badaniach edukacyjnych analizach statystycznych. Po drugie, dokonuje przeglądu badań opierających się nie tylko na wynikach egzaminów zewnętrznych, lecz także na rezultatach badań międzynarodowych zrealizowanych z udziałem naszego kraju. Jej głównymi odbiorcami mają być beneficjenci badań, którzy chcą korzystać z nich w sposób bardziej świadomy i krytyczny. Pozycja ta może być też przydatna dla osób, które planują lub nadzorują realizację badań edukacyjnych.

Z pewnością książka nie stanowi samodzielnego podręcznika, choć może być uzupełnieniem podręczników analiz statystycznych, gdzie rzadko kiedy uwzględniana jest specyfika badań edukacyjnych. Nasze przykłady dotyczą zawsze rzeczywistych badań na uczniach lub szkołach, a metody opisano, opierając się na własnych doświadczeniach z analiz edukacyjnych prowadzonych na polskich danych. Mamy nadzieję, że będzie to cenna wiedza, przydatna czytelnikom zainteresowanym takimi badaniami, w szczególności badaniami związanymi z polskimi egzaminami zewnętrznymi.

Książka ta stanowi więc wprowadzenie do analiz statystycznych wykorzystywanych w badaniach edukacyjnych. Oczywiście mowa tu o badaniach mających charakter ilościowy, wykorzystujących informacje wyrażone w postaci liczb. Taki charakter mają wyniki egzaminów zewnętrznych. Podręczników badań ilościowych jest sporo, jednak specyfiką tej publikacji jest opis zróżnicowanych zagadnień, które stanowią domenę różnych dyscyplin. Na przykład pomiar umiejętności i skalowanie jego wyników są często omawiane w podręcznikach psychologii, ale już kwestia badań reprezentatywnych to domena socjologów, a wnikliwe opisy analiz metodami regresji liniowej można znaleźć głównie w podręcznikach ekonometrii.

Wszystkie te zagadnienia są jednak kluczowe dla badań edukacyjnych, które z natury mają charakter interdyscyplinarny. Powoduje to, że zrozumienie wyników i metodologii badań edukacyjnych jest często trudne i nawet eksperci mogą być mało pomocni, bowiem ich wiedza ograniczona jest do jednej dyscypliny. Mamy nadzieję, że interdyscyplinarny charakter przedstawionego materiału, choć na pewno stawia dodatkowe wymagania Czytelnikowi, to jednak będzie pomocny nie tylko przy interpretacji rezultatów analiz, lecz także przy projektowaniu czy nadzorowaniu własnych badań.

Niewielka objętość tej książki sprawia, że opis niektórych zagadnień jest zdawkowy i z pewnością niepełny. Zachęcamy Czytelnika do podążania tropem przywołanych terminów i odwołań do literatury. Wiedza dotycząca badań ilościowych w edukacji jest naprawdę rozległa i niemożliwe jest jej zawarcie, nawet w postaci skrótowej, w jednej książce. Niech pozycja ta będzie zachętą do dalszego poszukiwania wiedzy, sięgania do wskazanych pozycji lub innych podręczników czy opisów badań.

Książka rozpoczyna się rozdziałem o pomiarze umiejętności i wiedzy uczniów. Opisane w nim kwestie mają charakter podstawowy, bowiem bez poprawnego i precyzyjnego pomiaru żadne badanie edukacyjne nie przyniesie spodziewanych korzyści. W rozdziale tym omówiono pojęcia i metody związane z klasyczną teorią testu – które można znaleźć w wielu pozycjach opublikowanych w Polsce – i z teorią odpowiedzi na zadanie testowe (IRT – *Item Response Theory*), która choć dominuje we współczesnych badaniach edukacyjnych, to nie doczekała się jeszcze w Polsce odpowiedniej publikacji. Mamy nadzieję, że jej opis przedstawiony w tej publikacji pomoże Czytelnikowi zrozumieć podstawowe koncepcje tej teorii i dostrzec jej intuicyjność i prostotę w podstawowych zastosowaniach.

Kolejny rozdział porusza kwestie dotyczące reprezentatywności badań edukacyjnych oraz możliwości dociekania na ich podstawie zależności przyczynowo-skutkowych. Są to kwestie podstawowe dla interpretacji każdego badania i choć mogą wydawać się dość abstrakcyjne, to w istocie bez porządnego zrozumienia koncepcji przedstawionych w tym rozdziale trudno jest określić, kiedy możemy ufać wynikom badań edukacyjnych, a kiedy ich rezultaty są w najlepszym wypadku dość przekonującą hipotezą wymagającą kolejnych badań.

Trzeci rozdział wprowadza metodę edukacyjnej wartości dodanej. Jest to metoda o tyle istotna, że stanowi według nas najlepszy sposób wykorzystania informacji płynącej z wyników polskich egzaminów zewnętrznych. Poświęcamy jej jednak stosunkowo niewiele miejsca, bowiem EWD doczekała się już wielu ogólnodostępnych publikacji opisujących dokładnie jej mechanizm i dających praktyczne przykłady zastosowań.

Czwarty rozdział ma charakter krytycznego przeglądu wyników badań w Polsce. Należy podkreślić, że w przeglądzie tym interesują nas bardziej zagadnienia metodologiczne, a nie sam temat badań. Chcieliśmy w ten sposób pokazać, jak

można wykorzystać wiedzę przekazaną w trzech pierwszych rozdziałach, aby samodzielnie ocenić jakość interesującego nas badania, a przez to poprawnie interpretować jego wyniki. Badania edukacyjne jak żadne inne wymagają krytycznego spojrzenia. Rzadko kiedy są w pełni przekonujące metodologicznie i dają rezultaty, którymi możemy się posłużyć bez żadnego „ale". Naszym krytycznym spojrzeniem nie chcemy jednak zniechęcać do realizacji czy wykorzystania badań ilościowych, których sami jesteśmy entuzjastami. Chcemy jedynie pokazać, że badania edukacyjne nie są prostym przedsięwzięciem, a wątpliwości i wyzwania metodologiczne są im nieodłączne. Z drugiej strony możemy dzięki nim wiele zyskać, co – mamy nadzieję – pokazuje ten rozdział.

Ostatni rozdział opisuje publicznie dostępne darmowe bazy, które zawierają dane o polskich uczniach. Omawiamy tu zarówno bazy danych z badań międzynarodowych, jak i zawierające wyniki egzaminów zewnętrznych, powstałe jako efekt prac realizowanych w Centralnej Komisji Egzaminacyjnej.

Książkę zamyka słowniczek podstawowych pojęć, który może być przydatny przy zrozumieniu tekstu posługującego się terminami z tak wielu dyscyplin. Prosimy Czytelnika, aby sięgał do niego jak najczęściej, nie zniechęcając się obcością wielu terminów, które z pewnością wkrótce staną się łatwe i zrozumiałe.

Mamy nadzieję, że niniejsza książka, choć z pewnością dość wymagająca przy pierwszym czytaniu, stanie się dla Czytelnika pozycją, do której będzie często sięgać, starając się zrozumieć i krytycznie spojrzeć na wyniki badań czy też planując własne przedsięwzięcia. Jesteśmy świadomi, że dla wielu czytelników jest to pozycja niełatwa, jednak zrozumienie współczesnych badań edukacyjnych wymaga sporego nakładu własnej pracy, bez którego jesteśmy tylko biernymi odbiorcami zmuszonymi do zbyt dużego kredytu zaufania dla badaczy.

Znaczną poprawę przystępności i klarowności wypowiedzi w porównaniu z pierwszą wersją tekstu zawdzięczamy Redaktorowi książki, któremu jesteśmy niezwykle wdzięczni za owocną współpracę. Wyrażamy także wdzięczność dr. hab. Romanowi Dolacie, koordynującemu krajowe prace nad edukacyjną wartością dodaną, za wsparcie merytoryczne oraz wiele bardzo przydatnych i cennych koleżeńskich rad. Przede wszystkim jednak dziękujemy naszym Rodzinom, bez których wsparcia i wyrozumiałości publikacja ta nigdy by nie powstała.

1. Zagadnienia pomiaru

Dla badań edukacyjnych pomiar wiadomości i umiejętności uczniów jest zagadnieniem podstawowym. Większość badań nad edukacją obiera wyniki uczniów jako zmienną objaśnianą. Czy są to porównawcze badania międzynarodowe, czy badania poziomu umiejętności uczniów w szkołach, dla wszystkich centrum analizy stanowią wyniki uczniów. Wykorzystując pomiar wiadomości i umiejętności uczniów, można podejmować próby wskazania, jak wyniki indywidualne powiązane są z interesującymi nas zmiennymi, np. w jakim stopniu osiągnięcia uczniów związane są ze statusem społeczno-ekonomicznym ich rodzin, jakością wykształcenia nauczycieli czy też od wprowadzeniem nowego programu nauczania. Do tych celów stosuje się często bardzo złożone metody statystyczne, zapominając jednak o rzeczy fundamentalnej – o jakości pomiaru. Nawet najbardziej wyrafinowane metody statystyczne i najbardziej przemyślane plany analizy stają się bezużyteczne, jeżeli pomiar najważniejszej zmiennej okazuje się pomiarem słabym. Dlatego też sporo miejsca poświęcimy problematyce pomiaru. Omawiamy metody pomiaru i narzędzia statystyczne, które są przydatne do kontrolowania jego jakości, a przez to do tworzenia dobrych narzędzi pomiarowych, czyli dobrych testów. Wiadomości te przydadzą się także osobom, które same nie prowadzą badań edukacyjnych, lecz chcą nabyć kompetencje, które pozwolą im na trafną ocenę przedstawianych im narzędzi pomiarowych lub wyników badań. Informacje te mogą przydać się każdemu praktykowi chcącemu kontrolować jakość wykorzystywanych przez siebie testów niezależnie od tego, czy są one tworzone samodzielnie, czy też uzyskane z innych źródeł. Poniższy rozdział wprowadza do statystycznej analizy wyników pomiaru i przekazuje podstawy dwóch koncepcji teoretycznych związanych z problematyką pomiaru – klasycznej teorii testu i tzw. *item response theory* (IRT lub w dosłownym tłumaczeniu – „teorii odpowiedzi na pytanie").

1.1. Problemy związane z pomiarem

Problematyka pomiaru wiadomości i umiejętności jest złożona i nie dotyczy wyłącznie zagadnień natury statystycznej. Jest ona szeroko opisywana w literaturze metodologicznej nauk społecznych, literaturze statystycznej, psychologicznej czy psychometrycznej. Podejścia w tych dziedzinach różnią się od siebie przede wszystkim rozłożeniem akcentów na różne elementy tej problematyki, poziomem sformalizowania oraz przedmiotem dociekań. Rozdział ten systematyzuje wiedzę na temat pomiaru i osadza ją we właściwym kontekście badań edukacyjnych. Czytelnicy, którzy po przeczytaniu tego rozdziału będą chcieli poszerzyć swoją wiedzę na ten temat i zapoznać się z podejściami w różnych dziedzinach nauk społecznych, mogą sięgnąć do literatury podanej na końcu tego rozdziału.

Można powiedzieć, że przed pomiarem edukacyjnym stoją cztery podstawowe problemy:

- problem trafności
- problem reprezentatywności
- problem rzetelności
- problem skalowania

Rozwiązanie problemów trafności i reprezentatywności należy przede wszystkim do konstruktorów testów i zadań, specjalistów przedmiotowych. Poświęcamy im kilka poniższych akapitów, jednak to dwoma ostatnimi problemami, mającymi charakter przede wszystkim statystyczny, będziemy zajmować się w tym rozdziale.

Trafność

Problematyka trafności pomiaru dotyczy zagadnień związanych z tym, czy dane wskaźniki, np. testy osiągnięć szkolnych, mierzą to, co zakłada się, iż mierzyć powinny. Problem trafności można dobrze opisać na prostym przykładzie. Zastanówmy się, czy średni wynik egzaminu gimnazjalnego w danej szkole może być trafnym wskaźnikiem jakości pracy zespołu nauczycielskiego. Niekoniecznie, możemy bowiem wyobrazić sobie gimnazjum, do którego trafili wyłącznie bardzo zdolni uczniowie, a nauczyciele nie pracowali z nimi zbyt dobrze. Łatwo przewidzieć wynik takiej sytuacji – pod koniec cyklu kształcenia wyniki uczniów były przeciętne, mimo iż mogły być znacznie wyższe. Możemy wyobrazić sobie również drugie gimnazjum, do którego trafili uczniowie o bardzo niskich uzdolnieniach, lecz systematyczna, wytrwała praca nauczycieli spowodowała, iż pod koniec cyklu kształcenia uczniowie w tej szkole uzyskali wyniki przeciętne, lecz nie bardzo słabe, jak można by oczekiwać. Patrząc jedynie na średnie wyniki egzaminów, musielibyśmy powiedzieć, iż jakość nauczania w dwóch rozpatrywanych placówkach jest podobna, co oczywiście nie

byłoby trafną diagnozą. Wyniki egzaminów mówią nam o poziomie wiedzy, jaki osiągnęli uczniowie pod koniec danego cyklu kształcenia, nie zaś o przyroście tej wiedzy, na którą, w co wierzymy, niebagatelny wpływ ma praca nauczycieli.

W kontekście pomiaru edukacyjnego najbardziej pomocne wydaje się psychologiczne pojmowanie trafności. Definicja psychologiczna składa się z dwóch części. O trafnym pomiarze mówi się wtedy, kiedy narzędzie pomiarowe wywołuje żądane procesy myślowe w głowie ucznia oraz daje możliwość zapisania dowodów, iż takie procesy myślowe zaistniały. Definicja ta dotyczy tak poszczególnych zadań, jak całego narzędzia pomiarowego. Pierwsza część mówi o tym, iż jeżeli chcemy mierzyć na przykład umiejętności matematyczne, to nasze narzędzie pomiarowe musi wywołać w uczniowskich głowach operacje matematyczne, a nawet bardziej precyzyjnie: jeżeli za pomocą jakiegoś zadania chcemy mierzyć np. znajomość i umiejętność wykorzystywania pojęcia procentu, to musimy wywołać takie procesy myślowe, które w rozwiązaniu zadania będą wykorzystywały owo pojęcie procentu. Zadanie będzie nietrafne, jeżeli do jego rozwiązania wystarczy umiejętność, na przykład, dodawania. Druga część definicji mówi o tym, iż narzędzie pomiarowe musi dawać empiryczne dowody na zaistnienie takiego procesu myślowego. Czyli rozwiązanie zadania musi jednoznacznie potwierdzać to, iż uczeń wykorzystał pojęcie procentu, a nie jedynie pojęcie sumy czy różnicy. Problem trafności testów jest niebagatelny, od niego zależeć będzie sensowność dalszych poczynań.

Reprezentatywność

Problem reprezentatywności, jeśli chodzi o testy osiągnięć szkolnych, sprowadza się w zasadzie do jednej kwestii: O co pytać w teście? Reprezentatywność jest zatem bardzo ściśle związana z problemem trafności. Każdy test ma ograniczoną długość, w związku z tym można zastosować jedynie ograniczoną ilość zadań. Jeżeli mówimy o testach osiągnięć szkolnych, czyli testach często poruszających szeroki wachlarz zagadnień, problem selekcji zadań jest niezwykle drażliwy. Co wybrać do testowania, a jakie zagadnienia można opuścić? Jaki okres historyczny jest ważniejszy? Jakie dzieło literackie można pominąć? Zawsze trzeba dokonywać wyboru. Nie istnieje test, który w pełni reprezentuje wiedzę, którą powinni dysponować uczniowie. Wybór zakresu wiadomości i umiejętności poddanych testowaniu powinien być kierowany z góry przyjętymi wymogami (np. podstawą programową nauczania na danym poziomie edukacji), a dobór zadań powinien następować wedle ściśle określonych, jawnych kryteriów. Trzeba jednak pamiętać o tym, że psychometryczna jakość pomiaru jest zagadnieniem odrębnym. Test może świetnie mierzyć poziom wiadomości i umiejętności, choć może nie być reprezentatywny. Podobnie test może być uznany za reprezentatywny, jednak jego własności psychometryczne powodują, że jest słabym narzędziem po-

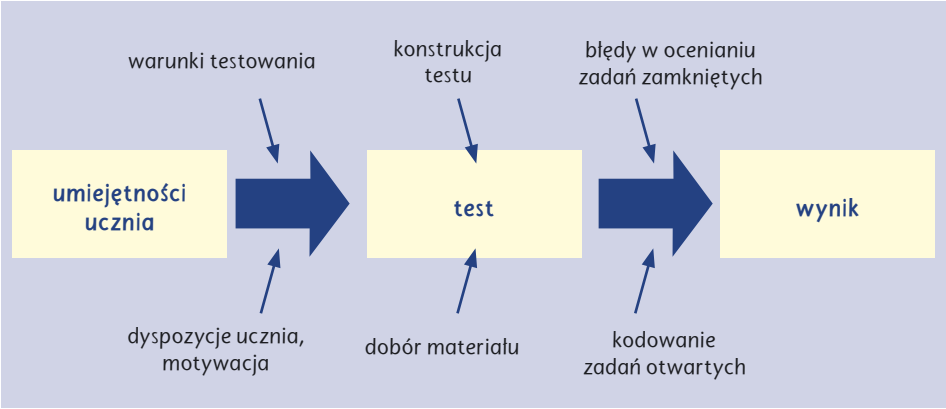
miarowym. Niestety, w gorączce dyskusji nad testami, także nad systemem egzaminów zewnętrznych, te dwie kwestie są często mylone. Ich rozróżnienie jest bardzo pomocne, bowiem pomaga określić, jakie cechy testu należy poprawić i pomaga dostrzec, że wysoka reprezentatywność testu nie gwarantuje, że jest on dobrym narzędziem pomiarowym.

Rzetelność

Podstawowym zadaniem pomiaru wiadomości i umiejętności jest określenie wiedzy ucznia w jakiejś dziedzinie. W edukacji często próbuje się określić poziom tych umiejętności na podstawie wyniku testu. Zastanówmy się jednak, co oprócz wiadomości i umiejętności ucznia może wpływać na wynik uzyskany w danym teście. Jednym z głównych czynników jest oczywiście dobór zadań (problem reprezentatywności). Uczeń może trafić na zadania reprezentujące wiedzę, którą powtarzał niedawno, ale może być też odwrotnie, może powtarzał dany wycinek materiału bardzo dawno temu albo trafił na zadania związane z materiałem, który był omawiany w szkole podczas jego choroby.

Testy edukacyjne, co podkreślaliśmy wcześniej, nie mogą mierzyć wszystkich wymaganych zagadnień jednocześnie. To, co robimy podczas pomiaru dokonywanego na podstawie testów, jest próbą określenia poziomu ogólnych umiejętności ucznia z danego przedmiotu na podstawie kilkunastu zadań wybranych według określonych zasad. Nie jest to zabieg skazany na klęskę. Jest bowiem wysoce prawdopodobne, iż uczeń potrafiący dzielić umie też dodawać. Jest wysoce prawdopodobne, iż uczeń, który zna pięć podstawowych lektur szkolnych, zna również szóstą lekturę obowiązkową. Tak samo, jak wysoce prawdopodobnym jest to, iż uczeń niepotrafiący dodawać nie potrafi też dzielić, oraz że uczeń, który nie czyta lektur szkolnych, nie zna szczegółów powieści „Nad Niemnem" Elizy Orzeszkowej. Innymi słowy, zbadawszy wiedzę ucznia w pewnych wycinkach materiału, możemy próbować przewidzieć, czy uczeń opanował pozostały materiał. Próbujemy tym samym stwierdzić, jaką wiedzę badany uczeń posiada w zakresie całego materiału. Nie mamy jednak pewności – wszystko jest jedynie wysoce prawdopodobne.

Problematyka rzetelności związana jest ściśle z problemem reprezentatywności, dotyczy jednak również innych własności pomiaru. Zastanówmy się, jakie inne czynniki mogą wpływać na wynik ucznia. Danego dnia uczeń może mieć złe samopoczucie, może być chory, może mieć osłabioną motywację, może mieć szczęście albo pecha podczas zgadywania w wypadku zadań zamkniętych. Niewątpliwie, co należy podkreślić, żaden znany nam pomiar wiadomości i umiejętności (nawet ten, który zawierałby cały materiał) nie jest pomiarem w pełni precyzyjnym. Każdy pomiar obarczony jest błędami. Problematyka rzetelności zajmuje się analizą tych błędów, czyli próbuje odpowiedzieć na pytanie, w jakim stopniu dany pomiar jest dokładny.



Rysunek 1.1. Źródła błędów w pomiarze edukacyjnym

Rzetelność testu dotyczy **precyzji**, z jaką jesteśmy w stanie zmierzyć poziom umiejętności i wiedzy ucznia na podstawie zadań dotyczących jedynie fragmentów materiału, który powinien być przez ucznia opanowany w całości.

Oczywiście dobór materiału nie jest jedynym źródłem błędów w pomiarze edukacyjnym. Rysunek 1.1 w schematyczny sposób wskazuje na główne źródła błędów mogące pojawić się w czasie pomiaru wiadomości i umiejętności. Uczeń manifestuje swe umiejętności, rozwiązując dane zadania w teście. Przełożenie umiejętności na wyniki testu może być jednak utrudnione zarówno przez warunki testowania (np. zbyt ciepło, zbyt głośno, zbyt wysokie ciśnienie, możliwość ściągania etc.), jak i indywidualne predyspozycje testowanego (czy uczeń jest zdrowy, czy wyspał się ostatniej nocy, jaką ma kondycję psychiczną) oraz motywację. Możemy przecież bez trudu wyobrazić sobie sytuację, w której uczeń posiada wymagane umiejętności, ale po prostu nie chce mu się rozwiązać zadań testowych.

Obok doboru materiału, sama konstrukcja testu może wpływać na wyniki ucznia, na przykład uczeń, który nigdy wcześniej nie rozwiązywał testu z zadaniami wielokrotnego wyboru, może mieć problemy z techniczną stroną wykonywania poleceń. Możliwości pojawiania się błędów nie kończą się wraz z oddaniem arkusza testowego. W tym momencie pojawia się nowa problematyka rzetelności punktowania (efekt egzaminatora), która uwidacznia się przede wszystkim w czasie oceniania zadań otwartych, czyli takich, w których uczeń nie wybiera jednej dobrej odpowiedzi spośród kilku mu podanych, lecz samodzielnie musi sformułować pisemną odpowiedź. Takie odpowiedzi muszą być sprawdzane przez wyszkolonych egzaminatorów na podstawie precyzyjnie sformułowanego klucza kodowego. Mimo wszystko ani staranie wykonany schemat odpowiedzi, ani doświadczenie i wyszkolenie egzaminatora nie gwarantują tego, iż wszystkie prace ocenione zostaną według identycznych kryteriów, czyli inaczej mówiąc, iż zostaną ocenione sprawiedliwie.

spis treści

Problem ten pojawia się przede wszystkim w przypadku zadań wymagających dłuższej wypowiedzi pisemnej i najdotkliwszy jest w przypadku zadań mierzących umiejętności humanistyczne, krótko mówiąc – wypracowań. Kogoś, kto chodził do szkoły, na pewno nie trzeba będzie przekonywać, iż dwóch różnych nauczycieli może postawić różne oceny za tę samą pracę pisemną. Niektórzy nauczyciele są bardziej surowi, niektórzy mniej, jedni lubią taki sposób pisania, inni zupełnie odmienny. W mniejszej, choć znaczącej skali, problem ten dotyczy również egzaminatorów i sprawdzanych przez nich prac.

Wynik ucznia zależy nie tylko od tego, co on napisze, ale również od tego, jak to, co napisał, będzie sprawdzane. Dlatego dobrym standardem we wszystkich badaniach edukacyjnych jest kontrola jakości sprawdzania prac, czyli tak zwanego efektu egzaminatora. W praktyce polega to na tym, iż dany odsetek prac jest wylosowywany (zazwyczaj około 10%) i sprawdzany jeszcze raz przez innych egzaminatorów. Na podstawie tego wyszukuje się nie tylko nierzetelnie oceniane zadania, ale również nierzetelnie oceniających koderów (osoby oceniające zadania). Możliwa jest też analiza statystyczna tego rodzaju efektów (patrz analizy efektu egzaminatora cytowane w rozdziale 4).

W aspekcie statystycznym problem rzetelności podejmowany jest przez klasyczną teorię testu, która stworzyła narzędzia do statystycznego ujęcia tego problemu. Dzięki tym narzędziom możemy w prosty sposób analizować precyzję, z jaką badane są umiejętności.

Są to kwestie niebagatelne dla badań edukacyjnych i należy to jeszcze raz podkreślić: jeżeli testy są nierzetelne, obarczone dużym błędem, to jakość analiz wykorzystujących wyniki testowania będzie również niska. Niska rzetelność testu powoduje błędne oszacowania korelacji lub współczynników regresji liniowej. Przy niskiej rzetelności będą one znacząco odbiegać od prawdziwych wartości. Trudno będzie też odnaleźć zależności statystyczne. W dalszej części będziemy zajmować się problematyką rzetelności i omawiać klasyczną teorię testu. W rozdziale poświęconym tej teorii będzie można znaleźć statystyczną definicję rzetelności oraz narzędzia, które pozwolą wykorzystać teoretyczną wiedzę w praktyce. Warto podkreślić, iż sprawa rzetelności w pomiarze wiadomości i umiejętności uczniów jest kwestią niezwykłej wagi – szczególnie w przypadkach testów przeznaczonych do użytku publicznego (między innymi narzędzia egzaminów zewnętrznych). Testy przeznaczone do użytku publicznego mogą być dopuszczone do użytku jedynie wtedy, kiedy mają zadowalającą i udokumentowaną rzetelność. Niska rzetelność wiąże się bowiem z nieprecyzyjnym, obarczonym błędem, ocenianiem wiadomości i umiejętności testowanych. Jeżeli test ma niską rzetelność, jego wynik końcowy zależał będzie w dużej mierze od szczęścia (lub pecha) ucznia rozwiązującego dany test. W skrajnych wypadkach może dochodzić do sytuacji, w których osoby o mniejszych wiadomościach i umiejętnościach będą miały wyższe wyniki końcowe testu od osób posiadających większe wiadomości i umiejętności. Nikogo nie

trzeba przekonywać, iż choćby w wypadku egzaminów zewnętrznych, gdzie ważą się przyszłe losy uczniów, jest to sytuacja niedopuszczalna. Problematyka rzetelności, wydawałoby się sprawa techniczna, ma zatem poważne implikacje społeczne. Pytanie o rzetelność testu jest w gruncie rzeczy pytaniem o sprawiedliwe ocenianie.

Zagadnienie skali

Ostatnie zagadnienie związane z pomiarem, którym chcemy się tu zająć, jest to problem skalowania, czyli tworzenia skal na podstawie wyników testu. Problem ten w zadowalający sposób został ujęty dopiero dzięki metodom skalowania opartym na teorii odpowiedzi na zadanie testowe (ang. *item response theory* – IRT). O tej metodzie będzie mowa w osobnym podrozdziale (1.3), a w tej części poruszymy tylko podstawowe kwestie.

Zastanówmy się, czym w istocie jest pomiar. Mówiąc najprościej, pomiar polega na przełożeniu rzeczywistości na liczby (numeryczne reprezentacje rzeczywistości). Pomiar to procedura, wedle której zgodnie z zasadami przyporządkowuje się wartości liczbowe właściwościom empirycznym. Pomiar jest zadaniem, w którym próbujemy przypisać znaczenie liczbom opisującym wynik ucznia na teście. Liczby przyporządkowujemy wedle pewnej reguły poszczególnym obiektom (wynikom testu). Reguła określa procedurę, jaką posłużono się w trakcie przyporządkowywania liczb lub obiektów. Reguła może być prosta, na przykład: przyporządkuj uczniom liczby od 1 do 6 zgodnie ze stopniem ich wiedzy na temat historii drugiej wojny światowej. Jeżeli uczeń ma małą wiedzę, przyporządkuj mu 1; jeżeli posiada dużą wiedzę, przyporządkuj mu liczbę 6. Mamy tutaj rzeczywistość – wiedzę ucznia na temat drugiej wojny światowej – i próbę jej przeniesienia na wartości numeryczne, czyli na standardową skalę ocen.

Podczas procedury skalowania mamy do dyspozycji zbiór obiektów, którym przysługuje pewna cecha. Ze względu na tę cechę możemy wyznaczyć określone relacje między obiektami. Skalowanie polega na tym, by do zbioru obiektów odpowiednio dobrać zbiór liczb. Odpowiednio, czyli w taki sposób, by relacje między liczbami były analogiczne do relacji między obiektami, ze względu na interesującą nas cechę. Jeżeli relacje między obiektami są podobne do relacji między wybranymi liczbami, to zachodzi tzw. izomorfizm.

Ten abstrakcyjny opis możemy zilustrować następującym przykładem. Mamy zbiór 3 uczniów: A, B i C. Uczeń A umie więcej od ucznia B, a uczeń B umie więcej od ucznia C. By tę relację „wymodelować" w zbiorze liczb, tworzymy trzyelementowy zbiór, w którym zachodzą analogiczne relacje. Umiejętności ucznia A opisujemy liczbą 3, umiejętności ucznia B liczbą 2, a umiejętności ucznia C liczbą 1. Widać, iż zachodzi tu wyraźne podobieństwo. 3 jest większe od 2 (uczeń A umie więcej niż uczeń B), 2 jest większe od 1 (uczeń B umie więcej od

ucznia C). Czyli relacje i związki odzwierciedlane przez system liczbowy są analogiczne do relacji i związków między rzeczywistymi pojęciami.

Kwestia izomorfizmu jest złożona i aby omówić ten problem dokładniej, należy odwołać się do typów skal pomiarowych. W naukach społecznych zasadniczo mówi się o czterech typach skal pomiarowych:

- skali nominalnej
- skali porządkowej
- skali interwałowej (przedziałowej)
- skali ilorazowej

W wypadku skali nominalnej o różnych obiektach możemy powiedzieć wyłącznie to, iż różnią się od siebie lub nie różnią się od siebie ze względu na daną cechę, nie możemy natomiast powiedzieć, jak bardzo dane obiekty różnią się od siebie. Skala nominalna tworzy nam zatem rozłączne kategorie obiektów: na przykład chłopcy i dziewczęta, ludzie zielonoocy, niebieskoocy i ludzie o oczach piwnych. Mówimy, że są chłopcy i dziewczęta, lecz nie stwierdzamy, że któraś płeć jest lepsza, większa etc. Spośród operacji arytmetycznych w tym wypadku między dwoma obiektami, dwojgiem ludzi, dwiema kategoriami możemy użyć tylko znaku równa się (=) lub nie równa się (≠). Dwójka osób jest tej samej płci lub nie jest tej samej płci. Nie wprowadza się tu żadnej gradacji.

Gdy wprowadzamy gradację, czyli mówimy, że coś jest większe, lepsze, mądrzejsze etc., przechodzimy do skali porządkowej. W wypadku skali porządkowej spośród operacji arytmetycznych możemy posługiwać się nie tylko znakami równości (=) i nierówności (≠), ale możemy również mówić o relacjach mniejszy (<) i większy (>) oraz mniejszy lub równy (≤) i większy lub równy (≥). Dobrym przykładem zastosowania takiej skali jest skala twardości minerałów, gdzie twardość minerałów określa się w stosunku do siebie. Mówi się, że jeden minerał jest twardszy od drugiego, ponieważ jeden minerał zarysowuje drugi, ale drugi nie zarysowuje pierwszego. Nie mówi się, o ile razy twardszy jest ten materiał od innego materiału; poprzestaje się na stwierdzeniu, że materiał jest twardszy. Wszędzie tam, gdzie możemy zastosować porządkowanie, gdzie mamy kilka obiektów i możemy powiedzieć, że coś jest większe, ładniejsze, mądrzejsze, z pewnością możemy używać skali porządkowej.

Jeżeli chcemy mówić, o ile coś jest twardsze, mocniejsze, cieplejsze etc., musimy odwołać się do skali interwałowej. W tej skali spośród operacji arytmetycznych dozwolone są dodatkowo dwa działania: dodawanie (+) i odejmowanie (-). Klasycznym przykładem użycia skali interwałowej jest pomiar temperatury w stopniach Celsjusza. Dysponujemy wtedy stałymi przedziałami pomiędzy danymi stopniami. Porównujemy temperaturę w stopniach Celsjusza w dwóch miejscach, powiedzmy, w Warszawie i we Wrocławiu. Jeżeli w Warszawie odczytujemy na termometrze +3°, a we Wrocławiu +12°, możemy stwierdzić, iż we Wrocławiu jest o 9° cieplej niż w Warszawie.

Nie możemy jednak powiedzieć, iż we Wrocławiu jest cztery razy cieplej, a to dlatego, że zero w tej skali jest umowne. Jeżeli przejdziemy na skalę Kelwina, która charakteryzuje się tymi samymi stopniami, lecz przesunięciem zera do -273°, to będziemy mieli do czynienia z temperaturą w Warszawie -270° oraz we Wrocławiu -261°. Czy teraz powiemy, że we Wrocławiu jest cztery razy cieplej? Jeżeli chcemy wypowiadać się, że coś jest dwa razy większe, cięższe, lepsze, musimy użyć skali ilorazowej. Na skali ilorazowej mierzymy na przykład wzrost, liczbę dzieci czy dochody. W każdym z tych wypadków możemy powiedzieć bez wahania, iż ktoś jest od kogoś dwa razy wyższy, iż ktoś ma dwa razy więcej dzieci od kogoś innego lub że zarabia trzy razy więcej pieniędzy. Robimy to bez problemu, ponieważ w każdym wypadku posiadamy jednoznaczne, niearbitralne zero.

Wróćmy teraz do naszej reguły przyporządkowywania: przyporządkuj uczniom liczby od 1 do 6 zgodnie ze stopniem ich wiedzy na temat historii drugiej wojny światowej. Jeżeli uczeń ma małą wiedzę, przyporządkuj mu 1; jeżeli dużą wiedzę, przyporządkuj mu liczbę 6. Jest to oczywiście prosta reguła tworząca skalę ocen szkolnych. Pytanie podstawowe brzmi: Jaka to jest skala? Otóż wszystko wskazuje na to, iż jest to skala porządkowa. Nie możemy bowiem powiedzieć na pewno, iż uczeń, który dostał 4, ma dokładnie dwa razy większą wiedzę od ucznia, który uzyskał 2. Co więcej, nie możemy też powiedzieć, że uczeń, który dostał 4, ma o tyle samo większą wiedzę od ucznia, który dostał 2, co uczeń, który dostał 5, od ucznia, który dostał 3. Poza tym trudno znaleźć takie miejsce skali, które odpowiadałoby definitywnie zerowym umiejętnościom. Czy uczeń, który dostał ocenę 1, naprawdę **nic** nie umie? Nie jest to skala interwałowa, gdyż różnica w wiadomościach między oceną 1 a 2 nie jest równa różnicy umiejętności między oceną 4 i 5. Bardzo trudno byłoby udowodnić, iż różnica umiejętności jest tutaj taka sama. Pozostaje zatem skala porządkowa. Czyli taka skala, w której możemy stwierdzić jedynie, iż uczeń A jest lepszy, gorszy lub (z pewnym marginesem błędu) ma takie same umiejętności, jak uczeń B.

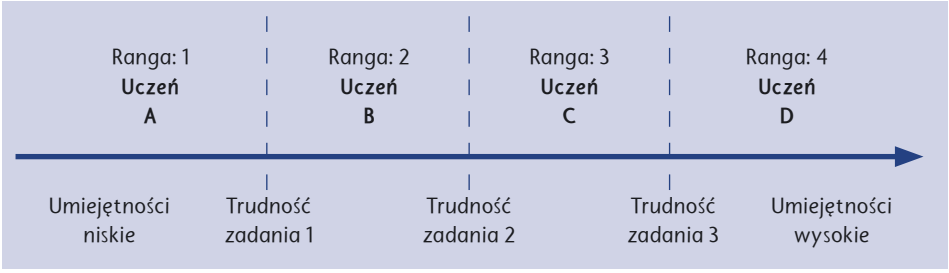
Ma to bardzo poważne implikacje, przede wszystkim matematyczne. Na skali porządkowej nie możemy dokonywać takich operacji, jak dodawanie, odejmowanie, mnożenie i dzielenie. Większość metod statystycznych ma zatem ograniczone zastosowanie. Warto zauważyć, iż w świetle tego, co napisaliśmy, nawet liczenie zwykłej średniej ocen musi budzić poważne kontrowersje. Choć liczenie średniej umożliwi uporządkowanie uczniów od najlepszego do najgorszego, to na pewno nie możemy powiedzieć, ile razy uczeń najlepszy jest lepszy od ucznia najgorszego, co sugerowałyby operacje na liczbach, których dokonujemy podczas liczenia średniej. Mimo iż jesteśmy w stanie to zrobić, używając prostego dzielenia, to z pomiarowego i teoretycznego punktu widzenia nie ma to najmniejszego sensu. W ostateczności może się nawet zdarzyć, że uczeń, który ma wyższą średnią, uzyskał w istocie niższe osiągnięcia niż ten o niższej średniej, ze względu na nieporównywalność skal i ocen między przedmiotami.

W praktyce badań edukacyjnych w kontekście pomiaru mamy kolejny problem – problem skalowania kumulatywnego. Większość badań nie jest oparta na pojedynczym zadaniu, lecz na testach skonstruowanych z pewnej liczby zadań. W tabeli 1.1 przedstawiono hipotetyczną sytuację, w której czterech uczniów (A, B, C, D) rozwiązuje test składający się z trzech (1, 2, 3) zadań. Zera oznaczają niepoprawne rozwiązanie zadania, jedynki rozwiązanie poprawne. Dla przykładu: uczeń A nie rozwiązał poprawnie ani jednego zadania, uczeń B poprawnie rozwiązał tylko zadanie 1. Zadania uporządkowano w kolumnie od najtrudniejszego do najłatwiejszego. Zadanie 3 jest najtrudniejsze, ponieważ rozwiązał je jeden uczeń, zadanie 1 jest najłatwiejsze, ponieważ rozwiązało je trzech uczniów.

Tabela 1.1. Hipotetyczne wyniki czterech uczniów rozwiązujących trzy zadania

Uczeń / Zadanie	A	B	C	D
3	0	0	0	1
2	0	0	1	1
1	0	1	1	1

Jest to klasyczny przykład skalowania kumulatywnego. Trudności zadań w tym przypadku mogą być granicami wyznaczającymi przedziały umiejętności. Pokazane zostało to na rysunku 1.2. Uczeń A posiada umiejętności mniejsze niż trudność zadania 1. Uczeń B posiada umiejętności wyznaczane przez trudność zadania 1 i zadania 2. Posiada umiejętności wyższe lub dokładnie na miarę zadania 1, lecz z całą pewnością niewystarczające do rozwiązania zadania 2. W statystyce taką skalę określa się mianem skalogramu Gutmana. Można by twierdzić, iż w przypadku takich wyników można mówić o skali interwałowej: uczeń C rozwiązał przecież dokładnie o jedno zadanie więcej niż uczeń B. Uczeń D rozwiązał natomiast dokładnie o jedno zadanie więcej niż uczeń C. Pewnie można by na to przystać, jeżeli z jakichś przyczyn chcielibyśmy mówić o liczbie



Rysunek 1.2. Czterech uczniów na kontinuum umiejętności podzielonym przez 3 zadania

rozwiązanych zadań, a nie o umiejętnościach czy o wiedzy ucznia. Mówienie o prostej sumie wykonanych zadań w kontekście pomiaru umiejętności może być jednak problematyczne.

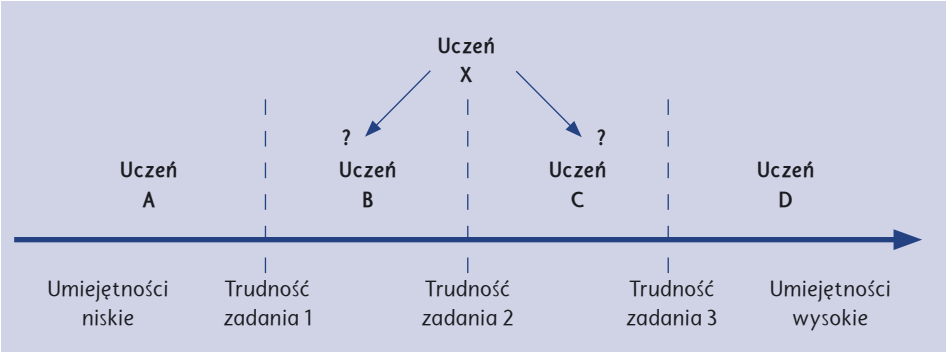
Przyjrzyjmy się tabeli 1.2, analogicznej do tabeli 1.1. Dokonano tutaj jedynie małej zmiany. Dodano ucznia X. Jak widać, uczeń X odpowiedział poprawnie na zadanie numer 2, ale niepoprawnie na zadania 3 i 1.

Tabela 1.2. Hipotetyczne wyniki pięciu uczniów rozwiązujących trzy zadania

Uczeń / Zadanie	X	A	B	C	D
3	0	0	0	0	1
2	1	0	0	1	1
1	0	0	1	1	1

Jak w tej sytuacji ocenić poziom umiejętności ucznia X? Czy o uczniu X możemy powiedzieć, iż posiadł więcej umiejętności od ucznia B? Skoro nie rozwiązał najłatwiejszego zadania, jest to wątpliwe. A czy możemy jego poziom umiejętności umieścić na równi z poziomem ucznia C? Przecież obydwaj uczniowie rozwiązali równie trudne zadanie. Lecz przecież uczeń C rozwiązał zadanie pierwsze i ma wyższy wynik sumaryczny. A może jego poziom umiejętności leży gdzieś pośrodku skali umiejętności? Ale bliżej ucznia B czy C?

Jak widać, sprawa ta jest skomplikowana. Dopiero przełożenie klasycznej teorii skalowania na język prawdopodobieństwa, dokonane około 35 lat temu, może dać nam zadowalające, uznane rozstrzygnięcia. Ponieważ takie przełożenie skalowania z języka prostego wyniku sumarycznego na język teorii prawdopodobieństwa odbyło się wraz ze stworzeniem koncepcji IRT, te rozwiązania zostaną bliżej opisane w części książki poświęconej temu podejściu.



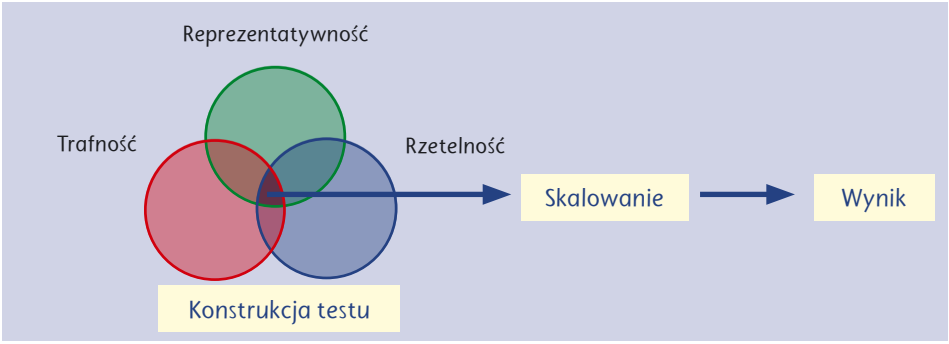
Rysunek 1.3. Pięciu uczniów na kontinuum umiejętności podzielonym przez 3 zadania, problem umiejscowienia ucznia X

Podsumowanie – problemy pomiaru

Gdy uświadomimy sobie podstawowe problemy pomiaru, okazuje się, że skonstruowanie dobrego narzędzia pomiarowego – dobrego testu – jest niełatwe. Na stworzenie dobrych narzędzi kognitywnych (tak niekiedy w sposób ogólny nazywa się testy mierzące umiejętności i wiedzę badanych) potrzeba miesięcy, a nawet lat ciężkiej pracy. Przykładowo, eksperci UNESCO oceniają, że stworzenie dobrego testu osiągnięć szkolnych powinno zająć nie mniej niż 20 tygodni, a są to kalkulacje zakładające istniejącą i sprawnie działającą infrastrukturę[1].

Rysunek 1.4 podsumowuje dotychczasowe rozważania, dając bezpośrednie odniesienie do konstrukcji testu oraz do wyniku będącego zwieńczeniem testowania.

Test powinien być tak skonstruowany, żeby zarówno on sam w całości, jak i jego poszczególne części czy zadania, zapewniały dostateczną reprezentatywność wiedzy, która ma być mierzona. Zadania mierzące tę wiedzę muszą być trafnie wybrane i sformułowane, a test i zadania w nim zawarte muszą zapewniać rzetelność oceniania wiedzy. Te trzy aspekty muszą być spełnione jednocześnie. Test bowiem może być dokładnie punktowany, ale reprezentatywność zadań może być niewystarczająca – taka sytuacja może się zdarzyć, gdy chcemy zbadać wiedzę ucznia o całym średniowieczu, a na teście dajemy dobrze skonstruowane i dokładnie punktowane zadania dotyczące jedynie panowania Mieszka I. Materiał może być też dobrze wybrany, ale zadania mogą być nietrafne, mogą na przykład mierzyć jedynie znajomość faktów, podczas gdy my chcielibyśmy poznać umiejętność ich wykorzystania. Trafne, ciekawe zadania mogą być, oczywiście,



Rysunek 1.4. Problematyka pomiaru w kontekście testów edukacyjnych

[1] Mówiąc o infrastrukturze, mamy na myśli na przykład stały zespół ekspertów pracujący nad testem, wypracowane procedury komunikowania się i współpracy między osobami układającymi zadania, przeprowadzającymi badania pilotażowe lub też prowadzącymi analizy statystyczne testujące właściwości psychometryczne tych zadań. Dodatkowo opracowane są też procedury administracyjne, a tworząca test instytucja ma doświadczenie przy publikacji i stosowaniu testów.

nierzetelne. Słaby schemat punktowania (model odpowiedzi) może sprawić, że części odpowiedzi nie będzie można w łatwy sposób zaklasyfikować jako dobrych lub złych.

Tym, co chcemy otrzymać dzięki pomiarowi edukacyjnemu, jest wynik testowania mówiący o umiejętnościach osób badanych. Należy pamiętać, iż nawet poprawnie skonstruowany test, jeżeli zostanie nieprawidłowo wyskalowany, może dać nam nieużyteczne wyniki. Nie należy zatem zapominać o procesie skalowania, po którym dopiero możemy mówić o jakiejś liczbowej reprezentacji umiejętności danego ucznia lub grupy uczniów.

Na szczęście nie jesteśmy pozbawieni pomocy w tym skomplikowanym procesie tworzenia narzędzi pomiarowych. Naszymi sprzymierzeńcami są tutaj statystyka i psychometria. W dalszej części tego rozdziału przedstawione zostaną najważniejsze koncepcje pomiarowe i najprzydatniejsze narzędzia związane z kontrolą jakości testowania i skalowania wyników.

1.2. Klasyczna teoria testu

Klasyczna teoria testu (*Classical Test Theory*), szczególnie w anglosaskiej terminologii, nazywana jest niekiedy teorią prawdziwego wyniku i błędu (*theory of true and error scores*) lub klasyczną teorią rzetelności (*classical reliability theory*). Nazwy te wiele mówią o całej teorii. Jest to koncepcja, która próbuje pogodzić na gruncie empirycznym i teoretycznym dwa ważne pojęcia: wynik prawdziwy i błąd pomiaru.

Teoria ta zakłada, iż istnieje pewien wynik, który określa prawdziwą wartość tego, co mierzymy. Do owego *prawdziwego wyniku* możemy dojść jedynie pośrednio poprzez empiryczne pomiary, które – jak wiemy – obarczone są błędami wynikającymi z niedoskonałości narzędzi pomiarowych. Żeby przybliżyć tę koncepcję, posłużmy się prostym przykładem. W klasycznej teorii testu zakłada się, że dany obiekt, np. długopis, ma swoją długość (czyli jakąś cechę), którą możemy zmierzyć linijką. Pomiar wykonany przy użyciu linijki nie jest jednak zbyt dokładny. Mierząc linijką, będziemy popełniać błędy – pewnie niewielkie, w granicach milimetra, ale prawdopodobnie nigdy nie uda nam się wskazać długości z większą precyzją, np. z dokładnością do tysięcznej części milimetra. Próbujemy tutaj dojść do *teoretycznej długości*, teoretycznej dlatego, że stojącej poza empirycznymi możliwościami pomiarowymi – potrafimy sobie wyobrazić długość, wiemy, iż możemy o niej myśleć jak o stałej, niezmiennej cesze, lecz nie możemy jej dokładnie zmierzyć.

Zakładamy, że podstawą wyników pomiarów jest jakiś *wynik idealny*, uzewnętrzniający się dzięki narzędziom (np. linijce), które stosujemy do pomiaru. Doświadczenie empiryczne pokazuje nam przecież, że wyniki pomiaru są różne, ale przecież utrzymują się (na pewno nieprzypadkowo) w pewnych granicach.

Podobnie jest w przypadku testów wiadomości i umiejętności – niedoskonałymi narzędziami próbujemy wypowiedzieć się o umiejętnościach ucznia. Sytuacja nie jest tu oczywiście tak przejrzysta, jak w wypadku linijki i długopisu. Trzeba jednak przyznać, iż gdy dajemy kolejne testy różnym uczniom, okazuje się, że występuje tutaj pewna powtarzalność, co do której nikogo nie trzeba przekonywać. Część uczniów będzie uzyskiwać regularnie niskie wyniki, a część uczniów – regularnie wysokie. Żaden test, nawet najlepszy, nie określi nam precyzyjnie wyniku, na jaki *zasługuje* uczeń, każdy wynik jest obarczony niepewnością pomiarową, tak jak pisaliśmy wcześniej: choćby spowodowaną wyborem zadań, sposobem ich sformułowania i oceniania.

Siła klasycznej teorii testu tkwi w tym, iż dzięki niej w stosunkowo prosty sposób możemy wypowiadać się za pomocą statystyki o dokładności narzędzia pomiarowego. Jest to zresztą podkreślone przez wielu badaczy, którzy często klasyczną teorię testu nazywają *klasyczną teorią rzetelności*. Dzięki klasycznej teorii testu nie uzyskamy wyniku, który bezbłędnie określałby umiejętności ucznia czy długość długopisu, możemy jednak dowiedzieć się, w jakim stopniu nasze narzędzia są niedokładne i jak duży może być nasz błąd – a to już naprawdę wiele.

Chociaż istnieją znacznie bardziej wyrafinowane techniki i koncepcje teoretyczne związane z testowaniem i mierzeniem umiejętności, klasyczna teoria testu nadal jest koncepcją chętnie wykorzystywaną przez wielu badaczy. Po pierwsze, właśnie ze względu na prostotę. Po drugie, ze względu na elastyczność. Klasyczna teoria testu jest też często punktem wyjścia do bardziej złożonych analiz, odrzucających kolejne nierzeczywiste założenia związane z teorią. Warto ją zatem poznać i wykorzystywać możliwości, jakie za nią stoją, do konstrukcji i analiz testów.

Pierwsza część tego podrozdziału będzie się składała z rozważań teoretycznych, momentami dość oderwanych od praktyki, niemniej jednak polecamy ją dla pełnego zrozumienia zagadnień poruszanych w drugiej części, gdzie skupimy się na praktycznym zastosowaniu klasycznej teorii testu. Warto poznać podstawowe założenia teorii po to, by uświadomić sobie jej ograniczenia i zrozumieć, jakie koncepcje przyświecały utworzonym na jej podstawie narzędziom mającym praktyczne zastosowanie.

Wynik prawdziwy i błąd pomiaru

W klasycznej teorii testu zakłada się, że faktycznie uzyskany wynik testu składa się z dwóch komponentów: *wyniku prawdziwego* i *błędu pomiaru*. Wynik prawdziwy utożsamia się ze średnią wyników uzyskanych przez ucznia, gdyby był testowany (teoretycznie) za pomocą tego samego narzędzia nieskończoną ilość razy. Musimy sobie wyobrazić hipotetyczną sytuację, w której uczeń dostaje test, rozwiązuje go, test zostaje sprawdzony, a wynik wyskalowany i zapisany. Następnie pamięć ucznia zostaje wymazana (tak, że badany nie pamięta już tego, że rozwiązywał jakiś test) i ponownie daje mu się do rozwiązania test, który już raz rozwiązywał (lecz

o tym nie pamięta). Uczeń go rozwiązuje, test zostaje sprawdzony i wyskalowany, a wyniki zapisane. Tę czynność powtarza się nieskończoną liczbę razy. Średni wynik wszystkich pomiarów będzie uznany za *wynik prawdziwy*.

Przypomnijmy raz jeszcze przykład długopisu mającego pewną długość, którą chcemy dokładnie zmierzyć za pomocą linijki. Jeżeli dokonamy kilku niezależnych pomiarów, okaże się, że wyniki nieznacznie różnią się od siebie. Raz odczytaliśmy trochę więcej na podziałce naszej miarki, raz trochę mniej. W klasycznej teorii testu zakłada się, że jeżeli mamy stałą miarę, to mierząc nieskończoną ilość razy (w praktyce: wiele razy) i uśredniając wyniki, dojdziemy do *wyniku prawdziwego*, właściwego dla danego narzędzia pomiarowego. Inaczej mówiąc, zakładamy, że błędy popełniane przy pomiarze mają charakter losowy, raz zaniżają, raz zawyżają wyniki, lecz przy odpowiednio dużej liczbie pomiarów błędy te znoszą się (ich średnia wynosi 0).

Warto zwrócić uwagę, iż *wynik prawdziwy,* rozumiany jako średnia pomiarów, jest pojęciem statystycznym i nie jest tożsamy z czymś, co moglibyśmy nazwać *cechą ukrytą obiektu*, czyli własnością ucznia warunkującą wszystkie uzyskane wyniki. Termin *wynik prawdziwy* odnosi się w tak przyjętej definicji tylko do konkretnej procedury pomiarowej. Zakładamy tutaj, że mierzony obiekt ma prawdziwy, właściwy sobie wynik danej procedury pomiarowej. Pomiar może być natomiast sam w sobie nietrafny, źle pomyślany (miarka może być nieprecyzyjnie wykonana, a test źle skonstruowany). Wynik procedury, w której wielokrotnie testujemy ucznia za pomocą źle przygotowanego narzędzia, też możemy nazwać *wynikiem prawdziwym*.

W postaci ogólnej (dlatego stosujemy wielkie litery) możemy zatem zapisać, iż wynik uzyskany podczas egzaminu jest sumą *wyniku prawdziwego* oraz błędu pomiaru:

$$X = T + E \qquad (1)$$

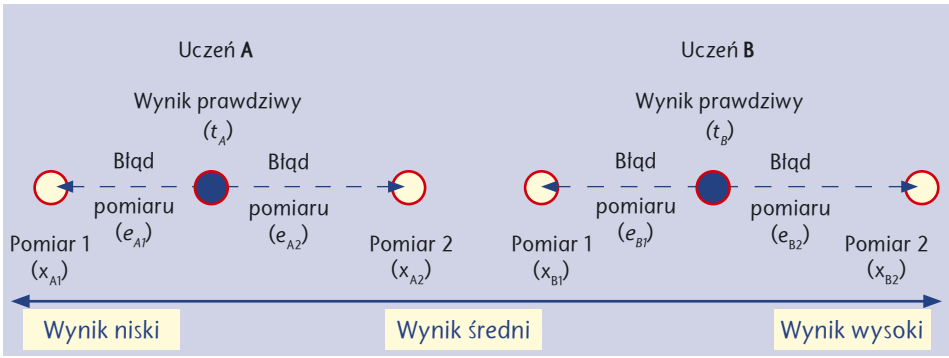
gdzie:

X – uzyskany wynik punktowy

T – wynik prawdziwy

E – błąd pomiaru

Powyższy wzór można dość łatwo przedstawić graficznie, co zostało zrobione na rysunku 1.5, na którym uwzględniono dwóch uczniów (uczeń A i B). Dla każdego z nich przedstawiono po dwa wyniki pomiarów oraz *wynik prawdziwy*. Wyniki pomiarów oznaczone zostały białymi kropkami i literą x z indeksem dolnym, w którym pierwszy znak określa ucznia, drugi numer pomiaru (na przykład wynik drugiego pomiaru dla ucznia A oznaczony został jako x_{A2}). Natomiast *wyniki prawdziwe* danych uczniów zaznaczone zostały kropkami czarnymi oraz literą t z pojedynczym indeksem oznaczającym ucznia (np. *wynik prawdziwy* dla ucznia B oznaczony został jako t_B). Błędy pomiaru oznaczone są przez przerywane strzałki i oznaczone symbolem e z indeksem analogicznym jak

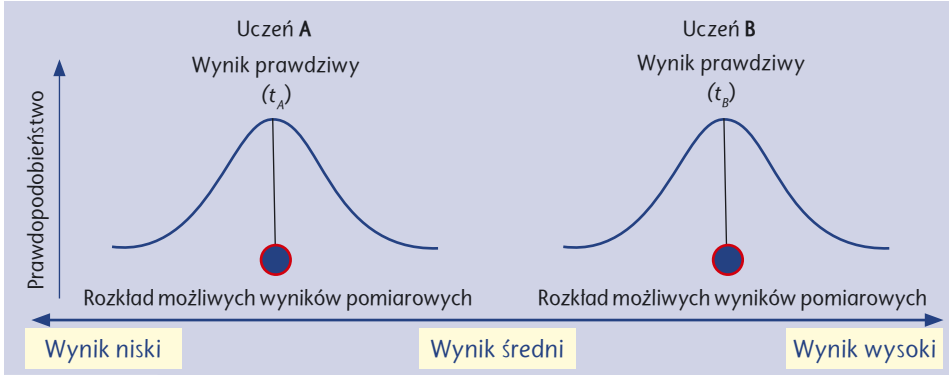


Rysunek 1.5. Pomiar, *wynik prawdziwy* i błąd pomiaru

w wypadku oznaczenia pomiaru. Na rysunku widać ponadto, iż wyniki ucznia A (tak wyniki pomiaru, jak *wynik prawdziwy*) lokują się poniżej średniej, natomiast wyniki ucznia B powyżej średniej.

Jeżeli wynik pomiaru jest wyższy niż *wynik prawdziwy*, mówimy, że błąd jest dodatni (tak jak e_{A2}). Jeżeli wynik pomiaru jest mniejszy niż *wynik prawdziwy*, wtedy błąd pomiaru określamy jako ujemny (tak jak e_{A1}).

Jako że wcześniej pisaliśmy, iż w klasycznej teorii testu mówi się o sytuacji teoretycznej, w której mamy do czynienia z nieskończoną liczbą pomiarów, dokonywanych za pomocą danego narzędzia, to trzeba zdawać sobie sprawę, iż rysunek 1.5 jest dużym uproszczeniem, gdyż pokazuje tylko po dwa pomiary dla każdego z badanych uczniów, a nie ich nieskończoną liczbę. Gdy chcemy teoretycznie poprawnie mówić o nieskończenie wielu pomiarach, należy odwołać się do rozkładu możliwych wyników pomiaru, a nie do poszczególnych pomiarów. Jeżeli błędy są losowe, oznacza to, że wyniki pomiarów mają określony rozkład prawdopodobieństwa. Dla ułatwienia często zakłada się, iż jest to rozkład



Rysunek 1.6. Rozkład prawdopodobieństwa pomiarów i *wynik prawdziwy*

normalny, co zostało przedstawione na rysunku 1.6, na którym nie zaznaczano już poszczególnych możliwych do uzyskania wyników, ale pokazano ich rozkłady za pomocą krzywej normalnej.

Krzywa normalna w tym wypadku mówi nam o prawdopodobieństwie uzyskania danego wyniku. Jeżeli pomiar powtarzalibyśmy nieskończenie wiele razy, to rozkład prawdopodobieństwa mówi nam, gdzie (na skali wyników) uzyskalibyśmy najwięcej wyników pomiarów, gdzie stosunkowo mało, a gdzie najmniej. W przypadku naszego rysunku tam, gdzie krzywa uzyskuje najwyższy punkt, czyli w okolicach *wyniku prawdziwego*, mielibyśmy najwięcej wyników (czyli *wynik prawdziwy* jest wynikiem najbardziej prawdopodobnym), najmniej natomiast na skrajach rozkładu – tam, gdzie krzywa staje się płaska. Kształty te są bardzo do siebie zbliżone i nie jest to przypadkowe. Niemal zawsze zakładamy, że pomiary obarczone błędem mają rozkład normalny, taki jak wiele losowych zjawisk w naturze. Najczęściej pomiary będą dawały wynik bliski *wynikowi prawdziwemu*, inaczej mówiąc, rzadko kiedy popełniamy duże błędy, choć i te się zdarzają. Co więcej, rozkłady są symetryczne wokół prawdziwej wartości, czyli tyle samo razy mylimy się na plus i tyle samo na minus, dzięki czemu możemy zakładać, że błędy pomiaru się znoszą.

Rozkład prawdopodobieństwa pomiarów wokół *wyniku prawdziwego* można oczywiście traktować jako rozkład błędów pomiaru, pamiętając o tym, że błąd pomiaru to po prostu różnica między wynikiem pomiaru a *wynikiem prawdziwym*. Jak widać na rysunku 1.6, zakładamy tutaj, że najbardziej prawdopodobne są małe błędy (w okolicach *wyniku prawdziwego*), najmniej prawdopodobne „grube" błędy – im dalej od *wyniku prawdziwego*, tym prawdopodobieństwo uzyskania danego wyniku mniejsze. Błąd w klasycznej teorii testu jest pojęciem fundamentalnym, wyjaśnijmy zatem, co teoria zakłada o jego naturze. Zakłada się zatem, że:

a) błędy pomiaru są losowe,
b) średnia błędów pomiaru równa jest zero,
c) błędy dla kolejnych testów (czy też poszczególnych zadań) są od siebie niezależne,
d) błędy nie są związane z *wynikiem prawdziwym* ucznia.

Przyjęcie takich założeń prowadzi do kilku implikacji. Błędy mają znak ujemny i dodatni – zakłada się, że średnia błędów równa jest zero. Błędy „znoszą się" nawzajem. Jeżeli będziemy przeprowadzać pomiar dostatecznie wiele razy, średnia błędów będzie asymptotycznie dążyła[2] do zera (punkty a i b). Kierunek błędów (plus lub minus) w jednym pomiarze lub w serii pomiarów nie przesądza o kierunku błędów w kolejnych pomiarach (punkty a, b i c). Błędy pomiaru danego narzędzia są takie same zarówno dla wyników wysokich, jak i dla niskich (punkt d).

[2] Przy coraz większej liczbie pomiarów średnia będzie coraz bliższa zera.

Mimo iż są to założenia znacznie upraszczające rzeczywistość pomiarową, są ważne, dzięki nim można wypowiadać się o jakości pomiarowej narzędzi. Formalnie rozkład indywidualnych błędów możemy zapisać jako: $N(0; \sigma_e^2)$. Taki zapis oznacza, iż błędy mają rozkład normalny (co symbolizuje litera *N*) o średniej równej 0 (pierwsze oznaczenie w nawiasie) oraz danej wariancji (oznaczonej jako: σ_e^2), czyli parametrze rozproszenia błędów (im rozproszenie większe, tym rozkład wyników szerszy).

Błędy systematyczne i niesystematyczne (losowe)

Błąd pomiaru w wyniku to wszystko to, co nie jest prawdziwym wynikiem – tak głosi definicja, warto jednak bardziej sprecyzować to stwierdzenie. Błędy pomiaru można podzielić na dwie kategorie: błędy systematyczne i błędy losowe.

Błędy **systematyczne** to wszystkie błędy, które wpływają na wyniki wszystkich badanych lub danych grup poddawanych pomiarowi tendencyjnie, czyli systematycznie zaniżają lub zawyżają wyniki. Pomiary konsekwentnie zawyżające lub zaniżające wyniki testowania są obarczone błędami systematycznymi.

Weźmy na przykład wagę. Jeżeli waga za każdym razem podaje wynik zaniżony o 10 gramów w stosunku do wagi obiektu, to jest to błąd systematyczny. Jeżeli stoper, którym mierzymy czas przebiegnięcia 100 metrów, śpieszy się – jednostki czasu są dla niego krótsze, stoper ten systematycznie podaje zawyżone, czyli w kontekście sportowej rywalizacji słabsze, wyniki. Jeżeli ważylibyśmy obiekty nieskończoną liczbę razy daną wagą lub mierzyli czas określonym stoperem, a błąd powtarzałby się w taki sam sposób (miał taki sam kierunek), mielibyśmy do czynienia z błędem systematycznym.

Podobnie, jeżeli test składa się z zadań obarczonych wrażliwością na płeć zdających (*gender bias*) i któraś z płci uzyskuje systematycznie gorsze wyniki, nie z powodu gorszych wiadomości i umiejętności, lecz np. z powodów motywacyjnych. Niektóre zadania mogą być po prostu zdecydowanie atrakcyjniejsze dla którejś płci. Rozwiązanie zadania mierzącego konkretną umiejętność może zależeć od wiadomości niezwiązanych z treścią nauczania, a mogących być z różnych powodów odmiennymi w poszczególnych grupach uczniów. Na przykład zadania związane z topografią Warszawy mogą być łatwiejsze dla uczniów z Warszawy. W takich przypadkach mówimy o błędach systematycznych.

W wypadku testowania błędy systematyczne mogą pojawiać się również z powodu nieprecyzyjnych schematów oceniania (modeli odpowiedzi), różnie zinterpretowanych przez różnych oceniających, z powodu różnych warunków testowania (np. jedna grupa używa kalkulatorów, a inna nie używa) oraz innych błędów w konstrukcji testu, związanych przede wszystkim z trafnością testu.

Klasyczna teoria testu **nie zajmuje się błędami systematycznymi**. Wszystkie narzędzia, jakie zostały stworzone na podstawie klasycznej teorii testów, opierają się na analizie drugiego typu błędów – błędów losowych. Dlatego definicję:

Błąd pomiaru w wyniku to wszystko to, co nie jest prawdziwym wynikiem, należy uzupełnić: *Błąd pomiaru to wszystko to, co jest losowe i nie jest prawdziwym wynikiem*. Jak pisaliśmy wcześniej, określenie *wynik prawdziwy* jest pojęciem statystycznym. Błąd systematyczny „przesuwa" nam *wynik prawdziwy* na danej skali, nie wpływa jednak na precyzję uzyskania tego „przesuniętego" wyniku. Jeżeli istnieje błąd systematyczny, systematycznie zaniżający wyniki, to i (w klasycznej teorii testu) *wynik prawdziwy* zostanie zaniżony; jeżeli mamy do czynienia z systematycznym błędem zawyżającym wyniki, to i *wynik prawdziwy* zostanie zawyżony.

W wypadku błędów niesystematycznych zakłada się, iż wpływają na wyniki losowo. Nawet jeżeli damy testowanemu dwa takie same testy jeden po drugim, jest prawdopodobne, iż uzyska różne wyniki. Oczywiście, aby być w zgodzie z teorią, w takim eksperymencie myślowym trzeba założyć, iż uczeń nie uczy się i nie męczy się podczas rozwiązywania testu (rozwiązania testów są niezależne). Ale nawet przy takich założeniach jest wysoce prawdopodobne, iż wyniki będą się różniły. Wtedy mówimy, iż zróżnicowanie otrzymanych wyników będzie jedynie skutkiem błędu losowego i to właśnie ten błąd jest przedmiotem analizy klasycznej teorii testu.

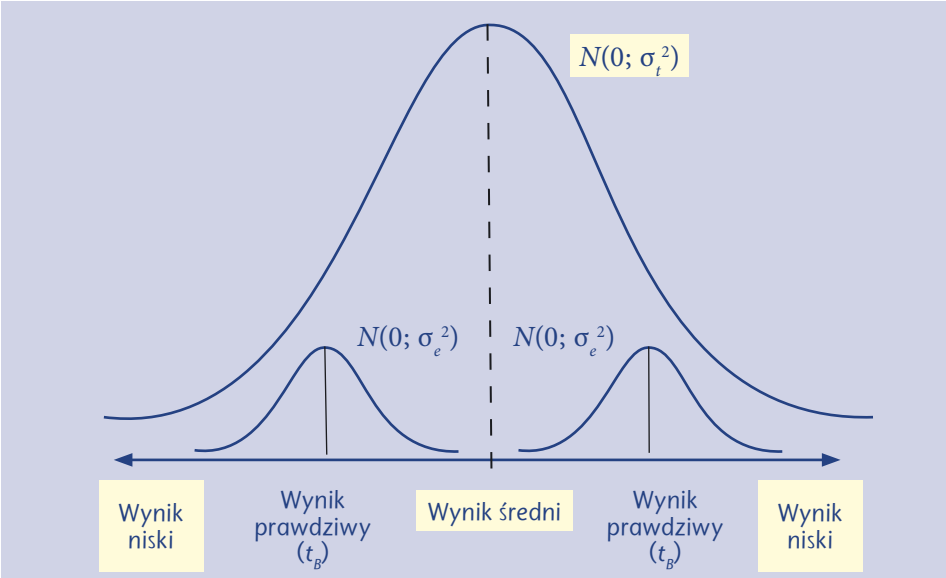
Rzetelność

Do tej pory rozważaliśmy przede wszystkim sytuację pomiarową, w której mierzyliśmy (badaliśmy) pojedynczego ucznia. Gdy jednak myślimy o teście, zazwyczaj dotyczy on większej liczby uczniów. Aby mówić o rzetelności, trzeba wprowadzić rozkład *wyników prawdziwych*. Został on przedstawiony na rysunku 1.7 narysowany ciągłą, grubą, czarną linią. Określone zostały jego parametry: średnia równa zero i wariancja σ_t^2 ($N(0; \sigma_t^2)$).

Średnia jest tu równa zero, co może dziwić – jest to konwencja wprowadzona przez statystyków, znacznie ułatwiająca obliczenia. Każdy rozkład łatwo można sprowadzić do średniej zero – odejmując od wszystkich wyników poszczególnych uczniów wynik średni testu. Wtedy uczniowie, którzy uzyskali wyniki poniżej średniej, charakteryzowani będą przez wyniki ujemne, a uczniowie powyżej średniej – przez wyniki dodatnie. Na rysunku 1.7 widać też rozkłady błędów losowych dla nieskończonej liczby powtórzeń testowania, dla dwóch wybranych uczniów.

Zakładając, że błędy są losowe, niezależne od siebie i niezależne od *wyniku prawdziwego*, możemy powiedzieć, iż wariancja, czyli zróżnicowanie wyników, dla całego testu ma dwa źródła: wariancję błędu i wariancję *wyniku prawdziwego*. Jako że błędy nie są zależne od *wyniku prawdziwego* i od siebie nawzajem, możemy zapisać całą wartość zmienności wyników uzyskanych analogicznie do sumy przedstawionej w równaniu 1: X = T + E, gdzie wynik pomiaru równy był *wynikowi prawdziwemu* plus błąd pomiaru. Równanie opisujące całkowite zróżnico-

spis treści



Rysunek 1.7. Rozkład *wyników prawdziwych*

wanie wyników można zapisać jako sumę zróżnicowania *wyników prawdziwych* i sumę zróżnicowania błędów:

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 \tag{2}$$

gdzie:

σ_x^2 – wariancja (zróżnicowanie) wyniku uzyskanego (pomiaru)
σ_t^2 – wariancja (zróżnicowanie) *wyniku prawdziwego*
σ_e^2 – wariancja (zróżnicowanie) błędu

Do tej pory o rzetelności mówiliśmy tylko ogólnikowo, jako o wskaźniku jakości narzędzia pomiarowego. Mówiliśmy o tym, że dobrze jest, gdy rzetelność jest wysoka – testy są precyzyjne, a ocenianie sprawiedliwe. Niska rzetelność łączy się natomiast z dużymi błędami pomiaru. Teraz możemy precyzyjnie zdefiniować to pojęcie. Rzetelność jest to stosunek zróżnicowania *wyniku prawdziwego* do zróżnicowania wyniku uzyskanego na podstawie testowania (będącego sumą zróżnicowania *wyniku prawdziwego* oraz zróżnicowania błędu pomiaru). W matematycznej formie możemy to zapisać następująco:

$$\rho = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} \tag{3}$$

gdzie, tak jak poprzednio:

σ_e^2 – wariancja *wyniku prawdziwego*
σ_t^2 – wariancja błędu

Rzetelność narzędzia pomiarowego jest zatem tym większa, im mniejsza jest wariancja błędu pomiaru, i przyjmuje wartości od 0 do 1 (gdzie 1 oznacza maksymalną rzetelność). Gdy rzetelność wynosi 1, mamy do czynienia z pomiarem idealnym, czyli takim, w którym wynik pomiaru jest tożsamy z *wynikiem prawdziwym*. Z prostego matematycznego przekształcenia wynika, iż wariancja błędu musi wynosić wtedy zero – pomiar jest bezbłędny. Oczywiście, w wypadku pomiaru edukacyjnego z takimi sytuacjami nie mamy do czynienia.

Łatwo zauważyć, iż aby obliczyć rzetelność, musimy znać wariancję błędu oraz wariancję *wyniku prawdziwego*. Dysponujemy jednak tylko wynikiem testowania i znamy tylko jego wariancję. Aby oszacować rzetelność, należałoby poznać przynajmniej wariancję błędu pomiaru. W praktyce badawczej musimy go oszacować, co często robi się przez powtórzenie pomiaru dwa razy na tej samej grupie uczniów, z użyciem tego samego narzędzia (tzw. test-retest). Takie podejście wydaje się najwłaściwsze, choć budzi ono wiele kontrowersji. Na przykład, w takich warunkach badawczych nie jesteśmy w stanie uniknąć tego, że uczniowie, rozwiązując test, uczą się. Użycie tego samego narzędzia pomiarowego w wypadku świata społecznego jest zawsze problematyczne, gdyż ludzie nie muszą reagować na serie podobnych pomiarów za każdym razem w taki sam sposób. Wcześniejsze pomiary mogą wpływać na pomiary późniejsze. Nieczęsto też dysponujemy możliwością powtórnego testowania tych samych uczniów z użyciem tego samego narzędzia. Na szczęście istnieją narzędzia statystyczne, które przynajmniej w przybliżeniu mogą oszacować rzetelność testu bez potrzeby powtórnego pomiaru, czyli wypowiedzieć się o rzetelności testu jedynie na podstawie jego jednokrotnego przeprowadzenia. Jedno z takich narzędzi opisujemy poniżej.

Klasyczna teoria testu jako pomoc w analizie testów i zadań

W tej części odniesiemy się do wskaźników opartych na klasycznej teorii testu, które są pomocne zarówno w analizie całych testów, jak i poszczególnych zadań. Wszystkie prezentowane wskaźniki łatwo otrzymuje się za pomocą popularnych programów statystycznych (SPSS, SAS, Stata, R) lub programów przeznaczonych do analiz testów. Wiedząc jednak, iż dostępność ich może być ograniczona, wszystkie przedstawione wskaźniki dla tego rozdziału prezentujemy w taki sposób, aby można było obliczyć je z łatwością za pomocą arkusza kalkulacyjnego (takiego jak MS Excel czy arkusz kalkulacyjny pakietu Open Office).

Wskaźniki, które zostaną omówione, to wskaźnik rzetelności alfa Cronbacha, łatwość i moc różnicująca. Ponadto przedstawiona zostanie prosta analiza graficzna. Wszystkie one są prostą, lecz nieocenioną pomocą zarówno dla konstruktorów testów, jak i dla wszystkich tych, którzy chcą kontrolować jakość testów.

Alfa Cronbacha – dolna granica rzetelności

Jak pisaliśmy wcześniej, oszacowanie rzetelności testu na podstawie wielokrotnego pomiaru nie dość, że budzi pewne kontrowersje, to często okazuje się niepraktyczne. Istnieją narzędzia statystyczne, które umożliwiają oszacowanie w przybliżeniu rzetelności testu na podstawie jego wyników. Jednym z takich narzędzi jest wskaźnik alfa Cronbacha, dzięki któremu możemy oszacować dolną granicę rzetelności. Nie potrafimy, co prawda, obliczyć dokładnej wartości rzetelności, ale dysponując dolną granicą rzetelności, możemy określić minimalną wartość rzetelności dla analizowanego przez nas testu. Możemy tym samym powiedzieć, iż test ma rzetelność nie mniejszą, niż mówi nam wartość oszacowania wskaźnika alfa Cronbacha. Oblicza się ją na podstawie poszczególnych zadań, traktując poszczególne zadania tak, jakby były osobnymi testami.

Poznanie rzetelności testu jest ważne, gdyż na jej podstawie możemy oceniać jakość naszego narzędzia pomiarowego. Testy o skrajnie niskiej rzetelności nie powinny być używane nie tylko w wypadku egzaminów zewnętrznych, ale również w praktyce szkolnej.

Obliczenie wskaźnika alfa Cronbacha nie jest trudne i sprowadza się do użycia jednego wzoru:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_{suma}^2}\right) \qquad (4)$$

gdzie:
k – liczba zadań w teście
σ_i^2 – wariancja dla *i*-tego zadania
σ_{suma}^2 – wariancja dla wyniku sumarycznego

Aby pokazać, jak obliczyć wskaźnik alfa Cronbacha, odwołajmy się do prostego przykładu. W tabeli 1.3 przedstawiono wyniki 5 uczniów, którzy pisali krótki, składający się z trzech zadań test. W wierszach możemy odczytać wyniki poszczególnych uczniów. W pierwszej kolumnie znajduje się numer identyfikacyjny każdego ucznia, w drugiej wyniki punktowe, jakie uczniowie uzyskali, rozwiązując zadanie pierwsze, w trzeciej – zadanie drugie, w czwartej – zadanie trzecie. W ostatniej kolumnie znajduje się suma punktów uzyskanych przez ucznia z trzech zadań. Ponadto w wierszu oznaczonym jako „Wariancja" podana została wariancja (patrz słowniczek) wyników punktowych – w kolejnych kolumnach dla poszczególnych odpowiedzi, a w ostatniej kolumnie dla sumy testu.

Jeżeli posługujemy się arkuszem kalkulacyjnym MS Excel, możemy użyć funkcji „Wariancja(zakres)". W komórce, w której chcemy uzyskać wartość

spis treści

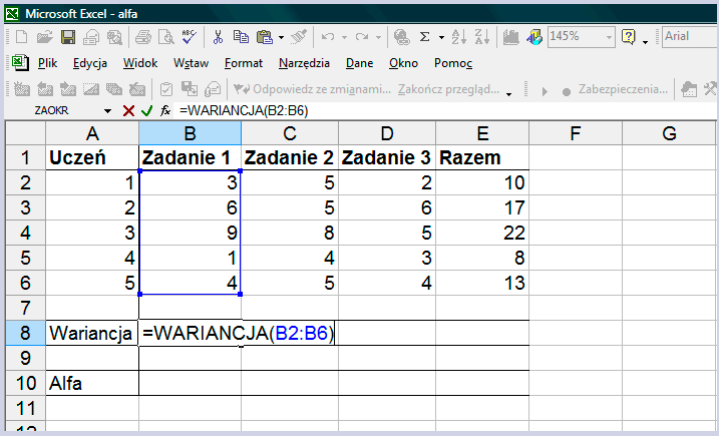
Tabela 1.3. Obliczanie wskaźnika alfa Cronbacha

Uczeń	Zadanie 1	Zadanie 2	Zadanie 3	Razem
1	3	5	2	10
2	6	5	6	17
3	9	8	5	22
4	1	4	3	8
5	4	5	4	13
Wariancja	9,3	2,3	2,5	31,5
Alfa	0,83			

wariancji danego zadania (lub sumy wyników), wpisujemy znak „=", a po nim wpisujemy „Wariancja", następnie otwieramy nawias i kursorem myszy zaznaczamy zakres interesujących nas wartości liczbowych, np. kolumnę z wynikami rozwiązań pierwszego zadania. Następnie zamykamy nawias i wciskamy klawisz *Enter*. Jeżeli zaznaczyliśmy liczby (3; 6; 9; 1; 4), to (tak jak w przykładzie) otrzymamy wynik 9,3.

Na rysunku 1.8 pokazano, jak na przykładzie wyników z tabeli 1.3 można obliczyć wariancję dla wyników zadania 1 za pomocą arkusza kalkulacyjnego MS Excel.

Procedurę powtarzamy dla wszystkich zadań oraz dla wyniku sumarycznego. Następnie podstawiamy do wzoru na wskaźnik alfa Cronbacha. Dla powyż-



Rysunek 1.8. Obliczanie wariancji za pomocą arkusza kalkulacyjnego MS Excel

szego przykładu podstawienie do wzoru na wskaźnik alfa Cronbacha wygląda następująco:

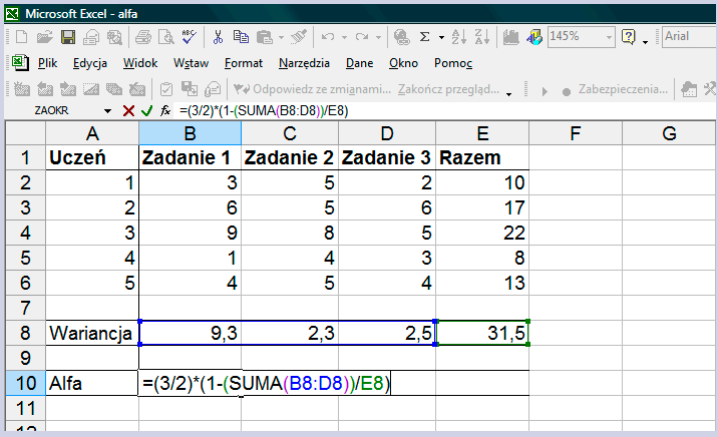
$$\alpha = \frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_{suma}^2}\right) = \frac{3}{3-1}\left(1-\frac{9,3+2,3+2,5}{31,5}\right) = \frac{3}{2}\left(1-\frac{14,1}{31,5}\right) = 0,83$$

Oczywiście nie trzeba dokonywać obliczeń „ręcznie" lub z użyciem kalkulatora. Dobrym narzędziem do tego typu obliczeń jest arkusz kalkulacyjny. Rysunek 1.9 przedstawia, w jaki sposób po obliczeniu potrzebnych wariancji został obliczony wskaźnik alfa Cronbacha.

Po obliczeniu wariancji dla wszystkich zadań zastosowana została formuła „=(3/2)*(1-SUMA(B8:D8))/E8)". W pierwszym nawiasie znajduje się formuła „k/(k-1)", czyli w naszym przypadku „3/(3-1)=3/2". Dalej mamy jeden minus suma wariancji zadań „(SUMA(B8:D8))" podzielona przez wariancję sumarycznego wyniku „(E8)".

Otrzymaliśmy wynik 0,83. Można go oceniać jako rzetelność przeciętną, jeszcze akceptowalną. W tabeli 1.4 przedstawiono przedziały wyników wskaźnika alfa Cronbacha z ich krótką interpretacją.

O wysokich wskaźnikach rzetelności mówimy, gdy wskaźnik alfa Cronbacha jest wyższy niż 0,9. W przedziale 0,70-0,89 uznaje się zwyczajowo, iż rzetelność nie jest wysoka, lecz jest do zaakceptowania. Gdy rzetelność testu znajduje się w przedziale 0,60-0,69, powinno to zaniepokoić konstruktora testu. Możliwe, że po usunięciu jakiegoś zadania (nieuwzględnieniu go w sumarycznym wyniku),



	A	B	C	D	E	F	G
1	**Uczeń**	**Zadanie 1**	**Zadanie 2**	**Zadanie 3**	**Razem**		
2	1	3	5	2	10		
3	2	6	5	6	17		
4	3	9	8	5	22		
5	4	1	4	3	8		
6	5	4	5	4	13		
7							
8	Wariancja	9,3	2,3	2,5	31,5		
9							
10	Alfa	=(3/2)*(1-(SUMA(B8:D8))/E8)					
11							

Rysunek 1.9. Obliczanie wskaźnika alfa Cronbacha za pomocą arkusza kalkulacyjnego MS Excel

Tabela 1.4. Wartości wskaźnika alfa Cronbacha i jego interpretacja

Alfa Cronbacha	Interpretacja
1,00-0,90	Wysoka rzetelność, test bardzo rzetelny
0,70-0,89	Rzetelność przeciętna, spójność testu do zaakceptowania
0,60-0,69	Rzetelność niska, w pewnych warunkach do zaakceptowania
0,00-0,59	Bardzo niska rzetelność, sugeruje, że test nie jest spójny i wskazuje na potrzebę jego analizy merytorycznej

rzetelność może wzrosnąć. Ponadto, przy małej liczbie zadań, czasami można zdecydować się na uznanie takiego testu za poprawny pod względem rzetelności. Trzeba pamiętać jednak o tym, iż precyzja pomiaru za pomocą takiego testu jest niewielka.

Test o stosunkowo niskiej rzetelności (0,60-0,69) może nadać się do tego, by sprawdzić, czy uczniowie osiągnęli dane umiejętności, czy nie. Innymi słowy, możemy z pewnym ryzykiem mówić o ocenie dwuwartościowej *umie – nie umie*. Lecz próba stworzenia z niego klasycznych 6 kategorii ocen jest ryzykowna. Przy tak niskiej rzetelności różnica pomiędzy dwoma stopniami (np. trójką i czwórką) może okazać się kwestią jedynie błędu pomiaru – czyli *de facto* zależeć od szczęścia lub pecha ucznia, a nie jego prawdziwej wiedzy czy prawdziwych umiejętności, które zamierzaliśmy mierzyć. Podobnie w wypadku rzetelności poniżej 0,6. W tego typu sytuacjach możemy się spodziewać, iż test obarczony jest dużym błędem, który sprawi, że nasze oceny będą niesprawiedliwe. Wymaga on zatem głębokiej analizy merytorycznej, poprawienia konstrukcji i/lub poprawienia poszczególnych zadań – lub w ogóle rezygnacji z niego jako narzędzia pomiarowego.

Łatwość

Oprócz analizy rzetelności, czyli wskaźnika dla całego testu, należy przeprowadzić analizę zadań – nie tylko wtedy, kiedy test wykazuje niską rzetelność, ale praktycznie zawsze. Z punktu widzenia pomiaru robimy to po to, by wyśledzić zadania słabe lub niepotrzebne pomiarowo, a z punktu widzenia dydaktyki dlatego, że analiza taka niesie nieocenioną ilość informacji o osiągnięciach uczniów.

Pierwszym wskaźnikiem, który wykorzystujemy przy analizie testu, jest wskaźnik łatwości. W klasycznej teorii testu łatwość definiuje się jako stosunek liczby punktów uzyskanych przez uczniów za zadanie do maksymalnej liczby punktów możliwych do uzyskania za to zadanie. Inaczej mówiąc, jest to średnia arytmetyczna wyników uczniów otrzymanych za zadanie podzielona przez maksymalną liczbę punktów za to zadanie. Wskaźnik łatwości jest wartością zawiera-

jącą się w przedziale od 0 do 1. Warto zauważyć, iż dla zadań punktowanych zero-jedynkowo (0 i 1) łatwość wyrażona w procentach jest jednocześnie informacją o odsetku uczniów, który poprawnie rozwiązał zadanie.

Łatwość w klasycznej teorii testu jest jednym ze wskaźników niezwykle wrażliwych na badaną grupę. Jest przecież oczywiste, iż jeżeli damy do rozwiązania ten sam test uczniom pierwszej klasy szkoły podstawowej i uczniom szóstej klasy, to łatwość zadania będzie dla tych dwóch badanych grup diametralnie różna. Jeżeli dany test ma być przeznaczony dla wszystkich pierwszoklasistów, to oczywiście łatwość poszczególnych zadań będzie różna w klasie bardzo dobrej i w klasie słabej. Jak widać, trudno mówić o łatwości zadania rozpatrywanej jako jego niezależna własność. Właściwie zawsze musimy mówić o łatwości w odniesieniu do konkretnej badanej grupy, która może być bardzo szeroka, na przykład mogą to być wszyscy polscy pierwszoklasiści. Gdy robimy test przygotowany specjalnie dla pewnej klasy, szkoły czy gminy, możemy wypowiadać się o łatwości danego zadania dla badanej grupy, ale nie możemy wnioskować na temat łatwości w innych grupach lub w populacji.

Jest to problem szczególnie ważny w aspekcie egzaminów zewnętrznych dla wszystkich uczniów. Teoretycznie bowiem przed użyciem zadania nie znamy jego trudności, a to uniemożliwia zaplanowanie właściwej struktury testu. Oczywiście są sposoby, by szacować łatwość zadań. Do tego celu należy posłużyć się losową próbą uczniów, którzy rozwiążą zadania przed przeprowadzeniem egzaminu na wszystkich uczniach. Szczegółowo zasady losowania uczniów i procedury szacowania wyników omówione zostaną w rozdziale 2. Wróćmy teraz do wskaźników opartych na klasycznej teorii testu.

W tabeli 1.5 przedstawiono zwyczajowo przyjęte przedziały wskaźnika łatwości. W pierwszej kolumnie mamy przedziały dla zadań, z których można było uzyskać zero punktów bądź jeden punkt (punktowanych zero-jedynkowo). W kolumnie trzeciej przedziały łatwości dla zadań wielopunktowych, dokładnie – zadań czteropunktowych.

Tabela 1.5. Wartości wskaźnika łatwości i ich interpretacja

Wskaźnik łatwości dla zadań otwartych	Interpretacja wartości wskaźnika	Wskaźnik łatwości dla zadań zamkniętych
0,00-0,19	zadanie bardzo trudne	0,00-0,39
0,20-0,49	zadanie trudne	0,40-0,62
0,50-0,69	zadanie umiarkowanie trudne	0,63-0,77
0,70-0,89	zadanie łatwe	0,78-0,92
0,90-1,00	zadanie bardzo łatwe	0,93-1,00

Wskaźnik łatwości umożliwia nam nie tylko stwierdzenie, iż dla danej populacji dane zadanie było łatwe bądź bardzo trudne. Wskaźnik łatwości może być też narzędziem do oceny testu. Dobry test, mogący mierzyć pełen przedział umiejętności, powinien zawierać zadania z całego ich zakresu. Jeżeli będzie zawierał zadania tylko trudne, będziemy mogli jedynie stwierdzić, że uczeń jest bardzo dobry lub nie jest bardzo dobry. Uczniowie bardzo dobrzy rozwiążą zadania bardzo trudne i będziemy mogli ich zidentyfikować. Jeżeli zadania będą bardzo trudne, to uczniowie średni odpowiedzą tak samo, jak uczniowie bardzo słabi. Nie będzie można wtedy dokonać rozróżnienia pomiędzy nimi. Obie grupy nie rozwiążą zadań bardzo trudnych. Nie będzie można wtedy dokonać rozróżnienia pomiędzy nimi.

Może oczywiście zdarzyć się sytuacja, że potrzebna jest nam właśnie taka informacja. Chcemy dowiedzieć się, którzy uczniowie są bardzo dobrzy, a reszta nas nie interesuje. Na przykład, gdy chcemy zakwalifikować najlepszych uczniów do konkursu. Wtedy świadomie test powinien składać się przede wszystkim z zadań trudnych. Możliwa jest też inna sytuacja – chcemy zidentyfikować uczniów słabych, na przykład po to, aby zapewnić im dodatkową pomoc pedagogiczną. Jeżeli chcemy wytypować takich uczniów za pomocą testu, wtedy rozsądnie jest skonstruować test w taki sposób, aby składał się z zadań przede wszystkim łatwych. W tym wypadku nie będą nas interesować informacje o uczniach dobrych i bardzo dobrych. Jeśli jednak chcemy mierzyć umiejętności i wiedzę w całej populacji uczniów o bardzo różnym poziomie osiągnięć, to test powinien zawierać odpowiednią liczbę zadań o każdym poziomie łatwości. Tak jest w wypadku egzaminów zewnętrznych, a przynajmniej sprawdzianu szóstoklasistów i egzaminu gimnazjalnego, gdzie dla każdego poziomu ucznia chcielibyśmy dokonać jak najbardziej precyzyjnego pomiaru.

Moc różnicująca

Kolejnym wskaźnikiem przydatnym w analizie zarówno zadań, jak i całego testu, jest moc różnicująca (synonimem dla tego pojęcia jest „moc dyskryminująca" – my jednak dla porządku terminologicznego zdecydowaliśmy się używać tej drugiej nazwy jedynie dla wskaźników opartych na modelach IRT).

Wskaźnik mocy różnicującej pokazuje nam, jak silnie zadanie związane jest z całym testem. Im zadanie jest silniej związane z wynikiem dla całego testu, tym jego zdolność przewidywania *wyniku prawdziwego* jest większa. Dobre pomiarowo zadania powinny być silnie związane z całym testem.

Jeżeli zadanie jest słabo (a w skrajnych przypadkach ujemnie) związane z całym testem, może to oznaczać, iż mierzy ono jakąś inną umiejętność, niż zaplanowali konstruktorzy testu, lub jest źle skonstruowane, może być nawet błędne pod względem merytorycznym. Brak związku zadania z całym testem oznacza, że uczniowie uzyskujący wysokie wyniki z całego testu (z innych za-

dań) nie uzyskują równie dobrych wyników przy rozwiązywaniu danego zadania i analogicznie – uczniowie uzyskujący słabe wyniki z całego testu nie uzyskują równie słabych wyników przy rozwiązywaniu zadania o niskiej mocy różnicującej.

W sytuacji skrajnej, lecz zdarzającej się w praktyce, gdy moc różnicująca danego zadania jest ujemna, oznacza to, iż uczniowie uzyskujący wysokie wyniki z całego testu rozwiązują źle zadanie o ujemnej mocy różnicującej, a uczniowie o niskich wynikach z całego testu rozwiązują to zadanie dobrze (dobrze według zamysłu autora lub schematu punktowania). Taka sytuacja, w której uczniowie dobrzy rozwiązują zadanie źle, a uczniowie słabi dobrze, jest z punktu widzenia pomiaru i zdrowego rozsądku niedopuszczalna. To tak, jakby waga pokazywała ludziom szczupłym i niewielkiego wzrostu, że ważą dużo, a ludziom wysokim i otyłym – że ważą mało. Takie zadania (o ujemnej mocy różnicującej) są rzadkie w testach edukacyjnych, choć zadania o niskiej mocy różnicującej zdarzają się często.

Jeżeli zadanie jest silnie związane z wynikiem całego testu, możemy mówić, iż dobrze różnicuje uczniów, czyli dobrze *przewiduje* ich umiejętności mierzone sumarycznym wynikiem testu. Przewidywanie łączy się z różnicowaniem. Gdybyśmy chcieli wypowiadać się o uczniach i ich umiejętnościach na podstawie rozwiązania jednego zadania ocenianego na 0 lub 1, moglibyśmy podzielić uczniów na dwie grupy – na tych, którzy zadania nie rozwiązali (0) i o nich sądzilibyśmy, iż są słabi ze względu na badaną umiejętność, oraz na tych, którzy zadanie rozwiązali poprawnie (1) i ich uznawalibyśmy za uczniów o wysokich umiejętnościach (mierzonych danym testem). Różnicowanie takie jest tym lepsze (lepsze, czyli popełniamy mniej błędów w ocenie ucznia), im wyniki tego zadania są silniej związane z wynikiem sumarycznym testu. Wynik sumaryczny testu jest bowiem wskaźnikiem *wyniku prawdziwego*, czyli wyniku określającego poziom umiejętności ucznia.

Wskaźnik mocy różnicującej można obliczyć na kilka sposobów. My odwołamy się do dwóch najprostszych metod. Moc różnicującą możemy traktować jako związek wyników zadania z wynikami całego testu, w którym mieści się zadanie – wskaźnik ten nazywać będziemy MOC(test). Związek ten możemy charakteryzować współczynnikiem korelacji Pearsona, czyli powszechnie używaną miarą związku liniowego między dwiema zmiennymi (patrz słowniczek). Drugi wskaźnik to związek wyników zadania z wynikami testu po wyłączeniu danego zadania, nazwany MOC(r-test)[3] i również często obliczany za pomocą korelacji Pearsona.

Istnieje kilka innych wskaźników mocy, które różnią się jedynie wykorzystaniem różnych współczynników korelacji. Zamiast współczynnika korelacji

[3] MOC(r-test) od „reszta testu", moc zadania dla sumy pozostałych pytań; po wyłączeniu zadania, dla którego analizowana jest moc.

Tabela 1.6. Wskaźniki mocy różnicującej dla hipotetycznych wyników punktowych

Uczeń	Zadanie 1	Zadanie 2	Zadanie 3	Razem
1	3	5	2	10
2	6	5	6	17
3	9	8	5	22
4	1	4	3	8
5	4	5	4	13
MOC(test)	0,99	0,88	0,79	
MOC(r-test)	0,97	0,79	0,64	

liniowej Pearsona używa się niekiedy korelacji biseryjnych (dają one bardziej precyzyjne wyniki dla zadań o dwupunktowej skali 0 i 1). Jeżeli Czytelnik korzysta z pakietu statystycznego, w którym istnieje możliwość użycia korelacji biseryjnych (dla zadań 0-1), namawiamy go, aby wykorzystał tę opcję. Podobnie, jeżeli zamawia analizy u wykwalifikowanych statystyków, powinien wiedzieć, iż istnieją takie możliwości. W innych wypadkach można posłużyć się korelacją Pearsona[4]. W praktyce bowiem różnice okazują się niewielkie.

Wróćmy do prostych wskaźników opartych na korelacji Pearsona, które bez problemu możemy wyliczyć samodzielnie. W tabeli 1.6 przedstawiono wartości wskaźników mocy różnicującej dla zadań z wcześniejszego przykładu. Jak widać, w naszym wypadku wartość MOC(test) jest bliska jedności, czyli maksymalnej wartości mocy różnicującej (teoretycznie wartości mocy różnicującej zawierają się w przedziale od -1 do 1).

Kolejne zadania coraz słabiej różnicują uczniów, chociaż nadal różnicowanie jest na względnie wysokim poziomie. Można też zobaczyć, że wartości wskaźnika MOC(r-test) dla każdego zadania są niższe niż MOC(test). W wypadku testów o dużej liczbie zadań (powyżej 25) różnice te są minimalne. Dla testów krótszych lepiej jest stosować wartość MOC(r-test), która nie jest zawyżana przez duży udział punktów za dane zadanie w całym teście. W długich testach sumaryczny wynik nie zależy tak bardzo od jednego zadania.

W tabeli 1.7 na następnej stronie przedstawione zostały interpretacje dla różnych przedziałów wskaźnika mocy różnicującej. Są to, oczywiście, tylko wskazówki, które mogą pomóc w wypadku interpretacji tego wskaźnika.

Wskaźniki obu mocy różnicujących bardzo łatwo jest policzyć samodzielnie, za pomocą arkusza kalkulacyjnego, tak jak w przypadku alfy Cronbacha, wyko-

[4] Współcześnie często stosuje się również wskaźnik mocy różnicującej oparty na korelacjach polichorycznych (dla zadań o większej liczbie punktów), które są dostosowane do zmiennych porządkowych. Problematyka korelacji polichorycznych wykracza jednak poza granice tej książki. Jeżeli jednak zamawiamy badania u wykwalifikowanych statystyków, możemy sobie życzyć, aby wykorzystywali oni najnowsze osiągnięcia statystyki.

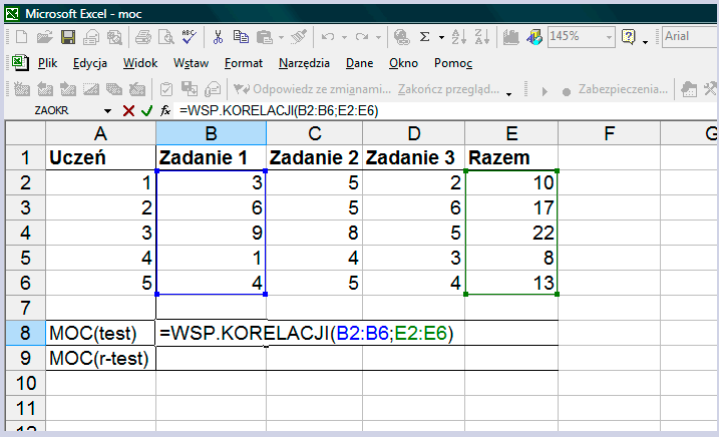
Tabela 1.7. Wartości wskaźnika mocy różnicującej i ich interpretacja

Moc różnicująca	Interpretacja mocy różnicującej zadania
< 0	Ujemna korelacja, zadanie definitywnie złe pomiarowo, możliwe błędy merytoryczne
0,00-0,19	Nie różnicuje, zadanie nie do zaakceptowania w takiej formie
0,20-0,39	Słabo różnicuje, warto przyjrzeć się bliżej temu zadaniu i je poprawić, raczej powinno zostać usunięte
0,40-0,59	Umiarkowanie różnicuje, zadanie przyzwoite pomiarowo
0,60-0,79	Dobrze różnicuje, zadanie dobre pomiarowo, takie wartości powinny pojawiać się najczęściej
0,80-1,00	Bardzo dobrze różnicuje, zadanie wzorcowo różnicujące

rzystując istniejące już formuły. Aby policzyć wskaźnik MOC(test), wystarczy użyć funkcji „=Wsp.korelacji(tablica1;tablica2)". Na rysunku 1.10 przedstawiono, jak łatwo uzyskać ten wynik dla dobrze znanego nam przykładu.

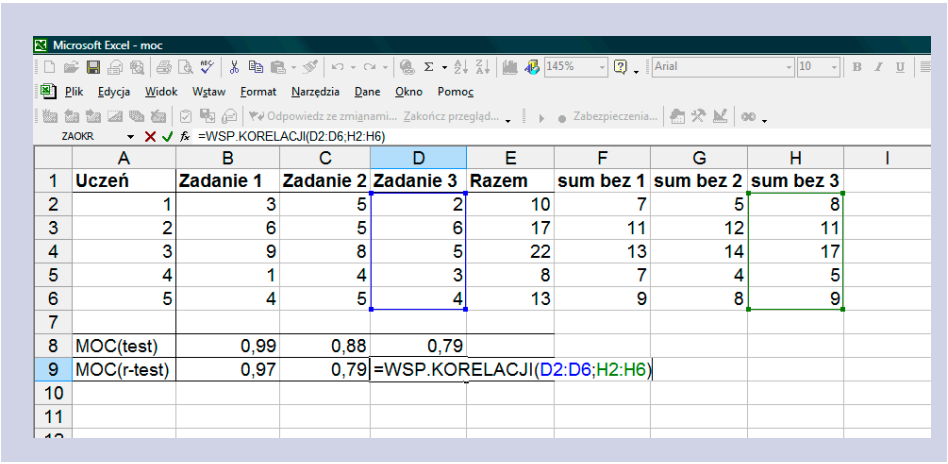
Wystarczy w komórce, w której chcemy uzyskać wynik, wpisać funkcję „=Wsp.korelacji", otworzyć nawias, następnie zaznaczyć zakres wyników odpowiedzi uczniów do pierwszego zadania (w przykładzie „B2:B6"), zamknąć nawias i nacisnąć klawisz *Enter*, a następnie wybrać wynik sumaryczny (w tym wypadku jest to kolumna „Razem" zakres „E2:E6"). Powtarzając tę procedurę, możemy obliczyć wskaźniki MOC(test) dla wszystkich zadań z testu.

Obliczanie wskaźnika MOC(r-test) jest tylko odrobinę bardziej skomplikowane. Wymaga utworzenia nowych kolumn, które będą zawierać wyniki sumaryczne



	A	B	C	D	E	F	G
1	**Uczeń**	**Zadanie 1**	**Zadanie 2**	**Zadanie 3**	**Razem**		
2	1	3	5	2	10		
3	2	6	5	6	17		
4	3	9	8	5	22		
5	4	1	4	3	8		
6	5	4	5	4	13		
7							
8	MOC(test)	=WSP.KORELACJI(B2:B6;E2:E6)					
9	MOC(r-test)						
10							
11							

Rysunek 1.10. Obliczanie wskaźnika MOC(test) za pomocą arkusza kalkulacyjnego MS Excel



	A	B	C	D	E	F	G	H	I
1	Uczeń	Zadanie 1	Zadanie 2	Zadanie 3	Razem	sum bez 1	sum bez 2	sum bez 3	
2	1	3	5	2	10	7	5	8	
3	2	6	5	6	17	11	12	11	
4	3	9	8	5	22	13	14	17	
5	4	1	4	3	8	7	4	5	
6	5	4	5	4	13	9	8	9	
7									
8	MOC(test)	0,99	0,88	0,79					
9	MOC(r-test)	0,97	0,79	=WSP.KORELACJI(D2:D6;H2:H6)					
10									
11									

Rysunek 1.11. Obliczanie wskaźnika MOC(r-test) za pomocą arkusza kalkulacyjnego MS Excel

bez uwzględnienia jednego zadania. Tak w prezentowanym przykładzie stworzono trzy nowe kolumny „sum bez 1", „sum bez 2", „sum bez 3", które oznaczają kolejno sumy punktów, nie licząc zadania pierwszego, drugiego i trzeciego. Obliczone zostały w prosty sposób: jako różnice sumy całkowitej („Razem") i wyników z kolejnych zadań. Gdy zostaną już utworzone kolumny zdające sprawę z wyników bez kolejnych zadań, otrzymanie wskaźników MOC(r-test) odbywa się w sposób analogiczny do wskaźników MOC(test), ale w tym wypadku obliczamy korelację między zadaniem a sumą testu z wyłączeniem wyników danego zadania.

Na rysunku 1.11 pokazany został przykład obliczania wskaźnika MOC(r-test) dla zadania trzeciego ze znanego przykładu. Analogicznie jak dla wskaźnika MOC(test), używamy funkcji „=Wsp.korelacji(tablica1; tablica2)". W przykładzie pokazane jest, jak obliczony został wskaźnik dla zadania trzeciego. Wpisano formułę „=Wsp.korelacji", otworzono nawias, następnie zaznaczono zakres „D2:D8" (wyniki dla zadania trzeciego), naciśnięto klawisz *Enter* i wybrano zakres „H2:H6" (wyniki dla sumy zadań testu bez zadania trzeciego), następnie kolejny raz naciśnięto klawisz *Enter*.

Warto wspomnieć na koniec o dość poważnym mankamencie wskaźnika mocy różnicującej, którym jest zaniżanie mocy dyskryminacyjnej dla zadań bardzo łatwych oraz zadań bardzo trudnych. Trzeba to mieć na uwadze, jeżeli oceniana jest jakość pomiarowa zadań o bardzo różnym poziomie łatwości.

Analiza graficzna

Bardzo częstą i bardzo dobrą praktyką w analizie zadań jest używanie analizy graficznej. Jest to prosta technika, oparta na współczynniku łatwości za-

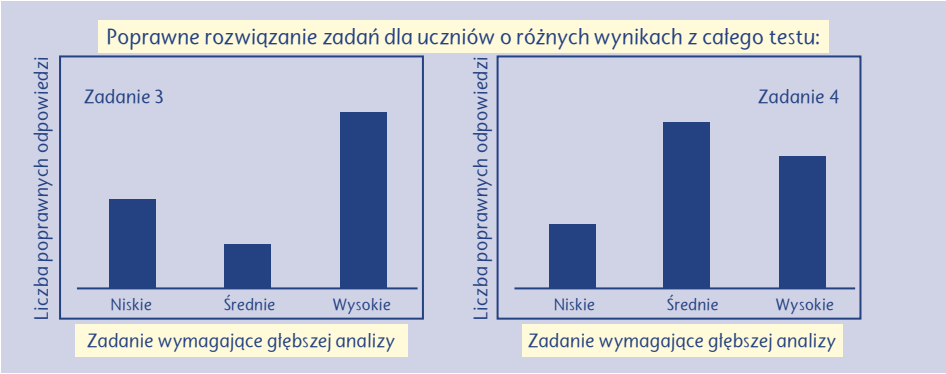
dań, dająca nam informacje dopełniające analizę mocy różnicującej. Zasada jest prosta. Obliczamy wynik dla całego testu. Następnie dzielimy uczniów na trzy grupy, według wyników uzyskanych w teście, tak, że w pierwszej grupie znajduje się 1/3 wszystkich uczniów, wybranych ze względu na najniższe wyniki. W grupie drugiej powinni znaleźć się uczniowie o średnich wynikach, a w grupie trzeciej uczniowie, którzy uzyskali najwyższe wyniki (1/3 wszystkich uczniów). W ten sposób mamy trzy równoliczne grupy (lub prawie równoliczne), dla których obliczmy łatwość dla poszczególnych zadań (czyli procent rozwiązania zadania). Po obliczeniu łatwości dla tych trzech grup najlepiej je przedstawić za pomocą wykresu słupkowego, tak jak zostało to przedstawione na rysunkach 1.12, 1.13 i 1.14.

Rysunki przedstawiają kilka typowych sytuacji, z którymi może zetknąć się osoba analizująca wyniki zadania. Na rysunku 1.12 przedstawiono dwie skrajne sytuacje.

Zadanie 1 wydaje się poprawne pod względem pomiarowym. Grupa uczniów o niskich wynikach z całego testu uzyskała wyniki niższe niż uczniowie średni i uczniowie, którzy uzyskali wysokie wyniki w całym teście – i ci uczniowie rozwiązali to zadanie najlepiej. W wypadku zadania drugiego jest dokładnie odwrotnie. Uczniowie najsłabsi uzyskali najlepsze wyniki dla tego zadania, uczniowie najlepsi (o najwyższych wynikach z tego testu) najrzadziej odpowiedzieli poprawnie. Jasne jest, iż zadanie to ma ujemną moc różnicującą i tym samym definitywnie jest zadaniem złym pomiarowo i warto się przyjrzeć dokładniej jego poprawności merytorycznej i formalnej.

Na rysunku 1.13 przedstawiono dwa profile zadań problematycznych. Jak widać, problematycznie układają się tutaj wyniki uczniów średnich. W zadaniu 3 uzyskali oni najniższe rezultaty spośród trzech analizowanych grup. W zadaniu 4 uzyskują wyniki najwyższe.

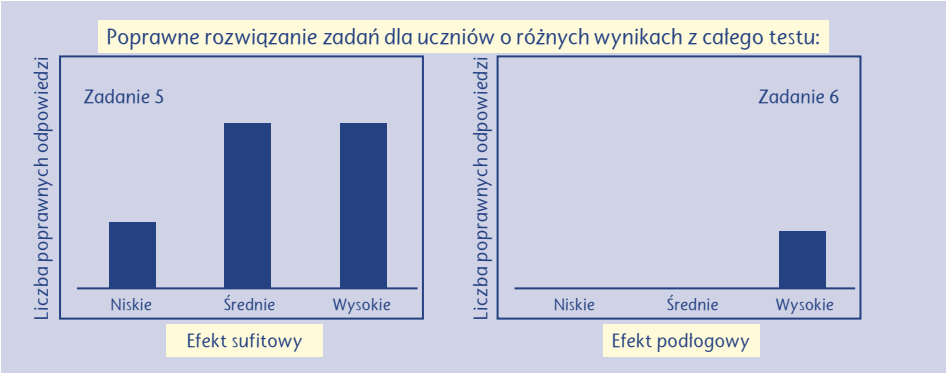
Rysunek 1.12. Analiza graficzna zadań – zadania o poprawnych właściwościach pomiarowych i niepoprawnych właściwościach pomiarowych (za: John Izard, UNESCO)



Rysunek 1.13. Analiza graficzna zadań – zadania o niejednoznacznych właściwościach pomiarowych (za: John Izard, UNESCO)

Może to świadczyć o pułapkach generowanych przez zadanie, czasami niepożądanych dojściach do jego rozwiązania, drogach na skróty, błędach w modelu odpowiedzi (w wypadku zadań otwartych) lub niejasnym sformułowaniu zadania. W takich wypadkach należy głębiej zastanowić się nad formą i sformułowaniem zadania oraz nad schematem punktowania (modelem odpowiedzi). Warto też zauważyć tutaj dużą zaletę analizy graficznej. Może ona pokazywać związki nieliniowe, czyli wychwycić nieregularności między tymi grupami uczniów i pokazać, jaką formę przybierają. Wskaźniki mocy różnicującej mogłyby w określonych sytuacjach nie wychwycić problemów odkrywanych przez analizę graficzną, co jeszcze dobitniej widać na przykładach przedstawionych na rysunku 1.14.

Mamy tutaj do czynienia z dwoma częstymi efektami napotykanymi w analizie zadań. W ilustracji wyników zadania 5 mamy do czynienia z tzw. efektem



Rysunek 1.14. Analiza graficzna zadań – efekt sufitowy i efekt podłogowy (za: John Izard, UNESCO)

sufitowym. Zadanie to jest tak łatwe, że zarówno uczniowie uzyskujący średnie wyniki z całego testu, jak i uczniowie uzyskujący wyniki wysokie dostają maksymalną liczbę punktów. Zatem zadanie to niewiele nam o nich powie. Jednak zadanie to nie musi być uważane za błędne. Tak jak pisaliśmy w punkcie dotyczącym łatwości zadań, jeżeli test skonstruowany jest przede wszystkim po to, by wychwycić uczniów najsłabszych, dominacja zadań o takim profilu odpowiedzi (lub podobnym) jest nawet wskazana; jeżeli chcemy stworzyć test badający szeroki zakres umiejętności, również może znaleźć się w nim kilka zadań o profilu podobnym do profilu charakteryzującego zadanie 5.

Podobnie ma się sprawa z zadaniem 6, w którym uwidacznia się tzw. efekt podłogowy. Widać wyraźnie, iż tylko niewielka grupa uczniów, i tylko spośród najlepszych, rozwiązała poprawnie to zadanie. Podobnie jak w wypadku zadania 5, o poprawności pomiarowej tych zadań świadczy kontekst ich użycia. Jeżeli zadanie 6 ma służyć wyszukaniu uczniów najlepszych, jeżeli test jest na przykład testem kwalifikującym do olimpiady przedmiotowej, to dominacja takich zadań ze względu pomiarowego jest nawet uzasadniona. Podobnie jak wcześniej, jeżeli test ma mierzyć cały przedział umiejętności, takie zadanie może w nim wystąpić.

1.3. Teoria odpowiedzi na zadanie testowe (item response theory, IRT)

W tej części omówione zostanie alternatywne do klasycznej teorii testu podejście pomiarowe: teoria odpowiedzi na zadanie testowe (ang. *item response theory,* IRT). Koncepcja ta ma wiele zalet, które przyczyniły się do jej dynamicznego rozwoju i zdobycia ogromnej popularności w rozmaitych zastosowaniach pomiarowych. Dzięki dorobkowi tej teorii, konstruktorzy testów i badacze edukacyjni w sposób bardziej adekwatny są w stanie analizować zarówno właściwości poszczególnych zadań testowych, jak i całych testów. Łatwiejsze stało się konstruowanie testów i dostosowanie ich do konkretnych potrzeb. Przy odpowiednim zaplanowaniu testowania możliwym stało również precyzyjne porównywanie różnych wersji testów oraz różnych badanych grup rozwiązujących podobne testy. W tej części opisane zostaną podstawowe zalety teorii odpowiedzi na zadanie testowe, wskazane zostaną również ważne problemy ujawniające się przede wszystkim wtedy, gdy badacze lub konstruktorzy testów przeceniają zalety opisywanej teorii i zbudowanych na niej narzędzi statystycznych.

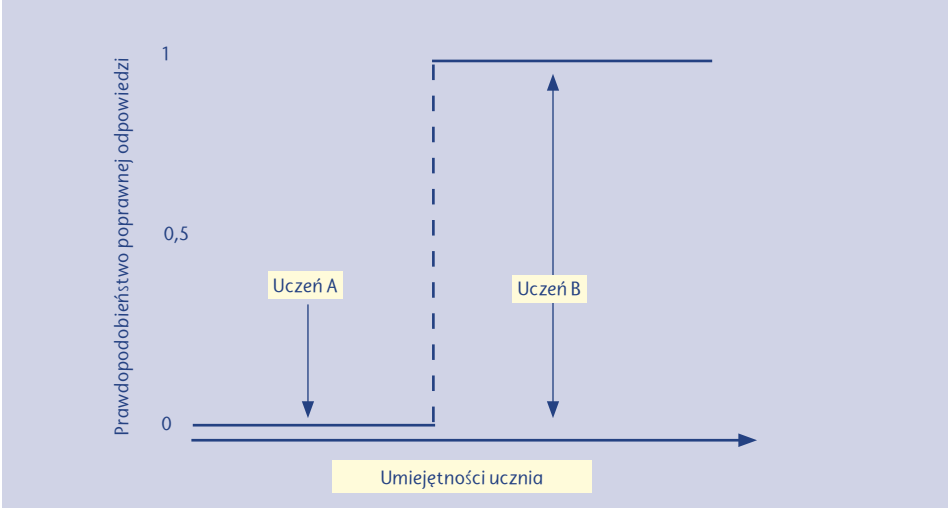
Podstawowe definicje

Gdy omawiane było *skalowanie* (patrz podrozdział 1.1 dotyczący skalowania), zmierzyliśmy się z niebagatelnym problemem: Jak na skali umiejętności umiejscowić ucznia, który wymyka nam się z prostego, kumulatywnego

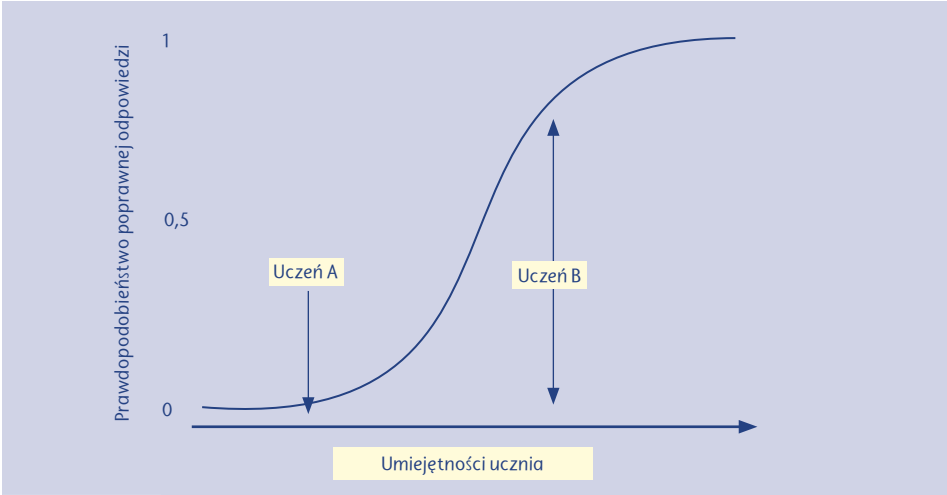
schematu? Przypomnijmy: mamy 3 zadania o określonych trudnościach. Jeżeli uczeń odpowiedział na wszystkie zadania lub na żadne z nich – sprawa jest prosta; podobnie, jeżeli odpowiedział na dwa łatwiejsze lub tylko na najłatwiejsze zadanie. Jeżeli patrzymy jedynie na wartości sumy punktów, coś niepokojącego dzieje się, gdy uczeń odpowiada prawidłowo na zadanie trudne, lecz błędnie na zadanie łatwiejsze. Intuicja podpowiada nam, że nie możemy zawsze w sposób trafny porównać uczniów o tej samej liczbie zdobytych punktów, gdyż jeden z nich mógł zdobyć punkty, rozwiązując zadania łatwe, drugi zaś – zadania trudne. W dużym skrócie można powiedzieć, iż właśnie z tego problemu narodziła się teoria odpowiedzi na zadanie testowe.

Zakładając prosty model kumulatywności (czyli proste dodawanie punktów), wypowiadamy się deterministycznie o umiejętnościach uczniów. Determinizm polega tutaj na założeniu, iż uczeń, który potrafi rozwiązać trudniejsze zadanie, będzie potrafił rozwiązać wszystkie zadania łatwiejsze z danego testu. Jeżeli chodzi o sytuacje kłopotliwe… no cóż – musimy udawać, iż one się nie zdarzają, lub traktować je jako coś, co w języku nauki nazywamy *anomalią*, czyli zachowaniem wykraczającym poza przyjęty model rzeczywistości. Deterministyczne założenie kumulatywności można przedstawić w prosty sposób w formie graficznej. Tak jak zostało to zrobione na rysunku 1.15.

Na rysunku przedstawiona została funkcja prawdopodobieństwa prawidłowej odpowiedzi na dane zadanie testowe przy założonym poziomie umiejętności, jaką przyjmujemy dla kumulatywnego, deterministycznego modelu skalowania (w skrócie: gdy po prostu dodajemy wyniki z poszczególnych zadań). Linia charakteryzuje nam prawdopodobieństwo poprawnej odpowiedzi na wybrane zadanie. Jak widać,



Rysunek 1.15. Deterministyczne założenie kumulatywności



Rysunek 1.16. Krzywa prawdopodobieństwa odpowiedzi na zadanie, która ma kształt spłaszczonej litery „S"

do pewnego poziomu umiejętności, np. reprezentowanego przez ucznia A, odpowiadający na zadanie ma zerowe prawdopodobieństwo poprawnej odpowiedzi. W pewnym punkcie wraz ze wzrostem umiejętności prawdopodobieństwo to zwiększa się raptownie do 1. Tak jak pokazane jest to na rysunku 1.15, uczeń A ma zerowe prawdopodobieństwo poprawnej odpowiedzi (nie ma możliwości na nie poprawnie odpowiedzieć), zaś prawdopodobieństwo poprawnej odpowiedzi dla ucznia B wynosi 1 (nie ma możliwości odpowiedzieć na nie błędnie).

Jeżeli głębiej się nad tym zastanowimy, intuicja podpowiada, że jest to mało realistyczne założenie. Zdrowy rozsądek podpowiada, iż uczeń o wysokim poziomie umiejętności powinien mieć wysokie prawdopodobieństwo odpowiedzi na dane zadanie, a uczeń o niskim poziomie umiejętności powinien mieć niską wartość prawdopodobieństwa. Jednak w żadnym wypadku prawdopodobieństwa te nie powinny równać się zero lub jeden. Zawsze możemy wyobrazić sobie sytuację, w której nawet bardzo dobry uczeń nie rozwiąże łatwego zadania, a stosunkowo słaby uczeń poprawnie zgadnie dobrą odpowiedź. Funkcja prawdopodobieństwa nie powinna zatem wyglądać jak schodki (tak jak na rysunku 1.15), lecz powinna być jakąś formą funkcji niemalejącej, gdzie uczeń o wysokich umiejętnościach ma wyższe prawdopodobieństwo poprawnej odpowiedzi niż uczeń o niskich umiejętnościach. Można na przykład przyjąć, iż prawdopodobieństwo odpowiedzi na dane zadanie powinna charakteryzować krzywa zbliżona kształtem do litery S. Tak jak zostało to przedstawione na rysunku 1.16.

Dobre zadanie powinno dobrze rozdzielać nam uczniów słabych i dobrych, czyli charakteryzować się niskim prawdopodobieństwem prawidłowego rozwiązania przez uczniów słabych i wysokim przez uczniów bardzo dobrych. Przy

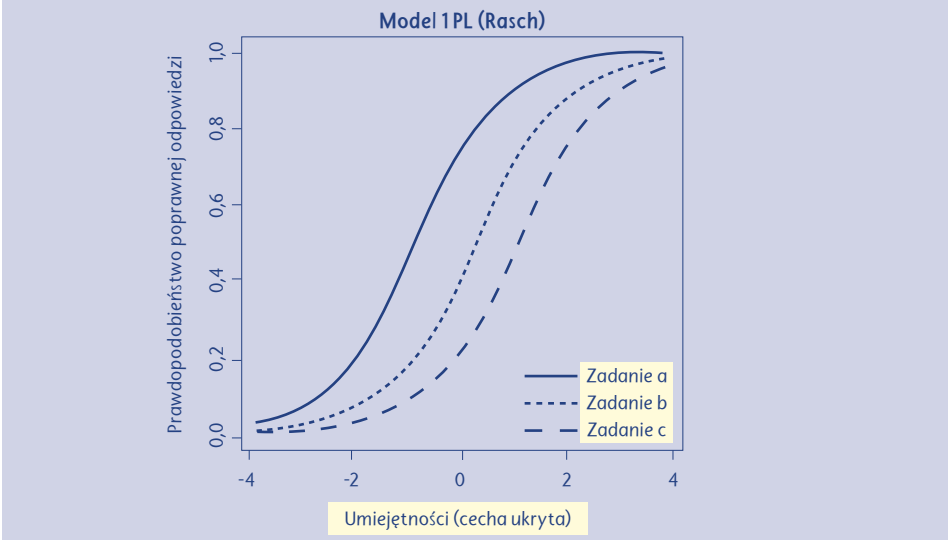
spis treści

czym dobrze by było, gdyby jak najmniejszy przedział umiejętności uczniów opisywany był za pomocą wartości prawdopodobieństwa bliskiej 0,5. Tak, aby fakt rozwiązania danego zadania jak najprecyzyjniej (choć nie z całkowitą pewnością) wskazywał nam, jak określić umiejętności ucznia.

Szczęśliwie składa się, że krzywą taką można opisać jako krzywą logistyczną. Szczęśliwie dlatego, że wzór na nią jest dość prosty, co dalej znacznie ułatwia cały szereg operacji matematycznych, w które nie będziemy tutaj wchodzić. Przytoczmy tylko wzór opisujący naszą krzywą:

$$f_i(\theta) = \frac{exp\,(\theta - \beta_i)}{1 + exp\,(\theta - \beta_i)} \qquad (5)$$

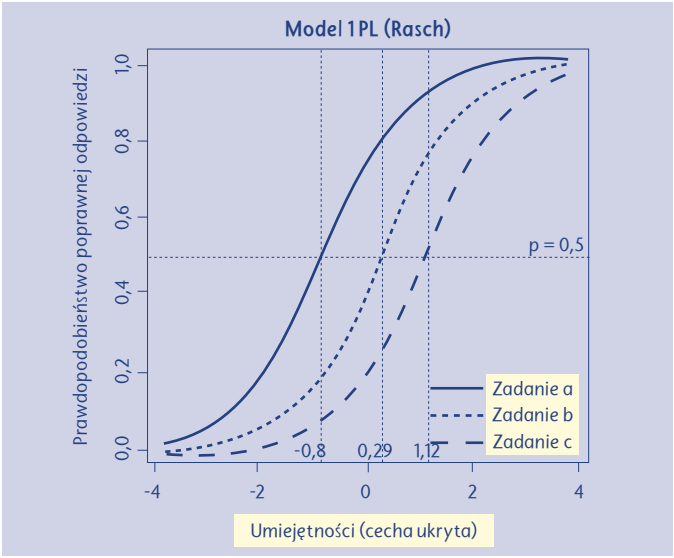
Wzór ten mówi tyle, że kontinuum umiejętności uczniów θ (teta) możemy opisać na podstawie funkcji $f_i(\theta)$ opisującej dane zadanie o indeksie *i* za pomocą wyrażenia $\frac{exp\,(\theta - \beta_i)}{1 + exp\,(\theta - \beta_i)}$, w którym wyrażenie β_i oznacza trudność zadania, zaś *exp* jest funkcją matematyczną, nadającą kształt krzywej[5]. Na rysunku 1.17 przedstawiono za pomocą omawianej funkcji trzy funkcje charakterystyczne dla trzech zadań. Wyrysowane za pomocą wzoru funkcji krzywe nazywamy krzywymi charakterystycznymi zadań. Dzięki nim możemy ustalić, jakie prawdopodobieństwo poprawnej odpowiedzi na zadanie mają uczniowie o różnych poziomach umiejętności. Krzywe charakterystyczne zadań różnią się jedynie poziomem trudności, czyli parametrem β_i, stąd też ogólna nazwa



Rysunek 1.17. Krzywe charakterystyczne trzech zadań o różnej trudności

[5] Funkcja *exp*(x) oznacza liczbę *e* podniesioną do potęgi *x*.

spis treści



Rysunek 1.18. Wyznaczanie trudności zadań na podstawie krzywych charakterystycznych

tego typu modeli: model jednoparametryczny nazywany w skrócie modelem 1PL. Łatwo zauważyć, iż zadanie **a** jest w tym układzie najłatwiejsze, zadanie **c** najtrudniejsze.

Warto podkreślić, iż istnieją dwa typy modelu jednoparametrycznego. Pierwszy z nich to model, w którym nachylenie dla wszystkich krzywych charakterystycznych zadań analizowanego testu równe jest 1. Drugi typ modelu jednoparametrycznego charakteryzuje się tym, iż nachylenie wszystkich krzywych charakterystycznych analizowanego testu jest takie samo dla każdego zadania, lecz wartość ta nie równa się 1. Wartość ta wyznaczana jest za każdym razem dla badanego testu i zależy od właściwości zadań, które na test się składają. Model o nachyleniu krzywych charakterystycznych dla wszystkich zadań równym 1 nazywany jest jednoparametrycznym modelem Rascha (lub w skrócie modelem Rascha). Model, w którym równe dla wszystkich zadań nachylenie jest szacowane dla danego testu, nazywany jest po prostu modelem jednoparametrycznym.

Zadania w tym przykładzie pokazują, że dla dowolnej wartości umiejętności prawdopodobieństwo prawidłowej odpowiedzi na zadanie trudniejsze jest niższe niż na zadanie łatwiejsze; i tak, jeżeli analizowalibyśmy ucznia, który na skali umiejętności posiada wartość 1, to prawdopodobieństwo odpowiedzi na zadanie *a* (czyli najłatwiejsze) będzie wynosiło około 0,9. Prawdopodobieństwo odpowiedzi na zadanie *b* – około 0,7, natomiast prawdopodobieństwo, iż uczeń rozwiąże poprawnie zadanie najtrudniejsze, czyli *c,* wynosi około 0,5.

Żeby podać charakterystykę liczbową zadania, czyli powiedzieć coś o zadaniu za pomocą jednej liczby, czyli wskaźnika trudności, musimy odwołać

się do pewnej arbitralnej konwencji. Otóż przyjęło się, że liczbową reprezentacją trudności zadania skalowanego w IRT będzie wartość umiejętności ucznia wtedy, gdy prawdopodobieństwo poprawnej odpowiedzi na zadanie wynosi dokładnie 0,5. Na rysunku 1.18 pozioma przerywana linia wyznacza prawdopodobieństwo poprawnej odpowiedzi na zadanie o wartości 0,5. Przecięcie linii prawdopodobieństwa z krzywą charakterystyczną zadania określa trudność tego zadania (zaznaczone pionowymi przerywanymi liniami). Jak widać, zadanie *a* ma trudność równą -0,84, zadanie *b* 0,29, a zadanie *c* jest najtrudniejsze i charakteryzuje się trudnością 1,12.[6]

Poziom umiejętności uczniów 1 PL (model Rascha)

Celem skalowania jest liczbowe określenie poziomu umiejętności ucznia na podstawie uzyskanych odpowiedzi na zadania. Wyznaczenie wartości liczbowej reprezentującej poziom umiejętności odbywa się poprzez skomplikowane procedury matematyczne. Obrazowo można powiedzieć, iż ustalanie poziomu umiejętność polega na przypisaniu uzyskanego empirycznie wyniku do najbardziej prawdopodobnej i najlepiej pasującej dla szacowanej skali liczby.

Połączenie prawdopodobieństwa poprawnej odpowiedzi na zadanie oraz wartości cechy ukrytej pod względem matematycznym nie jest bardzo skomplikowane. Możemy poznany w poprzednim podpunkcie wzór (5) doprowadzić do następującej postaci:

$$P(X_{ij} = 1 \mid \theta_i; \beta_j) = \frac{exp(\theta_j - \beta_i)}{1 + exp(\theta_j - \beta_i)} \qquad (6)$$

Wzór ten mówi nam tyle, że prawdopodobieństwo zdarzenia *X* dla *i*-tego ucznia przy *j*-tym zadaniu oznaczone jako 1 (czyli konwencjonalnie zdarzenie, które uznajemy za sukces, a w naszym wypadku za prawidłowe rozwiązanie zadania) dla danej wartości umiejętności ucznia (θ_i) oraz danego parametru trudności zadania *j* (β_j) można opisać na podstawie znanej nam funkcji $\frac{exp(\theta_j - \beta_i)}{1 + exp(\theta_j - \beta_i)}$.

W tabeli 1.8 przedstawiono kilka możliwych kombinacji poziomu umiejętności ucznia, trudności zadania, które rozwiązuje, oraz prawdopodobieństwa prawidłowego rozwiązania tego zadania dla ucznia charakteryzującego się daną umiejętnością obliczoną na podstawie funkcji, którą przytaczaliśmy powyżej.

[6] Skala trudności teoretycznie zawiera się w przedziale od minus nieskończoności do plus nieskończoności. W zdecydowanej większości wypadków trudności pytań zawierają się w przedziale od -4 do 4. Metryka skali wynika bezpośrednio z przekształceń matematycznych stosowanych podczas modelowania.

spis treści

Tabela 1.8. Umiejętności ucznia, trudność zadania oraz wynikające z nich wartości prawdopodobieństwa poprawnej odpowiedzi na zadanie

Poziom umiejętności ucznia	Trudność zadania	Prawdopodobieństwo prawidłowej odpowiedzi
-1	-1	0,5
0	0	0,5
2	2	0,5
-1	0	0,27
0	1	0,27
1	2	0,27
-1	-2	0,73
0	-1	0,73
1	0	0,73
-1	1	0,12
0	2	0,12
1	3	0,12
-1	-3	0,88
0	-2	0,88
1	-1	0,88

Jak łatwo zauważyć, tam, gdzie trudność zadania ma taką samą wartość jak poziom umiejętności ucznia, prawdopodobieństwo poprawnej odpowiedzi wynosi 0,5. Oczywiście, łączy to nas bezpośrednio z definicją trudności zadania podaną wcześniej.

Ważne jest też, iż możemy obliczyć łączne prawdopodobieństwo, czyli prawdopodobieństwo uzyskania danego wyniku przez ucznia o danym poziomie umiejętności. Na przykład załóżmy, że test składa się z trzech zadań. Pierwsze o trudności -1, drugie o trudności 0, trzecie o trudności 1. Prawdopodobieństwo, iż uczeń odpowie dobrze na pierwsze zadanie, jest iloczynem prawdopodobieństwa poprawnej odpowiedzi na pierwsze zadanie i prawdopodobieństwa odpowiedzi błędnej na zadanie drugie i trzecie (prawdopodobieństwo błędnej odpowiedzi = 1 – prawdopodobieństwo poprawnej odpowiedzi). Prawdopodobieństwo odpowiedzi na pierwsze dwa zadania to iloczyn prawdopodobieństwa poprawnej odpowiedzi na dwa pierwsze zadania oraz prawdopodobieństwa błędnej odpowiedzi na trzecie zadanie. Prawdopodobieństwo odpowiedzi na trzy zadania to iloczyn prawdopodobieństw poprawnych odpowiedzi dla tych trzech zadań. Zostało to przedstawione w tabeli 1.9.

Nie wgłębiając się zbytnio w podstawy statystyczne, określanie umiejętności poszczególnych uczniów w tym modelu polega na wyliczaniu na podstawie uzyskanych wyników tego, jakie byłyby najbardziej wiarygodne wartości cechy ukrytej ucznia, aby ten uzyskał dany wynik. Oczywiście nie mamy danych trudności zadań oraz umiejętności uczniów, trzeba je dopiero wyliczyć. Stosuje się do tego takie procedury, które „szukają" jednocześnie trudności zadań oraz wartości umiejętności uczniów, które byłyby najbardziej prawdopodobne dla posiadanych danych. Poszukuje się takich parametrów, czyli nieznanych zmiennych w równaniu, które

Tabela 1.9. Prawdopodobieństwo trzech poprawnych odpowiedzi

Poziom umiejętności ucznia	Prawdopodobieństwo poprawnej odpowiedzi na zadanie			Łączne prawdopodobieństwo odpowiedzi na trzy zadania		
	Zad. 1 (-1)	Zad. 2 (0)	Zad. 3 (1)	Poprawne tylko: Zad. 1	Poprawne tylko: Zad. 1, Zad. 2	Poprawne: Zad. 1, Zad. 2, Zad. 3
-3	0,12	0,05	0,02	0,11172	0,00588	0,0001
-2	0,27	0,12	0,05	0,22572	0,03078	0,0016
-1	0,50	0,27	0,12	0,3212	0,1188	0,0162
0	0,73	0,50	0,27	0,26645	0,26645	0,0986
1	0,88	0,73	0,50	0,1188	0,3212	0,3212
2	0,95	0,88	0,73	0,03078	0,22572	0,6103
3	0,98	0,95	0,88	0,00588	0,11172	0,8193

maksymalizowałyby w modelu wiarygodność tego, co zdarzyło się w danych. Stąd nazwa tego typu szacowania: estymacja metodą największej wiarygodności (*Maximum Likelihood Estimation*).

Cała sztuka oszacowania poziomu umiejętności ucznia polega na wyznaczeniu funkcji wiarygodności uzyskania określonych wyników ze względu na parametr modelu oznaczający poziom umiejętności na podstawie zastanych danych. Jeżeli uda się to zrobić, oszacowanie wartości umiejętności danego ucznia jest banalnie proste. Odpowiedź znajduje się poprzez porównanie wiarygodności uzyskania danego wyniku z wiarygodnością uzyskania takiej odpowiedzi dla różnych wartości umiejętności. Polega to na odnalezieniu wartości cechy ukrytej, która daje największą wartość wiarygodności uzyskania takiego, a nie innego wyniku.

Ideę szacowania poziomu umiejętności można przedstawić za pomocą tabeli 1.9. Załóżmy, iż chcemy oszacować poziom umiejętności ucznia, który poprawnie rozwiązał pierwsze dwa zadania, natomiast dla zadania trzeciego podał błędną odpowiedź. Patrzymy zatem na przedostatnią kolumnę (Łączne prawdopodobieństwo odpowiedzi na trzy zadania; Poprawne tylko: Zad. 1, Zad. 2). Szukamy zdarzenia najbardziej prawdopodobnego (precyzyjniej: o najwyższej wiarygodności). Najwyższe łączne prawdopodobieństwo dla zdarzenia, w którym uczeń rozwiązał dwa pierwsze zadania poprawnie, a trzecie niepoprawnie, wynosi 0,3212. Jest to wartość wyznaczona dla poziomu umiejętności 1 (pierwsza kolumna). Uczeń, który odpowiedział na dwa pierwsze zadania poprawnie, ma zatem poziom umiejętności najprawdopodobniej równy 1. Podobną procedurę możemy oczywiście stosować również dla innych par odpowiedzi.

spis treści

Trzeba raz jeszcze podkreślić, iż to, co przedstawiliśmy w tym punkcie, jest bardzo uproszczonym obrazem szacowania umiejętności uczniów za pomocą skalowania Rascha. I w żadnym wypadku nie można traktować go jako pełnoprawnego opisu, jest to raczej przedstawienie pewnego sposobu myślenia o wyniku i podkreślenie jego probabilistycznego charakteru. Nie jest naszym zadaniem wykładać w tej publikacji podstaw zaawansowanej teorii estymacji.

Wyznaczanie wartości liczbowych parametrów trudności zadań lub wartości umiejętności pozostawmy specjalistom i ich komputerom. Ważne, że zrozumieliśmy logikę tego sposobu wyznaczania wartości liczbowych, wartości, które, co warto podkreślić, jako pochodne prawdopodobieństwa można umieszczać przynajmniej na skali przedziałowej.

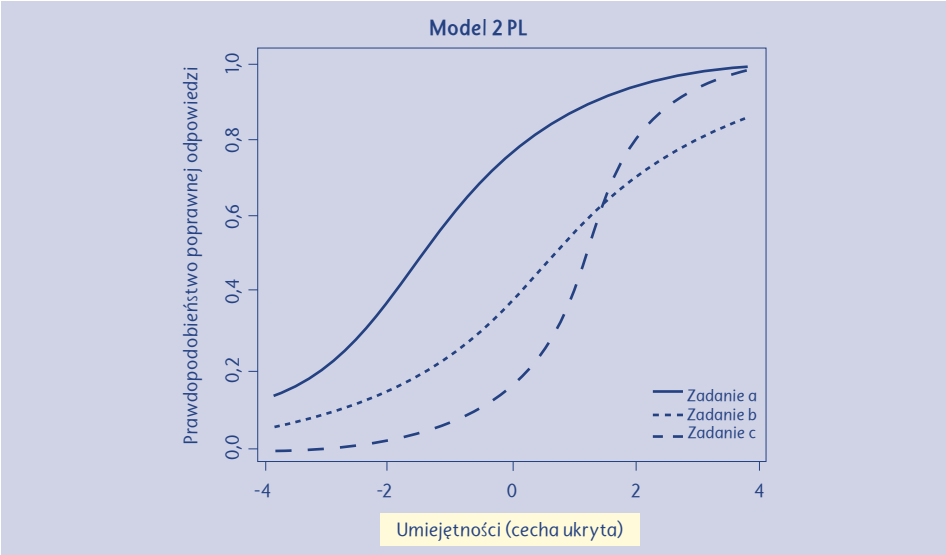
Warto zwrócić uwagę na skalę umiejętności (cechy ukrytej). Teoretycznie wartości parametrów trudności mogą zawierać się w przedziale $(-\infty; +\infty)$, w praktyce okazuje się, iż przybierają one zazwyczaj wartości od -4 do +4. Przy czym oczywiście 0 nie oznacza tutaj tego, że uczeń nie posiada żadnych umiejętności. Skala taka jest stosowana ze względu na prostotę matematyczną. Często jednak w publikacjach wyników przelicza się tę skalę na bardziej przyjazną dla oka, w której średnia wynosi 500, a odchylenie standardowe 100. Tak jest na przykład w badaniach PISA, PIRLS czy TIMSS. Warto zwrócić uwagę, iż nie są to punkty w klasycznym znaczeniu, które rozpowszechnione jest przez klasyczną teorię testu. Mimo iż mówi się, iż dany kraj uzyskał średnią 576 punktów, to wynik ten nie jest sumą dobrych odpowiedzi, tylko matematycznym przekształceniem skali IRT.

Model dwuparametryczny

Na rysunku 1.17 i 1.18 zakładaliśmy, że nasze zadania różnią się jedynie trudnością, czyli przesunięciem na osi umiejętności. W rzeczywistości tak być nie musi. Zadania mogą różnić się również nachyleniem krzywej. Żeby uwzględnić takie zjawisko w modelu matematycznym, wystarczy do wcześniej prezentowanego równania dodać jeszcze jeden parametr tak, że:

$$f_i(\theta) = \frac{exp\left(a_i(\theta - \beta_i)\right)}{1 + exp\left(a_i(\theta - \beta_i)\right)} \qquad (6)$$

Parametr a_i nazywamy dyskryminacją i mówi nam on, jak bardzo krzywe prawdopodobieństwa dla danych zadań są strome. Jest to parametr, który w klasycznej teorii testu odpowiada parametrowi mocy różnicującej i w taki sposób można go też interpretować (patrz klasyczna teoria testu). Na rysunku 1.19 przedstawiono trzy zadania o różnych parametrach trudności i różnych parametrach dyskryminacji. Tak jak poprzednio, zadanie *a* jest najłatwiejsze, zadanie *b* jest trudniejsze od zadania *a*, najtrudniejszym zadaniem jest zadanie *c* (podobnie jak w modelu jednoparametrycznym łatwość zadania określamy dla prawdopodobieństwa



Rysunek 1.19. Krzywe charakterystyczne trzech zadań o różnej trudności i różnej dyskryminacji

równego 0,5). Nie jest to regułą, lecz w tym konkretnym przypadku okazało się, że najtrudniejsze zadanie ma też najwyższy parametr mocy dyskryminacyjnej, czyli krzywa charakterystyczna jest najbardziej stroma. Natomiast najłatwiejsze zadanie *a* okazało się mieć krzywą charakterystyczną bardziej stromą niż zadanie *b* (zadanie środkowe pod względem trudności). Najmniej stroma spośród przedstawionych krzywych jest krzywa charakterystyczna zadania *b*, dlatego też zadanie to ma najniższą moc dyskryminacyjną.

Użycie modelu dwuparametrycznego jest niezwykle przydatne dla analizy poszczególnych zadań i dla całego testu. Mając wskaźnik wyznaczający trudność zadania i moc dyskryminacyjną, możemy oceniać jakość pomiarową zadań. Dla ułatwienia interpretacji podane zostaną słowne oceny parametru dyskryminacji zaproponowane przez Franka Bakera (Baker 2001):

Tabela 1.10. Kategoryzacja wartości dyskryminacji

Ocena dyskryminacji	Przedział wartości
Brak	0
Bardzo niska	0,01-0,34
Niska	0,35-0,64
Umiarkowana	0,65-1,34
Wysoka	1,35-1,69
Bardzo wysoka	> 1,70
Doskonała (teoretyczna)	+ nieskończoność

Można próbować stworzyć podobną klasyfikację dla parametru trudności. Taka klasyfikacja została przedstawiona w tabeli 1.11. Jest to jednak o wiele bardziej arbitralny podział niż poprzedni.

Tabela 1.11. Kategoryzacja wartości parametru trudności

Ocena trudności	Przedział wartości
Bardzo łatwe	< (-1,5)
Łatwe	(-1,5) do (-0,5)
Średnie	od (-0,5) do 0,5
Trudne	od 0,5 do 1,5
Bardzo trudne	> 1,5

Wskaźnik dyskryminacji oparty na teorii odpowiedzi na zadanie testowe jest znacznie lepszą miarą niż analogiczny do niego wskaźnik mocy różnicującej zbudowany na podstawie klasycznej teorii testu, który ma zasadniczą wadę polegającą na zaniżaniu mocy dyskryminacyjnej dla zadań o bardzo dużej lub bardzo małej łatwości (lub trudności).

Warto zauważyć również, iż uwzględnienie mocy dyskryminacyjnej zadań nastręcza pewnych problemów. Może dojść do takiej sytuacji, iż krzywe charakterystyczne będą się przecinały, tak jak krzywe charakterystyczne zadania *b* i zadania *c*. Umieszczenie tych dwóch zadań w jednym teście jest problematyczne. Okazuje się bowiem, iż prawdopodobieństwo uzyskania poprawnej odpowiedzi na zadanie *b* było wyższe niż na zadanie *c*, ale tylko wśród grupy uczniów, których poziom umiejętności nie przekraczał około 1,75. Po przekroczeniu tego progu sytuacja się odwraca. Zadania mają zatem różną trudność względem siebie w zależności od poziomu umiejętności ucznia, który je rozwiązuje. Jest to nie tylko niepokojące ze względów teoretycznych, lecz także ze względów matematycznych, w które nie będziemy się zagłębiać.

W praktyce są dwie strategie rozwiązywania problemu przecinających się krzywych. Po pierwsze, wyrzuca się zadania, których krzywe charakterystyczne się przecinają, a przynajmniej jedno z nich. Po drugie, poprzestaje się na matematycznym opisie rzeczywistości, takim jak przedstawiony w pierwszym prezentowanym wzorze, czyli nie pozwala się na to, aby zadania różniły się pod względem dyskryminacji – stosuje się model z jednym parametrem. Jeżeli jednak uda się dobrać zadania w taki sposób, by nie przecinały się ich krzywe, często oszacowania umiejętności uczniów będą znacznie bardziej dokładne, ponieważ model, dzięki któremu skalujemy umiejętności, będzie lepiej dopasowany niż prostsze modele (np. model Rascha).

Należy wspomnieć, iż istnieje jeszcze kilka alternatywnych modeli IRT, na przykład model, w którym uwzględnia się dodatkowy parametr, czyli parametr korygujący możliwość zgadywania c_i. Wtedy mówimy o modelu trzyparametrycznym (3PL).

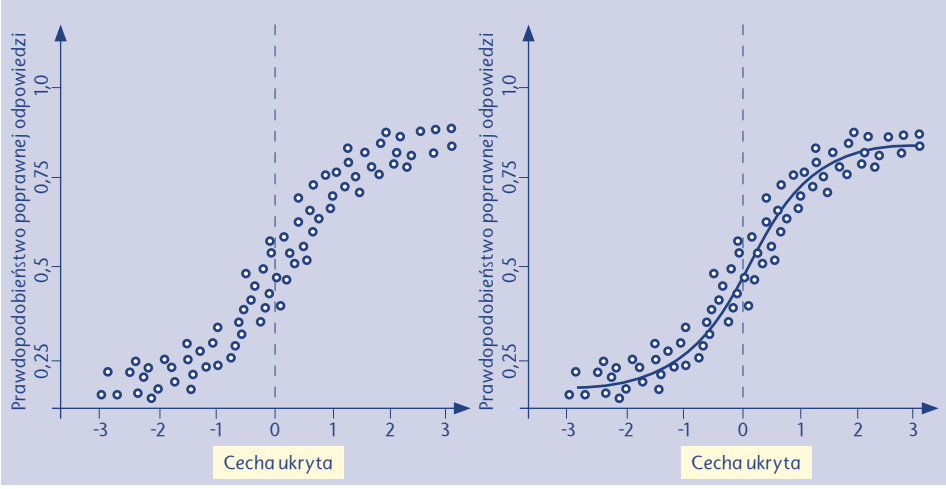
Są też modele czteroparametryczne (4PL), w których uwzględnia się parametr pomyłki. Są to już wyrafinowane modele, które rzadko stosuje się w kontekście szkolnym do skalowania – ze względu na ich złożoność i różnorodne problemy teoretyczne i obliczeniowe. W kontekście edukacyjnym pod względem popularności bezkonkurencyjnym pozostaje model jednoparametryczny, najczęściej używany w postaci modelu Rascha.

Dopasowanie krzywej charakterystycznej do modelu

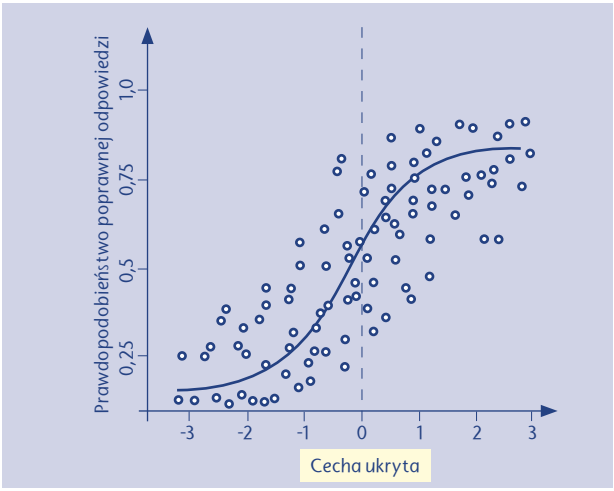
Warto podkreślić, iż krzywa charakterystyczna zadania jest matematycznym dopasowaniem modelu do istniejących danych. Na podstawie wszystkich zadań w danym teście, dzięki skomplikowanym procedurom estymacji (zostawmy to komputerom) szacowane jest prawdopodobieństwo poprawnej odpowiedzi na dane zadanie (oś pionowa – rysunek 1.17 i 1.18) w stosunku do wartości cechy ukrytej (oś pozioma – rysunek 1.17 i 1.18).

Na rysunku 1.20 (z prawej strony) pokazano przykład bardzo dobrego dopasowania. Widać wyraźnie, iż rozmieszczenie punktów wskazujących na prawdopodobieństwo poprawnej odpowiedzi dla danej cechy ukrytej jest bardzo dokładnie dopasowane do krzywej. W praktyce może dochodzić również do innych sytuacji: dopasowanie może być słabe. Taką sytuację ilustruje rysunek 1.21.

Punkty symbolizujące prawdopodobieństwo uzyskania poprawnej odpowiedzi na zadanie dla określonych wartości cechy ukrytej mogą nie pokrywać się dokładnie z krzywą teoretyczną, mimo iż ta jest najlepszym z możliwych dopasowań parametrycznych dla danego modelu. W takich sytuacjach często zmuszeni jesteśmy odrzucić takie zadanie, by wyznaczyć wartość cechy ukrytej dla ucz-



Rysunek 1.20. Dopasowanie krzywej charakterystycznej – przykład dobrego dopasowania. Przerywaną linią zaznaczono środek skali



Rysunek 1.21. Dopasowanie krzywej charakterystycznej – przykład słabego dopasownia. Przerywaną linią zaznaczono środek skali

nia rozwiązującego cały test. Pozostawienie zadania słabo dopasowanego może prowadzić do poważnych błędów w procesie szacowania parametrów modelu, w tym parametrów określających trudności pozostałych zadań, oraz parametrów określających umiejętności poszczególnych testowanych uczniów.

Są dwa sposoby oceniania dopasowania zadań. Pierwszy z nich polega na zastosowaniu statystyki chi-kwadrat, która często podawana jest przez programy statystyczne szacujące parametry zadań. Im większa wartość tej statystyki, tym zadanie jest lepiej dopasowane. Część programów statystycznych oferuje również test istotności statystycznej oparty na tej statystyce. Jednak właściwości tego testu w zastosowaniu dla modeli IRT nie zostały jeszcze dobrze zbadane i należy je traktować z daleko idącą ostrożnością.

Drugi sposób polega na porównywaniu dopasowania uzyskanego za pomocą modelu parametrycznego z dopasowaniem na podstawie modelu nieparametrycznego, czyli takiego, który nie używa krzywej logistycznej, lecz najlepiej dopasowanej ze wszystkich możliwych krzywych (więcej o dopasowaniu nieparametrycznym będziemy pisali w kolejnych podrozdziałach).

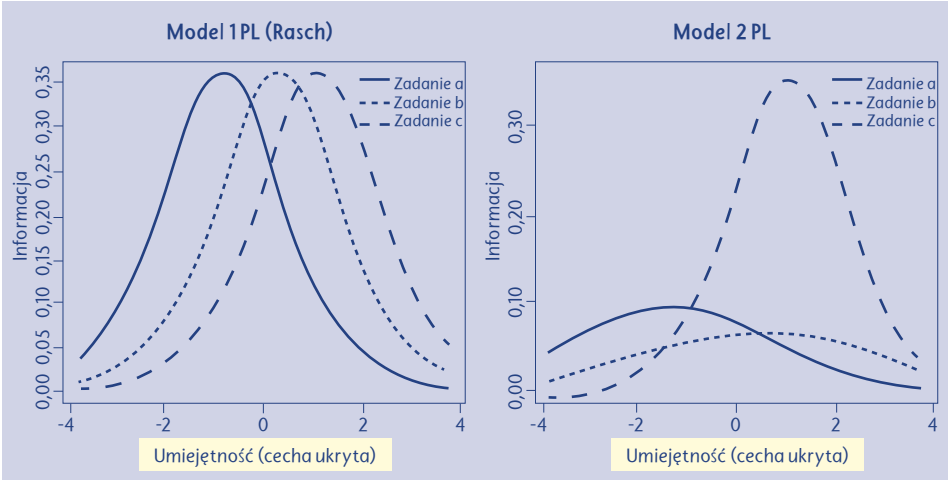
Niestety, w obu wypadkach analizujący zadania skazani są na na arbitralność. Sami muszą określić, co to znaczy mała wartość testu chi-kwadrat, lub za pomocą rysunku stwierdzić, czy dopasowanie parametryczne jest wystarczająco różne od dopasowania nieparametrycznego, po to, by podjąć decyzję o pozostawieniu bądź odrzuceniu danego zadania.

Krzywa informacyjna zadania

Informacja wedle klasycznej definicji to funkcja precyzji, z jaką można estymować parametry modelu statystycznego. Jeżeli szacujemy parametr z dużą

precyzją, to uzyskujemy o nim precyzyjną informację, inaczej mówiąc, możemy powiedzieć, iż uzyskujemy o nim dużo informacji. Jeżeli szacujemy parametr z małą precyzją, uzyskujemy o nim mało informacji. Używając metod statystycznych, jesteśmy w stanie obliczyć ilość informacji, którą uzyskujemy za pomocą konkretnego zadania, dla każdej wartości cechy ukrytej. Graficzna reprezentacja ilości informacji dla różnych wartości cechy ukrytej w teorii odpowiedzi na zadania testowe reprezentowana jest przez krzywą informacyjną zadania.

Na rysunku 1.22 po lewej stronie przedstawiono krzywe charakterystyczne trzech zadań o równym parametrze dyskryminacji, ale różnej trudności (są to krzywe informacyjne dla zadań omawianych w poprzednich podrozdziałach). Po prawej stronie rysunku 1.22 przedstawiono krzywe informacyjne dla zadań o różnej trudności i różnej wartości parametru dyskryminacji (krzywe dla zadań wykorzystanych w haśle „Model dwuparametryczny" patrz rysunek 1.19). Pole pod krzywą informacyjną określa, na ile udaje się zredukować niepewność w pomiarze umiejętności uczniów, a zatem, na ile uda się uzyskać wiarygodną informację o umiejętnościach uczniów w danych przedziałach. Jak widać, im wyższy parametr dyskryminacji, tym informacja staje się bardziej precyzyjna dla określonego przedziału (pole pod krzywą informacyjną jest większe). I tak zadanie *c* jest trudne oraz ma wysoką dyskryminację, dlatego precyzyjnie mierzy wysokie wartości umiejętności. Inaczej mówiąc, zadanie to pozwala precyzyjnie oceniać umiejętności uczniów testowanych na danym poziomie. Zadanie *b* jest średnie pod względem trudności oraz charakteryzuje się słabą dyskryminacją, co ma wpływ na kształt krzywej informacyjnej. Jak widać, to zadanie daje nam pewną ilość informacji (określaną przez pole pod krzywą) dla szerokiego przedziału umiejętności, lecz w żadnym określonym przedziale umiejętności nie jest



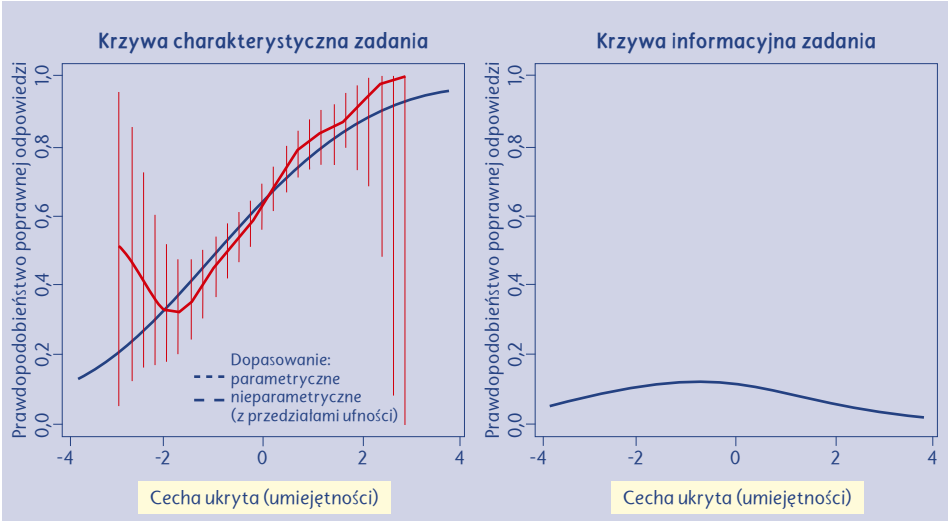
Rysunek 1.22. Krzywe informacyjne dla modelu z jednym parametrem (lewa strona) oraz dla modelu z dwoma parametrami (prawa strona)

ona zbyt duża. Zatem takie zadanie nie przyniesie nam istotnej wiedzy o umiejętnościach ucznia.

Dopasowanie nieparametryczne

Aby przyjrzeć się dokładnie danemu zadaniu, warto przyjrzeć się dokładnemu dopasowaniu krzywej charakterystycznej za pomocą modelu nieparametrycznego. Dopasowanie nieparametryczne zaznaczone jest czerwoną linią i precyzyjnie określa, jakie jest prawdopodobieństwo poprawnej odpowiedzi dla ucznia o danym poziomie umiejętności. Model parametryczny „nagina" rzeczywistość do zadanych parametrów, przez co nie pokazuje wszystkich aspektów analizowanej zależności (prawdopodobieństwa poprawnej odpowiedzi względem umiejętności). Model nieparametryczny opisuje po prostu idealnie zależności w danych, bez przyjmowania jakichkolwiek założeń o ich kształcie. Na rysunku 1.23 poniżej przedstawiono przykład dopasowania parametrycznego i nieparametrycznego. Jak widać, dopasowanie parametryczne nie pokaże, iż dla niskich wartości umiejętności prawdopodobieństwo poprawnej odpowiedzi jest podejrzanie wysokie.

Dodatkowo w przedziale (-3; -2) prawdopodobieństwo poprawnej odpowiedzi spada wraz ze wzrostem umiejętności (sic!). Oznacza to, iż bardzo słabi uczniowie zgadują, jaka jest prawidłowa odpowiedź, i mogą na tym zyskać, w porównaniu z uczniami słabymi. Może to być spowodowane na przykład niedociągnięciami związanymi z „martwymi dystraktorami". Dodatkowo na rysunku dopasowania nieparametrycznego pokazano błąd oszacowania – pionowe



Rysunek 1.23. Krzywe charakterystyczne odpowiedzi na zadania (dopasowanie parametryczne oraz nieparametryczne) oraz krzywa informacyjna zadania

linie. Im linie dłuższe, tym oszacowanie dla danego obszaru umiejętności gorsze. Koresponduje to oczywiście z krzywą informacyjną zadania. W przedziale, gdzie pionowe linie są krótkie, krzywa informacji przyjmuje wyższe wartości; tam, gdzie linie są dłuższe – mniejsze wartości, czyli informacja jest mniej istotna (pole pod krzywą jest mniejsze).

Analiza testu jako całości

Do tej pory elementy teorii odpowiedzi na zadanie testowe rozważane były przede wszystkim w kontekście analizy pojedynczych zadań. Dzięki krzywej charakterystycznej zadania, parametrom trudności i dyskryminacji, krzywej informacyjnej i dopasowaniu nieparametrycznemu można precyzyjnie ocenić jakość pomiarową zadania i zadecydować o tym, czy zadanie spełnia oczekiwania konstruktorów testu. W tej części spojrzymy na test jako na całość. Postępujemy w ten sposób zgodnie z logiką analizy prawdziwych testów, gdzie najpierw analizuje się poszczególne zadania, sprawdza, czy spełniają one podstawowe kryteria (założone przez testujących), sprawdza się, jak związane są z całym testem, wychwytuje się zadania z wyraźnymi błędami, a następnie patrzy się na właściwości całego testu.

Krzywa charakterystyczna testu

W poprzednich punktach szczegółowo omawialiśmy krzywą charakterystyczną zadania. Analogiczną konstrukcją w wypadku analizy całego testu jest krzywa charakterystyczna testu. Zarówno kształt krzywej, jak i jej interpretacja są analogiczne, jak w wypadku krzywej charakterystycznej zadania. Istotną zmianą jest zmiana osi pionowej. O ile w wypadku krzywej charakterystycznej zadania na osi pionowej dane było prawdopodobieństwo poprawnej odpowiedzi, to w tym wypadku oś pionowa charakteryzowana jest przez *wynik prawdziwy* ucznia. Pojęcie „wynik prawdziwy" (*true score*) nie jest w tym kontekście związane z klasyczną teorią testu (tożsamość nazwy jest trochę myląca, lecz większość autorów posługuje się takim nazewnictwem).

Gdy chcemy poznać wyniki sumaryczne dla danego testu w prosty sposób, dla każdego ucznia dodajemy punkty z poszczególnych zadań. Zakładając, iż zadania punktujemy w skali 0-1, suma punktów odpowiada liczbie poprawnie rozwiązanych zadań. Natomiast *wynik prawdziwy* w wypadku analizy całego testu zdefiniowany został w teorii odpowiedzi na zadanie testowe, jako suma prawdopodobieństwa poprawnych odpowiedzi na poszczególne zadania. Załóżmy, że mamy 6 zadań. Uczeń odpowiedział poprawnie na 5 pierwszych zadań, na ostatnie (zadanie najtrudniejsze) udzielił odpowiedzi niepoprawnej. Jego sumaryczny wynik to suma punktów, w tym wypadku 5. *Wynik prawdziwy*, wedle definicji teorii odpowiedzi na zadania testowe, to suma prawdopodobieństwa wszystkich

Tabela 1.12. Wynik sumaryczny i *wynik prawdziwy*

Zadanie	Punkty	Prawdopodobieństwo poprawnej odpowiedzi
1	1	0,99
2	1	0,94
3	1	0,92
4	1	0,70
5	1	0,65
6	0	0,40
	Wynik sumaryczny	*Wynik prawdziwy*
Suma	5	4,6

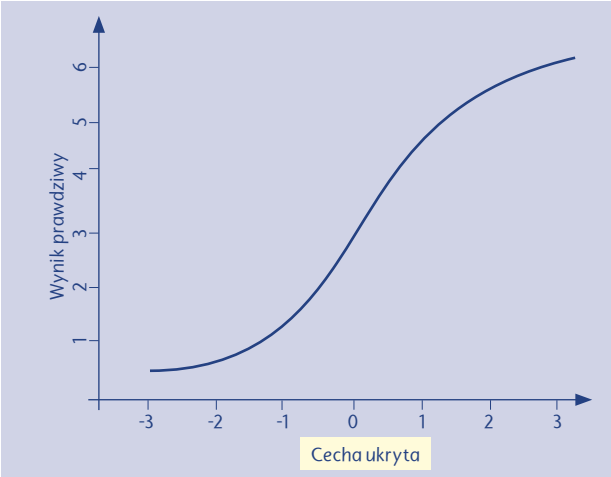
poprawnych odpowiedzi (szacowanych dla ucznia o danej umiejętności). Omawiany przykład przedstawiony został w tabeli 1.12.

Jak widać, sumaryczny wynik ucznia wynosi 5, podczas gdy *wynik prawdziwy* w tym konkretnym wypadku wynosi 4,6. *Wynik prawdziwy* w teorii odpowiedzi na zadanie testowe ma łatwą interpretację. Należy zauważyć tylko, iż prawdopodobieństwa poprawnej odpowiedzi na dane zadania są takie same dla wszystkich uczniów o tym samym poziomie umiejętności. Załóżmy, iż uczeń, który odpowiadał na 6 analizowanych zadań, charakteryzował się poziomem umiejętności równym 1. Teoretyczny – a zarazem najbardziej prawdopodobny – wynik, jaki powinni otrzymać wszyscy uczniowie o poziomie umiejętności 1, wynosi 4,6. *Wynik prawdziwy* w teorii odpowiedzi na zadanie testowe jest zatem przewidywanym wynikiem punktowym dla uczniów o danym poziomie umiejętności.

Krzywa charakterystyczna testu jest funkcją, która „przekłada" wartości estymowanej cechy umiejętności na *wyniki prawdziwe*. Przykładowa krzywa charakterystyczna testu przedstawiona została na rysunku 1.24.

Krzywa charakterystyczna testu może przybierać nieregularne kształty, zbliżone czasem do krzywych uzyskanych dzięki dopasowaniu nieparametrycznemu. Analizując jej kształt, możemy wypowiadać się o trudności całego testu, patrząc, jak środek krzywej przesunięty jest w stosunku do skali cechy ukrytej. Możemy analizować również, jak dobrze test różnicuje uczniów o różnych poziomach umiejętności, patrząc na nachylenie krzywej. O teście opisanym przez krzywą charakterystyczną z rysunku 1.24 możemy powiedzieć, iż najsilniej różnicuje on uczniów o wartości cechy ukrytej w okolicach zera (krzywa wtedy jest najbardziej stroma), słabo natomiast różnicuje uczniów o bardzo wysokiej wartości cechy ukrytej (umiejętności) oraz bardzo niskich wartościach (krzywa w tym punkcie jest najbardziej płaska).

Dzięki krzywej charakterystycznej testu możemy również zrównywać testy lub poszczególne ich formy, wiedząc jedynie, jakie wartości umiejętności (cechy ukrytej) wiążą się z punktowymi *wynikami prawdziwymi* (co zapewnia nam



Rysunek 1.24.
Krzywa charakterystyczna testu

krzywa charakterystyczna testu) dla dwóch testów. Dzięki prostemu przełożeniu wyników jednego testu na wartości cechy ukrytej, a następnie wyrażeniu wartości cechy ukrytej za pomocą *wyników prawdziwych* drugiego testu, uzyskujemy wyniki dla dwóch testów wyrażone na skali jednego wybranego testu.

Zrównywanie takie jest możliwe, kiedy szacowana wartość cechy ukrytej dla obydwu testów wyrażona jest w tej samej skali. Możliwe jest to wyłącznie w trzech sytuacjach:

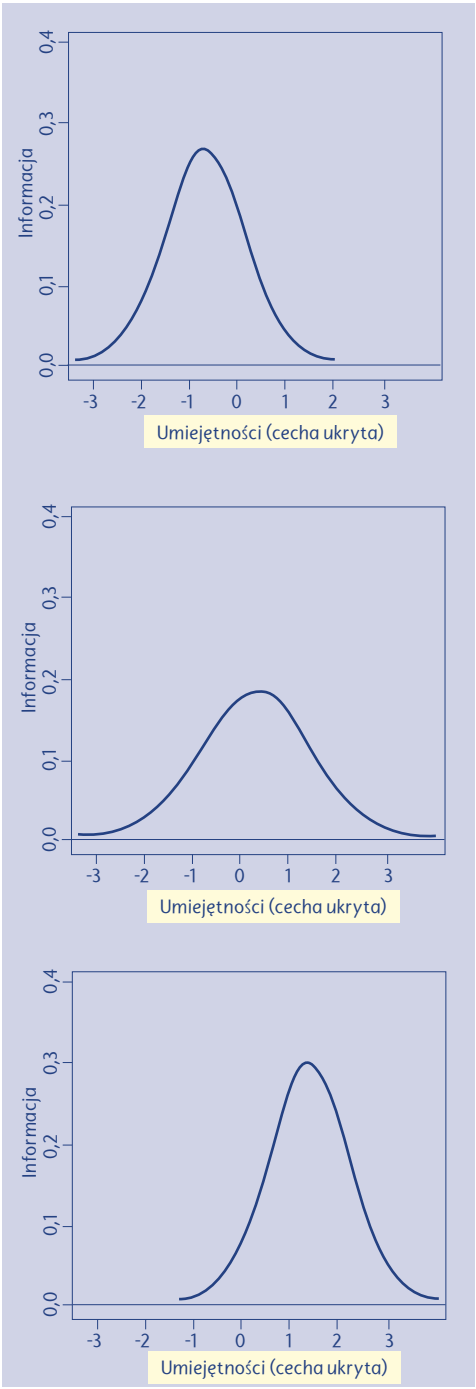
- gdy obydwa testy rozwiązywane były przez tych samych uczniów lub część uczniów rozwiązywała obydwa testy
- gdy w obydwu testach znajdowała się pewna liczba zadań wspólnych
- gdy testowaliśmy dwie losowo dobrane grupy uczniów z tej samej populacji (to rozwiązanie daje najmniej precyzyjne zrównywanie)

W innym wypadku procedury estymacyjne nie dają nam gwarancji uzyskania takiej samej metryki dla szacowanych wartości cechy ukrytej (umiejętności) dla dwóch osobno skalowanych testów. Problem ten rozwinięty zostanie szczegółowiej w ostatniej części tego rozdziału.

Krzywa informacyjna testu

Krzywą informacyjną można wyznaczyć nie tylko dla jednego zadania, lecz także dla całego testu. Wtedy krzywa informacyjna pokazuje, w jakich obszarach test niesie najwięcej informacji (podobnie jak wcześniej pole pod krzywą informacyjną zadania). Innymi słowy, możemy się dowiedzieć, dla jakiego przedziału umiejętności dane narzędzie pomiarowe (test) dawało będzie najbardziej precyzyjne oszacowania cechy ukrytej (umiejętności).

Krzywa informacyjna testu jest sumą informacji wszystkich zadań. Oznacza to, iż większa liczba zadań umieszczonych w teście skutkować będzie zazwyczaj



Rysunek 1.25. Krzywe informacyjne trzech testów

większą ilością informacji. Ale nie tylko – ilość informacji, jakie uzyskamy na podstawie danego testu, zależała będzie również od jakości pomiarowej zadań (a przede wszystkim parametru dyskryminacji tych zadań). Na rysunku 1.25 przedstawiono trzy różne krzywe informacyjne dla trzech różnych testów.

Na rysunku 1.25 możemy dostrzec, iż trzy testy dają podobną ilości informacji (mają podobne pole pod krzywą) z tym, że każdy z nich najwięcej informacji podaje dla innego przedziału cechy ukrytej (umiejętności). Używając pierwszego (od lewej) testu, najprecyzyjniej szacować będziemy wyniki dla uczniów o niskich wartościach cechy ukrytej, środkowy test najlepiej szacował będzie średnie umiejętności, ma on też najszerszy zakres. Ostatni test daje nam najwięcej informacji o uczniach posiadających wysokie wartości cechy ukrytej.

Strategie budowania testów

Wszystkie poznane do tej pory narzędzia statystyczne, oparte na teorii odpowiedzi na zadanie testowe, w znacznym stopniu mogą ułatwić tworzenie dobrych pod względem psychometrycznym testów. Wielką zaletą teorii odpowiedzi na zadanie testowe jest to, iż narzędzia, które na jej podstawie powstały, umożliwiają projektowanie testów o możliwych do przewidzenia właściwościach psychometrycznych.

Sama teoria do takiego zadania jest jednak niewystarczająca. Aby móc efektywnie układać dobre testy, potrzebny jest tzw. bank zadań, czyli

zbiór zadań już wcześniej wypróbowanych, zawierający informację o parametrach zadań oszacowanych za pomocą modeli IRT. Taki bank zadań powinien powstawać latami, gromadząc sukcesywnie zadania i ich charakterystyki wykorzystywane we właściwym egzaminowaniu oraz zadania i ich charakterystyki z przeprowadzanych sukcesywnie badań pilotażowych. Bank zadań może być również stworzony *ad hoc*, co zazwyczaj jest strategią mniej efektywną (aczkolwiek od czegoś należy zacząć) na podstawie szeroko zakrojonych badań pilotażowych, w których (na losowych próbach) sprawdza się dużą liczbę zadań po to, by wybrać z nich mniejszą część do testowania właściwego.

Posiadając zbiór udokumentowanych psychometrycznie zadań, czyli bank zadań oraz narzędzia statystycznie oparte na teorii odpowiedzi na zadanie testowe, można stworzyć test odpowiadający potrzebom i oczekiwaniom konstruktorów. Robi się to, przewidując właściwości planowanego testu na podstawie zadań o znanej charakterystyce. Przewidywanie takie polega na wyborze zadań, a następnie (na podstawie parametrów poszczególnych zadań) wyliczeniu charakterystyk dla całego testu – chodzi tu przede wszystkim o krzywą charakterystyczną testu i krzywą informacyjną testu.

Po wyliczeniu charakterystyk testu sprawdzamy, czy są one zgodne z założeniami. Jeżeli nie, to możemy „wymienić" część zadań, by polepszyć charakterystyki testu, a następnie jeszcze raz oszacować ogólny obraz testu – tym razem ulepszonego. Znając wartości parametrów zadań, wszystkie te czynności możemy zrobić za pomocą komputera, bez kolejnych badań próbnych.

Oczywiście, po stworzeniu testu za pomocą komputera, należy przeprowadzić badanie pilotażowe, aby utwierdzić się w przekonaniu, iż test „działa" we właściwy sposób. Sprawdzamy, że zadania zostały odpowiednio przepisane na arkusz testowy, nie zachodzą pomiędzy nimi nieprzewidziane interakcje (np. w tekście jednego z poleceń znajduje się odpowiedź na inne zadanie).

Podstawowym pozostaje pytanie: Jakie charakterystyki powinien posiadać dobry test mierzący wiadomości i umiejętności? Nie jest to pytanie łatwe. Każdy test ma jakiś cel, którego autor powinien być świadomy. Zastosowań testów jest bardzo wiele. Test wiedzy i umiejętności może być egzaminem wstępnym na prestiżową uczelnię, może służyć jako podstawa do ewaluacji placówek edukacyjnych, celem testu może być wyłonienie przyszłych olimpijczyków lub wyszukanie uczniów o szczególnych potrzebach edukacyjnych itp. Wydaje się jednak, iż ze względu na wymienione cele każdy test można próbować zaklasyfikować do jednego z trzech rodzajów: *testu przesiewowego, testu szerokiego zasięgu* lub *testu selekcyjnego*.

- Testy przesiewowe (*screening tests*) – czyli testy, których celem jest precyzyjnie orzekanie, czy dany uczeń uzyskał minimalną (ustaloną przez specjalistów z danej dziedziny) wartość umiejętności. Testy takie powinno się stosować, gdy chcemy, aby egzamin zdał duży procent zdających (np. egzaminy maturalne na poziomie podstawowym). Celem takiego testu nie jest orzekanie o tym, jak wysoką wiedzę i umiejętności ma dany uczeń, lecz czy osiągnął wymagane mini-

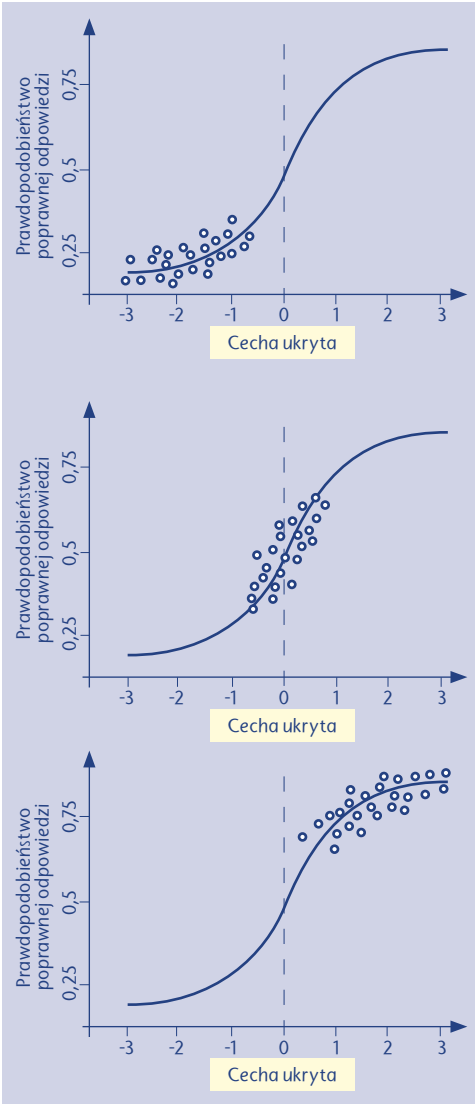
mum. Innymi słowy, chodzi o rozróżnienie zdających posiadających niskie oraz bardzo niskie wiadomości i umiejętności. Środek krzywej charakterystycznej testu powinien znaleźć się na poziomie umiejętności, korespondującym z progiem zaliczenia testu. Krzywa informacyjna powinna osiągnąć maksimum dla punktu wyznaczonego przez poziom umiejętności korespondujący z progiem zaliczenia testu. Wartości parametrów trudności poszczególnych zadań powinny być jak najbardziej zbliżone do progu zdawalności. Jeżeli jest możliwość wyboru między wieloma zadaniami o różnych charakterystykach, powinny zostać wybrane zadania o najwyższych parametrach dyskryminacji.

- Testy szerokiego zasięgu – używa się ich wtedy, gdy konstruktorowi zależy na zdobyciu jak największej ilości informacji o całej badanej grupie. Testy tego typu stosuje się między innymi w międzynarodowych badaniach osiągnięć szkolnych lub wtedy, gdy wyniki testowe służyć mają zewnętrznej ewaluacji placówek edukacyjnych, lub gdy przeprowadza się badania eksploracyjne. Taki typ testu stosowany jest również w wypadku sprawdzianu po szkole podstawowej i egzaminów gimnazjalnych. W takich przypadkach wymaga się, by zarówno krzywa charakterystyczna, jak i krzywa informacyjna testu ulokowane były na środku skali. Ponadto, co do krzywej charakterystycznej, pożądane jest, aby pole pod jej powierzchnią było jak najszersze, przy zachowaniu stosunkowo wysokiej zawartości informacyjnej. Istnieją dwie strategie prowadzące do osiągnięcia takiego efektu. Pierwsza strategia polega na próbie dobrania zadań z całego zakresu skali. Druga strategia proponuje wybranie przede wszystkim zadań ze skrajów skali kosztem zadań o średniej trudności (oczywiście, pozostawiając część zadań ze środka skali trudności). Te dwie strategie mogą wykazywać się różnym poziomem efektywności dla różnych sytuacji (np. różna długość testu, jakość zadań, jakimi konstruktor dysponuje, złożoność testu etc.).
- Testy selekcyjne – czyli testy przeznaczone dla pomiaru wysokich umiejętności. Ten typ testu jest lustrzanym odbiciem testów przesiewowych. Stosuje się je wszędzie tam, gdzie próg zaliczenia jest wysoki, a konstruktorom chodzi przede wszystkim o dobre zróżnicowanie uczniów o wysokich umiejętnościach (np. egzaminy wstępne na uczelnie wyższe, egzaminy maturalne na poziomie rozszerzonym etc.). Środek krzywej charakterystycznej w tego typu teście powinien znaleźć się na wysokim (wyznaczonym przez specjalistów) poziomie umiejętności. Krzywa informacyjna powinna osiągnąć maksimum dla punktu wyznaczonego przez wysoki poziom umiejętności. Wartości parametrów trudności poszczególnych zadań powinny być jak najbardziej zbliżone do wartości umiejętności, które szczególnie interesują konstruktorów testu. Jeżeli jest możliwość wyboru między wieloma zadaniami o różnych charakterystykach, powinny zostać wybrane zadania o najwyższych parametrach dyskryminacji.

Oczywiście można wyobrazić sobie testy, które nie będą dawały się łatwo zakwalifikować do jednej z tych trzech kategorii. Konstruktor takiego testu w tym wypadku będzie musiał iść na pewne kompromisy. To nie zmienia faktu,

iż podane trzy strategie pozostają dobrym punktem wyjścia do konstruowania testów o pożądanych właściwościach psychometrycznych, mogących swą jakością zadowolić zarówno konstruktorów testów, jak i ich zleceniodawców.

Pułapki teorii odpowiedzi na zadanie testowe



Rysunek 1.26. Ilustracja założenia niezależności od grupy badawczej. Na rysunku trzy dopasowania dla trzech grup (por. rysunek 1.20 i 1.21)

Warto pod koniec tego podrozdziału poruszyć problematykę związaną z największymi zagrożeniami związanymi z teorią odpowiedzi na zadanie testowe. Największe zagrożenia biorą się z bezkrytycznie przyjmowanych założeń teoretycznych modeli, które, jak się okazuje, często nie są spełniane w praktyce badawczej. Mowa tutaj o założeniu niezależności szacowania umiejętności od badanej grupy (*group invariant*) oraz niezależności szacowania umiejętności od stosowanych zadań (*item invariant*). Pierwsze założenie mówi o tym, iż poprawne szacowanie parametrów zadań nie jest zależne od badanej grupy (nieważne, czy testujemy wyłącznie uczniów słabych, wyłącznie uczniów średnich, czy wyłącznie uczniów bardzo dobrych pod względem badanych umiejętności, – otrzymane przez nas parametry zadań, a co za tym idzie, również całych testów, będą identyczne).

Sytuację taką przedstawia rysunek 1.26, na którym przedstawiona jest krzywa charakterystyczna odpowiedzi dla jednego zadania. Jako że kształt krzywej jest znany (jest wyrażany s-kształtną krzywą logarytmiczną), w procesie estymacji dopasowywane jest tylko przesunięcie krzywej względem osi wyznaczającej cechę ukrytą, czyli umiejętności ucznia (parametr trudności), a w modelu dwu-

parametrycznym (2PL) wyznacza się również nachylenie krzywej, czyli parametr dyskryminacji. Do wyznaczenia tych dwóch parametrów, jak się wydaje, niepotrzebna jest grupa uczniów o umiejętnościach pokrywających cały zakres skali, bo jak pokazuje rysunek 1.26, testowanie na dowolnej grupie uczniów „dopasuje" krzywą w identyczny sposób. Drugą stroną niezależności szacowania umiejętności od grupy jest niezależność szacowania umiejętności od zastosowanych zadań. O ile takie założenie byłoby spełnione, to nawet zupełnie różnymi testami (oczywiście z tej samej dziedziny) można badać umiejętności dwóch uczniów i następnie bez żadnego problemu je porównywać.

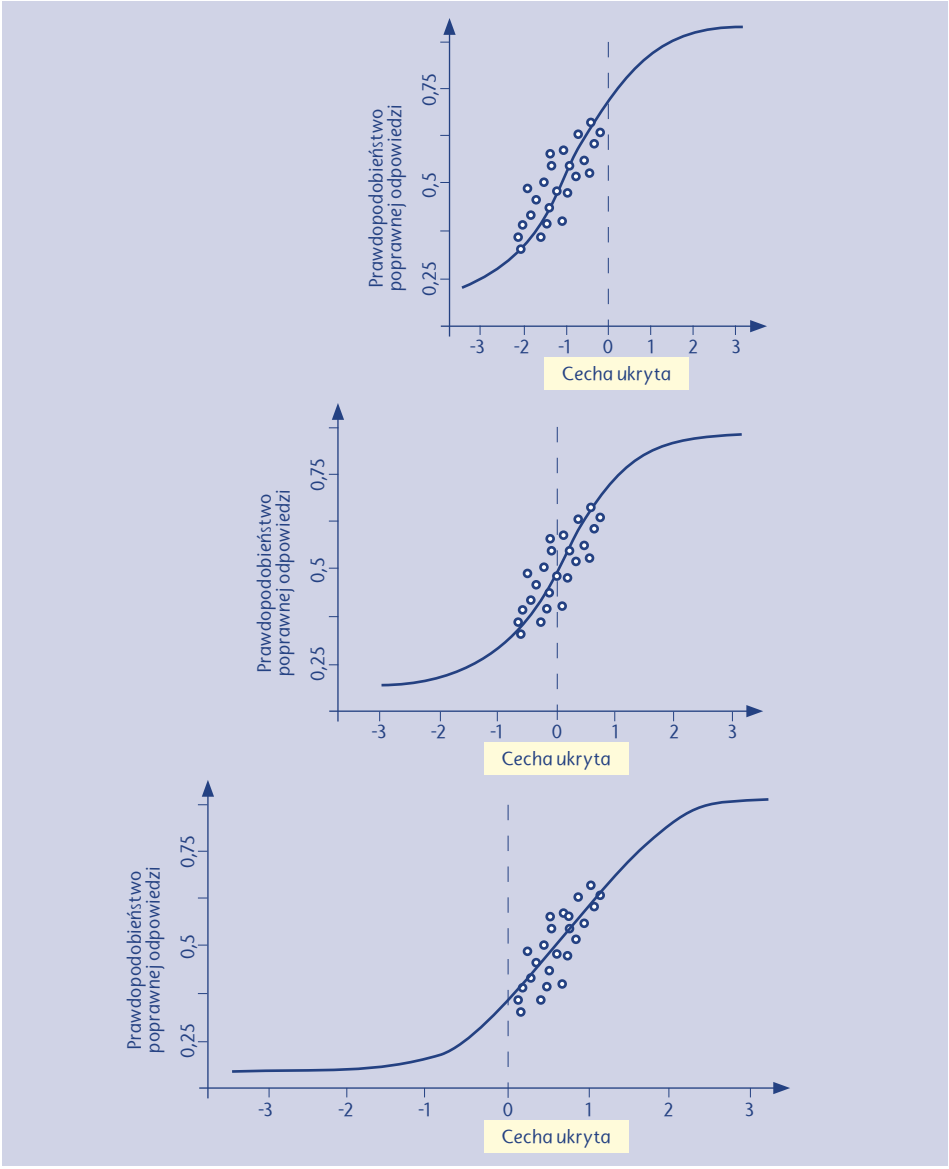
Oba założenia tworzą bardzo kuszącą obietnicę teorii. Gdyby obietnica była prawdziwa, nie trzeba by przeprowadzać dużych badań pilotażowych. Testując zadania na jakiejkolwiek grupie, otrzymalibyśmy właściwe parametry zadań. Porównywanie różnych badanych grup byłoby niezwykle proste, nie trzeba by było zrównywać testów, wyniki byłyby bowiem bezpośrednio porównywalne, a skoro tak, to można byłoby łatwo mierzyć absolutny przyrost wiedzy w kolejnych latach testowania.

Niestety, obietnica ta nie jest prawdziwa. W założeniach niezależności szacowania umiejętności od grupy i od badanych zadań kryje się kolejne ukryte założenie, iż skala cechy ukrytej w procesie estymacji jest skalą stałą, absolutną i niezależną od procesów estymacji. W rzeczywistości tak nie jest. Skala w modelach IRT jest za każdym razem wyznaczana. Z powodu tego, że zadania nie są nigdy dopasowane idealnie, proces kalibracji (czyli ustalania skali) daje za każdym razem nieco inne rezultaty.

Co dzieje się, gdy wyznaczane w procesie estymacji skale różnią się od siebie? Doskonale przedstawia to rysunek 1.27, gdzie parametry zadania, czyli parametry określające kształt krzywej charakterystycznej, estymowane były na trzech różnych skalach – bardzo wąskiej, średniej i szerokiej[7]. Widać, iż w zależności od skali, która została przyjęta, krzywa charakterystyczna zmienia swoje położenie względem zera (czyli środka skali zaznaczonego przerywaną linią). W zależności od przyjętej skali, analizowane zadanie okazuje się raz zadaniem łatwym, raz średnim, a raz trudnym.

Zawodne procesy estymacji nie pozwolą na spełnienie obietnic danych przez teorię. Nie jesteśmy jednak na całkiem straconej pozycji. Skale te mają ciekawą właściwość, są „identyczne co do przekształcenia liniowego". Oznacza to, iż możemy je zrównywać (czyli sprowadzać do jednej skali) w bardzo łatwy sposób, używając kilku prostych działań: mnożenia, dodawania i odejmowania. Skale dwóch testów możemy zrównywać, tak jak powiedziane było to w punkcie dotyczącym krzywej charakterystycznej testu: gdy obydwa testy rozwiązywane były przez tych samych uczniów lub część uczniów rozwiązywała

[7] Różnice szacowanych skal wynikają przede wszystkim z różnego doboru zadań i poziomów testowanych umiejętności. Przez to dobór innych zadań do testu ma wpływ na skalę, na której szacowana jest krzywa charakterystyczna danego zadania.



Rysunek 1.27. Krzywa charakterystyczna i problemy skali

obydwa testy, gdy w obydwu testach znajdowała się pewna liczba zadań wspólnych lub gdy testowaliśmy dwie losowo dobrane grupy uczniów z tej samej populacji. Niezależność skali co do przekształcenia liniowego umożliwia w tych warunkach efektywne zrównywanie dające porównywalność. Wcześniejsze metody, czyli zapewnienie porównywalności za pomocą klasycznej teorii testu, były niezwykle skomplikowane i przynosiły wątpliwe efekty.

1.4. Cele a metody pomiaru: od egzaminów zewnętrznych po badania międzynarodowe

W tej części zawarto kilka uwag dotyczących praktycznych kwestii wpływających na wybór metody pomiaru umiejętności i wiedzy uczniów, a także na sposoby ich prezentacji. Omawiamy poniżej, jak wybór metody pomiaru zależy od celu, dla którego pomiaru dokonujemy, a przede wszystkim od tego, czy wyniki wykorzystywane będą na poziomie ucznia, czy też agregowane i prezentowane dla szerszych grup. W praktyce często mamy do czynienia z sytuacją, gdy wyniki testów mają być zarówno prezentowane dla pojedynczych uczniów, jak i porównywane między większymi grupami. Takim przykładem są wyniki egzaminów zewnętrznych, które przede wszystkim dostarczają wyników dla poszczególnych uczniów, ale i służą porównaniom między placówkami czy regionami, a także są podstawą badań istotnych dla polityki edukacyjnej państwa. Warto więc przemyśleć, kiedy możemy stosować wyniki egzaminów zewnętrznych bez żadnej korekty, a kiedy powinny być one poddane odpowiedniemu skalowaniu i standaryzacji, zanim przeprowadzone zostaną jakiekolwiek porównania.

Zależność między celami a metodami pomiaru

Pomiar umiejętności i wiedzy uczniów ma zazwyczaj dwa różne cele. Bardzo często za pomocą testów chcemy ocenić umiejętności i wiedzę każdego ucznia. Dokładniej mówiąc, naszym celem jest prezentacja osobnego wyniku dla każdego ucznia, który jak najbardziej precyzyjnie oddaje jego prawdziwy poziom wiadomości i umiejętności. Wyniki na poziomie szkół czy całej populacji są tutaj sprawą drugorzędną. Nieco inne podejście powinno być stosowane, gdy naszym celem jest określenie przeciętnego poziomu umiejętności i wiedzy dla grup uczniów czy też całych populacji. W takim wypadku wynik indywidualny ucznia nie jest prezentowany, stanowiąc jedynie pośrednią informację pomagającą określić wynik dla szerszej grupy uczniów.

System egzaminów zewnętrznych ma przede wszystkim na celu określenie wyniku dla każdego ucznia, należy więc do pierwszej grupy testów wymienionej powyżej. Jednak często jest też wykorzystywany do określenia średniego wyniku szkoły, gminy, powiatu czy też nawet porównywania regionów (wyników między rejonami poszczególnych OKE czy też województwami) i subpopulacji uczniów całego kraju (np. dziewczęta i chłopcy, uczniowie ze wsi a uczniowie z miast). Co więcej, wyniki egzaminacyjne wykorzystywane są do tworzenia bardziej złożonych miar mających np. określić jakość nauczania w danej szkole. Taką miarą jest – przykładowo – wskaźnik EWD (edukacyjnej wartości dodanej) gimnazjum, który uwzględnia poziom wiadomości i umiejętności ucznia pod koniec nauki w szkole podstawowej (wynik sprawdzianu), oceniając, ile zyskał on podczas nauki w gimnazjum.

O ile proste modele prezentowane w tym rozdziale są często podstawą do oceny poziomu umiejętności i wiadomości poszczególnych uczniów, to już podobne oceny dla większych grup zyskują na precyzji, gdy pod uwagę brany jest model wyników dla całej populacji uczniów. Załóżmy, że uczeń uzyskał 30 punktów na sprawdzianie szóstoklasistów. Jak wiemy, wynik ten jest miarą prawdziwych umiejętności ucznia obarczoną błędem pomiaru, czyli niepewnością związaną z zestawem rozwiązywanych zadań, dyspozycją ucznia danego dnia etc. (patrz dyskusja na początku rozdziału). Co więcej, 30 punktów może być dość szeroką kategorią zawierającą uczniów o bardzo różnym poziomie wiadomości i umiejętności, których nie możemy rozróżnić ze względu na ograniczoną liczbę zadań w teście (np. 30 punktów na sprawdzianie z 2007 roku uzyskało ponad 20 tysięcy szóstoklasistów).

Ten sam uczeń poddany identycznemu testowi uzyskałby zapewne nieco inny wynik. Najczęściej bardzo bliski poprzedniemu, jednak z pewnym prawdopodobieństwem nieco inny. Jeśli moglibyśmy podawać wyniki z dokładnością do kilku cyfr po przecinku, to na pewno kolejne testy pokazywałyby inne wartości, nawet gdyby po zaokrągleniu uczeń otrzymywał zawsze 30 punktów. Co więcej, moglibyśmy też rozróżnić wyniki poszczególnych uczniów i tak na przykład, zamiast 20 tysięcy szóstoklasistów z tym samym wynikiem równym 30 punktów, mielibyśmy tysiąc szóstoklasistów z wynikiem równym 30,27 punktu.

W zależności od tego, jak wykorzystujemy ten wynik, takie rozróżnienie – związane z większą precyzją pomiaru – może mieć spore znaczenie lub też być zupełnie zbędne z praktycznego punktu widzenia. Dla indywidualnego wyniku sprawdzianu udostępnianego uczniowi i nauczycielom taka dokładność nie jest konieczna. Wiemy przecież, że błąd pomiaru wynosi kilka punktów, więc w porównaniu z nim ułamki są naprawdę sprawą pomijalną. Jeśli jednak wynik stanowi podstawę kolejnych obliczeń, np. średniej czy EWD szkoły, to już obliczenie wyniku ucznia z jak największą dokładnością zwiększy istotnie precyzję oszacowań tych wskaźników, dając dodatkowo znacznie lepszą podstawę do określenia zakresu, w jakim znajduje się prawdziwy wynik (przedziału ufności). Intuicyjnie, o ile korzyści płynące z dużej dokładności są niewielkie przy pojedynczym wyniku, to nabierają znaczenia, gdy korzystamy z kilkudziesięciu wyników obliczonych z większą precyzją.

Jak w takim razie możliwe jest zwiększenie precyzji pomiaru i prezentowanie wyników z dokładnością do kilku cyfr po przecinku, skoro testy zawsze składają się z kilkudziesięciu zadań? Można argumentować, że znając pewne charakterystyki uczniów, jesteśmy w stanie ocenić, z jakim prawdopodobieństwem uczeń z zaobserwowanym wynikiem 30 punktów posiada rzeczywistą wiedzę i umiejętności na wyższym, podobnym czy też niższym poziomie.

Teoria stojąca za tym rozróżnieniem jest dość złożona i nie ma tu miejsca, aby omawiać ją szczegółowo, jednak poniższy przykład powinien pomóc w zro-

zumieniu różnicy. Załóżmy, że mamy do czynienia z dwojgiem uczniów. Obydwoje uzyskali po 30 punktów z testu z matematyki. Dziewczynka uzyskała jednak wyższy od chłopca wynik z testu z fizyki, pochodzi przy tym z rodziny inteligenckiej, a chłopiec z robotniczej. Jest więc prawdopodobne, ale oczywiście nie do końca pewne, że jej prawdziwy poziom umiejętności jest wyższy niż chłopca, bowiem na podstawie wyników innych uczniów wiemy, że osoby, które uzyskały wyższe wyniki z testu z fizyki, uzyskują też średnio wyższe wyniki z testu matematycznego, a dzieci z rodzin inteligenckich wyższe wyniki niż dzieci z rodzin robotniczych. Z drugiej strony, dziewczynki uzyskały przeciętnie niższe wyniki z tego testu niż chłopcy, co sugeruje, że to jednak chłopiec może mieć wyższy poziom rzeczywistych umiejętności.

Trudno wyważyć te argumenty bez odpowiedniego modelu statystycznego. Model ten powinien uwzględniać wyniki wszystkich uczniów i szacować rzeczywisty poziom wiadomości i umiejętności zarówno na podstawie wyników testu, jak przeciętnych różnic w wynikach dzieci – zależnie na przykład od płci, pochodzenia społecznego, a także wyników z innych testów (w praktyce wykorzystuje się wszelkie dodatkowe informacje o cechach zgromadzone przy okazji pomiaru). Taki złożony model statystyczny będzie w stanie „wyważyć" prawdopodobieństwa związane z uzyskaniem danego wyniku zależnie od bogatego zestawu cech ucznia i podać najbardziej prawdopodobny wynik z dokładnością do kilku miejsc po przecinku.

Teraz powinno być już jasne, dlaczego podobne podejście nie jest wskazane przy określaniu wyników poszczególnych uczniów. Choć modele te mogą być stosowane także w tym wypadku, to tutaj często preferowany jest wynik szacowany za pomocą prostego modelu IRT, bez uwzględnienia cech ucznia, lub nawet stanowiący prostą sumę poprawnych odpowiedzi, tak jak w polskich egzaminach.

Przyczyna jest bardzo prosta – nikt nie chciałby, żeby jego końcowy wynik zależał od płci czy pochodzenia społecznego, a także od wyniku z innego testu. Na pewno chłopcu z powyższego przykładu nie podobałaby się ani korekta wyniku „w dół" ze względu na pochodzenie społeczne, ani ze względu na gorszy wynik z fizyki; dziewczynce też nie podobałaby się korekta ze względu na niższe wyniki innych dziewcząt. Stąd wyniki egzaminów obliczane są bez korekty ze względu na cechy ucznia, uwzględniając jedynie odpowiedzi na zadania testowe z danego przedmiotu.

Zupełnie inną sytuację mamy jednak, gdy szacujemy wyniki uczniów bez intencji ich publikacji. Obliczamy wynik dla każdego z nich, lecz raczej celem określenia osiągnięć dla szerszej grupy. Skoro nie ma potrzeby podawania wyniku ucznia, to wiedząc, że jest on obarczony błędem pomiaru, warto uwzględnić wszystkie informacje, jakie posiadamy i określić najbardziej prawdopodobny wynik każdego ucznia, zwiększając precyzję pomiaru. Uwzględnienie takich cech jak płeć, wyniki z innych testów czy pochodzenie społeczne, pozwoli zwiększyć

precyzję ocen uczniów, a przez to też precyzję ocen grup uczniów – czy to szkół, czy to regionów.

Korekta wyników o cechy uczniów jest zawsze wskazana, gdy prowadzimy badania edukacyjne, a więc naszym celem nie jest ocena uczniów, ale raczej odniesienie uzyskanych wyników do ich cech lub charakterystyk szkół czy efektów programów edukacyjnych. Tego rodzaju modele są też stosowane w międzynarodowych porównawczych badaniach umiejętności i wiedzy uczniów, które obecnie wyznaczają standardy metodologiczne badań edukacyjnych. Z tego względu omówimy krótko metody pomiaru stosowane w badaniu PISA, w którym Polska uczestniczyła we wszystkich edycjach.

Pomiar w badaniach międzynarodowych

Badanie PISA obejmuje populacje 15-latków we wszystkich krajach członkowskich OECD (Organizacji Współpracy Gospodarczej i Rozwoju, skupiającej najbardziej rozwinięte demokratyczne gospodarki świata), a także w kilkudziesięciu krajach partnerskich. Metodologia pomiaru stosowana w tym badaniu jest bardzo bliska metodom stosowanym w innych znanych badaniach międzynarodowych w rodzaju TIMSS (badanie umiejętności matematycznych i w naukach przyrodniczych wśród 4- i 8-klasistów, w którym niestety Polska nie uczestniczy) oraz PIRLS (badanie umiejętności czytania wśród 4-klasistów, w którym Polska uczestniczyła w 2006 roku). Metodologia tych badań rozwinięta została w pracach nad amerykańskim systemem ogólnonarodowych testów NAEP, a potem dostosowana do badań międzynarodowych. Obecnie badania te wykorzystują najbardziej zaawansowane metodologicznie sposoby pomiaru i raportowania umiejętności uczniów, które stanowią punkt odniesienia dla wszystkich innych badań edukacyjnych.

Badania NAEP, PISA, TIMSS czy PIRLS podają statystyki dla wyników uczniów zarówno na skali punktowej, jak i na skali porządkowej, umożliwiającej jakościową interpretację wyników. Przykładowo, badanie NAEP wyróżnia 3 poziomy umiejętności i wiedzy uczniów: podstawowy, biegły, zaawansowany. PISA 2006 podaje wyniki w naukach przyrodniczych, stosując 6 opisowych poziomów wraz z kategorią uczniów, którzy nie osiągnęli nawet poziomu najniższego. Dzięki temu wyniki można odnieść do realnych umiejętności uczniów. I tak na przykład, wynik na poziomie 2 w naukach przyrodniczych w badaniu PISA 2006 oznacza, że uczeń posiada wiedzę z nauk przyrodniczych umożliwiającą wyjaśnianie prostych zjawisk w znanym kontekście. Wynik na poziomie 5 oznacza już, że uczeń potrafi w sposób naukowy interpretować złożone zjawiska, stosować posiadaną wiedzę i naukowe metody poznania, oceniać wartość informacji z punktu widzenia nauk przyrodniczych, krytycznie analizować nieznane mu zjawiska i budować ich własną naukową interpretację.

spis treści

Podstawowy cel pomiaru w badaniach międzynarodowych został już opisany powyżej. Zakłada się, że określenie wyniku poszczególnego ucznia nie jest celem badania. Zależy nam na uzyskaniu jak najbardziej precyzyjnej oceny wyników uczniów w danej populacji (najczęściej w danym kraju) lub też w subpopulacji (np. dziewczynki w Polsce). Można wykazać, że korzystanie z wyników stanowiących proste zsumowanie poprawnych odpowiedzi czy też rezultatów uzyskanych z prostych modeli IRT, nie biorących pod uwagę cech uczniów, daje błędne oceny średnich wyników w populacji, a tym bardziej błędne oceny zróżnicowania tych wyników (na przykład ich wariancji).

Z tego wynika, że prostsze metody dają też błędne oszacowania procentu uczniów na określonym poziomie umiejętności (np. procentu uczniów na poziomie określonym jako „podstawowy"), różnic w średnich wynikach między grupami (na przykład chłopców i dziewczynek, uczniów ze wsi i z miasta) czy też zależności między wynikami a interesującymi nas zmiennymi (na przykład wynikami a pochodzeniem społecznym ucznia, czyli zmiennymi wykorzystywanymi w niemal każdym badaniu edukacyjnym). Z tego względu stosuje się korektę wyników, wykorzystując dla całej populacji model opierający się o bogaty zestaw cech ucznia i jego rodziny, wyników w testach z różnych dziedzin, a także odpowiedzi na zadania dotyczące na przykład strategii uczenia się czy stosunku do nauki. Cechy te zbierane są poprzez dodatkowy kwestionariusz, który wypełniany jest po teście wiadomości i umiejętności, a większość z nich stanowi dobre predyktory wyników ucznia.

Ze względu na to, że wiedza, jaką badają NAEP, PISA, PIRLS i TIMSS, jest niezwykle rozległa, a ograniczenia budżetowe są zawsze dość rygorystyczne, opracowano metody, w których uczniowie rozwiązują różne zestawy zadań, tzw. *booklets* (dosłownie „książeczki" czy „broszury"). Każdy uczeń dostaje więc tylko część zadań przygotowanych z danej dziedziny, na przykład jedynie zadania z *booklets* numer 1, 3 oraz 7. Następnie szacowane są parametry statystyczne zadań z każdej z *booklets* metodami IRT. Ponieważ testy są losowo przypisywane uczniom, to parametry te nie powinny się różnić między zestawami zadań. Jeżeli jednak różnice wystąpią, to są korygowane. Jest to możliwe, ponieważ uczniowie rozwiązują te same zadania w różnych konfiguracjach *booklets*.

Dodatkowo 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 etc.) stało się tak, że jakieś zadanie okazuje się znacznie trudniejsze lub łatwiejsze tylko w jednym kraju, to jest ono usuwane, zanim skalowane są wyniki 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, bowiem najprostsze statystyki pokażą, że nawet najlepiej czytający uczniowie w tych krajach nie są

w stanie go rozwiązać, podczas gdy nie sprawia ono trudności uczniom polskim, którym zarówno święty Mikołaj, jak i zima, są zapewne dużo bliższe.

W końcu uzyskujemy w pełni porównywalne wyniki testów, które można poddać dalszej „obróbce". W praktyce proces przygotowania zadań jest tak drobiazgowy, wspierany także badaniami próbnymi w większości krajów, że w każdej edycji badania PISA usuwanych jest na koniec dosłownie tylko kilka zadań. Pozostałe zadania okazują się niemal identyczne pod względem właściwości psychometrycznych we wszystkich krajach.

Stoimy wciąż przed dwoma problemami. Po pierwsze, każdy uczeń odpowiedział jedynie na część z całej baterii zadań. Po drugie, wynik każdego ucznia oddany jest na skali z zaokrągleniem do pełnej liczby i bez uwzględnienia jego cech indywidualnych. Oba problemy rozwiązuje model tzw. *plausible values* (dosłownie model „wiarygodnych wartości"), gdzie uczniom przypisywanych jest 5 równie prawdopodobnych wyników, po wzięciu pod uwagę odpowiedzi na wszystkie zadania testowe, a także związków odpowiedzi na testy z cechami uczniów w całej populacji. Ten dość skomplikowany model statystyczny, przypisujący 5 wyników każdemu uczniowi, odzwierciedla zarówno niepewność pomiaru, jak i niepewność związaną z uwzględnieniem cech indywidualnych. Wykazano jednak, że posługując się pięcioma *plausible values*, można odtworzyć rozkład prawdziwych umiejętności w całej populacji[8].

Tak więc nie tylko średnie wyniki w całej populacji, ale i wyniki dla podgrup, wariancja tych wyników oraz dalsze analizy odnoszące osiągnięcia uczniów do ich 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 tego, że każda z *booklets* posiada zadania wspólne z innym *booklets*, a model

[8] Można tego dowodzić zarówno analitycznie, posługując się matematycznymi modelami, jak i poprzez symulację, gdzie zmienną ukrytą (określającą prawdziwy poziom wiedzy ucznia) odzwierciedlają obserwowalne odpowiedzi na zadania testowe, a związki między odpowiedziami a cechami ucznia mają z góry założone charakterystyki. W ten sposób badacz zna w pełni strukturę danych i sprawdza, czy model potrafi odtworzyć wartości zmiennej ukrytej (znanej nam w symulacji, ale nie uwzględnianej bezpośrednio w analizie statystycznej, tak jak to jest w rzeczywistości, gdzie zmiennej tej nie obserwujemy). Okazuje się, że model *plausible values* jest w stanie w pełni odtworzyć nie tylko średnie wartości zmiennej ukrytej, ale i cały jej rozkład, co nie jest możliwe w wypadku prostszych modeli IRT (por. Margaret Wu (2005), „The role of plausible values in large-scale surveys", Studies In Educational Evaluation, vol. 31(2-3), s. 114-128; por. także przykłady w: OECD (2009), *PISA Data Analysis Manual*, dostępne za darmo na stronach www.pisa.oecd.org; omówienie różnic między wynikami analiz z różnie oszacowanymi wynikami uczniów można znaleźć w: Jakubowski M. (2009), *Impact of IRT scaling methodology on secondary analysis*, wersja robocza dostępna na www.wne.uw.edu.pl/mjakubowski).

statystyczny równocześnie bierze pod uwagę wszystkich uczniów i wszystkie zadania testowe[9].

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

W odniesieniu do systemu egzaminów zewnętrznych doświadczenia badań międzynarodowych pokazują, w jakim kierunku system ten można rozwijać. Dotyczy to nie tylko samych modeli statystycznych skalujących odpowiedzi uczniów z testów, ale także samego przygotowania testów i szczegółowego procesu ich walidacji. Oczywiście system egzaminacyjny ma przede wszystkim dawać oceny umiejętności poszczególnych uczniów, stąd np. model *plausible values* jest tu niepotrzebny, ale już do badań odnoszących wyniki uczniów do ich cech lub cech szkół czy programów edukacyjnych takiego rodzaju podejście w mniejszym lub większym stopniu powinno być stosowane.

Wszelkie badania opierające się na prostszych modelach pomiaru wiadomości i umiejętności ucznia mogą być porównane z metodologią badań międzynarodowych. Dzięki takiemu porównaniu dowiemy się, w jakich aspektach wyniki opierające się na prostszych metodach są mało wiarygodne lub wręcz zniekształcone w spodziewanym kierunku (np. wariancja surowych, niepoddanych skalowaniu wyników egzaminów będzie z pewnością zaniżona w stosunku do prawdziwej wariancji umiejętności). Dokumentacja badań międzynarodowych jest ogólnie dostępna, choć jedynie w języku angielskim, na stronach internetowych tych badań, a także w wielu ogólnodostępnych publikacjach przybliżających tę tematykę[10]. Zachęcamy do ich lektury i krytycznego spojrzenia na badania egzaminów zewnętrznych w Polsce z perspektywy bardziej zaawansowanej metodologii stosowanej w badaniach międzynarodowych.

[9] Dokładniej, model bezpośrednio szacuje wyniki w kilku dziedzinach (por. OECD (2009), *PISA Technical Report*, rozdział 9).

[10] Informacje o badaniu NAEP dostępne są na stronie http://nces.ed.gov/nationsreportcard/, gdzie znaleźć można zarówno informacje o wynikach, jak i historię badań oraz pełną dokumentację techniczną. Dokumentacja PISA jest udostępniana na stronach www.pisa.oecd.org, a dokumentacja TIMSS oraz PIRLS na stronie http://timss.bc.edu/. Większość artykułów, także przeglądowych i popularnonaukowych, zawarta jest w bazach ERIC dostępnych pod adresem www.eric.ed.gov.

Literatura

Pomiar

Babbie, E. (2005), *Badania społeczne w praktyce*, Wydawnictwo Naukowe PWN: Warszawa.

Brzeziński, J. (2007), *Metodologia badań psychologicznych*, Wydawnictwo Naukowe PWN: Warszawa.

Brzeziński, J. (red.) (2006), *Metodologia badań psychologicznych*: *wybór tekstów*, Wydawnictwo Naukowe PWN: Warszawa.

Choynowski, M. (1971), *Pomiar w psychologii,* w: Kozielecki, J. (red.), *Problemy psychologii matematycznej* (s. 43-63), Wydawnictwo Naukowe PWN: Warszawa.

Lissowski G., Haman J., Jasiński M. (2008), *Podstawy statystyki dla socjologów*, Wydawnictwo Naukowe Scholar: Warszawa.

Nachmias, Ch., Nachmias, F. (2001), *Metody badawcze w naukach społecznych*, Wydawnictwo Zysk i S-ka: Poznań.

Niemierko, B. (2004), *Pomiar wyników kształcenia*, Wydawnictwa Szkolne i Pedagogiczne: Warszawa.

Walenta, K. (1971), *Podstawowe pojęcia w teorii pomiaru*, w: Kozielecki, J. (red.), *Problemy psychologii matematycznej* (s. 43-63), Wydawnictwo Naukowe PWN: Warszawa.

Klasyczna teoria testu

Ferguson, G. A., Takane Y. (2007), *Analiza statystyczna w psychologii i pedagogice*, Wydawnictwo Naukowe PWN: Warszawa.

Izard, J., *Quantitative research methods in educational planning. Module 7: Trial testing and item analysis in test construction*, UNESCO International Institute for Educational Planning: http://www.iiep.unesco.org/capacity-development/training/training-materials/quantitative-research.

Salkind N. J. (red.) (2007), *Encyclopedia of Measurment and Statistics*, SAGE: Thousand Oaks-London-New Delhi.

Verchelst, N. (2007), *Klasyczna teoria testu*, Biuletyn Badawczy CKE „Egzamin", 9/2007, www.cke.edu.pl.

IRT

Aranowska, E. (2005), *Pomiar ilościowy w psychologii*, Wydawnictwo Naukowe Scholar: Warszawa.

Backer, F. B. Backer (2001), *The Basics of Item Response Theory*, ERIC: Wisconsin.

Boomsman, A., Van Dujin M. A. J., Snijders T. A. B. (2000), *Essays on Item Response Theory*, New York: Springer.

Davier von, M., Carstensen, C. H. (red.) (2007), *Multivariate and mixture distribution Rasch models: Extensions and applications*, New York: Springer.

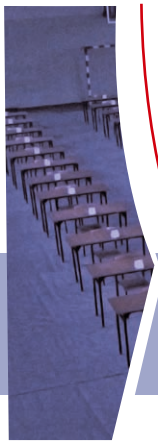
Linden van der, W. J. i Hambleton, R. K. (red.) (1997), *Handbook of Modern Item Response Theory*, New York: Springer.

Masters, G. N., Keeves, J. P. (red.) (1999), *Advances in measurement in educational research and assessment*, New York: Pergamon (Elsevier Science).

Szaleniec, H. (2007), *Refleksje na temat rozwoju systemu egzaminacyjnego*, XIII Konferencja Diagnostyki Edukacyjnej „Uczenie się i egzamin w oczach uczniów", Łomża, 5-7.10.2007.
Verchelst, N. (2007), *Probabilistyczna teoria testu*, Biuletyn Badawczy CKE „Egzamin", 9/2007, www.cke.edu.pl.

Międzynarodowe badania umiejętności i wiedzy uczniów

www.pisa.oecd.org – międzynarodowa strona programu OECD PISA zawierająca darmowe publikacje oraz bazy danych
http://timss.bc.edu/ – międzynarodowa strona programów TIMSS oraz PIRLS zawierająca darmowe publikacje oraz bazy danych



2. Wnioskowanie o populacji

W poprzednim rozdziale skupiliśmy się na kwestiach pomiaru wiedzy i umiejętności, które z oczywistych względów stanowią centralne zagadnienie w badaniach edukacyjnych. Badania edukacyjne są jednak najczęściej prowadzone celem wnioskowania o osiągnięciach uczniów i ich uwarunkowaniach w szerszej populacji. Inaczej mówiąc, badacza rzadko kiedy interesują wyniki pojedynczego ucznia. Nie powinno go także zadowalać jedynie określenie poziomu osiągnięć w zebranej próbce uczniów. Celem badania powinno być wnioskowanie o osiągnięciach w całej docelowej populacji, na przykład populacji 10-latków czy też populacji uczniów trzecich klas gimnazjów. Co więcej, badacz najczęściej chciałby poznać wpływ, jaki na osiągnięcia mają wybrane cechy (zmienne).

W niniejszym rozdziale omawiamy możliwość wnioskowania na podstawie danych z losowej próby, czyli grupy uczniów lub szkół losowo wybranych do badania. Wnioskować chcemy o zależnościach w populacji, czyli o wszystkich uczniach lub szkołach o tych samych cechach, na przykład uczniach w tym samym wieku czy uczących się w tym samym roczniku; gimnazjach publicznych; uczniach liceów ogólnokształcących itp. Poruszamy także kwestie interpretacji obserwowanych związków między zmiennymi jako przyczynowo-skutkowych.

O ile sposoby uzyskania danych reprezentatywnych dla populacji są dość dobrze opisane i poprawnie przeprowadzone, badania często można uogólniać na szersze grupy uczniów czy szkół, to już wnioskowanie z zebranych danych o przyczynach i skutkach jest bardzo trudne. Omawiamy te kwestie w kontekście informacji o uczniach, które zbierane są w szkołach, a także z perspektywy oceny rozwiązań funkcjonujących w szkołach. Zagadnienia te są jednak podstawowe także dla innych dyscyplin, zainteresowani czytelnicy mogą więc poszukiwać dodatkowych informacji w podręcznikach badań socjologicznych, psychologicznych czy ekonometrycznych.

Poniższa dyskusja może być pomocna dla osób, które rozważają przeprowadzenie własnych badań. Jednak większość z nas będzie po prostu zastanawiać się, czy wyniki opracowane w konkretnym badaniu w szkołach można odnieść do problemów, jakie nas interesują. Czy stanowić one mogą wskazówkę, jak postępować? Czy można, opierając się na wynikach danego badania, zmienić sposób nauczania lub organizacji własnej szkoły? W każdej sytuacji, niezależnie od jakości przeprowadzonego badania terenowego i dalszej analizy zebranych w nim wyników, powstawać będą wątpliwości, w jakim stopniu wyniki te można uogólnić także na nasz użytek. Warto wiedzieć, kiedy badanie daje podstawę do takich uogólnień. Jak duże jest prawdopodobieństwo tego, że w naszej sytuacji interesujące nas czynniki odgrywają podobną rolę. Zrozumienie różnic między metodami zbierania danych i ich analizy może nam w tym bardzo pomóc.

Podstawową kwestią do rozważenia jest to, czy mamy do czynienia z próbą losową, czy też z danymi zebranymi w sposób nielosowy. Jedynie próba losowa daje podstawę do wnioskowania o zależnościach zachodzących w całej populacji. Nie znaczy to jednak, że każda próba losowa dobrze opisuje każdą populację. Wiele zależy też od metody przeprowadzonego losowania, a następnie od sposobu analizy danych. O tych kwestiach piszemy w podrozdziale 2.1.

Dane zbierane w sposób nielosowy nie powinny być z zasady uogólniane na całą populację. W edukacji często jednak dane takie już posiadamy i chcemy wykorzystać ich potencjał. Często też naszym celem nie jest uzyskanie reprezentatywności dla całej populacji, ale zrozumienie zależności zachodzących w konkretnej grupie. W podrozdziale 2.2 omawiamy kwestie dotyczące tego rodzaju danych.

W podrozdziale 2.3 poruszamy kwestie możliwości oceny zależności przyczynowo-skutkowych. W podrozdziałach 2.4 oraz 2.5 opisujemy natomiast dwa podejścia do badań edukacyjnych, które zyskują ostatnio coraz większe znaczenie. Badania eksperymentalne znane są od lat, jednak w edukacji rzadko kiedy były wykorzystywane z uwagi na kwestie praktyczne i natury etycznej. W podrozdziale 2.4 krótko omawiamy podstawowe schematy eksperymentalne, a przede wszystkim zalety i wady eksperymentów w porównaniu z badaniami reprezentatywnymi. W podrozdziale 2.5 przedstawiamy metody quasi-eksperymentalne czy też tzw. naturalne eksperymenty, które od kilkunastu lat są coraz szerzej doceniane. Nie są to tak naprawdę metody zbierania danych, lecz raczej sposoby analizy, jednak wiąże się z nimi nieco inne podejście także do samych zbiorów danych.

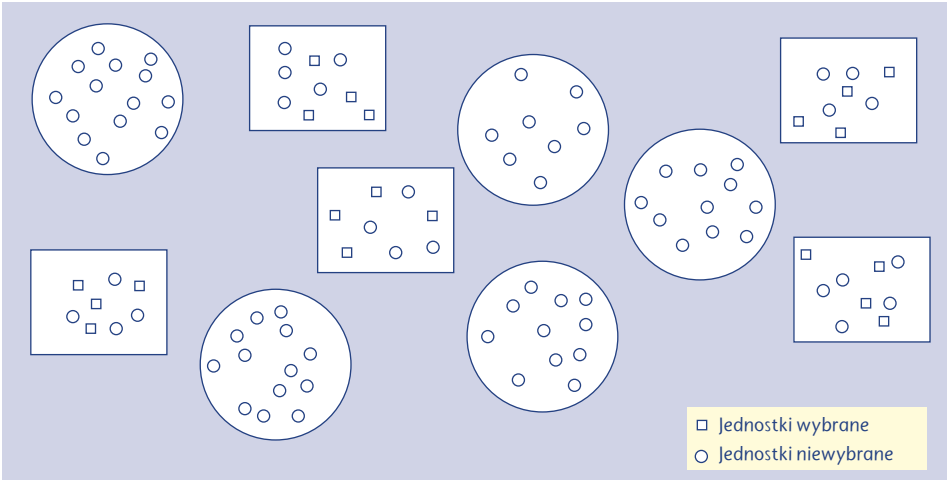
W kolejnym podrozdziale dotykamy kwestii ważnej dla badaczy edukacyjnych, którzy często mają dostęp do danych dla całej populacji. Nie można w takiej sytuacji mówić o próbie i podstawowe pojęcia statystyki tracą swój pierwotny sens. Wprowadzamy abstrakcyjne pojęcie superpopulacji, wyjaśniając, w jakich sytuacjach analizy statystyczne dla całej populacji mają sens, a kiedy powinniśmy się ograniczyć do opisu danych.

Ostatni podrozdział 2.7 dotyka kwestii badań prowadzonych na niewielką skalę, na przykład w kilku czy w kilkunastu szkołach. Badania tego typu mają zawsze nieco inny charakter i nie mogą stanowić podstawy do wnioskowania o procesach zachodzących w całej populacji uczniów, nauczycieli czy szkół. Mają one jednak swoje zalety i często dostarczają nam wyników, które dopełniają badania na większych próbach. Omawiamy krótko kwestie, które stanowią o jakości tego rodzaju badań i możliwości porównania ich wyników z innymi pracami.

2.1. Próby losowe – kiedy badanie można nazwać reprezentatywnym?

Najczęstsze w ilościowych badaniach edukacyjnych są badania wykorzystujące losową próbę uczniów, rodziców, nauczycieli lub szkół. Są one najczęściej podobne do badań sondażowych, których wyniki tak lubią ogłaszać media. Badania edukacyjne mają jednak kilka cech, które je wyróżniają, powodując też, że często ich analiza jest trudniejsza niż w wypadku podobnych badań na innych populacjach.

Badania edukacyjne prowadzone są najczęściej w szkołach. Ma to swoje zalety i wady. W porównaniu z badaniami dorosłych, badania uczniów są stosunkowo tanie, jeżeli koszt policzymy jako stosunek wydanych środków do końcowej liczby obserwacji w bazie danych. Wynika to z tego, że najczęściej na pierwszym etapie losowane są szkoły, wewnątrz których na drugim etapie losowani są do badania uczniowie. Taki schemat losowania nazywamy losowaniem wielopoziomowym. Taką właśnie sytuację przedstawia rysunek 2.1, gdzie przedstawiona została populacja



Rysunek 2.1. Losowanie wielopoziomowe (za: Snijders i Bosker 1999)

zagnieżdżona w 10 grupach. Wybrane jednostki oznaczone zostały przez symbol prostokąta (tak na poziomie grup, jak na poziomie jednostek). W tym schemacie najpierw wylosowano pięć grup, a później z każdej grupy po cztery jednostki.

Koszt realizacji badania związany jest głównie z dotarciem do szkoły, a nie ze zbieraniem danych od uczniów. Jeśli porównamy taki schemat badania z badaniem gospodarstw domowych, to oczywiście jest ono tańsze. W szkole możemy przebadać kilkudziesięciu uczniów, podczas gdy w gospodarstwie tylko kilka osób. Nie znaczy to jednak, że liczba końcowych obserwacji (liczba uczniów) decyduje w pełni o ilości dostępnych informacji. Uczniowie z jednej szkoły mają często bardzo podobne charakterystyki, stąd próba 30 uczniów z jednej szkoły nie jest równoważna próbie 30 uczniów z 30 szkół. Odwołajmy się do prostego przykładu: losujemy najpierw pewną liczbę szkół, a potem badamy wszystkich uczniów z wylosowanych szkół. Okazuje się, że wyniki uczniów, ich status społeczny czy niektóre inne cechy są podobne dla uczniów chodzących do tych samych szkół. Na przykład przeciętnie rodzice uczniów szkół publicznych są mniej zamożni niż rodzice uczniów uczęszczających do szkół prywatnych.

Częstym błędem popełnianym w badaniach edukacyjnych jest zrównanie liczby obserwacji z prób wylosowanych przez losowanie proste i wielopoziomowe. Losowanie proste jest dokonywane najczęściej z listy zawierającej docelowe obserwacje do uwzględnienia w próbie lub też pełnej listy obserwacji w populacji, na przykład przy losowaniu do sondażu wyborczego z listy wyborców. Losowanie wielopoziomowe to w badaniach edukacyjnych najczęściej takie, gdzie najpierw losuje się szkołę, potem klasę, a na końcu uczniów z wylosowanej klasy. Inaczej mówiąc, jest to losowanie, gdzie zamiast końcowej obserwacji (ucznia) losujemy najpierw jednostkę, do której on przynależy razem z innymi potencjalnymi obserwacjami (szkoła lub klasa).

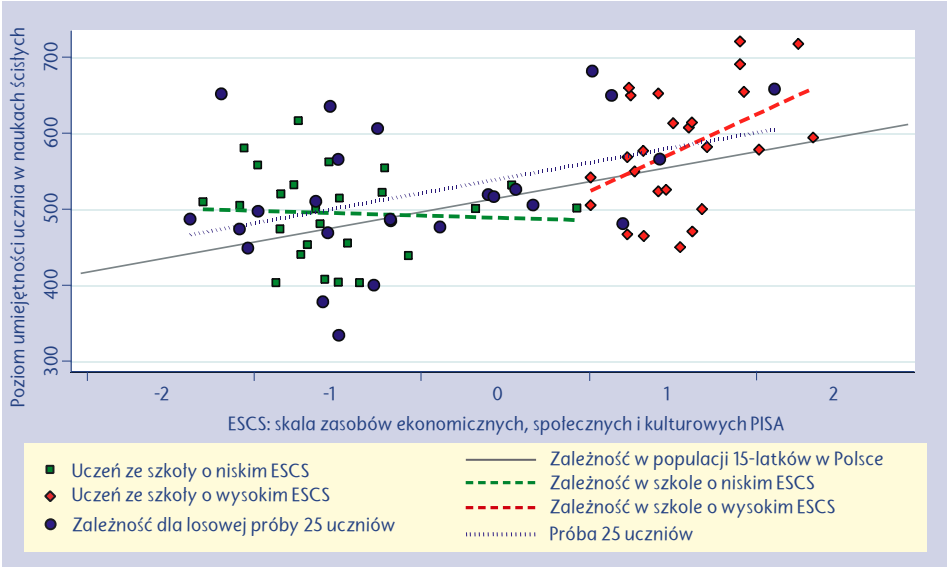
Ilość informacji zebrana w losowaniu wielopoziomowym jest znacznie mniejsza, ze względu na podobieństwo uczniów danej szkoły, które mierzymy tzw. współczynnikiem korelacji wewnątrzszkolnej. O ile znamy wartość takiego współczynnika, na przykład z wcześniejszych badań, to możemy w oparciu o odpowiednie formuły obliczyć wielkość próby niezbędnej do uzyskania interesującej nas precyzji. W oparciu o tego typu obliczenia można stwierdzić, że liczebność ogólna polskiej próby uczniów w doborze zespołowym lub wielopoziomowym powinna być określona na około 3-6 tysięcy uczniów w nie mniej niż 80-100 szkołach, przy czym dokładnie można to określić, znając szczegóły badania. Próba o podobnej precyzji, lecz losowana w sposób prosty, bezpośrednio z listy wszystkich uczniów w Polsce, mogłaby być znacznie mniejsza. Byłaby jednak nieporównanie droższa i znacznie trudniejsza do realizacji, bowiem zamiast około 100 szkół, musielibyśmy zapewne odwiedzić co najmniej kilkaset placówek.

Różnice między próbami wylosowanymi w różnych schematach łatwo pokazać na przykładach. Załóżmy, że chcemy dowiedzieć się, jaki wpływ na wyniki uczniów ma pozycja społeczna ich rodziców. Posłużymy się skrajnym przykładem,

który pokazuje jednak klarownie różnice między badaniami edukacyjnymi a innymi. Przyjmijmy, że nasza próba może wynieść 50 uczniów (co jest oczywiście za małą liczbą dla jakiegokolwiek badania reprezentatywnego). W jednym schemacie losowania wybieramy dwie szkoły, a następnie z każdej z nich jedną klasę i 25 uczniów. W drugim wybieramy 50 uczniów z 50 szkół, czy też 50 gospodarstw domowych, a w nich przeprowadzamy test wiadomości i umiejętności z jednym z dzieci i zbieramy informacje o pozycji społecznej rodziców.

Losując 50 gospodarstw domowych, uzyskamy zapewne dość zróżnicowaną próbę rodziców o bardzo różnym poziomie edukacji i rozmaitej sytuacji zawodowej. Nawet przy tak niewielkiej próbie uzyskamy pewien obraz zróżnicowania pozycji społecznej rodzin oraz umiejętności i wiadomości dzieci. Jeśli jednak wylosujemy jedynie 2 szkoły i 2 klasy (a tym bardziej, jeśli dobierzemy je bez losowania), to istnieje duża szansa, że będą to 2 grupy uczniów z rodzin o różnym statusie społecznym. Wewnątrz wylosowanych klas uczniowie będą jednak zapewne pochodzili z dość podobnych rodzin. Przykładowo, jedna ze szkół będzie się znajdować w inteligenckiej dzielnicy, a druga w dzielnicy robotniczej lub też jedna z klas będzie gromadzić uczniów zdolniejszych, a druga uczniów słabszych, z problemami w nauce. Czy w ten sposób możemy ocenić zależność między pozycją społeczną rodziców a wynikami uczniów? Poniższy rysunek, opierający się na prawdziwych danych, wyraźnie pokazuje interesujący nas problem.

Na rysunku 2.2 przedstawiono wyniki uczniów z próby badania PISA 2006 w Polsce. Na osi poziomej zaznaczono wartości skali ESCS mierzącej zasoby



Rysunek 2.2. Szacowanie zależności między pochodzeniem społecznym a wynikami uczniów. Różnica między próbą uczniów a próbą szkół i uczniów

ekonomiczne, społeczne i kulturowe rodziny ucznia. Na osi pionowej zaznaczono wynik uzyskany przez ucznia w teście z nauk przyrodniczych. Punktami oznaczono obserwacje wylosowane na dwa sposoby. Niebieskie kropki to 25 uczniów wylosowanych z pełnej próby. To zapewne 25 uczniów z 25 różnych szkół, choć może się zdarzyć, że wylosowano 2 lub więcej uczniów z tej samej szkoły. Widać, że obserwacjom tym odpowiadają bardzo różne wartości skali ESCS i wyniki testu. Zielone kwadraty to 25 uczniów jednej z dwóch celowo dobranych szkół, tej o niższym przeciętnym statusie społeczno-ekonomicznym rodzin uczniów. Czerwone romby to natomiast uczniowie z drugiej dobranej szkoły, o przeciętnie wysokim statusie społeczno-ekonomiczym rodzin. Widać na pierwszy rzut oka ogromną różnicę w statusie rodzin uczniów obu placówek, a także w wynikach.

Interesuje nas oszacowanie zależności między skalą ESCS, czyli statusem społeczno-ekonomicznym rodziny ucznia, a osiągnięciami uczniów. Cienka ciągła linia pokazuje, jak zależność ta wygląda w całej próbie PISA 2006, reprezentatywnej dla populacji 15-latków w Polsce. Jest ona niemal równoległa do niebieskiej, kropkowanej linii oszacowanej dla losowej próby 25 uczniów. Tak więc nawet posługując się tą mikroskopijną próbą 25 uczniów, jesteśmy w stanie ocenić niemal bezbłędnie relację między statusem społecznym a wynikami uczniów. Nie jest to przypadek. Co prawda, w każdej 25-osobowej próbie uczniów relacja wyglądałaby inaczej, ale w większości prosta miałaby nachylenie zbliżone do tego w całej populacji.

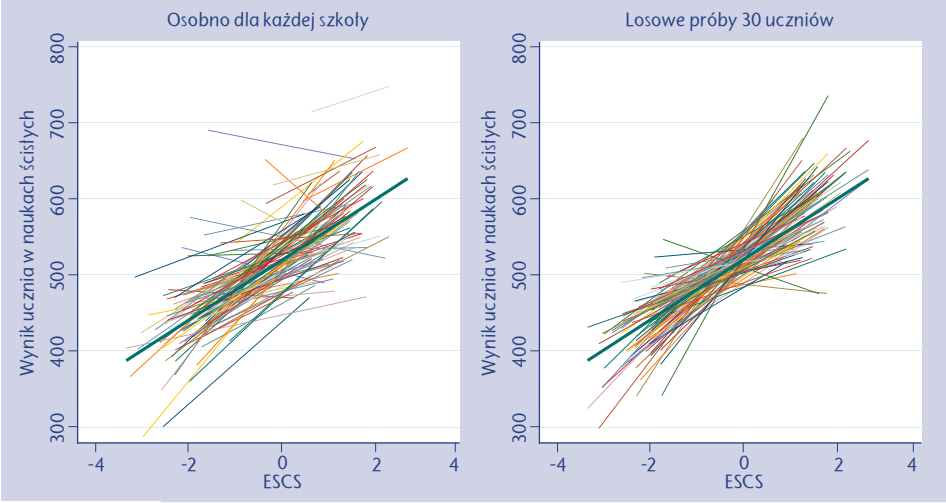
Spójrzmy teraz na relacje w dwóch dobranych szkołach. W jednej z nich, tej o niskim poziomie ESCS, korelacja jest negatywna. W drugiej, tej o wysokim ESCS, pozytywna i nieco silniejsza niż w całej populacji. Oczywiście, szkoły te zostały specjalnie dobrane, żeby ukazać skrajną sytuację. Często jednak zależności wewnątrzszkolne mogą odbiegać od zależności w całej populacji. Ma to kilka konsekwencji, które należy mieć na uwadze, planując badanie czy też analizując dane z już zrealizowanego badania. Po pierwsze, kilkaset obserwacji z kilku szkół nie zawiera zazwyczaj podobnej ilości informacji statystycznej, co kilkaset obserwacji wylosowanych z kilkudziesięciu szkół. Jak już powiedziano powyżej, ze względów finansowych i organizacyjnych wskazane jest losowanie szkół a potem uczniów, należy jednak unikać sytuacji, gdy liczba szkół jest niewielka, nieproporcjonalnie mniejsza niż liczba uczniów.

Czasami, gdy chcemy na przykład badać wpływ składu uczniowskiego szkoły na wyniki ucznia, liczba obserwacji w każdej szkole musi być odpowiednia do oszacowania tych relacji dla każdej placówki. Wciąż jednak potrzebujemy sporej liczby szkół, aby móc wnioskować o populacji uczniów, a jeszcze większej liczby, jeśli chcemy wnioskować o populacji szkół lub nauczycieli (na przykład sposobach organizowania pracy szkół czy metodach nauczania).

Standardowym przykładem wyważenia liczby szkół i uczniów są międzynarodowe badania umiejętności uczniów w rodzaju PISA, PIRLS czy nieobecnego

w Polsce TIMSS. W badaniu PISA 2006 w większości krajów wylosowano około 5000 uczniów, pochodzących z od stu kilkudziesięciu do dwustu kilkudziesięciu szkół. Przeciętnie wylosowano około 30 uczniów na placówkę. Niektóre kraje, takie jak Kanada, Hiszpania czy Włochy, wylosowały znacznie większe próby (w Kanadzie ponad 20 tys. uczniów z około 900 szkół), jednak powodem była tu potrzeba uzyskania wyników reprezentatywnych także dla regionów (prowincji) każdego z tych krajów. Większość krajów zebrała próby o wielkości podobnej do polskiej. Nieco mniej liczne są próby w badaniach PIRLS i TIMSS, gdzie losuje się najpierw szkoły, a potem w zależności od planu losowania danego kraju losuje się jedną lub dwie klasy. W ten sposób zyskujemy możliwość analizy relacji zachodzących w klasach. Podobne schematy losowania są najczęściej stosowane w badaniach mających stanowić podstawę do uogólnienia wniosków dla całej populacji[11].

Jakie niebezpieczeństwo grozi, gdy będziemy analizować dane z pojedynczych szkół, pokazano już na rysunku 2.2. Były to jednak celowo dobrane szkoły, zobaczmy więc, jak podobne relacje wyglądają wewnątrz 100 losowo dobranych szkół (w każdej wylosowano średnio 30 uczniów) i porównajmy z relacjami oszacowanymi dla 30 losowo dobranych uczniów z całej populacji (najczęściej z 30 różnych szkół). W obu wypadkach liczba obserwacji jest niewielka, co dobrze pokazuje także problem szacowania jakichkolwiek zależności dla niewielkich prób. Spójrzmy na wykres poniżej (rysunek 2.3). Pokazano na nim po lewej stronie 100 zależności oszacowanych niezależnie w każdej ze 100 szkół, jedynie na podstawie



Rysunek 2.3. Zależności między statusem społeczno-ekonomicznym rodziny ucznia a jego osiągnięciami oszacowane osobno dla 30 uczniów z każdej z wylosowanych 100 szkół (po lewej) oraz 100 razy dla 30 uczniów, wylosowanych po jednym uczniu z 30 różnych szkół (po prawej)

[11] Przykłady schematów doboru próby dla Polski można znaleźć w rozdziale 5.

danych dla 30 uczniów z tej placówki, a po prawej 100 zależności oszacowanych dla 30 uczniów wylosowanych pojedynczo z 30 różnych szkół. Gruba linia pokazuje zależność w całej populacji, a oba wykresy mają identyczną skalę, żeby nie zniekształcać porównania. Zależności oszacowano metodą regresji liniowej, gdzie do danych dopasowywana jest najlepiej je opisująca prosta, w tym wypadku prosta określająca zależność między statusem społeczno-ekonomicznym a wynikami ucznia.

Widać, że jeśli chodzi o określenie zależności na poziomie uczniów, znacznie bardziej spójne oceny daje nam bezpośrednie losowanie uczniów z całej populacji. Jak napisano powyżej, w badaniach najczęściej obierany jest złoty środek w postaci losowania odpowiednio dużej liczby szkół a potem uczniów. Często jednak z braku funduszy lub też ze względu na łatwość organizacyjną, badania realizowane są na niewielkiej liczbie szkół, lecz dużej liczbie uczniów (na przykład wszyscy uczniowie każdej placówki). Wykresy pokazują, że efektywna liczba informacji zdobyta w taki sposób jest znacznie niższa. Potwierdzają to także odpowiednie statystyki.

W naszym przykładzie przeciętna ocena współczynnika regresji określającego zależność między wynikami a statusem rodziny ucznia wyniosła dla prób 30 uczniów z 30 różnych szkół niemal tyle samo, co prawdziwa wartość w populacji (równa 41 punktów i odpowiadająca różnicy w punktach związanej ze zmianą wartości indeksu PISA ESCS o jedno odchylenie standardowe). Średnia ocena zależności uzyskana odrębnie w każdej z wylosowanych szkół wyniosła o 6 punktów mniej. Co więcej, wariancja, a więc zróżnicowanie ocen uzyskanych w każdej ze 100 próbek, była niemal o 1/3 większa dla oszacowań dokonanych osobno w każdej szkole niż dla oszacowań dokonanych na próbkach 30 uczniów z 30 szkół (439 w stosunku do 342 punktów). Oznacza to dokładnie tyle, że gdy opieramy się na zależnościach w pojedynczych szkołach, prawdopodobieństwo popełnienia znaczącego błędu w ocenie zależności w całej populacji jest dużo większe, niż gdybyśmy opierali się na podobnej wielkości próbie uczniów, ale wylosowanej ze znacznie większej liczby szkół.

Czy sytuacja byłaby nieco lepsza, gdybyśmy losowali kilka szkół? Nie, także wtedy, o ile liczba tych szkół byłaby wciąż niewielka, bezpośrednie losowanie uczniów daje dużo lepsze rezultaty. Niezbędne jest więc takie wyważenie próby, aby zarówno liczba szkół, jak i liczba uczniów odpowiadały zarówno wymogom analizy, jak i możliwościom organizacyjno-finansowym.

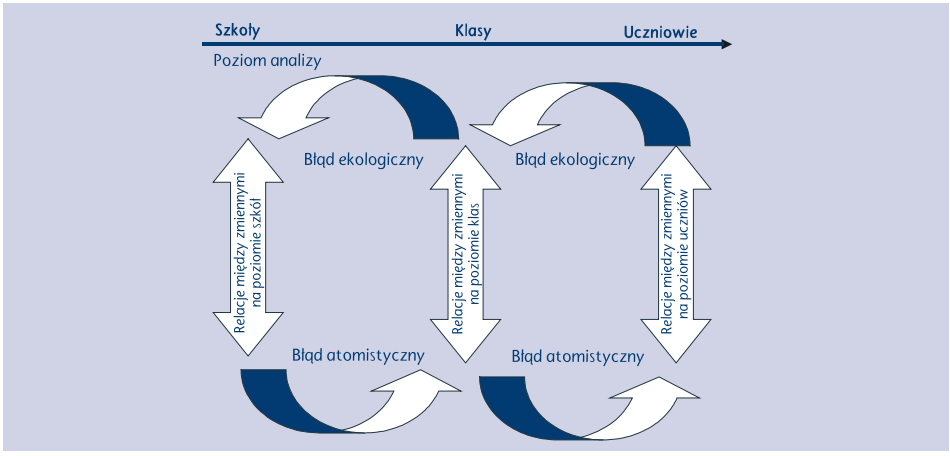
Przykłady te zmuszają do zastanowienia się nad wiarygodnością badań prowadzonych przez dobranie kilkunastu szkół i przetestowanie wszystkich uczęszczających do nich uczniów, zamiast losowania dużej liczby szkół, a potem przebadania odpowiedniej liczby uczniów w każdej szkole. Pokazują, że wnioskowanie o nawet najprostszych zależnościach może być obarczone sporym błędem wynikającym jedynie z niepoprawnie dobranego schematu losowania. Należy to mieć na uwadze, analizując wyniki badań oświatowych i oceniając ich wiarygodność.

Błąd ekologiczny i atomistyczny

Często przywoływanymi błędami wnioskowania w naukach społecznych są błąd ekologiczny i błąd atomistyczny. Błędy te dotyczą nieuprawnionego wnioskowania o relacjach z jednego poziomu na temat relacji z drugiego poziomu analizy, czyli na przykład poziomu szkół i poziomu uczniów. Problem ten ilustruje rysunek 2.4 dla trzech przykładowych poziomów analizy: szkół, klas i uczniów. Kwestie te są o tyle istotne dla badań edukacyjnych, że prawie zawsze mamy tu do czynienia z co najmniej dwoma poziomami analiz: szkołą i uczniami, a często dochodzą dodatkowe, na przykład poziom klas czy gmin.

W wypadku błędów ekologicznego i atomistycznego nie chodzi o nieuprawnione połączenie poziomów, lecz o bezpośrednie przenoszenie właściwości z jednego poziomu na drugi poziom. Chodzi więc o nieuzasadnione wnioskowanie na temat zależności na jednym poziomie analizy na podstawie podobnej relacji, lecz na innym poziomie analizy.

Sztandarowym przykładem jest tu praca Robinsona[12], który pokazał, iż nieuprawnionym jest przenoszenie związków z wyższego poziomu na niższy. Robinson pokazał, że choć związek między procentowym udziałem ludności afroamerykańskiej w danym dystrykcie a procentowym udziałem analfabetów wydawał się wysoki (wysoka korelacja związana z tym, że w dystryktach, gdzie zamieszkiwało wielu czarnoskórych Amerykanów, procent analfabetów był wysoki), to okazało się jednak, że na poziomie indywidualnym związek ten był wyjątkowo słaby. Analfabetami w tych dystryktach byli mniej więcej w takim samym odsetku czarnoskórzy, jak biali. Inaczej mówiąc, dystrykty z dużą liczbą



Rysunek 2.4. Ilustracja przedstawiająca wnioskowania obarczone błędem ekologicznym i atomistycznym

[12] Robinson *W. S., Ecological Correlations and the Behavior of Individuals. American Sociological Review 15: 351–357, 1950.*

spis treści

analfabetów były częściej zamieszkane przez ludność afroamerykańską, jednak mieszkający tam biali także byli częściej niż gdzie indziej analfabetami.

Tak więc porównanie między dystryktami, gdzie wysoki procent analfabetyzmu wiązał się z wysokim procentem ludności czarnoskórej tam zamieszkałej, wcale nie dowodzi faktu, że podobna zależność istniała na poziomie indywidualnym. Biorąc pod uwagę różnice w analfabetyzmie między dystryktami, relacja ta była znacznie słabsza. Inaczej mówiąc, patrząc na tę zależność osobno wewnątrz każdego dystryktu i uśredniając ją pomiędzy wszystkimi dystryktami, wcale nie potwierdzono, że to Afroamerykanie są częściej analfabetami niż biali.

Odwrotna sytuacja też może mieć miejsce – zależność odkryta na niższym poziomie wcale nie musi być odzwierciedlana na wyższym poziomie. Dobrym przykładem, który jest szczególnie istotny w Polsce, jest różnica między gminami miejskimi i wiejskimi. Jeśli spojrzymy na przykład na zależność między wielkością klas a wynikami uczniów na poziomie gmin, to okaże się, że jest ona silnie pozytywna. Mniejsze klasy są w szkołach wiejskich, a większe w szkołach miejskich, gdzie również wyniki są wyższe. Jeśli jednak spojrzymy na różnice wewnątrz poszczególnych gmin, a tym bardziej wewnątrz poszczególnych szkół, to okaże się, że zależności takowej brak lub co najwyżej jest ona lekko negatywna. Uczniowie chodzący do większych klas w tej samej szkole wcale nie uzyskują lepszych wyników, o ile mają podobne pochodzenie społeczne, umiejętności i pozostałe warunki nauczania.

Na poziomie gmin czy szkół często obserwujemy zależności pozorne wynikające z różnic w statusie społeczno-ekonomicznym, zasobach kulturowych i dochodach ludności. Zależności bardzo często okazują się zupełnie różne wewnątrz gmin czy szkół, gdzie uczniowie najczęściej pochodzą z podobnego środowiska. Do kwestii tych jeszcze powrócimy w rozdziale 3 i 4.

Paradoks Simpsona

Rozwinięciem błędu atomistycznego jest tzw. paradoks Simpsona, który także dotyczy interpretacji zależności stwierdzanych na różnych poziomach analizy. Przyjrzyjmy się zmyślonym danym dotyczącym programu przekonywania uczniów do segregacji odpadów (Julious, Mullee 1994)[13] prezentowanych w tabeli 2.1. Uczniowie podzieleni zostali na cztery grupy wedle dwóch kategorii: szkoły, do której uczęszczają (1 i 2), oraz wedle sposobu przekonywania (program A i program B). Dla każdej z grup podany został procent uczniów, którzy zadeklarowali segregację odpadów (uznajemy ten fakt jako sukces) oraz stosunek uczniów przekonanych do wszystkich uczniów.

Jak widać, program A w każdej grupie (ze względu na rodzaj szkoły) jest bardziej efektywny. W przypadku szkoły 1 program A był skuteczny w 93%,

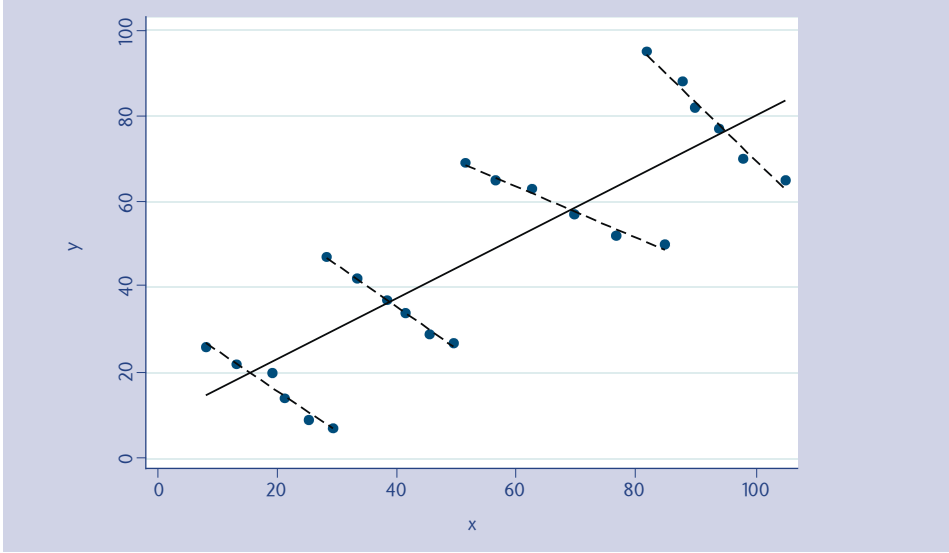
[13] Julious S., Mullee M., (1994). *Confounding and Simpson's paradox*. „BMJ" 309 (6967), strony 1480-1481

Tabela 2.1. Procent sukcesów w segregacji odpadów dla dwóch programów w czterech grupach uczniów (za: Julious, Mullee 1994)

	Program A	Program B	Razem
Szkoła 1	Grupa 1 93% (81/87)	Grupa 2 87% (234/270)	82% (315/357)
Szkoła 2	Grupa 3 73% (192/263)	Grupa 4 69% (55/80)	72% (247/343)
Razem	78% (273/350)	83% (289/350)	

a program B w 87%. W przypadku szkoły 2 program A był skuteczny dla 73% uczniów, zaś program B jedynie dla 69% przypadków. Pomimo takich wyników, dla wszystkich badanych uczniów zależność była przeciwna. Można to obliczyć, posługując się liczbami uczniów podanymi w nawiasach. Program A okazał się skuteczny jedynie dla 78% uczniów, podczas gdy program B aż dla 83%. Wnioski, które możemy wyciągnąć na podstawie dwóch badanych szkół, rozpatrywanych osobno, nie są takie same, jak wnioski, jakie moglibyśmy uzyskać, rozpatrując wszystkich uczniów razem.

Paradoks Simpsona oraz błędy atomistyczny i ekologiczny są w istocie rzeczy blisko ze sobą związane. Ilustruje to rysunek 2.5. Cała populacja licząca 24 jednostki została podzielona na 4 grupy. Chociaż w całej populacji związek między x i y jest dodatni, to w każdej z czterech wyszczególnionych podgrup, jeżeli analizujemy je osobno, sprawa ma się inaczej. Korelacja między x i y w każdej z grup jest ujemna.



Rysunek 2.5. Różnica w zależnościach wewnątrz grup, między grupami i w całej populacji

Często wyniki różnią się między poziomem szkół i uczniów, a od nas zależy, który poziom uznamy za bardziej adekwatny. Za każdym razem, gdy analizujemy dane dla szkół i uczniów, należy rozważyć podobne kwestie i podać wyniki na odpowiednim poziomie analizy. Często wyniki na obu poziomach mogą być interesujące, nawet gdy wzajemnie sobie przeczą. Kwestie te zawsze obecne są w badaniach edukacyjnych, z tego też względu opracowano specjalne metody potrafiące „rozdzielić" poziomy, analizując zależności osobno dla każdego poziomu w jednym modelu. Są to tzw. modele wielopoziomowe, o których będziemy jeszcze w tej książce mówić.

Losowość doboru próby i reprezentatywność danych

Omówione powyżej kwestie są dość złożone, może więc należałoby zacząć od kwestii zasadniczej, wydającej się oczywistą, ale często nieodpowiednio rozumianej w badaniach edukacyjnych. Jest to zasada w pełni losowego doboru próby.

Wiemy już, że poprawny schemat badania edukacyjnego polega w najprostszej wersji na losowaniu szkół, w kolejnym kroku ewentualnie klas, a potem uczniów. Często dochodzi do tego tzw. warstwowanie, czyli na przykład odrębne losowanie szkół wiejskich czy miejskich, czy też szkół prywatnych i publicznych. Jednak nawet przy najbardziej wyrafinowanym schemacie losowania podstawowe jest zachowanie losowości doboru na każdym etapie. Jedynie w wypadku uczniów jednej klasy można włączyć ich w całości do próby. Zarówno dobieranie szkoły, jak i klasy lub uczniów wewnątrz szkoły muszą być w pełni losowe.

Nie są to pryncypia łatwe do zachowania przy realizacji badania. Niektóre szkoły odmawiają udziału, a inne, nawet niewylosowane, same chętnie się do badania zgłaszają. Niezgodne z zasadami jest jednak zastępowanie szkół odmawiających udziału innymi szkołami, o ile wcześniej nie stworzono zapasowej listy szkół podobnych[14]. Karygodne jest zastępowanie ich szkołami zgłaszającymi się, ale niewylosowanymi. Podobnie nieprawidłowe jest dobieranie klas czy uczniów wedle ich dostępności czy w zależności od zgody dyrektora.

Przed wkroczeniem do szkoły ankieterzy muszą posiadać odpowiednie instrumenty do losowania klas wewnątrz szkoły, a także do losowania uczniów, a wszelkie odstępstwa muszą być odnotowywane. Znowu karygodne, a niestety częste w praktyce, jest zastępowanie innymi klas czy uczniów wylosowanych, ale niemożliwych do przebadania.

Kwestie te dotykają najczęstszych błędów popełnianych w praktyce badań oświatowych. Ich szczegółowa znajomość niezbędna jest raczej osobom prowa-

[14] Wielu specjalistów uważa nawet, że nie powinno się tworzyć „rezerwowych" list szkół do użycia w przypadku odmowy udziału w badaniu, a w razie odmowy po prostu badamy mniejszą liczbę szkół. W praktyce często takich list się używa, poświęcając nieco pryncypia doboru próby dla uzyskania jej docelowej wielkości.

dzącym badania niż tym, którzy tylko są zainteresowani wynikami, dlatego nie będziemy ich już dalej rozwijać. Ważne jest jednak, aby w pełni rozumieć te zagadnienia i móc samemu ocenić, czy badanie, które dostarczyło nam wniosków, którymi chcemy się posłużyć, zostało przeprowadzone z zachowaniem losowego doboru próby. Czy liczba szkół jest wystarczająca do oceny zależności w sposób wiarygodny? Czy dobór jednostek badania był w pełni losowy? To podstawowe pytania, jakie należy sobie zadać, planując badanie czy też korzystając z wyników już przeprowadzonego badania.

Krótko mówiąc, jeżeli obserwacje, które poddano badaniu, nie zostały dobrane losowo, to wszelkie wyniki mogą być łatwo podważone i nie mogą być uogólniane na całą populację. Przykłady odstępstw od losowości doboru próby, pochodzące z rzeczywistych doświadczeń przy realizacji badań edukacyjnych, przedstawiono poniżej. Warto zrozumieć zarówno pryncypia, jak i zapoznać się z przykładami, bo często odstępstwa od losowości doboru próby nie są oczywiste i łatwo uznać za reprezentatywne badanie, które nie spełnia podstawowych wymogów losowego doboru próby.

Przeglądając wyniki badań edukacyjnych czy też korzystając z informacji o sondażach w mediach, często natrafiamy na określenie „badanie reprezentatywne". Jest to jedno z najczęściej nadużywanych pojęć w naukach społecznych. Ze względów marketingowych firmy badające rynek i opinie zawsze twierdzą, że ich badania są reprezentatywne, choć potem okazuje się, że na przykład sondaż przedwyborczy mylnie wskazał zwycięzcę. Kiedy więc możemy powiedzieć, że badanie jest reprezentatywne i co to właściwie znaczy?

Najkrócej mówiąc, badanie jest reprezentatywne, jeśli w oparciu o dane w nim zebrane jesteśmy w stanie odtworzyć wartości interesujących nas cech populacji, z odpowiednią dla naszych celów precyzją. Często jednak nie znamy prawdziwych wartości tych cech, a badanie zostało przeprowadzone właśnie po to, aby je określić. Wtedy reprezentatywność badania może oznaczać, że podstawowe charakterystyki, przede wszystkim tzw. charakterystyki demograficzne, mają w naszej próbie podobne wartości do tych w populacji.

W każdym niemal badaniu posiadamy (a jeśli nie, to znaczy, że zostało ono błędnie zaplanowane) podstawowe informacje o płci, wieku, kategorii wielkości miejscowości zamieszkania, a także o wykształceniu respondentów. Są to cechy, o których mamy dokładne dane dla całej populacji. Wiemy, ile jest w Polsce osób określonej płci i w danym wieku. Wiemy, ile osób mieszka na wsi, a ile w małych i dużych miastach. Wiemy też, ile jest osób z wykształceniem wyższym, a ile z zawodowym. Co więcej, posiadamy takie dane w zestawieniach krzyżowych, czyli wiemy na przykład, ile jest kobiet z wykształceniem wyższym, a ile z zawodowym. Dodatkowo, informacje te możemy uzyskać w rozbiciu na województwa lub regiony statystyczne. Są one dostępne w publikacjach i bazach danych Głównego Urzędu Statystycznego. Możemy więc je wykorzystać, aby określić, czy dane badanie potrafi odtworzyć rozkład cech w populacji.

Mówiąc najprościej, posiadając dane dla całej Polski i sprawdzając reprezentatywność badania dla całego kraju, możemy sprawdzić, czy procent kobiet z wykształceniem wyższym, w wieku 40-65 lat, mieszkających w województwie lubelskim jest taki sam w naszej próbie, jak w rzeczywistości. Oczywiście wartości będą się nieco różniły. Rzetelnie przeprowadzone badanie powinno jednak prezentować odpowiednie porównania poparte testami statystycznymi (na przykład na różnice między średnimi wartościami z próby zebranej w badaniu a wartościami w populacji, na przykład między procentem kobiet i mężczyzn w próbie a w populacji). Jeśli dla odpowiednio dużej liczby cech nasza próba ma podobne rozkłady jak w populacji, to możemy stwierdzić, że badanie jest reprezentatywne dla tej populacji.

W badaniach edukacyjnych szkół i uczniów lista cech demograficznych jest nieco krótsza. Zróżnicowanie wieku jest mniejsze, często, jak na przykład w badaniu PISA, uczniowie są z tego samego rocznika. Co więcej, rzadko kiedy posiadamy wiarygodną informację o wykształceniu rodziców ucznia. Najczęściej informacja ta zbierana jest wśród uczniów i z doświadczenia wiemy, że odbiega od rzeczywistych wartości. Można to łatwo stwierdzić, porównując na przykład rozkłady wykształcenia podawane przez uczniów w badaniu PISA z rzeczywistymi rozkładami wykształcenia dorosłych posiadających dzieci uczęszczające do gimnazjum czy o zbliżonym wieku, pochodzącymi z innych badań. Różnice są spore.

Każde badanie edukacyjne powinno jednak bez problemu zgromadzić informacje o płci ucznia, typie miejscowości zamieszkania (wystarczy informacja o lokalizacji szkoły) i regionie. Możemy wykorzystać przynajmniej te dane do sprawdzenia reprezentatywności próby z badania. Rzetelnie przeprowadzone badanie powinno taką analizę reprezentatywności zawierać w dokumentacji lub przynajmniej prezentować rozkłady podstawowych charakterystyk demograficznych z próby. Sprawdzenie reprezentatywności względem cech demograficznych jest sprawą naprawdę prostą i powinno być podstawowym krokiem przed każdą analizą danych czy wykorzystaniem końcowych wyników badania.

Niekiedy okazuje się, że badanie, z którego korzystamy, nie przechodzi powyżej opisanych testów. Niestety, osoby często korzystające z badań wiedzą, że nawet często cytowane i uznane badania okazują się przy bliższym spojrzeniu niereprezentatywne. Należy wtedy upewnić się, że przeprowadzając testy, postępujemy zgodnie z metodami przewidzianymi do analizy tego zbioru danych[15]. Po

[15] Jeśli na przykład w zbiorze istnieją wagi probabilistyczne, to powinniśmy je uwzględnić w naszych analizach. Wagi te niemal zawsze dostępne są dla danych pochodzących z losowania wielopoziomowego, a obecnie ich wykorzystanie umożliwiają niemal wszystkie popularne pakiety statystyczne. Nieuwzględenienie wag powoduje, że wszelkie oszacowania będą obarczone błędem. Drugim błędem często popełnianym przy analizie danych o strukturze wielopoziomowej jest brak korekty błędów standardowych, a przez to brak precyzji testów statystycznych, ze względu na pogrupowanie jednostek. W naszym przypadku najczęściej będzie to pogrupowanie uczniów w szkołach, o którym już kilkakrotnie wspominaliśmy. Coraz więcej pakietów statystycznych umożliwia prawidłową analizę także pod tym względem.

drugie, jeśli wciąż wydaje nam się, że rozkład cech z próby znacząco odbiega od tego w populacji, należy zastanowić się, czy na pewno porównujemy te same populacje. Być może zebrana próba jest na przykład reprezentatywna dla uczniów trzecich klas szkół podstawowych, a my porównujemy ją z rozkładem cech dla populacji 12-latków.

Reprezentatywność jest pojęciem określonym względem pewnej populacji. Nie ma badania uczniów, które byłoby reprezentatywne dla dowolnej ich grupy. Przykładowo, badanie PISA jest reprezentatywne dla populacji 15-latków w każdym kraju, podczas gdy badanie TIMSS jest reprezentatywne dla 4- i 8-klasistów.

Jeśli wciąż jednak wątpimy, że nasza próba ma podobny rozkład danej cechy jak w populacji, którą ma reprezentować, to musimy się zastanowić, czy cecha ta może mieć istotne znaczenie w naszej analizie. Jeśli na przykład interesuje nas wpływ statusu społecznego uczniów na ich wyniki, to nie musimy się martwić, że nasza próba zawiera nieco więcej dziewcząt. Chyba że wiemy z innych źródeł, że chłopcy i dziewczęta różnią się pod względem relacji między statusem a wynikami. Jeśli dodatkowo relacja ta jest odmienna względem płci w analizach z naszej próby, to mamy powód do obaw. Musimy wtedy albo uwzględnić płeć ucznia w analizach, albo też zastosować bardziej skomplikowane metody ponownego ważenia obserwacji tak, aby uzyskać odpowiedni balans względem płci w naszej próbie[16].

Na pewno więcej obaw można mieć pod względem różnic między próbą a populacją, względem procentu uczniów ze szkół w różnych lokalizacjach. Jeśli nasza próba posiada na przykład więcej uczniów ze szkół wiejskich i zastosowanie wag tego nie zmienia, to większość analiz może być podważona, bowiem lokalizacja szkoły jest zawsze silnie związana zarówno z osiągnięciami uczniów, jak i z cechami ich rodzin czy dostępem do zasobów kulturowych. W takiej sytuacji wszelkie analizy odnoszące się do wymienionej powyżej zmiennej muszą być interpretowane bardzo ostrożnie.

Reprezentatywność oznacza wreszcie to, że wnioski z próby mogą być uogólniane na całą populację. Jeśli w sondażu przedwyborczym kandydat A uzyskuje 19% z błędem w granicach 2%, to oczekujemy, że w całej populacji wyborców ma on takie właśnie poparcie – w okolicach 19%. Jeśli w pilotażu egzaminu zewnętrznego stwierdzamy różnicę w poziomie umiejętności matematycznych między chłopcami i dziewczętami wynoszącą 2,5 punktu z przedziałem ufności 2,2 do 2,8, to oczekujemy, że egzamin przeprowadzony w całej populacji stwierdzi także różnicę w przedziale 2,2-2,8. Jeśli jednak badamy uczniów szkół wiejskich i stwierdzamy, że różnica między osiągnięciami dzieci rodziców z wykształceniem wyższym a z wykształceniem zawodowym wynosi 6 punktów, to

[16] Intucyjnie, przez ważenie obserwacji można rozumieć nadawanie większej wagi w obliczeniach obserwacjom rzadszym w naszej próbie niż w populacji i zmniejszanie wagi obserwacji częstszych w próbie niż w populacji. Tematyka ta jest dość skomplikowana i wykracza poza zakres tego opracowania, warto jednak pamiętać, że wagi powinny być stosowane w niemal każdym badaniu reprezentatywnym, w szczególności jeśli pochodzi ono z losowania wielopoziomowego.

nie możemy twierdzić, że różnica ta w całej populacji uczniów wynosi właśnie tyle. Z pewnością jest ona inna na wsi i w mieście, i próba uczniów ze szkół wiejskich w żaden sposób nie może nam posłużyć do oceny tej różnicy dla populacji uczniów ze szkół miejskich.

Reprezentatywność gwarantuje więc, że dla badanej populacji jesteśmy w stanie odtworzyć interesujące nas rozkłady charakterystyk, różnice w poziomach umiejętności czy siłę zależności między wybranymi cechami. Nie gwarantuje jednak, o czym niestety często zapominają osoby korzystające z wyników badań czy też sami badacze prezentujący ich wyniki w mediach, że wnioski możemy uogólniać na dowolne populacje. Nie jest to prawomocne, podobnie jak wtedy, gdy mamy do czynienia z nielosowo dobraną próbą. O tym piszemy już jednak w kolejnym podrozdziale.

2.2. Próby nielosowe: dobór celowy i dane zastane

Podstawową zaletą losowego doboru próby jest to, że w ten sposób zapewniamy zrównoważenie cech zarówno obserwowalnych, jak i nieobserwowalnych. Inaczej mówiąc, dzięki losowości doboru próby mamy jednakowe szanse wylosowania osoby o danej wartości cechy, którą potrafimy zmierzyć, ale co ważniejsze, także jednakowe prawdopodobieństwo wylosowania osoby o wartości cechy, której nie potrafimy zmierzyć (taką cechę zwyczajowo nazywany nieobserwowalną). Dzięki w pełni losowemu doborowi możemy zagwarantować, że nasza próba będzie reprezentatywna, tzn. że będzie miała taki sam rozkład najważniejszych cech, jak w populacji.

Oczywiście w praktyce każde badanie terenowe narażone jest na odstępstwa od pryncypiów, co zaburza losowość doboru, a w konsekwencji jego reprezentatywność. Istnieje jednak szereg sposobów, aby odstępstwa te miały minimalny wpływ na reprezentatywność naszej próby (m.in. dodatkowe ważenie obserwacji). Zachowanie podstawowych zasad doboru i realizacji próby powoduje, że dla odpowiednio dużej liczby zebranych obserwacji dane mają charakterystyki zbliżone do populacji. Jeśli jednak na którymś z etapów nie zachowaliśmy zasady losowości doboru próby i nie możemy tego skorygować, to wszelkie wnioskowanie w oparciu o zebrane dane ma wątpliwą wartość. Przede wszystkim może być łatwo podważone i uznane za wynikające z nielosowości zebranych danych.

Istnieją właściwie dwie sytuacje, gdy dobór nielosowy ma uzasadnienie. Pierwsza związana jest z analizami populacji nie do końca poznanych. Czasem na przykład nie chcemy określić wartości nasilenia problemu, o którym wiemy, że istnieje, ale raczej szukamy listy problemów uczniów, nie wiedząc nawet, czy one występują. Wtedy możemy prowadzić badania mające naturę „rozpoznawczą", niepretendujące do miana reprezentatywnych. Podobnie, jeśli chcemy zro-

zumieć zagadnienie dotyczące grupy uczniów, o której niewiele wiemy. Wtedy możemy prowadzić wywiady na przykład metodą kuli śniegowej, gdzie respondent wskazuje kolejne osoby do badania.

Są to sytuacje częste w badaniach jakościowych, których niniejsza publikacja nie dotyczy. Warto jednak podkreślić, że w wypadku takich badań niemożliwe jest uogólnianie zebranych w nich wniosków. Służą one głębszemu poznaniu określonych zagadnień, co nie jest możliwe w badaniach ilościowych, szczególnie jeśli mamy problemy z kwantyfikacją ważnych cech. Jednak karygodne jest, niestety dość częste w naukach społecznych, uogólnianie wniosków z wywiadów przeprowadzonych w 5 zaprzyjaźnionych szkołach na populację uczniów całego kraju. Szczególnie, jeśli podawane są liczbowe wartości określające na przykład nasilenie pewnych zjawisk.

Druga sytuacja, ściślej związana z interesującymi nas tutaj badaniami ilościowymi, występuje, gdy poddajemy analizie dane już zastane czy też mamy szansę realizacji badania tylko na z góry założonej grupie uczniów. Przykładowo, w 5 szkołach przeprowadzono badanie ankietowe wszystkich uczniów. Sięgnięcie po wyniki ankiet daje nam szansę uzyskania danych do analizy bez dodatkowych kosztów. Jeśli nie posiadamy środków na realizację własnego badania, to zapewne będziemy chcieli skorzystać z takiej możliwości. Trudno dyskutować z argumentem, że „lepszy rydz niż nic". Trzeba jednak podkreślić, że każde tego rodzaju badanie nie może być tak po prostu uogólnione na szerszą populację uczniów. Jeśli było to badanie w 5 szkołach, to wnioski płyną tylko w stosunku do tych 5 szkół, co gorsza, zapewne w stosunku do uczniów, którzy już do nich nie uczęszczają.

Poważniejszy argument dotyczy w tym wypadku zasadności stosowania jakichkolwiek narzędzi analizy statystycznej. Jeśli bowiem dane nie pochodzą z losowego doboru próby, to wszelkie podstawowe pojęcia statystyki tracą sens. Jest to argument dość złożony, wrócimy do niego w części traktującej o badaniach danych z całej populacji.

Czy można jednak coś zrobić, żeby zwiększyć wiarygodność rezultatów uzyskanych na dobranej nielosowo próbie? Jak upewnić się, chcąc skorzystać z rezultatów takiego badania, czy płynące z niego wnioski możemy wykorzystać do własnych dociekań? Czy badanie stwierdzające efektywność jakiejś metody nauczania, ale opierające się na analizach nauczyciela w 3 zaprzyjaźnionych szkołach, można uogólnić na własną sytuację? Tutaj wkraczamy w sferę sztuki, a nie nauki, jednak nie tylko nasza intuicja może nam podpowiedzieć, czy wyniki z takiego badania mogą być dla nas pomocne. Powinniśmy postępować podobnie jak w wypadku analizy reprezentatywności losowo dobranej próby. Należy sprawdzić, czy rozkłady najważniejszych cech w próbie z tego badania są podobne do rozkładów w populacji lub też choćby w naszej placówce.

Trzeba też głębiej zastanowić się, jakie różnice mogą występować między taką próbą, a interesującą nas grupą uczniów czy nauczycieli. Jeśli różnice te związane są z poziomem umiejętności uczniów, ich pochodzeniem społecznym

lub innymi cechami mogącymi wpływać na przykład na efektywność zastosowania badanych metod nauczania, to wtedy należy sugestie płynące z takiego badania traktować ostrożnie. Podobną analizę musimy też jednak przeprowadzić, próbując przełożyć wyniki badania reprezentatywnego na nasze własne problemy. Po prostu badanie reprezentatywne dla uczniów z całej Polski może być mało przydatne, jeśli uczymy w miejskiej szkole uczniów wybitnych lub głównie uczniów z problemami w nauce. Wtedy musimy szukać analiz z badania reprezentatywnego dla populacji, które przeprowadzono z myślą o takich podgrupach uczniów. Tylko takie analizy mogą być dla nas pomocne.

Dla osób korzystających z wyników badań, chcących zastosować płynące z nich sugestie w swojej pracy, szukających inspiracji do wprowadzania nowych metod nauczania, których efektywność potwierdzają ponoć wyniki badań, najważniejsze nie jest zapewne zrozumienie, jak dobrze zaplanować realizację badania, ale czy wyniki badania, które nas zainteresowały, są wiarygodne. Opisane powyżej kwestie miały temu właśnie służyć. Tylko rozumiejąc, jak powinno się prawidłowo zrealizować badania, można też określić ich wiarygodność. Do pełnego zrozumienia tych kwestii potrzeba głębokiej specjalistycznej wiedzy. Można też jednak przejrzeć dokumentację badania, zwracając uwagę na opisane powyżej zagadnienia.

Poniżej podajemy przykłady losowego i nielosowego doboru próby. Przeanalizowanie podobnych kwestii dla dowolnego badania nie powinno stwarzać trudności nawet osobom nieobeznanym z podstawami statystyki. Przykłady te odpowiadają na ogólne pytanie o to, czy próba została dobrana losowo, a dokładniej, czy na każdym etapie dobierania próby badane jednostki były wybierane drogą w pełni losową.

Przykłady w pełni losowego doboru próby:

- Do badania z wcześniej przygotowanej listy wszystkich szkół w województwie wylosowano kilkadziesiąt placówek. Następnie w każdej placówce wylosowano po 2 klasy, stosując wcześniej przygotowane losowe listy klas dla każdej szkoły. W każdej wylosowanej klasie wybrano losowo 20 uczniów, których poddano badaniu. Taka próba powinna być reprezentatywna dla populacji uczniów tego województwa.
- Listę szkół podstawowych z całej Polski podzielono na 4 „warstwy" według kategorii komisji egzaminacyjnych: szkoły wiejskie, szkoły z małych miast, szkoły ze średnich miast, szkoły z dużych miast. Docelową próbę około 5000 uczniów z 200 szkół podzielono między warstwy tak, aby w każdej z nich znalazł się taki procent próby, jak w całej populacji uczniów w Polsce (na przykład odsetek uczniów szkół wiejskich w próbie był taki sam jak odsetek uczniów szkół wiejskich w populacji). W szkołach wylosowano następnie po 25 uczniów klas szóstych lub poddano badaniu wszystkich, jeśli było ich mniej. Taka próba powinna być reprezentatywna dla populacji 6-klasistów w Polsce.

■ W gminie stworzono listę wszystkich uczniów uczęszczających do kilkunastu gimnazjów publicznych. Następnie z listy tej wylosowano 500 uczniów i poddano badaniu kwestionariuszowemu ich samych oraz rodziców. Taka próba powinna być reprezentatywna dla populacji gimnazjalistów tej gminy.

Przykłady nielosowego lub nie do końca poprawnego doboru próby:

■ Do badania z wcześniej przygotowanej listy wszystkich szkół w województwie wylosowano kilkadziesiąt placówek. Następnie w każdej placówce badaniu poddano uczniów klas z pierwszą i ostatnią literą. Błąd: klasy nie były losowane i wyniki badania będą mało reprezentatywne, bowiem często klasy A i klasy oznaczone ostatnią literą są wyjątkowe.

■ Listę szkół z całej Polski podzielono na 2 grupy: szkoły wiejskie oraz szkoły miejskie. Następnie z każdej grupy wylosowano po 20 szkół i poddano badaniu wszystkich ich uczniów. Błędy: liczba szkół może być zbyt niska, aby zapewnić odpowiednią wariancję cech uczniów; trudno też ocenić, jaka będzie ostateczna liczba badanych uczniów; liczba uczniów wiejskich będzie znacznie niższa, więc analizy będą wymagały skomplikowanego ważenia obserwacji, dodatkowo obniżając precyzję wyników.

■ W gminie rozesłano po szkołach ankietę do uczniów i rodziców. Błąd: na ankietę będą odpowiadać tylko niektórzy uczniowie i rodzice, o podobnych cechach; liczba ankiet z innych grup będzie bardzo zaniżona, a wnioski z badania mogą być mało reprezentatywne.

■ Przebadano kilkadziesiąt tysięcy dzieci w przedszkolach, które zgłosiły się do udziału w badaniu. Błąd: do badania zgłoszą się przedszkola i dzieci, które różnią się zapewne od przedszkoli i dzieci, które do badania się nie zgłosiły; wyniki będą mało reprezentatywne dla populacji przedszkolaków w kraju; znacznie bardziej wiarygodne wyniki uzyskano by przy dziesięciokrotnie mniejszej próbie, lecz dobranej losowo.

■ Przeprowadzono wywiady z uczniami i nauczycielami kilkunastu szkół prywatnych oraz publicznych celem porównania organizacji, metod i efektywności w tych placówkach. W badaniu wzięły udział szkoły z jednego miasta, które wyraziły zgodę. Błąd: takie badanie nie jest reprezentatywne zarówno ze względu na nielosowy dobór szkół, a przez to i uczniów, jak i ze względu na liczną próbę.

2.3. Badania reprezentatywne a określanie związków przyczynowo-skutkowych

Metody badań reprezentatywnych mają gwarantować, że uzyskane dane będą odzwierciedlały rozkład opinii lub cech w populacji. Nie gwarantują one jednak w żaden sposób, że dzięki zebranym danym będziemy w stanie ocenić efektywność narzędzi polityki edukacyjnej lub nowych metod nauczania. Jedy-

ne, co najczęściej możemy określić na podstawie danych z badań reprezentatywnych, to powiązania między interesującymi nas zmiennymi.

Możemy zwiększyć prawdopodobieństwo wniosku, że obserwowane powiązania wynikają z celowego działania, na przykład wprowadzenia reformy nauczania, stosując wyrafinowane narzędzia analizy i bogate zbiory danych, dzięki którym możemy kontrolować cechy uczniów, rodziców, nauczycieli i szkół. Wciąż jednak będą to tylko obserwowane powiązania i zawsze pozostanie wątpliwość, czy ich przyczyną są interesujące nas działania, czy też wynikają ze współwystępowania pewnych cech uczniów (szkół) i obserwowanych efektów.

Kwestie te najlepiej zrozumieć na przykładzie. Zacznijmy od klasycznego problemu oceny wpływu wielkości klasy na osiągnięcia uczniów, który doczekał się na świecie olbrzymiej literatury badawczej. Chcemy dowiedzieć się, czy przyrostowi wiedzy uczniów sprzyjają mniej liczne klasy, czy też może odwrotnie, klasy większe. Popularne intuicje skłaniają większość ludzi do twierdzenia, że klasy mniejsze są oczywiście lepsze, bowiem nauczyciel może w nich poświęcić więcej czasu pojedynczemu uczniowi, uczniowie nie są anonimowi i łatwiej jest zorganizować pracę w klasie.

Jest jednak sporo prac przekonujących, że większe klasy mogą być korzystne dla uczniów. Niektórzy psychologowie przekonują, że postępy uczniów w dużym stopniu zależą od tego, jakich mają kolegów w klasie, a przede wszystkim od tego, czy znajdą wśród rówieśników osoby o podobnym poziomie zdolności i wiadomości. W przypadku gdy uczeń posiada w klasie podobnie zdolną koleżankę albo kolegę, jest pewne prawdopodobieństwo, że przyrost jego wiedzy będzie większy, niż gdyby w klasie byli tylko uczniowie od niego znacznie mniej lub znacznie bardziej zdolni. Oczywiście w klasie większej łatwiej jest znaleźć taką osobę. Stąd większe klasy mogą przynosić uczniom wymierne korzyści.

Z drugiej strony istnieją zapewne negatywne efekty większych klas związane z trudnościami w utrzymaniu porządku czy też ograniczoną możliwością pracy nauczyciela z pojedynczym uczniem etc. W istocie, efekty te działają w przeciwnych kierunkach i mogą się albo znosić, albo też któryś z nich przeważy, powodując, że obserwowany wpływ wielkości klasy będzie dodatni lub ujemny[17].

Jak tę kwestię rozstrzygnąć? Oczywiście prowadząc odpowiednie badania empiryczne. Temat wpływu wielkości klasy jest tak popularny, bowiem

[17] W ciekawej pracy holenderskich badaczy wykorzystano metodę zmiennych instrumentalnych, aby rozdzielić te efekty, potwierdzając, że na wpływ wielkości klasy składają się dwa efekty: pozytywny związany z większym prawdopobieństwem znalezienia koleżanki/kolegi o podobnym IQ, a także negatywny związany z trudniejszymi warunkami nauczania w większych klasach (Dobbelsteen S., Levin, J., Oosterbeek H. (2002), *The Causal Effect of Class Size on Scholastic Achievement*: *Distinguishing the Pure Class Size Effect from the Effect of Changes in Class Composition*, „Oxford Bulletin of Economics and Statistics", vol. 64(1), s. 17-38).

łatwo o dane pozwalające na jego badanie. Mając wyniki egzaminów, można od razu policzyć wielkość klasy (sumując liczbę uczniów zdających egzamin w klasie) i rozpocząć badanie. Można więc stwierdzić, że każde badanie reprezentatywne mierzące poziom wiedzy uczniów i gromadzące dane o wielkości klasy pozwala przeprowadzić odpowiednie testy. Czy jednak każde badanie może dostarczyć nam odpowiedzi na pytanie, czy wielkość klasy jest powiązana z wynikami uczniów? Doświadczenie kilkuset badań empirycznych dotyczących tej kwestii pokazuje, że nie, a wręcz określenie prawdziwej zależności przyczynowo-skutkowej między liczbą uczniów a ich osiągnięciami jest bardzo trudne.

Problemy z określeniem wpływu wielkości klasy na wyniki uczniów można pokazać na prostych przykładach. Choć z założenia są to sytuacje przesadzone, to jednak ukazują problemy, które występują w większości podobnych badań. Przykładowo, przypuśćmy, że chcemy porównać wpływ wielkości klasy na podstawie wyników w dwóch szkołach. Jedna to szkoła w niewielkiej miejscowości, druga to szkoła na dynamicznie rozwijającym się wielkomiejskim osiedlu. W pierwszej z nich klasy będą zazwyczaj mniejsze, z prostego powodu małej liczby uczniów w roczniku. W drugiej klasy będą zazwyczaj duże z powodu rosnącej liczby uczniów spowodowanej coraz większą liczbą rodzin osiedlających się w okolicy. W większości wypadków szkoły te będą się też różniły wynikami uczniów.

Uczniowie w mniejszych miejscowościach uzyskują zazwyczaj niższe wyniki niż uczniowie w dużych miastach, co potwierdzają zarówno wyniki egzaminów zewnętrznych, jak i na przykład badanie PISA w Polsce. Jest to powodowane nie tyle tym, że szkoły i nauczyciele na wsi prezentują przeciętnie niższą jakość nauczania, co niestety wiele osób w naszym kraju przyjmuje za pewnik, ale że uczniowie na wsi pochodzą częściej z rodzin rolniczych lub robotniczych, o niższym kapitale edukacyjnym. Co z tego wynika dla naszej próby oceny wpływu wielkości klasy na wyniki uczniów? Porównując klasy mniejsze ze szkoły wiejskiej z większymi klasami ze szkoły miejskiej, stwierdzimy zapewne pozytywną zależność. Jednak nie ma ona nic wspólnego z prawdziwym wpływem wielkości klasy, a jedynie odzwierciedla silne powiązanie wielkości klasy i wyników z innymi czynnikami, takimi jak pochodzenie społeczne uczniów.

Najprostsze metody statystyczne w rodzaju testu na różnice średnich wyników między szkołami z różnymi wielkościami klas czy testu na współczynnik korelacji między wielkością klasy a wynikami, dadzą zazwyczaj mylny obraz prawdziwej zależności, ze względów wymienionych powyżej. W badaniach edukacyjnych często jednak korzysta się z metod biorących pod uwagę wpływ innych obserwowalnych zmiennych. Na przykład, w powyżej opisanym przypadku możemy wziąć pod uwagę pochodzenie społeczne uczniów, choćby korzystając z popularnej metody regresji liniowej. Metoda ta szacuje zależność między wiel-

kością klasy a wynikami uczniów, biorąc pod uwagę różnicę w pochodzeniu społecznym uczniów. Inaczej mówiąc, metodą tą możemy ocenić, jaka jest ta relacja, zakładając, że wszyscy uczniowie mają podobne pochodzenie (oczywiście pod warunkiem, że nasz model statystyczny jest poprawnie zdefiniowany, co nie zawsze ma miejsce).

Podobne metody mogą być z powodzeniem stosowane, jeśli różnice między osobami i grupami, które mają wpływ na interesujące nas zmienne, są obserwowalne. Obserwowalność w praktyce oznacza, że możemy te cechy zmierzyć i wykorzystać w naszym modelu statystycznym, kontrolując ich wpływ na wyniki oraz zmienną, której wpływ oceniamy (w naszym wypadku wielkość klasy). Niestety w edukacji, ale też i ogólniej w naukach społecznych, to cechy nieobserwowalne mają często zasadniczy wpływ na wyniki. Są to najczęściej cechy, których po prostu nie możemy zmierzyć, sprawia to bardzo duże trudności, lub też których po prostu nie posiadamy w bazie danych do naszego badania i rzadko kiedy istnieje możliwość zebrania takich informacji.

Podstawową zmienną, którą chcielibyśmy „mieć" w każdym badaniu, jest wrodzony poziom zdolności czy też potencjał edukacyjny ucznia. Po pierwsze, bardzo trudno tego typu pojęcia zdefiniować. Po drugie, każdy pomiar tego rodzaju charakterystyki jest zazwyczaj specyficzny dla danej definicji i niekoniecznie mierzy potencjał, który ma znaczenie w naszej sytuacji. Po trzecie, taki pomiar jest bardzo kosztowny, więc rzadko kiedy zbiory danych zawierają tego typu dane. Po czwarte, w wielu przypadkach dokonanie pomiaru takiej charakterystyki jest niemożliwe, bowiem badania mają charakter jednorazowy, a pomiar potencjału musiałby być dokonany kilka lat wcześniej (najlepiej wielokrotnie).

W wypadku badań wykorzystujących wyniki egzaminów zewnętrznych rzadko kiedy będziemy posiadali tego typu dane. Często wystarczyć musi nam jedynie końcowy wynik. Wyjątkiem jest sytuacja, gdy wyniki egzaminów łączymy z własnymi testami zdolności uczniów lub też wtedy, gdy jeden wynik egzaminu jesteśmy w stanie połączyć z innym wynikiem. W tej drugiej sytuacji, o ile egzaminy mierzą podobne umiejętności i wiadomości, możemy zakładać, że kontrolujemy poziom umiejętności i wiadomości ucznia na wcześniejszym etapie, oceniając wpływ interesujących nas zmiennych na poziom umiejętności i wiadomości ucznia na etapie późniejszym.

Jeśli jesteśmy to w stanie zrobić i poprawnie zdefiniujemy nasz model statystyczny, to wariancja w wynikach, którą objaśniamy, zależy głównie od czynników, które uległy zmianie między pomiarami, a więc między pierwszym a kolejnym egzaminem. Tego rodzaju wariancja jest w znacznie mniejszym stopniu uwarunkowana czynnikami, które mogą zaburzać ocenę interesujących nas zmiennych. Taki dokładnie zamysł leży u podstaw metody edukacyjnej wartości dodanej, którą opisujemy w kolejnym rozdziale. O ile metoda ta w standardowej wersji nie jest w stanie określić zależności przyczynowo-skutkowych, to

jednak daje znacznie lepsze ich przybliżenie, niż metody niebiorące pod uwagę poziomu wiadomości i umiejętności uczniów na wcześniejszych etapach nauczania.

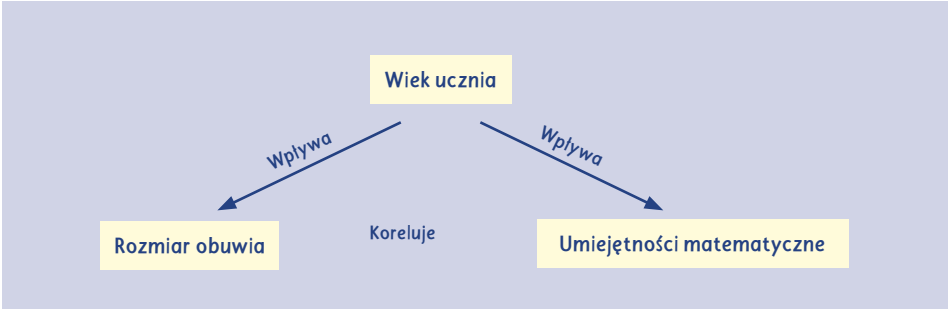
Największym wyzwaniem dla badań edukacyjnych jest jednak często zależność interesujących nas zmiennych oraz wyników uczniów od zmiennych, których w większości przypadków nie można zmierzyć. Zdolności ucznia lub na przykład tzw. „klimat szkolny" są zmiennymi trudnymi do zdefiniowania i pomiaru. Badacze jednak opracowali podejścia, które choć zawsze będą poddawane krytyce i dyskusji, to jednak prowadzą do kwantyfikacji tych pojęć.

Pewnych zachowań nie da się jednak w łatwy sposób przenieść na skalę liczbową. Przykładowo, jak zmierzyć popularność placówki szkolnej? Żeby to zrobić, musielibyśmy przeprowadzić sondaż reprezentatywny dla wszystkich okolicznych rodzin, które potencjalnie mogą posłać dzieci do tej placówki. Jest to nie tylko trudne ze względu na koszty, ale i precyzję takiego pomiaru preferencji rodzin. Sondaże przedwyborcze są dobrym przykładem, jak trudno jest zmierzyć prawdziwe preferencje przyszłych głosujących. Można się spodziewać, że w wypadku szkół byłoby to jeszcze trudniejsze.

O ile trudno jest wskazać zmienne, które uznać można za w pełni niemierzalne, to jednak w wielu wypadkach problemy praktyczne (koszty, trudność dotarcia do respondentów, niemożność cofnięcia się w czasie), jak i teoretyczne (problemy z definicją i jej operacjonalizacją), powodują, że w wielu badaniach edukacyjnych nie posiadamy pomiaru najważniejszych zmiennych. Jak wpływać to może na wyniki?

Wróćmy do przykładu badania wpływu wielkości klasy. Załóżmy, że mamy do czynienia z dwoma miejskimi szkołami. Jedna z nich jest postrzegana jako szkoła bardzo dobra i w istocie taką jest. Druga wręcz przeciwnie, jest szkołą słabą i taką ma opinię. W pierwszej z nich będziemy mieli znacznie większe klasy, ze względu na dużą liczbę chętnych do korzystania z placówki o wysokiej jakości nauczania. W drugiej odwrotnie. Korelacja między wynikami uczniów a wielkością klas oszacowana dla tych dwóch szkół będzie silnie dodatnia. Czy znaczy to, że większe klasy podnoszą wyniki uczniów? Nie. Znaczy to jedynie tyle, że dobre placówki mają większe klasy, ponieważ mają dużo chętnych.

Czy możemy w łatwy sposób kontrolować zależność między popularnością placówki a wynikami uczniów? Nie. Liczba chętnych przypadająca na jedno miejsce mogłaby posłużyć jako miara popularności, ale co, jeśli rodzice i uczniowie mylnie oceniają jakość nauczania? W ten sposób szkoła z większymi klasami, popularna, miałaby też wyższe wyniki, ale nie ze względu na jakość nauczania, tylko ze względu na to, że najlepsi uczniowie chcą do niej uczęszczać, wierząc, że jest to szkoła dobra. Tego rodzaju „samospełniające się przepowiednie" powodują, że badania edukacyjne opierające się na miarach korelacyjnych czy też zwy-



Rysunek 2.6. Przykład korelacji pozornej (za Babbie, 2005)

kłej regresji są często obarczone bardzo dużym błędem, którego ani rozmiaru, ani kierunku często nie jesteśmy w stanie ocenić.

Babbie (2005) przedstawia dowcipny przykład zależności pozornej. Jeżeli patrzymy na populację uczniów, istnieje silny pozytywny związek między umiejętnościami matematycznymi a rozmiarem obuwia. Czy to znaczy, iż wielkość stopy wpływa na umiejętności uczniów? Oczywiście, że nie. Jest to klasyczny problem korelacji pozornej. Istnieje bowiem trzecia zmienna, która wyjaśnia związek wielkości stopy z umiejętnościami matematycznymi – jest nią wiek.

Im więcej uczeń ma lat, tym większy rozmiar buta. Jednocześnie im dziecko starsze, tym połączenia neuronowe w jego mózgu gęstsze i bardziej sprawne, zatem umiejętności matematyczne wyższe.

Najpewniejszym sposobem uzyskania oszacowań interesujących nas zależności bliskich rzeczywistym związkom przyczynowo-skutkowym jest przeprowadzenie eksperymentu lub wykorzystanie metod, które na podstawie istniejących danych „odtwarzają" sytuację eksperymentalną. O metodach tych piszemy w dwóch kolejnych punktach poniżej. W tej chwili spójrzmy na oceny zależności między wielkością klasy a wynikami uczniów, uzyskane dla danych z Polski różnymi metodami, dla 3 różnych zbiorów danych z innym pomiarem umiejętności (sprawdzian szóstoklasistów, egzamin gimnazjalny po uwzględnieniu wyników sprawdzianu, czyli metoda EWD, oraz badanie PISA) – tabela 2.2. Przykład badania tego rodzaju przedstawiamy dokładniej w rozdziale 4.

Widać wyraźnie, jak różne są to oceny. Naiwne metody szacujące związek między wielkością klasy a wynikami wskazują na pozytywny efekt, sprzeczny z intuicją. Jednak zastosowanie odpowiednich metod zbliżających nas do oceny „prawdziwej" zależności między wielkością klasy a wynikami uczniów daje oceny bliskie zera lub nawet negatywne, zgodne z popularną intuicją. Czy wielkość klasy ma rzeczywisty wpływ na wyniki uczniów, jest wciąż kwestią otwartą. Wyniki różnią się też zapewne w zależności od poziomu nauczania, średniej wielkości klas, a także przedmiotu nauczania i kraju.

W badaniach edukacyjnych poprawny pomiar, ale i poprawne zaplanowanie tzw. strategii identyfikacyjnej, która pozwoli ocenić prawdziwe związki przyczynowo-skutkowe między zmiennymi, są najważniejsze dla wiarygodności wyników. Niestety, bardzo często proste zależności stwierdzane standardowymi metodami mają charakter pozorny. Inaczej mówiąc, są to jedynie zależności, które obserwujemy, a które niewiele nam mogą powiedzieć o tym, czy na przykład zastosowanie podobnych rozwiązań przyniesie spodziewane korzyści w naszym przypadku.

Tabela 2.2. Przykładowe oceny wpływu wielkości klasy na wyniki uczniów w oparciu o różne zbiory danych i metody badawcze

Badanie wpływu wielkości klasy na wyniki sprawdzianu szóstoklasistów w oparciu o metody quasi-eksperymentalne Źródło: Jakubowski M., Sakowski P. (2006), *Quasi-experimental estimates of class size effect in primary schools in Poland*, „International Journal of Educational Research", vol. 45(3)			
		W oparciu o zwykłą regresję liniową	W oparciu o naturalny eksperyment
Zmiana w wyniku sprawdzianu szóstoklasistów względem procentowej zmiany w wielkości klasy		1,12	-0,75
Badanie wpływu wielkości klasy w oparciu o modele EWD Źródło: Jakubowski M. (2008), *Zastosowanie modeli EWD do ewaluacji polityki edukacyjnej*			
	Regresja liniowa bez uwzględnienia wyników sprawdzianu	Regresja EWD z uwzględnieniem wyników sprawdzianu	Regresja EWD z efektami stałymi (efekty zmian wielkości klas w tej samej szkole)
Zmiana wyniku egzaminu gimnazjalnego przy zmianie wielkości klasy o 1 ucznia	0,476	0,106	-0,015 (nieistotne statystycznie)
Badanie wpływu wielkości klasy w oparciu o wyniki uczniów w naukach przyrodniczych w PISA 2006 Źródło: obliczenia własne			
	Regresja wielkości klasy na wynik ucznia	Regresja wielkości klasy na wynik ucznia przy kontroli zasobów społeczno-ekonomicznych rodziny ucznia	
Zmiana wyniku ucznia przy zmianie wielkości klasy o 1 ucznia	2,08	0,94 (nieistotne statystycznie)	

Uwagi: w tabeli podano jedynie wybrane wyniki z cytowanych prac; wszystkie efekty są statystycznie istotne na poziomie 0,05, jeśli nie podano inaczej.

2.4. Eksperymenty

Według wielu badaczy najlepszym sposobem badania zależności przyczynowo-skutkowych jest przeprowadzenie eksperymentu. Pod pojęciem badania eksperymentalnego kryje się wiele podejść, które jednak opierają się na tej samej koncepcji: porównania wartości zmiennej wynikowej w grupie poddanej „bodźcowi" i w grupie kontrolnej, przy czym obie grupy dobierane są losowo.

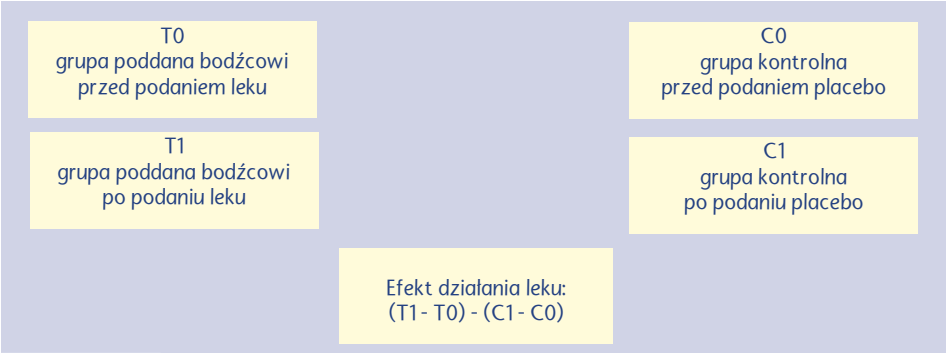
Wartość tej podstawowej zasady łatwo zrozumieć, posługując się prostym przykładem. Załóżmy, że chcemy ocenić, czy dana substancja medyczna skutecznie leczy określoną chorobę. Możemy podać lekarstwo grupie osób, która sama zgłosi chęć korzystania z niego, a następnie porównać ich stan zdrowia przed kuracją i po. Okazuje się, że stan zdrowia tych osób polepszył się. Czy oznacza to, że lekarstwo poskutkowało? Proste intuicje sugerują, że niekoniecznie. Być może stan zdrowia tych osób polepszył się samoistnie? Być może samo spotkanie z lekarzem i świadomość poddania leczeniu zapewniły poprawę zdrowia?

Jak możemy więc określić, czy na zanik choroby miało jednak wpływ podanie lekarstwa, czy też inne procesy? Wystarczy, że dobierzemy dwie grupy chorych, jednej podamy lekarstwo, a drugiej podamy tzw. placebo, czyli coś udającego lekarstwo, lecz niezawierającego interesującej nas substancji. W ten sposób, niejako oszukując pacjentów, możemy ocenić czysty wpływ podania leku na chorobę.

Czy dwie grupy mogą dobrać się samodzielnie? Nie, muszą one zostać dobrane losowo. Po pierwsze, dzięki losowaniu cechy pacjentów w obu grupach będą także rozłożone losowo. W ten sposób wykluczymy możliwość, że w jednej z grup znajdą się pacjenci starsi, określonej płci, o większej podatności na działanie leku itp. A więc wykluczymy potencjalne czynniki powiązane z działaniem leku, które mogłyby zaburzyć obraz efektywności jego działania przy porównaniu grup.

Jeśli dwie grupy zostały dobrane w pełni losowo, a przez to posiadają identyczne cechy obserwowalne i nieobserwowalne, i jeśli pacjenci nie są świadomi tego, czy podano im prawdziwy lek, czy tylko placebo, to zwyczajne porównanie stanu zdrowia w obu grupach przed podaniem i po podaniu leku (lub placebo) pozwoli określić zależność przyczynowo-skutkową między podaniem leku a stanem zdrowia pacjentów. To stanowi o przewadze eksperymentu nad innymi schematami badań. Jeśli uda się spełnić podstawowe warunki, to wystarczy proste porównanie zmiennej wynikowej między grupami oraz przed i po, żeby ocenić zależności przyczynowo-skutkowe. Schemat porównania obrazuje rysunek 2.7.

Efekt działania leku można obliczyć bardzo prosto, jako różnicę w stanie zdrowia w grupie poddanej bodźcowi przed podaniem (T0) i po podaniu leku (T1) minus podobna różnica w grupie kontrolnej przed podaniem (C0) i po



Rysunek 2.7. Podstawowy schemat eksperymentalny

podaniu placebo (C1), a więc jako (T1 - T0) - (C1 - C0). Taki schemat zapewnia, że jeśli dla obu grup pacjentów możliwe jest na przykład samowyleczenie, to ten efekt zostanie wzięty pod uwagę (jeśli w obu grupach stan zdrowia poprawi się o 5 punktów, to różnica ta zniknie po odjęciu zmian w grupach od siebie). Bierzemy pod uwagę jedynie różnicę w zmianach stanu zdrowia między grupami. Z tego względu kluczowe jest, aby obie grupy miały identyczne cechy, co zapewnia dobór losowy.

Taki schemat stoi u podstawy współczesnej medycyny, czy też ogólniej nauk eksperymentalnych, gdzie stanowi podstawową metodę zdobywania naukowej wiedzy o świecie. Czy tego rodzaju schemat można zastosować do badań edukacyjnych? Czy można go łatwo przełożyć na ocenę polityki edukacyjnej? Czy możliwe jest spełnienie podstawowych założeń? Czy wnioski z eksperymentu mogą posłużyć reformowaniu całego systemu edukacji? Niestety, odpowiedź na powyższe pytania brzmi w większości przypadków „raczej nie". Dlaczego? Z jakiego względu może być on stosowany z powodzeniem w medycynie lub biologii, ale nie w edukacji? Czy istnieją sytuacje, kiedy możemy się nim posłużyć?

Przede wszystkim, rzadko kiedy w edukacji możemy przeprowadzić eksperyment. Z bardzo prostej przyczyny: uczniowie to nie pacjenci zgłaszający się po terapię lub też opłacani za udział w eksperymencie. Pomijając studentów uczelni wyższych, którzy są głównym „materiałem" eksperymentalnych badań w psychologii czy ekonomii, to uczniowie szkół powszechnych nie mogą być poddawani eksperymentom bez zgody rodziców, a i nawet wtedy byłby to w wielu przypadkach zabieg dość kontrowersyjny i trudny do pełnej realizacji. Większość uczniów w krajach rozwiniętych, w tym w Polsce, to uczniowie szkół publicznych. Trudno więc dopuścić możliwość, że mogliby uczestniczyć w eksperymencie bez ich woli, trudno też wyobrazić sobie, że mogliby zostać za udział w nim opłacani. Co więcej, uczniowie nie decydują w pełni o swym losie i o zgodę trzeba pytać rodziców.

spis treści

Trudno sobie wyobrazić sytuację, gdy testując nowe rozwiązanie edukacyjne, o którym jesteśmy przekonani, że ma pozytywny wpływ na rozwój uczniów, świadomie dobieramy grupę poddaną „bodźcowi", czyli korzystającą z nowego rozwiązania, oraz grupę kontrolną, spodziewając się, że ta druga na tym straci. Trudno wyobrazić sobie troskliwych rodziców, którzy by się świadomie zgodzili na to, że ich dziecko nie skorzysta z nowego rozwiązania. Z pewnością jedni rodzice zgodziliby się na poddanie eksperymentowi (zastosowanie wobec ich dzieci nowego sposobu nauczania), a inni woleliby pozostać przy starych metodach. Jednak dając taki wybór, dobieramy grupy eksperymentalne nielosowo, tracąc główną zaletę tego schematu badania. W takim wypadku, a jest to bardzo częsty przypadek w badaniach edukacyjnych, nie można już mówić o pełnej porównywalności dwóch grup poddanych badaniu. Różnica w rozwoju uczniów między tymi grupami może być interpretowana jako efekt nowych metod, ale też jako efekt innego rozwoju grup uczniów o odmiennych cechach.

Inne kwestie związane są z problemami natury praktycznej, wynikającymi ze sposobów organizacji pracy szkół. Uczniowie najczęściej odbywają zajęcia w ustalonych grupach i trudno byłoby dokonać losowania i ponownego podziału na grupy. Co więcej, w systemie szkolnictwa publicznego trudno odmówić uczniowi zmiany grupy, szczególnie jeśli rodzic wyrazi podejrzenie, że jego dziecko zostało celowo umieszczone w grupie bez dostępu do nowych metod nauczania (czyli w naszej grupie kontrolnej). To powoduje, że w praktyce stworzenie i zachowanie losowego doboru grup eksperymentalnych jest w edukacji niezwykle trudne, a wręcz niewykonalne.

Bardzo często w badaniach edukacyjnych teoretycznie prawidłowy schemat eksperymentalny aplikowany jest do grup, które dobierane są w sposób nielosowy. Często są to zaprzyjaźnione (lub prowadzone osobiście przez badacza) szkoły lub klasy. Często brakuje grupy kontrolnej (wtedy właściwie nie mamy już nawet do czynienia ze schematem eksperymentalnym) lub jest ona dobierana *ad hoc*, po prostu jako reszta uczniów, dla których dostępny jest pomiar. Następnie autorzy badania interpretują wyniki jako zależności przyczynowo-skutkowe, dodatkowo uogólniając je na całą populację uczniów. Takie badania są niestety liczne, ale wnioski z nich płynące są najczęściej zupełnie niewiarygodne, wbrew przekonaniu autorów.

Inne problemy znane są psychologii, gdzie schemat eksperymentalny stanowi dla wielu także główną metodę badania empirycznego. Znane są efekty związane z tym, że uczestnicy mają świadomość uczestniczenia w eksperymencie. Jest to prawdopodobne także w edukacji, gdzie uczniowie, mając świadomość, że stosowane są nowe rozwiązania, których oceną zainteresowani są nauczyciele, mogą na nie reagować pozytywnie lub negatywnie, inaczej niż zareagowaliby, nie mając świadomości uczestniczenia w eksperymencie.

Być może jeszcze większe znaczenie ma świadomość nauczycieli. Mogą się oni bardziej lub mniej starać zależnie od tego, czy chcieliby, aby nowe rozwią-

zania zostały ocenione pozytywnie, czy też negatywnie. Ma to znaczenie szczególnie, gdy wpływa to na ich warunki pracy (większy lub mniejszy wysiłek, stres itp.) lub gdy z eksperymentatorami są związani emocjonalnie, co ma często znaczenie w badaniach na małą skalę. Samo uczestnictwo w eksperymencie może powodować, że uczestniczący chcą uzyskać wyniki potwierdzające badaną hipotezę, o ile domyślają się, jaki efekt jest spodziewany. Podobnego rodzaju efekty są szeroko opisane w literaturze psychologicznej (na przykład efekty Hawthorne'a czy Pigmaliona).

Ogólnie mówiąc, w naukach społecznych, gdzie materiałem doświadczalnym są ludzie, trudno oczekiwać podobnej jakości wyników, jak w naukach eksperymentalnych typu biologia, gdzie testom poddawane są myszy. Nawet myszy mogą wpływać na wyniki eksperymentów, a ludzie mają w tym względzie o wiele większe możliwości.

Czy więc schemat eksperymentalny nie ma żadnego znaczenia w edukacji? Niekoniecznie. Po pierwsze, przeprowadzanie eksperymentów na małą skalę jest możliwe i ma sens, o ile nie zmienia to zasadniczo zachowania uczniów, rodziców i nauczycieli. Nauczyciele mogą eksperymentować z technikami lub programem nauczania w sposób niezauważalny dla uczniów. Jeśli sami nie są pewni, czy dane metody są korzystne lub niekorzystne, to także nie stoi za tym problem moralny nieinformowania o badaniu rodziców czy uczniów.

W gruncie rzeczy nasza praca to zawsze pewne działania na pograniczu eksperymentu. Prawdziwe badanie zaczyna się jednak wtedy, gdy świadomie zastosujemy z pełnym rygorem schemat eksperymentalny. Losowo wyznaczymy grupy uczniów poddane i niepoddane nowemu rozwiązaniu, następnie dokonamy pomiaru ich wiedzy i umiejętności co najmniej przed eksperymentem i po jego zakończeniu, choć warto też rozważyć schematy dokładające pomiar w trakcie. Tego rodzaju badania są możliwe na małą skalę i dla grup uczniów poddanych eksperymentowi mogą stanowić dobrą wskazówkę, czy dane rozwiązanie przynosi korzyści.

Problem stanowi tu uogólnienie wyników na większą populację. Po pierwsze, jeśli były to niewielkie grupy, w dodatku grupy podzielone losowo na poddaną „bodźcowi" oraz eksperymentalną, ale już nie wybrane losowo z szerszej populacji, to nie jest prawomocne uogólnienie wyników na całą populację uczniów. Pojęcie reprezentatywności badania ma takie samo zastosowanie do badań eksperymentalnych i podobne konsekwencje niezachowania losowego doboru próby, co omówiliśmy już powyżej.

Po drugie, istnieją badania eksperymentalne na dużą skalę. Najbardziej znanym tego rodzaju eksperymentem jest projekt STAR przeprowadzony w latach dziewięćdziesiątych w stanie Tennessee w USA[18]. W projekcie tym poddano eksperymentowi 7000 uczniów z 79 szkół, w których przypisano ich w pełni losowo

[18] Patrz http://www.heros-inc.org/star.htm

do trzech typów klas: małych (między 13 a 17 uczniami), przeciętnych (między 22 a 25 uczniami) oraz przeciętnych z dodatkową pomocą nauczyciela. Co istotne, nie tylko uczniowie, ale i nauczyciele zostali przypisani do klas w sposób losowy. Projekt rozpoczął się w przedszkolach i zakończył pierwszą fazę po 4 latach, w trzeciej klasie szkoły podstawowej. Jest jednak wciąż kontynuowany i ostatnie badania dotyczyły już edukacji uczniów objętych tym projektem na poziomie wyższym.

Projekt ten stał się tematem setek prac badawczych i szerokiej dyskusji na temat zasadności badań eksperymentalnych na dużą skalę. Oficjalne wyniki sugerowały, że mniejsze klasy, a dokładniej mniejsza liczba uczniów na nauczyciela, znacząco poprawiają wyniki uczniów. Co więcej, wskazywano też, że efekty te widoczne są po wielu latach, nawet w procencie uczniów kończących *college*. Krytycy podkreślali, że nauczyciele uczący mniejsze klasy mogli dodatkowo starać się o podniesienie wyników uczniów, mając na uwadze, że przyszłe zmniejszenie klas poprawi zatrudnienie nauczycieli i prawdopodobnie wpłynie też na podniesienie ich pensji, co było nagłaśnianie przez związki zawodowe. Z tego samego względu nauczyciele uczący klasy „przeciętne" nie mieli zachęty do prowadzenia zajęć na wysokim poziomie.

Znane są także problemy ze zmianą grup przez uczniów, których rodzice protestowali przeciwko przypisaniu do klas większych. Władzom szkolnym nie udało się utrzymać składu wielu klas i uczniowie zmieniali przypisane grupy, zakłócając losowość ich doboru. Krótko mówiąc, okazało się, że świadomość uczestniczenia w eksperymencie stanowi główne wyzwanie dla jego powodzenia. W praktyce zachowanie losowego składu grup eksperymentalnych i ograniczenie celowych zachowań osób w nim uczestniczących nie jest możliwe. Co prawda, powstały analizy stosujące wyrafinowane techniki statystyczne korygujące zachwiania w losowym składzie grup eksperymentalnych, to jednak niemożliwe jest skorygowanie wyników na przykład o celowe zachowania nauczycieli. Powoduje to, że wyniki projektu STAR są szeroko podważane, gdyż tak kosztowne i trudne przedsięwzięcie nie przyniosło przekonującej odpowiedzi na jedno z zasadniczych pytań w systemie edukacji: jaka wielkość klasy sprzyja najbardziej rozwojowi ucznia?

Powyższa dyskusja i przykłady sugerują, że być może największą korzyść schemat eksperymentalny niesie dla badań edukacyjnych jako rodzaj wzorca metody umożliwiającej wyznaczenie zależności przyczynowo-skutkowych, których określenie ma na celu większość badań edukacyjnych. Eksperyment to punkt odniesienia dla każdego przemyślanego badania empirycznego. Zawsze powinniśmy postawić sobie pytanie o to, czy nasze dane i metoda badania pozwalają odtworzyć sytuację bliską eksperymentalnej. Jakiego rodzaju zniekształcenia obrazu relacji przyczynowo-skutkowej może powodować nielosowy dobór grup, które porównujemy? Jakie reakcje osób poddanych rozwiązaniom, które nas interesują, mogą wpływać na wyniki. Czy analizowane grupy są szczególne pod

określonymi względami i jak to wpływa na możliwość uogólnienia wyników? Takie i podobne pytania zawsze powinny towarzyszyć badaniom empirycznym, szczególnie jeśli ich ambicją jest doradzenie rozwiązań mających przynosić korzyści uczniom. Popularne i niezwykle wpływowe jest też w ostatnich latach podejście, które szuka sytuacji eksperymentalnych w rzeczywistości. Są to tzw. eksperymenty naturalne lub quasi-eksperymenty, o których piszemy w kolejnym punkcie.

2.5. Naturalne eksperymenty

Zaletą eksperymentu jest to, że posiadamy grupy osób poddane i niepoddane działaniom interesującego nas czynnika, które przy okazji dobrane są w sposób w pełni losowy, a przez to można oczekiwać, że mają identyczne cechy ukryte i obserwowalne. Jest to idealna sytuacja do oceny efektów działania programu edukacyjnego lub też nowego sposobu organizowania nauczania, jednak, jak stwierdziliśmy powyżej, jest to też sytuacja w edukacji niemal niemożliwa do osiągnięcia w praktyce w sposób zaplanowany. Co jednak, gdy samo życie losowo poddaje grupy osób działaniom interesującego nas czynnika? Czy możemy taką sytuację wykorzystać do oceny interesujących nas zagadnień?

Posłużmy się przykładami, aby lepiej ukazać sytuacje, jakie nazywamy naturalnymi eksperymentami. Załóżmy, że jesteśmy w stanie śledzić losy wszystkich uczniów urodzonych w danej miejscowości. Liczba urodzeń w każdym roku jest nieco różna. Wyłączając trendy demograficzne, pozostają losowe wahania liczby urodzeń, które przekładają się następnie na wielkość roczników szkolnych, a te z kolei wpływają na wielkość klas. Porównując zmiany w liczbie urodzeń między latami, można określić, jaki wpływ miały one na zmiany w wielkości klas.

Ta część wariancji wielkości klas, która powiązana jest z liczebnością rocznika, ma więc charakter losowy, niezależny z pewnością od władz oświatowych. Można więc wykorzystać to źródło wariancji do oceny, jak wielkość klasy wpływa na wyniki uczniów. Trzeba jednak wyłączyć pozostałą wariancję, która może być powodowana świadomymi wyborami nauczycieli, rodziców i uczniów. Powstały metody, które w dość prosty sposób to umożliwiają. Są to metody zmiennych instrumentalnych, których przykład wykorzystania do oceny wpływu wielkości klasy na wyniki uczniów przedstawiamy w rozdziale 4.

Inny przykład, ze słynnego badania amerykańskiego. Załóżmy, że przymus uczęszczania do szkoły definiowany jest wedle roku urodzenia, a nie wieku, jaki uczeń ma w momencie rozpoczęcia roku szkolnego. Natomiast czas obowiązkowej nauki w szkole definiowany jest według wieku ucznia. W ten sposób uczniowie urodzeni w ostatniej ćwiartce roku kalendarzowego zmuszeni są do uczęszczania do szkoły o 3 miesiące dłużej niż pozostali (zakładając, że szkoła zaczyna

się w październiku). Powoduje to losową wariancję w długości kształcenia (bowiem trudno uznać datę urodzenia za czynnik związany jakkolwiek z polityką edukacyjną), którą słynni amerykańscy badacze Joshua Angrist i Alan Krueger wykorzystali do oceny korzyści finansowych z wydłużonej edukacji[19]. Porównali oni dochody osób urodzonych w ostatniej ćwiartce roku kalendarzowego z pozostałymi na przestrzeni kilkudziesięciu lat, znajdując silne potwierdzenie tezy o korzystnym wpływie edukacji na dochody, a nawet przeliczając dodatkowe miesiące nauki na konkretne kwoty.

Mówimy tu o tzw. naturalnych eksperymentach, zwanych też czasem quasi-eksperymentami, gdzie z przyczyn losowych, niezależnych od uczestników zdarzenia, zachodzą zmiany w organizacji szkoły, sposobach nauczania itp. Można wtedy argumentować, że nie tylko mamy rodzaj eksperymentu, lecz jest to też eksperyment w naturalnych warunkach, które często łatwo można uogólnić na szerszą populację. Co więcej, „naturalność" tej sytuacji powoduje, że jej uczestnicy nie są świadomi uczestniczenia w zaplanowanym eksperymencie, bo też najczęściej w nim nie uczestniczą, a dane analizowane są po pewnym czasie, gdy osoby te nie mają już żadnego wpływu na to, jakie zostaną wyciągnięte z nich wnioski. Z tego względu naturalne eksperymenty są uważane przez wielu, włączając piszących te słowa, za metodę doskonalszą od planowanego eksperymentu, szczególnie jeśli chodzi o badania edukacyjne.

Istnieje właściwie tylko jedno, choć zasadnicze, ograniczenie dla tych metod. Naturalne eksperymenty są rzadkie, a ich „odnalezienie" wymaga sporej inwencji badacza. Przedstawione wyżej przykłady badań to prace znane i niezwykle cenione wśród ekonomistów zajmujących się edukacją, opublikowane w najlepszych czasopismach. Podobne pomysły są jednak dość rzadkie i oczywiście nie każdy program edukacyjny można w ten sposób ocenić. W części prezentującej wyniki badań omawiamy prace wykorzystującą podobne metody do oceny wpływu wielkości klasy na wyniki uczniów w Polsce.

2.6. Badania z użyciem danych z całej populacji

Posługując się wynikami egzaminów, często wykorzystujemy dane dla całej populacji uczniów. Przykładowo, mamy wyniki wszystkich szkół w Polsce i dokonujemy analizy statystycznej tych danych. Wydawałoby się, że ponieważ posiadamy informację o wynikach wszystkich szkół, mamy też pełną pewność, w sensie statystycznym, dotyczącą tych wyników. Inaczej mówiąc, nie ma sensu analizować tych danych, posługując się testami statystycznymi. Przykładowo, jeśli chcemy porównać dwa województwa, to prostu porównujemy średnie wyniki egzaminów uzyskane przez uczniów z nich pochodzących i wnioskujemy w ten

[19] Angrist, J. D., Krueger, A. B. (1991), *Does Compulsory School Attendance Affect Schooling and Earnings*?, „The Quarterly Journal of Economics", vol. 106(4), s. 979-1014.

sposób o różnicy w umiejętnościach i wiadomościach. Czy takie podejście jest prawidłowe? Odpowiedź nie jest, wbrew pozorom, prosta i dotyka założeń leżących u podstaw statystyki.

Z jednej strony, wynik ucznia można traktować jako pewnik, traktując ocenę z testu wykonanego danego dnia jako cel sam w sobie i ostateczny rezultat pomiaru jego wiedzy. 30 punktów z egzaminu to 30 punktów i nie ma potrzeby zastanawiać się, w jakim stopniu pewne jest, że uczeń posiada rzeczywiście wiadomości i umiejętności na tym poziomie, bowiem interesuje nas tylko jego wynik z danego dnia i w określonym teście. W takiej sytuacji można też przypuszczać, że możemy porównać średnie wyniki dla województw, traktując podobnie wyniki wszystkich uczniów zdających określony egzamin w danym dniu i w danym województwie.

Tego rodzaju podejście dominuje w raportowaniu wyników egzaminów i ma to swoje głębokie uzasadnienie. Egzamin jest sprawdzianem danego dnia w oparciu o dany test. Nie interesuje nas to, jaką uczeń wiedzę posiada w rzeczywistości, o ile nie potrafi jej wykazać na egzaminie. Odpowiada to założeniu o jednorazowym pomiarze umiejętności podczas egzaminów i podawaniu wyników na skali w prosty sposób oddający liczbę poprawnych odpowiedzi. Porównanie wyników województw nie jest z tego punktu widzenia obarczone niepewnością, bowiem egzaminowi zostali poddani wszyscy uczniowie.

Co jednak, gdy interesuje nas tak naprawdę, czy uczniowie dwóch województw różnią się pod względem wiadomości i umiejętności, a nie jedynie wyniku egzaminu z danego dnia? Czy wtedy możemy po prostu porównać średnie w województwach? Istnieją co najmniej dwa argumenty, które powodują, że nawet przy porównaniach wyników dla całych populacji stosuje się testy statystyczne odzwierciedlające niepewność tych porównań. Pierwsze źródło niepewności wynika z szeroko pojętej rzetelności pomiaru (patrz rozdział 1). Jak wiemy, każdy test obarczony jest błędem pomiaru, czy mówiąc inaczej, nie ma stuprocentowej rzetelności. Ten sam uczeń uzyskałby różne wyniki na tym samym teście, choć jego prawdziwy poziom wiadomości i umiejętności nie uległ zmianie. Kwestie te szeroko omawia rozdział 1.

Drugie źródło niepewności związane jest z pojęciem superpopulacji. Jeśli porównujemy szkoły, regiony, kraje, to można argumentować, że nawet jeśli posiadamy wyniki wszystkich uczniów zdających egzamin określonego dnia, to powinniśmy stosować podobne testy statystyczne, jak w wypadku losowych prób uczniów. Argument ten opiera się na koncepcji superpopulacji, a jego zasadność zależy od kontekstu, w jakim dokonujemy porównań. Krótko mówiąc, koncepcja superpopulacji zakłada, że grupa uczniów (szkół), na której dokonujemy pomiaru, stanowi jedynie próbkę wszystkich potencjalnych uczniów (szkół). Podobnie jak losowa próba z pełnej populacji danego rejonu, tak cała populacja uczniów tego rejonu z danego momentu czasu daje jedynie pewien pomiar możliwości wszystkich potencjalnych uczniów tego rejonu.

Choć argument i koncepcja mogą wydawać się dość przewrotne, to jednak mają głęboki sens, szczególnie w badaniach, gdzie dokonujemy porównań cech uważanych za permanentne lub przynajmniej oddziałujące przez dłuższy czas. Dotyczy to na przykład badań jakości nauczania czy społeczno-kulturowych uwarunkowań wyników uczniów. W takiej sytuacji nie interesuje nas jedynie jakość nauczania uzyskana z pojedynczą grupą uczniów czy też oszacowana dla niej relacja między uwarunkowaniami społeczno-kulturowymi a wynikami. Interesuje nas, jaki charakter miałaby taka relacja także wtedy, gdyby szkoła pracowała z podobną, ale jednak nie identyczną grupą uczniów.

Tak więc interpretacja i sposób wykorzystania wyników dla całej populacji zależą od kontekstu. Jeśli chodzi nam o porównanie wyników odzwierciedlających ogólny potencjał danej szkoły, regionu, kraju, to pojęcie superpopulacji ma głęboki sens. Jeśli chcemy porównać wyniki konkretnej grupy z danego dnia, to pojęcie to nie ma tu zastosowania, choć i tak wynik może być podawany jako wartość z przedziałem ufności, ze względu na błąd pomiaru wspomniany powyżej.

Dobrym przykładem zastosowania teorii superpopulacji są miary edukacyjnej wartości dodanej, najczęściej liczone dla całej populacji szkół w danym kraju. EWD często liczona jest przez modele wielopoziomowe z efektami losowymi, gdzie zakłada się, że efekty szkół, czyli to, co w praktyce uznajemy za miary EWD szkół, mają rozkład normalny, tak jak wiele cech w przyrodzie (na przykład wzrost). Można się spierać, że mamy tu do czynienia z pełną populacją szkół, więc nie ma sensu zakładać, że efekty te mają jakikolwiek rozkład o naturze losowej, skoro są dokładnie znane. Inaczej mówiąc, można twierdzić, że szkoła uzyskała określoną EWD w konkretnym roku z konkretnym rocznikiem uczniów i w określonym przedmiocie, nie ma więc żadnego sensu zakładanie, że wynik ten jest losowy (dokładniej, że jest jedynie zaobserwowaną realizacją zmiennej losowej o rozkładzie normalnym). Trzeba jednak zauważyć, że w przypadku EWD nie ograniczamy się do pomiaru jakości nauczania szkoły w odniesieniu do konkretnej grupy uczniów czy danego testu, ale raczej chcemy poznać miarę jakości nauczania, która dotyczyłaby dowolnej grupy uczniów i dowolnego testu mierzącego umiejętności i wiedzę z interesujących nas przedmiotów. Inaczej mówiąc, chcemy poznać jakość nauczania trwale charakteryzującą daną szkołę, która widoczna byłaby także z inną grupą uczniów[20].

Można więc argumentować, że „zaobserwowana" EWD to tylko jedna z możliwych do uzyskania przez daną placówkę miar jakości nauczania. Z innymi uczniami i nieco innym testem szkoła uzyskałaby zapewne nieco inną EWD, choć również bliską prawdziwej wartości określającej jakość jej pracy. Daje to podsta-

[20] Inaczej argumentując, jeśli EWD szkoły byłaby zupełnie inna z innymi uczniami, to podważałoby to sens całego podejścia. EWD liczona jest właśnie po to, aby uzyskać miarę jakości pracy szkoły niezależną od grupy uczniów, którzy w szkole tej akurat się uczą, ale zależną od wysiłków jej pracowników. Oczywiście, że obserwowalna EWD może zmieniać się zależnie od składu uczniowskiego, ale powinny być to zmiany natury losowej, a nie systematycznej.

wę do szacowania EWD modelami statystycznymi na danych z całej populacji oraz podawanie wartości EWD w przedziale ufności. W istocie uwzględnienie losowej natury EWD ma głęboki sens, jako że miara ta powinna odzwierciedlać trwałe, niezależne od składu uczniowskiego i narzędzia testowego, cechy danej placówki, które obserwujemy jedynie z pewnym błędem.

2.7. Badania na małych grupach

Bardzo często badania edukacyjne prowadzone są na niewielkich grupach uczniów kilku klas lub szkół. Także w tym wypadku możliwe jest stosowanie narzędzi statystycznych, dzięki którym można przeprowadzać odpowiednie testy potwierdzające prawomocność wniosków, jakie chcemy wyciągnąć. Taka możliwość istnieje jednak tylko wtedy, gdy dobór uczniów do badania jest w pełni losowy. Jeśli badaniu poddano nawet kilkuset uczniów z kilku szkół, które wybrane zostały nielosowo, na przykład dzięki znajomościom udało się po prostu uzyskać zgodę dyrektora szkoły, to wnioski ani nie mogą być poddane testom statystycznym, ani też nie powinny być uogólniane poza grupę uczniów poddaną badaniu.

Można więc powiedzieć, że badania na niewielkich próbach mogą być podstawą do wyciągania wniosków przydatnych dla innych, ale tylko wtedy, gdy próby zostały dobrane w pełni losowo. Nawet jeśli planujemy lokalne badanie, a środki pozwalają jedynie na poddanie testom kilkudziesięciu-kilkuset uczniów, to postarajmy się o ich losowy dobór (na przykład wylosowanie klas, a potem uczniów w klasach). Choć z pewnością nieco skomplikuje to planowanie i realizację badania, to przyniesie korzyść w postaci znacznie wyższej wiarygodności wyników. Jeśli chcemy posłużyć się wynikami już przeprowadzonego badania, to koniecznie sprawdźmy, czy było ono przeprowadzone na losowej próbie. Jeśli nie, to wyniki mogą być zupełnie nieadekwatne do naszej sytuacji i w każdym wypadku trudno uznać je za wiarygodne.

Badania na niewielką skalę mają jednak ważną zaletę: najczęściej badacz świetnie zna kontekst i może łączyć analizy ilościowe z jakościowymi. Choć niewielkie próby powodują, że analizy statystyczne będą siłą rzeczy dość ograniczone, to szczególnie w edukacji zrozumienie kontekstu jest zawsze rzeczą trudną do przecenienia. Badania na niewielkich grupach, uzupełniane o bardziej szczegółowe analizy jakościowe (indywidualne wywiady pogłębione, badania fokusowe, analiza dokumentów, obserwacja itp.) mogą stanowić wstępny lub nawet końcowy etap badań ilościowych na większą skalę. Na wstępnym etapie stanowią swego rodzaju pilotaż lub też po prostu próbę dookreślenia pytań badawczych, skonfrontowania ich z rzeczywistością. Na etapie końcowym badania jakościowe mogą świetnie uzupełniać wyniki ilościowe o głębsze interpretacje, wskazanie prawdopodobnych wyjaśnień zauważonych zależności itp.

Mało przydatne jest przy tym wprowadzanie sztucznych granic między badaniami ilościowymi i jakościowymi. Sprawnie skoordynowane, świetnie się uzupełniają. Badania jakościowe nigdy nie mogą stanowić podstawy do uogólnień o procesach, poziomie wiedzy lub opiniach w całej populacji uczniów, nauczycieli czy szkół. Za to badania ilościowe często grzeszą brakiem znajomości kontekstu, powierzchownością wyników i ich interpretacji.

Na małych grupach szczególnie przydatny może być podstawowy schemat eksperymentalny. O ile losowo dobierzemy grupy eksperymentalne, to nawet niewielka liczba uczestniczących w badaniu uczniów może dać bardzo ciekawe i statystycznie istotne wyniki. Prawidłowo przeprowadzony eksperyment daje rezultaty, które wymagają użycia najprostszych metod statystycznych (testy na różnice w średnich, proste statystyki), niewymagających dużej liczby obserwacji, w odróżnieniu od technik zaawansowanych. Należy jedynie pamiętać o zachowaniu losowości doboru grup i upewnieniu się, że potencjalne różnice między nimi zostały zniwelowane do minimum. Niestety, wiele badań nie przestrzega tych podstawowych zasad, więc warto upewnić się, wykorzystując ich wyniki, czy rzeczywiście są one prawomocne.

Literatura

Babbie, E. (2005), *Badania społeczne w praktyce*, Wydawnictwo Naukowe PWN: Warszawa.

Fink, A. (2009), *How to Conduct Surveys. A Step-by-Step Guide*, SAGE Publications, wyd. IV

Ross, Kenneth. R, *Module 3: Sample design for educational survey research*, UNESCO IIEP, www.iiep.unesco.org/capacity-development/training/training-materials/quantitative-research

Snijders, T. A. B., Bosker, Roel, J. (2004), *Multilevel Analysis*, Sage, Thousand Oaks – London – New Dehli.

Wolf, Richard M., *Module 4: Judging educational research based on experiments and surveys*, UNESCO IIEP, www.iiep.unesco.org/capacity-development/training/training-materials/quantitative-research

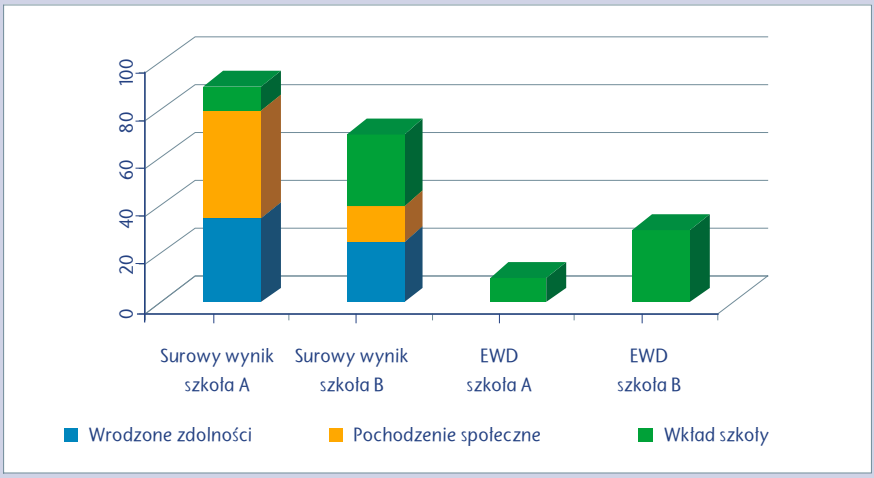
3. Ewaluacja metodami edukacyjnej wartości dodanej

W tej części opisujemy modele edukacyjnej wartości dodanej (EWD), które mogą być stosowane w badaniach edukacyjnych o różnorodnej tematyce zarówno na podstawie danych z całej populacji uczniów, jak i z reprezentatywnych prób. Modele EWD są znacznie bardziej wiarygodne niż typowe analizy poziomu kształcenia nieuwzględniające „wyjściowego" poziomu umiejętności uczniów. Przy ewaluacji polityki edukacyjnej interesuje nas bowiem wpływ, jaki mają oceniane programy edukacyjne na *przyrost* wiedzy uczniów, a nie na jej *poziom*. Poziom wiedzy związany jest bowiem przede wszystkim z pochodzeniem społecznym i wrodzonymi zdolnościami ucznia, a także z wcześniej zdobytymi umiejętnościami i wiadomościami, na które w danym momencie nie możemy wpływać. Przyrost wiedzy zależy już jednak w dużej mierze od tego, co działo się w trakcie nauki na określonym poziomie.

Modele EWD są narzędziami statystycznymi mającymi charakter ilościowy. Nie zastąpią one nigdy metod ewaluacji narzędziami jakościowymi, które co prawda nie dają podstawy do obiektywnej analizy porównawczej jakości pracy szkół, ale za to pozwalają wniknąć głębiej w procesy działające w szkołach[21]. W praktyce ważne jest stosowanie obydwu narzędzi, które dają wzajemnie uzupełniającą się wiedzę.

Różnice między oceną na podstawie końcowych wyników egzaminacyjnych a w oparciu o miary EWD ukazuje wykres poniżej. Trzeba od razu podkreślić, że wartości na wykresie są całkiem hipotetyczne, jednak zapewne bliskie prawdziwym zależnościom. W istocie jest bowiem tak, że osiągnięcia ucznia w znacznym stopniu zależą od jego wrodzonych zdolności oraz od innych cech związanych

[21] Metody jakościowe opisane są w książce: Ruchała, K. (2008), *Metodologia badań nad edukacją*, Wydawnictwa Akademickie i Profesjonalne: Warszawa.



Rysunek 3.1. Różnice w ocenie szkoły na podstawie surowego wyniku egzaminu i na podstawie EWD

z pochodzeniem społecznym jego rodziny. Potwierdzają to dziesiątki badań, gdzie status społeczno-ekonomiczny ucznia jest zawsze jednym z najsilniejszych predyktorów jego osiągnięć, podobnie zresztą, jak wcześniejsze osiągnięcia ucznia, zazwyczaj niezależnie od przedmiotu i rodzaju testu. W praktyce trudno jest ocenić, jaka część wyniku zależy od cech wrodzonych i nabytych ucznia, a jaka od pracy szkoły, bowiem wiele z tych zależności nie jest bezpośrednio obserwowalnych[22]. Ważne jest jednak to, że wrodzone i nabyte cechy ucznia z pewnością mają nie mniejsze znaczenie niż wysiłki nauczycieli.

Wpływa to znacząco na porównania według końcowych wyników uzyskanych na egzaminie. Na wykresie poniżej (rysunek 3.1) przedstawiono średnie wyniki szkoły A i B oraz miary EWD obliczone dla tych szkół. Przedstawiono też hipotetyczny „skład" surowego wyniku, który uzależniono od trzech składników: wrodzonych zdolności, pochodzenia społecznego, a także jakości nauczania (wkładu szkoły). Nieuwzględnienie wcześniejszego wyniku uczniów, który zależy przede wszystkim od ich pochodzenia społecznego oraz zdolności, powoduje, że lepsza wydaje się szkoła A, której uczniowie uzyskali wyższe wyniki, głównie dzięki wyższemu poziomowi wrodzonych zdolności i korzystniejszemu pochodzeniu społecznemu.

[22] Jak na przykład oddzielić cechy wrodzone od nabytych? Zazwyczaj szkoły lepsze gromadzą też zdolniejszych lub bardziej zmotywowanych uczniów. Jak więc ocenić wkład szkoły, skoro wiemy, że uczniowie grupują się (dokonują samodzielnego wyboru, kierują nimi rodziny lub też przydzielają ich nauczyciele) w placówkach czy klasach według zdolności?

Uwzględnienie wcześniejszego wyniku powoduje, że oceniamy głównie wkład szkoły, który jest większy w wypadku placówki B. Takie porównanie umożliwiają miary EWD, dzięki czemu szkoły oceniane są niezależnie od wrodzonych i nabytych zdolności uczniów. Oczywiście sytuacja tu przedstawiona jest dość abstrakcyjna, a dodatkowo zakłada, że metody EWD są w stanie całkowicie wyłączyć wpływ wrodzonych i nabytych cech ucznia niezależnych od wysiłków szkoły. W praktyce nie jest to nigdy w pełni możliwe, jednak badania potwierdzają, że metody EWD są obecnie najlepszym sposobem dokonywania oceny „wkładu szkoły" w wyniki egzaminacyjne uczniów. W dużym stopniu metody potrafią wyłączyć wpływ czynników od szkoły niezależnych, choć wiele zależy tu od jakości pomiaru umiejętności i wiedzy uczniów, a także od poprawności modeli statystycznych EWD[23].

Oczywiście zadaniem nauczycieli i szkoły jest przezwyciężanie barier w uczeniu się związanych z pochodzeniem społecznym i wrodzonymi talentami, a także wyrównywanie umiejętności i wiedzy, z którymi uczniowie „przychodzą" do szkoły. O ile istnieją chlubne przykłady pokonywania barier pochodzenia społecznego, a także wyrównywania braków w zdolnościach ciężką pracą, to jednak pochodzenie społeczne, wrodzone zdolności i wcześniej nabyte umiejętności i wiadomości są zawsze najlepszymi predyktorami obecnych wyników uczniów. W badaniu PISA około 15% wariancji wyników uczniów w Polsce wyjaśnianych jest pochodzeniem społecznym, a modele EWD objaśniają około 50-60% osiągnięć uczniów w gimnazjum poziomem wiadomości i umiejętności po szkole podstawowej (mierzonych wynikami sprawdzianu szóstoklasistów). Biorąc pod uwagę, że znaczna część nieobjaśnionej wariancji związana jest z losowym błędem pomiaru, to staje się oczywiste, że obecni nauczyciele i szkoła mają wpływ na stosunkowo niewielką część końcowego wyniku ucznia. Stąd ocena ich pracy na podstawie całego wyniku z egzaminu nie ma większego sensu i potrzebne są modele EWD, które oceniają wyniki pracy, biorąc pod uwagę jedynie tę część wyniku, na którą szkoła może mieć wpływ.

W Polsce opracowano i udokumentowano metodę EWD dla gimnazjów (patrz Dolata R. (red.), (2007), *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie egzaminów zewnętrznych*, CKE, a także informacje na stronie www.ewd.edu.pl). Wariant tej metody przedstawiono w niniejszym rozdziale, podając przykłady jej wykorzystania w ewaluacji wewnątrzszkolnej.

[23] Najpełniejsze jak dotąd omówienie tych kwestii wraz z odwołaniami do właściwej literatury można znaleźć w monografii RAND Corporation poświęconej modelom EWD (por. McCaffrey D., Lockwood J., Koretz M., Hamilton L. (2005), *Evaluating Value-Added Models for Teacher Accountability*, Rand Corporation MG-158), a także w niedawnej publikacji OECD poświęconej także tym modelom (OECD (2008), *Measuring Improvements in Learning Outcomes*: *Best Practices to Assess the Value-Added of Schools*, Paris).

Poza wzięciem pod uwagę wejściowego i wyjściowego poziomu umiejętności, zaletą ewaluacji metodami EWD wykorzystującymi wyniki egzaminów zewnętrznych jest możliwość ilościowej analizy znacznej liczby danych, nawet dla całej populacji. Dzięki temu uzyskane rezultaty są bardzo precyzyjne, a wnioski można uogólniać na uczniów, nauczycieli i szkoły w całej Polsce (lub mniejsze populacje, dla których dokonano analizy). W ten sposób metody EWD dają podstawę do oceny polityki edukacyjnej na każdym poziomie decyzyjnym: ogólnokrajowym, regionalnym, a nawet gminy lub szkoły. W Polsce powstało kilka prac dających przykład wykorzystania metody EWD do oceny efektywności wydatków oraz wpływu wielkości klas na przyrost wiedzy uczniów w polskich gimnazjach. Prace te omawiamy w rozdziale 4, jednak czytelnicy zainteresowani zastosowaniem modeli EWD do analiz statystycznych mogą sięgnąć do oryginalnych artykułów[24].

Warto nadmienić, że zarówno wydatki na gimnazja, jak i liczebność uczniów w klasach, dotyczą finansowych nakładów na edukację gimnazjalną. Ewaluacja metodą EWD może jednak dotyczyć dowolnych aspektów polityki edukacyjnej, o ile możliwe są one do skwantyfikowania. W polskim kontekście nasuwają się liczne pomysły na podobne badania dotyczące na przykład wpływu wykształcenia nauczycieli, segregacji uczniów między szkołami i klasami, wpływu różnic w realizowanych programach i metodach nauczania itp. Ważnych kwestii do badania jest sporo, warto więc spojrzeć, jak można wykorzystać modele EWD do tego celu.

3.1. Modele EWD – wprowadzenie

Idea leżąca u podstaw metod edukacyjnej wartości dodanej jest dość prosta. Według tego podejścia, dla oceny efektów kształcenia na danym etapie, należy przy ocenie końcowych wyników ucznia wziąć pod uwagę jego wyniki, zanim rozpoczął naukę na tym etapie. Mogłoby się wydawać, że najprostszym sposobem obliczenia EWD jest odjęcie wyników pomiarów umiejętności danego ucznia sprzed rozpoczęcia i po zakończeniu nauki na danym etapie. Różnica miałaby wskazywać na przyrost wiedzy ucznia. Jednak takie podejście nie jest obecnie polecane. Po pierwsze, skale pomiarowe dwóch testów musiałyby być identyczne, co jest w praktyce niemal niemożliwe do osiągnięcia, szczególnie dla pomiarów

[24] Patrz: Jakubowski, M. (2008), *Zastosowanie modeli EWD do ewaluacji polityki edukacyjnej*, gdzie opisano modele statystyczne leżące u podstaw zaprezentowanych modeli. Dokładniejsze wyniki wraz z opisem kontekstu badania, metodologii, danych oraz interpretacją rezultatów, można znaleźć w pracach: Jakubowski, M. (2007), *Czy wydatki na gimnazja są efektywne*?, „Gospodarka Narodowa", 11-12/2007, a także Jakubowski, M. (2008), *Teaching quality and decentralization*. Wszystkie te prace dostępne są na stronie www.wne.uw.edu.pl/mjakubowski w zakładce „publikacje".

odległych w czasie, gdzie zarówno zestaw zadań, jak i umiejętności oraz inne cechy rozwojowe ucznia są zapewne bardzo różne. Po drugie, wykazano, że dynamika rozwoju i jej powiązania z pomiarem umiejętności powodują, że proste różnice w wynikach są kiepskim wskaźnikiem przyrostu wiedzy uczniów, szczególnie jeśli chcemy mierzyć przyrost dla szerszej populacji[25].

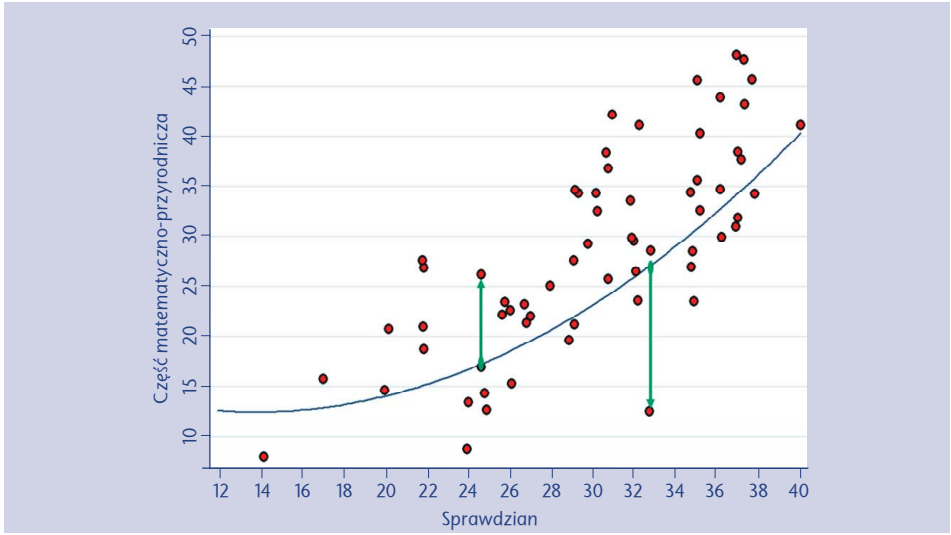
Z tych względów popularne jest obecnie podejście, w którym metoda regresji liniowej służy do określenia, w jakim stopniu grupy uczniów wykazują różny przyrost wiedzy. Dla pojedynczych uczniów bardziej wskazane jest analizowanie wyników na tych samych skalach pomiaru. Jeśli takimi skalami nie dysponujemy, tak jak w wypadku egzaminów zewnętrznych w Polsce, to bezpośrednie porównywanie wyników uczniów z różnych egzaminów nie ma żadnego sensu i może być bardzo mylące[26].

Metoda regresji liniowej jest w swojej istocie niezwykle prosta, choć już same modele statystyczne służące do oceny EWD są zazwyczaj dość złożone. Wyobraźmy sobie, że dla znacznej liczby uczniów posiadamy pomiary wiedzy z dwóch momentów w czasie. Chcemy ocenić, dla których uczniów przyrost wiedzy był ponadprzeciętny lub też poniżej przeciętnej. W tym celu szacujemy równanie regresji dla jak najszerszej grupy uczniów. W najprostszym wypadku, gdy chcemy wykorzystać jedynie dwa wyniki testów, po prostu określamy prostą lub krzywą, która najlepiej opisuje relację między nimi. Takie dopasowanie krzywej określającej relację między wynikami sprawdzianu szóstoklasistów oraz wynikami egzaminu gimnazjalnego w części matematyczno-przyrodniczej dla jednego z roczników uczniów pokazuje wykres poniżej (rysunek 3.2).

Nieliniowość tej zależności wynika z tego, że skale wyników egzaminów są zupełnie różne. Jednak (nawet jeśli wyniki byłyby odniesione do tej samej skali) tego typu nieliniowość może wynikać choćby z tego, że uczniowie o wyższym potencjale początkowym szybciej zdobywają wiedzę na kolejnym etapie nauczania. Trudno określić, który z tych powodów jest ważniejszy w tym wypadku, niemniej jednak widać z tego przykładu, że liczenie zmiany w poziomie wiedzy

[25] McCaffrey D., Lockwood J., Koretz M., Hamilton L. (2005), *Evaluating Value-Added Models for Teacher Accountability*, Rand Corporation MG-158.

[26] Wyniki polskich egzaminów nie są bezpośrednio porównywalne ani między latami w tym samym przedmiocie, ani między przedmiotami. Każdy egzamin posiada swoją własną skalę i odejmowanie wyników od siebie nie ma większego sensu. Widać to choćby po różnicach w podstawowych wskaźnikach statystycznych dla wyników egzaminów: średniej, medianie czy odchyleniu standardowym, które co roku są inne dla każdego przedmiotu. Tak więc ani 30 punktów nie oznacza tego samego na sprawdzianie w roku 2007 i 2008, ani też 30 punktów nie oznacza tego samego na egzaminie gimnazjalnym w 2008 roku w części humanistycznej i części matematyczno-przyrodniczej. Co więcej, różnica między 30 a 25 punktami nie oznacza tego samego w obu częściach egzaminu gimnazjalnego czy w tej samej części między latami. Zaletą opisywanych modeli EWD jest to, że nie wymagają porównywalności skal punktowych między egzaminami, choć zrównanie skal zwiększyłoby zapewne ich rzetelność, a już na pewno ułatwiłoby interpretację.



Rysunek 3.2. Dopasowanie zależności między wynikami sprawdzianu a wynikami egzaminu gimnazjalnego w części matematyczno-przyrodniczej

uczniów, a następnie EWD szkół, przez proste odejmowanie wyników egzaminów nie ma tu najmniejszego sensu.

Kropki na wykresie to wyniki uczniów danej szkoły. Jej EWD w najprostszej metodzie liczymy jako średnią różnicę między wynikiem uzyskanym a wynikiem przewidywanym, który jest niczym innym, jak punktem leżącym na krzywej, a który można interpretować jako przeciętny wynik egzaminu gimnazjalnego uzyskany przez uczniów o tym samym rezultacie ze sprawdzianu. Różnice te odzwierciedlają dla dwóch przykładowych uczniów pionowe strzałki. Jeden z nich uzyskał wynik znacznie wyższy niż oczekiwany (jego „kropka" leży nad krzywą przewidywanych wyników), podczas gdy drugi uzyskał wynik niższy (jego „kropka" leży pod krzywą przewidywanych wyników). Średnie z takich różnic można traktować jako miarę przeciętnego przyrostu wiedzy w danej szkole. W tej szkole większość uczniów uzyskała wyniki egzaminu wyższe od przewidywanych prezentowaną krzywą. Jeśli krzywa ta została oszacowana dla całej populacji uczniów w Polsce, a tak powinno być i jest w tym wypadku, to można powiedzieć, że uczniowie tej szkoły uzyskali wyższe wyniki na egzaminie, niż ich rówieśnicy o tych samych wynikach sprawdzianu. Stąd EWD tej szkoły będzie zapewne dodatnia.

Dopasowana krzywa opisuje przeciętny wynik egzaminu gimnazjalnego w części matematyczno-przyrodniczej, jaki uzyskali uczniowie o danym wyniku sprawdzianu, przy założeniu, że wraz z wynikiem sprawdzianu przewidywany wynik egzaminu gimnazjalnego rośnie w sposób uporządkowany i przewidywalny, tzn. niemożliwe jest, że uczniowie z wyższym wynikiem ze sprawdzianu

uzyskaliby średnio niższe wyniki egzaminu lub też, że mamy do czynienia z nagłymi skokami (na przykład między kolejnymi wynikami sprawdzianu 28, 29, 30 punktów różnica w średnich wynikach egzaminu wynosi 20, 21, 27). Inaczej mówiąc, zakładamy, że zależność tę da się dobrze opisać prostą lub krzywą. W tym wypadku potwierdza to dobre dopasowanie krzywej do danych, poświadczone odpowiednimi testami statystycznymi[27].

Niezwykle ważne jest to, że miara EWD jest miarą relatywną. Wyniki uczniów z egzaminu gimnazjalnego porównujemy z wynikami ich rówieśników, którzy uzyskali podobny wynik na sprawdzianie. Stąd dany uczeń może być ponadprzeciętny lub poniżej przeciętności – w zależności od grupy uczniów, z którą akurat przyszło mu się porównywać. Podobnie jego szkoła może mieć niższą lub wyższą EWD tylko w odniesieniu do szkół i uczniów z tego samego rocznika.

EWD jest miarą relatywną, nie może służyć do określenia efektywności nauczania czy postępów uczniów względem stałego układu odniesienia, tak jak na przykład mierzymy temperaturę względem „stopnia 0". Jak już wspomnieliśmy, tworzenie skal pomiarowych umiejętności i wiadomości uczniów ze stałym, wspólnym układem odniesienia jest teoretycznie możliwe, lecz w praktyce trudno osiągalne i przez wielu uważane za wątpliwe.

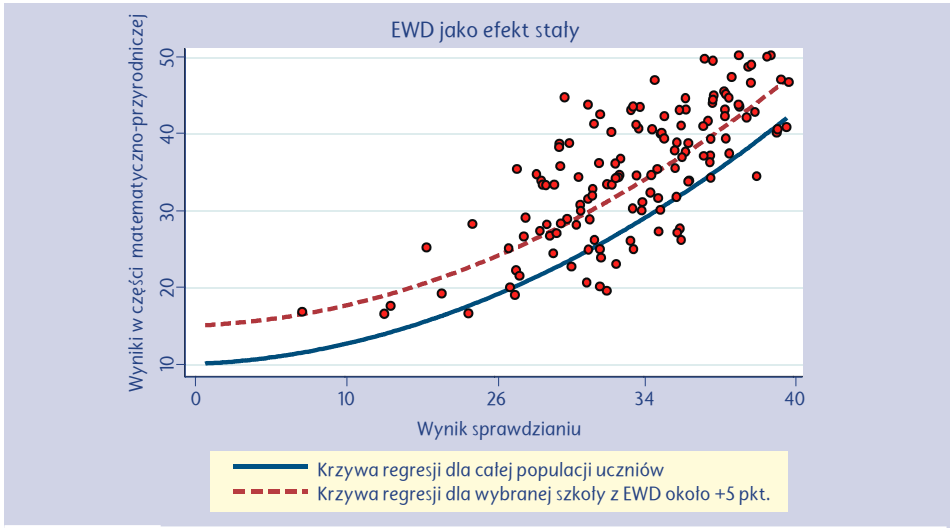
Metody EWD używane w badaniach edukacyjnych najczęściej wykorzystują modele efektów stałych lub losowych (tzw. modele wielopoziomowe). Zazwyczaj oba modele w podstawowych wersjach i dla dużych prób prowadzą do podobnych wniosków, przedstawiamy więc jedynie podstawową ideę leżącą u ich podstaw, bez specjalnego rozróżnienia między nimi[28]. Podstawowe założenia modeli EWD pokazuje wykres na sąsiedniej stronie (rysunek 3.3). Przedstawiono na nim ponownie zależność dla całej populacji oraz jej dopasowanie dla pojedynczej szkoły. Uczniowie tej szkoły uzyskali ponadprzeciętne, wyższe niż przewidywane wyniki na egzaminie gimnazjalnym, z tego względu dopasowana dla nich zależność leży powyżej krzywej dla populacji. Pionowa różnica w położeniu krzywych odzwierciedla EWD mierzoną w skali punktowej egzaminu gimnazjalnego jako tzw. „efekt stały".

EWD może służyć porównaniu przeciętnego przyrostu wiedzy między gimnazjami. Odpowiednie modele zostały opracowane dla danych z każdego rocznika przez zespół EWD powołany przez Centralną Komisję Egzaminacyjną i kierowany przez Romana Dolatę[29]. Opracowano także model podsumowujący dane z kilku, dokładniej trzech, roczników. Przykładowe wyniki pokazuje wy-

[27] Krzywa została dopasowana metodą regresji liniowej z wynikiem sprawdzianu i jego kwadratem jako predyktorami wyników egzaminu gimnazjalnego.

[28] Dokładniejszy opis tych modeli, a także analizę różnic w kontekście EWD liczonego dla polskich gimnazjów, przedstawiono w pracy: Jakubowski, M. (2007), *Analiza zmienności w czasie wskaźników edukacyjnej wartości dodanej*, Biuletyn Badawczy CKE „Egzamin" 14/2007, www.cke.edu.pl

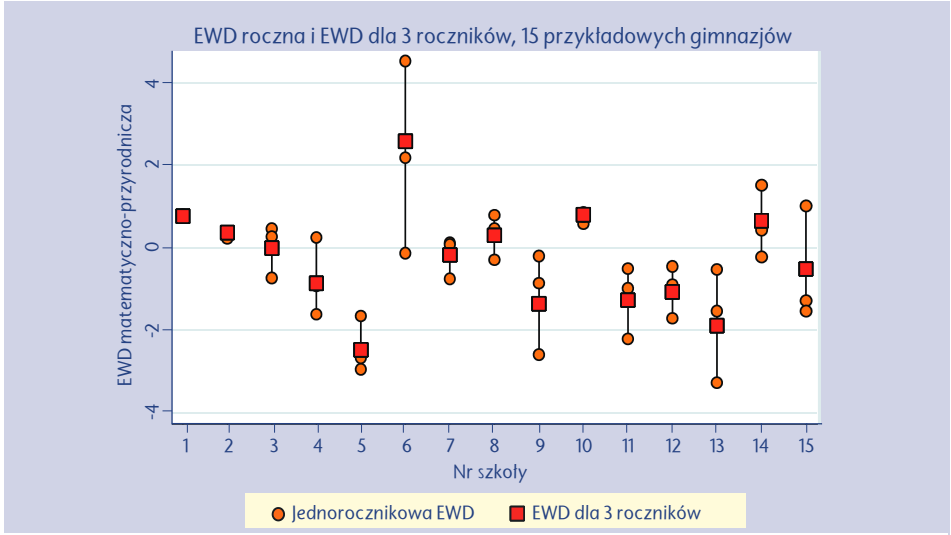
[29] Dokumentacja metod EWD dla polskich gimnazjów dostępna jest na stronie www.ewd.edu.pl



Rysunek 3.3. Dopasowanie zależności między wynikami sprawdzianu a wynikami egzaminu gimnazjalnego w części matematyczno-przyrodniczej dla całej populacji oraz po uwzględnieniu przesunięcia dla konkretnej szkoły

kres poniżej (rysunek 3.4). Kropki oznaczają tu EWD jednorocznikową, a kwadraty to EWD liczona łącznie dla danych z trzech kolejnych roczników.

Jak widać, różnice w ocenach między rocznikami są dość znaczne dla niektórych gimnazjów (w szczególności dla gimnazjum nr 6) i niewielkie dla innych



Rysunek 3.4. Porównanie EWD dla jednego i trzech roczników dla przykładowych gimnazjów

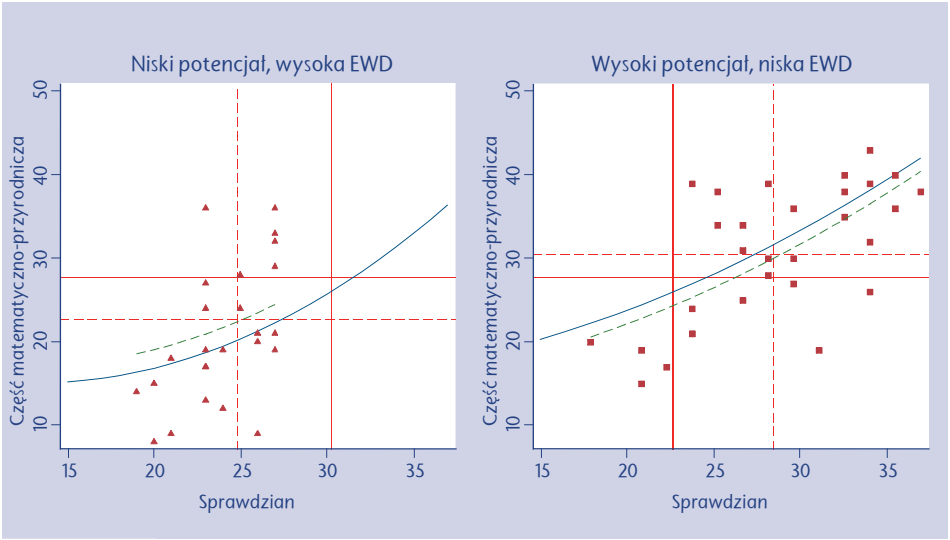
(szczególnie gimnazja 1, 2 oraz 10). Może to być spowodowane różną wielkością tych placówek. Dla małej liczby uczniów EWD szacowana jest z o wiele mniejszą precyzją niż dla szkół większych, stąd też może się znacznie różnić między rocznikami. Z tego powodu EWD podawana jest w przedziałach ufności, które określają jej precyzję, a dokładnie, w jakim przedziale wartości znajduje się z bardzo wysokim prawdopodobieństwem „prawdziwa" miara EWD.

Trzyrocznikowa EWD może służyć jako bardziej stabilna i precyzyjna miara, obliczana dla co najmniej kilkudziesięciu, a czasem i kilkuset uczniów. Jej precyzja jest wystarczająco wysoka dla oceny długofalowej jakości nauczania w gimnazjum. Oczywiście, EWD z różnych roczników mogą się znacząco różnić, choćby ze względu na zmiany w składzie nauczycielskim, zmiany metod nauczania itp. Wyjaśnienie tych różnic powinno być jednym z celów badań prowadzonych w oparciu o EWD.

Jak więc wykorzystać EWD do badań edukacyjnych? Podstawowe metody łatwo wyjaśnić na prostych wykresach, posługując się pojedynczymi szkołami, czy nawet klasami, czego przykład pokazujemy poniżej. Trzeba jednak podkreślić, że w praktyce jedynie duża liczba obserwacji, zarówno na poziomie uczniów, jak i na poziomie szkół, gwarantuje, że nasze wyniki będą wiarygodne i mogą być uogólniane na szerszą populację. Choć EWD może być stosowana nawet dla danych z pojedynczych placówek, to w tym wypadku służy raczej analizom pomocnym w bieżącej ewaluacji nauczania w danej szkole. Ewaluacja polityki edukacyjnej czy metod nauczania wymaga większych, losowo dobieranych, reprezentatywnych prób lub też danych z pełnej populacji jednego lub kilku roczników. Należy więc potraktować przykłady wskazane poniżej jako ukazujące intuicję leżącą u podstaw badania metodami EWD. Przykłady zastosowań różnych metod EWD dla danych z całej populacji opisano w przywołanych wcześniej artykułach, omawiamy je krótko także w rozdziale 4.

Zacznijmy od przykładu dwóch gimnazjów, które znacząco różnią się pod trzema względami: wyniku sprawdzianu uczniów do nich uczęszczających, końcowego wyniku egzaminu gimnazjalnego, a także pod względem EWD. Na wykresie na sąsiedniej stronie (rysunek 3.5) osobno przedstawiono rozrzuty wyników uczniów tych dwóch hipotetycznych szkół, przy czym uczniów gimnazjum z niższym wynikiem końcowym oznaczono trójkątami (wykres po lewej), a uczniów gimnazjum z wyższym wynikiem egzaminu kwadratami (wykres po prawej).

Wykres zawiera także pionowe linie określające wyniki egzaminów i EWD. Linie ciągłe oznaczają średnie wyniki w całej populacji. Linie przerywane oznaczają średnie wyniki dla danej szkoły. Przykładowo, dla szkoły po lewej, o niskim potencjale, średnia jej uczniów ze sprawdzianu leży poniżej średniej w populacji (na wykresie linia ją opisująca jest przesunięta w lewo), podobnie linia oznaczająca średnią z egzaminu leży poniżej średniej dla populacji. Dla szkoły po prawej mamy sytuację odwrotną, tutaj wyniki egzaminów są wyższe niż przeciętna w kraju.



Rysunek 3.5. Przykład gimnazjów o różnym potencjale i różnym EWD

Inaczej jest z EWD. Z wykresu wynika, że uczniowie szkoły po lewej uzyskali wyniki wyższe niż oczekiwane, stąd dodatnia EWD zaznaczona przesunięciem przerywanej krzywej dopasowanej dla tej placówki względem dopasowania dla całej populacji oznaczonego linią ciągłą. Porównując ze szkołą po prawej, widać jednak, że i tak końcowy wynik z egzaminu jest niższy. Uczniowie ze szkoły po prawej uzyskali wynik wyższy niż pozostali uczniowie w Polsce, zarówno ze sprawdzianu, jak i z egzaminu. Jednak EWD tej placówki jest niższa, bowiem uczniowie ci względem uczniów o podobnych wynikach sprawdzianu uzyskali na egzaminie mniej punktów, co pokazuje przesunięcie w dół przerywanej krzywej opisującej wyniki w tej szkole względem ciągłej krzywej przewidującej wyniki egzaminu względem wyników sprawdzianu dla całego kraju.

Wykres ten jasno pokazuje różnice między analizą w oparciu o surowe wyniki egzaminacyjne, a analizą metodami EWD. Załóżmy, że w szkole po lewej klasy są niewielkie, a w szkole po prawej klasy są bardzo duże, czyli na przykład, że mamy do czynienia ze szkołą wiejską i miejską. Odnosząc wyniki egzaminu do wielkości klasy, stwierdzilibyśmy dodatnią zależność: im większa klasa, tym wyższy wynik. Patrząc jednak na EWD, widzimy, że zależność jest odwrotna: im większa klasa, tym niższa EWD. Oczywiście EWD jest tu bardziej wiarygodnym wskaźnikiem, bowiem odnosi się do okresu nauczania, dla którego wielkość klasy jest liczona i bierze pod uwagę wyniki uczniów przed rozpoczęciem nauki w gimnazjum.

Choć nie jest pewne, że w oparciu o analizę EWD uzyskamy wyniki niezbicie świadczące o istnieniu lub braku jakiejś relacji, to z pewnością są one bar-

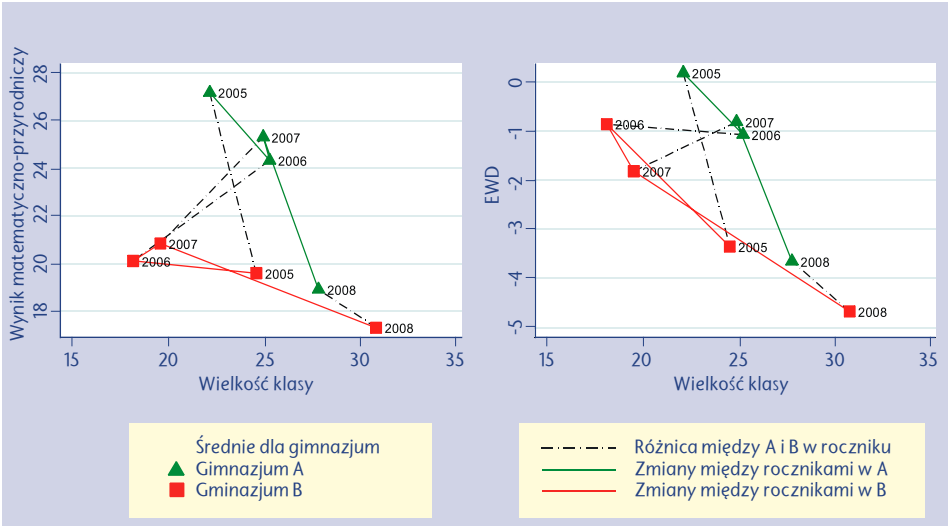
dziej wiarygodne niż proste korelacje z końcowymi wynikami egzaminów, bez uwzględnienia wyników wejściowych. Nadal jednak ostrożność jest wskazana. Może na przykład w szkole po lewej uczniowie korzystali z dodatkowych kursów, innego programu nauczania itp., czyli czynników, które mogłyby podnieść ich wyniki niezależnie od wielkości klasy. Tylko kontrolując wpływ tych czynników w regresji metodą EWD możemy ocenić, jaki wpływ miała sama wielkość klasy na przyrost wiedzy uczniów. Nieuwzględnienie jakiegokolwiek istotnego czynnika w regresji spowoduje, że uzyskane wyniki będą odbiegały od prawdziwych zależności przyczynowo-skutkowych. Jak można się domyślać, tego typu metody niemal zawsze pomijają jakieś ważne czynniki, stąd wskazane jest krytyczne spojrzenie na rezultaty nimi uzyskiwane.

Jednym ze sposobów na uniknięcie potencjalnych przekłamań w wynikach analizy związanych z różnicami między szkołami jest spojrzenie na zmiany w wynikach wyłącznie wewnątrz szkół. Inaczej mówiąc, wykorzystanie metod opierających się na wariancji wewnątrz obserwowanych jednostek, a nie pomiędzy nimi. W żargonie ekonometrii są to modele z efektami stałymi, a w nomenklaturze badań edukacyjnych często mówi się o wariancji wewnątrzszkolnej i międzyszkolnej. Idea jest prosta w swoich podstawowych założeniach: zamiast badać różnice między szkołami, które zależą od wielu czynników, których nie potrafimy w pełni kontrolować, należy spojrzeć na różnice wewnątrz szkół, gdzie środowisko społeczne i edukacyjne jest podobne i nie zmienia się istotnie nawet między rocznikami. Stąd łatwiej jest wyizolować interesujące nas zależności, patrząc na różnice wewnątrz rocznika w danej szkole lub między rocznikami w danej szkole.

Przykładowo, wpływ zmiany podręcznika na wyniki uczniów najlepiej analizować, porównując grupy uczniów korzystające z różnych podręczników wewnątrz tych samych szkół, następnie niejako uśredniając informację z dużej liczby placówek. Porównanie między szkołami obarczone byłoby błędem związanym ze wszystkimi innymi czynnikami wpływającym na poziom nauczania w tych placówkach, a już wykorzystanie końcowych wyników bez modeli EWD prowadziłoby z pewnością do błędnych wniosków. Podobnie wpływ zmian w wielkościach klas najlepiej analizować, patrząc na różnice w postępie (miary EWD) klas różnej wielkości w tej samej placówce, a nawet na zmiany EWD kolejnych roczników, gdzie zapewne wariancja wielkości klas jest większa.

Oczywiście nie chodzi o to, żeby skupić się jedynie na zmianach w jednej szkole, zamiast porównań między kilkuset szkołami. Chodzi o to, żeby podstawę oceny stanowiły tylko zmiany *wewnątrz* tych szkół, bez patrzenia na różnice *między* nimi, które następnie uśredniamy dla kilkuset szkół, co daje precyzyjną ocenę wpływu tych zmian dla całej populacji.

Różnice między podejściem wykorzystującym głównie wariancję między szkołami a podejściem skupiającym się na wariancji wewnątrz szkół ukazuje



Rysunek 3.6. Szacowanie efektu wielkości klasy w oparciu o dane dla kilku roczników

wykres poniżej (rysunek 3.6). Pomijamy na nim wariancję wewnątrzszkolną na poziomie uczniów (zróżnicowanie cechy w danym roczniku i konkretnej szkole) na rzecz zmian wartości danej cechy między latami w tej samej szkole. Inaczej mówiąc, patrzymy, jak zmiany wartości danej cechy w tej samej szkole wpływają na zmiany w średnich wynikach uczniów i jej EWD.

Na kolejnym etapie możemy takie zmiany policzyć dla wszystkich szkół i uśrednić dla całego kraju. W ten sposób nasze wyniki nie będą narażone na krytykę, że wynikają z pozornej korelacji związanej z różnicami między szkołami, których nie jesteśmy w stanie kontrolować. Wręcz przeciwnie, nasze wyniki opierają się jedynie na zmianach wartości interesujących nas zmiennych w tych samych placówkach, bez porównywania szkół między sobą. Wyniki te są więc w pełni odporne na wszelkie różnice między szkołami, bowiem przy ich obliczaniu po prostu szkół nie porównujemy. Tego typu metody nazywane są metodami efektów stałych, opierającymi się na wariancji „wewnętrznej", czy też metodami panelowymi, jeśli posiadamy obserwacje tej samej jednostki (szkoły lub ucznia) z kilku momentów w czasie. Wróćmy jednak do przykładu. Wykres pokazuje dwie placówki na przestrzeni 4 lat, o różnej EWD i różnej wielkości klas. Szkoła A oznaczona jest trójkątami, a szkoła B kwadratami.

Wykres po lewej stronie pokazuje wyniki egzaminu gimnazjalnego w każdej szkole (oś pionowa), osobno dla każdego rocznika, i odnosi je do wielkości klasy (oś pozioma). Wykres po prawej w podobny sposób odnosi przeciętne wielkości klas do EWD obliczonej dla różnych roczników w tych samych szkołach. Linie przerywane pokazują różnice między wynikami dwóch szkół osobno w każdym

spis treści

roczniku. Linie ciągłe pokazują różnice między wynikami roczników osobno w każdej ze szkół. Tak więc linie ciągłe odzwierciedlają zależności wewnątrz szkół, a linie przerywane między szkołami.

Wykres może wydawać się dość chaotyczny przy pierwszym spojrzeniu, jednak dość jasno pokazuje, jakie różnice bierzemy pod uwagę, szacując model w oparciu o wariancję między szkołami, a jakie – biorąc pod uwagę wariancję między rocznikami w tej samej szkole. Linie przerywane odpowiadają zależnościom między szkołami z tego samego rocznika. Zarówno dla surowego wyniku egzaminu (panel po lewej stronie), jak i dla EWD linie te mają zarówno dodatnie, jak i ujemne nachylenie, choć dla EWD tylko w jednym wypadku mamy do czynienia z zależnością pozytywną (rocznik 2007). Te różnice mogą być związane z cechami nieuwzględnionymi w tak prostej analizie, na przykład lokalizacją szkoły czy też jej popularnością (proszę zauważyć, że szkoła A ma niemal zawsze wyższe wyniki i EWD niż szkoła B, więc różnice w nachyleniu krzywych wynikają jedynie z tego, że w dwóch rocznikach szkoła A miała mniejsze klasy – 2005 i 2008 – a w dwóch większe). Z pewnością nie dają wiarygodnej podstawy do oceny zależności między wielkością klasy a wynikami uczniów, a biorąc pod uwagę ich różne nachylenie, sugerują raczej brak jakiejkolwiek zależności.

Linie przerywane odpowiadają natomiast zmianom między rocznikami, osobno w każdej ze szkół. W niemal wszystkich wypadkach wzrost wielkości klasy wiąże się ze spadkiem końcowego wyniku z egzaminu, ale też EWD szkoły, choć w wypadku EWD zmiany te wydają się lepiej widoczne. Inaczej mówiąc, w 7 na 8 przypadków proste mają nachylenie ujemne, tylko dla gimnazjum B między 2006 i 2007 niewielka zmiana w wyniku końcowym jest dodatnia i połączona ze wzrostem wielkości klas.

Widać więc, że patrząc jedynie na wariancję wewnątrz szkół, stwierdzilibyśmy silny ujemny wpływ wielkości klasy. Trzeba też podkreślić, że jeśli patrzymy tylko na wariancję w wielkości klas i wynikach w pojedynczych szkołach, to nawet badanie w oparciu o końcowy wynik egzaminu, bez miary EWD, daje dobrą podstawę do oceny zależności, o ile skład społeczny i potencjał edukacyjny roczników są między latami podobne. EWD poprawia tu na pewno wiarygodność modelu, jednak w rzeczywistości różnice między rocznikami w szkole są dużo mniejsze niż między uczniami z różnych szkół w danym roczniku, a przez to nawet porównywanie końcowych wyników między rocznikami tej samej placówki może stanowić satysfakcjonującą strategię badawczą, o ile dane EWD nie są dostępne. Oczywiście przykład powyższy nie może być uogólniany na całą populację. Takie uogólnienia możliwe są jedynie w oparciu o badania na danych reprezentatywnych dla populacji uczniów. Przykład takich badań omówiono w rozdziale 4.

3.2. Przykłady wewnątrzszkolnych lub wewnątrzgminnych analiz EWD

W tej części przedstawiamy kilka przykładów prostych analiz na poziomie szkoły i gminy. Więcej przykładów, jak i dokładny opis, w jaki sposób można ich dokonywać w oparciu o model EWD oszacowany dla całej Polski, można znaleźć w ogólnodostępnej broszurze autorstwa Ewy Stożek[30]. Poniżej przedstawiamy jedynie kilka wybranych przykładów analiz tego typu, odsyłając zainteresowanych wykorzystaniem EWD w ewaluacji wewnątrzszkolnej do wspomnianego tekstu.

Na stronie www.ewd.edu.pl dostępny jest też kalkulator EWD, który można wykorzystać do podobnych analiz (sposób jego działania opisuje dokładnie cytowana praca). Warto od razu podkreślić różnicę między wskaźnikiem EWD liczonym za pomocą kalkulatora, czyli w oparciu o model dla całego kraju, a liczonym metodą EWD dla innych danych. Przy wykorzystaniu kalkulatora EWD stanowi relatywną ocenę postępu grup uczniów względem innych uczniów całego kraju. W praktyce oznacza to, że wartość zero przypisana jest szkole, w której przyrost wiedzy uczniów nie odróżnia się od przeciętnego w kraju. Jeśli podobne modele oszacowano by dla mniejszej populacji, na przykład dla wszystkich uczniów danej gminy, to pomijając mniejszą precyzję, także interpretacja będzie odmienna. Takie modele mierzyć będą EWD względem uczniów i szkół tylko tej gminy, nie mogą stanowić podstawy porównań z uczniami z innych rejonów kraju. Z tego względu zalecamy wykorzystanie kalkulatora i dokumentacji metody EWD dla gimnazjów, dzięki czemu własne analizy będą posiadały wygodny układ odniesienia w postaci osiągnięć uczniów i szkół w całym kraju.

Tabela 3.1 oraz wykresy poniżej pokazują EWD obliczoną dla 3 szkół jednej gminy na podstawie wyników sprawdzianu szóstoklasistów oraz osiągnięć w części humanistycznej egzaminu gimnazjalnego z 2007 roku. Każda szkoła posiada trzy klasy. EWD obliczona została w oparciu o model dla całego kraju (czyli na przykład przy wykorzystaniu wspomnianego powyżej kalkulatora EWD), a więc wartość 0 stanowi przeciętną EWD dla Polski. Tabela zawiera EWD wraz z 95-procentowymi przedziałami ufności określającymi jej precyzję. EWD obliczono dla szkół, dla klas, a także dla szkół w podziale na płeć. Porównania EWD szkół i klas można dokonać, posługując się przedziałami ufności. Równoważne jest to wykonywaniu prostych testów statystycznych na różnice EWD między grupami. Dwie grupy mają różną EWD, o ile ich przedziały ufności nie posiadają wspólnych wartości, na przykład jeśli dolna granica przedziału ufności dla jednej z grup jest wyższa niż górna granica dla innej grupy, to ich EWD jest różna, wyższa w pierwszej z grup. W ten prosty sposób

[30] Stożek, E. (2008), *Analiza wyników egzaminów zewnętrznych z wykorzystaniem metody EWD w ewaluacji wewnątrzszkolnej*, dostępne pod adresem www.ewd.edu.pl/broszura_ewd.pdf

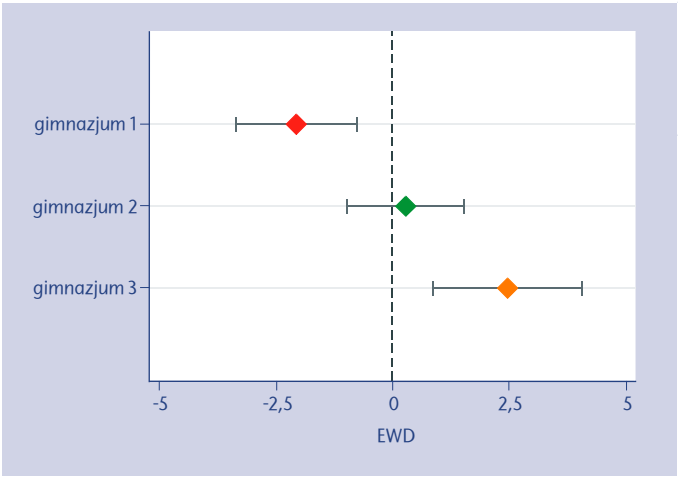
Tabela 3.1. Oszacowania EWD w części humanistycznej dla trzech przykładowych szkół oraz w podziale na klasy i płeć

Szkoła oraz grupa uczniów, dla której obliczono EWD		Wartość punktowa EWD	Dolna granica przedziału ufności EWD	Górna granica przedziału ufności EWD
gimnazjum 1	cały rocznik	-2,04	-3,34	-0,75
	klasa A	-0,82	-3,07	1,43
	klasa B	-3,95	-6,50	-1,39
	klasa C	-1,45	-3,54	0,64
	chłopcy	-1,95	-3,72	-0,18
	dziewczęta	-2,12	-4,05	-0,19
gimnazjum 2	cały rocznik	0,29	-0,96	1,54
	klasa A	2,16	0,29	4,03
	klasa B	-0,33	-2,82	2,16
	klasa C	-1,21	-3,44	1,03
	chłopcy	-1,40	-3,27	0,47
	dziewczęta	1,72	0,11	3,33
gimnazjum 3	cały rocznik	2,47	0,87	4,07
	klasa A	1,18	-1,84	4,19
	klasa B	1,58	-2,00	5,15
	klasa C	4,33	2,58	6,08
	chłopcy	0,46	-2,07	2,99
	dziewczęta	4,54	2,72	6,35

można dokonywać analiz EWD dla dowolnie zdefiniowanych grup, o ile tylko posiadamy dla nich wynik sprawdzianu i egzaminu gimnazjalnego każdego ucznia. Kalkulator EWD umożliwia ich wprowadzenie, a następnie podobną do przedstawionej tu analizę.

Powtórzmy więc, jeśli przedziały ufności „zazębiają się", tzn. posiadają wspólne wartości, to nie można prawomocnie powiedzieć, że określone grupy (szkoły, klasy, dziewczynki i chłopcy) różnią się pod względem przyrostu wiedzy mierzonego EWD. Jeśli przedziały ufności nie posiadają wspólnych wartości, to można stwierdzić, że mają różne EWD. Nie należy natomiast wyciągać żadnych wniosków z porównania punktowych wartości EWD. W rzeczywistości informacja, że jedna szkoła posiada EWD równą +1, a druga -1 niewiele nam mówi. Jeśli przedział ufności EWD dla pierwszej szkoły wynosi od -1 do +3, a przedział ufności EWD drugiej szkoły wynosi od -2 do 0, to przedziały te zazębiają się i jest wysoce prawdopodobne, że prawdziwy przyrost wiedzy w obu szkołach jest podobny.

Różnice EWD między szkołami, klasami, a także ze względu na płeć, przedstawiają wykresy poniżej. Wykres (rysunek 3.7) porównuje przedziały ufności



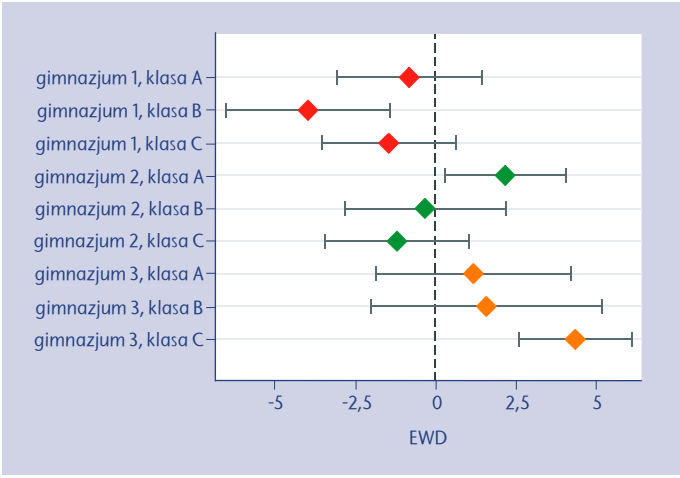
Rysunek 3.7.
Porównanie EWD przykładowych gimnazjów

EWD między gimnazjami. Widać, że nie sposób odróżnić EWD gimnazjum 2 od EWD dla pozostałych dwóch szkół. W obu wypadkach, przy porównaniu gimnazjum 1 i 2 oraz 2 i 3, przedziały ufności posiadają wspólne wartości. Można jednak z dużą pewnością stwierdzić, że gimnazjum 1 ma niższą EWD niż gimnazjum 3. Przedziały ufności EWD tych szkół nie posiadają wspólnych wartości, co więcej, są widocznie od siebie oddalone.

Na wykresie zaznaczono także wartość EWD równą 0, a więc średnią dla tego rocznika w kraju. Widać, że gimnazjum 1 ma niższą EWD niż średnia krajowa, a gimnazjum 3 wykazuje się ponadprzeciętnym przyrostem wiedzy uczniów w skali całego kraju. Gimnazjum 2 jest w istocie szkołą z przeciętnymi postępami w nauce. Wykres ten pokazuje, że liczenie EWD w oparciu o model dla całego kraju ma tę zaletę, że wyniki mogą być odniesione do krajowej przeciętnej. Jeśli porównalibyśmy tylko szkoły 2 i 3, to można by odnieść wrażenie, że gimnazjum 2 charakteryzują niskie postępy uczniów. Tymczasem jest to szkoła przeciętna, bowiem szkoła numer 3 jest placówką ponadprzeciętną.

Wykres kolejny (rysunek 3.8) przedstawia porównanie EWD między klasami. Precyzja ocen EWD jest tutaj znacznie mniejsza, bowiem liczba obserwacji jest niewielka. W praktyce, dla analiz wewnątrzszkolnych, należałoby zastosować mniej konserwatywne przedziały ufności, na przykład 90%, dzięki czemu łatwiej będzie opisać obserwowane różnice. Czytelnicy, wykonując własne analizy, mogą jednak przyjąć mniej lub bardziej bezpieczne założenia, zależnie od tego, czy porównania służą własnym celom (można stosować przedziały 90%), czy też mają być podstawą szerszej dyskusji (lepiej stosować bezpieczne przedziały 95%).

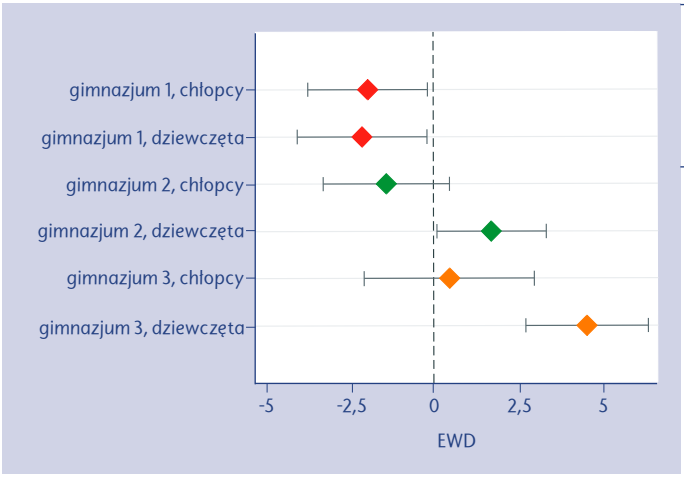
W tym przykładzie pozostaniemy przy poziomie 95%, przy którym w żadnej ze szkół nie można stwierdzić, że któraś z klas wyróżnia się niższym bądź wyższym przyrostem wiedzy. Porównując między szkołami, można jednak zauważyć,



Rysunek 3.8. Porównanie EWD klas w przykładowych gimnazjach

że na przykład klasa A w szkole 1 ma podobną jakość nauczania, jak klasa A w szkole 2, jednak klasa B w szkole 1 ma już jakość niższą. Klasy A w szkole 2 i C w szkole 3 wyróżnia wysoki przyrost wiedzy uczniów i warto by przeprowadzić lokalną analizę wyjaśniającą sukcesy w pracy z uczniami tych klas. Tego typu analizy mogą być bardzo przydatne dla poszukiwania wzorców nauczania w swojej okolicy.

Ostatni wykres (rysunek 3.9) ukazuje różnice w EWD w podziale na płeć. Jedynie w gimnazjum 1 dziewczynki uzyskały podobną punktową ocenę EWD jak chłopcy. W pozostałych szkołach punktowa EWD dziewcząt jest wyższa, choć przedziały ufności posiadają niewielką część wspólną i trudno jednoznacznie orzec, że różnice są statystycznie istotne (prawdopodobnie są na poziomie



Rysunek 3.9. Porównanie przyrostu wiedzy ze względu na płeć w przykładowych gimnazjach

90%). Nawet zakładając ostrożne 95-procentowe przedziały ufności, można jednak stwierdzić, że choć chłopców we wszystkich trzech szkołach charakteryzuje podobny przyrost wiedzy, to już dla dziewczynek widoczne są spore różnice. Dziewczynki w szkole 2 posiadają znacznie wyższą EWD niż w szkole 1, a już różnicę w EWD dziewczynek między gimnazjum 1 a 3 można określić jako edukacyjną przepaść. Z całą pewnością gimnazjum 3 wyróżnia jakość pracy z dziewczynkami, w tym względzie inne placówki mogą skorzystać z doświadczeń tej szkoły. Można jednak stwierdzić, że wszystkie trzy szkoły przeciętnie lub słabo radzą sobie w pracy z chłopcami.

Jak widać z tabeli oraz wykresów, przedziały ufności EWD dla większości szkół i klas posiadają wspólne wartości. Pokazuje to jedynie, że trudno jest porównywać jakość nauczania dla danych z jednego, najczęściej mało licznego, rocznika uczniów. Tym bardziej trudno jest wyciągać wnioski na podstawie różnic EWD między klasami, które z założenia są mało liczne w kategoriach statystycznych. Stąd warto sięgnąć po wskaźniki EWD sumujące obserwacje z kilku roczników, na przykład trzyrocznikowe EWD. W niniejszym tekście nie omawiamy wykorzystania danych z kilku roczników. Podstawowe metody analizy wskaźników EWD są jednak takie same, jak dla danych z jednego rocznika.

Oczywiście, wnioski przedstawione powyżej dla 3 przykładowych szkół powinny zostać pogłębione o szczegółowe analizy wyników uczniów, wsparte studiami jakościowymi. Rozmowy z nauczycielami czy analiza metod i programów nauczania stosowanych w różnych grupach mogą nam umożliwić zrozumienie źródeł zaobserwowanych różnic. Metoda EWD stanowi obiektywny wskaźnik różnic w przyrostach wiedzy między grupami uczniów, który może być elastycznie stosowany nawet w pojedynczych placówkach. Jednak wskaźnik ten powinien stanowić jedynie wstęp do pogłębionej analizy służącej poprawie nauczania w naszych szkołach. Zachęcamy do lektury wspomnianej broszury o wykorzystaniu EWD w analizach wewnątrzszkolnych, gdzie przygotowano więcej przystępnych przykładów analiz metodami EWD, a także do lektury innych materiałów ze stron www.ewd.edu.pl. Zachęcamy też do pogłębiania wiedzy płynącej z analiz EWD własnymi studiami.

Podsumowanie

Powyżej przedstawiono podstawy metody EWD, posługując się prostymi przykładami. Metody te mogą być stosowane na niewielką skalę, ale ich prawdziwy potencjał leży w możliwościach wykorzystania do ewaluacji polityki edukacyjnej wyników egzaminów zewnętrznych dostępnych dla niemal pełnej populacji uczniów. W rozdziale pokazano, w jaki sposób zmienia się interpretacja wyników w zależności od tego, czy analizowane są różnice między szkołami, wewnątrz szkół, czy też między rocznikami w tych samych placówkach.

Metody EWD to potężne narzędzie ewaluacji polityki edukacyjnej państwa. Szeroka dostępność danych powoduje, że w oparciu o nie może być prowadzona bieżąca ocena działań oświatowych państwa, szkół i samorządów, ugruntowana w dużych zbiorach informacji na temat osiągnięć uczniów, a przez to precyzyjna i możliwa do uogólnienia na całą populację i podgrupy. Zastosowanie metod EWD wymaga jednak świadomości ich ograniczeń. Nie każdy istotny statystycznie wynik w regresji metodą EWD należy interpretować jako zależność przyczynowo-skutkową. Określenie takich zależności wymaga przemyślanej strategii badawczej, przetestowania leżących u jej podstaw założeń, a także sprawdzenia odporności wyników na zmiany w specyfikacji modelu i w próbie. Poprawnie stosowane metody EWD stanowią znacznie bardziej wiarygodne narzędzie ewaluacji niż typowe metody korelacyjne.

Literatura

Biuletyn Badawczy CKE „Egzamin", 8/2006 *Edukacyjna wartość dodana* (dostępne pod adresem www.cke.edu.pl/).

Biuletyn Badawczy CKE „Egzamin", 14/2007 *Edukacyjna wartość dodana – cz. 2* (dostępne pod adresem www.cke.edu.pl/).

Dolata, R. (red.), (2007), *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie wyników egzaminów zewnętrznych*, wyd. CKE: Warszawa.

Dolata, R. *Edukacyjna wartość dodana, czyli jak wykorzystywać wyniki egzaminów zewnętrznych do oceny efektywności nauczania*, dostępne pod adresem codn.codn.edu.pl/struktura/prkzso/doc/czytelnia/ewd.pdf

Jakubowski, M. *Zastosowanie modeli EWD do ewaluacji polityki edukacyjnej*, wystąpienie podczas XIV Krajowej Konferencji Diagnostyki Edukacyjnej, Opole 2008, dostępne pod adresem www.wne.uw.edu.pl/mjakubowski w zakładce „publikacje".

Jakubowski, M. *Czy wydatki na gimnazja są efektywne?,* „Gospodarka Narodowa", 11-12/2007.

McCaffrey D., Lockwood J., Koretz M., Hamilton L. (2005), *Evaluating Value-Added Models for Teacher Accountability*, Rand Corporation MG-158)

OECD (2008), *Measuring Improvements in Learning Outcomes*: *Best Practices to Assess the Value-Added of Schools*, Paris).

4. Przegląd wyników badań w Polsce

Badania edukacyjne to skomplikowane pod względem metodologicznym przedsięwzięcie. Z jednej strony pojawiają się w nich trudności związane z pomiarem i jego problematyką, z drugiej strony mamy do czynienia ze skomplikowaną, hierarchiczną strukturą danych. Łatwo o wnioski opierające się na pozornych zależnościach, wynikających z pominiętych związków między zmiennymi. Badania edukacyjne często nękają problemy z trafnością wskaźników. Mamy podstawowe kłopoty z ujęciem niektórych terminów, takich jak jakość, efektywność czy nierówność. Brak mocnych teorii edukacyjnych powoduje problemy z ustaleniem porządku przyczynowości, termin ten jest nadużywany lub też zakładany *implicite*, kiedy autorzy w sposób nieuprawniony odnoszą wyniki do całej populacji lub też postulują, że określona zmienna powoduje zmiany w wynikach uczniów, choć stwierdzony związek jest w istocie jedynie zależnością zaobserwowaną w danych.

Ważne jest, aby analizując wyniki badań, zachować świadomość licznych pułapek metodologicznych, które „czyhają" na badaczy zajmujących się edukacją. Z tego względu poniższego przeglądu dokonano z perspektywy metodologicznej. Mniej ważna jest tu tematyka samych badań, a raczej metodologia i wiarygodność uzyskanych wyników. W ten sposób chcemy pokazać, jak analizować zastane rezultaty badań, na jakie kwestie zwrócić uwagę, jak ocenić wiarygodność i przydatność wyników dla własnych potrzeb. W przeglądzie tym odnosimy się do informacji zawartych w trzech pierwszych rozdziałach, gdzie opisano zagadnienia metodologii badań edukacyjnych.

4.1. Badania wyników egzaminów całej populacji uczniów

Można podzielić badania wyników egzaminów w Polsce na dwie kategorie. Pierwszą stanowią badania charakteryzujące trendy lub zróżnicowanie w całości systemu edukacyjnego. Jest to spojrzenie „z lotu ptaka" na dane z całej populacji uczniów. Drugą kategorię stanowią analizy efektywności pracy szkół, gdzie zwraca się uwagę na indywidualne lub szkolne uwarunkowania uzyskanych wyników. Trzeba przy tym przyznać, że opublikowano jak dotąd niewiele prac analizujących wyniki egzaminów w całej populacji, nie wliczając corocznych raportów Centralnej Komisji Egzaminacyjnej, które nie są materiałami *stricte* badawczymi. Jedyną znaną nam pozycją książkową, w której takie analizy są na szerszą skalę przeprowadzane, jest *Szkoła – segregacje – nierówności* napisana przez Romana Dolatę. W przeglądzie badań będziemy często odwoływać się do tej pozycji, posiłkując się także wynikami analiz publikowanymi w Biuletynach Badawczych CKE.

System edukacyjny „z lotu ptaka"

Ciekawym sposobem patrzenia na system edukacyjny jest perspektywa geograficznego zróżnicowania wyników. O taką analizę (wykorzystującą dodatkowo dane sondażowe) pokusili się Konrad Ł. Czapiewski i Przemysław Śleszyński na łamach Biuletynu Badawczego CKE „Egzamin" (11/2007). Podstawowym podziałem, do jakiego odwołali się autorzy, był podział centrum – peryferie, gdzie przez centrum rozumie się rozwinięte obszary aglomeracji miejskich, peryferiami zaś są pozostałe, mniej zurbanizowane obszary. Nietrudno się domyślić, że wyniki egzaminacyjne „centrum" okazały się znacznie wyższe niż na peryferiach. Najlepsze wyniki obserwowane były w dużych aglomeracjach miejskich, takich jak aglomeracja warszawska, poznańska czy krakowska. Najgorsze wyniki odnotowywano w Polsce północnej i zachodniej oraz na Ziemiach Odzyskanych. Z perspektywy tego podejścia badawczego wyróżniał się jedynie Śląsk – wewnątrz aglomeracji śląskiej badacze odnaleźli ponadprzeciętne zróżnicowanie wyników.

Trzeba powiedzieć, iż perspektywa geograficzna jest ciekawa, lecz ma znaczne ograniczenia. Na jej podstawie można znakomicie budować obraz systemu, jednak moc wyjaśniająca takich analiz jest znikoma. Zróżnicowanie geograficzne jest odbiciem skomplikowanych związków między społecznymi i ekonomicznymi przyczynami zróżnicowania indywidualnych wyników. Wyniki te z pewnością nie powinny służyć analizom efektywności rozmaitych rozwiązań czy wnioskowaniu o zależnościach między zmiennymi. Dywagacje na temat przyczyn regionalnego zróżnicowania wyników są jedynie domysłami, być może źródłem interesujących hipotez, ale na pewno nie pewnej wiedzy na temat tych zależności.

Tego typu nadużycie można zauważyć w innych pracach jednego z autorów (patrz J. Śleszyński i współpracownicy, *Ekonomiczne uwarunkowania wyników sprawdzianu szóstoklasistów i egzaminu gimnazjalnego przeprowadzonych w latach 2002-2004*, Biuletyn Badawczy CKE „Egzamin" 2/2004, a także inne artykuły z udziałem tego autora w tymże numerze Biuletynu), gdzie zróżnicowanie geograficzne autor „objaśnia" dziesiątkami różnych wskaźników, liczonych najczęściej na poziomie gmin. Jak sam tytuł wskazuje, autor postuluje, że korelacje między tymi wskaźnikami a wynikami mają charakter uwarunkowań. Niestety, jest to spore nadużycie, biorąc pod uwagę poziom agregacji danych oraz proste metody korelacyjne wykorzystane przez autorów. Choć w samym tekście autorzy bardzo skrupulatnie posługują się określeniem „współwystępowanie", z którego nie wynika, że dana zmienna wpływa bezpośrednio na wyniki egzaminów, to już interpretacja wyników często sugeruje, że w istocie zmiana wartości zmiennej objaśniającej powoduje zmiany wyników.

Przykładowo, analizując relację między wydatkami na płace nauczycieli a wynikami uczniów, autorzy stwierdzają: *Poziom wynagrodzeń pozytywnie <u>korelował</u> z wynikami zarówno sprawdzianu szóstoklasistów, jak i egzaminu gimnazjalnego. (…). Generalnie, wniosek o <u>wpływie</u> wysokości wynagrodzenia pracowników szkół (w tym nauczycieli) na osiągnięcia edukacyjne uczniów jest logiczny, gdyż wynika z faktu, że za lepszą płacę można zatrudnić lepszych specjalistów, a równocześnie wyższe wynagrodzenie ma charakter motywacyjny. Należy mieć zatem nadzieję, że podwyżki dla nauczycieli wdrożone w latach 2002-2004 odegrają tu bardzo pozytywną rolę.* (Podkreślenia własne). Różnica między słowami „korelował" i „wpływa" jest kluczowa. Choć autorzy piszą o współwystępowaniu, to tytuł sugeruje uwarunkowania, skąd blisko do tego, żeby wnioskować o reformach oświaty na podstawie zapewne pozornych zależności.

W wypadku płac nauczycieli wiemy, że zależą one przede wszystkim od stopnia awansu zawodowego, jak i od zasobności i szczodrości samorządów. Zarówno przeciętny stopień awansu nauczycieli, jak i budżety samorządowe są zapewne silnie powiązane z takimi charakterystykami, jak poziom wykształcenia lokalnej ludności czy poziom podatków płaconych przez ludność i lokalne przedsiębiorstwa, a te cechy są też zapewne związane z wynikami uczniów. Stąd płace nauczycieli, podobnie jak procent ludzi noszących okulary w lokalnej społeczności, mogą być pozornie związane z wynikami uczniów, równocześnie nie mając żadnego bezpośredniego wpływu na ich poziom. Zarówno płace, jak i liczba ludzi noszących okulary zależą bowiem od tych samych zmiennych, które warunkują osiągnięcia uczniów, przede wszystkim od statusu społeczno-ekonomicznego ich rodzin.

Najprostsze miary powiązania zmiennych opierające się na korelacji są niemal zawsze obarczone błędem związanym z pominięciem innych zależności między zmiennymi. Jednak jak opisano w rozdziale 2, nawet bardziej zaawansowane techniki w rodzaju regresji wielu zmiennych nie mogą stanowić pewnej

podstawy wnioskowania o bezpośrednim wpływie zmiennych. Należy zawsze o tym pamiętać, próbując wyciągać z analiz empirycznych wnioski dotyczące polityki edukacyjnej czy choćby własnej praktyki zawodowej.

Jednym z innych niedociągnięć metodologicznych częstych w podobnych badaniach jest sposób interpretacji testów statystycznych. Typowe testy zakładają pewne ryzyko popełnienia błędu, tzw. poziom istotności, który zazwyczaj przyjmowany jest na poziomie 5%. Jeśli jednak wykonywana jest większa liczba testów na zależnych od siebie zmiennych, to standardowe statystyki, według których liczone są poziomy istotności, ulegają znaczącej zmianie. Mówiąc prościej, wykonując 100 testów na tym samym zbiorze danych, na przykład na różnicach średnich między dwiema grupami, i przyjmując poziom istotności 5%, powinniśmy znaleźć istotne różnice w średnich w 5 przypadkach, nawet jeśli w rzeczywistości różnic tych nie ma, co wynika z założonego ryzyka popełnienia błędu i teorii statystycznej leżącej u podstaw.

W ten sam sposób autorzy omawianego tekstu, testując 100 zależności, znajdą zapewne 5 istotnych statystycznie. Autorzy nie zastosowali korekty testów, dzięki której zmalałoby zakładane ryzyko popełnienia błędu, przez co prezentowane przez nich rezultaty są dość mało wiarygodne, nawet w warstwie opisowej. Niestety, jest to często spotykany błąd. Dobrze, jeśli autorzy sami podają liczbę wykonanych porównań (testów). Gorzej, jeśli wykonano ich setki, a prezentowane są jedynie wyniki dla różnic potwierdzonych jako statystycznie istotne. W takiej sytuacji prezentowane są często wyniki przypadkowe, których istotność wynika jedynie z losowego charakteru obserwowanych zmiennych, a nie z rzeczywistych zależności.

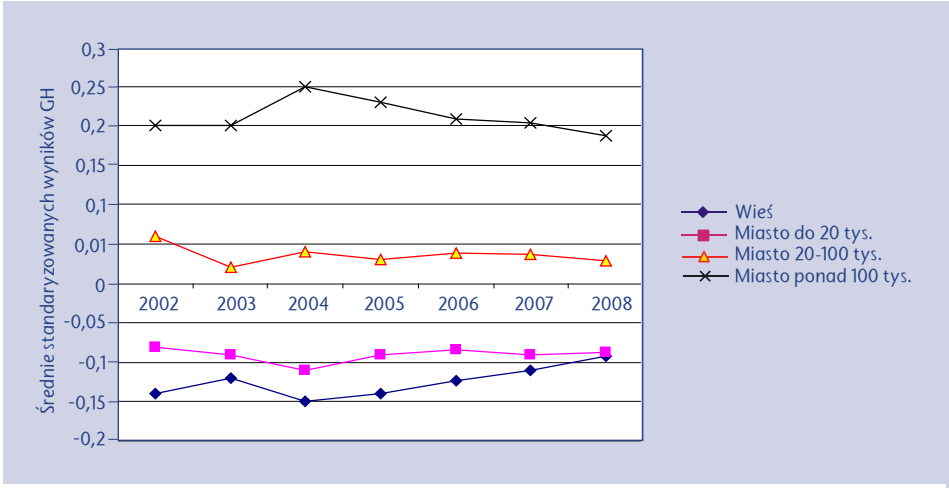
Warto również wspomnieć o czynniku, który może utrudniać stworzenie dobrego obrazu geograficznego zróżnicowania umiejętności uczniów mierzonych na podstawie wyniku egzaminacyjnego. W kolejnych latach funkcjonowania systemu egzaminacyjnego dostrzegalny był czynnik *efektu egzaminatora* na poziomie okręgowych komisji egzaminacyjnych[31]. Najprościej mówiąc, egzaminatorzy z różnych OKE oceniali prace egzaminacyjne z różną surowością. W pewnych OKE po prostu trudniej było zdobyć lepszy wynik niż w innych. Różnice te na poziomie indywidualnym nie były duże, nie miały też znaczenia dla porównań wyników uczniów czy szkół w danym regionie, jednak mają znaczenie w porównaniach całych regionów czy miast, gdzie znacząco mogą wpływać na rezultaty.

Drugim podejściem do analizy zróżnicowania regionalnego wyników egzaminacyjnych jest spojrzenie na rezultaty uczniów ze względu na lokalizację szkoły w podziale na 4 kategorie: miasto powyżej 100 tysięcy mieszkańców, miasto od 20 do 100 tysięcy, miasto poniżej 20 tysięcy oraz obszary wiejskie. Prezentacje

[31] Dubiecka, A., Szaleniec, H., Węziak, D. (2006), *Efekt egzaminatora w egzaminach zewnętrznych* w: *O wyższą jakość egzaminów szkolnych*, cz. I, *Etyka egzaminacyjna i zagadnienia ogólne*, red. B. Niemierko, M. K. Szmigel, Lublin.

wyników w takim podziale można znaleźć w corocznie publikowanych raportach Centralnej Komisji Egzaminacyjnej, a szczegółową analizę z kilku ostatnich lat zaprezentował Roman Dolata (2008) we wspominanej już książce. Najwyższe wyniki egzaminacyjne, chodzi tutaj o sprawdzian po szkole podstawowej oraz egzamin gimnazjalny, można zaobserwować w dużych miastach powyżej 100 tysięcy mieszkańców, niższe w miastach od 20 do 100 tysięcy, jeszcze niższe w miastach poniżej 20 tysięcy, a najniższe na wsi.

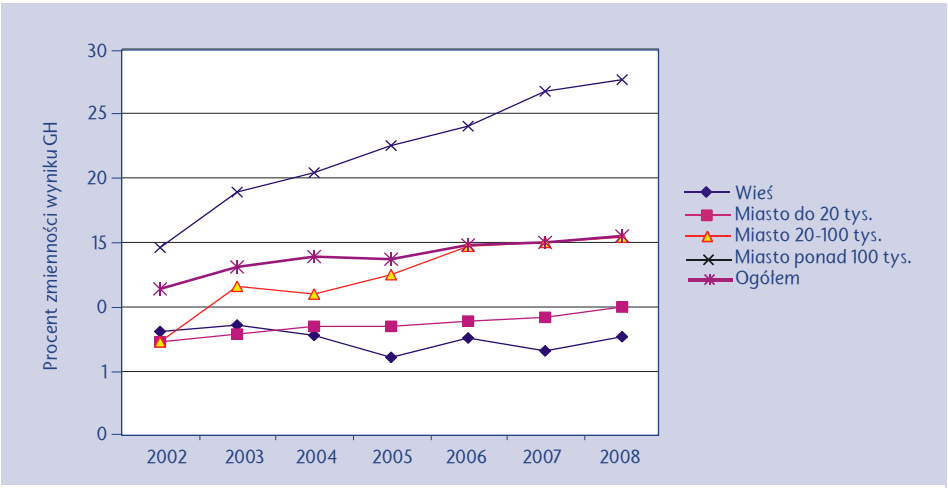
Nie może to budzić wielkiego zdziwienia. Ciekawa jest natomiast dynamika zmian dystansów między tymi czterema kategoriami lokalizacji (patrz Dolata 2008). Autor posłużył się tu wynikami wystandaryzowanymi tak, aby średnia w każdym roku wyniosła 0, a odchylenie standardowe 1, dzięki czemu wyniki są porównywalne między latami[32]. Jeżeli chodzi o wyniki szkół podstawowych, to różnice między osiąganymi wynikami na przestrzeni lat praktycznie nie zmieniają się. Znaczną przewagę nad pozostałymi kategoriami mają duże miasta, najsłabiej wypadają obszary wiejskie, trochę lepiej małe miasta i miasta od 20 do 100 tysięcy mieszkańców. Bardzo podobny obraz rysują wyniki egzaminu gimnazjalnego, ale w tym wypadku warto podkreślić, iż różnica między wsią a małym miastem (do 20 tysięcy) jest praktycznie niedostrzegalna. Z tym, że w wypadku części humanistycznej egzaminu gimnazjalnego można zaobserwować lekki



Rysunek 4.1. Standaryzowane wyniki szkół według wielkości miejscowości, na podstawie: Dolata (2008)

[32] Tego typu standaryzacja powoduje, że o ile relatywne różnice między grupami zachowują sens, więc analizy autora są jak najbardziej prawidłowe, to już niemożliwe jest oszacowanie ogólnego trendu w wynikach polskich uczniów. Według nas nie jest to jednak możliwe ani w oparciu o oryginalne wyniki egzaminów, ani też dowolne ich przekształcenie, ze względu na brak odniesienia zadań testowych do jakiejkolwiek wspólnej skali pomiarowej.

spis treści



Rysunek 4.2. Międzyszkolne zróżnicowanie wyników egzaminu w części humanistycznej w zależności od lokalizacji gimnazjum, na podstawie: Dolata (2008)

trend spadkowy różnicy między wynikami w tych czterech kategoriach. W wypadku części matematyczno-przyrodniczej egzaminu jest natomiast odwrotnie – dostrzegalny jest minimalny trend powiększający różnicę wyników między tymi kategoriami.

Dolata (2008) przedstawia również ciekawe analizy zróżnicowania międzyszkolnego. Pokazuje, jak na przestrzeni lat zróżnicowanie między szkołami zmienia się w omawianych 4 kategoriach. Otóż szkoły najbardziej różnią się między sobą w dużych miastach, w pozostałych kategoriach (wieś; miasto do 20 tys.; miasto 20 do 100 tys.) zróżnicowanie międzyszkolne jest znacznie mniejsze. Dla wsi i mniejszych miast zróżnicowanie międzyszkolne jest stosunkowo podobne. Niepokoi natomiast silny trend wskazujący na dynamiczne różnicowanie się szkół na terenach dużych miast. Na przykład w latach 2002-2008 zróżnicowanie między szkołami w dużych miastach wzrosło prawie dwukrotnie.

W książce Dolaty można znaleźć również ciekawe wnioski wyprowadzone na podstawie analiz wyników egzaminacyjnych dotyczących segregacji wewnątrzszkolnych. Przytoczmy jego najważniejsze wnioski (2008: 134-135):

- *Przejście uczniów ze szkół podstawowych do gimnazjów wiąże się w segmencie placówek wiejskich ze spadkiem zróżnicowania międzyszkolnego. Oznacza to, że w wypadku placówek wiejskich przejście uczniów ze szkoły podstawowej do gimnazjum oznacza „lepsze wymieszanie" dzieci o równym poziomie uprzednich osiągnięć i większe prawdopodobieństwo spotkania się w jednej szkole ucznia słabego i dobrego.*
- *Przejście uczniów ze szkół podstawowych do gimnazjów wiąże się w segmencie szkół zlokalizowanych w dużych miastach ze wzrostem zróżnicowania*

międzyszkolnego. Oznacza to, że w wypadku gimnazjów w dużych miastach przejście uczniów ze szkoły podstawowej do gimnazjum oznacza nasilenie segregacji dzieci ze względu na poziom uprzednich osiągnięć szkolnych i niższe prawdopodobieństwo spotkania się w jednej szkole ucznia słabego i dobrego.

- *W gimnazjach – szczególnie miejskich – uczniowie poddawani są znaczącej segregacji wewnątrzszkolnej ze względu na uprzednie osiągnięcia.*
- *W segmencie gimnazjów zlokalizowanych w średnich i dużych miastach trzy lata nauki w gimnazjum oznaczają – w stosunku do poziomu segregacji ze względu na uprzednie osiągnięcia szkolne uczniów na progu gimnazjum – wzrost międzyszkolnego i międzyoddziałowego zróżnicowania wyników nauczania. W wypadku średnich miast łącznie zróżnicowanie międzyszkolne i międzyoddziałowe jest o blisko 5-6 pp silniejsze niż poziom segregacji ze względu na uprzednie osiągnięcia, w segmencie gimnazjów zlokalizowanych w dużych miastach efekt różnicowania przewyższa poziom segregacji o 7-10 pp.*

Trzeba podkreślić, że wnioski te mają charakter opisowy. Autor stwierdza jedynie zależności istniejące w danych, oddzielając warstwę opisową od części analizującej politykę edukacyjną w oparciu o te dane, gdzie więcej jest stwierdzeń mających charakter postulatów niż opisu danych. Taki sposób postępowania gwarantuje, że czytelnicy nie są „zwodzeni" pozornymi zależnościami przedstawianymi jako związki przyczynowo-skutkowe. Krótko mówiąc, tekst ten stanowi dobry przykład analizy, gdzie oddzielono to, co wiemy, od tego, co uważamy. Nie jest to w badaniach edukacyjnych łatwe, ale analizując ich wyniki, zawsze powinniśmy podjąć samodzielną próbę oddzielenia stwierdzeń tego rodzaju. Badania opisane powyżej, podsumowujące dane dla całego kraju, mają najczęściej charakter opisowy i nie powinny być inaczej wykorzystywane.

Badania efektywności szkół

Przykład analiz w oparciu o całą populację uczniów stanowią inne badania wykorzystujące wyniki egzaminów zewnętrznych, ale mające mniej opisowy, a bardziej spekulatywny charakter. W analizach tych najczęściej próbuje się wyjaśniać różnice w wynikach uczniów lub efektywności szkół mierzonej wskaźnikami EWD, opierając się na zmiennych konstruowanych z baz CKE lub OKE. Z oczywistych względów możliwości takich badań są ograniczone, bowiem bazy te zawierają, oprócz wyniku i oznaczeń szkoły oraz klasy, jedynie informacje na temat płci, dysleksji, statusu olimpijczyka oraz datę urodzenia ucznia.

Pojawił się także nurt badań, w którym dane egzaminacyjne łączone są z danymi z innych baz w Polsce. Najwięcej prac opiera się tu na danych GUS, w szczególności z Banku Danych Regionalnych GUS, gdzie gromadzone są wskaźniki na poziomie gmin, powiatów, województw oraz regionów i podregionów według

nomenklatury GUS. Przykład tego rodzaju badania opisano już powyżej (prace J. Śleszyńskiego i współpracowników), w tej części jednak omawiamy badania, których celem jest zrozumienie głębszych zależności związanych z efektywnością szkół, a nie jedynie opis związków między zmiennymi.

Nie doczekaliśmy się jeszcze bardziej dogłębnych analiz w oparciu o dane z Systemu Informacji Oświatowej oraz wyniki egzaminacyjne. Jest to spowodowane niedostosowaniem obydwu systemów i związanymi z tym trudnościami z łączeniem danych. Należy oczekiwać, że wraz z upływem czasu ulegnie to zmianie i baza SIO, jako najbogatsze źródło informacji o organizacji szkół, zacznie być intensywniej wykorzystywana do badań edukacyjnych.

Jak już wspomniano i dokładnie opisano w rozdziale 3, wykorzystując wskaźnik edukacyjnej wartości dodanej, a raczej modele regresyjne odpowiadające metodzie liczenia EWD, badacze są w stanie mierzyć relatywny przyrost wiedzy uczniów, zarówno indywidualnie, jak i w grupach – przede wszystkim szkołach, ale i na przykład w gminach, co daje szerokie możliwości analiz różnego typu. Również w tym wypadku można powołać się na publikację Dolaty (2008), w której – wykorzystując wyniki egzaminacyjne oraz metodologię edukacyjnej wartości dodanej – autor analizuje wpływ składu klasy na efektywność nauczania w gimnazjum. Wyniki sugerują, iż możemy z pewną dozą ostrożności stwierdzić, iż słabi uczniowie trafiający do klas o korzystnym składzie mają wyższe prawdopodobieństwo uzyskania wysokich wyników, niż inni uczniowie o podobnym potencjale, którzy trafili do klas słabszych. Ostrożność jest tu wskazana, bowiem, jak wykazano w dziesiątkach prac opierających się na danych z innych krajów, skład osobowy klas lub szkół zależy w dużym stopniu od zmiennych trudnych do kontrolowania w badaniach statystycznych, na przykład popularności szkoły czy świadomych wyborów rodziców, uczniów, nauczycieli lub dyrektorów, które mają wpływ na to, jak uczniowie są dzieleni na klasy lub nawet do jakich szkół uczęszczają.

Innym przykładem badania opierającego się na metodzie EWD i odnoszącego dane egzaminacyjne do wskaźników spoza systemu edukacji jest badanie wpływu wydatków na ucznia gimnazjum na przyrost wiedzy uczniów[33]. Badanie opiera się na modelach wielopoziomowych, w tym modelach edukacyjnej wartości dodanej. Wykorzystano przy tym dane dla trzech roczników uczniów, gdzie wyniki egzaminacyjne uczniów zrównano do skali wyników z 2007 roku. Uwzględnienie danych z kilku roczników umożliwiło zwiększenie precyzji oszacowań a także skorzystanie nie tylko z wariancji wydatków i wyników między gminami, ale i między tymi samymi szkołami między latami. Zmiany między latami wykorzystano też do oceny efektywności dla szkół podstawowych, gdzie co prawda zastosowanie modelu

[33] Por. Jakubowski, M. (2008), *Teaching quality and decentralization* (dostępne na www.wne.uw.edu.pl/mjakubowski w zakładce „publikacje"). Pierwsze wyniki opublikowane zostały w polskim artykule *Efektywność wydatków na gimnazja* („Gospodarka Narodowa", 11-12/2007).

EWD nie jest możliwe, bowiem brakuje pomiaru „na wejściu", jednak wykorzystanie wariancji między latami zwiększa wiarygodność wyników.

Skorzystanie z danych dla trzech roczników powoduje, że badanie to prezentuje wyniki z ogromną precyzją, co jest główną zaletą pracy z danymi dla całej populacji. W analizach wykorzystano wyniki ponad miliona 300 tys. uczniów z ponad 6 tys. gimnazjów położonych w 2466 gminach (liczba gmin w Polsce jest nieco większa, jednak dla części gmin niemożliwe było połączenie danych CKE i GUS lub też danych między latami, ze względu na zmiany w kodach i jednostkach administracyjnych).

Rezultaty tego badania okazały się podobne do uzyskanych na danych jednorocznych. Nie stwierdzono, żeby wydatki wpływały znacząco czy to na przeciętne wyniki uczniów, czy to na wyrównywanie wyników względem osiągnięć u progu gimnazjum. Badanie dostarcza też jednak ciekawych spostrzeżeń *à propos* innych wskaźników. Okazuje się, że w każdym wypadku, nawet regresji EWD, dochody na mieszkańca gminy (odzwierciedlające w przybliżeniu zamożność mieszkańców), wielkość szkoły, a także średni wynik uczniów gimnazjum ze sprawdzianu, są pozytywnie i silnie związane z wynikami uczniów lub ich przyrostem w czasie nauki w gimnazjum.

Pokazuje to, że wszelkie korelacje wyników z innymi zmiennymi mogą być pozorne, o ile zmienne te skorelowane są z dochodami lub wielkością miejscowości, a przez to z wykształceniem i zasobami kulturowymi społeczności lokalnej czy też z popularnością placówki. Wyniki te dobitnie pokazują, że proste miary korelacyjne dają mało wiarygodne rezultaty, o ile oceniać chcemy wpływ wybranych czynników na wyniki. Badanie to zawiera także wyniki dotyczące pozytywnego związku z partycypacją w edukacji przedszkolnej oraz negatywnego efektu zwiększania zróżnicowania wyników uczniów, jednak rezultaty te należy traktować z ostrożnością z podobnych względów, jak wymienione powyżej.

Powstało też sporo prac, które nie wykorzystują metod EWD, jednak stosują metody regresji liniowej celem oceny zależności między wskaźnikami na poziomie gmin lub szkół a wynikami uczniów, po uwzględnieniu wpływu wybranych zmiennych kontrolnych. W dwóch nieopublikowanych dotąd pracach Mikołaj Herbst oraz Jan Herczyński wykorzystali dane egzaminacyjne do oceny zależności między wynikami uczniów oraz wybranymi wskaźnikami[34].

W pracy *Is Large More Effective than Small Is Beautiful? Size and Performance of Primary Schools in Poland* stwierdzono ciekawy, nieliniowy związek między wynikami uczniów szkół podstawowych a wielkością szkoły. Do pewnego momentu wraz z wielkością szkoły rosną wyniki uczniów, jednak od pewnego momentu wyniki te zaczynają spadać. Choć trudno uznać, że wpływ wielkości szkoły na wyniki uczniów może być w pełni wiarygodny sposób oceniony zastosowanymi w tym badaniu metodami, a wręcz doświadczenia dziesiątków badań

[34] Prace te dostępne są na stronie www.esep.pl

spis treści

tego rodzaju, głównie z USA, pokazują, że niezbędne są tu metody quasi-eksperymentalne, to wyniki te są zgodne z intuicją i w pewnym stopniu są zgodne także z wynikami analiz nad wpływem wielkości klas, o których piszemy poniżej.

W innej pracy ci sami autorzy zastanawiają się nad wpływem możliwości wyboru między gimnazjami na jakość ich pracy (*School Choice and Student Achievement*). Badanie to stanowi ciekawą próbę analizy kwestii konkurencji między szkołami, a jego główną zaletą jest odniesienie tego typu dyskusji, tak popularnej w krajach zachodnich, do polskiego kontekstu. Autorzy znajdują słaby pozytywny związek między możliwością wyboru szkoły a wynikami uczniów, równocześnie jednak stwierdzają też pozytywny związek między możliwością wyboru a poziomem segregacji międzyszkolnej. Zależności te mogą być jednak także efektem powiązania wyników uczniów ze zmiennymi nieuwzględnionymi w modelu, przede wszystkim z jakością nauczania, ale i popularnością placówek. Brakiem tej pracy, jak i wielu innych badań opierających się na jednorazowych testach umiejętności uczniów, jest zrównanie wyników uczniów z jakością pracy szkoły. Analizy EWD pokazują, że zależności między zmiennymi znacząco zmieniają się po uwzględnieniu wcześniejszych wyników uczniów, a surowy wynik egzaminu odzwierciedla bardziej środowisko społeczno-ekonomiczne i skład uczniowski szkół niż jakość pracy placówek.

W pracy zbiorowej autorstwa Pawła Swianiewicza, Mikołaja Herbsta i Wojciecha Marchlewskiego *Finansowanie i realizowanie zadań oświatowych na obszarach wiejskich* (raport na zlecenie Związku Gmin Wiejskich RP, 2005) drobiazgowo omówiono kwestie związane z finansowaniem i dostępnością edukacji, w szczególności na terenach wiejskich, ale i z odniesieniem do innego typu gmin. Praca ta zawiera też analizy wyników egzaminów zewnętrznych metodami regresji liniowej. Nie są to analizy EWD, bowiem autorzy nie korzystali z danych z połączonymi wynikami sprawdzianu i egzaminu gimnazjalnego, a dodatkowo zmienne zagregowano do średniego poziomu gmin, co zmniejsza znacząco precyzję w porównaniu z analizami na poziomie uczniów. Niemniej jednak bogactwo materiału i drobiazgowość analizy powodują, że jest to praca godna polecenia, stanowiąca ciekawy materiał opisujący zależności na poziomie gmin między wynikami uczniów a organizacją i finansowaniem ze strony samorządowej.

W oparciu o dane egzaminacyjne możliwe jest także przeprowadzenie analiz zbliżających nas do poznania prawdziwych zależności przyczynowo-skutkowych. W Polsce znana nam jest jedna praca tego rodzaju wykorzystująca wyniki egzaminów zewnętrznych. Maciej Jakubowski oraz Paweł Sakowski zastosowali do polskich danych metody quasi-eksperymentalne, wykorzystując dane o wynikach klas w szkołach podstawowych z okręgu OKE Warszawa do oceny wpływu wielkości klasy na wyniki uczniów (*Quasi-experimental estimates of class size effect in primary schools in Poland*, „International Journal of Educational Research", vol. 45(3), s. 202-215, a także Jakubowski, M. *Małe, średnie*

czy też duże? O wpływie liczby uczniów w klasie na jakość nauczania, Biuletyn Badawczy CKE „Egzamin", 2/2004).

W pracy tej zastosowano do polskich danych błyskotliwe pomysły ze znanej pracy Caroline Hoxby[35] oraz jednego z najczęściej cytowanych badań edukacyjnych autorstwa Joshua Angrista i Victora Lavy'ego[36]. Założeniem każdej z metod jest wykorzystanie w wielkościach klas wariancji, którą można uznać za w pełni losową, niezależną zarówno od obserwowalnych czynników, takich jak pochodzenie społeczne uczniów czy poziom ich umiejętności, jak i nieobserwowalnych, takich jak wybory rodziców i dyrektorów szkół czy też trudne do zmierzenia zdolności uczniów lub atmosfera w klasie. Jeżeli jesteśmy w stanie określić losową wariancję w wielkościach klas, oddzielić ją od wariancji nielosowej, a następnie powiązać ze zmianami w wynikach uczniów, to uzyskany wynik świadczyć będzie o zależności przyczynowo-skutkowej między wynikami a liczbą uczniów w klasie.

W badaniu dla danych polskich zastosowano dwie quasi-eksperymentalne strategie, których porównanie daje dodatkową gwarancję, że obserwowane efekty mają naturę przyczynowo-skutkową. Po pierwsze, oszacowano efekt zmiany wielkości klas w tych samych placówkach między latami. Dzięki temu wykluczono wpływ na oszacowane zależności zmiennych różnicujących wyniki między szkołami. Dodatkowo zastosowano metodę zmiennych instrumentalnych[37], gdzie średnia wielkość klasy w szkole posłużyła jako „instrument" dla faktycznej wielkości klasy, dzięki czemu wykluczono wariancję w wielkości klas powiązaną z procesami przydzielania uczniów o różnych poziomach umiejętności do klas różnej wielkości. Inaczej mówiąc, wyłączono w ten sposób wpływ procesów segregacji wewnątrzszkolnej na szacowane zależności.

---

[35] Hoxby, C. M. (2000), *The Effects Of Class Size On Student Achievement*: *New Evidence From Population Variation*, „The Quarterly Journal of Economics", vol. 115(4), s. 1239-1285.

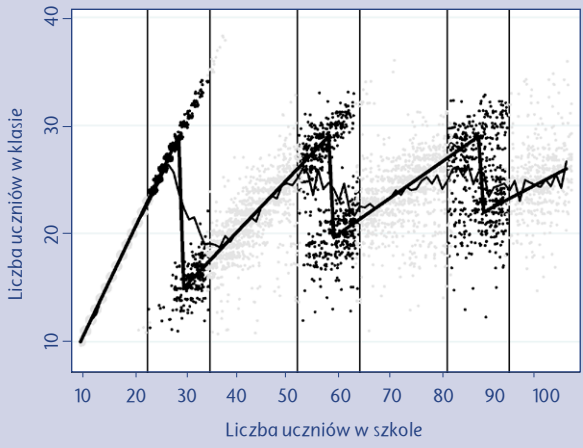
[36] Angrist, J. D., Lavy, V. (1999), *Using Maimonides' Rule To Estimate The Effect Of Class Size On Scholastic Achievement*, „The Quarterly Journal of Economics", MIT Press, vol. 114(2), s. 533-575.

[37] Metoda zmiennych instrumentalnych opiera się na założeniu, że posiadamy lub jesteśmy w stanie skonstruować zmienną, która powiązana jest ze zmienną uwzględnioną w modelu regresji, jednak ma charakter egzogeniczny, tzn. jej wariancja ma charakter losowy. Inaczej mówiąc, zmienna ta wpływa na zmienną zależną (tutaj wynik ucznia) jedynie poprzez zmienną niezależną (tutaj wielkość klasy), dla której jest instrumentem. Intuicyjnie, w metodzie tej zmienna instrumentalna stanowi sposób na oddzielenie nielosowej i losowej wariancji w interesującej nas zmiennej zależnej (tutaj wielkości klasy) i odniesienie jej do wariancji w zmiennej zależnej (tutaj wyniki uczniów), a więc stanowi metodę quasi-eksperymentalną. Także w tej metodzie poszukujemy losowej wariancji w objaśnianej zmiennej, podobnie jak w eksperymencie przypisujemy losowo obserwacje do grupy poddanej działaniu bodźca i do grupy kontrolnej, zapewniając w ten sposób podobieństwo obu grup pod każdym względem oprócz poddania działaniu bodźca. Podejście to stanowi obecnie dominującą metodę określania zależności przyczynowo-skutkowych w ekonometrii (por. Angrist, J., Krueger, A. (2001), *Instrumental variables and the search for identification: From supply and demand to natural experiments*, „Journal of Economic Perspectives", vol. 15(4), s. 69-85).

Taka strategia ma charakter quasi-eksperymentu, o ile zmiany w wielkościach klas między rocznikami w danej szkole mają charakter losowy (por. rozdział 2). Można spodziewać się, że zmiany te związane są przede wszystkim z różnicami w liczbie urodzeń w danej okolicy czy rejonie. Z pewnością jednak na zmiany te mają wpływ migracje oraz zmiany w popularności placówek, które przy tej strategii badawczej mogą zniekształcać wyniki.

Druga zastosowana metoda ma charakter jedynie quasi-eksperymentalny. W pierwszym kroku określono dla polskich danych próg wielkości klas, dla którego szkoły najczęściej tworzą nowe oddziały. Okazuje się, że choć nie istnieją przepisy wyraźnie określające wielkość klasy w Polsce, to jednak wymogi nakładane przez kuratoria oświaty oraz przekonanie dyrektorów i samorządów o uciążliwości bardzo dużych klas powodują, że klasy osiągają przeciętnie maksymalnie 29 uczniów[38]. Inaczej mówiąc, empirycznie stwierdzono, że nowe klasy tworzone są, gdy liczebność roczników osiąga około 29, 58, 87 uczniów.

Widoczne jest to na wykresie poniżej (rysunek 4.3), na którym zaznaczono faktyczne wielkości klas (oś Y) dla odpowiadającej im liczby uczniów w danym roczniku (oś X). Dla liczby uczniów w roczniku w okolicach 29, 58, 87 istnieją zarówno szkoły, które już stworzyły nowe klasy, jak i takie, które jeszcze ich nie stworzyły. Oczywiście szkoły te różnią się pod względem średniej wielkości klasy, co wykorzystano w analizie. O ile przyjmiemy, że szkoły te nie różnią się pod innymi względami, to porównanie wyników szkół, które utworzyły nowe klasy



Rysunek 4.3. Liczba uczniów w roczniku w szkole a liczba uczniów w klasach w obszarach nieciągłości (ciemniejsze punkty) oraz poza obszarami nieciągłości (punkty jasnoszare)

[38] W oryginalnej pracy Angrista i Lavy'ego wykorzystano przepis prawny istniejący w Izraelu, nazwany zasadą Majmonidesa, określający maksymalną wielkość klas na poziomie 40 uczniów.

spis treści

i które jeszcze ich nie utworzyły, dostarcza quasi-eksperymentalnych oszacowań wpływu wielkości klasy na wyniki uczniów.

Tego rodzaju oszacowania uzyskano, analizując jedynie szkoły w tzw. przedziałach nieciągłości, które dość arbitralnie określono na 24-34, 53-63 i 72-92 uczniów w roczniku, a więc w okolicach plus minus 5 uczniów od 29 i jego wielokrotności. Obliczenia powtórzono jednak dla innych „progów" tworzenia nowych klas oraz rozmiarów obszarów „nieciągłości", uzyskując zbliżone rezultaty. W badaniu zastosowano metodę zmiennych instrumentalnych, gdzie wykluczono wariancję w wielkości klas niezwiązaną z „zasadą" tworzenia nowych klas po przekroczeniu pewnego progu (wartości tej zmiennej pokazano na wykresie jako linię ciągłą). Jako kluczową zmienną kontrolną zastosowano liczbę uczniów w danym roczniku w danej szkole, dzięki czemu porównania dotyczą jedynie szkół o podobnej wielkości, wykluczając wszelkie różnice z tym związane. Tego typu podejście, bardzo popularne w ostatnich latach, nazywane jest *regression discontinuity design*, czyli w wolnym tłumaczeniu „regresją nieciągłości". Intuicyjnie, metoda ta wykorzystuje gwałtowne skoki w wartościach zmiennej objaśniającej (niezależnej), które o ile mają naturę losową, to można je odnieść do zmian w wartościach zmiennej objaśnianej (zależnej), oceniając w ten sposób związek przyczynowo-skutkowy między tymi zmiennymi. Ograniczeniem tego podejścia jest przede wszystkim rzadkość tego typu rzeczywistych sytuacji. Dodatkowo efekty można określić tylko dla obszarów nieciągłości. Zmiany wartości zmiennych poza tymi obszarami nie są bowiem brane pod uwagę, co zazwyczaj wyklucza możliwość oszacowania związku dla dużej części populacji.

Obydwie metody skonfrontowano z wynikami uzyskanymi dla typowych metod regresyjnych, które wskazywały, zgodnie z przypuszczeniami, na pozytywną zależność między liczbą uczniów w klasie a ich wynikami. Po zastosowaniu obydwu metod quasi-eksperymentalnych uzyskano natomiast negatywne oceny tej zależności. Uzyskane wyniki autorzy zinterpretowali, odnosząc je do rzeczywistej różnicy w wynikach uczniów miast i wsi. W ten sposób stwierdzają, że dalsze zwiększenie liczebności klas w szkołach wiejskich, chociażby w wyniku popularnej swojego czasu racjonalizacji sieci szkolnej, do poziomu klas w szkołach miejskich (dla danych z roku 2002 oznaczałoby to zmianę z 18 do 24 uczniów w klasie) spowodowałoby zwiększenie różnicy w wynikach między miastem i wsią o dalsze kilkanaście lub nawet kilkadziesiąt procent. Choć obliczenia te można uznać za opierające się na dość abstrakcyjnych założeniach, to odniesienie efektów stwierdzonych w badaniu do rzeczywistych różnic uzmysławia wpływ rozwiązań przyjmowanych w systemie edukacji na rzeczywiste zróżnicowanie wyników.

Opisane powyżej badanie quasi-eksperymentalne wyraźnie pokazuje potencjalne problemy związane ze stosowaniem prostych technik statystycznych w celu stwierdzenia zależności pomiędzy nakładami edukacyjnymi

a osiągnięciami uczniów. W badaniu stwierdzono, że zwykłe regresje błędnie wskazywały na pozytywną zależność pomiędzy wielkością klasy a wynikami uczniów. Zastosowanie bardziej wiarygodnych technik quasi-eksperymentalnych spowodowało odwrócenie znaku tej zależności. W badaniach edukacyjnych, nawet prowadzonych w oparciu o doskonałe bazy danych i precyzyjny pomiar umiejętności, często będziemy mieli do czynienia z sytuacją, gdy różne metody dają niespójne wyniki. Zastosowanie odpowiednich technik statystycznych lub też wykorzystanie źródeł losowych zmian w wartościach analizowanej zmiennej otwiera możliwość uzyskania bardziej spójnych, a przez to bardziej wiarygodnych wyników. Warto szukać tego typu analiz, a także mieć nadzieję, że z biegiem czasu w Polsce pojawi się coraz więcej badań tego typu.

4.2. Badania wyników egzaminów zewnętrznych na próbach reprezentatywnych

Z bardzo prostej przyczyny badania efektywności nauczania, szczególnie badania EWD, są trudne do zrealizowania na całej populacji, nawet gdy mamy dostęp do baz z wynikami egzaminów zewnętrznych. Bazy te nie zawierają zazwyczaj interesujących nas zmiennych, a jedynie wynik ucznia, płeć, ewentualnie datę urodzenia czy informację o dysleksji. Potrzebne są zatem dodatkowe badania terenowe, gdzie do wyników egzaminów zewnętrznych dołączane są informacje z kwestionariuszy zrealizowanych na próbie uczniów, rodziców lub nauczycieli. Przykłady takich badań prezentujemy poniżej.

Badania tego typu mogą być realizowane na małych próbach, w określonym regionie lub choćby w jednej gminie. Jednak w każdym wypadku ważne jest zachowanie zasady losowego doboru próby. Niezależnie od zasięgu badania, a także wielkości próby, jednostki do badania powinniśmy dobierać losowo, zachowując reprezentatywność dla określonej w badaniu populacji. Ewentualnie możemy analizować dane dla całej populacji, na przykład wszystkich 15-latków w gminie, odwołując się do wspominanej wcześniej teorii superpopulacji, dbając jednak, aby wszyscy uczniowie wzięli udział w badaniu. Dowolność udziału lub duża liczba odmów powoduje, że nasze badanie przestaje być reprezentatywne, a wnioski można będzie łatwo podważyć.

Badania edukacyjne na reprezentatywnych próbach są kosztowne, co związane jest ze specyficznym sposobem losowania. Mowa tutaj o zespołowym i wielostopniowym doborze próby. Gdy chcemy dowiedzieć się czegoś o uczniach, najprostszy schemat zespołowy polega na wylosowaniu danej liczby szkół z listy wszystkich szkół istniejących na interesującym nas obszarze. W wylosowanych szkołach przeprowadza się badania ankietowe nie tylko wśród uczniów, lecz także wśród ich rodziców i nauczycieli.

Losowanie wielostopniowe jest rozszerzeniem badania zespołowego. W tym wypadku losujemy zespoły – na przykład szkoły, ankietujemy nie wszystkich uczniów, ale znowu dobraną losowo próbkę – losujemy pewną liczbę uczniów z każdej szkoły. Odpowiednie kalkulacje pokazują, że liczebność ogólna próby dla populacji uczniów w Polsce w doborze zespołowym lub wielopoziomowym powinna być określona na poziomie około 3-6 tysięcy uczniów w nie mniej niż 80-100 szkołach, przy czym dokładnie można to określić, jeśli znane są szczegóły badania. Nie licząc badań międzynarodowych, takich jak PISA i PIRLS, o których piszemy w kolejnym podrozdziale, stosunkowo niewiele badań w Polsce mogło poszczycić się taką liczebnością próby. Poniżej dokonujemy przeglądu wybranych wyników tych badań.

Jednym z badań spełniających minimalne wymagania związane z doborem próby było badanie zrealizowane na zamówienie Centralnej Komisji Egzaminacyjnej przez firmę Pentor Research International w maju i czerwcu 2006 roku. Przeprowadzone zostało kompleksowe badanie zogniskowane na społecznych, psychologicznych, ekonomicznych czynnikach wpływających na zróżnicowanie wyników egzaminów zewnętrznych. Projekt badawczy zakładał analizy wyników uczniów zdających egzamin gimnazjalny. Zwieńczeniem tych prac była seria artykułów, które ukazały się w trzech kolejnych numerach Biuletynu Badawczego CKE „Egzamin": 11/2007; 12/2007; 13/2007. Bliższe informacje na temat tego badania można znaleźć też w rozdziale 5, gdzie opisano publicznie dostępną bazę danych z tego badania. Poniżej przyjrzyjmy się najciekawszym wynikom prezentowanym w przytaczanych numerach Biuletynu Badawczego CKE „Egzamin".

Przedstawione przez Macieja Jakubowskiego wyniki badań świadczą o mocnym powiązaniu wyników egzaminacyjnych uczniów z czynnikami ekonomicznymi, dochodami gospodarstw domowych, z sytuacją zawodową rodziców i z poziomem ich wykształcenia (*Wpływ czynników ekonomicznych na wyniki egzaminów zewnętrznych*, Biuletyn Badawczy CKE „Egzamin", 11/2007). Wyniki badania sugerują, że wyższe dochody i wykształcenie rodziców ucznia powiązane są z lepszymi warunkami materialnymi kształcenia (wyposażeniem w pomoce naukowe, komputer, dostęp do Internetu czy do dóbr kultury). Czynniki te związane są nie tylko z poziomem osiągnięć uczniów, lecz także z przyrostem wiedzy w gimnazjum. Uczniowie z zamożniejszych i lepiej wykształconych rodzin osiągają lepsze wyniki z egzaminu gimnazjalnego niż ich koledzy o podobnym wyniku ze sprawdzianu.

Jest to bardzo ważna obserwacja, która pokazuje, że polskie gimnazja w niewielkim stopniu niwelują dysproporcje w osiągnięciach uczniów, wynikające z różnic dotyczących pochodzenia społecznego. Poza tym analizy pokazują, że dochody związane są z poziomem oraz przyrostem wiedzy uczniów nawet po wyłączeniu wpływu wykształcenia. Kwestią otwartą pozostaje jednak sposób, w jaki zasobność materialna wpływa na wyniki. Przedstawione analizy mają

charakter wstępny i jedynie sugerują dalsze kierunki poszukiwań, celem bliższego określenia relacji między dochodami a wynikami uczniów.

Analizy Renaty Siemieńskiej oraz Ireneusza Białeckiego (*Wpływ rodziny i grup rówieśniczych na wyniki egzaminów zewnętrznych*, Biuletyn Badawczy CKE „Egzamin", 12/2007) również potwierdzają to, iż wyższe dochody i wykształcenie rodziców są silnie powiązane z osiągnięciami egzaminacyjnymi uczniów. Ciekawostką jest, że nie stwierdzono istnienia zależności między długością spędzania czasu z rodzicami a sukcesami egzaminacyjnymi. Autorzy stwierdzają natomiast, że w rodzinach, w których dzieci uważają, iż są traktowane w sposób partnerski i cieszą się pewnym zakresem swobody, uczniowie osiągają lepsze wyniki w testach. Znów jest to tylko zależność statystyczna, ale niewątpliwie może stanowić ciekawe źródło hipotez dla bardziej szczegółowych badań.

Ciekawym problemem okazuje się aktywność pozaszkolna, która – według Siemieńskiej i Białeckiego – *nie stanowi dodatkowego „kapitału" najlepszych uczniów*. Lecz jak dodaje Rafał Piwowarski (*Aktywność społeczno-kulturalna ucznia a wyniki egzaminów zewnętrznych*, Biuletyn Badawczy CKE „Egzamin", 12/2007): *możemy jednak z całą pewnością powiedzieć, że istnieją znaczne różnice między uczniami, którzy osiągnęli najlepsze i najgorsze wyniki z egzaminu gimnazjalnego, a szeroko rozumianą aktywnością pozaszkolną młodzieży. Uczniowie lepsi znacznie częściej biorą udział w zajęciach pozaszkolnych, kursach i konkursach. Częściej również wyjeżdżają – zwłaszcza za granicę.*

Ciekawe wyniki przyniosły też dociekania Grażyny Wieczorkowskiej i Małgorzaty Siarkiewicz (*Inteligencja ucznia a wynik egzaminu zewnętrznego*, Biuletyn Badawczy CKE „Egzamin", 13/2007). Przeprowadzone przez te badaczki analizy wykazały silny związek poziomu inteligencji z wynikami egzaminów, choć najlepszym predyktorem nie okazała się wcale inteligencja. Zmiennymi lepiej wyjaśniającymi wyniki egzaminu okazały się oceny nauczycielskie. Rezultat ten pokazuje silne powiązanie tematyki badanej podczas egzaminów zewnętrznych z kwestiami nauczanymi w szkołach, których pojmowanie zależy tylko częściowo od szeroko pojętej inteligencji ucznia.

Zaprezentowany materiał badawczy ma w dużej mierze charakter opisowy. Poddany analizie zbiór danych może być wykorzystany do dalszych, bardziej szczegółowych analiz, odpowiadających na ważne dla systemu oświaty pytania. Postawiono szereg hipotez, jednak ich sprawdzenie wymaga bardziej precyzyjnego określenia zmiennych i prawdopodobnie zastosowania nowoczesnych metod zastąpienia braków danych, szczególnie w wypadku analiz dotyczących dochodów, gdzie liczba dostępnych obserwacji jest znacznie mniejsza niż wielkość próby. Niewątpliwie jednak najważniejszą kwestią są problemy metodologiczne dotyczące mocnego powiązania zmiennych, z których wiele jest pozornie skorelowanych z wynikami uczniów i ukrywa oddziaływanie zmiennych wymagających osobnego zdefiniowania lub też niedostępnych w zbiorze danych. Niewątpliwie w dalszych analizach powinno

się brać pod uwagę również czynniki pozaekonomiczne, opisane w innych raportach z tego badania.

Badania na mniejszych populacjach

Badania reprezentatywne dla całej populacji polskich uczniów są, jak pisaliśmy w pierwszej części tego rozdziału, drogie, wymagające dużej liczby obserwacji. Jakkolwiek badania na mniejszych populacjach nie dają nam tak pewnych informacji i nie umożliwiają uogólnienia wniosków na cały kraj, to można dzięki nim poznać i zrozumieć przynajmniej wycinek rzeczywistości, zweryfikować wstępne hipotezy lub potraktować je jako badania pilotażowe przed badaniem ogólnopolskim. Oczywiście jest też możliwe, iż badaczy interesuje jedynie mały wycinek geograficzny, powiat, gmina, miasto, województwo czy obszar przynależny do konkretnej okręgowej komisji egzaminacyjnej. Ważne jest to, że ograniczenie populacji, o której chcemy się wypowiadać, nie umniejsza wcale poprawności metodologicznej ani nie podważa wniosków odnoszących się do tej populacji, jeżeli próba pozostaje losowa. Tutaj ważne jest tylko to, aby nie formułować na podstawie takich badań sądów, które kategorycznie odnosiłyby się do całej populacji.

Jednym z przykładów badań przeprowadzonych na nieogólnopolskiej próbie uczniów było badanie: *Dlaczego uczniowie nie lubią matematyki? Pedagogiczno-społeczne uwarunkowania osiągnięć egzaminacyjnych uczniów w zakresie umiejętności matematycznych* opracowane przez Zofię Hryhorowicz, Adelę Łabuzińską oraz Bartosza Mikę, a opublikowane w Biuletynie Badawczym CKE „Egzamin", 15/2008. Badanie przeprowadzone było na reprezentatywnej próbie dla okręgu poznańskiego[39]. Wyników tych nie można niestety uogólnić na cały teren OKE Poznań, a tym bardziej dla całej Polski[40]. Trzeba zaznaczyć, iż autorzy nie uwzględnili w analizach pogrupowania uczniów w szkołach (patrz rozdział 2), co nie podważa głównych wyników, jednak prezentowana przez autorów precyzja oszacowań jest z pewnością dość znacznie zawyżona. W badaniu wzięło udział ponad 2,5 tysiąca uczniów ze szkół podstawowych, gimnazjów i szkół ponadgimnazjalnych oraz ponad stu nauczycieli i dyrektorów szkół.

Zaletą tego badania jest to, iż projektując proces badawczy, założono wielostronne podejście do problemu nauczania matematyki w szkołach. Problem

[39] Gorzów Wlkp., Konin, Kalisz, Koszalin, Leszno, Piła, Poznań, Szczecin, Zielona Góra.

[40] W 2006 roku Centralna Komisja Egzaminacyjna zrealizowała ogólnopolskie badania umiejętności matematycznych trzecioklasistów. Rezultaty tego badania, niestety mające głównie charakter opisowy, ale dostarczające wielu cennych informacji i spostrzeżeń, można znaleźć na stronach CKE (www.cke.edu.pl), gdzie dostępna jest także książka M. Dąbrowskiego analizująca odpowiedzi uczniów na pytania testowe (*Pozwólmy dzieciom myśleć! O umiejętnościach matematycznych polskich trzecioklasistów*).

niechęci i niskich wyników uczniów z matematyki jest mało poznany, a opisywane badania pokazały go w nowym świetle. Podstawową metodą badawczą był audytoryjny wywiad kwestionariuszowy uczniów oraz indywidualne kwestionariusze nauczycieli i dyrektorów. Badania te miały również komponent jakościowy, bowiem badacze przeprowadzili szereg obserwacji lekcji, a część uczniów poproszona została również o rozwiązanie specjalnie przygotowanych zadań. Analizowane były również wyniki egzaminów gimnazjalnych.

Analizując dane, jeżeli chodzi o warunki nauczania w okręgu poznańskim, badacze stwierdzili, że:

- im wyższy poziom edukacyjny, tym mniejsza liczba obowiązkowych godzin matematyki,
- zajęcia dydaktyczno-wyrównawcze realizowane są zgodnie z obowiązującą siatką godzin,
- uczniom uzdolnionym, szczególnie na wyższych etapach edukacyjnych, stwarza się warunki do poszerzania swoich zainteresowań podczas zajęć nieobowiązkowych,
- największą liczbę uczestników konkursów odnotowano w szkołach podstawowych,
- nauczyciele części szkół rozumieją potrzebę popularyzowania wiedzy matematycznej (Biuletyn Badawczy CKE „Egzamin", 15/2008, s. 17).

Jednego z ciekawszych wyników tego badania dostarczyła analiza dyferencjału semantycznego (precyzyjniej: pewnego wariantu klasycznego dyferencjału), czyli pytania ankietowego wymagającego, aby uczeń określił swoją postawę na skali, gdzie podane są jedynie skrajne twierdzenia. Badania pokazały, iż uczniowie nie są jednoznacznie negatywnie nastawieni do matematyki. Mało tego, są przekonani o jej użyteczności nie tylko w szkole, ale i w życiu codziennym. Wyniki badania pokazują też, jak wielką satysfakcję sprawia poprawne rozwiązanie zadania, co może być istotnym czynnikiem motywacyjnym. Zdecydowana większość uczniów przychyla się bowiem do twierdzenia, iż rozwiązanie trudnego zadania matematycznego sprawia im dużą przyjemność. Są to stosunkowo nieoczekiwane wyniki, które pokazują, iż nauczanie matematyki powinno być silnie popierane przez samych uczniów.

Jednocześnie przychylność do matematyki wyraźnie się zmniejsza wraz z kolejnymi etapami edukacji. Uczniowie szkół podstawowych lubią matematykę najbardziej, podczas gdy uczniowie szkół ponadgimnazjalnych są już częściej nastawieni do niej negatywnie. Na podstawie tych badań można zatem postawić hipotezę, iż polska szkoła systematycznie zniechęca uczniów do matematyki: im dłużej się jej uczą, tym bardziej nie chcą się jej uczyć. Oczywiście, taką tezę należałoby zweryfikować w oparciu w właściwe badania podłużne, mierzące postawy i opinie tych samych uczniów na przestrzeni lat lub porównanie matematyki z innymi przedmiotami szkolnymi.

Dodatkowych informacji dostarczyły obserwacje lekcji. Przykład tego typu analizy stanowi ciekawe połączenie badań ilościowych i jakościowych. Choć badanie samo sobie nie dostarcza zbyt wielu konkluzji, obarczone jest też paroma niedociągnięciami metodologicznymi (choćby nieuwzględnienia struktury losowania), to stawia wiele ciekawych hipotez i pobudza do refleksji nad sposobami nauczania matematyki w Polsce.

4.3. Porównawcze badania międzynarodowe

Badanie PISA

Badanie PISA (*Programme for International Student Assessment* – Program Międzynarodowej Oceny Umiejętności Uczniów) to jedyne międzynarodowe badanie umiejętności i wiadomości uczniów obecne w Polsce od samego początku jego istnienia. Badanie to zarządzane jest przez Organizację Współpracy Gospodarczej i Rozwoju (OECD) zrzeszającą trzydzieści najwyżej rozwiniętych demokratycznych państw świata, w tym Polskę. PISA realizowana jest jednak w znacznie większej liczbie krajów, które zgłaszają do niej akces, pragnąc porównać osiągnięcia własnych uczniów z krajami OECD.

Badanie PISA realizowane jest w cyklu 3-letnim, począwszy od 2000 roku. W każdym z cykli oceniane są trzy kluczowe domeny: czytanie, matematyka i nauki ścisłe, jednak jedna z nich oceniana jest ze znacznie większą precyzją. I tak w 2000 roku szczegółowej ocenie poddano czytanie, w 2003 matematykę, a w 2006 nauki ścisłe. W praktyce oznacza to, że wybrana dziedzina oceniana jest na podstawie około 100 pytań dotyczących wielu jej aspektów. Następnie wyniki podawane są na trzech skalach, osobno dla czytania, matematyki i nauk przyrodniczych, ale dla głównej dziedziny także na podskalach.

Przykładowo, w 2006 roku, gdy badano przede wszystkim nauki ścisłe, podano wyniki nie tylko dla nauk przyrodniczych, matematyki oraz czytania, ale i dla poddziedzin nauk przyrodniczych, gdzie wydzielono trzy skale dotyczące umiejętności rozpoznawania zagadnień naukowych, wyjaśniania zjawisk przyrodniczych w sposób naukowy, a także interpretacji i wykorzystywania wyników i dowodów naukowych. Dodatkowo przedstawiono także wyniki w podziale na dziedziny wiedzy z zakresu nauk przyrodniczych, gdzie także wyodrębniono trzy skale dotyczące wiedzy z zakresu układów nieożywionych, układów ożywionych oraz układów Ziemi i kosmosu. W edycji PISA 2009 badane w najszerszym zakresie będzie ponownie czytanie, a w edycji PISA 2012 najdokładniejszy pomiar będzie dotyczył ponownie matematyki etc.

Oprócz pomiaru wiedzy i umiejętności uczniów, badanie PISA gromadzi także bogate informacje dotyczące motywacji, opinii i cech rodziny ucznia.

Niemal wszystkie zebrane dane publikowane są w ogólnie dostępnych bazach danych (patrz rozdział 5), co umozliwia prowadzenie niezależnych analiz. Sama baza danych dla Polski stanowi niezwykle bogaty materiał do analiz, niestety, rzadko wykorzystywany w naszym kraju. Niewiele powstało prac badawczych analizujących wyniki i kontekst osiągnięć uczniów polskich. Ciekawy materiał zawierają raporty krajowe z każdej edycji badania PISA, które stanowią w dużej mierze materiał opisowy prezentujący uzyskane wyniki, choć zawarte są tam także wstępne analizy z uwypukleniem pozycji Polski[41].

Wyniki badania PISA są dość ciekawe same w sobie, dlatego poniżej prezentujemy je w kontekście osiągnięć uczniów polskich. Następnie omawiamy wyniki badania porównującego osiągnięcia polskich uczniów na przestrzeni kilku edycji PISA. Badanie to zostało zainicjowane przez Bank Światowy, zainteresowany przyczynami znaczącego wzrostu poziomu umiejętności polskich uczniów mierzonych w PISA między pierwszą edycją w 2000 a ostatnią w 2006 roku. Wzrost ten spostrzeżony został w wielu krajach i wzbudził chęć poznania przyczyn tak znacznego polepszenia się wyników polskich uczniów.

Badanie PISA realizowane jest w każdym kraju na osobnej próbie reprezentatywnej dla populacji 15-latków, jednak w oparciu o te same zasady realizacji badania oraz te same kwestionariusze i pomiar wiadomości i umiejętności uczniów. PISA posiada swoistą koncepcję umiejętności i wiadomości uczniów, które określane są wspólnym terminem *literacy* (tłumaczone dość dowolnie jako „rozumowanie" lub też „biegłość"). Organizatorzy PISA twierdzą, że badanie to wyróżnia się pomiarem wiadomości i umiejętności niezbędnych uczniom w przyszłości w życiu dorosłym, na rynku pracy lub do tego, aby w pełni swobodnie funkcjonować w społeczeństwie.

Metody pomiaru przedstawiliśmy w rozdziale 1. Są one podobne dla wszystkich międzynarodowych badań umiejętności, także dla PIRLS oraz TIMSS. To, co wyróżnia PISA, to właśnie definiowanie mierzonych umiejętności w 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 15-latkowie.

Dane PISA można analizować w dwóch, równie ważnych dla każdego systemu edukacyjnego, wymiarach. Pierwszy z nich dotyczy poziomu umiejętności uczniów. Mierzy się go przede wszystkim przez średni wynik uczniów w danym

[41] Patrz informacje na stronie www.ifispan.waw.pl/ifis/badania/program_pisa Instytutu Filozofii i Socjologii PAN realizującego program PISA w Polsce, gdzie dostępne są m.in. krajowe raporty PISA.

kraju, ale i przez procent uczniów, którzy osiągnęli odpowiednio wysoki poziom umiejętności (w PISA 2006 zdefiniowano na przykład 6 poziomów umiejętności w naukach 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 (mierzy poziom segregacji uczniów między szkołami ze względu na poziom wiedzy) lub też moc zależności między osiągnięciami a statusem społeczno-ekonomicznym rodziny ucznia (im silniejsza, tym większe nierówności ze względu na pochodzenie społeczne). Wyniki te przedstawiamy poniżej dla Polski, opierając się na oficjalnych raportach oraz własnych analizach baz danych PISA.

Wykres (rysunek 4.4) pokazuje pozycję Polski w odniesieniu do krajów OECD, zarówno pod względem średniego wyniku w naukach przyrodniczych w PISA 2006, jak i pod względem zróżnicowania wyników. Jak widać, Polska plasuje się pod względem wyniku (oś pionowa) blisko średniej dla OECD wynoszącej 500 punktów (dokładnie Polska uzyskała 498 punktów). Bardzo podobne wyniki, nierozróżnialne statystycznie od średniej dla OECD, uzyskały takie kraje jak Węgry, Szwecja, Dania czy Francja.

Wykres pokazuje jednak wyraźnie, że wyróżnia nas niskie zróżnicowanie wyników. Odchylenie standardowe wyników uczniów w naukach przyrodniczych wyniosło w Polsce w 2006 roku 90 punktów. Jest to mniej niż średnia OECD równa 95, znacznie mniej niż w krajach anglosaskich (USA, Wielka

Rysunek 4.4. PISA 2006, nauki przyrodnicze: zróżnicowanie i średni poziom wyników w krajach OECD i w Polsce

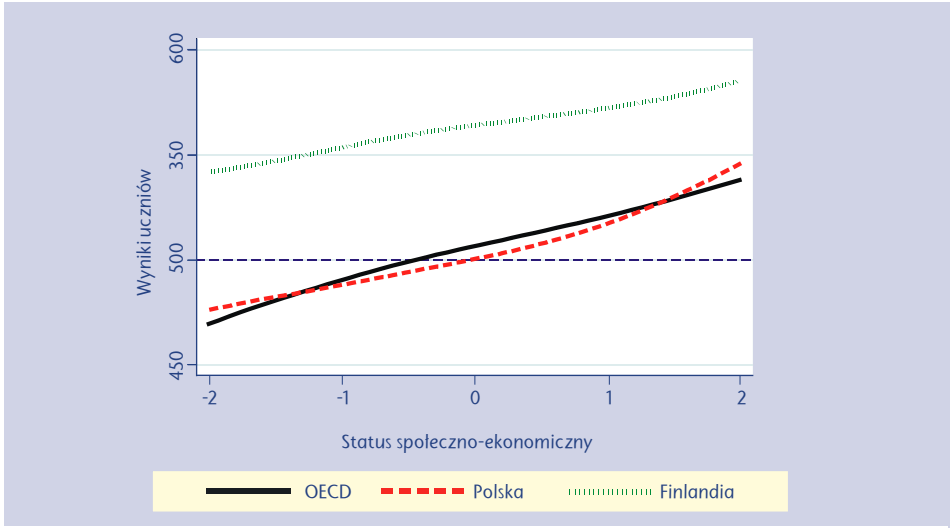
Brytania, Nowa Zelandia), które tradycyjnie charakteryzuje spore zróżnicowanie osiągnięć uczniów, ale i mniej niż w krajach skandynawskich. Trzeba tu jednak wziąć pod uwagę fakt, że duże zróżnicowanie w tych krajach wynika w dużym stopniu z uwzględnienia w badaniu imigrantów, których w szkołach w Polsce jest znikoma liczba.

Spoglądając na wyniki w podskalach dla nauk przyrodniczych, można zauważyć, że polscy uczniowie znacznie słabiej niż średnia dla OECD wypadli w rozpoznawaniu zagadnień naukowych, trochę gorzej w interpretacji i wykorzystywaniu wyników i dowodów naukowych, a nieco lepiej w wyjaśnianiu zjawisk przyrodniczych w sposób naukowy. W odniesieniu do zakresów wiedzy polscy uczniowie najlepiej wypadli w zakresie układów ożywionych.

Niewielkie zróżnicowanie wyników, które niewątpliwie stanowi powód do chwały, przekłada się też, niestety, na znikomą (w porównaniu z innymi krajami OECD) liczbę uczniów, którzy dysponują umiejętnościami na najwyższym poziomie. W Polsce na najwyższych poziomach 5 i 6 według kategorii PISA znalazło się jedynie 6,8% uczniów, w porównaniu do 9% średnio w OECD i ponad 20% w Finlandii. Co prawda posiadamy też mniej niż przeciętnie w OECD uczniów na najniższych poziomach, ale daleko nam tutaj do Finlandii, gdzie słabych uczniów jest znacznie mniej (w Polsce w naukach przyrodniczych na poziomie 1 i poniżej znajduje się 17% uczniów, średnio w OECD jest to 19%, a w Finlandii jedynie 4%). Niewątpliwie brak odpowiedniej liczby uczniów o najwyższych umiejętnościach budzi wątpliwości co do tego, czy w systemie naszego szkolnictwa gimnazjalnego dostatecznie wspieramy najbardziej zdolnych uczniów. Spora liczba uczniów o najniższym poziomie wiedzy pokazuje też, że mimo znacznej poprawy, nadal mamy wiele do zrobienia, jeśli chodzi o uczniów najsłabszych.

Nieco inny obraz wyłania się, gdy weźmiemy pod uwagę status społeczno-ekonomiczny rodzin uczniów. W tym wypadku Polska nieco odstaje od innych członków OECD, jako kraj o wciąż stosunkowo niższym niż przeciętna krajów zamożnych poziomie wykształcenia i innej strukturze zawodowej. Z badań IALS przeprowadzonych w Polsce w 1994 roku wiemy też, że poziom umiejętności w populacji dorosłych jest niższy niż w takich krajach, jak Kanada lub Szwajcaria, co zapewne również przekłada się na relację między pochodzeniem społecznym a wynikami uczniów.

Badanie PISA umożliwia analizowanie tej relacji w oparciu o szereg wskaźników, z których najczęściej używanym jest indeks ESCS mierzący nie tylko pozycję społeczno-zawodową rodziców ucznia, lecz także zasoby kulturowe i ekonomiczne jego rodziny. Relację między tym wskaźnikiem a osiągnięciami uczniów w naukach przyrodniczych ukazano na wykresie poniżej. Relację tę oszacowano tylko dla krajów OECD i po wyłączeniu z analiz imigrantów, aby zapewnić większą porównywalność z danymi dla Polski.



Rysunek 4.5. PISA 2006: relacja między statusem społeczno-ekonomicznym rodziny ucznia a jego osiągnięciami w naukach przyrodniczych

Wykres (rysunek 4.5) ukazuje relację między statusem społeczno-ekonomicznym rodziny ucznia a jego osiągnięciami w naukach przyrodniczych w krajach OECD, a także osobno dla Polski oraz dla Finlandii. Finlandia wyróżnia się tu pod dwoma względami, podstawowymi dla oceny systemów edukacji. Po pierwsze, wyraźnie widać, że niezależnie od statusu rodziny ucznia wyniki 5-latków w Finlandii są wyższe niż przeciętnie w krajach OECD (około 500 punktów, oznaczone poziomą, przerywaną, niebieską linią), a także w Polsce.

Finlandia od lat przoduje w rankingach PISA, potwierdzając niezwykle wysoką efektywność swojego systemu kształcenia powszechnego. Co więcej, uczniowie z rodzin o niskim statusie społeczno-ekonomicznym uzyskują w Finlandii wyniki znacznie bliższe uczniom z rodzin o statusie wyższym. Inaczej mówiąc, relacja między statusem a osiągnięciami jest bardziej „płaska" w Finlandii niż przeciętnie w OECD. Świadczy to nie tylko o wysokiej efektywności systemu kształcenia w Finlandii, ale i o jego dużej mocy wyrównywania szans edukacyjnych.

Z wykresu wynika też, że uczniowie w Polsce uzyskują podobne wyniki do swoich rówieśników w krajach OECD, bliskie dla każdego poziomu statusu społeczno-ekonomicznego ich rodzin. Widoczne są jednak niewielkie różnice, które mogą stanowić temat bardziej dogłębnych badań (niestety, nie znamy takiego przykładu badań w Polsce). Otóż uczniowie z rodzin o niskim, ale i o wysokim statusie, uzyskują nieco wyższe wyniki, niż ich rówieśnicy z rodzin o podobnym statusie w innych krajach OECD. Niestety, uczniowie z rodzin o „przeciętnym" statusie uzyskują wyniki wyraźnie niższe.

spis treści

Choć pełna porównywalność skali mierzącej status rodzin dla Polski i bardziej rozwiniętych krajów OECD może budzić pewne wątpliwości, to zakładając, że jednak taka porównywalność w dużym stopniu istnieje, można pokusić się o postawienie tezy, że polskie gimnazja, w porównaniu ze szkołami kształcącymi 15-latków w innych krajach OECD, sprzyjają uczniom z rodzin zarówno o niskim, jak i o bardzo wysokim statusie. To ciekawa teza, stanowiąca jednak raczej hipotezę do dalszych badań niż wniosek badawczy. Niemniej te proste analizy pokazują potencjał, jaki niosą ze sobą badania PISA. Należy mieć nadzieję, że będzie on częściej wykorzystywany w badaniach w naszym kraju.

Polska wyróżniła się w badaniach PISA jednym z większych wzrostów w wynikach uczniów między rokiem 2000 a 2006. Skupimy się tutaj na przyroście umiejętności uczniów w czytaniu, bowiem tylko ta domena jest w pełni porównywalna między edycją PISA 2000 i PISA 2006. Wnioski zaprezentowane poniżej pochodzą z raportu Banku Światowego poszukującego przykładów korzystnych rozwiązań w systemach edukacji na całym świecie[42].

Badanie PISA 2000 przeprowadzono w Polsce jeszcze w starym systemie, gdzie 15-latkowie znajdowali się w szkołach różnego typu: liceach, technikach oraz szkołach zawodowych. Wyniki PISA 2000 były dla Polski wysoce niekorzystne zarówno pod względem niskiego średniego wyniku w zakresie umiejętności czytania, jak i dużego zróżnicowania umiejętności uczniów. Szczególnie niekorzystnie wypadli uczniowie szkół zawodowych, z których wielu nie mogło się pochwalić nawet podstawowym poziomem umożliwiającym czytanie ze zrozumieniem prostych tekstów. Wyniki PISA 2003 a następnie 2006 udokumentowały wysoki wzrost umiejętności polskich uczniów w zakresie czytania. Powstaje uzasadnione pytanie, czym spowodowany był ten wzrost i czy rzeczywiście miał miejsce, czy też poprawa wyników związana była ze zmianami w metodologii badania.

W omawianym raporcie wykorzystano między innymi metody dopasowania prób za pomocą tzw. *propensity score reweighting*. Dzięki tym metodom możliwe jest dokonanie symulacji wyników uczniów, którzy uczęszczali do szkół zawodowych w 2000 roku, gdyby posiadali poziom wiedzy podobnych uczniów (ze względu na takie cechy, jak płeć czy pochodzenie społeczne) zbadanych w 2006 roku. Inaczej mówiąc, metody te symulują zmianę w osiągnięciach uczniów różnych rodzajów szkół związaną z wprowadzeniem jednolitego systemu edukacji w ramach gimnazjum. Symulacja możliwa jest dzięki temu, że dane PISA posiadają dokładne charakterystyki uczniów i ich rodzin, w dużym stopniu określające zarówno prawdopodobieństwo uczęszczania w 2000 roku do danego typu szkoły, jak i uzyskiwane w danej edycji PISA wyniki. Porównanie zaprezentowano w tabeli 4.1 na następnej stronie.

[42] Jakubowski, M., Patrinos, H. A., Porta, E. E., Wiśniewski, J. (2008), *The Impact of the 1999 Education Reform in Poland*, raport opracowany na zlecenie Banku Światowego (w publikacji).

Tabela 4.1. Zmiana w wynikach polskich uczniów między edycjami PISA – porównanie osiągnięć w czytaniu w oparciu o próby uczniów o zrównoważonych charakterystykach

Typ szkoły	Faktyczne wyniki polskich uczniów PISA 2000 (średnia OECD = 500)	Faktyczne wyniki polskich uczniów PISA 2006 (średnia OECD = 500)	Symulowana zmiana w wynikach między PISA 2000 a PISA 2006
Wszystkie szkoły	480	514	37,3* (6,4)
Uczniowie szkół zawodowych z 2000 roku	358	–	114,6* (6,9)
Uczniowie techników z 2000 roku	480	–	23,4* (5,9)
Uczniowie liceów ogólnokształcących z 2000 roku	544	–	2,1 (5,8)

Uwagi: W nawiasach podano błędy standardowe. * wynik istotny statystycznie na poziomie 0,01.

Wyniki te wyraźnie pokazują, że największymi beneficjentami zmian systemu edukacji w Polsce byli uczniowie szkół zawodowych. Symulowany przyrost umiejętności czytania mierzonych przez PISA 2000 oraz PISA 2006 wyniósł dla tej grupy uczniów ponad 100 punktów, co jest zmianą olbrzymią, większą niż jedno odchylenie standardowe wyników uczniów w krajach OECD. Co więcej, zmiana ta była także korzystna dla uczniów techników, którzy zyskali około 20 punktów, czyli 1/5 odchylenia standardowego wyników. To także zmiana, która w badaniach edukacyjnych uznawana jest zazwyczaj za bardzo znaczną. Uczniowie liceów, zgodnie z intuicją, nie zyskali na reformie oświaty. Niemniej jednak wprowadzenie jednorodnego systemu edukacji dla 15-latków w Polsce przyniosło wzrost przeciętnego poziomu umiejętności o ponad 1/3 odchylenia standardowego wyników. Potwierdza to sukces reformy edukacji w tym względzie.

Badanie nad edukacją obywatelską (Civic Education Study)

O badaniu *Civic Education Study* warto pisać przynajmniej z kilku powodów. Po pierwsze, było to jedno z pierwszych międzynarodowych badań edukacyjnych

spis treści

przeprowadzonych na taką skalę, w której Polska wzięła udział. Po drugie, badanie to charakteryzuje się ciekawą i złożoną strukturą, nie było to wyłącznie badanie sondażowe oraz mierzące poziom umiejętności uczniów. Włożono dużo wysiłku, aby przeanalizować i porównać kontekst badania w różnych krajach oraz opisać różniące się między krajami warunki instytucjonalne. Po trzecie, w najbliższych latach planowane jest powtórzenie tego badania; będziemy mogli zatem zobaczyć, jak wiele się zmieniło w przeciągu ostatnich dziesięciu lat, jeżeli chodzi o wychowanie obywatelskie. Po czwarte, badanie to eksploruje niezwykle ważny dla społeczeństwa wycinek wiedzy i umiejętności młodych obywateli, często niesłusznie bagatelizowany i pomijany podczas analiz wiedzy i osiągnięć uczniów.

Przechodząc do konkretów, celem badań zorganizowanych przez Międzynarodowe Towarzystwo Badania Osiągnięć Szkolnych (*The International Association for the Evaluation of Educational Achievement* – IEA) było określenie poziomu wiedzy i postaw obywatelskich. Badania miały służyć zarówno praktykom z obszaru szkolnictwa, jak i decydentom; przyniosły bowiem szczegółowy porównawczy materiał empiryczny z zakresu wychowania obywatelskiego w skali nieznanej do tej pory w Europie. Szczegółowo rzecz ujmując, można powiedzieć za Anną Wiłkomirską (w: Dolata 2004: 9), że *badaczy interesowało, w jakim stopniu szkoła włącza się lub może włączać w proces edukacji obywatelskiej oraz na ile jej oferta może być kompensacyjna, a na ile komplementarna wobec oddziaływań innych źródeł socjalizacji: rodziny, rówieśników, mediów i różnych organizacji młodzieżowych*.

Badanie CivEd (dalej będziemy używać tego skrótu od *Civic Education Study*) planowane było od 1994 roku. W proces ten zaangażowanych było pięćdziesiąt jeden krajów, z czego dwadzieścia osiem zdecydowało się na udział w badaniach. Projekt ten składał się z trzech elementów rozłożonych chronologicznie na 3 etapy – badania kontekstowo-instytucjonalne, badanie populacji 14-latków i badania młodzieży starszej – 17-latków).

Od 1996 do 1998 przeprowadzona została pierwsza faza CivEd, której celem było systematyczne opisanie procesu edukacji obywatelskiej w krajach uczestniczących w badaniu w postaci studium przypadku (kraju). Dzięki temu opisany został szerszy kontekst dla badania w danym kraju. *Zespoły narodowe opracowały analizy wielu problemów: statusu edukacji obywatelskiej w szkołach, priorytetów programowych, które pozwalają przygotować młodzież do zadań obywatelskich, konstrukcji samych programów nauczania w omawianym przedmiocie, rozwijania społecznej tożsamości – narodowej, lokalnej, ponadnarodowej oraz stosunku do ważnych międzynarodowych konfliktów. Badano również poglądy obywateli na temat pożądanego sposobu kształcenia obywatelskiego oraz jaki obraz społeczeństwa obywatelskiego jest kreowany w podręcznikach i innych materiałach dydaktycznych*. (Wiłkomirska w: Dolata 2004: 11).

W drugiej fazie CivEd skupiono się na badaniu wiedzy i postaw związanych z wychowaniem obywatelskim 14-latków, co w wypadku Polski ma niebagatelne znaczenie z uwagi na to, iż przebadany został tym samym ostatni rocznik ósmoklasistów. Główne badanie drugiego etapu (poprzedzone wcześniejszymi badaniami pilotażowymi) przeprowadzone zostało w 1999 roku.

W Polsce, by zapewnić reprezentatywność i porównywalność z innymi krajami, przebadano 3 376 uczniów w 179 szkołach. Tak jak i w innych tego typu projektach, stopa realizacji próby badawczej była w Polsce wysoka. Aż 90% wylosowanych szkół wzięło udziału w badaniu, a w szkołach, w których odbyło się testowanie, do testu i ankiety przystąpiło 94% uczniów. Warto podkreślić, iż międzynarodowe minimum wynosiło 75% i w wypadku aż 10 krajów nie udało go się osiągnąć.

W sumie we wszystkich krajach, w których przeprowadzone zostało badanie CivEd, 90 000 14-latków rozwiązało test mierzący wiedzę i wypełniło ankietę badającą postawy. Równolegle z badaniami uczniów przeprowadzono wywiady z nauczycielami trzech przedmiotów: wiedzy o społeczeństwie, historii oraz języka polskiego, oraz z dyrektorami. Jednak wyniki ankiet nauczycielskich i dyrektorskich przeanalizowane zostały w sposób szczątkowy. Skupiono się na wynikach młodzieży, gdzie głównym narzędziem badawczym był kwestionariusz ucznia, który z grubsza można podzielić na dwie części. W pierwszej części uczniowie rozwiązywali test wiadomości i umiejętności interpretacji przekazu politycznego. W drugiej części wypełniali ankietę dotyczącą opinii i postaw związanych z życiem obywatelskim.

Na podstawie pytań zawartych w teście mierzono trzy obszary wiadomości i umiejętności obywatelskich: 1) ogólną wiedzę obywatelską, 2) wiedzę dotyczącą pojęć związanych przede wszystkim z funkcjonowaniem demokracji, 3) umiejętność rozumienia przekazów politycznych. Ogólny wskaźnik wiedzy obywatelskiej obejmował wszystkie pytania. Test skalowany był za pomocą IRT, precyzyjniej: modelu Rascha (patrz rozdział 1), utworzono trzy niezależne skale umiejętności i każdą z nich przekształcono w taki sposób, by średnia międzynarodowa uczniów ze wszystkich krajów wynosiła 100 przy odchyleniu standardowym 20.

W tabeli 4.2 przedstawiono główne wyniki badania na próbie 14-latków. Podano wyniki wyskalowane, a w nawiasach wartość błędu standardowego, który wyraża niepewność oszacowania: im wartość większa, tym niepewność większa i na odwrót – im mniejsza wartość błędu standardowego, tym niepewność oszacowania mniejsza. Warto podkreślić, że w każdym z badanych przypadków precyzja oszacowania jest zadowalająca.

Jak widać, polscy uczniowie wykazali się niezwykle wysokim poziomem wiedzy (tak ogólnej, jak pojęciowej) oraz stosunkowo wysokim (powyżej średniej 100) poziomem umiejętności rozumienia przekazów politycznych. Tak wysokie wyniki okazały się sporym zaskoczeniem dla polskich badaczy.

Tabela 4.2. Wyniki testu wiedzy i umiejętności dla *Civic Education Study* (1999) dla 28 krajów, populacja 14-latków. Podano wyniki wyskalowane, a w nawiasach błędy standardowe (za: Dolata 2005: 29)

Kraj	Ogólna wiedza obywatelska	Wiedza pojęciowa	Rozumienie przekazów politycznych
Polska	111 (1,7)	112 (1,3)	106 (1,7)
Finlandia	109 (0,7)	108 (0,7)	110 (0,6)
Cypr	108 (0,5)	108 (0,5)	108 (0,5)
Grecja	108 (0,8)	109 (0,7)	105 (0,7)
Hongkong	107 (1,1)	108 (1,0)	104 (1,0)
USA	106 (1,2)	102 (1,1)	114 (1,2)
Włochy	105 (0,8)	105 (0,8)	105 (0,8)
Słowacja	105 (0,7)	107 (0,7)	103 (0,7)
Norwegia	103 (0,5)	103 (0,5)	103 (0,4)
Czechy	103 (0,8)	103 (0,8)	102 (0,8)
Australia	102 (0,8)	99 (0,7)	107 (0,8)
Węgry	102 (0,6)	102 (0,6)	101 (0,7)
Słowenia	101 (0,5)	102 (0,5)	99 (0,4)
Dania	100 (0,5)	100 (0,5)	100 (0,5)
Niemcy	100 (0,5)	99 (0,5)	101 (0,5)
Rosja	100 (1,3)	102 (1,3)	96 (1,3)
Anglia	96 (0,6)	96 (0,6)	105 (0,7)
Szwecja	99 (0,8)	97 (0,8)	102 (0,7)
Szwajcaria	98 (0,8)	96 (0,8)	102 (0,8)
Bułgaria	98 (1,3)	99 (1,1)	95 (1,3)
Portugalia	96 (0,7)	97 (0,7)	95 (0,7)
Belgia	95 (0,9)	94 (0,9)	96 (0,9)
Estonia	94 (0,5)	94 (0,5)	95 (0,5)
Litwa	94 (0,7)	94 (0,7)	93 (0,7)
Rumunia	92 (0,9)	93 (1,0)	90 (0,7)
Łotwa	92 (0,9)	92 (0,9)	92 (0,8)
Chile	88 (0,7)	89 (0,6)	88 (0,8)
Kolumbia	86 (0,6)	89 (0,8)	84 (1,2)

Jak pisaliśmy wcześniej, oprócz wiedzy i umiejętności badano również postawy uczniów. Analizę danych zorganizowano wokół kilkunastu podstawowych skal postaw obywatelskich:

- skala konwencjonalnych zachowań obywatelskich
- skala zaangażowania obywatelskiego
- skala odpowiedzialności ekonomicznej państwa
- skala odpowiedzialności socjalnej państwa
- skala zaufania do instytucji życia publicznego
- skala patriotyzmu
- skala praw imigrantów
- skala praw kobiet
- skala planowanych działań politycznych

- skala przekonania o sile wspólnych działań
- skala otwartego klimatu

Szczegółowe omówienie wyników analiz opartych na tych skalach można znaleźć w książce *Młodzi obywatele. Wyniki międzynarodowych badań młodzieży* napisanej wspólnie przez Romana Dolatę, Krzysztofa Kosełę, Annę Wiłkomirską oraz Annę Zielińską (2004). Poświęćmy natomiast odrobinę miejsca metodologicznym aspektom tej analizy.

Kwestionariusz w badaniu CivEd skonstruowany był w taki sposób, iż w kolejnych jego częściach uczniowie odpowiadali na baterię pytań dotyczących podobnych zagadnień. Założono tutaj, iż badani posiadają pewne postawy, które można analizować na podstawie empirycznych manifestacji w postaci odpowiedzi na konkretne pytania. Postawy są tu czymś więcej niż pojedynczymi odpowiedziami na pytania i charakteryzować mają światopogląd ucznia. Postawę taką traktuje się tutaj jako cechę ukrytą ucznia, co od razu przywołuje nam terminologię skalowania IRT.

Skale te tworzymy podobnie jak w wypadku testów wiadomości i umiejętności, przy czym niezbędny jest tutaj bardziej skomplikowany model IRT, zwany *Partial Credit Model*, który umożliwia analizę pytań o większej liczbie odpowiedzi na skali porządkowej. Prosty model Rascha (patrz rozdział 1) zakłada, iż mamy dwie możliwości: uczeń odpowiedział dobrze – i wtedy dostaje punkt, lub uczeń odpowiedział źle – nie dostaje punktu. Podczas zadawania pytania związanego z postawami często stosuje się formę porządkową (zdecydowanie nie zgadzam się, nie zgadzam się, trudno powiedzieć, zgadzam się, zdecydowanie zgadzam się etc.). *Partial Credit Model* umożliwia analizę takich odpowiedzi, matematyczny opis tego procesu wykracza jednak poza ramy tej publikacji[43]. Warto wiedzieć, że tak jak w wypadku skalowania prostymi modelami IRT, otrzymujemy tutaj wyniki z przedziału od -4 do 4, które (zazwyczaj) sytuują respondenta na kontinuum cechy ukrytej. W wypadku testu umiejętności wyniki zbliżające się do -4 oznaczały bardzo niskie umiejętności, a wyniki zbliżające się do +4 – bardzo wysokie.

W wypadku postaw mamy do czynienia z niską wartością charakteryzującą postawę, czyli na przykład słabym patriotyzmem, lub z wysoką wartością charakteryzującą postawę, czyli na przykład silnym patriotyzmem. Tak jak w wypadku testów, skalę tę można łatwo transformować. Na przykład twórcy badania CivEd postanowili podawać wyniki danych postaw na skali o średniej 10 i odchyleniu standardowym 2. Jest to, oczywiście, jedynie operacja na liczbach, która

[43] Tego typu model stosuje się także do skalowania odpowiedzi z testów wiadomości i umiejętności, kiedy możliwe jest częściowe rozwiązanie zadania. Na przykład badanie PISA zastosowało taki model w 2006 roku. Szczegółowy opis modelu i odwołania do literatury można znaleźć w ogólnodostępnej publikacji OECD PISA 2006 Technical Report, do ściągnięcia ze stron www.pisa.oecd.org.

nie wpływa na interpretację, a w odczuciu autorów przyczynia się do większej czytelności podawanych wyników.

Pod koniec opisu tego badania warto wspomnieć o trzecim etapie badania CivEd (2000), które było analogiczne do badania przeprowadzonego w drugiej fazie, lecz tym razem zbadano populację 17-latków. Tak jak w drugiej fazie, głównym narzędziem badawczym był kwestionariusz ucznia składający się z pytań badających postawy oraz testu, a w wypadku 17-latków wzbogacono go o elementy wiedzy ekonomicznej.

Niestety, z wielu powodów nie udało się przeprowadzić tej fazy we wszystkich wcześniej badanych krajach (w tej fazie wzięło udział 16 krajów), a realizacja badania w wypadku części z tych krajów pozostawiała wiele do życzenia. W konsekwencji nie można przedstawić podobnego rzetelnego rankingu, jak w wypadku 14-latków (tabela 4.2). W Polsce jednak realizacja trzeciej fazy odbyła się pomyślnie, tak jak w wypadku fazy drugiej, co otworzyło możliwości do ciekawych porównań między dwoma badanymi kohortami – 14-latków i 17-latków. Z braku miejsca nie omawiamy tych rezultatów, ale zachęcamy zatem Czytelnika do sięgnięcia do źródła, czyli polskiej analizy wyników CivEd w przywoływanej już książce *Młodzi obywatele. Wyniki międzynarodowych badań młodzieży.*

Badanie PIRLS

PIRLS to międzynarodowe badanie jednej z najważniejszych umiejętności potrzebnej do sprawnego funkcjonowania człowieka w dwudziestym pierwszym wieku – umiejętności czytania. Jak pisze Krzysztof Konarzewski, koordynator merytoryczny badania PIRLS w Polsce: *Wszelkie społeczeństwo opiera się na wiedzy, która krąży między ludźmi, jest wymieniana w codziennych spotkaniach. Dziś, inaczej niż przed wiekami, przeważająca część tej wiedzy zawiera się w tekstach pisanych. To nadaje wyjątkową wartość zdolności czytania. Ta zdolność jako dyspozycja jednostki ludzkiej jest warunkiem jej rozwoju umysłowego i samorealizacji – kto nie umie czytać, nie odniesie sukcesu w żadnej dziedzinie życia. Ta zdolność jako cecha społeczeństwa jest warunkiem jego rozwoju, zarówno gospodarczego, jak i duchowego* (2007: 7).

W badaniu PIRLS mierzy się biegłość w czytaniu wśród dzieci 10-letnich; w roku 2006 przeprowadzono je w 45 krajach. Jest to badanie, które odbywa się w cyklu 5-letnim, tak że następny cykl badania zaplanowano na 2011 rok. Warto podkreślić, iż PIRLS 2011 będzie realizowane łącznie z niezależnym badaniem TIMSS 2011 (tzw. mały TIMSS dla 10-latków), w którym zostaną także zmierzone umiejętności matematyczne oraz znajomość przyrody.

W 2006 roku w Polsce test wykonało 4 854 uczniów z 250 oddziałów klasy trzeciej w 148 szkołach podstawowych. Test, jako że objął klasę trzecią,

mówi nam o umiejętnościach uczniów, którzy kończą etap kształcenia zintegrowanego.

Test polegał na przeczytaniu tekstu i odpowiedzi na zadane pytania. Do każdego tekstu zadawano przeciętnie 12 pytań, ogółem 126. Połowa z nich wymagała wyboru jednej z czterech odpowiedzi, połowa zaś – krótkiej odpowiedzi pisemnej.

Wyniki testu wyskalowane zostały metodami IRT tak, że średnia dla wszystkich uczniów we wszystkich krajach wynosiła 500, a odchylenie standardowe 100. Stworzono kilka skal pomiarowych, z czego trzy prezentujemy w tym opracowaniu w tabeli 4.3: wynik ogólny oraz skale umiejętności wyszukiwania informacji i interpretowania tekstu. W nawiasach (podobnie jak w wypadku badania CivEd) podane zostały błędy standardowe. Należy też zwrócić uwagę, iż dla kilku krajów nie została podana wartość skali „interpretowanie" – nie pozwolił na to sposób realizacji badania w tych krajach (zbyt małe wielkości próby).

Wynik lepszy od wyniku dzieci z Polski uzyskały dzieci z 28 krajów, gorszy – z 16 krajów. Jak widać, Polska znajduje się w środku stawki, lecz niepokojące jest to, iż większość krajów Europy Zachodniej przewyższa nas pod względem badanych umiejętności. Gorszy wynik niż Polska osiągnęły: Afryka Południowa, Maroko, Kuwejt, Katar, Indonezja, Iran oraz Trynidad i Tobago. Jeżeli patrzeć na nas jako na jeden z 11 krajów dawnego bloku wschodniego, musimy ulokować siebie na siódmej pozycji, przed uczniami z Rumunii czy Macedonii, ale za uczniami rosyjskimi czy węgierskimi.

Oprócz badań kognitywnych przeprowadzone zostały również badania ankietowe rodziców i uczniów. Jednym z najciekawszych wyników, co podkreśla Konarzewski (2007: 5), mówi nam o tym, iż *Pod względem poczucia bezpieczeństwa w szkołach polscy trzecioklasiści uplasowali się na piątej pozycji, zaraz po krajach skandynawskich,* ex aequo *z Bułgarią*. Interesujące są informacje o wielkości oddziałów szkolnych. Polską szkołę charakteryzują *niewielkie oddziały klasowe (według GUS – średnio 20 uczniów w oddziale). Mniej uczniów przypadających na jednego nauczyciela mają tylko szkoły w Danii, Norwegii i na Węgrzech. Polska szkoła nie odbiega od średniej pod względem wyposażenia, a przewyższa ją pod względem wykształcenia nauczycieli* – Konarzewski (2007).

Z innych ciekawych wniosków z badania PIRLS 2006 wymienić można obserwację, że biegłość w czytaniu rośnie w oddziałach, w których czyta się dłuższe teksty prozatorskie, czy na przykład, że biegłość w czytaniu rośnie, gdy dzieci czytają nie tylko w szkole, ale i poza nią. Zachęcamy do zapoznania się ze szczegółowymi rezultatami przedstawionymi w raporcie końcowym, który jest ogólnodostępny na stronach internetowych CKE. Ogólnodostępna jest także międzynarodowa baza danych, zawierająca również wyniki polskich uczniów, o czym piszemy w rozdziale 5.

Tabela 4.3. Wyniki badania PIRSL 2006 dla 45 krajów. Podano wyniki wyskalowane, a w nawiasach błędy standardowe (za: Konarzewski 2007)

Kraj	Wynik ogólny	Wyszukiwanie	Interpretowanie
Rosja	565 (3,4)	562 (3,4)	563 (3,2)
Hongkong	564 (2,4)	558 (2,5)	566 (2,4)
Kanada, Alberta	560 (2,4)	553 (2,6)	564 (2,3)
Singapur	558 (2,9)	560 (3,3)	556 (2,7)
Kanada, Kolumbia	558 (2,6)	551 (2,8)	562 (2,5)
Luksemburg	557 (1,1)	565 (1,2)	548 (0,9)
Kanada, Ontario	555 (2,7)	543 (3,1)	563 (2,9)
Węgry	551 (3,0)	544 (2,8)	554 (3,0)
Włochy	551 (2,9)	544 (2,8)	556 (2,9)
Szwecja	549 (2,3)	550 (2,4)	546 (2,2)
Niemcy	548 (2,2)	555 (2,6)	540 (2,2)
Bułgaria	547 (4,4)	538 (4,2)	553 (4,4)
Belgia (flamandzka)	547 (2,0)	545 (1,9)	547 (1,8)
Niderlandy	547 (1,5)	551 (2,0)	542 (1,5)
Dania	546 (2,3)	551 (2,7)	542 (2,3)
Kanada, Nowa Szkocja	542 (2,2)	533 (2,2)	548 (2,0)
Łotwa	541 (2,3)	534 (2,5)	545 (1,9)
Stany Zjednoczone	540 (3,5)	532 (3,3)	546 (3,3)
Anglia	539 (2,6)	533 (2,8)	543 (2,4)
Austria	538 (2,2)	544 (2,1)	530 (2,2)
Litwa	537 (1,6)	531 (1,9)	540 (1,6)
Tajpej	535 (2,0)	541 (2,0)	530 (1,9)
Kanada, Quebec	533 (2,8)	533 (2,7)	531 (2,7)
Nowa Zelandia	532 (2,0)	524 (2,3)	538 (2,2)
Słowacja	531 (2,8)	529 (2,8)	531 (2,8)
Szkocja	527 (2,8)	525 (2,8)	528 (2,6)
Francja	522 (2,1)	523 (2,1)	518 (2,3)
Słowenia	522 (2,1)	519 (2,1)	523 (2,0)
POLSKA	519 (2,4)	516 (2,4)	522 (2,3)
Hiszpania	513 (2,5)	508 (2,5)	515 (2,6)
Izrael	512 (3,3)	507 (3,2)	516 (3,6)
Islandia	511 (1,3)	516 (1,2)	503 (1,3)
Mołdawia	500 (3,0)	486 (2,9)	515 (2,9)
Belgia (Walonia)	500 (2,6)	501 (2,6)	497 (2,5)
Norwegia	498 (2,6)	502 (2,3)	495 (2,4)
Rumunia	489 (5,0)	489 (5,2)	490 (5,3)
Gruzja	471 (3,1)	478 (3,3)	461 (3,5)
Macedonia	442 (4,1)	446 (3,8)	439 (4,0)
Trinidad i Tobago	436 (4,9)	438 (4,7)	437 (5,0)
Iran	421 (3,1)	428 (3,3)	418 (3,3)
Indonezja	405 (4,1)	409 (3,9)	404 (4,1)
Katar	353 (1,1)	361 (1,2)	…
Kuwejt	330 (4,2)	337 (3,9)	…
Maroko	323 (5,9)	336 (6,2)	…
Afryka Południowa	302 (5,6)	307 (5,3)	…

spis treści

Studium przypadku – eksperyment kwidzyński

Czasami analizowanie pewnych problemów może nie być możliwe za pomocą prób reprezentatywnych dla całej populacji. Badania studiów przypadku skupiają się na jakimś konkretnym, ciekawym z naszej perspektywy przypadku. Badania takie nie są w żadnej mierze reprezentatywne i nie możemy wnioskować, iż badane zjawiska przełożone na całą populację dadzą takie same efekty, jak w mniejszej skali czy na jednym konkretnym przypadku. Studia przypadku nie dają pewności, lecz dostarczają argumentów w dyskusji, ponadto dostarczają cennego materiału do analizy mechanizmów, które mogą wystąpić również w szerszej skali. Z drugiej strony, są cenne dla zainteresowanych właśnie danymi konkretnymi przypadkami, ujawniają bowiem mechanizmy działań i ich skutki. Trzeba jednak jeszcze raz podkreślić, że przy interpretacji ich wyników należy zawsze pamiętać o tym, że uogólnianie wniosków jest tu działalnością spekulatywną, o czym, niestety, często zapominają autorzy i interpretatorzy tego typu badań.

Przykładem takiego badania jest opisywany przez Dolatę (2008) eksperyment kwidzyński. W mieście tym w 1993 stworzony został oświatowy rynek publiczny, funkcjonujący z pewnymi zmianami do dzisiejszego dnia. Od 1996 roku eksperymentowi kwidzyńskiemu towarzyszy monitorowanie jakości kształcenia (Dolata i inni, 2001), co umożliwiło wnikliwą analizę tego przypadku.

W Kwidzynie – jak pisze Dolata (2008) – *nie zrezygnowano w pełni z rejonizacji (co zresztą w świetle polskich rozwiązań prawnych jest niemożliwe), a jedynie wprowadzono zasadę, że szkoła w ramach wolnych miejsc może przyjmować uczniów spoza rejonu. Rodziców zachęcano do aktywnego wyboru szkoły, a od 1995 roku rodzice dzieci w wieku szkolnym otrzymują symboliczny dokument – Kwidzyński Bon Oświatowy. W momencie zapisu dziecka do szkoły dokument ten jest deponowany w wybranej przez rodziców szkole. Zmiana szkoły wiąże się z przeniesieniem dokumentu do nowej placówki. Choć Bon nie jest* de facto *bonem pieniężnym, to ma pełnić funkcję psychologiczną.*(…) *Rynek szkół podstawowych w Kwidzynie jest dość bogaty. Przed reformą ustroju szkolnego działało 8 szkół podstawowych, po reformie – 5 placówek. Ponieważ Kwidzyn nie jest miastem bardzo rozległym, to nawet bez transportu samochodowego każda rodzina ma do wyboru kilka szkół. Z pewnością nie mamy do czynienia z monopolem. Z kolei niż demograficzny sprawia, że szkoły nie są przepełnione i mogą przyjmować praktycznie wszystkich chętnych spoza rejonu.*

Okazało się, że wykorzystanie przez rodziców wyboru szkół w skali całego miasta jest stosunkowo duże – około 40% uczniów szkół podstawowych uczęszcza do szkoły nierejonowej. Okazuje się również, iż popularność szkół pokrywa się w pewnym stopniu z oceną jakości szkół wydaną na podstawie badań monitorujących. W latach 1996-2000 dostrzeżono też ewidentny wzrost nasilenia segregacji międzyszkolnych. Choć, jak pisze autor studium – *Oczywiście, nie możemy mieć pewności, że wzrost ten jest wynikiem działania sił rynkowych.*

spis treści

Ale z pewnością jest to raczej argument za tezą, że tworzenie publicznego rynku oświatowego pociąga za sobą ryzyko wzrostu segregacji.

Okazuje się również, że mechanizmy rynkowe w niewielkim stopniu przyczyniły się do rywalizacji, która mogłaby podnieść efektywność kształcenia. Badania nie przyniosły dowodów na to, że zmiany w udziale w rynku poszczególnych szkół są skorelowane ze zmianami w jakości nauczania. Nie zaobserwowano też danych, które wskazywałyby na wpływ rynku na efektywność kształcenia mierzoną testami osiągnięć szkolnych.

Badania jakościowe na małych, nielosowych próbach

Wiele osób ulega złudzeniu, że badania jakościowe to badania łatwe. Nie ma tam skomplikowanych wzorów, setek liczb, karkołomnych obliczeń, skomplikowanych tabelek, długo wypracowywanych kwestionariuszy, programów komputerowych przetwarzających zebrane dane. W badaniu jakościowym wystarczy iść w teren, aby zrobić wywiady z interesującą nas grupą, a następnie opisać nasze wrażenia. Nic bardziej mylnego! Zaplanowanie i przeanalizowanie materiału z badań jakościowych może być znacznie bardziej czasochłonne i skomplikowane, niż materiału z badań ilościowych.

Badania jakościowe wymagają znacznie bardziej złożonej metodologii zbierania danych, a sposób ich analizy nie odbiega wcale wyrafinowaniem od skomplikowanych analiz statystycznych, choć rzeczywiście jest mniej zmatematyzowany. Myliłby się również ktoś, kto uważałby, iż w badaniach jakościowych nie używa się metod statystycznych i analitycznych programów komputerowych. O badaniach jakościowych powstało już wiele pozycji i nie jest tutaj naszym celem prezentowanie metodologii i technologii prowadzenia takich badań, w tym podrozdziale chcemy wskazać przykład ciekawego badania, gdzie podjęto próbę połączenia badań jakościowych z analizą wyników egzaminów zewnętrznych. Jest to studium Krzysztofa Konarzewskiego pod tytułem *Przygotowanie uczniów do egzaminu: pokusa łatwego zysku*. Raport z tych badań jest dostępny na stronie Instytutu Spraw Publicznych (www.isp.org.pl).

Autor raportu zadał sobie pytanie o wpływ systemu egzaminacyjnego na osiągnięcia uczniów i pracę nauczycieli. Samo bowiem wprowadzenie egzaminów miało – wedle autorów reformy – „uruchomić proste rezerwy tkwiące w szkołach", czyli chodziło o pełniejsze wykorzystanie czasu na lekcjach, staranniejszą realizację programu i bardziej przemyślane ocenianie uczniów. Po wprowadzeniu egzaminów pojawił się jednak problem nazywany ogólnie „nauczaniem pod testy", gdyż przygotowanie do egzaminów może służyć nie tylko usystematyzowaniu nauki, ale i jej uschematyzowaniu i spłaszczeniu. Jak pisze autor, *najczarniejszy scenariusz przewiduje coraz więcej przygotowań do egzaminu, coraz lepsze wyniki egzaminacyjne i coraz gorsze wykształcenie absolwentów*. Jak do tego nie dopuścić – to główne pytanie, jakie zadaje sobie autor.

Aby to zbadać, Konarzewski dokonał analizy testów gimnazjalnych. Założył, a później sprawdził to metodami statystycznymi, iż egzaminy gimnazjalne mogą składać się z dwóch części, a w zasadzie z dwóch oddzielnych testów. Pierwsza część mierzy dyspozycje łatwo wyuczalne, druga – trudno wyuczalne. Pierwsza promuje schematyzm i odtwórcze myślenie, druga jest nieschematyczna i nastawiona na problemowy charakter rzeczywistości. Z perspektywy nauczyciela bardziej korzystne dla uzyskania wysokich wyników może być nauczanie tylko pod zadania mierzące dyspozycje łatwo wyuczalne. Szczególnie, jeśli wyniki egzaminacyjne postrzegane są jako ważny aspekt oceny jego pracy. Czy jest tak zatem w rzeczywistości?

Przeprowadzenie wiarygodnego badania jakościowego wymagało wyboru odpowiednich szkół, a precyzyjniej: klas szkolnych. Schemat wyboru nie mógł ograniczać się jedynie do wyboru klas, w których wyniki egzaminu gimnazjalnego okazywały się znacznie lepsze lub znacznie gorsze dla dwóch typów zadań egzaminacyjnych (łatwych – odtwórczych, i trudnych – twórczych). Czasami poziom klasy nie pozwala po prostu na uzyskanie wysokich wyników w zadaniach twórczych. Autorowi badania chodziło o zbadanie takich nauczycieli, którzy przyjęli strategię nauczania odtwórczych czynności, mimo iż potencjał klasy pozwalał na znacznie ambitniejsze działania, czyli jak nazwał to autor – nie oparli się pokusie łatwego zysku nauczania pod testy. Jak zatem odnaleźć takich nauczycieli?

Konarzewski rozwiązał ten problem, korzystając z faktu, iż daną klasę uczą różni nauczyciele, a uczniowie w klasie zdają dwie części egzaminu gimnazjalnego. Można zatem wyobrazić sobie taką sytuację, w której jeden nauczyciel na swoim przedmiocie naucza rzeczy odtwórczych, a drugi na swoim przedmiocie opiera się tej pokusie. Do badania wytypowani zostali nauczyciele, którzy uczyli takie klasy, które na jednej części egzaminu gimnazjalnego uzyskiwały wysokie wyniki w pytaniach odtwórczych i niskie dla twórczych, a w drugiej części wysokie wyniki dla pytań twórczych, a niskie dla odtwórczych (lub w odwrotnej kolejności). Dzięki takiemu zabiegowi badacze, którzy wyruszyli w teren, dotarli do takich szkół, gdzie nauczyciele przedmiotów humanistycznych i matematycznych mieli diametralnie różne podejście do nauczania.

Jako że budżet badania był ograniczony, przeprowadzono je w Warszawie poprzez wywiady z 15 nauczycielami języka polskiego i z 16 nauczycielami matematyki wybranymi na podstawie opisanego schematu. Ważne jest to, że wywiady prowadzone były na podstawie wcześniej stworzonego scenariusza. Spisane wywiady zostały wprowadzone do komputera, a ich treść została zakodowania za pomocą programu komputerowego. W dalszych analizach umożliwiło to użycie prostych analiz statystycznych. Wyniki okazały się ciekawe, choć odnoszą się przede wszystkim do psychologicznego aspektu funkcjonowania nauczyciela w szkole. Po analizie i interpretacji danych, według autora „pokusie łatwego zysku" ulega ten, kto:

- kultywuje ideologiczną poprawność i zgodę interpersonalną oraz dba, by zachowywać się zgodnie z zajmowaną pozycją,
- nie dostrzega u młodzieży, a zwłaszcza u swoich uczniów, żadnych zalet – nie lubi ich lub nie ceni,
- nie zachowuje dystansu wobec instytucji zewnętrznego egzaminu testowego,
- nie ocenia pozytywnie samego siebie w roli nauczyciela-dydaktyka. (Konarzewski 2008: 29).

Badanie to stanowi ciekawy przykład łączenia analiz jakościowych prowadzonych na niewielkich próbach z analizami ilościowymi. Nie uniknęło jednak problemów związanych z tego typu badaniami. Niewielka próba powoduje, że wyniki są z pewnością zależne od doboru szkół do badania. Co więcej, uzyskane rezultaty interpretowane są w oparciu o kilka istotnych korelacji, które mogą być nie tylko pozorne, ale i przypadkowe. Niewątpliwie badanie to stawia wiele ciekawych hipotez, jakie można by poddać badaniom na szerszą skalę. Jego wartością nie są być może same wyniki, które łatwo można podać w wątpliwość, ale sposób postawienia problemu i śmiałość hipotez możliwych do wykorzystania w kolejnych badaniach. Miejmy nadzieję, że będzie ich coraz więcej.

Literatura:

Białecki, I., Renata, S. (2007), *Wpływ rodziny i grup rówieśniczych na wyniki egzaminów zewnętrznych*, Biuletyn Badawczy CKE „Egzamin", 12/2008, www.cke.edu.pl.

Białecki, I., Renata, S. (2007), *Aspiracje zawodowe uczniów a wyniki egzaminów zewnętrznych*, Biuletyn Badawczy CKE „Egzamin", 13/2008, www.cke.edu.pl.

Coleman, J. S. (1986), *Individual Interests and Collective Action*, „The American Journal of Sociology", vol. 91(6), May, s. 1309-1335.

Czapiewski, K. Ł., Śleszyński P. (2007), *Geografia zróżnicowania wyników egzaminów zewnętrznych*, Biuletyn Badawczy CKE „Egzamin", 11/2008, www.cke.edu.pl.

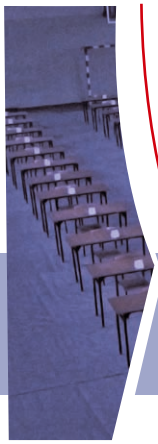
Dolata, R. (2008), *Szkoła – segregacje – nierówności*, Wydawnictwo Uniwersytetu Warszawskiego: Warszawa.

Dolata, R., Koseła, K., Wiłkomirska, A., Zielińska, A. (2004), *Młodzi obywatele. Wyniki międzynarodowych badań młodzieży*, Wydawnictwo Uniwersytetu Warszawskiego: Warszawa.

Dolata, R., Murawska, B., Putkiewicz, E., Żytko M. (2001), *Monitorowanie osiągnięć szkolnych jako metoda doskonalenia edukacji*, Wydawnictwo Akademickie Żak: Warszawa.

Dubiecka, A., Szaleniec, H., Węziak, D. (2006), *Efekt egzaminatora w egzaminach zewnętrznych*, w: *O wyższą jakość egzaminów szkolnych*, cz. I, *Etyka egzaminacyjna i zagadnienia ogólne*, Niemierko, B. (red.), Szmigel, M. K., Lublin.

Herbst, M., Herczyński, J., *Is Large More Effective than Small Is Beautiful? Size and Performance of Primary Schools in Poland*, praca niepublikowana, dostępne na www.esep.pl.
Herbst, M., Herczyński, J., *School Choice and Student Achievement*, praca niepublikowana, dostępne na www.esep.pl.
Hryhorowicz, Z., Łabuzińska, A., Mika, B. (2007), *Dlaczego uczniowie nie lubią matematyki? Pedagogiczno-społeczne uwarunkowania osiągnięć egzaminacyjnych uczniów w zakresie umiejętności matematycznych*, Biuletyn Badawczy CKE „Egzamin", 15/2008, www.cke.edu.pl.
Jakubowski, M. (2004), *Małe, średnie czy też duże? O wpływie liczby uczniów w klasie na jakość nauczania*, Biuletyn Badawczy CKE „Egzamin", 2/2004, www.cke.edu.pl.
Jakubowski, M. (2007), *Efektywność wydatków na gimnazja*, „Gospodarka Narodowa", 11-12/2007.
Jakubowski, M. (2008), *Wpływ czynników ekonomicznych na wyniki egzaminów zewnętrznych*, Biuletyn Badawczy CKE „Egzamin", 11/2008, www.cke.edu.pl.
Jakubowski, M. (2009), *Teaching quality and decentralization* (w publikacji), dostępne na www.wne.uw.edu.pl/mjakubowski w zakładce „publikacje".
Jakubowski, M., Patrinos, H. A., Porta, E. E., Wiśniewski, J. (2008), *The Impact of the 1999 Education Reform in Poland*, raport opracowany na zlecenie Banku Światowego (w publikacji).
Konarzewski, K. (2007), *Jak czytają dzieci w Polsce i na świecie PIRLS 2006*, Centralna Komisja Egzaminacyjna: Warszawa.
Konarzewski, K. (2008), *Przygotowanie uczniów do egzaminu: pokusa łatwego zysku* (raport z tych badań jest dostępny na stronie Instytutu Spraw Publicznych www.isp.org.pl).
MEN (praca zbiorowa), *Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2006 w Polsce*, Ministerstwo Edukacji Narodowej.
Piwowarski, R. (2007), *Aktywność społeczno-kulturalna ucznia a wyniki egzaminów zewnętrznych*, Biuletyn Badawczy CKE „Egzamin", 12/2008, www.cke.edu.pl.
Sakowski, P., Jakubowski, M. (2006), *Quasi-experimental estimates of class size effect in primary schools in Poland*, „International Journal of Educational Research", vol. 45(3), s. 202-215.
Swianiewicz, P., Herbst, M., Marchlewski, W. (2005), *Finansowanie i realizowanie zadań oświatowych na obszarach wiejskich*, raport na zlecenie Związku Gmin Wiejskich RP.
Wieczorkowska, G., Siarkiewicz, M. (2007), *Inteligencja ucznia a wynik egzaminu zewnętrznego*, Biuletyn Badawczy CKE „Egzamin", 13/2008, www.cke.edu.pl.



5. Bazy danych

W tym rozdziale przedstawiamy najciekawsze publicznie dostępne bazy danych. Dzięki nim każda osoba mająca chęci i kompetencje do tego (jeżeli Czytelnik dotarł do tego rozdziału, to niewątpliwie takie chęci i kompetencje posiada), może analizować samodzielnie wyniki badań przeprowadzanych na tysiącach uczniów. Może samodzielnie formułować hipotezy i empirycznie je weryfikować, używając reprezentatywnych danych zebranych w studiach, zaplanowanych i zrealizowanych z należytą dokładnością i starannością, nie wydając przy tym milionów na realizację swoich badań. Oczywiście w bazach, które będziemy prezentować, nie odnajdziemy wszystkich zmiennych, które mogłyby interesować badacza, w kontekście zagadnień edukacyjnych, lecz i tak to, czym dysponujemy dzięki zrealizowanym badaniom, może przekroczyć nawet bardzo śmiałe oczekiwania. Ilość ciekawych informacji, jakie możemy uzyskać, nie odchodząc od podłączonego do Internetu komputera, jest wprost zadziwiająca.

Oczywiście czasem nic nie zastąpi własnego badania terenowego, a my dalecy jesteśmy od namawiania do używania jedynie publicznie dostępnych zbiorów danych. Eksploracja tego typu danych może służyć nie tylko do analizy zastanych informacji, lecz może być również pomocna w fazie projektowania własnych badań. Używając udostępnionych baz danych, możemy przeprowadzić wstępną weryfikację hipotez lub dokonać obliczeń, które ułatwią nam decyzję o doborze schematu losowania we własnych badaniach.

Rozdział ten składa się z dwóch części. W pierwszej prezentujemy bazy danych badań międzynarodowych, w których uczestniczyła również Polska, w drugiej części skupimy się na ogólnodostępnych bazach danych z wynikami badań zrealizowanych w ramach projektów Centralnej Komisji Egzaminacyjnej w Polsce. W chwili publikacji tej książki bazy te powinny być udostępnione na stronach CKE w przystępnej postaci wraz z dokumentacją i podręcznikami użytkownika.

Niektóre z dostępnych baz danych wyposażone są w przeglądarki umożliwiające analizowanie zbiorów on-line, bez potrzeby użycia specjalistycznych programów statystycznych. Przeglądarkom tym poświęcimy szczególną uwagę, gdyż są to narzędzia, które umożliwiają analizę ciekawych informacji nawet mniej wprawnemu użytkownikowi komputera. Nie będzie to jednak zestaw szczegółowych instrukcji niezbędnych do wykonania obliczeń (Czytelnik bez trudu znajdzie je na stronach internetowych, na których publikowane są te zbiory danych), lecz raczej przegląd możliwości, jakie dają owe narzędzia.

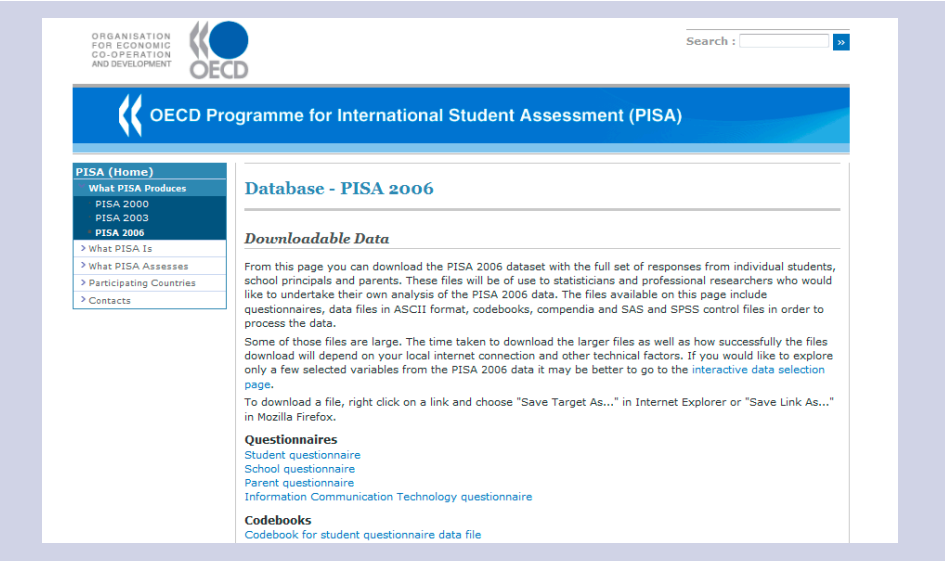
5.1. Bazy międzynarodowe

Badanie PISA (www.pisa.oecd.org)

Baza danych z badania PISA to największa z dostępnych baz danych o tematyce edukacyjnej, z jakiej można bezpłatnie czerpać informacje na temat ponad 70 krajów (w tym Polski) zebrane przez ostatnich 10 lat. Informacje o badaniu PISA (o tym, jakich zmiennych można się w tej bazie spodziewać oraz o podstawowych wynikach otrzymanych dzięki analizie danych dla Polski) Czytelnik znajdzie w rozdziale 4.

Dostęp do bazy danych można znaleźć pod adresem: http://pisa2006.acer.edu.au/.

Rysunek 5.1. Główna strona witryny (http://pisa2006.acer.edu.au/) zawierającej dostęp do bazy danych PISA



Rysunek 5.2. Zawartość zakładki *Download the PISA 2006 data set – for professional researchers*

Stronę główną tej witryny prezentujemy na rysunku 5.1.

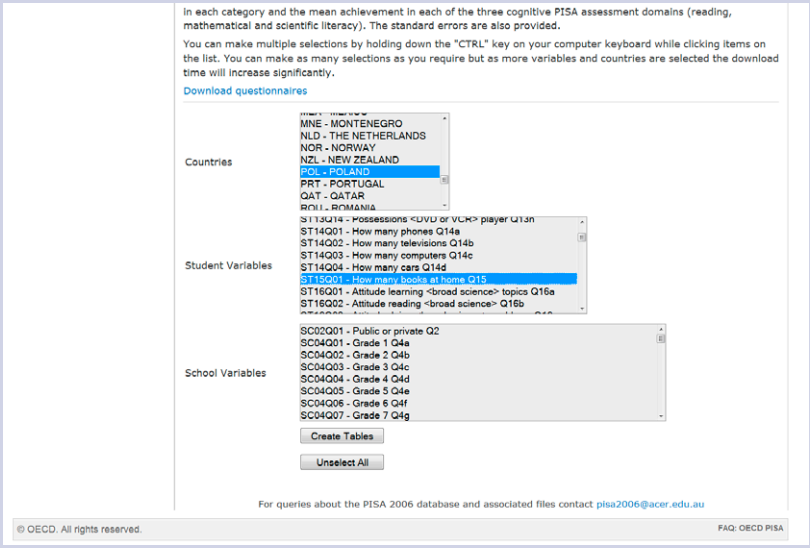
Najpierw musimy zdecydować, z którego roku chcemy analizować dane. W tej chwili do wyboru mamy trzy możliwości (w lewej części strony, rysunek 5.1 – rok 2000, 2003 i 2006, a w kolejnych latach dodawane będą wyniki nowych cykli). Po decyzji, z którego roku dane chcemy wykorzystać, musimy zdecydować, jak chcemy analizować dane z wybranego roku. Witryna oferuje trzy możliwości dostępu do informacji zawartych w bazach danych.

Jako pierwszą można wybrać opcję: *Download the PISA 2006 data set – for professional researchers,* która umożliwia ściągnięcie całego zbioru i przeprowadzenie analiz za pomocą własnego programu statystycznego. Jest to opcja rekomendowana dla zawansowanych użytkowników. Zachęcamy jednak do odwiedzenia tej zakładki nawet użytkowników nieuważających się za profesjonalnych badaczy. Oprócz danych z tego badania dostępne są formy kwestionariuszy wypełnianych przez respondentów, książki kodowe, kompendia, w których znajdują się podstawowe informacje o wynikach, i materiały techniczne, do których należy zajrzeć w miarę potrzeby. Zawartość zakładki *Download the PISA 2006 data set – for professional researchers* przedstawiona została na rysunku 5.2.

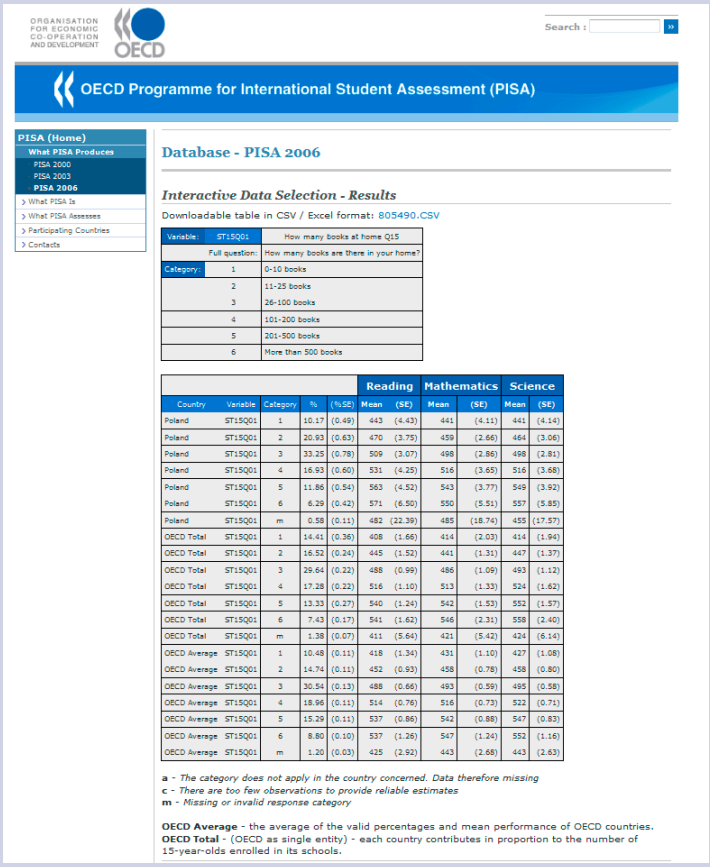
Osobom zainteresowanym analizą zmiennych z tej bazy polecamy zapoznanie się z kwestionariuszami tak, aby poznać, jakie pytania stoją za poszczególnymi zmiennymi, które mogą być później analizowane. Przejrzenie kwestionariusza jest także dobrym sposobem na szybkie zorientowanie się, jakich informacji możemy się spodziewać po analizowanym przez nas zbiorze danych.

Drugi sposób na dostanie się do zasobów zawartych w analizowanym zbiorze danych to wybór zakładki: *Interactive data selection – to explore the PISA 2006 dataset*, ze strony głównej witryny. Ta strona umożliwia obliczanie wyników dla trzech umiejętności mierzonych przez badanie PISA w dowolnie zdefiniowanych przez użytkownika grupach (patrz rysunek 5.3. *Interactive data selection – to explore the PISA 2006 dataset*). Mamy tutaj do dyspozycji trzy okna. W pierwszym oknie wybieramy kraj, w którym chcemy przeprowadzić analizę, w drugim i trzecim oknie możemy wybrać zmienną, na podstawie której utworzone zostaną grupy. Przy czym w drugim oknie znajdziemy zmienne charakteryzujące uczniów (wykształcenie rodziców, status społeczny, nastawienie do nauki etc.), a w trzecim zmienne charakteryzujące szkoły (zasoby materialne szkoły, informacje o nauczycielach etc.). Na rysunku 5.3 można zobaczyć przykład, w którym wybraliśmy Polskę – jako kraj, dla którego wykonywać będziemy obliczenia – oraz zmienną charakteryzującą ucznia: liczba książek w domu.

Rezultaty, jakie otrzymamy z tak zaprojektowanej analizy, przedstawiono na rysunku 5.4. W pierwszej tabelce dostajemy opis zmiennej, którą wybraliśmy do grupowania. Pod nazwą zmiennej (*How many books at home Q15* – Ile książek w domu) otrzymujemy informację, jak dokładnie brzmiało pytanie (*How many books are there in your home?* – Ile książek znajduje się w twoim domu?) oraz możliwe do wyboru kategorie (1 – od 0 do 10 książek; 2 – od 11 do 25 książek; etc.). Natomiast kategoria *m* oznacza brak danych (od angielskiego *missing*).



Rysunek 5.3. Okno *Interactive data selection – to explore the PISA 2006 dataset*



ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT
OECD
Search :

OECD Programme for International Student Assessment (PISA)

PISA (Home)
What PISA Produces
PISA 2000
PISA 2003
PISA 2006
What PISA Is
What PISA Assesses
Participating Countries
Contacts

Database - PISA 2006

Interactive Data Selection - Results

Downloadable table in CSV / Excel format: 805490.CSV

Variable:	ST15Q01	How many books at home Q15
	Full question:	How many books are there in your home?
Category:	1	0-10 books
	2	11-25 books
	3	26-100 books
	4	101-200 books
	5	201-500 books
	6	More than 500 books

					Reading		Mathematics		Science	
Country	Variable	Category	%	(%SE)	Mean	(SE)	Mean	(SE)	Mean	(SE)
Poland	ST15Q01	1	10.17	(0.49)	443	(4.43)	441	(4.11)	441	(4.14)
Poland	ST15Q01	2	20.93	(0.63)	470	(3.75)	459	(2.66)	464	(3.06)
Poland	ST15Q01	3	33.25	(0.78)	509	(3.07)	498	(2.86)	498	(2.81)
Poland	ST15Q01	4	16.93	(0.60)	531	(4.25)	516	(3.65)	516	(3.68)
Poland	ST15Q01	5	11.86	(0.54)	563	(4.52)	543	(3.77)	549	(3.92)
Poland	ST15Q01	6	6.29	(0.42)	571	(6.50)	550	(5.51)	557	(5.85)
Poland	ST15Q01	m	0.58	(0.11)	482	(22.39)	485	(18.74)	455	(17.57)
OECD Total	ST15Q01	1	14.41	(0.36)	408	(1.66)	414	(2.03)	414	(1.94)
OECD Total	ST15Q01	2	16.52	(0.24)	445	(1.52)	441	(1.31)	447	(1.37)
OECD Total	ST15Q01	3	29.64	(0.22)	488	(0.99)	486	(1.09)	493	(1.12)
OECD Total	ST15Q01	4	17.28	(0.22)	516	(1.10)	513	(1.33)	524	(1.62)
OECD Total	ST15Q01	5	13.33	(0.27)	540	(1.24)	542	(1.53)	552	(1.57)
OECD Total	ST15Q01	6	7.43	(0.17)	541	(1.62)	546	(2.31)	558	(2.40)
OECD Total	ST15Q01	m	1.38	(0.07)	411	(5.64)	421	(5.42)	424	(6.14)
OECD Average	ST15Q01	1	10.48	(0.11)	418	(1.34)	431	(1.10)	427	(1.08)
OECD Average	ST15Q01	2	14.74	(0.11)	452	(0.93)	458	(0.78)	458	(0.80)
OECD Average	ST15Q01	3	30.54	(0.13)	488	(0.66)	493	(0.59)	495	(0.58)
OECD Average	ST15Q01	4	18.96	(0.11)	514	(0.76)	516	(0.73)	522	(0.71)
OECD Average	ST15Q01	5	15.29	(0.11)	537	(0.86)	542	(0.88)	547	(0.83)
OECD Average	ST15Q01	6	8.80	(0.10)	537	(1.26)	547	(1.24)	552	(1.16)
OECD Average	ST15Q01	m	1.20	(0.03)	425	(2.92)	443	(2.68)	443	(2.63)

a - *The category does not apply in the country concerned. Data therefore missing*
c - *There are too few observations to provide reliable estimates*
m - *Missing or invalid response category*

OECD Average - the average of the valid percentages and mean performance of OECD countries.
OECD Total - (OECD as single entity) - each country contributes in proportion to the number of 15-year-olds enrolled in its schools.

Rysunek 5.4. Wyniki PISA dla polskich uczniów o różnych zasobach książek

Druga tabela w oknie wyników jest podstawowym elementem analizy. Warto na początku zwrócić uwagę, iż statystyki obliczane są w trzech wersjach – co oznaczone jest w pierwszej kolumnie tabeli. Po pierwsze, dla wybranego kraju (*Poland*). Po drugie, dla wszystkich uczniów z badanych krajów OECD jednocześnie (*OECD Total*). W tym wypadku mamy do czynienia ze średnią wyników wszystkich badanych uczniów. Po trzecie, jako wynik średni kraju OECD (*OECD Average*). Tutaj mamy do czynienia ze średnią średnich wyników w krajach OECD – najpierw dla każdego kraju liczona jest średnia, a potem te średnie są dodawane i dzielone przez liczbę krajów OECD, czyli obliczana jest średnia ze średniej. Pierwszy wynik średni możemy zatem interpretować jako wyniki średnie uczniów dla całego OECD, drugi jako średnie wyniki dla przeciętnego kraju OECD.

W wytworzonej przez system tabeli odnajdujemy informację o rozkładzie procentowym kategorii (%) – czyli, dla przykładu, widzimy, że około 10,17% ucz-

spis treści

niów w Polsce pochodzi z domów, w których nie ma więcej niż 10 książek. W kolejnej kolumnie poznajemy błąd standardowy dla wyliczonych proporcji (*% SE*) mówiący nam, jak dokładne jest oszacowanie umieszczone we wcześniejszej kolumnie (czyli 10,17%). W naszym przypadku dla kategorii od 0 do 10 książek w Polsce wynosi ono 0,49 – jest to nieznaczna liczba w porównaniu z 10,17%, zatem możemy stwierdzić, iż oszacowanie tego procentowego udziału uczniów jest dosyć dokładne. Skoro błąd standardowy wynosi 0,49, to możemy twierdzić, że przedział (9,21; 11,13) zapewne zawiera prawdziwy wynik.[44]

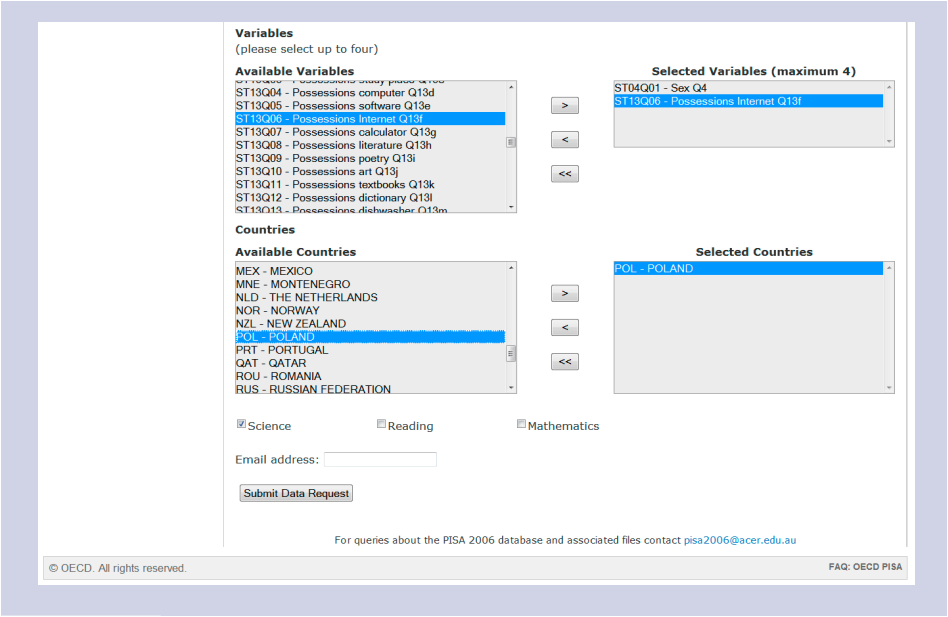
W kolejnych sześciu kolumnach podawane są wyniki testów PISA dla trzech umiejętności: czytania (*reading*), matematyki (*mathematics*) i nauk przyrodniczych (*science*). Dla każdej umiejętności podane są średnie wyniki (*mean*) i błędy standardowe (*SE*). Jak widać w tabeli 5.4, odkryliśmy silny związek między liczbą książek a wynikiem testów PISA. Im więcej książek w domu dziecka, tym, średnio rzecz biorąc, jego wynik jest wyższy. Taka zależność dostrzegana jest zarówno w Polsce, jak i w innych krajach OECD. Uczniowie w Polsce, w których domach jest nie więcej niż 10 książek, uzyskują średni wynik z czytania 443, z matematyki 441 i z nauk przyrodniczych 441. Natomiast ich koledzy, w których domach znajduje się ponad 500 książek, otrzymują kolejno: z czytania 571, z matematyki 550 oraz z nauk przyrodniczych 557. Różnice te są duże i znaczące statystycznie[45].

Oczywiście nie oznacza to, że książki w magiczny sposób zwiększają umiejętności uczniów, posiadanie książek przez rodziców również nie musi oznaczać tego, że uczniowie czytają te książki. Odkryliśmy w tej chwili pewną zależność statystyczną, którą należałoby dalej wyjaśniać. Możemy na tym etapie postawić jedynie pewne hipotezy, na przykład, że domy, w których jest więcej książek, to domy bogatsze, gdzie rodzice przywiązują większą wagę do edukacji swoich dzieci, a zatem motywują dzieci do uczenia się, lub rodzice mający dużo książek to rodzice wykształceni, którzy mogą pomagać uczniom w nauce, mogą ich uczyć sami etc.

Trzecią możliwością analizowania danych z witryny pisa2006.acer.edu.au jest użycie modułu: *Multi-dimensional Data Request* (prośba o dane wielowymiarowe). Dzięki tej opcji użytkownik może analizować bardziej złożone problemy, używając analizy wielozmiennowej, czyli takiej, w której analizuje się kilka

[44] Dokładniej, w 95% przypadków tak obliczony przedział ufności zawierać będzie prawdziwy wynik. Obliczamy to ze wzoru (x-1,96 × SE; x+1,96 × SE), gdzie x to estymowany wynik, 1,96 to stała obliczona dla rozkładu normalnego, a SE to błąd standardowy. Jeżeli chcielibyśmy obliczyć przedział ufności dla 90%, powinniśmy użyć stałej 1,64.

[45] Tak samo jak dla obliczenia przedziałów ufności dla udziału procentowego, możemy dla poszczególnych wyników obliczyć przedziały ufności: (x-1,96 × SE; x+1,96 × SE). Jeżeli Czytelnik to zrobi dla poszczególnych wyników testów umiejętności PISA dla różnych kategorii liczby książek, odkryje, że w zdecydowanej większości nie pokrywają się one ze sobą. Na przykład wynik z czytania dla pierwszej kategorii (0-10 książek) zawiera się w przedziale (434; 451) przy ustalonym 95-procentowym poziomie ufności, wynik z ostatniej kategorii (powyżej 500 książek) w przedziale (558; 583).



Rysunek 5.5. Okna dialogowe w zakładce: *Multi-dimensional Data Request*

zmiennych wyjaśniających jednocześnie. Przejdźmy do prostego przykładu. Na rysunku zamieszczone zostały okna dialogowe, które pojawiają się po otwarciu zakładki *Multi-dimensional Data Request*. W pierwszym oknie z lewej strony mamy podane zmienne, których możemy użyć w analizie. Jednocześnie możemy wybrać do 4 zmiennych. Wybór jest niezwykle prosty, wybieramy kursorem myszy interesującą nas zmienną, następnie naciskamy przycisk strzałki umieszczony po prawej stronie kursora. W przykładzie prezentowanym poniżej wybraliśmy dwie zmienne (okno wybranych zmiennych to pierwsze okno z prawej strony): płeć (*SEX Q4*) i posiadanie dostępu do Internetu (*Possessions Internet Q13f*).

W dolnym lewym oknie wybraliśmy kraj, jaki nas interesuje. Wybraliśmy Polskę i – podobnie jak w wypadku wyboru zmiennych – myszką nacisnęliśmy przycisk strzałki z prawej strony okna, tak aby oznaczenie Polski znalazło się w prawym dolnym oknie wybranych krajów (można zaznaczyć kilka krajów). Teraz pozostało nam zdecydować, czy chcemy analizować wyniki z nauk przyrodniczych (*Science*), czytania (*Reading*) czy matematyki (*Mathematics*). W naszym wypadku wybraliśmy jedynie nauki przyrodnicze. Aby otrzymać interesujące nas wyniki, musimy podać swój adres e-mailowy. Po wprowadzeniu adresu i jego akceptacji wyniki pojawią się na naszej skrzynce pocztowej w ciągu kilku minut. Zostają one dostarczone w postaci pliku *Excel*. Plik ten zawiera dwie zakładki. W pierwszej zakładce znajdziemy opis stworzonych

Tabela 5.1. Pierwsza zakładka z pliku, który przesyła portal PISA – opis kategorii (tłumaczenie własne)

Variable (Zmienna)	*Variable Description* (Opis zmiennej)	*Category* (Kategoria)	*Category Label* (Etykieta Kategorii)
ST04Q01	*Are you female or male*? (Czy jesteś kobietą, czy mężczyzną?)	1	*Female* (Kobieta)
ST04Q01	*Are you female or male*? (Czy jesteś kobietą, czy mężczyzną?)	2	*Male* (Mężczyzna)
ST13Q06	*A link to the Internet* (Podłączenie do Internetu)	1	*Yes* (Tak)
ST13Q06	*A link to the Internet* (Podłączenie do Internetu)	2	*No* (Nie)
		7	*N/A* (Nie dotyczy)
		8	*Invalid* (Błędnie wypełnione)
		9	*Missing* (Brak danych)

wedle naszej prośby kategorii. Dla prezentowanego przykładu otrzymany opis kategorii prezentujemy w tabeli 5.1.

Serwis podaje nam opis zmiennych, które wybrane zostały do analizy. W pierwszej kolumnie mamy nazwę zmiennej (będącą jednocześnie numerem pytania w kwestionariuszu), w drugiej kolumnie opis zmiennej (niekiedy treść pytania). W naszym wypadku pierwsza zmienna charakteryzuje płeć: „Czy jesteś kobietą, czy mężczyzną?". Druga dotyczy posiadania podłączenia do Internetu. W trzeciej kolumnie znajdują się numeryczne kategorie odpowiedzi, a w czwartej ich słowne reprezentacje. Zatem kategoria 1 dla pytania o płeć oznacza kobietę, kategoria 2 mężczyznę. W wypadku pytania o podłączenie do Internetu kategoria 1 oznacza odpowiedź twierdzącą, a kategoria 2 przeczącą. Pozostałe kody 7, 8 i 9 oznaczają różne sytuacje, w których nie udzielono odpowiedzi.

Ta zakładka jest nam niezbędna, abyśmy mogli przeanalizować wyniki umieszczone w zakładce drugiej nadesłanego nam pliku. Przykład wyników zawartych w drugiej zakładce przedstawiamy w tabeli 5.2.[46]

W pierwszej kolumnie podany został kraj, w którym przeprowadzona została analiza (oczywiście jest to Polska, którą wybraliśmy). Następne dwie kolumny charakteryzują nam grupy, które zostały stworzone na naszą prośbę.

[46] Z tabeli dla przejrzystości usunięte zostały kategorie oznaczające braki danych.

spis treści

Tabela 5.2. Druga zakładka z pliku, który przesyła portal PISA – wyniki analiz

Country (Kraj)	ST04Q01	ST13Q06	*N*	*%*	*% SE*	*Mean*	*SE*
Poland (Polska)	1	1	1411	24,2	0,83	517	3,16
Poland (Polska)	1	2	1354	25,0	0,82	478	3,05
Poland (Polska)	2	1	1455	25,8	0,68	518	3,45
Poland (Polska)	2	2	1191	22,5	0,66	480	3,10

Stworzone zostały cztery grupy, ponieważ wybraliśmy dwie zmienne mające po dwie kategorie (płeć: mężczyzna, kobieta; podłączenie do Internetu: tak, nie). W pierwszym (po nagłówku) wierszu mamy zatem do czynienia z następującymi grupami uczniów: kobiety posiadające dostęp do Internetu (1,1), w drugim wypadku: kobiety nieposiadające dostępu do Internetu (1,2), w trzecim: mężczyźni posiadający dostęp do Internetu (2,1), a w ostatniej – mężczyźni nieposiadający dostępu do Internetu (2,2).

Kolejne kolumny zawierają informacje o liczebności kategorii w próbie (*N*), estymowanym udziale procentowym danej kategorii w populacji (*%*), procentowym błędzie standardowym tego oszacowania (*% SE*), szacowanym średnim wyniku z nauk przyrodniczych dla tej grupy (*Mean*) oraz błędzie standardowym tego oszacowania (*SE*). Jak widać, w grupie kobiet posiadających podłączenie do Internetu (1,1) w badaniu znalazło się 1411 respondentów, co odpowiada około 24% populacji 15-latków. Błąd oszacowania tego udziału jest niewielki, bo wynosi jedynie 0,83. Natomiast średni wynik z zadań przyrodniczych wynosi 517 przy błędzie standardowym 3,16.

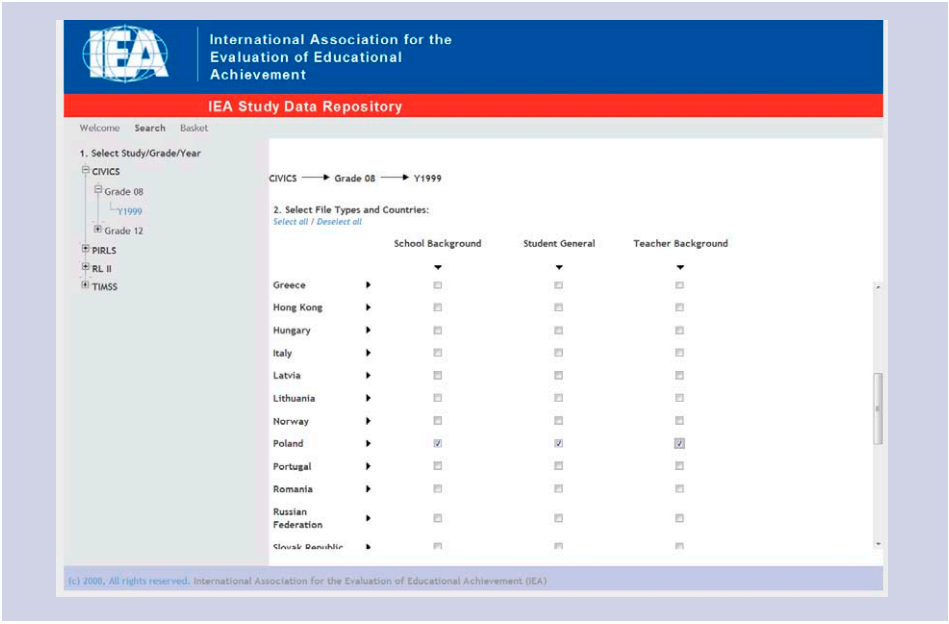
Gdy przeanalizujemy wszystkie wyniki zawarte w tabeli 5.2, widać wyraźnie, iż posiadanie lub nieposiadanie dostępu do Internetu jest silnie związane z uzyskanymi wynikami, a płeć nie odgrywa tu ważnej roli. Chłopcy i dziewczynki posiadający lub nieposiadający w domu podłączenia do Internetu uzyskują bardzo podobne wyniki (statystycznie nierozróżnialne), jest ich też podobna liczba w każdej kategorii. Różnice uwidaczniają się dopiero, gdy porównujemy wyniki grupy posiadającej podłączenie do Internetu i nieposiadającej takiego podłączenia.

Jak widać, nawet bez skomplikowanego oprogramowania statystycznego można z powodzeniem przeprowadzać nawet dosyć złożone analizy statystyczne. Zachęcamy do eksplorowania tych danych, z których niewątpliwie da się uzyskać wiele ciekawych informacji.

Badanie CivEd (http://www.iea.nl/cived.html)

Badanie CivEd i najważniejsze jego wyniki przedstawione zostały w rozdziale 4. Organizatorzy badania udostępnili zbiory zebranych przez siebie danych na stronie http://rms.iea-dpc.org/#. Podobnie jak w wypadku badań PISA, znajdziemy tam dokumentację techniczną, książki zmiennych (*codebooks*) oraz same zbiory danych: plik z danymi zebranymi od ucznia, pliki z badań nauczycieli oraz z charakterystykami szkół. Dane te są dostępne w osobnych plikach dla każdego z uczestniczących w badaniu krajów.

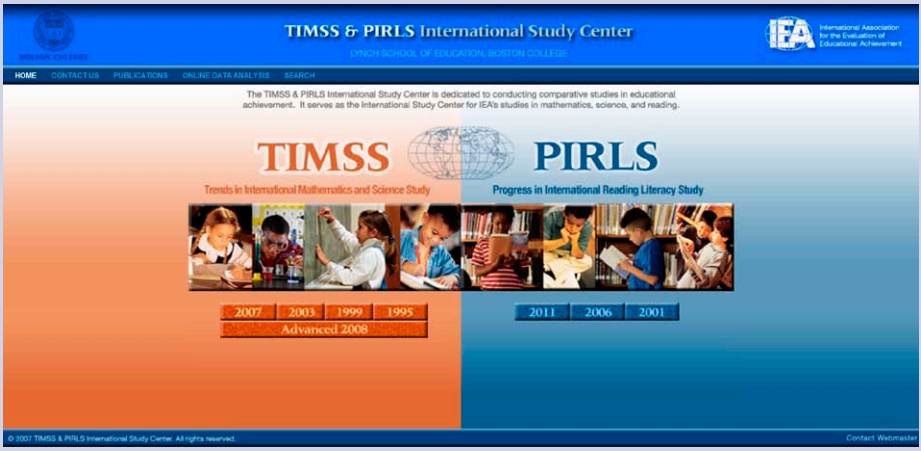
Niestety, witryna daje nam jedynie dostęp do bazy danych w postaci plików gotowych do odczytu za pomocą programów statystycznych (takich jak SPSS czy SAS). Nie zostały tutaj przygotowane żadne narzędzia, które umożliwiałyby analizowanie wyników on-line. IEA opracowała co prawda darmowy program do analizy danych (który można znaleźć na ich stronie), lecz jest to program skierowany do zaawansowanych użytkowników.



Rysunek 5.6. Witryna, z której można ściągnąć dane dotyczące badania CivEd (http://rms.iea-dpc.org/#)

Badania PIRLS i TIMSS (http://timss.bc.edu/

W rozdziale 4 omawialiśmy pokrótce badania PIRLS 2006, czyli badania umiejętności czytania uczniów czwartych klas szkół podstawowych. W rozdziale



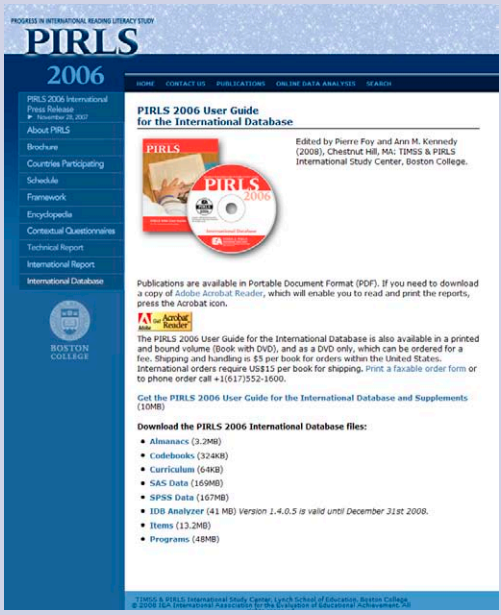
Rysunek 5.7. Strona główna badań TIMSS i PIRLS (http://timss.bc.edu)

tym wzmiankowaliśmy jedynie o analogicznym projekcie TIMSS realizowanym również przez IEA. O ile celem badania PIRLS jest porównanie kompetencji związanych z umiejętnością czytania, to celem badania TIMSS (*Trends in International Mathematics and Science Study*) jest porównywanie zdolności matematycznych oraz wiedzy i umiejętności z zakresu nauk przyrodniczych uczniów klas czwartych i ósmych. PIRLS oraz TIMSS przypominają w swojej konstrukcji badanie PISA z tym, że badaniu podlegają uczniowie młodsi, a i metodologia, założenia, cele badań oraz uczestnicząca grupa krajów są nieco odmienne.

Po wyborze badań i roku ich realizacji (rysunek 5.7) pojawia się strona poświęcona wybranemu cyklowi badania. Aby przejść do zakładki, w której znajdują się zbiory danych, wybieramy zakładkę *International data base* (ostatnia pozycja na liście z lewej strony). Po jej wyborze pojawia się strona, z której można pobrać dokumentację techniczną, książkę kodów oraz interesujące nas bazy danych w formacie plików obsługiwanych przez profesjonalne pakiety statystyczne SPSS oraz SAS (patrz rysunek 5.8). Strony pozwalające na pobieranie danych z badania TIMSS I PIRLS działają w taki sam sposób.

Realizatorzy badań, podobnie jak w wypadku badania PISA, umożliwiają użytkownikowi strony internetowej analizę danych on-line. Niestety, do tej pory dostępne są jedynie dane do 2001 roku, czyli badania, w których Polska nie brała jeszcze udziału. Aby dostać się do modułu analizy danych on-line, na stronie głównej badań należy wybrać przycisk *online data analysis,* po naciśnięciu którego zostajemy przekierowani na stronę http://lighthouse.air.org/timss/. Dzięki tej stronie będziemy mogli przeprowadzić proste analizy statystyczne.

spis treści

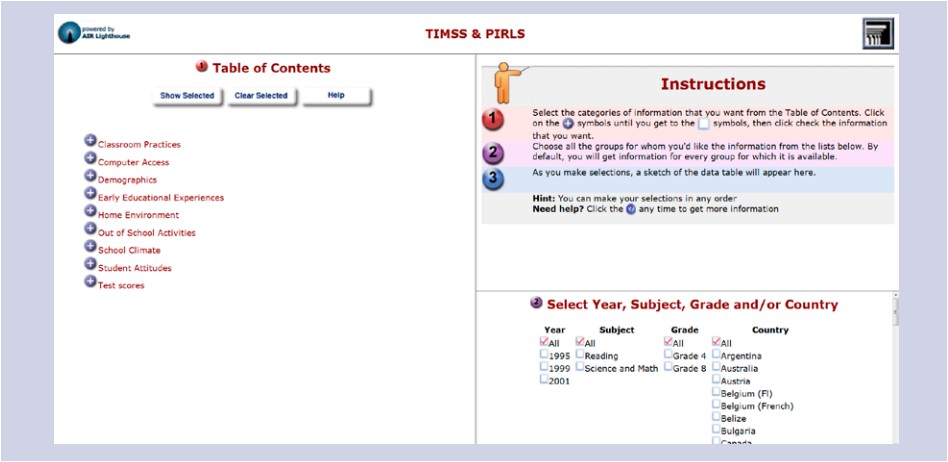


Rysunek 5.8.
Zakładka strony badań PIRLS 2006 (http://timss.bc.edu), z której możemy pobrać pliki zawierające dane z badania PIRLS 2006 (strona z badaniami TIMSS wygląda analogicznie)

By przystąpić do analizy, w dolnym prawym oknie (patrz rysunek 5.9) musimy wybrać rok badania (*Year*), interesujący nas przedmiot (*Subject*), poziom nauczania (*Grade*) oraz kraj (*Country*), w którym chcemy przeprowadzić nasze analizy. Niestety, jak pisaliśmy już wcześniej, w portalu dostępne są jedynie dane do roku 2001, czyli nie znajdziemy w nich Polski, która nie brała udziału w tych wczesnych edycjach badania. W przykładzie prezentowanym w tej książce wybraliśmy badania TIMSS z 1999 roku, uczniów po ośmiu latach kształcenia (*Grade 8*), a krajem, w którym będziemy przeprowadzać analizy, są Czechy.

W lewej części okna (patrz rysunek 5.9) znajduje się lista kategorii, po rozwinięciu której będziemy mogli dostać się do interesujących nas zmiennych: *Classroom Practices* (metody nauczania), *Computer Access* (dostęp do komputera), *Demographics* (zmienne demograficzne), *Early Educational Experiences* (wczesne doświadczenia edukacyjne), *Home Environment* (środowisko domowe), *Out of School Activities* (zajęcia pozaszkolne), *School Climate* (klimat szkolny), *Student Attitudes* (postawy uczniów) i *Test scores* (wyniki testów).

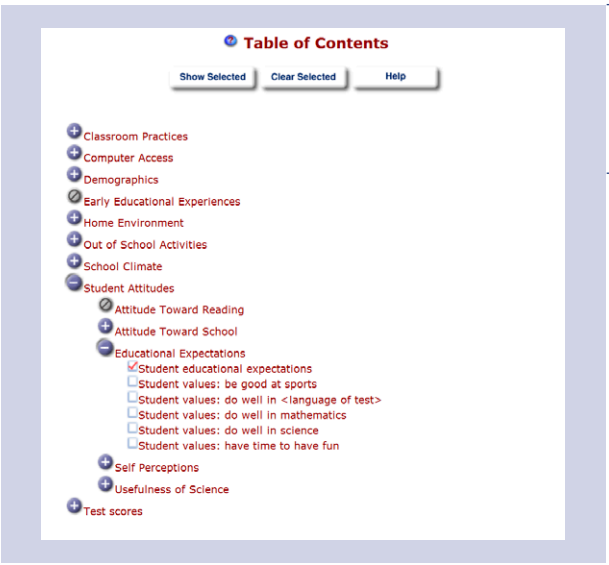
Jako że interesowały nas postawy uczniów, rozwinęliśmy zakładkę *Student Attitudes* (klikając na znaczek plus przy tej kategorii). Następnie wybraliśmy zakładkę zawierającą zmienne o oczekiwaniach edukacyjnych (*Educational Expectations)*, by w końcu wybrać interesującą nas zmienną opisującą oczekiwania edukacyjne uczniów (*Students' Educational Expectations).*



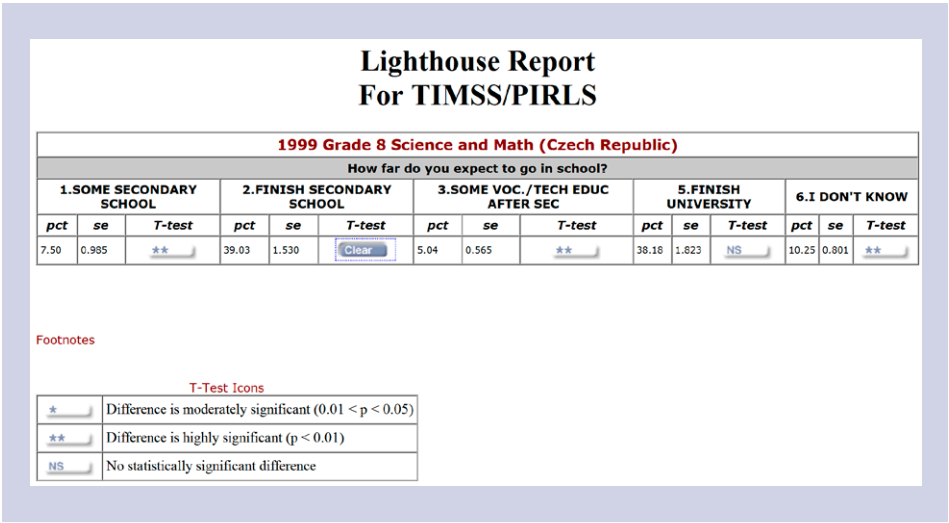
Rysunek 5.9. Strona umożliwiająca analizę danych PIRLS i TIMSS (http://lighthouse.air.org/timss/)

Nawet po tak pobieżnym przejrzeniu tematów Czytelnik może zorientować się, że odnalezienie wielu zmiennych opisujących interesujące problemy jest bardzo proste.

Po wyborze zmiennych możemy przystąpić do analizy na dwa sposoby. Możemy stworzyć tabelę (naciskając na przycisk *Create table*, w prawej górnej części okna) lub wykres słupkowy (naciskając na przycisk *Create graph*, w prawej górnej części okna).



Rysunek 5.10. Wybór zmiennych na stronie umożliwiającej analizę danych PIRLS i TIMSS (http://lighthouse.air.org/timss)



Lighthouse Report For TIMSS/PIRLS

1999 Grade 8 Science and Math (Czech Republic)

How far do you expect to go in school?

1.SOME SECONDARY SCHOOL			2.FINISH SECONDARY SCHOOL			3.SOME VOC./TECH EDUC AFTER SEC			5.FINISH UNIVERSITY			6.I DON'T KNOW		
pct	*se*	*T-test*	*pct*	*se*	*T-test*	*pct*	*se*	*T-test*	*pct*	*se*	*T-test*	*pct*	*se*	*T-test*
7.50	0.985	**	39.03	1.530	Clear	5.04	0.565	**	38.18	1.823	NS	10.25	0.801	**

Footnotes

T-Test Icons

*	Difference is moderately significant ($0.01 < p < 0.05$)
**	Difference is highly significant ($p < 0.01$)
NS	No statistically significant difference

Rysunek 5.11. Tabela wygenerowana przez system analizy on-line przedstawiająca rozkład odpowiedzi dotyczących aspiracji edukacyjnych uczniów (http://lighthouse.air.org/timss)

Zacznijmy od stworzenia tabeli. Po przyciśnięciu przycisku „stwórz tabelę" (*Create table*) w nowym oknie pokazuje się tabela, w której zawarte są wyniki odpowiedzi uczniów na temat ich oczekiwań edukacyjnych. W szarym polu tabeli (patrz rysunek 5.11) podana została dokładna treść pytania, jakie zostało zadane uczniom: Twoim zdaniem, jakie wykształcenie uzyskasz? (*How far do you expect to go in school?*).

Poniżej szarego pola znajdują się kategorie, które mogli wybrać uczniowie: 1. Nieukończone średnie (*Some secondary school*), 2. Ukończone średnie (*Finish secondary school*), 3. Dalsze kształcenie zawodowe po szkole średniej (*Some Voc./Tech educ after sec*), 4. Ukończone wyższe (*Finish University*), 5. Nie wiem (*I don't know*). Pod każdą z tych kategorii znajdują się trzy pola. W pierwszym z nich podawany jest procent zaznaczeń danej kategorii (*pct*), w drugiej błąd standardowy (*se*). Jak widać na rysunku 5.11, naukę w szkole średniej, ale bez jej ukończenia, deklarowało 7,5% czeskich uczniów, przy błędzie standardowym 0,985; skończenie szkoły średniej 39,03%, przy błędzie standardowym 1,530; dalsze kształcenie zawodowe po szkole średniej 5,04%, przy błędzie standardowym 0,565; ukończenie szkoły wyższej 38,18%, przy błędzie standardowym 1,823; a 10,25% czeskich uczniów nie potrafiło odpowiedzieć na to pytanie.

Pod każdą z kategorii odpowiedzi znajduje się jeszcze trzecie pole o nazwie *t-test*. W przeciwieństwie do pozostałych pól, jest to pole interaktywne. Gdy naciśniemy przycisk znajdujący się pod tym polem, uruchomimy ocenę statystycznej istotności na podstawie *t-testu*. W przykładzie prezentowanym na

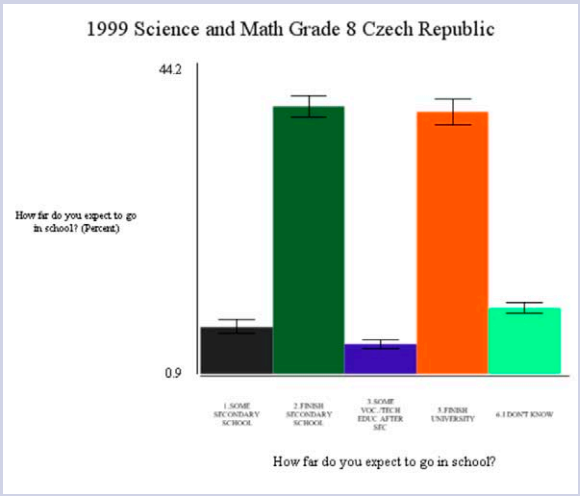
rysunku 5.11 nacisnęliśmy przycisk w polu *t-test* pod kategorią „Ukończone średnie". Pozwala to sprawdzić, czy odsetek uczniów, który wybrał tę odpowiedź, jest znacząco większy lub znacząco mniejszy niż w wypadku innych kategorii.

W chwili naciśnięcia przycisku pod polem *t-test* w kategorii „Ukończone średnie", pod innymi kategoriami pojawią się oznaczenia symbolizujące istotność różnic: ** – dwie gwiazdki oznaczają różnice istotne przy poziomie istotności 95% (czyli w 95% przypadków wnioskowanie nasze jest poprawne[47]), * – jedna gwiazdka oznacza różnice istotne przy poziomie istotności 90% (czyli w 90% przypadków wnioskowanie nasze jest poprawne), oznaczenie *NS* mówi nam o tym, iż różnice nie są istotne statystycznie.

W naszym przykładzie widać wyraźnie, iż odpowiedź „Ukończone średnie" (2.) była zdecydowanie częściej wybierana niż odpowiedzi 1., 3. i 6. (pod tymi kategoriami obserwujemy dwie gwiazdki oznaczające istotne różnice). Natomiast nie ma istotnej różnicy między procentem uczniów deklarujących zakończenie edukacji na szkole średniej i na etapie szkoły wyższej (przy kategorii 5. ukończenie szkoły wyższej widnieje napis *NS*). Tutaj porównywaliśmy kategorię „Ukończone średnie" (była to nasza kategoria odniesienia), z pozostałymi kategoriami. Oczywiście nic nie stoi na przeszkodzie, aby wybrać inną kategorię odniesienia (referencyjną). Wystarczy kliknąć na pole *t-test* pod inną kategorią.

Drugim sposobem analizy, jaki umożliwia portal, jest analiza za pomocą wykresu (po wyborze zmiennych należy wybrać opcję *Create graph*). Przykład takiej graficznej prezentacji wyników umieszczony został na rysunku 5.12. Jest to klasyczny wykres słupkowy, który prezentuje nam rozkład odpowiedzi uczniów (ten sam, co w wypadku tabeli). Dodatkowo na wykresie przedstawione są przedziały ufności odpowiedzi, reprezentowane przez oznaczenia przypominające odwrócone o 90 stopni wielkie litery H. Boki tej odwróconej litery zawierają 95-procentowe przedziały ufności szacowane na podstawie próby. Na podstawie naszych szacowań nie możemy bezbłędnie stwierdzić, że 38,18% uczniów w całej populacji uważa, iż uzyska wykształcenie wyższe. Ale możemy powiedzieć, że z 95-procentowym poziomem ufności w całej populacji wynik ten oscyluje w przedziałach wyznaczonych na rysunku. Warto zauważyć, iż po analizie wykresu dochodzimy do podobnych konkluzji, co po analizie tabeli: procent uczniów oczekujących ukończenia szkoły wyższej jest podobny do procentu uczniów oczekujących ukończenia szkoły średniej – przedziały ufności tych dwóch kategorii częściowo pokrywają się. Nie możemy zatem powiedzieć, czy więcej uczniów oczekuje ukończenia uniwersytetu, czy szkoły średniej. Statystycznie

[47] Precyzyjniej mówiąc, w 95% przypadków podejmujemy właściwą decyzję, odrzucając hipotezę zerową na rzecz hipotezy konkurencyjnej. W tym wypadku hipoteza zerowa mówi nam, iż procent wyborów dwóch kategorii jest taki sam. Hipoteza konkurencyjna mówi o tym, że procent odpowiedzi na dwie kategorie jest różny.



Rysunek 5.12.
Wykres wygenerowany przez system analizy on-line przedstawiający rozkład odpowiedzi dotyczących aspiracji edukacyjnych uczniów (http://lighthouse.air.org/timss)

rzecz biorąc, nie ma istotnej różnicy między tymi wynikami. Możemy jednak stwierdzić, że znacznie mniej czeskich uczniów nie zamierza ukończyć szkoły średniej, kontynuować nauki zawodowej po szkole średniej lub też nie potrafi odpowiedzieć na to pytanie.

5.2. Polskie bazy danych (www.cke-efs.pl)

Polskie bazy danych, które prezentujemy w tej części, dostępne będą na stronie www.cke-efs.pl. Na razie dane dostępne są wyłącznie w postaci plików obsługiwanych przez specjalistyczne programy statystyczne (SPSS i Stata), lecz w przyszłości witryny obsługujące te dane będą rozbudowywane w taki sposób, aby na wzór badań PISA można było za pośrednictwem strony internetowej nie tylko eksplorować bazę danych, ale także przeprowadzać za pośrednictwem Internetu nawet bardziej złożone analizy. W tej części przedstawiamy opis i strukturę tych baz danych.

Baza Pentor-CKE

Opublikowane bazy danych powstały dzięki badaniom zleconym przez Centralną Komisję Egzaminacyjną w ramach projektu: *Badania dotyczące wyników egzaminów zewnętrznych*. Projekt ten realizowany był w latach 2004-2006 przy współudziale Europejskiego Funduszu Społecznego i objął swym zasięgiem liczne badania związane między innymi ze: zrównywaniem wyników egzaminów,

komunikowaniem wyników z użyciem wskaźnika edukacyjnej wartości dodanej, wykorzystywaniem wyników przez szkoły itp.[48]

Pentor RI z zespołem ekspertów zaplanował, a następnie przeprowadził badania w gimnazjach. Dane zostały zebrane i zapisane w postaci elektronicznej. Pentor RI przygotował też pierwsze wersje wyników. W dalszej fazie nad wynikami badań pracowało kilkunastu ekspertów, uczestniczących także w fazie planowania badania. Zwieńczeniem tych prac była seria artykułów, które ukazały się w trzech kolejnych numerach Biuletynu Badawczego Centralnej Komisji Egzaminacyjnej „Egzamin": 11/2007; 12/2007; 13/2007. Szczegóły dotyczące badania przytaczaliśmy już w rozdziale 4. Bogactwo danych i łączące się z tym ogromne możliwości analizowania różnorodnych problemów związanych z polską szkołą i systemem egzaminacyjnym okazały się tak duże, iż zdecydowano się na udostępnienie bazy danych szerszej rzeszy odbiorców tak, aby jej potencjał został w pełni wykorzystany.

Publikacja bazy danych wymagała zintegrowania jej poszczególnych części, pracy nad jej „wyczyszczeniem", stworzenia spójnego systemu etykietowania, precyzyjnego opisu poszczególnych elementów oraz przeważenia zbioru tak, aby poszczególne jego pytania mogły być używane w celach opisowych i traktowane jako reprezentatywne dla populacji uczniów trzecich klas gimnazjów. Zwieńczeniem prac jest na nowo przygotowana baza danych wraz z pełną dokumentacją, która ma być udostępniona w Internecie.

W ramach badania przebadano dyrektorów, nauczycieli, rodziców i uczniów szkół gimnazjalnych. Dane mają charakter ilościowy. Zebrane zostały za pomocą kwestionariuszy audytoryjnych (w wypadku uczniów) i kwestionariuszy do samodzielnego wypełnienia (w wypadku pozostałych badanych). Największym modułem badawczym w tym badaniu było badanie uczniów, które można podzielić na dwie części. Część pierwszą wypełniali wszyscy uczniowie z wybranych klas – był to krótki kwestionariusz do samodzielnego wypełnienia (około 15 minut) o nazwie „Twoi przyjaciele". Tę część nazywać będziemy dalej **badaniem podstawowym.**

Po zakończeniu badania podstawowego ankieter robił krótką przerwę. Uczniowie wytypowani do drugiej części badania pozostawali w sali, pozostali opuszczali salę – ich udział w badaniu dobiegł końca. Druga część, nazywana dalej **badaniem zasadniczym,** obejmująca 20 wybranych uczniów, została przeprowadzona metodą wywiadu audytoryjnego (1,5 h), który zawierał:

- test na inteligencję (test matryc Ravena),
- kwestionariusz zasadniczy, zadający szczegółowe pytania dotyczące warunków edukacyjnych, statusu społecznego oraz „życia w szkole",
- kwestionariusz umiejętności poznawczych.

[48] Na podstawie: Lisiecka, Z. Biuletyn Badawczy Centralnej Komisji Egzaminacyjnej „Egzamin", 11/2007.

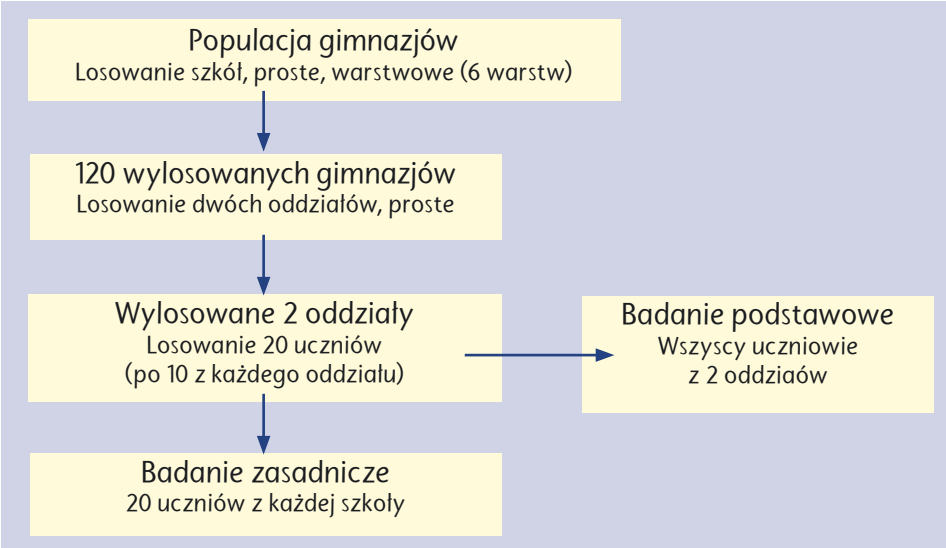
Dobór próby do badania miał charakter doboru wielopoziomowego warstwowego. Na pierwszym poziomie losowane były gimnazja, na drugim poziomie uczniowie. Warstwy na poziomie szkół stworzone zostały ze względu na dwa kryteria:

1) lokalizacja szkoły – miasto i wieś
2) klasyfikacja ze względu na średni wynik szkoły osiągnięty na egzaminie gimnazjalnym w roku szkolnym 2004/2005 w trzech równych przedziałach: szkoły najlepsze – wynik powyżej średniej, średnie oraz słabsze – poniżej średniej.

W wyniku zastosowania tych kryteriów otrzymano warstwy zaprezentowane w sposób pokazany w tabeli 5.3. Przydział szkół do poszczególnych warstw (alokacja próby) został ustalony arbitralnie: 20 szkół dla każdej warstwy. Miało to zapewnić odpowiednią liczbę szkół dla przeprowadzenia porównań między szkołami o różnych wynikach egzaminu i różnej lokalizacji. Trzeba zauważyć, iż alokacja ta, mimo iż arbitralna, nie odbiega też znacznie od udziału szkół w populacji. Gimnazja miejskie według operatu z 2005 roku stanowiły około 49,3% populacji szkół, gimnazja wiejskie zaś około 50,7%. Natomiast podział ze względu na wyniki był podziałem na trzy równe części, co nie zaburzało proporcjonalności na poziomie szkół. Tabela 5.3 przedstawia schemat badania i alokację próby.

Tabela 5.3. Założony podział na warstwy i planowane liczebności wylosowanych uczniów

	Szkoły miejskie W populacji: 3205 (50,7%) (60 szkół, 1200 uczniów)	Szkoły wiejskie W populacji: 3114 (49,3%) (60 szkół, 1200 uczniów)
Wynik powyżej średniej (40 szkół, 800 uczniów)	20 szkół, 2 klasy do badania podstawowego oraz 20 uczniów do badania zasadniczego	20 szkół, 2 klasy do badania podstawowego oraz 20 uczniów do badania zasadniczego
Wynik średni (40 szkół, 800 uczniów)	20 szkół, 2 klasy do badania podstawowego oraz 20 uczniów do badania zasadniczego	20 szkół, 2 klasy do badania podstawowego oraz 20 uczniów do badania zasadniczego
Wynik poniżej średniej (40 szkół, 800 uczniów)	20 szkół, 2 klasy do badania podstawowego oraz 20 uczniów do badania zasadniczego	20 szkół, 2 klasy do badania podstawowego oraz 20 uczniów do badania zasadniczego



Rysunek 5.13. Schemat losowania

W kolejnych krokach: (a) z każdej wylosowanej szkoły do badania podstawowego losowano po dwa oddziały z klas III (jeżeli nie było przynajmniej dwóch oddziałów, wybierano jeden); (b) następnie spośród uczniów uczęszczających do wylosowanych klas losowo dobierano 20 uczniów do badania zasadniczego. Po dziesięciu uczniów z każdej z dwóch wylosowanych wcześniej klas, jeżeli w szkole znajdowały się co najmniej dwie klasy, lub dwudziestu uczniów z jednej klasy – jeżeli w szkole znajdował się tylko jeden oddział klasy III. Ogólny schemat losowania przedstawiony został na rysunku 5.13.

Tak zaplanowanie badanie wymaga od badacza dużej wprawy w analizowaniu danych. Baza danych jednak rekompensuje te trudności swoją bogatą zawartością i mnogością zmiennych, które mogą posłużyć do weryfikacji wielu hipotez. Aby analizować dane zawarte w tej bazie w poprawny sposób, rekomendujemy zapoznanie się z podręcznikiem użytkownika bazy, dostępnym również na stronach CKE. W podręczniku tym omówiono na prostych przykładach, w jaki sposób poprawnie analizować dane zawarte w bazie.

Baza EWD

Udostępniona baza danych, która zawiera wyniki egzaminów zewnętrznych umożliwiające oszacowanie edukacyjnej wartości dodanej, powstała jako efekt prac prowadzonych przez grupę badawczą kierowaną przez Romana Dolatę przy Centralnej Komisji Egzaminacyjnej. Grupa ta, w składzie (kolejność alfabetyczna): Roman Dolata, Maciej Jakubowski, Ewa Kędracka, Przemysław Majkut,

Artur Pokropek, Anna Rappe, Ewa Stożek, Krystyna Szmogel, odpowiada za przygotowanie danych, modeli statystycznych oraz szkoleń z analizy efektywności pracy szkół metodą edukacyjnej wartości dodanej.

Wyniki prac tej grupy opublikowane zostały w książce pod redakcją Romana Dolaty oraz specjalnych numerach biuletynów CKE. Powstała też publikacja dokumentująca prace nad modelem EWD w Polsce w języku angielskim. Publikacje te można znaleźć na stronie http://www.ewd.edu.pl/publikacje_ewd.html. Istnieje też bardzo bogata literatura dotycząca metod modelowania EWD w oparciu o doświadczenia innych krajów. Głównym źródłem wiedzy są tu raporty OECD (2008) oraz RAND (2005). Przykłady modelowania EWD w oparciu o polskie dane opisano w rozdziałach 3 oraz 4.

Wyjątkową cechą udostępnianych zbiorów danych jest to, że dla każdego ucznia zawierają połączone wyniki sprawdzianu oraz egzaminu gimnazjalnego. Łączenie zostało dokonane na poziomie Okręgowych Komisji Egzaminacyjnych, które co roku przekazywały grupie badawczej dane uczniów, dla których scalenie wyników obu egzaminów zewnętrznych było możliwe. Ze względu na to, że brakuje w polskim systemie egzaminacyjnym indywidualnego identyfikatora dla każdego ucznia, łączenie wykonano głównie w oparciu o imię i nazwisko oraz płeć uczniów, które na szczęście rzadko podawane były błędnie, dzięki czemu tylko dla stosunkowo niewielkiej części uczniów danych nie połączono.

Trzeba też wziąć pod uwagę, że łączenie na poziomie OKE powoduje, że uczniowie, którzy między sprawdzianem szóstoklasistów a egzaminem gimnazjalnym zmienili miejsce zamieszkania na leżące w innej OKE, nie mogli zostać uwzględnieni. Przy łączeniu nie brano też pod uwagę uczniów piszących inne niż standardowe arkusze egzaminacyjne, bowiem wyniki tych uczniów nie są w pełni porównywalne. Powoduje to, że łącznie przez 3 lata dla około 8% uczniów danych nie zdołano połączyć. Nie jest to duża liczba, biorąc pod uwagę, że sporo uczniów, dla których wyników nie udało się połączyć, pisało arkusze niestandardowe i pominięcie tych uczniów jest uzasadnione. Jednak największą pominiętą grupę stanowią uczniowie drugoroczni, co trzeba mieć na uwadze, interpretując dane. Poza tymi przypadkami braki danych można też uznać w większości za losowe, bowiem wynikają z błędnie zapisanych danych uczniów w systemie. Jedynie kwestia braku uczniów, którzy zmienili rejon OKE, powoduje nielosowy dobór związany między innymi z migracją. Jest to jednak znikomy procent uczniów i dla analiz na poziomie szkół nie powinien mieć większego znaczenia. Z ogólnej liczby gimnazjów w Polsce (nawet włączając na przykład szkoły przyszpitalne) baza przygotowana do liczenia EWD zawiera ponad 98% szkół.

Baza, z której wylosowano udostępniane dane, została też ograniczona do szkół, dla których możliwe jest prowadzenie porównywalnych, merytorycznie i statystycznie uprawionych analiz. Po pierwsze, wykluczono gimnazja niepubliczne. Po drugie, ze względu na to, że losowano nie tylko szkoły, lecz także klasy, dla których potencjalne analizy powinny być również wiarygodne, wykluczono

klasy małe, z mniej niż 15 uczniów (około 2% uczniów), oraz klasy bardzo duże, z więcej niż 35 uczniów (mniej niż 1% uczniów). Po trzecie, założono, że udostępnione zostaną dane jedynie dla szkół, dla których posiadamy odpowiednią liczbę obserwacji z każdego roku. Takie kryterium jest niezbędne, o ile dane mają być wykorzystywane do oceny zmian między latami w poszczególnych szkołach. Za minimalną liczbę przyjęto 10 obserwacji w każdym roku[49]. Po czwarte, przyjęto, że ze wszystkich 3 lat każde gimnazjum powinno posiadać co najmniej 50 obserwacji. Po piąte, wykluczono z bazy niewielką liczbę obserwacji dla uczniów z brakiem wyniku lub wynikiem równym zero. Przyjęcie wszystkich powyższych kryteriów zmniejszyło znacząco liczbę szkół (o 20%), jednak liczba uczniów zmniejszyła się o mniej niż 8%. Uzyskane w ten sposób dane nie są co prawda reprezentatywne dla całej populacji gimnazjalistów w Polsce, jednak umożliwiają prowadzenie miarodajnych analiz efektywności nauczania w gimnazjach metodami EWD.

Do bazy z wynikami egzaminacyjnymi dołączono zmienne z Systemu Informacji Oświatowej (SIO) Ministerstwa Edukacji Narodowej oraz Banku Danych Regionalnych Głównego Urzędu Statystycznego (BDR GUS). Dane SIO łączone były za pomocą kodu gimnazjum przypisanego przez OKE. Kod ten powinien znajdować się także w bazach SIO, niestety w wielu wypadkach nie był tam jednak zapisany lub zapisany był błędnie. Dla takich szkół dane łączono po nazwie szkoły i miejscowości, w której znajduje się placówka. Dane GUS połączono, wykorzystując kod nadawany przez GUS gminom, który zawarty jest także w kodzie szkoły przypisywanym przez OKE. W tym wypadku także występowały różnice wynikające z innego kodowania OKE w dużych miastach oraz ze zmian kodów GUS nie zawsze odzwierciedlonych w kodach OKE. Dla wszystkich szkół i gmin różnice te udało się jednak wyjaśnić i dane zostały połączone.

Dołączone dane SIO dotyczą poziomu szkoły, a dane BDR GUS poziomu gminy (lub miasta-powiatu). Wysoki poziom agregacji danych BDR GUS powoduje, że nie zawsze są to zmienne dobrze określające sytuację danej szkoły. Jeśli na przykład gmina przeznacza bardzo różne kwoty podlegającym jej placówkom, to średnie wydatki na poziomie gminy przypisane szkole są kiepskim wskaźnikiem jej rzeczywistych zasobów. Można jednak uznać, że w większości wypadków gminy rozdysponowują środki na podobnych zasadach dla wszystkich szkół. Można też traktować te zmienne jako odzwierciedlające poziom finansowania lub inne cechy na poziomie lokalnym, czyli jako wskaźniki otoczenia szkoły, w jakim funkcjonuje. Ostateczna interpretacja zależy od użytkowników da-

[49] Liczbę obserwacji (liczbę uczniów zdających w danym roku egzamin gimnazjalny) obliczono jako średnią z liczby prawidłowych wyników w części humanistycznej oraz w części matematyczno-przyrodniczej. W większości szkół liczba zdających obie części egzaminu była identyczna, jednak w kilku przypadkach różniła się. Jeśli w jednej z części liczba zdających wynosiła w danym roku 10, ale w innej 9, to taka szkoła nie była brana pod uwagę przy losowaniu prezentowanej bazy danych, ponieważ średnia liczba uczniów wyniosła w tej szkole 9,5.

nych, a przede wszystkim od tego, jakie zmienne i w jakim celu wykorzystywano w konkretnym badaniu.

Przykłady badań w oparciu o dane EWD dla gimnazjów omówiono już w rozdziałach 3 oraz 4. Przykładowe analizy można znaleźć w książce pod redakcją Romana Dolaty *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie wyników egzaminów zewnętrznych*, w pracy Macieja Jakubowskiego *Zastosowanie modeli EWD do ewaluacji polityki edukacyjnej* lub w innych omówionych w rozdziale 4. Powstał także podręcznik użytkownika opisujący dokładnie strukturę i proces tworzenia bazy, a także zawierający przykłady poprawnej analizy danych metodami EWD.

Baza „Trzecioklasiści"

Projekt *Badanie umiejętności podstawowych uczniów trzecich klas szkoły podstawowej* został rozpoczęty w 2005 roku. Badania te, zlecone przez Centralną Komisję Egzaminacyjną w Warszawie, zostały zrealizowane przy współfinansowaniu przez Europejski Fundusz Społeczny. Celem badania było zebranie informacji na temat podstawowych umiejętności (matematycznych i językowych) uczniów kończących pierwszy etap kształcenia w szkole podstawowej. Podstawową grupą badanych byli uczniowie klas trzecich. Oprócz nich badaniem objęto także ich rodziców, nauczycieli i dyrektorów szkół, do których oni uczęszczali. Badanie obejmuje trzy obszary zagadnień:

- umiejętności uczniów klas trzecich,
- motywacja uczniów klas trzecich do uczenia się,
- wybrane czynniki wpływające na poziom umiejętności uczniów.

Zebrane podczas badania dane zostały następnie przeanalizowane przez zespół projektujący badanie, czego efektem jest raport *Badanie umiejętności podstawowych uczniów trzecich klas szkoły podstawowej. Raport z badania ilościowego*, pod redakcją Mirosława Dąbrowskiego i Małgorzaty Żytko.

Badanie zostało zrealizowane w czerwcu 2006 roku. Samo zbieranie informacji i badanie uczniów trwało dwa dni. Uczniowie uczestniczyli wówczas w dwóch godzinnych seriach testów dziennie. Także wówczas rozmawiano z pracownikami szkoły oraz zbierano pozostałe informacje.

W badaniu ilościowym uczestniczyły następujące grupy:

- uczniowie klas trzecich:
  zostali oni poddani testom sprawdzającym badającym umiejętności matematyczne, czytanie ze zrozumieniem i pisanie; wypełniali także kwestionariusz „Ja i moja nauka";
- rodzice/opiekunowie badanych uczniów:
  tej grupy badanych dotyczył jedynie kwestionariusz „Rodzic/opiekun o uczniu";
- nauczyciele badanych klas:

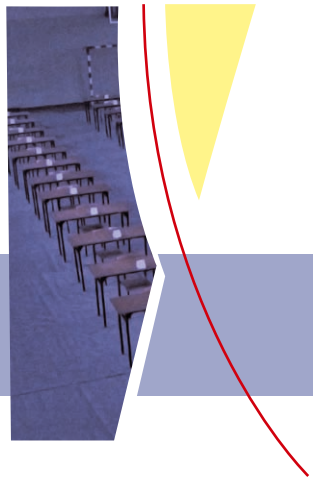
kwestionariusze, które były skierowane do nauczycieli, to „Edukacja matematyczna w klasach 1-3", „Edukacja językowa w klasach 1-3", „Nauczyciel o klasie", „Nauczyciel o uczniach"; proszono także nauczycieli o wyrażenie opinii o poszczególnych testach;

- dyrektorzy szkół, w których przeprowadzono badania:

do tej grupy respondentów zastosowano kwestionariusz „Dyrektor o szkole".

Dodatkowe informacje o szkole zostały uzyskane za pomocą kwestionariusza kierowanego do ankieterów („Ankieter o szkole"). Proszono w nim, by opisał on warunki, w jakich uczą się uczniowie klas trzecich, budynek szkolny, jego stan techniczny.

W bazie zawarto wszystkie odpowiedzi z ankiet i testów umiejętności. Przygotowaną do wykorzystania bazę dopełniono odpowiednią dokumentacją zawierającą podręcznik użytkownika, który oprócz dokładnego opisu realizacji badania oraz zawartości bazy, prezentuje także przykładowe analizy.



6. Słowniczek

błąd losowy
Jest to błąd mający charakter losowy, niezwiązany z działaniem jakiegokolwiek systematycznego mechanizmu. Z definicji nie można przewidzieć wartości błędu losowego w kolejnych pomiarach. Można natomiast modelować losowe błędy za pomocą rozkładów statystycznych, na przykład rozkładu normalnego, gdzie określa się średnią wartość błędu (najczęściej przyjmuje się dla wygody 0) oraz jego wariancję. Każdy pomiar obarczony jest błędem losowym.

błąd pomiaru
Każdy pomiar obarczony jest błędem. W wypadku pomiaru osiągnięć szkolnych błąd pomiaru wynika z działania takich czynników, jak m.in: dyspozycja ucznia w danym dniu, zgadywanie odpowiedzi w zadaniach zamkniętych, efekt egzaminatora w zadaniach otwartych, ściąganie czy przypadkowość w doborze zadań testowych.

błąd próbkowania
Wszelkie obliczenia opierające się na losowej próbie uczniów obarczone są błędem związanym z tym, że próba stanowi tylko jedną z możliwych reprezentacji wartości zmiennych w całej populacji. Jeśli badamy na przykład reprezentatywną ogólnopolską próbę uczniów, starając się na jej podstawie określić poziom umiejętności w całym kraju, to błąd próbkowania związany jest z tym, że wylosowana grupa różni się nieco w swoich charakterystykach od wszystkich uczniów w kraju. Różnice te powinny być niewielkie w badaniach reprezentatywnych, wykorzystujących losowe próby. Mogą być znaczne, jeśli przyjęto niepoprawny schemat losowania, a tym bardziej, gdy próba nie była dobierana w pełni losowo.

błąd standardowy
Statystyczny sposób wyrażania *błędu próbkowania* (zob.) przy obliczaniu parametrów populacyjnych. Przykładowo błąd standardowy dla średniej zapisujemy jako: $BS(E)=\frac{\sigma}{\sqrt{n}}$, gdzie σ jest odchyleniem standardowym, a n liczebnością próby. Za pomocą błędu standardowego w prosty sposób można wyznaczać przedziały ufności.

błąd systematyczny
Błąd systematyczny wynika z wadliwej procedury pomiaru, zwykle prowadzi do „przesunięcia" wyniku pomiaru na danej skali w jedną stronę (jego zawyżenia lub zaniżenia). Jeżeli istnieje błąd

systematyczny, systematycznie zaniżający wyniki, to i *wynik prawdziwy* (zob.) zostanie zaniżony; jeżeli mamy do czynienia z systematycznym błędem zawyżającym wyniki, to i *wynik prawdziwy* zostanie zawyżony.

cecha ukryta (nieobserwowalna)
Cecha obiektu, która nie jest nam bezpośrednio dana, a jedynie obserwowana pośrednio na podstawie obserwowalnych wartości innych wskaźników z nią powiązanych. Przykładem cechy ukrytej jest na przykład konstrukt „umiejętności matematyczne ucznia". Umiejętności matematyczne ucznia nie są dane badaczowi bezpośrednio. Inaczej mówiąc, nie istnieje obserwowalna (możliwa do bezpośredniego zmierzenia) zmienna, która by takie umiejętności odzwierciedlała. Badacz jedynie wnioskuje na temat tych umiejętności na podstawie empirycznych manifestacji tej cechy, prościej mówiąc, na podstawie rozwiązanych przez ucznia zadań matematycznych.

chi-kwadrat
To rozkład statystyki testowej często używany w różnych zastosowaniach statystycznych. Na podstawie rozkładu chi-kwadrat jesteśmy w stanie testować hipotezy statystyczne, na przykład hipotezę mówiącą o tym, iż model jest dobrze dopasowany do danych, lub hipotezę o niezależności statystycznej dwóch zmiennych. Testy oparte na statystyce chi-kwadrat są jednymi z najczęściej używanych testów w zastosowaniach statystycznych.

efekt egzaminatora
To *błąd systematyczny* (zob.), czyli nielosowy, pomiaru, który wynika bezpośrednio z różnic w ocenianiu między egzaminatorami. Ma miejsce wyłącznie w wypadku oceniania zadań otwartych. Można wyróżnić dwa typy efektu egzaminatora: prosty efekt egzaminatora i interakcyjny efekt egzaminatora. Prosty efekt egzaminatora polega na systematycznym zaniżaniu lub zawyżaniu wyników przez konkretnego egzaminatora w stosunku do innych egzaminatorów (egzaminatorzy są albo zbyt surowi, albo zbyt łaskawi). Interakcyjny efekt egzaminatora polega na specyficznym ocenianiu konkretnych odpowiedzi na zadania (zachodzi interakcja między egzaminatorem a pracą) – konkretne odpowiedzi na zadania mogą wywoływać u danych egzaminatorów mechanizm nieuzasadnionego zawyżania lub zaniżania wyników (na przykład gdy egzaminator upodobał sobie jeden typ odpowiedzi). W polskim systemie zbadane zostały efekty egzaminatora związane ze sposobami oceniania (tradycyjne a elektroniczne) lub przynależnością do zespołu egzaminatorów.

efekt podłogowy
Efekt podłogowy zachodzi wtedy, gdy pojedyncze zadanie lub cały test okazują się zbyt trudne dla grupy uczniów słabych. Gdy pojawia się efekt podłogowy, nie jesteśmy w stanie dobrze zróżnicować uczniów słabych, tj. dokonać na przykład rozróżnienia pomiędzy uczniami słabymi a uczniami bardzo słabymi, wymagającymi szczególnej, dodatkowej pomocy. Podobnie jak w wypadku *efektu sufitowego* (zob.), występowanie efektu podłogowego może być niepożądane, tak jest najczęściej, lub może być pomijalne, na przykład gdy celem jest jedynie zidentyfikowanie najlepszych uczniów.

efekt sufitowy
Efekt sufitowy zachodzi wtedy, gdy pojedyncze zadanie lub cały test okazują się zbyt łatwe dla grupy najzdolniejszych uczniów. Jeżeli znaczna część uczniów zdobędzie maksymalną liczbę punktów z danego testu, uniemożliwi to precyzyjne określenie poziomu ich umiejętności. Dowiemy się, że ta grupa uczniów umie dużo, lecz nie będziemy w stanie rozróżnić poziomu umiejętności tych uczniów, na przykład oddzielić uczniów bardzo dobrych od wybitnych. W zależności od rodzaju testu, efekt sufitowy może stanowić problem lub też nie. Jeśli test ma jedynie określić, czy uczeń osiągnął pewien, dość nisko określony poziom wiedzy, to efekt sufitowy nie ma znaczenia. Podobnie, gdy chcemy jedynie uczniów poklasyfikować na słabych, przeciętnych

i ponadprzeciętnych. Dla większości testów chcemy jednak rozróżnić umiejętności jak największej liczby uczniów, wtedy efekt sufitowy obniża jakość pomiaru.

estymacja
Metoda statystyczna wyznaczania interesującego nas *parametru* (zob.) w *populacji* (zob.) – na przykład średnia, odchylenie standardowe – na podstawie próby z tej populacji. Rozróżniamy estymację punktową, czyli na przykład średni wynik w populacji albo EWD szkoły, a także estymację przedziałową, na przykład przedział ufności dla średniej albo przedział ufności dla EWD szkoły.

korelacja
Miara powiązania dwóch zmiennych. Mówiąc, że „wyniki sprawdzianu silnie korelują z wynikami egzaminu gimnazjalnego", mamy na myśli, że za pomocą wyników sprawdzianu można w dużym stopniu przewidzieć wynik egzaminu gimnazjalnego. Korelacja dodatnia oznacza w tym wypadku, że przeciętnie wraz ze wzrostem wyników sprawdzianu wzrastają wyniki egzaminu gimnazjalnego. Korelacja bliska zera (mniejsza niż 0,3-0,4) oznacza, że brak jest związku między dwiema zmiennymi. Analizując dane statystyczne, mówiąc o korelacji, najczęściej mamy na myśli tzw. współczynnik korelacji liniowej Pearsona, który mierzy siłę zależności liniowej między dwiema ciągłymi zmiennymi. Istnieją jednak też inne współczynniki korelacji dostosowane do różnego rodzaju zmiennych, na przykład korelacja rangowa dla zmiennych mierzonych na skali porządkowej. Współczynnik korelacji może przyjmować wartości od minus 1 do 1. Znak oznacza kierunek zależności, a wartość współczynnika określa siłę powiązania między zmiennymi.

krzywa informacyjna
Jest to funkcja, która wyraża ilość informacji w danym przedziale umiejętności, jaką otrzymujemy na podstawie pojedynczego zadania. Krzywa informacyjna estymowana jest na podstawie teorii odpowiedzi na zadanie testowe.

krzywa logistyczna
Krzywa będąca wykresem funkcji postaci $f(x)=\frac{1}{1+e^{-x}}$ (w najprostszej jednoparametrycznej formie) znajduje szerokie zastosowanie nie tylko w naukach przyrodniczych i technicznych, lecz także w naukach społecznych czy w medycynie. Jest podstawową funkcją, którą wykorzystuje się w modelowaniu opartym na teorii odpowiedzi na zadanie testowe.

łatwość zadania (testu)
Parametr określający, w jakim stopniu zadanie jest łatwe. Konwencjonalnie pod pojęciem łatwości zadania kryje się *parametr estymowany* (zob.) na podstawie klasycznej teorii testu – procent poprawnych odpowiedzi na zadanie.

„martwy" dystraktor
To taki dystraktor, czyli błędna odpowiedź w zadaniu wyboru, który bardzo rzadko w porównaniu z innymi dystraktorami wybierany jest przez rozwiązujących dane zadanie. Zazwyczaj jest to dystraktor, który jednoznacznie kojarzy się uczniom jako odpowiedź niewłaściwa. Martwe dystraktory osłabiają psychometryczne właściwości zadań zamkniętych, gdyż uczeń nieznający odpowiedzi na zadanie, eliminując jeden z dystraktorów, zwiększa swoje prawdopodobieństwo poprawnej odpowiedzi, ograniczając zgadywanie do pozostałych odpowiedzi do wyboru.

moc dyskryminująca; wskaźnik dyskryminacji
Moc dyskryminująca jest wskaźnikiem przydatnym w analizie zarówno zadań, jak i całego testu. Synonimem dla tego pojęcia jest *moc różnicująca* (zob.) – my jednak dla porządku terminologicznego zdecydowaliśmy się używać tej drugiej nazwy jedynie dla wskaźników opartych na klasycznej teorii testu, mimo iż są wyrazem analogicznych zjawisk, a jedynie sposób ich obliczania jest odmienny.

moc różnicująca; wskaźnik mocy różnicującej
Moc różnicująca jest wskaźnikiem przydatnym w analizie zarówno zadań, jak i całego testu. Synonimem dla tego pojęcia jest „*moc*" *dyskryminująca* (zob.) – my jednak dla porządku terminologicznego zdecydowaliśmy się używać tej drugiej nazwy jedynie dla wskaźników opartych na modelach IRT, mimo iż są wyrazem analogicznych zjawisk, a jedynie sposób ich obliczania jest odmienny. Wskaźniki mocy różnicującej pokazują nam, jak silnie zadanie związane jest z całym testem. Im pytanie silniej związane z wynikiem dla całego testu, tym jego moc różnicująca (inaczej moc prawdziwego wyniku) jest większa. Dobre pomiarowo zadania powinny mieć dużą moc różnicującą, czyli być silnie związane z całym testem.

odchylenie standardowe
Jest to pierwiastek z *wariancji* (zob.) danej zmiennej. Obok wariancji jest to najczęściej używany parametr zróżnicowania wartości zmiennej. Odchylenie standardowe często wykorzystuje się do opisu *rozkładu normalnego* (zob.) z powodu wygodnej interpretacji. Na przykład: jeżeli test na inteligencję ma średnią 100 i odchylenie standardowe 15, oznacza to, iż około 68% *populacji* (zob.) ma wyniki między 85 punktami a 115 punktami, około 95,5% populacji wyniki między 70 punktami a 130 punktami. Jeżeli chcielibyśmy dowiedzieć się, w jakim przedziale znajduje się 90% populacji, musimy przemnożyć odchylenie standardowe przez 1,64. W naszym wypadku $15 \times 1{,}64 = 24{,}6$. Zatem w odległości +/– 24,6 punktu od średniej będzie się zawierało 90% wyników (czyli między 75,4 a 124,6). Jeżeli chcemy uzyskać informacje o 95% wyników, odchylenie standardowe należy przemnożyć przez 1,96. W badaniach edukacyjnych często wyraża się efekty zmiennych w wartościach odchylenia standardowego, jako sposób na porównanie siły zależności między różnymi zmiennymi mierzonymi na skalach o nieporównywalnych wartościach lub też między badaniami opierającymi się na różnej metodologii i odmiennych danych.

parametr
Parametr jest liczbową charakterystyką opisującą elementy *populacji* (zob.). Parametr opisuje cechy populacji bezpośrednio lub pośrednio (poprzez model statystyczny). Przykładami parametrów opisujących populację bezpośrednio są średnia i *wariancja* (zob.). Innym typem parametrów są parametry opisujące właściwości populacji na podstawie przyjętego modelu. Są to na przykład wartości współczynników *regresji* (zob.) czy parametr *trudności* (zob.) lub *dyskryminacji* (zob.) w modelach IRT. Parametry nie są dane, trzeba je wyznaczać (estymować) na podstawie skończonego zbioru elementów (próby).

populacja
Jest to zdefiniowany zbiór obiektów interesujących badacza. Przykładem może być populacja Polaków zamieszkałych na stałe w Polsce w wieku od 15 do 65 lat. Badaczy edukacyjnych interesują przede wszystkim populacje uczniów, ewentualnie szkół, rodziców lub nauczycieli. Na przykład populacja uczniów uczących się w klasie trzeciej gimnazjum czy populacja rodziców uczniów szkół podstawowych lub też populacja nauczycieli uczących matematyki w gimnazjum. Interesującą nas populację powinniśmy określić przez wiek, poziom nauczania, ale także i geograficznie, na przykład populacja rodziców w Polsce, populacja nauczycieli województwa małopolskiego czy też populacja 14-latków zamieszkujących kraje Unii Europejskiej. Analizując wyniki badań lub też omawiając własne wyniki analiz, należy zawsze odnosić się do danej populacji. Porównując wyniki badań, należy mieć na uwadze, że mogą one dotyczyć różnych populacji, co powoduje różnice w rezultatach. Bezpośrednio można porównywać jedynie wyniki badań reprezentatywnych dla tych samych populacji.

przedział ufności
Aby wyznaczyć 90-procentowy przedział ufności wokół średniej, stosujemy wzór $E(X) \pm 1{,}64BS(E)$, jeżeli chcemy wyznaczyć 95-procentowy przedział ufności, stosujemy formułę $E(X) \pm 1{,}96BS(E)$. Definicja 95-procentowego przedziału ufności mówi tyle, iż w wyznaczonym przez nas przedziale

ufności w 95% przypadków obliczeń będzie się zawierał prawdziwy wynik. Innymi słowy, jeżeli losowalibyśmy z danej *populacji* (zob.) próbę 100 razy i na jej podstawie konstruowali przedział ufności, to w 95 przypadkach wynik w populacji zawierałby się w skonstruowanych przez nas przedziałach ufności. Jeżeli konstruujemy 90-procentowy przedział ufności, wynik prawdziwy będzie zawierał się w tych przedziałach w 90% przypadków.

przedział ufności dla wskaźnika EWD

Wyznaczone przedziały ufności możemy traktować jako regułę decyzyjną. Jeżeli chcemy w sposób odpowiedzialny formułować na podstawie wskaźnika EWD (edukacyjnej wartości dodanej) oceny typu: *szkoła A lepiej uczy w zakresie sprawdzanym przez egzamin gimnazjalny od szkoły B*, to warto wiedzieć, jakie jest ryzyko popełnienia błędu. Przedziały ufności pozwalają nam to ryzyko oszacować. Jeżeli wyznaczymy 95-procentowe przedziały ufności dla wskaźnika EWD dla porównywanych szkół i przedziały te są rozłączne, to stwierdzenie, że szkoły te różnią się pod względem wartości wskaźnika EWD, obarczone jest niewielkim ryzykiem sformułowania nietrafnej oceny. Gdy przedziały częściowo pokrywają się, formułowanie wszelkich porównań jest ryzykowne i należałoby się od nich powstrzymać.

Sugerowanym współczynnikiem ufności dla wskaźnika EWD jest 95%, bowiem w tym wypadku błędna ocena może wiązać się z dotkliwymi konsekwencjami. Inaczej mówiąc, porównując wskaźnik EWD placówek, często chcemy dokonać porównania ich jakości nauczania i wskazana jest tu duża ostrożność, szczególnie jeśli wyniki porównań dostępne są publicznie lub też wykorzystuje się je jako podstawę istotnych decyzji. Jeśli jednak nasze analizy mają charakter wewnętrzny, nastawiony na przykład na określenie postępów grup uczniów i poprawę jakości pracy z nimi, to wskazane jest stosowanie niższego współczynnika ufności, na przykład 90%. Przyjęcie niższego współczynnika ufności dla tych grup związane jest z mniej dotkliwymi konsekwencjami sformułowania nietrafnej oceny.

regresja liniowa

Metoda statystyczna określenia zależności między *zmienną zależną* (zob.) (inaczej: zwaną zmienną objaśnianą), a *zmiennymi niezależnymi* (zob.) (inaczej: zmiennymi objaśniającymi). Posługując się równaniem regresji, można przewidywać – w sensie statystycznym – wartości zmiennej zależnej na podstawie wartości zmiennych niezależnych. Nazwa regresja liniowa nawiązuje do tego, że w równaniu określamy wartości zmiennej zależnej liniową kombinacją zmiennych niezależnych, czyli na przykład $y = a + b \times x + c \times z$. W typowych zastosowaniach szacuje się równanie regresji metodą najmniejszych kwadratów, tak aby jak najlepiej opisywało ono zależności między zmiennymi w próbie.

Regresję stosuje się na przykład w modelach EWD dla gimnazjum, gdzie zmienną zależną jest wynik egzaminu gimnazjalnego, a zmiennymi niezależnymi są wyniki sprawdzianu oraz oznaczenia dla płci i dysleksji uczniów (por. zmienne kontrolne). Najczęściej w metodzie regresji analizujemy współczynniki równania regresji, które oddają przeciętną zmianę w wartościach zmiennej zależnej powiązaną ze zmianami o jedną jednostkę w wartościach zmiennej niezależnej. Przykładowo, w regresji EWD współczynnik dla wyników sprawdzianu oznacza, o ile punktów zmieni się przewidywany wynik egzaminu gimnazjalnego wraz ze zmianą o 1 punkt wyniku sprawdzianu ucznia. Jednak w wypadku modelu EWD dla gimnazjów metoda regresji liniowej służy jedynie dopasowaniu zależności między wynikami egzaminów. Współczynniki przy zmiennych mają tu mniejsze znaczenie, bowiem interesują nas przede wszystkim wartości przewidywane oraz tzw. reszty, na podstawie których obliczamy wskaźnik EWD placówki. Istnieją jednak metody wykorzystujące regresję liniową, gdzie wskaźnik EWD szkoły szacuje się bezpośrednio jako zmienną w równaniu, na przykład metoda regresji liniowej z efektami stałymi.

reprezentatywne (badanie)

To badanie opierające się na losowej próbie, pozwalające na odtworzenie określonych parametrów z *populacji* (zob.), na przykład średniego wyniku uczniów czy też rozkładów ze względu na płeć lub wykształcenie rodziców. Badanie jest reprezentatywne tylko względem określonej populacji,

warto mieć to na uwadze, porównując badania różnych populacji, na przykład porównując wyniki badania reprezentatywnego dla 15-latków z wynikami badania reprezentatywnego dla trzecich klas szkoły podstawowej. Różnice w wynikach związane są tu z różnicami między tymi populacjami, a niekoniecznie z innymi cechami badania.

reprezentatywność (pomiaru)
Stopień dokładności odzwierciedlenia cech przedmiotu badań i ich różnorodności. Problem reprezentatywności, jeśli chodzi o testy osiągnięć szkolnych, sprowadza się w zasadzie do kwestii: O co pytać w teście? Pomiar jest reprezentatywny, jeżeli pokrywa w miarę swoich możliwości (barierą jest liczba zadań) przestrzeń szczegółowych treści i umiejętności, jakie mają zostać zmierzone w procedurze pomiarowej. Innymi słowy, w pomiarze edukacyjnym reprezentatywność jest to zgodność między tym, co było przedmiotem pomiaru (tym, co było nauczane), a tym, co zostało zastosowane w narzędziu pomiarowym (zawartość testu).

reszta w modelach EWD
Różnica między uzyskanym przez ucznia wynikiem na egzaminie gimnazjalnym a wynikiem przewidywanym na podstawie równania *regresji* (zob.). Ponieważ reszta jest wartością nieprecyzyjną (obarczoną zarówno błędem pomiaru na sprawdzianie, jak i błędem pomiaru na egzaminie gimnazjalnym oraz błędem dopasowania krzywej regresji), nie należy jej wykorzystywać jako indywidualnej miary postępu ucznia w czasie nauki w gimnazjum. Reszty należy używać jedynie do analizowania EWD szkoły lub innej badanej grupy. Wspomniane błędy mają w większości charakter losowy i niwelują się dla liczniejszych grup, stąd szacowanie EWD szkoły ma już sens. Im większa liczba uczniów, tym błąd pomiaru jest tu mniejszy.

rozkład normalny
Większość analiz statystycznych może być aplikowana jedynie dla zmiennych, które rozkładają się w *populacji* (zob.) w sposób losowy, a w wypadku zmiennych ciągłych najczęściej zakłada się, że mają one rozkład normalny. W rozkładzie tym najwięcej obserwacji ma wartości bliskie średniej i liczba obserwacji maleje w miarę oddalania się od niej. Przy tym rozkład jest symetryczny, a więc jest tyle samo obserwacji powyżej, jak i poniżej średniej. Przykładem zmiennej o rozkładzie normalnym jest wzrost. Zmienną o rozkładzie normalnym (lub do niego zbliżonym) o średniej μ i wariancji σ^2, lub odchyleniu standardowym σ, zapisuje się następująco: $N(\mu;\sigma^2)$ lub $N(\mu;\sigma)$. Rozkład normalny opisuje się za pomocą funkcji, której kształt przypomina dzwon, stąd też angielska nazwa *bell curve*.

rzetelność
Rzetelność narzędzia pomiarowego dotyczy precyzji, z jaką jest ono w stanie zmierzyć badaną cechę (na przykład poziom umiejętności i wiedzy ucznia na podstawie zadań dotyczących jedynie fragmentów materiału, który powinien być przez ucznia opanowany w całości). Jeżeli narzędzie pomiarowe jest rzetelne, to kolejne pomiary będą dawały ten sam rezultat (lub bardzo zbliżony).

skala centylowa
Skala pozycyjna wyników uporządkowanych rosnąco: aby ustalić wartości centyli, porządkujemy wyniki od najniższych do najwyższych, dzielimy ich zbiór na 100 równych części, a następnie określamy położenie granic między tymi częściami na skali wyników i oznaczamy je jako centyle C1, C2, C3, C4, C5,... C99. Liczba centyli wynosi 99. Grup centylowych, czyli grup wyników leżących między centylami, jest 100. Numer kolejny najbliższego centyla stanowi rangę centylową danego wyniku testowania, to jest procent wyników niższych od danego. Na przykład wynik szkoły w 45. centylu oznacza, że 44% szkół uzyskało wynik niższy od tej szkoły.

skala ilorazowa
Na skali ilorazowej mierzymy na przykład wzrost, liczbę dzieci czy dochody. W każdym z tych wypadków możemy powiedzieć bez wahania, iż ktoś jest od kogoś dwa razy wyższy, iż ktoś ma

dwa razy więcej dzieci od kogoś innego lub że zarabia trzy razy więcej pieniędzy. Robimy to bez problemu, ponieważ w każdym wypadku posiadamy jednoznaczne, a nie arbitralne zero, stanowiące punkt odniesienia dla każdego pomiaru.

skala interwałowa
Jeżeli chcemy mówić, o ile coś jest twardsze, mocniejsze, cieplejsze etc. i chcemy wyrazić wielkość tej różnicy za pomocą liczb, musimy odwołać się do skali interwałowej. Klasycznym przykładem użycia skali interwałowej jest pomiar temperatury w stopniach Celsjusza.

skala nominalna
Skala nominalna dzieli obiekty na grupy, których nie da się w żaden sposób uporządkować czy stosować do nich działań arytmetycznych. Na przykład chłopcy i dziewczęta czy ludzie zielonoocy, niebieskoocy i ludzie o oczach piwnych. Jedyna operacja, jaką możemy przeprowadzać na zmiennej mierzonej na tej skali, to operacja porównywania – sprawdzać, czy dwa obiekty są ze sobą tożsame, czy też nie (na przykład czy dwie jednostki mają taką samą płeć, czy też nie).

skala porządkowa
Skala porządkowa występuje wszędzie tam, gdzie możemy zastosować porządkowanie: mając kilka obiektów, możemy powiedzieć, że coś jest większe, ładniejsze, mądrzejsze, równe itp.

skala staninowa
Nazwa od ang. *standard nine* czyli „standardowa dziewiątka". Dziewięciostopniowa znormalizowana skala wyników. Kolejne stopnie tej skali zawierają odpowiednio 4%, 7%, 12%, 17%, 20%, 17%, 12%, 7%, 4% wyników uporządkowanych rosnąco. Skala pozwala pozycjonować wynik ucznia lub też wynik szkoły (na przykład średni wynik lub wskaźnik EWD) i porównywać z wynikami innych uczniów (innych szkół). Kolejne stopnie skali przyjęto nazywać: najniższy, bardzo niski, niski, niżej średni, średni, wyżej średni, wysoki, bardzo wysoki, najwyższy. Przykładowo, wynik szkoły w staninie 8. (bardzo wysokim) oznacza, że 4% szkół uzyskało wynik wyższy, 7% – porównywalny, a 89% – wynik niższy iż ta szkoła. Skala ta jest przydatna do opisu wyników egzaminów, jeżeli nie są one wyrażone na standaryzowanej skali, tak jak w wypadku polskich egzaminów. Ze względu jednak na to, że zawiera ona tylko 9 kategorii, jest mało dokładna i nie powinna być używana do obliczeń statystycznych, a jedynie służyć jako jeden ze sposobów komunikowania wyników egzaminów.

skalogram (Gutmana)
Składa się z zestawu dychotomicznych pytań o prostej, jednowymiarowej, kumulatywnej strukturze. Pozytywna odpowiedź na trudniejsze pytanie implikuje odpowiedzi na pytanie łatwiejsze. Utworzona za pomocą skalogramu Gutmana skala posiada charakter skali porządkowej.

skalowanie
Procedura, za pomocą której odzwierciedlamy własności obiektów i zależności zachodzące między tymi obiektami przy wykorzystaniu liczb. Prosty przykład skalowania to przypisywanie uczniom ocen. Opisujemy umiejętności uczniów za pomocą zestawu liczb (najczęściej sześciu). Od tego prostego systemu skalowania wymagamy, by zachowane zostały zależności wynikające z własności obiektów, tj. uczeń słabszy dostaje niższą wartość liczbową oceny od ucznia lepszego. Uczeń słaby posiadł mniejszą wiedzę od ucznia dobrego, co odzwierciedlamy w systemie liczbowym, gdyż 3 jest mniejsze od 5.

skalowanie kumulatywne
Procedura skalowania, w której dla określenia poziomu danej cechy używa się więcej niż jednego pomiaru. Zakłada się tu, że pomiar użyty w skalowaniu ma charakter porządkowy (określany jest przez skalę porządkową). Zakłada się tutaj, iż jeżeli uczeń rozwiąże zadanie trudne, musi

rozwiązać również zadanie łatwe, oraz jeżeli uczeń nie rozwiązał zadania łatwego, to nie rozwiąże również zadania trudnego – przy założeniu, że zadania te mierzą te same osiągnięcia.

trafność
Najkrótsza definicja trafności mówi o tym, że pomiar jest trafny, jeżeli mierzy to, co zostało założone jako cel pomiaru. W naukach edukacyjnych często posługujemy się psychologiczną definicją trafności. O trafnym pomiarze mówi się wtedy, kiedy narzędzie pomiarowe wywołuje żądane procesy myślowe w głowie ucznia oraz daje możliwość zapisania dowodów, iż takie procesy myślowe zaistniały.

trudność zadania (testu)
Parametr (zob.) określający, w jakim stopniu zadanie jest trudne. Konwencjonalnie pod pojęciem trudności zadania kryje się parametr *estymowany* (zob.) na podstawie teorii odpowiedzi na zadanie testowe (parametr beta), lecz może on oznaczać również procent błędnie rozwiązanych zadań przez daną grupę uczniów w klasycznej teorii testu.

wariancja
Parametr (zob.) określający zróżnicowanie wartości zmiennej. Na przykład, jeżeli mamy do czynienia z klasą, w której wyniki uczniów ze sprawdzianu są bardzo zbliżone do siebie (powiedzmy, wszyscy uczniowie osiągają wyniki bliskie klasowej średniej), to wariancja wyników sprawdzianu w tej klasie jest bardzo mała. Jeżeli wyniki w klasie są bardzo zróżnicowane, tzn. odnotowano zarówno dużo wyników słabych, jak i dużo wyników bardzo dobrych, przy niewielkiej liczbie wyników średnich, to wariancja będzie duża. Wariancję liczymy w następujący sposób $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, gdzie n to liczba obserwacji, x_i wartość zmiennej dla danej obserwacji, a $\bar{x}$ średnia wartość zmiennej dla wszystkich badanych. Jest to zatem średnia sumy kwadratów różnic między pojedynczymi wartościami zmiennej a jej średnią. Tak zdefiniowaną wariancję określa się mianem wariancji z populacji. Istnieje również drugi sposób obliczania wariancji: $\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, czyli suma kwadratów różnic między wynikiem danego ucznia a średnią dzielona jest w tym wypadku przez liczbę badanych pomniejszoną o jeden. W takiej sytuacji mówi się o wariancji z próby.

współczynnik ufności
Prawdopodobieństwo, że oszacowany *przedział ufności* (zob.) będzie zawierał prawdziwą wartość *parametru* (zob.) w *populacji* (zob.). Dla współczynnika ufności 95-procentowy przedział ufności średniej z próby będzie w 95 na 100 przypadków zawierał prawdziwą wartość średniej w populacji.

wynik prawdziwy
Wynik prawdziwy zrównuje się ze średnią wyników otrzymanych przez ucznia, gdyby był on testowany (teoretycznie) za pomocą tego samego narzędzia nieskończoną liczbę razy.

zmienna niezależna (objaśniająca)
Zmienna lub zmienne, za pomocą których objaśniamy *zmienną zależną* (*objaśnianą*) (zob.). W matematycznym zapisie zmienne niezależne pojawiają się po prawej stronie równania, często oznaczane jako x_1, x_2,... x_n. W badaniach egzaminacyjnych są to często zmienne, za pomocą których pragniemy objaśnić na przykład umiejętności uczniów. Przykładowo: wykształcenie rodziców, status społeczny, motywacja albo wcześniejsze wyniki ucznia (jak w modelu EWD).

zmienna zależna (objaśniana)
Zmienna, którą objaśniamy za pomocą innych zmiennych. W matematycznym zapisie występuje po lewej stronie równania, najczęściej oznaczona jako *y*. W badaniach edukacyjnych bardzo czę-

sto zmienną zależną jest zmienna określająca umiejętności uczniów, na przykład wyniki egzaminów, które próbuje się opisać za pomocą zmiennych niezależnych (objaśniających), na przykład płeć, status społeczny rodziny ucznia czy też wcześniejsze jego wyniki. Zależnie od typu zmiennej zależnej stosujemy różnego rodzaju modele statystyczne. Dla zmiennych ciągłych, a za takie najczęściej przyjmujemy wyniki uczniów z egzaminów, używamy *regresji liniowej* (zob.) i pochodnych modeli, na przykład hierarchicznych modeli liniowych czy regresji z efektami stałymi. Jeśli zmienną zależną jest sukces ucznia na egzaminie (na przykład dwuwartościowa zmienna: 0 – „zdał", 1 – „nie zdał"), to używamy modeli regresji zwanych „probit" lub „logit". Popularne pakiety statystyczne zawierają odpowiednie modele niemal dla każdego rodzaju zmiennej zależnej.

Maciej Jakubowski

Absolwent Instytutu Socjologii Uniwersytetu Warszawskiego, adiunkt na Wydziale Nauk Ekonomicznych tej uczelni (doktorat w 2006 roku). Analityk w zespole zarządzającym programem PISA w OECD. Zainteresowania naukowe – badania z zakresu ekonomii edukacji i rynku pracy, a przede wszystkim metody statystyczne stosowane do ewaluacji projektów edukacyjnych. Autor kilkudziesięciu artykułów naukowych oraz rozdziałów w książkach opublikowanych w Polsce i zagranicą. Współautor metody edukacyjnej wartości dodanej dla polskich gimnazjów. Szczęśliwy mąż i ojciec. Pasjonuje się podróżami (próbował przejechać Afrykę Wschodnią na rowerze), muzyką (grał jazz na kontrabasie) oraz sportem (wspinanie i narciarstwo wysokogórskie).

Artur Pokropek

Absolwent Wydziału Pedagogicznego Uniwersytetu Warszawskiego oraz Instytutu Socjologii tej uczelni, doktorant na Wydziale Pedagogicznym UW oraz pracownik Instytutu Filozofii i Socjologii Polskiej Akademii Nauk. Pracuje w Zespole Interdyscyplinarnych Studiów nad Edukacją (ZISE). Współpracował przy międzynarodowych badaniach edukacyjnych, m.in. PISA, PIAAC, TALIS. Współpracuje z Centralną Komisją Egzaminacyjną w ramach grupy badawczej EWD. Zainteresowania badawcze – socjologia, ekonomia, polityka edukacji, metodologia badań społecznych, statystyka, psychometria. Szczęśliwy mąż, zapalony fotograf i zwolennik dobrej muzyki.