

Jacek Haman, Krzysztof Bulkowski

## Metodologia badania OECD PISA<sup>1</sup>

Pierwsza edycja badania OECD PISA przeprowadzona została w 2000 roku, ostatnia (na moment pisanego tego tekstu) – w roku 2025. Prace nad realizacją każdej edycji zaczynają się na około 2 lat przed okresem terenowej realizacji badania, kończą się ponad rok po nim. Stabilność rozwiązań metodologicznych jest kluczowa dla zapewnienia porównywalności wyników badania w czasie, ale – zwłaszcza w przypadku badania już mającego za sobą ponad ćwierć wieku historii – nie może ona oznaczać zamknięcia metodologii na jakąkolwiek ewolucję, ignorującego zmiany warunków prowadzenia badania, rozwój metod czy dostępność narzędzi.

Badanie OECD PISA prowadzone jest cyklicznie, dotychczas zasadniczo co trzy lata (edycje 2000–2018); ze względu na pandemię edycja planowana na rok 2021 została zrealizowana z rocznym opóźnieniem (2022), ostatnia edycja (2025) została przeprowadzona po upływie trzech lat, natomiast kolejne planowane są już w cyklu czteroletnim. Zasadniczym założeniem badania jest zapewnienie jak najlepszej porównywalności wyników, zarówno między uczestniczącymi w badaniu krajami/regionami, jak i między poszczególnymi edycjami. Realizacja tego założenia wymaga ścisłego przestrzegania procedur badawczych i rygorystycznego przestrzegania zasad metodologii oraz zachowania ciągłości rozwiązań metodologicznych (a także organizacyjnych) w czasie. Z drugiej strony jednak w ciągu ponad ćwierćwiecza funkcjonowania projektu badawczego zmieniające się warunki jego prowadzenia (np. liczba uczestniczących krajów), rozwój technik badawczych (możliwość przejścia z realizacji testów na papierze na badania przy użyciu komputerów, stosowania testów adaptacyjnych itp.), a także poszerzanie zakresu badania o kolejne dziedziny uzupełniające, powodowały ciągłą ewolucję metodologii badania.

W niniejszym opracowaniu przedstawiono główne założenia metodologii badania PISA – przede wszystkim te niezmiennie, stanowiące o ciągłości badania, ale również historię najważniejszych modyfikacji i ich znaczenie dla interpretacji wyników PISA, a także dla porównywania wyników w czasie.

---

<sup>1</sup> Niniejsze opracowanie stanowi kompilację wcześniejszych tekstów jego autorów, zawartych w raportach z kolejnych edycji badania OECD PISA: Biedrzycki i in., 2010; Federowicz, 2013; Federowicz i Sitek, 2017; Haman, 2020; Bulkowski i Sitek, 2020; Haman i Bulkowski, 2024, również Haman i Ostrowska, 2014, zaktualizowanych i uzupełnionych z uwzględnieniem aspektów związanych z edycją OECD PISA 2025.

## Populacja i próba

### Populacja docelowa badania PISA

W badaniu OECD PISA biorą udział uczniowie (a więc osoby objęte systemem szkolnym) w wieku 15 lat, a dokładniej – uczniowie, którzy ukończyli 15 lat w roku poprzedzającym badanie. Operacyjna definicja „piętnastolatka” przyjmowana w badaniu nie jest zatem w pełni zgodna z naturalną interpretacją słowa „piętnastolatek”: część uczestników badania w chwili wypełniania testów PISA ma już ukończonych lat szesnaście. Co do zasady badanie obejmuje uczniów wszystkich typów szkół istniejących w danym systemie szkolnym. Wyłączeni z badanej populacji są jedynie uczniowie opóźnieni w czasie trwania nauki (względem typowej sytuacji) o więcej niż 2 lata (czyli uczniowie klas VI i niższych w wypadku polskiej szkoły podstawowej lub ekwiwalentu tej klasy w innych systemach szkolnych), uczniowie nieznający w wystarczającym stopniu języka, w którym prowadzone jest badanie, oraz uczniowie, którzy ze względu na specjalne potrzeby edukacyjne mogliby mieć istotną trudność w wypełnieniu testów. Wyjątkowo dopuszczalny jest również pewien minimalny poziom wyłączeń (do 0,5% populacji) ze względów techniczno-organizacyjnych (np. uczniowie bardzo małych lub trudno dostępnych szkół, których włączenie do badań wiązałoby się z nieproporcjonalnie wysokimi kosztami); takie wyłączenia muszą być jednak ściśle uzgodnione przez krajowy zespół badawczy z konsorcjum międzynarodowym.

W Polsce od 2022 roku zasadniczą większość badanej populacji stanowią uczniowie klas I i II szkół ponadpodstawowych – liceów ogólnokształcących, techników, szkół branżowych I stopnia, jak również ogólnokształcących szkół muzycznych II stopnia i innych ogólnokształcących szkół artystycznych, oraz piętnastoletni uczniowie klas VIII i VII szkół podstawowych. Podobną strukturę miała populacja badana w pierwszej edycji PISA w roku 2000<sup>2</sup>. W latach 2003–2015 dominującą grupą w badanej populacji byli uczniowie klas III gimnazjów (i w niewielkim odsetku – niższych klas gimnazjów oraz pierwszych klas szkół ponadgimnazjalnych), natomiast w edycji 2018 – odbywającej się w warunkach trwającej reformy struktury szkół (likwidacja gimnazjów) – dominującą grupę stanowili uczniowie III klas wygaszanych gimnazjów (z udziałem również uczniów szkół ponadgimnazjalnych oraz podstawowych). Z kolei nieudana ostatecznie próba przesunięcia wieku startu edukacji szkolnych z 7 na 6 lat z lat 2012–2015 skutkowałą znaczącym wzrostem odsetka uczniów klas II szkół ponadpodstawowych w populacji badanej w edycjach 2022 i – zwłaszcza – 2025.

---

<sup>2</sup> Z zastrzeżeniem, że w roku 2000 naukę na poziomie ISCED3c prowadziły Zasadnicze Szkoły Zawodowe, a nie – jak współcześnie – Szkoły Branżowe I stopnia, a w kategorii średnich szkół zawodowych oprócz techników funkcjonowały także licea zawodowe.

## Próba

Badanie PISA jest realizowane na reprezentatywnej próbie losowej. Schemat doboru próby uczniów ma charakter dwustopniowego doboru warstwowego z zastosowaniem losowania systematycznego, w którym pierwszym stopniem doboru był wybór szkoły, drugim zaś – losowanie uczniów z uprzednio wylosowanych szkół (również w trybie systematycznego losowania warstwowego). Pierwszy etap losowania – losowanie szkół – realizowany jest przez konsorcjum międzynarodowe prowadzące badanie; rolą polskiego zespołu PISA jest dostarczenie listy identyfikatorów szkół wraz z danymi niezbędnymi do przeprowadzenia losowania. Drugi etap – losowanie uczniów – realizowany jest przez krajowy zespół PISA przy użyciu narzędzia dostarczonego przez konsorcjum (co w szczególności oznacza, że krajowy zespół nie ma wpływu na wynik losowania, natomiast nie ma potrzeby przekazywania na zewnątrz dodatkowych danych polskich uczniów).

System warstwowania próby szkół uzgadniany jest przez krajowe zespoły badawcze z konsorcjum międzynarodowym i dostosowywany w każdym przypadku do lokalnych warunków (a więc lokalnych potrzeb oraz możliwości związanych z dostępnością niezbędnych danych). W każdym przypadku w losowaniu szkół stosowany jest podział na warstwy jawne (*explicite*) – w Polsce były one wyznaczane ze względu na typ szkoły. W ramach tych warstw losowanie prowadzone jest osobno. Dzięki stosowaniu metody losowania systematycznego możliwe jest wykorzystanie również warstw *implicite* (określających uporządkowanie operatu losowania) – przykładowo w Polsce w 2025 roku były nimi (w hierarchii od najważniejszej do najmniej ważnej): w warstwie szkół zawodowych podział na szkoły branżowe I stopnia i technika; we wszystkich warstwach: publiczność (szkoły publiczne, szkoły prywatne), poziom urbanizacji powiatu, w którym ulokowana jest szkoła; skład szkoły ze względu na płeć (szkoły z dominującym udziałem dziewcząt, szkoły mieszane, szkoły z dominującym udziałem chłopców); ostatnią cechą (ciągłą) porządkującą operat losowania była wielkość szkoły (liczba uczniów). Prawdopodobieństwo wylosowania danej szkoły do próby zależne jest od szacowanej liczby jej piętnastoletnich uczniów (tzw. próba PPS – *probability proportional to size*); pewne odstępstwa od tej reguły dotyczą jedynie szkół małych (dla zachowania równego prawdopodobieństwa wylosowania do próby poszczególnych uczniów) oraz bardzo małych (ze względów logistycznych).

Losowanie uczniów do badania również opiera się na schemacie losowania systematycznego z warstwami *implicite* (uporządkowanie operatu ze względu na płeć i klasę); zgodnie z tym schematem przydziela się uczniom również konkretne warianty zestawów testowych, a także determinuje się udział ucznia w ewentualnych modułach opcjonalnych badania.

Co do zasady liczba uczniów losowanych w poszczególnych szkołach jest (w danej edycji badania i w danym kraju) stała – w Polsce w ostatnich edycjach było to 46 uczniów; jeśli w wylosowanej szkole nie uczyło się tylu piętnastolatków, do próby włączani byli wszyscy piętnastoletni uczniowie danej szkoły. W efekcie prawdopodobieństwo wylosowania do próby każdego ucznia należącego do badanej populacji w danym kraju jest w przybliżeniu takie samo (pozostałe, niewielkie zróżnicowanie tych prawdopodobieństw jest ostatecznie niwelowane w procedurze ważenia danych).

Liczbę szkół losowanych do próby ustala się tak, by – biorąc pod uwagę czynniki takie jak przewidywane poziomy realizacji próby (odsetek wylosowanych uczniów faktycznie biorących udział w badaniu) i odsetek szkół małych w ogóle szkół – osiągnąć ostatecznie przyjętą minimalną liczbę uczniów uczestniczących w badaniu. Wielkość ta była różna w różnych edycjach badania – od około 4000 uczniów w edycji PISA 2000, do ponad 7000 w ostatnich edycjach badania, co wymagało uprzedniego wylosowania od 150 do blisko 300 szkół.

Ze względu na to, że nie we wszystkich wylosowanych do próby szkołach udaje się zrealizować badanie, każda szkoła w próbie ma wskazane dwie szkoły rezerwowe, w których można przeprowadzić badanie w sytuacji odmowy udziału szkoły z próby zasadniczej. Natomiast w przypadku próby uczniów nie przewiduje się możliwości zastępowania wylosowanego ucznia – w przypadku braku zgody rodziców, odmowy ze strony ucznia lub nieobecności w dniu badania wylosowany uczeń definitywnie „wypada” z próby (tzw. realizacja próby „do wyczerpania”).

## Dobór próby a błędy losowe i nielosowe

Próba w badaniu PISA jest **próbą losową**, co oznacza, że dla każdego ucznia z badanej populacji można wyznaczyć znane i niezerowe prawdopodobieństwo, że zostanie on wylosowany do próby. Próba losowa jest z definicji **próbą reprezentatywną**, jeśli prawdopodobieństwo trafienia do niej jest równe dla wszystkich członków populacji; jeśli nie jest – staje się reprezentatywna po nadaniu jej elementom odpowiednich **wag**, odwrotnie proporcjonalnych do prawdopodobieństwa trafienia do próby (tzw. ważenie poststratyfikacyjne). Wyniki uzyskane na próbie reprezentatywnej można uogólniać na populację, z której została ona wylosowana, jednakże są one obarczone błędami losowymi. Wielkość tych błędów oczywiście nie jest znana, ale opierając się na zasadach rachunku prawdopodobieństwa, można określić wielkość, której błąd losowy nie przekroczy z określonym prawdopodobieństwem (zwykle 95%). Wyznacza się w ten sposób przedziały ufności dla poszczególnych wyników, których szerokość (a ściślej jej połowę) traktuje się zwykle jako

dokładność oszacowania<sup>3</sup>. Dokładność oszacowań (a zatem potencjalna wielkość błędów losowych) zależy od dwóch<sup>4</sup> czynników: wielkości próby (im większa próba, tym większa dokładność oszacowań) oraz schematu jej doboru, który może się przyczynić zarówno do zwiększenia, jak i zmniejszenia dokładności w porównaniu z tzw. prostą próbą losową.

Schemat doboru próby uczniów w badaniu PISA ma charakter dwustopniowego doboru warstwowego z zastosowaniem losowania systematycznego, w którym pierwszym stopniem doboru był wybór szkoły, drugim zaś – losowanie uczniów z uprzednio wylosowanych szkół. Zastosowanie doboru warstwowego umożliwia zmniejszenie skali błędów losowych<sup>5</sup>, natomiast dwustopniowy charakter losowania prowadzi do ich zwiększenia w porównaniu z hipotetycznym badaniem realizowanym na próbie losowej prostej. Wielkość błędów losowych w porównaniu z próbą prostą jest wypadkową tych dwóch czynników, przy czym znacznie silniejszy jest efekt związany z doborem wielostopniowym: błędy losowe wyników uzyskiwanych na liczącej ok. 5000 elementów próbie mają skalę odpowiadającą kilkusetelementowej próbie prostej. Nie oznacza to oczywiście, że lepszą próbą byłaby właśnie próba prosta: ze względów organizacyjnych znacznie łatwiejsze i tańsze do przeprowadzenia jest badanie na próbie 5000 uczniów zgrupowanych w 200 szkołach niż, powiedzmy, 800 uczniów, z których każdy jest uczniem innej szkoły. Niezależnie od tego badanie wielu uczniów w tej samej szkole jest niezbędne, by móc w badaniu nie tylko charakteryzować zatomizowaną społeczność uczniów, lecz także wypowiadać się o systemie edukacyjnym, w skład którego wchodzi całe społeczności szkolne.

Zastosowanie złożonego schematu doboru próby oznacza jednak, że do określania dokładności oszacowań nie można stosować standardowych technik obliczeniowych – proste wykorzystanie metod zaimplementowanych w typowych programach statystycznych prowadziłoby do niedoszacowania błędów. Z tego względu błędy losowe wyników w badaniu PISA szacowane są z wykorzystaniem metod replikacyjnych, a ściślej – techniki *balanced random replicates* w variancie Faya. Metoda ta pozwala na wyznaczanie przedziałów ufności oraz weryfikację hipotez statystycznych z uwzględnieniem specyfiki przyjętego schematu doboru próby. Mówiąc o reprezentatywności próby oraz dokładności oszacowań, należy pamiętać o kilku zasadach:

<sup>3</sup> Dokładność oszacowania można podawać również jako błąd standardowy (*standard error*, s.e.); połowa szerokości przedziału ufności odpowiada na ogół 1,96 błędowi standardowemu.

<sup>4</sup> Zależy ona oczywiście również od rozkładu populacyjnego badanej cechy, ta jednak jest czynnikiem niezależnym od charakterystyki próby losowej.

<sup>5</sup> Należy przy tym pamiętać, że uwzględnienie lub nieuwzględnienie jakiegokolwiek cech w warstwowaniu próby nie decyduje o tym, czy próba jest „reprezentatywna” ze względu na daną cechę; reprezentatywność próby ma zapewnić przede wszystkim jej losowy charakter.

- Dokładność oszacowań zależy od liczebności próby, natomiast jest praktycznie niezależna od tego, jaką część populacji stanowi próba<sup>6</sup>. W konsekwencji, przy tym samym schemacie doboru próby, próba o liczebności 5000 osób da taką samą dokładność oszacowania niezależnie od tego, czy wylosowana ona była z populacji liczącej sto tysięcy, milion czy dziesięć milionów ludzi.
- Schemat doboru próby, a także sposób jej warstwowania czy ewentualne sub- i nadreprezentacje, o ile są przeprowadzone prawidłowo i właściwie uwzględnione przy ważeniu próby, nie mają wpływu na jej reprezentatywność (choć mają wpływ na wielkości błędów losowych). W konsekwencji w pełni dopuszczalne jest porównywanie wyników uzyskanych z prób wylosowanych przy użyciu różnych schematów doboru czy wykorzystujących różne warstwowanie. Pomimo tego w kolejnych cyklach badania PISA dąży się do utrzymania stałych schematów doboru i warstwowania próby, jednakże ewentualne ich zmiany nie stanowią przeszkody w porównywaniu wyników badania PISA z różnych lat.

## Realizacja badania a błędy nielosowe

Oprócz błędów losowych wyniki każdego badania mogą być obciążone różnorodnymi błędami nielosowymi (systematycznymi). Ich źródłem mogą być: nietrafność narzędzi pomiarowych, błędy proceduralne w realizacji badania, a przede wszystkim – niepełna realizacja próby. W przeciwieństwie do błędów losowych, których wartości wprawdzie nie znamy, ale ich skalę możemy przewidzieć, błędy nielosowe nie poddają się opisowi probabilistycznemu i nie da się łatwo wskazać, jaką wielkość czy kierunek mogą osiągnąć. Błędów takich nie da się nigdy całkowicie uniknąć, jednak staranne przygotowanie i realizacja badania ma służyć ich minimalizacji.

Standardy techniczne badania PISA wymagają osiągnięcia minimum 80% poziomu realizacji próby i warunek taki w Polsce zawsze był spełniany. Wielkość tę warto porównać z poziomami realizacji prób w typowych badaniach społecznych na próbach reprezentatywnych dorosłej ludności Polski, które w przypadku najlepiej realizowanych badań akademickich zwykle nie przekraczają 45%, a w przypadku badań komercyjnych (np. sondaże przedwyborcze) nie osiągają nawet 30% założonej próby.

Z różnych względów można się spodziewać, że uczniowie wylosowani do próby, ale niebiorący udziału w badaniu, mają przeciętnie niższy poziom umiejętności niż uczniowie ostatecznie

<sup>6</sup> Pomijamy tutaj sytuację, gdy próba stanowi rzeczywiście znaczącą część (np. 80%) populacji.

wypełniający testy (absencja jest bardziej prawdopodobna np. w przypadku uczniów słabiej zmotywowanych bądź z innych względów częściej opuszczających lekcje). Na tyle, na ile różnice w prawdopodobieństwie udziału wylosowanych uczniów w badaniu można powiązać z ich znanymi cechami (jak płeć, klasa lub szkoła, do której uczęszczają), wpływ niepełnej realizacji próby jest ograniczany poprzez stosowanie, procedur „ważenia postrealizacyjnego”, jednakże nie mogą one zniwelować jej wpływu na wyniki całkowicie. Ocena wielkości wpływu niepełnej realizacji próby na uzyskane wyniki badania była przedmiotem analiz polskiego zespołu PISA w czasie poprzednich edycji badania. Uzyskane przez nas wyniki były szerzej przedstawione w raportach z PISA 2006, 2009 i 2012 (Bartnik i in., 2007, s. 7–8; Biedrzycki i in., 2010, s. 16–18; Federowicz, 2013, s. 14–15). Dzięki wykorzystaniu danych o wynikach na sprawdzianie po szkole podstawowej uczniów wylosowanych do badania i porównaniu uczniów, którzy ostatecznie wzięli udział w badaniu lub nie, udało się oszacować, że efekt niepełnej realizacji próby dla średnich wyników polskich uczniów wynosi od około 1 do 3 pkt (o tyle prawdopodobnie średnie wyniki byłyby niższe, gdyby w badaniu wzięli udział wszyscy wylosowani uczniowie), a więc był mniejszy niż skala przewidywanych błędów o charakterze czysto losowym.

W kwestii porównań międzynarodowych brak jest danych dotyczących wpływu niepełnej realizacji próby dla średnich wyników uzyskanych we wszystkich krajach biorących udział w badaniu, można jednak przypuszczać, że podobne mechanizmy jak w Polsce mogą działać także w pozostałych krajach. Ponieważ poziom realizacji próby w Polsce jest podobny jak w innych krajach uczestniczących w badaniu (wyłączając kraje, w których nie zostały spełnione standardy wyznaczone przez OECD), można oczekiwać, że również wielkość efektów niepełnej realizacji próby będzie podobna. Tak więc niepełna realizacja próby nie stanowi istotnej przeszkody dla dokonywania międzynarodowych porównań wyników – choć oczywiście stanowi ona pewien dodatkowy czynnik niepewności.

## Pomiar i skalowanie<sup>7</sup>

Wyniki badania PISA – a więc oszacowania kompetencji w poszczególnych obszarach badania – przypisywane są poszczególnym uczniom na podstawie procedury skalowania, uwzględniającej liczbę poprawnie rozwiązanych zadań oraz ich trudność. Zastosowanie tych procedur daje szereg istotnych korzyści, w porównaniu z ocenianiem kompetencji przez np. zwykłe zliczanie odsetka poprawnie rozwiązanych zadań. W szczególności możliwe jest ocenianie na tej samej skali (a więc

<sup>7</sup> Tekst z tej sekcji zawiera fragmenty powtórzone bezpośrednio lub z pewnymi modyfikacjami za rozdziałami jednego ze współautorów w raportach z badania PISA 2009 (Biedrzycki i in., 2010) i PISA 2012 (Federowicz, 2013), PISA 2015 (Federowicz i Sitek, 2017) oraz PISA 2018 (Sitek i Ostrowska, 2020).

zapewnienie porównywalności wyników) uczniów, którzy rozwiązywali częściowo różne zestawy zadań testowych (a w pewnym zakresie nawet uczniów, którzy rozwiązywali całkowicie różne zestawy zadań), bez przyjmowania trudnego do spełnienia założenia o identycznym poziomie ich trudności. Dzięki temu w badaniu PISA możliwe jest wykorzystanie znacznie większej liczby różnorodnych zadań, niż gdyby wszyscy uczniowie mieli wykonywać te same zestawy testowe; możliwe jest również porównywanie wyników różnych edycji badania, pomimo że jedynie część zadań (tzw. zadania kotwiczące) jest powtarzanych w kolejnych edycjach.

Gdy w 2000 roku przeprowadzano pierwszą edycję badania OECD PISA, skalowanie wyników testów oparte na modelach teorii odpowiedzi na pytanie testowe (IRT) było w Polsce nowością. Od tego czasu IRT było wykorzystywane w szeregu badań krajowych i międzynarodowych, jak również było przedmiotem wielu publikacji, zarówno o charakterze wprowadzającym, jak i bardzo zaawansowanym. Z tego względu poniższe omówienie skalowania w badaniu PISA dzieli się na trzy sekcje:

- Sekcja pierwsza przeznaczona jest przede wszystkim dla czytelników niemających na co dzień do czynienia z pomiarem w paradygmacie IRT – przedstawiamy w niej najważniejsze elementy tej koncepcji, niezbędne dla zrozumienia istoty pomiaru kompetencji w badaniu PISA; zainteresowanych jej głębszym poznaniem odsyłamy do dostępnej w Polsce literatury, w szczególności do książek (Jakubowski i Pokropek, 2009; Pokropek, 2015).
- W drugiej sekcji podajemy najważniejsze informacje na temat szczegółowych rozwiązań modeli IRT stosowanych w badaniu OECD PISA z podkreśleniem tych elementów, które zostały zmodyfikowane w ostatnich edycjach badania OECD PISA.
- W sekcji trzeciej omawiamy jeden ze sposobów prezentacji wyników dotyczących umiejętności uczniów – podział skali na poziomy umiejętności. Uwzględnienie tego sposobu w analizie pozwala na interpretowanie wyników PISA w kategoriach praktycznych, opisywalnych w naturalnym języku umiejętności i kompetencji.

## Założenia ogólne – podstawy modeli IRT

W badaniu PISA skalowanie wyników testu opiera się na teorii odpowiedzi na pytanie testowe (IRT – *Item Response Theory*), a ściślej – na uogólnionym modelu Rascha. Koncepcja ta odwołuje się do następujących założeń:

- To, czy dany uczeń rozwiąże prawidłowo dane zadanie, jest zdarzeniem losowym.
- Prawdopodobieństwo zajścia tego zdarzenia determinowane jest przez dwa czynniki:
  - poziom umiejętności ucznia,
  - poziom trudności zadania.

Zakłada się przy tym określoną postać funkcji wiążącej prawdopodobieństwo rozwiązania zadania o danej trudności z poziomem umiejętności ucznia. W modelu Rascha jest to, zasadniczo, funkcja logistyczna; poszczególne warianty modeli IRT różnią się od siebie głównie uwzględnianiem pewnych dodatkowych parametrów tej funkcji – obok samej trudności zadania także jego mocy dyskryminacyjnej<sup>8</sup>, a w przypadku zadań zamkniętych – prawdopodobieństwo udzielenia poprawnej odpowiedzi poprzez zgadywanie. Zwyczajowo określa się poziom trudności zadania i poziom umiejętności badanego na tej samej skali, przyjmując, że badany o poziomie kompetencji  $k$  rozwiąże zadania o trudności  $k$  z prawdopodobieństwem równym  $\frac{1}{2}$ .

Określenia „poziom umiejętności” i „poziom trudności” odpowiadają podstawowemu zastosowaniu modelu Rascha – pomiarowi kompetencji. W innych jego zastosowaniach mówimy w tym miejscu o natężeniu cechy oraz o charakterystyce konkretnego jej wskaźnika. Model Rascha w badaniu PISA używany jest – poza samym badaniem kompetencji – do konstruowania szeregu skal cech kontekstowych.

- Zarówno poziom umiejętności poszczególnych badanych, jak i poziom trudności poszczególnych zadań (i ewentualnie ich moc dyskryminacyjna) traktowane są jako zmienne ukryte (latentne) – ich estymacja jest celem procesu skalowania.

W procesie skalowania jednocześnie szacowane są poziomy trudności zadań oraz kompetencje badanych – polega to na poszukiwaniu (za pomocą przede wszystkim procedur iteracyjnych) takich kombinacji ich wartości, które z największym prawdopodobieństwem prowadzą do uzyskania zaobserwowanych wyników badania (estymacja metodami największej wiarygodności). Drugim elementem procesu skalowania jest ocena zgodności założeń modelu z danymi obserwowanymi. Przykładowo, może się okazać, iż bardzo trudno jest utrzymać założenie, że szanse na rozwiązanie

<sup>8</sup> Moc dyskryminacyjna zadania określa, jak silna jest zależność prawdopodobieństwa rozwiązania zadania od poziomu umiejętności ucznia: im jest ona wyższa, tym szybciej wzrasta prawdopodobieństwo rozwiązania wraz ze wzrostem umiejętności, natomiast w przypadku zadań o słabej mocy dyskryminacyjnej różnica szans na rozwiązanie zadania między uczniami o wysokich i niskich kompetencjach może być niewielka.

danego zadania wynikają z poziomu tej samej umiejętności, która odpowiada za umiejętność poprawnego rozwiązania pozostałych zadań. W takiej sytuacji może się okazać, że trafniejsze wyniki uzyskamy, pomijając w analizie dane odnoszące się do tego zadania.

Estymacja trudności zadań może być dokonywana na całości danych z badania, możliwe jest jednak także wykorzystanie danych o poziomie trudności zadań oszacowanych uprzednio. Możliwość ta wykorzystywana jest na kilka sposobów; w szczególności:

- Użycie zadań kotwiczących o trudności oszacowanej już w poprzednich cyklach badania PISA pozwala zakotwiczyć skale PISA względem wcześniejszych edycji badania, a tym samym osiągnąć porównywalność i współmierność wyników kolejnych cykli PISA. Osiągnięcie tego efektu wymaga jednak wyskalowania zadań kotwiczących na odpowiednio bogatym materiale – z tego względu pełna porównywalność wyników kolejnych edycji badania PISA dla danej dziedziny możliwa jest jedynie od momentu, gdy dana dziedzina była głównym przedmiotem edycji (jak np. rozumienie czytanego tekstu w edycji PISA 2000, matematyka – PISA 2003, rozumowanie w naukach przyrodniczych – PISA 2006); porównywanie wyników wcześniejszych edycji wiąże się z większym ryzykiem błędów losowych.
- Skalowanie trudności zadań odbywa się wyłącznie z użyciem wyników pochodzących z podstawowej populacji badanych, a więc – populacji piętnastolatków. W badaniach uzupełniających projekt międzynarodowy – jak w prowadzonych w części poprzednich edycji polskich badaniach uczniów szkół ponadgimnazjalnych – wykorzystywane są trudności zadań oszacowane w międzynarodowej części badania. W ten sposób jednocześnie osiągane są dwa cele: ocena umiejętności w dodatkowych populacjach na tych samych skalach co w przypadku piętnastolatków; a jednocześnie odseparowanie podstawowego badania międzynarodowego od dodatkowych elementów badania specyficznych dla poszczególnych krajów.
- Możliwe jest szacowanie parametrów zadań z wykorzystaniem jedynie części badanej próby, a następnie skalowanie umiejętności wszystkich badanych uczniów w oparciu o tak uzyskane parametry zadań. Metoda ta była wykorzystywana w części poprzednich edycji badania PISA ze względu na mniejsze wymagania co do mocy obliczeniowej.

Istotną korzyścią płynącą z zastosowania modelu Rascha jest również możliwość oceny na tej samej skali badanych, którzy wykonywali częściowo różne zestawy zadań. W ten sposób możliwe jest

wykorzystanie w badaniu znacznie większej liczby zadań, a więc zbadanie szerszego spektrum podobszarów poszczególnych umiejętności. Co więcej – możliwe jest celowe przydzielanie różnym uczniom zadań o innym poziomie trudności, dostosowanym do ich poziomu umiejętności (testowanie adaptatywne). Wdrażanie takiego rozwiązania rozpoczęto w edycji badania 2018, a od edycji 2025 stosowane jest we wszystkich dziedzinach.

Skale (umiejętności badanych i trudności zadań) w modelu Rascha mają charakter skal przedziałowych. Pozwalają zatem na interpretowanie i porównywanie wielkości różnic między poszczególnymi wynikami (np. między średnimi dla krajów, średnimi dla typów szkół, wynikami poszczególnych badanych). Skale te nie mają jednak obiektywnego punktu zerowego – a zatem nie jest możliwe określanie proporcji między wynikami. Tak więc, przykładowo, bezsensowne byłoby stwierdzenie, że „kraj A uzyskał wynik o 20% lepszy od kraju B”. Jednocześnie poziom umiejętności wyrażony w punktach PISA ma charakter relatywny (i nie odnosi się do żadnych obiektywnie zdefiniowanych oczekiwań co do tego, co powinni wiedzieć lub umieć badani). Skale skonstruowane są w ten sposób, by wartość 500 pkt odpowiadała średniej wyników krajów OECD w badaniu PISA 2000 oraz by jeden punkt odpowiadał jednej setnej odchylenia standardowego wyników w populacji krajów OECD w badaniu PISA 2000 (choć ze względu na dokładności oszacowania wsteczną porównywalność wyników PISA należy ograniczyć do edycji, w której dana dziedzina była dziedziną wiodącą).

Probabilistyczny charakter *Item Response Theory* oznacza także, że przy interpretacji wyników badania bierze się pod uwagę, że dwóch uczniów o tym samym rzeczywistym poziomie umiejętności może uzyskać w teście różne wyniki i vice versa – dwie osoby, które uzyskały taki sam wynik, mogą w rzeczywistości mieć różne poziomy umiejętności. Innymi słowy, probabilistyczny charakter odpowiedzi na bodziec testowy jest drugim, obok reprezentatywnego charakteru badania (błędy związane z próbą), źródłem błędów losowych w wynikach badania PISA. Sposobem uwzględniania tych błędów w analizie jest wykorzystanie do szacowania poziomów umiejętności uczniów estymatorów wartości prawdopodobnych (*plausible values*, PV). Należy jednak mieć świadomość, że to, czym różnią się modele IRT od innych podejść do testowania (jak np. klasyczna teoria testu), to nie sam fakt, że wyniki pomiaru obarczone są niepewnością – od tej bowiem nie jest wolny żaden model. Ważne jest to, że modele IRT pozwalają – w pewnym zakresie – oszacować skalę tej niepewności i zamknąć ją w modelu probabilistycznym.

Celem badania OECD PISA jest znalezienie najlepszych oszacowań przeciętnych wyników dla określonych populacji i dokonywanie porównań międzygrupowych: między uczniami z różnych

krajów, szkół różnych typów, między dziewczętami a chłopcami itd. W przeciwieństwie do egzaminów celem badania nie jest dostarczanie danych o wynikach konkretnych uczniów biorących w nim udział. Z tego względu zarówno organizacja testu, jak i procedury skalowania dobrane są taki sposób, aby zminimalizować błędy oszacowań średnich i wariancji poziomów umiejętności wyznaczanych dla podzbiorowości uczniów, nawet jeśli dzieje się to kosztem większych błędów oszacowań indywidualnych.

Możliwość uzyskania poprawy jakości oszacowań dla zbiorowości kosztem dokładności oszacowań indywidualnych może się wydawać paradoksalna. Przykładem takiego mechanizmu – na poziomie organizacji testu – jest wspomniana już parokrotnie zasada rozdzielania pełnej puli zadań między poszczególne zestawy testowe. Łatwo zauważyć, że jest to działanie korzystne z punktu widzenia oceny przeciętnego poziomu umiejętności dla zbiorowości: dzięki wykorzystaniu większej liczby zadań możemy uwzględnić więcej częściowych składowych umiejętności czy wiedzy wchodzącej w skład danej dziedziny, zaś z drugiej strony ewentualne specyficzne czynniki sprzyjające lub utrudniające rozwiązanie danego zadania, niepowiązane z badaną dziedziną, będą miały mniejszy systematyczny wpływ na ogólny wynik. Równie łatwo przy tym zauważyć, że zasada ta utrudnia porównywanie wyników poszczególnych uczniów, którzy rozwiązywali przecież odmienne zestawy zadań<sup>9</sup>.

Na poziomie skalowania przykładanie większej wagi do poprawności oszacowań zbiorowych niż indywidualnych przejawia się, z jednej strony, w stosowaniu estymatorów *plausible values*, pozwalających na lepsze oszacowanie wariancji oraz błędu losowego dla średnich, kosztem dodatkowego błędu losowego na poziomie oszacowań indywidualnych, z drugiej zaś strony – w uwzględnianiu w modelach skalowania (a ściślej w modelach wyznaczania rozkładów prawdopodobieństwa, z których losowane są wartości *plausible values*) danych kontekstowych, pochodzących z kwestionariuszy ucznia. Działanie takie byłoby absolutnie nie do przyjęcia, gdyby celem badania była ocena indywidualnego poziomu umiejętności danego ucznia: trudno byłoby zaakceptować procedurę, w której ocena ucznia zależałaby nie tylko od odpowiedzi, jakich udzieli na egzaminie, lecz także od np. wykształcenia jego rodziców. W sytuacji, gdy celem badania jest ocena przeciętnych wyników dla zbiorowości, postępowanie takie pozwala jednak zmniejszyć poziom błędów losowych oszacowań. Co więcej, oszacowania poziomów poszczególnych umiejętności podawane są – na podstawie danych kontekstowych oraz wyników testów z zakresu

<sup>9</sup> Warto jednak zwrócić uwagę, że właśnie dzięki zastosowaniu modeli IRT porównywanie indywidualnych wyników uczniów, którzy rozwiązywali różne zestawy zadań, jest możliwe – jednakże byłoby ono obciążone większą niepewnością niż w przypadku porównywania wyników uczniów rozwiązujących dokładnie te same zadania ze względu na konieczność uwzględnienia dodatkowego składnika błędu losowego – tzw. błędu łączenia.

pozostałych umiejętności – nawet w przypadku uczniów, którzy w ogóle nie rozwiązywali zadań z zakresu danej umiejętności. Na poziomie indywidualnym takie oszacowania (które w tym wypadku lepiej byłoby nazywać przewidywaniami) są, oczywiście, zwykle obciążone dużym błędem; jednakże przy szacowaniu parametrów poziomu umiejętności dla zbiorowości lepsze (dokładniejsze) oszacowania uzyskuje się, uwzględniając zarówno oszacowania pochodzące z rzeczywistego testowania danej umiejętności, jak i te wyznaczone jedynie na podstawie danych kontekstowych i z pozostałych testów.

Dla celu losowania wartości *plausible values* przyjmuje się, że dla każdej kombinacji branych pod uwagę zmiennych kontekstowych<sup>10</sup> rozkład poziomu umiejętności uczniów jest rozkładem normalnym. Model łączący wartości zmiennych kontekstowych ze średnią i wariancją tego rozkładu wyznaczany jest za pomocą modeli strukturalnych (modeli regresji zmiennych ukrytych), co z kolei, zgodnie z regułami rachunku prawdopodobieństwa, pozwala wyznaczyć dla każdego badanego warunkowe rozkłady prawdopodobieństwa poziomu umiejętności ze względu na jego wartości zmiennych kontekstowych oraz wyniki testu. Z tego rozkładu losowanych jest po 10 *plausible values* dla każdej z mierzonych kompetencji.

Modele łączące wartości zmiennych kontekstowych z rozkładami umiejętności były wyznaczane dla każdego kraju biorącego udział w badaniu osobno – wynika to nie tylko z faktu, że zależności tego typu mogą być silnie warunkowane kulturowo, ale że formalnie te same cechy kontekstowe mogą w różnych krajach reprezentować niekoniecznie porównywalne stany faktyczne (przykładowo – za wykształceniem rodziców na tym samym formalnie poziomie ISCED mogą stać systemy edukacyjne bardzo silnie różniące się poziomem nauki). Oznacza to także, że ten etap procedury skalowania nie ma wpływu na oszacowania średnich osiągnięć na poziomie całego kraju.

## Skalowanie wyników badania PISA: rozwiązania szczegółowe i ich ewolucja

Choć ogólne założenia stojące za skalowaniem wyników OECD PISA są – podobnie jak zasady doboru próby – stałe, to szczegółowe rozwiązania podlegają pewnej ewolucji, związanej z jednej strony z rozwojem teorii skalowania, z drugiej zaś – ze wzrostem dostępnej mocy obliczeniowej komputerów (skalowanie danych z tak dużego badania jest niezwykle złożone obliczeniowo). Największe zmiany w tym zakresie zostały wprowadzone w edycji PISA 2015, a ich zakres i konsekwencje zostały bardziej szczegółowo omówione w raporcie z tego badania (Federowicz i Sitek, 2017, s. 10–13); tutaj powtarzamy jedynie najważniejsze informacje.

<sup>10</sup> Ścisłej rzecz biorąc, zmienne kontekstowe są poddawane analizie głównych składowych, po czym w modelowaniu uwzględniane są składowe odpowiedzialne łącznie za 80% ich łącznej wariancji dla danego kraju.

- Do skalowania wyników PISA stosowany jest zasadniczo jednoparametryczny model IRT (z uwzględnieniem jedynie trudności zadania), w przypadku zadań zamkniętych uwzględniający również prawdopodobieństwo przypadkowego wybrania prawidłowej odpowiedzi. Od cyklu PISA 2015 wykorzystywany jest również model dwuparametryczny (trudność oraz moc dyskryminacji)<sup>11</sup>.
- Do edycji PISA 2012 szacowanie trudności zadań prowadzone było na podpróbie, do której włączano losowo po 500 przypadków z każdego kraju/regionu (do roku 2009 – wyłącznie z krajów OECD), jedynie z bieżącego cyklu badania. Od cyklu PISA 2015 szacowanie trudności zadań przeprowadzane jest na bazie wyników uzyskanych przez wszystkich uczniów z wszystkich dotychczasowych edycji badania.

Ponadto od edycji PISA 2015 wprowadzono zmiany związane ze sposobem traktowania zadań, do których uczeń nie dotarł (zadania, na które nie udzielono żadnej odpowiedzi, znajdujące się na końcu zestawu testowego) oraz z zadaniami, których relatywny poziom trudności w danym kraju odbiegał od poziomu trudności wyznaczonego wspólnie dla całego badania (o ile we wcześniejszych edycjach badania możliwe było jedynie pominięcie takiego zadania przy skalowaniu wyników uczniów z danego kraju, od edycji 2015 możliwe jest również wykorzystanie danych o jego rozwiązaniu do szacowania poziomu umiejętności uczniów, jednakże w taki sposób, że nie ma ono wpływu na średnią wyników uczniów danego kraju na skali międzynarodowej).

Zmiany szczegółowych rozwiązań modeli skalowania stosowanych w PISA prowadzą do poprawy dopasowania modeli do danych, a więc do coraz lepszego wyskalowania wyników kolejnych cykli badania. „Lepsze” metody skalowania to jednak również „inne” metody: każda zmiana – nawet na lepsze – musi być zatem rozpatrywana także jako pewne naruszenie ciągłości badania i potencjalne zagrożenie dla porównywalności wyników w czasie. Porównanie wyników skalowania wcześniejszych edycji badania PISA za pomocą modeli stosowanych oryginalnie, jak i udoskonalonych modeli wprowadzonych w kolejnych cyklach badania wskazuje jednak, że zmiany są niewielkie, a średni wynik polskich uczniów osiągany przy różnych modelach skalowania nie różni się o więcej niż (w zależności od dziedziny) 2 do 4 pkt.

<sup>11</sup> Model dwuparametryczny stosowano w przypadku tych zadań, dla których uzyskiwano dzięki temu znaczącą poprawę dopasowania; jeśli wyniki uzyskane w modelu jednoparametrycznym były dopasowane w stopniu zadowalającym, pozostawano przy modelu jednoparametrycznym.

## Poziomy umiejętności

Wartości punktowe, będące miarami umiejętności uczniów i trudności zadań, nie mają bezpośredniej interpretacji innej niż poprzez odniesienie do średniej i odchylenia standardowego w populacji: nie mają one żadnego zakotwiczenia w obiektywnych umiejętnościach czy wymaganiach. Gdyby komunikowanie wyników badania ograniczyć do wyskalowanych wyników, można by wprawdzie wskazywać, jacy uczniowie wykazali się wyższymi, a jacy niższymi kompetencjami w danej dziedzinie, czy nawet porównywać wielkości różnic umiejętności (np. średnich wyników różnych krajów, średnich wyników chłopców i dziewcząt itp.), nic by jednak nie można było powiedzieć o tym, co uczniowie umieją, jakie problemy potrafią rozwiązywać, a jakie sprawiają im trudności. Tę informację można byłoby uzyskać, analizując rozwiązania konkretnych zadań, ale, po pierwsze, treść większości zadań PISA nie jest ujawniana publicznie (ze względu na ich wykorzystanie w kolejnych cyklach badania), po drugie zaś i tak pozostawałaby kwestia interpretacji syntetycznego wyniku całego testu.

Rozwiązanie tego problemu możliwe jest dzięki temu, że trudności zadań i umiejętności uczniów metody IRT określają na tej samej skali. O uczniu, którego poziom umiejętności został oszacowany na, powiedzmy, 550 pkt, zakłada się, że z prawdopodobieństwem 50% rozwiąże zadanie o poziomie trudności 550. Jeśli wskażemy, jakich konkretnych umiejętności wymaga rozwiązanie zadań na określonym poziomie trudności, będziemy również w stanie określić, jakich umiejętności oczekujemy od uczniów o poziomie kompetencji wystarczającym do rozwiązania tego zadania z odpowiednio dużym prawdopodobieństwem.

Na tej zasadzie wyznaczone zostały **poziomy umiejętności** (*proficiency levels*): zakresy wyników punktowych z przypisanymi charakterystycznymi dla nich umiejętnościami. W badaniu PISA 2000 zdefiniowanych zostało pięć<sup>12</sup> poziomów umiejętności w dziedzinie rozumienia czytanego tekstu; w kolejnych edycjach badania definiowane były poziomy umiejętności dla pozostałych dziedzin; jednocześnie system poziomów był rozbudowywany poprzez definiowanie dodatkowych poziomów powyżej i poniżej poziomów zdefiniowanych pierwotnie (z zachowaniem zdefiniowanych w poprzednich cyklach wartości granicznych poszczególnych poziomów). Proces definiowania poziomów umiejętności dla dziedziny czytania w badaniu PISA 2000 przebiegał następująco:

<sup>12</sup> Lub sześć, jeśli uwzględnimy także „poziom 0”, a więc uczniów, którzy nie osiągnęli nawet poziomu pierwszego.

- Określanie wartości granicznych rozdzielających kolejne pięć poziomów (oraz minimalnego wyniku dla najniższego poziomu 1.); z wyjątkiem najwyższego, piątego poziomu (nieograniczonego z góry), wszystkie poziomy wyznaczone były jako zakres wyników o szerokości ok. 72 pkt. Z punktu widzenia użytkownika danych PISA można by przyjąć, że granice te zostały określone w sposób arbitralny, choć, w istocie, stały za tym dość precyzyjne reguły matematyczne ze stosunkowo niewielką liczbą arbitralnie przyjętych parametrów<sup>13</sup>. Przykładowo, 2. poziom umiejętności został zdefiniowany jako wyniki w zakresie od 408 do 480 pkt.
- Do poszczególnych poziomów umiejętności zostały przypisane zakresy trudności zadań charakteryzujących dany poziom. Granice zakresu trudności zadań odpowiadających danemu poziomowi dobrane są w taki sposób, aby uczniowie o wyniku na jego dolnej granicy byli przeciętnie w stanie rozwiązać połowę z zadań przypisanych do tego poziomu umiejętności. Można zatem przyjąć, że uczniowie z wynikami w zakresie danego poziomu umiejętności z dużym prawdopodobieństwem (nawet dla najsłabszych z nich przekraczającym 50%) rozwiązują zadania przypisane do danego poziomu<sup>14</sup>.
- Zespół ekspertów PISA dokonał analizy umiejętności niezbędnych do rozwiązania poszczególnych zadań przypisanych do kolejnych poziomów umiejętności i na tej podstawie określił zakres umiejętności, których opanowanie odpowiada osiągnięciu danego poziomu.

Analogiczna procedura była stosowana w kolejnych edycjach do definiowania poziomów umiejętności dla pozostałych dziedzin, a także dla definiowania dodatkowych poziomów (powyżej piątego lub poniżej pierwszego). Zakresy punktowe poziomów umiejętności (a także ich szerokość) dla poszczególnych dziedzin są różne i różne są – oczywiście – charakterystyczne dla danego poziomu umiejętności. Z tego względu szczegółowe charakterystyki poziomów umiejętności podane są w opracowaniach poświęconych założeniom teoretycznym badania dla poszczególnych dziedzin.

<sup>13</sup> Są one opisane w *PISA 2015 Technical Report* (OECD, 2017), s. 279–281, nie mają jednak bezpośredniego znaczenia dla interpretacji poziomów umiejętności.

<sup>14</sup> Najlepsi uczniowie z danego poziomu (a więc o wyniku bliskim granicy tego i następnego poziomu) rozwiązywali ok. 70% zadań odpowiadających temu poziomowi (i blisko 50% zadań z następnego poziomu). Zauważmy przy tym, że – zgodnie z potocznym rozumieniem „opanowania umiejętności” – jeśli z zadaniem o poziomie trudności 408 powiążemy jakieś konkretne, wystarczające dla rozwiązania tego zadania umiejętności, to o uczniu z wynikiem 408 raczej nie powiemy, że umiejętności te opanował (50% szans na rozwiązanie danego zadania to zbyt mało). W konsekwencji zakres trudności zadań służących do scharakteryzowania danego poziomu umiejętności ma granice punktowe ułożone niżej niż granice zakresu poziomu umiejętności określonego dla uczniów. (Zakresy punktowe poziomów umiejętności podawane w raportach z badania zawsze dotyczą uczniów, a nie zadań).

Odwoływanie się w relacjonowaniu wyników badania PISA do poziomów umiejętności pozwala na powiązanie wyników punktowych z konkretnymi kompetencjami, które posiadają uczniowie – ze zdolnością do rozwiązywania określonych typów problemów. Wykorzystując poziomy umiejętności, należy jednak pamiętać, że:

- Liczba poziomów umiejętności i, do pewnego stopnia, zakresy punktowe wyników decydujące o przypisaniu ucznia do danego poziomu są zasadniczo wynikiem arbitralnej decyzji ekspertów PISA.
- Przypisanie zadań do zdefiniowanych uprzednio poziomów umiejętności zależy od trudności zadań określonej obiektywnie (powiązanej – za pośrednictwem modeli IRT – z odsetkiem badanych, którzy umieli rozwiązać dane zadanie).
- Opis umiejętności charakteryzujących dany poziom jest wynikiem analizy treści zadań dokonanej przez ekspertów dziedzinowych PISA.

Przyjmuje się, że umiejętności odpowiadające poziomowi 2. określają minimum tego, co w danej dziedzinie jest niezbędne do radzenia sobie przez ucznia w dalszej edukacji oraz dorosłym życiu. Nie oznacza to jednak, że określony wynik punktowy jednoznacznie dzieli populację na tych, którzy sobie poradzą, i tych, którzy sobie nie poradzą. Granice przedziałów punktowych są do pewnego stopnia umowne i przybliżone, a opisu kompetencji powiązanych z danym wynikiem punktowym nie należy traktować jako bezwzględnie trafnego opisu konkretnych umiejętności posiadanych lub nie przez każdego ucznia, który dany wynik by uzyskał. Poziomy umiejętności są bardzo użytecznym narzędziem pozwalającym zinterpretować wyniki PISA w kategoriach praktycznych, opisywalnych w naturalnym języku umiejętności i kompetencji. Stosując je, należy jednak brać pod uwagę, że o ich zakresie decydowały zarówno kryteria obiektywne, jak i arbitralne decyzje ekspertów.

## Wykorzystanie testu adaptatywnego w badaniu OECD PISA

Realizacja badania PISA od edycji 2015 wyłącznie za pomocą narzędzi komputerowych umożliwiła w badaniu PISA 2018 wprowadzenie kolejnej „nowości metodologicznej” – **testowania adaptatywnego**. Rozwiązanie to wprowadzono wtedy jedynie dla dziedziny głównej, czyli rozumienia czytanego tekstu. W badaniu 2022 testowanie adaptatywne objęło także matematykę, zaś w PISA 2025 – rozumowanie w naukach przyrodniczych.

Uczniowie rozwiązujący zadania PISA w edycjach 2000–2012 korzystali z tradycyjnych, papierowych zeszytów testowych. Od edycji 2006 stopniowo wprowadzane były – jako opcjonalne komponenty badania – elementy testów komputerowych; w Polskiej edycji badania pojawiły się one pierwszy raz w badaniu PISA 2009. Przełom nastąpił w edycji PISA 2015, kiedy w większości krajów (w tym w Polsce) badanie głównych dziedzin przeprowadzone zostało wyłącznie za pomocą narzędzi komputerowych. Choć zmiana narzędzia poprzedzona była szczegółowymi analizami oraz bardzo rozbudowanym badaniem próbnym, stanowiła ona istotną cezurę. Jakkolwiek, ogólnie rzecz biorąc, badanie OECD PISA zachowało swoją ciągłość i porównywanie wyników „papierowych” edycji 2000–2012 oraz „komputerowych” 2015–2025 jest dopuszczalne, to należy pamiętać, że zmiana ta wprowadziła dodatkowe źródło niepewności i takie porównania powinny być dokonywane ze szczególną ostrożnością. Problemy z tym związane – a także prawdopodobne konsekwencje dla wyników uzyskanych przez polskich uczniów – szczegółowo omówione zostały w raporcie z badania PISA 2015 w Polsce<sup>15</sup>.

Zastąpienie papierowych zeszytów testowych komputerami daje jednak bardzo wymierne korzyści. Nie polegają one – podkreślimy – na powierzeniu komputerom oceny rozwiązań: jak dotąd ocena rozwiązań zadań otwartych dokonywana jest przez ludzi, specjalistów w konkretnych dziedzinach<sup>16</sup>. Stosowanie testów komputerowych – poza niewątpliwym uproszczeniem logistyki realizacji badania – pozwala na znaczne zwiększenie różnorodności zestawów zadań rozwiązywanych przez poszczególnych uczniów, a w konsekwencji – na zwiększenie liczby zadań, a także wykorzystanie większej liczby wariantów ich rotacji, co pozwala na zmniejszenie błędów związanych z czynnikami specyficznymi dla poszczególnych zadań.

Zastosowanie testu komputerowego pozwala także na wykorzystanie w badaniu PISA testowania adaptatywnego, a więc dostosowania poziomu trudności pytań do szacowanego poziomu umiejętności ucznia.

To, że dostosowanie poziomu trudności zadania do poziomu umiejętności może być celowe, podpowiada intuicja i doświadczenie. Informacja, że uczeń o niskim poziomie umiejętności nie potrafił rozwiązać kolejnego zadania bardzo trudnego bądź też że uczeń bardzo zdolny poradził

<sup>15</sup> Pośrednie analizy wskazywały, że zmiana narzędzia, w przypadku polskich uczniów, mogła prowadzić do obniżenia przeciętnych wyników od kilku do kilkunastu punktów (zob. Federowicz i Sitek, 2017, s. 25–27).

<sup>16</sup> Od kilku lat OECD prowadzi badania dotyczące możliwości wykorzystania algorytmów opartych na sztucznej inteligencji w procesie kodowania odpowiedzi otwartych, jednak rozwiązania takie nie są jeszcze wdrażane do badania. Obecnie automatycznie kodowane są jedynie braki odpowiedzi oraz proste odpowiedzi w pytaniach polegających np. na wpisaniu przez ucznia konkretnej liczby. Jeśli dana odpowiedź pojawiała się kilkakrotnie w poprzednich edycjach badania i była oceniana zawsze w ten sam sposób przez osoby kodujące, program automatycznie nadaje jej odpowiedni kod.

sobie z kolejnym łatwym zadaniem, w bardzo niewielkim stopniu uzupełnia naszą wiedzę o ich poziomie umiejętności. Jeśli chcemy precyzyjniej określić poziom umiejętności ucznia, o którym (wstępnie) wiemy, że umie dużo, powinniśmy dać mu do rozwiązania zadanie trudne, zaś uczniowi słabemu dać zadanie łatwe. Intuicja ta ma pełne potwierdzenie we własnościach statystycznych modeli IRT: dodanie kolejnego zadania będzie miało największy wpływ na zmniejszenie błędu losowego pomiaru, jeśli będzie to zadanie o poziomie trudności odpowiadającym poziomowi umiejętności badanego (a więc takiego, dla którego badany ma 50% szans na podanie poprawnego rozwiązania). Przypomnijmy jeszcze raz, że dzięki zastosowaniu do skalowania umiejętności uczniów modelu IRT, fakt, że różni uczniowie rozwiązują różne zestawy zadań – także zestawy różniące się poziomem trudności – nie stanowi przeszkody do określenia ich umiejętności na tej samej skali.

Szczegółowe rozwiązania procedury testowania adaptatywnego można prześledzić na przykładzie adaptatywnego testu rozumienia czytanego tekstu z PISA 2018. W kolejnych edycjach i w odniesieniu do innych dziedzin stosowane procedury mogły być poddawane niewielkim modyfikacjom<sup>17</sup>.

W PISA 2018 procedura ta przebiegała następująco: Każdemu badanemu przypisano początkowo pięć podzestawów zadań z danej dziedziny: podzestaw startowy (core, 7–10 zadań), dwa warianty (łatwy i trudny) podzestawu etapu I (stage I, 12–15 zadań) oraz dwa warianty (łatwy i trudny) podzestawu etapu II (stage II, 12–15 zadań). Z tych podzestawów uczeń ostatecznie otrzymał do rozwiązania trzy – startowy, jeden z podzestawów etapu I oraz jeden z etapu II.

Po wykonaniu zadań z podzestawu startowego program sterujący przebiegiem badania<sup>18</sup> na podstawie rozwiązań zadań zamkniętych wstępnie przypisywał badanemu kategorię poziomu umiejętności (niską, średnią lub wysoką)<sup>19</sup>. Uczniowie z grupy „niskiej” z prawdopodobieństwem 90% jako kolejny podzestaw otrzymywali podzestaw łatwy, zaś z prawdopodobieństwem 10% – trudny; dla uczniów z grupy „wysokiej” prawdopodobieństwo to wynosiło odpowiednio 10%

<sup>17</sup> Procedura testowania adaptatywnego dla dziedziny rozumowania w naukach przyrodniczych jest praktycznie identyczna jak dla dziedziny rozumienia czytanego tekstu; w przypadku testu umiejętności matematycznych stosowane były trzy – a nie dwa – poziomy trudności podzestawów (trudny, średni i łatwy). Pewne różnice (wynikające z liczby wykorzystywanych zadań) w procedurach wiązały się także z tym, czy dana dziedzina była w danej edycji dziedziną wiodącą, czy nie.

<sup>18</sup> W edycjach 2018 i 2022 program ten działał na komputerze, przy którym pracował uczeń; w edycji 2025, gdy testowanie odbywało się online – na serwerze konsorcjum międzynarodowego. Jest to jednak różnica czysto techniczna, bez znaczenia dla samej zasady prowadzenia badania.

<sup>19</sup> Przykładowo, w edycji 2022 zadania zamknięte stanowiły przeciętnie 91% zadań z etapu startowego oraz 64% zadań z etapu I. Zadania otwarte oczywiście nie mogły być uwzględniane przy przypisywaniu uczniów do grup poziomów umiejętności tworzonych na użytek testu adaptatywnego (są one oceniane na dalszym etapie opracowywania wyników badania). Należy przy tym podkreślić, że wstępna ocena umiejętności ucznia dokonywana automatycznie na użytek testowania adaptatywnego nie ma wpływu na ostateczny wynik PISA dla danego ucznia, choć, oczywiście, ponieważ oba te oszacowania opierają się częściowo na tych samych danych, ich wartości są ze sobą silnie skorelowane.

i 90%, zaś dla uczniów z grupy „średniej” – po 50%. Po wykonaniu zadań z podzestawu etapu I badani ponownie przypisywani byli do kategorii poziomów umiejętności (na podstawie liczby poprawnych odpowiedzi na zadania zamknięte z, łącznie, podzestawów startowego i etapu I); warianty podzestawów etapu II były przypisywane im w sposób analogiczny jak podzestawów etapu I. W konsekwencji choć każdy uczeń mógł trafić na zadanie o dowolnym poziomie trudności, to zastosowany schemat doboru zadań zwiększał prawdopodobieństwo, że uczeń rozwiązywał głównie zadania dostosowane do jego poziomu umiejętności.

Korzyścią z testowania adaptatywnego jest zmniejszenie błędu losowego oszacowania poziomu umiejętności ucznia w porównaniu z nieadaptatywnym testem o tej samej liczbie zadań – innymi słowy, pozwala on zwiększyć precyzję pomiaru bez zwiększania liczby zadań. Jakże zaś są koszty tego rozwiązania? Z jednej strony, bardziej skomplikowany schemat testowania oznacza większą wrażliwość na błędy nielosowe, związane z niedoskonałym spełnianiem przez własności psychometryczne zadań założeń modeli IRT (a należy pamiętać, że realnie stosowane zadania nigdy nie będą w sposób idealny spełniały założeń modelu). Z drugiej strony, zastosowanie testu adaptatywnego stwarza pewną trudność praktyczną dla osób analizujących wyniki badania: ponieważ grupy uczniów rozwiązujących konkretne zadania z założenia różnią się przeciętnym poziomem umiejętności, traci sens bezpośrednio analizowanie rozkładów odpowiedzi (czy też odsetków poprawnych rozwiązań) poszczególnych zadań. W konsekwencji interpretacja wyniku badania oraz czytelne jego przedstawienie innym odbiorcom stają się znacznie trudniejsze. Problem ten można częściowo rozwiązać poprzez wyliczanie – na bazie stosowanego modelu IRT – hipotetycznego rozkładu odpowiedzi dla całej badanej próby (jaki byłby odsetek poprawnych rozwiązań, gdyby zadanie dostali do rozwiązania wszyscy badani), taki rozkład jednak byłby obciążony dodatkowymi błędami, chociażby związanymi z nieidealną zgodnością własności zadania z założeniami modelu IRT.

Jak wynika z obliczeń Kentaro Yamamoto, Hyo Jeong Shin i Lale Khorramdela (2019), zastosowanie testu adaptatywnego w PISA 2018 pozwoliło zmniejszyć błąd losowy pomiaru przeciętnie o około 4,5%; dla uczniów o wyniku bardzo wysokim (rzędu 700 pkt) lub bardzo niskim (rzędu 300 pkt) zysk ten był większy i mógł sięgać 10%. Trzeba wyraźnie stwierdzić, że wobec potencjalnych komplikacji związanych z interpretacją wyników badania korzyść ta jest niewielka (zwłaszcza że błąd pomiaru jest odpowiedzialny tylko za część całkowitego błędu losowego badania). Z drugiej strony należy pamiętać, że zastosowany schemat alokacji zadań był dość konserwatywny – przy bardziej agresywnym podejściu zapewne i korzyść mogłaby być znaczniejsza. Zastosowanie testu adaptatywnego w badaniu PISA należy traktować zatem raczej jako ostrożny eksperyment –

i w poszerzeniu doświadczenia metodologicznego, a nie w samej poprawie dokładności wyników badania należy się doszukiwać jego głównej wartości.

## Pomiar danych kontekstowych<sup>20</sup>

W badaniu PISA, oprócz pomiaru umiejętności piętnastolatków w trzech podstawowych dziedzinach, we wszystkich krajach realizowany jest również pomiar danych kontekstowych, które mogą pomóc w zrozumieniu głównych wyników. W tym celu wykorzystuje się kwestionariusze wypełniane przez uczniów oraz dyrektorów szkół, a w niektórych krajach również przez nauczycieli i rodziców. Choć zwykle dane kwestionariuszowe wykorzystywane są właśnie jako kontekst do pomiaru umiejętności, a więc jako źródło zmiennych wyjaśniających zróżnicowanie poziomu umiejętności, to dane o postawach i opiniach młodzieży, a także o odczuciach związanych ze szkołą mogą być interesujące same w sobie. Dane zbierane w kwestionariuszach PISA nie dotyczą wyłącznie szkoły lub systemu edukacji. Na ich podstawie wnioskować można również o przekonaniach młodzieży, jej podejściu do spraw prywatnych i społecznych, co ma przełożenie na główny cel badania PISA, czyli ocenę przygotowania młodzieży do dorosłego życia.

## Założenia pomiaru za pomocą kwestionariuszy

Założenia pomiaru za pomocą kwestionariuszy (*questionnaire framework*) stanowią podstawę dla wszystkich działań związanych z konstruowaniem treści pytań i tworzeniem wskaźników z tej części badania. Od 2012 roku przyjęto zasadę, że reguły tworzenia kwestionariuszy dla kolejnych edycji wiążą się ściśle z potrzebą monitorowania zmian zachodzących na poziomie systemów edukacji, szkół oraz uczniów, które mogą mieć przełożenie na wiedzę i umiejętności uczniów (OECD, 2013). Ze względu na możliwości dokonywania porównań między edycjami, większość pytań kwestionariuszowych zadaje się w tej samej formie w kolejnych edycjach, z dodatkiem pytań specyficznych dla danej edycji badania potrzebnych dla głębszego podejścia do zrozumienia wyników głównej dziedziny danej edycji badania (w ostatnich edycjach: 2018 – czytania, 2022 – matematyki, a w 2025 – rozumowania w naukach przyrodniczych). Szczegółowe informacje o założeniach pomiaru za pomocą kwestionariuszy dostępne są w raportach technicznych z badania PISA (OECD, 2026), a sam proces konstruowania kwestionariuszy i szczegółowy proces skonstruowanych wskaźników zawarty jest w raporcie technicznym z badania PISA 2022 (OECD, 2024).

<sup>20</sup> Ponieważ generalne zasady dotyczące pomiaru danych kontekstowych są wspólne dla kolejnych edycji badania PISA, niniejszy rozdział stanowi częściowo powtórzenie rozdziału (jednego ze współautorów tego rozdziału) o danych kwestionariuszowych z raportu PISA 2018 (Bulkowski i Sitek, 2020).

Założenia dotyczące zagadnień poruszanych w kwestionariuszach są podzielone na cztery kategorie, związane z różnymi typami mierzonych konstruktów i możliwością wnioskowania o poszczególnych czynnikach mających związek z głównym przedmiotem badania, czyli pomiarem umiejętności:

- konstrukty dotyczące wskaźników niekognitywnych i metakognitywnych,
- konstrukty dotyczące pochodzenia ucznia,
- konstrukty dotyczące nauczania i uczenia się,
- konstrukty dotyczące polityk szkół i zarządzania nimi.

Poniżej przedstawiono krótkie opisy każdej z czterech kategorii.

### *Konstrukty dotyczące wskaźników niekognitywnych i metakognitywnych*

Do tej kategorii należą umieszczone w kwestionariuszu ucznia zagadnienia związane z postawami, przekonaniem, motywacjami i aspiracjami uczniów oraz ich zachowania związane z uczeniem się, takie jak czas poświęcany na naukę. Często mierzone są one w odniesieniu do głównej dziedziny badania. Wyniki te same w sobie stanowią dużą wartość, lecz w kontekście równoległego mierzenia umiejętności mogą być cenne w wyjaśnianiu różnicowania w osiągnięciach.

### *Konstrukty dotyczące pochodzenia ucznia*

W ramach tej kategorii zbierane są dane dotyczące statusu społeczno-ekonomicznego ucznia, w tym statusu zawodowego i wykształcenia rodziców, stanu posiadania członków gospodarstwa domowego, w tym dóbr kultury, zasobności gospodarstwa domowego, zasobów edukacyjnych w domu i liczby książek w domu. Pozwala to nie tylko na scharakteryzowanie poszczególnych uczniów, lecz także na przypisanie szkołom uśrednionych charakterystyk uczniów, które stanowią kontekst dla podejmowanych działań edukacyjnych. W ramach tej kategorii gromadzone są również informacje o dotychczasowej i planowanej karierze edukacyjnej oraz zamierzeniach dotyczących przyszłych wyborów zawodowych.

### *Konstrukty dotyczące nauczania i uczenia się*

W ramach tej kategorii zbierane są dane związane z efektywnością pracy nauczycieli, uczeniem się uczniów i organizacją nauczania. Poszczególne części kwestionariuszy szkół i kwestionariuszy uczniów koncentrują się na kwalifikacjach nauczycieli, praktykach nauczania, klimacie w szkole,

czasie nauki oraz możliwościach uczenia się w szkole i poza nią. Pytania dotyczące nauczania i uczenia się w znaczniej większości dotyczą głównej dziedziny.

### **Konstrukty dotyczące polityk szkół i zarządzania nimi**

W ramach tej kategorii zbierane są dane dotyczące różnych czynników związanych z efektywnością szkoły. Są to m.in. rozwój zawodowy nauczycieli, zarządzanie szkołą, zaangażowanie rodziców, klimat szkolny (np. wysokie oczekiwania w zakresie osiągnięć uczniów) oraz wykorzystanie ewaluacji do poprawiania warunków i efektów kształcenia. W tej kategorii mieszczą się również wskaźniki dotyczące zasobów i sposobów wsparcia szkoły związanych z główną dziedziną badania: rozumieniem czytanego tekstu. Należą do nich np. biblioteki szkolne, urządzenia cyfrowe oraz szkolne zasady i praktyki dotyczące ich używania i rozwijania umiejętności czytania.

### **Kwestionariusze używane w badaniu**

Przedstawione powyżej założenia związane ze zbieraniem danych w czterech kategoriach realizowane są poprzez udostępnianie kwestionariuszy do samodzielnego wypełniania różnym grupom osób zaangażowanych w edukację szkolną uczniów. Dyrektorzy szkół i uczniowie są głównymi respondentami w badaniu, a kwestionariusze przeznaczone dla nich są obowiązkowe do zrealizowania w każdym kraju biorącym udział w badaniu. Oprócz tego w każdej edycji badania PISA każdy z krajów może zdecydować o przeprowadzeniu dodatkowego pomiaru za pomocą kilku innych, opcjonalnych kwestionariuszy, jak kwestionariusz rodziców, nauczycieli (adresowany do nauczycieli przedmiotów powiązanych z główną dziedziną danej edycji badania), dodatkowe kwestionariusze skierowane do uczniów (jak kwestionariusz dotyczący finansów osobistych, kariery edukacyjnej, technologii informacyjno-komunikacyjnych).

Wersje międzynarodowe kwestionariuszy są opracowywane przez konsorcjum organizujące badanie. Każde z pytań podlega procedurze adaptacji, uwzględniającej kontekst kulturowy i związany z kształtem systemu edukacji. Adaptacje mogą być wprowadzane jednak jedynie według ściśle ustalonych reguł, by zachować porównywalność wyników pomiędzy krajami. Wszystkie tłumaczenia podlegają wieloetapowej procedurze kontroli i weryfikacji. Ponadto, każdy kraj ma prawo do dodania kilku swoich pytań (umieszczanych w kwestionariuszu po pytaniach wspólnych dla wszystkich krajów). Polska regularnie korzysta z takiej możliwości, a pytania dodatkowe opracowują eksperci z Instytutu Badań Edukacyjnych – Państwowego Instytutu Badawczego.

### **Kwestionariusz ogólny ucznia**

Podstawowa część kwestionariusza ucznia, realizowana we wszystkich krajach, przeprowadzana jest bezpośrednio po zakończeniu dwugodzinnej sesji testowej. Na odpowiedzi uczniów na tę część kwestionariusza przeznaczono około 35 minut. W Polsce kwestionariusze, podobnie jak zadania w części kognitywnej, przedstawiano uczniom na komputerze. Zagadnienia poruszane w pytaniach w tym kwestionariuszu dotyczyły m.in.: postaw i poglądów uczniów dotyczących ich życia, rodziny oraz doświadczeń związanych ze szkołą, uczenia się w szkole, planu lekcji oraz czasu, jaki poświęcają na naukę, oraz kwestii związanych z główną dziedziną danej edycji badania (jak poglądy ucznia związane z daną dziedziną nauki oraz jej miejsce w nauczaniu w szkole).

### **Kwestionariusz szkoły**

Obowiązkowy dla wszystkich krajów kwestionariusz szkoły wypełniany jest przez dyrektorów szkół lub osoby wyznaczone przez dyrektorów. Zagadnienia poruszane w pytaniach w tym kwestionariuszu dotyczyły m.in.: ogólnych informacji o szkole, kierowania szkołą, nauczycieli pracujących w szkole, oceniania i ewaluacji, działań podejmowanych wobec niektórych grup uczniów, atmosfery (klimatu) szkoły. Kwestionariusz jest wypełniany online.

### **Wskaźniki, indeksy, skale wyliczane na podstawie danych kontekstowych**

Charakterystyką badania PISA, jak również dobrym standardem, który powinien być stosowany we wszystkich tego typu badaniach, jest udostępnianie wszystkich danych (poza danymi mogącymi pozwolić na identyfikację uczestników) w postaci zbiorów danych jednostkowych. Dzięki temu można analizować odpowiedzi respondentów na każde pojedyncze pytanie obecne w kwestionariuszach. Wiele z pytań zawartych w kwestionariuszach zostało zaprojektowanych jednak w formie swego rodzaju zestawów, z myślą o analitycznym zagregowaniu ich w celu oddania poziomu wartości jakiejś nieobserwowalnej bezpośrednio cechy, tzw. cechy ukrytej (np. poczucie przynależności do szkoły, przekonanie o samodzielności w uczeniu się matematyki, status społeczno-ekonomiczny). Stoi za tym założenie, że jeśli zadajemy respondentom pytania dotyczące podobnych kwestii, ich odpowiedzi są wskaźnikami (przejawami) badanych zjawisk lub cech. Z tego względu na podstawie poszczególnych pytań wyliczane są dodatkowe, zdefiniowane przed badaniem, zmienne pochodne (wskaźniki, indeksy oraz skale), które reprezentują poszczególne mierzone cechy. W badaniu PISA przyjęto trzy rodzaje zmiennych pochodnych:

- proste wskaźniki wynikające z przekształceń arytmetycznych lub rekodowania jednego lub kilku pytań,
- skale opracowane przy użyciu modeli IRT (teorii odpowiedzi na pytanie testowe),
- indeksy złożone – z oddzielną autorską metodą wyliczania – oparte na kombinacji dwóch lub większej liczby wskaźników.

Skale wyliczane w oparciu o metody IRT zostały wystandaryzowane – wszystkie skale są przygotowane w ten sposób, że dla pierwszej edycji badania, w której wykorzystano daną skalę średnia dla krajów OECD wynosi 0, a odchylenie standardowe 1. Jakkolwiek nie w każdym przypadku możliwe jest w pełni wiarygodne porównywanie wyników dla poszczególnych miar w różnych edycjach badania, zakotwiczenie skal w wynikach konkretnej edycji pozwala na przynajmniej częściową ocenę trendów i zmian w czasie.

## Materiały i opracowania OECD

Zebrane w trakcie badania dane są dostępne do pobrania na [stronie Organizacji Współpracy Gospodarczej i Rozwoju \(OECD\)](#) wraz z pełną dokumentacją procedur stosowanych na każdym etapie procesu badawczego. Na międzynarodowej stronie badania udostępniono również narzędzia badawcze – treść pytań zawartych w kwestionariuszach, a także przykłady zadań rozwiązywanych przez uczniów. Kilka takich zadań jest udostępnianych po każdej edycji badania. OECD publikuje również raport międzynarodowy – poszczególne tomy raportu prezentowane będą sukcesywnie w kolejnych miesiącach po publikacji wyników głównych<sup>21</sup>.

---

<sup>21</sup> Więcej szczegółów dotyczących organizacji i metodologii PISA 2025 (wraz z dodatkowymi przypisami dotyczącymi specyficznych sytuacji w niektórych krajach) znajduje się w raporcie technicznym z badania PISA 2025 (w całości dostępnym w 2027 roku) oraz w aneksach do raportu międzynarodowego. Niektóre rozdziały raportu międzynarodowego są dostępne na stronie OECD od września 2026.

## Bibliografia

- Bartnik, E., Czetwertyńska, G., Czetwertyński, P., Cyngot, D., Federowicz, M., Grabowska, A., Haman, J., Komorek, J., Marciniak, Z., Ostrowska, B., Sitek, M., Sułowska, A., (2007) *Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2006 w Polsce*. Ministerstwo Edukacji Narodowej
- Biedrzycki, K., Cyngot, D., Federowicz, M., Grzęda, M., Haman, J., Karpiński, Z., Lipiec, M., Marciniak, Z., Ostrowska, E. B., Sawiński, Z., Sitek, M., Spalik, K., Sułowska, A., Swat-Pawlicka, M., Walicki, P., Wierzchowski, T. (2010). *Program Międzynarodowej Oceny Umiejętności Uczniów OECD PISA. Wyniki badania 2009 w Polsce*. Ministerstwo Edukacji Narodowej.
- Bulkowski, K., Sitek, M. (2020). Komponent kwestionariuszowy badania PISA: dobrostan uczniów i klimat szkoły. W: M. Sitek, E. B. Ostrowska (red.), *PISA 2018 Czytanie, rozumienie, rozumowanie* (s. 236–255). Instytut Badań Edukacyjnych.
- Federowicz, M. (red.). (2013). *OECD PISA 2012. Wyniki badania 2012 w Polsce*. Ministerstwo Edukacji Narodowej.
- Federowicz, M., Sitek, M. (red.). (2017). *Program Międzynarodowej Oceny Umiejętności Uczniów. Wyniki badania PISA 2015 w Polsce*. Instytut Badań Edukacyjnych.
- Haman, J. (2020). Metodologia badania PISA 2018. W: M. Sitek, E. B. Ostrowska (red.), *PISA 2018 Czytanie, rozumienie, rozumowanie* (s. 19–38). Instytut Badań Edukacyjnych.
- Haman, J., Bulkowski, K. (2024). Metodologia badania PISA 2022. W: J. Kaźmierczak, K. Bulkowski (red.), *Polscy piętnastolatki w perspektywie międzynarodowej. Wyniki badania PISA 2022* (s. 19–40). Instytut Badań Edukacyjnych.
- Haman, J., Ostrowska, E. B. (2014). OECD PISA methodology and the interpretation of Polish students' results in the area of science. *Edukacja Biologiczna i Środowiskowa*, 1(13), 29–41.
- Jakubowski, M., Pokropek, A. (2009). *Badając egzaminy: podejście ilościowe w badaniach edukacyjnych*. Centralna Komisja Egzaminacyjna.
- OECD (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD Publishing. <https://dx.doi.org/10.1787/9789264190511-en>

OECD (2017). *PISA 2015 Technical Report*. OECD Publishing.

<https://www.oecd.org/pisa/data/2015-technical-report/>

OECD (2023a). *PISA 2022 Assessment and Analytical Framework*. OECD Publishing.

<https://doi.org/10.1787/dfe0bf9c-en>

OECD (2024). *PISA 2022 Technical Report*. OECD Publishing. <https://doi.org/10.1787/01820d6d-en>

OECD (2026). *PISA 2025 Technical Report* [w przygotowaniu]. OECD Publishing.

Pokropek, A. (red.). (2015). *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii: teoria i zastosowania*. Instytut Badań Edukacyjnych.

Sitek, M., Ostrowska, E. B. (red.). (2020). *PISA 2018 Czytanie, rozumienie, rozumowanie*. Instytut Badań Edukacyjnych.

Yamamoto, K., Shin, H. J., Khorramdel, L. (2019). *Introduction of Multistage Adaptive Testing Design in PISA 2018*. OECD Working Paper No. 209.