

Wykorzystanie testu adaptatywnego w badaniu OECD PISA 2018¹

Realizacja badania PISA od edycji 2015 wyłącznie za pomocą narzędzi komputerowych umożliwiła w badaniu PISA 2018 wprowadzenie kolejnej nowości metodologicznej – testowania adaptatywnego.

W edycji PISA 2000–2012 uczniowie rozwiązywali zadania, korzystając z tradycyjnych, papierowych zeszytów testowych. Od edycji 2006 stopniowo wprowadzane były – jako opcjonalne komponenty badania – elementy testów komputerowych; w Polskiej edycji 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 istotną cezurę. Jakkolwiek, ogólnie rzecz biorąc, badanie OECD PISA zachowało swoją ciągłość i porównywanie wyników „papierowych” edycji 2000–2012 oraz „komputerowych” 2015 i 2018 jest dopuszczalne, to należy pamiętać, że zmiana ta wprowadziła dodatkowe źródło niepewności i porównania takie powinny być dokonywane ze szczególną ostrożnością. Problemy z tym związane – a także prawdopodobne konsekwencje dla wyników uzyskanych przez polskich uczniów – szczegółowo omówiliśmy w raporcie z badania PISA 2015 w Polsce.

Zastąpienie papierowych zeszytów testowych komputerami daje jednak bardzo wymierne korzyści. Nie polegają one – podkreślmy – na powierzeniu komputerom oceny rozwiązań: w dalszym ciągu ocena rozwiązań zadań otwartych dokonywana jest przez ludzi, specjalistów w konkretnych dziedzinach. Stosowanie testów komputerowych – poza niewątpliwym uproszczeniem logistyki realizacji badania – umożliwia znaczne zwiększenie różnorodności zestawów zadań rozwiązywanych przez poszczególnych uczniów, a w konsekwencji – na zwiększenie liczby zadań i wykorzystanie większej liczby wariantów ich rotacji. Pozwala to na zmniejszenie błędów związanych z czynnikami specyficznymi dla poszczególnych zadań lub dla pozycji zajmowanej w teście.

Zastosowanie testu komputerowego pozwala także na wykorzystanie w badaniu PISA testowania adaptatywnego, a więc dostosowania poziomu trudności pytań do szacowanego poziomu umiejętności ucznia. Z możliwości tej po raz pierwszy skorzystano w bieżącej edycji badania; podobna procedura była już jednak stosowana w badaniu kompetencji osób dorosłych PIAAC organizowanego przy OECD.

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 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 naj-

¹ Opracował Jacek Haman

większy 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.

W badaniu OECD PISA 2018 testowanie adaptatywne zostało zastosowane jedynie dla wiodącej dziedziny – rozumienia czytanego tekstu. W tym celu każdemu badanemu przypisane było początkowo pięć podzestawów zadań: 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 otrzymywał do rozwiązania trzy – startowy, jeden z podzestawów etapu I oraz jeden z etapu II.

Po wykonaniu zadań z podzestawu startowego program sterujący przebiegiem badania, na podstawie rozwiązań zadań zamkniętych wstępnie przypisywał badanemu kategorię poziomu umiejętności („niską”, „średnią” lub „wysoką”). Uczniowie z grupy „niskiej” z prawdopodobieństwem 90% jako kolejny podzestaw otrzymywali podzestaw „łatwy”, zaś z prawdopodobieństwem 10% – „trudny”; dla uczniów z grupy „wysokiej” prawdopodobieństwa te wynosiły odpowiednio 10% 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 podzestawów startowego i etapu I łącznie); 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 rozwiązywać będzie głównie zadania dostosowane do jego poziomu umiejętności.

Korzyścią z testowania adaptatywnego jest zmniejszenie błędu losowego oszacowania poziomu umiejętności ucznia w porównaniu do nieadaptatywnego testu 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.

Każde nowe rozwiązanie metodologiczne wiąże się z pewnym ryzykiem dla badania i wymaga od odbiorców jego wyników ciągłego doskonalenia się w metodach analizy danych i dbałości, by z nowości wynieść oczekiwaną korzyść, nie narażając się na błędy związane ze specyfiką nowych narzędzi. Testowanie adaptatywne wprowadzane jest do badania OECD PISA bardzo ostrożnie – zaczynając od jednej dziedziny i z wykorzystaniem stosunkowo prostych schematów przydzielania badanym zadań o różnej trudności. Takie podejście minimalizuje ryzyko, a jednocześnie pozwala na stopniowe udoskonalanie metodologii badania. OECD PISA od blisko dwudziestu lat wyznaczają standardy metodologiczne w badaniach edukacyjnych i ważne jest, by same korzystały z rozwoju nowych metod badawczych i możliwości, jakie badaczom daje postęp techniczny.