Aleksandra Jasińska-Maciążek
Instytut Badań Edukacyjnych

Nowa-stara formuła sprawdzianu
Porównawcza analiza psychometryczna egzaminów z 2014 i 2015 roku

W kwietniu 2015 roku szóstoklasiści po raz pierwszy przystąpili do sprawdzianu w nowej formule. Zmiany w sposobie przeprowadzania tego egzaminu wprowadzono Rozporządzeniem Ministra Edukacji Narodowej z 25 kwietnia 2013 r. (Dz. U. 2013 r. poz. 520). Sprawdzian w nowej formule został podzielony na dwie części. Pierwsza obejmuje zadania z języka polskiego i z matematyki, druga z języka obcego nowożytnego. Pomiarem umiejętności w zakresie języka obcego szóstoklasiści zostali objęci po raz pierwszy. Wydłużono czas pisania sprawdzianu. Na pierwszą część, obejmującą język polski i matematykę, przeznaczono 80 minut (do 2014 roku było to 60 minut), a na test z języka obcego 45 minut. Zmieniono sposób prezentowania wyników, zastępując sumę punktów wynikiem wyrażonym w procentach, co było m.in. motywowane zapewnieniem możliwości konstrukcji testów o różnej maksymalnej liczbie punktów do zdobycia w kolejnych latach. Wyniki z pierwszej części są także dodatkowo prezentowane osobno dla języka polskiego i osobno dla matematyki. Zmieniono również definicję mierzonych umiejętności, a także rozbudowano katalog form zadań mogących się pojawić na egzaminie. Zadania otwarte sprawdzianu w nowej formule podlegają regułom oceny holistycznej, jednak taki sposób oceniania był już stosowany wcześniej (np. w ocenie wypracowania). Konsekwencje niektórych z tych zmian zostaną poddane analizie i omówione w dalszej części referatu.

Analiza właściwości psychometrycznych sprawdzianu przeprowadzonego w 2015 roku jest niezbędna przed podjęciem prac nad wskaźnikami edukacyjnej wartości dodanej (EWD) dla II etapu edukacyjnego. Pierwsze modele EWD dla szkół podstawowych w Polsce zostały policzone z wykorzystaniem wyników sprawdzianu z 2014 roku oraz wyników Ogólnopolskiego Badania Umiejętności Trzecioklasistów (OBUT), zrealizowanego w 2011 roku (Dolata i in., w przygotowaniu). Zmiana koncepcji testów wykorzystanych do obliczania wskaźników EWD może utrudniać zachowanie ciągłości wypracowanych modeli i wymaga bardziej wnikliwej analizy konsekwencji dla skal pomiarowych. Analizy prezentowane w tekście mają więc na celu rozpoznanie ewentualnych zmian skal umiejętności mierzonych sprawdzianem w kontekście prac nad modelami EWD dla II etapu kształcenia. Wskaźniki te są liczone tylko dla umiejętności ogólnych związanych z językiem polskim i matematyką, dlatego w artykule skupiono się wyłącznie na pierwszej części sprawdzianu w nowej formule, pomijając test z języka obcego nowożytnego. Sprawdzian w nowej formule porównano z egzaminem przeprowadzonym rok wcześniej, jako że dopiero od tamtego roku liczone są modele EWD dla szkół podstawowych.

Na potrzeby referatu wybrano i przeanalizowano kilka najważniejszych z punktu widzenia porównania skal pomiarowych sprawdzianu z 2014 i 2015 roku problemów: porównano założone przez twórców egzaminów struktury mierzonych przez testy umiejętności; dokonano analizy jednowymiarowości skali umiejętności, w świetle zmiany polegającej na wyszczególnieniu części z języka polskiego i z matematyki w nowej formule sprawdzianu; oceniono jakość skal pomiarowych, analizując dopasowanie danych do przyjętego modelu, rzetelność testów oraz parametry rozkładów wyników; przeanalizowano dopasowanie testów do badanych populacji w zakresie precyzji szacowania wyników uczniów o różnym poziomie umiejętności. Wyniki te pozwolą ocenić zasadność kontynuacji wypracowanych modeli EWD dla II etapu edukacyjnego, wykorzystujących wyniki sprawdzianu oraz porównywalność interpretacji wskaźników.

Metoda

W analizach wykorzystano wyniki z dwóch sesji egzaminacyjnych: sprawdzianu z 2014 i 2015 roku. Za teorię pomiaru wspierającą prezentowane analizy przyjęto model Rascha – jeden z modeli funkcjonujących w ramach teorii IRT (*item response theory*) (Rasch, 1980)[1], która dostarcza narzędzia do testowania właściwości psychometrycznych zadań i skal. Model Rascha wybrano dlatego, że jako jedyny z modeli IRT umożliwia wykorzystanie wyników w modelach EWD dla szkół podstawowych. Po pierwsze, ma on teoretyczne podstawy do tego, by skale utworzone z jego wykorzystaniem traktować jako interwałowe, co jest jednym z założeń modeli EWD (Ballou, 2009; Reardon i Raudenbush, 2009). Po drugie, spośród różnych modeli IRT tylko w modelu Rascha suma punktów zdobyta w teście jest statystyką dostateczną dla oszacowania poziomu osiągnięć (wyniku wyskalowanego) (De Ayala, 2009). Ma to decydujące znaczenie w przypadku modeli EWD dla II etapu edukacyjnego, bowiem dane umożliwiające obliczenie wskaźników EWD są przechowywane w szkołach w postaci sumy punktów. Aby szkoły mogły korzystać z tych wskaźników, należało wybrać taki model, który nie wymagał znajomości pełnego rekordu odpowiedzi.

W prezentowanych analizach wykorzystano zatem jednowymiarowy model Rascha. Analizy wykonano w środowisku R z wykorzystaniem pakietu TAM (Kiefer, Robitzsch i Wu, 2015) stosując metodę estymacji brzegowej największej wiarygodności (*marginal maximum likelihood*, MML).

Struktura testu

Pierwsze pytanie dotyczy tego, czy modyfikacja struktury egzaminu nie spowodowała znaczącej zmiany mierzonych umiejętności. Jest to o tyle ważne, że wskaźniki EWD są miarą efektywności nauczania w zakresie mierzonym testem, którego wyniki zostały wykorzystane do ich obliczenia (Dolata i in., 2013, rozdz. 2; Pokropek, 2013).

Sprawdzian do 2014 roku w założeniu mierzył pięć umiejętności ponadprzedmiotowych, kształtowanych w ramach obowiązkowych zajęć edukacyjnych, tj.:

[1] Więcej informacji na temat tej metody zainteresowany czytelnik może znaleźć w polskojęzycznych pracach Macieja Jakubowskiego i Artura Pokropka (2009), Bartosza Kondratka i Artura Pokropka (2013), Aleksandry Jasińskiej i Michała Modzelewskiego (2014) oraz w książce pod redakcją Henryka Szaleńca (2009).

czytanie, pisanie, rozumowanie, korzystanie z informacji i wykorzystywanie wiedzy w praktyce (*Informator z aneksem dla uczniów ze specyficznymi trudnościami w uczeniu się*, 2009). Pierwsza część sprawdzianu w nowej formule ma natomiast mierzyć stopień opanowania wymagań określonych w podstawie programowej kształcenia ogólnego z zakresu języka polskiego i matematyki. Zadania mogą być oparte na tekstach lub informacjach z zakresu historii lub przyrody (Centralna Komisja Egzaminacyjna, 2013). Wydaje się zatem, że skale zostały zdefiniowane w trochę inny sposób. Przyjrzyjmy się jednak arkuszom testowym i kartotekom sprawdzianów w 2014 i 2015 roku.

Tabela 1. Porównanie struktury testów

Sprawdzian 2014				Sprawdzian 2015				
	Umiejętność	L. zad.	L. pkt		Umiejętność	L. zad.	L. pkt	
JĘZYK POLSKI	Czytanie	10 WW	10	JĘZYK POLSKI	Odbiór wypowiedzi i wykorzystanie zawartych w nich informacji	11 WW 1 KO	14	21
					Analiza i interpretacja tekstów kultury			
	Pisanie	1 KO 1 RO	10 (w tym: 8 za RO)		Tworzenie wypowiedzi	1 RO	7	
MATEMTYKA	Wykorzystywanie wiedzy w praktyce	4 WW 2 KO	8	MATEMATYKA	Sprawność rachunkowa	3 WW	3	20
	Korzystanie z informacji	4 WW	4		Wykorzystywanie i tworzenie informacji	3 WW 1 RO	5	
	Rozumowanie	2 WW 1 KO 1 RO	8		Modelowanie matematyczne	2 WW 1 RO	5	
	Razem	10 WW 3 KO 1 RO	20		Rozumowanie i tworzenie strategii	3 WW 1 RO	7	
RAZEM		26 zad.	40	RAZEM		27 zad.	41	

W 2014 roku pierwszych 10 zadań zamkniętych wielokrotnego wyboru (WW) odwoływało się do umiejętności czytania. Były to zadania do dwóch tekstów: popularnonaukowego oraz literackiego (wiersza) i według kartoteki testu mierzyły m.in. umiejętności takie jak: odczytywanie głównej myśli utworu, odczytywanie i porównywanie informacji zawartych w tekście, wnioskowanie na ich podstawie, charakteryzowanie bohatera tekstu, odczytywanie przenośnych znaczeń, rozpoznawanie cech charakterystycznych tekstu, bezpośredniego zwrotu do odbiorcy czy określenie funkcji epitetów. W 2015 roku pierwszych 11 zadań WW oraz jedno zadanie otwarte krótkiej odpowiedzi (KO; zad. 12) z części sprawdzianu z języka polskiego odwoływało się do umiejętności ogólnych określonych jako odbiór wypowiedzi i wykorzystanie zawartych w nich informacji oraz analiza i interpretacja tekstów kultury. Były to zadania do dwóch tekstów: popularnonaukowego i komiksu. Wśród umiejętności szczegółowych znalazły się takie jak: określanie tematu oraz głównej myśli tekstu, wyszukiwanie informacji wyrażonych wprost i ukrytych, wyciąganie wniosków na temat przesłanek zawartych w tekście, charakteryzowanie i ocenianie bohaterów. Jedno

zadanie odwoływało się do umiejętności rozpoznawania funkcji składniowych wyrazów użytych w wypowiedziach (podmiot) oraz rozpoznawania formy przypadków, czyli umiejętności z zakresu świadomości językowej. Porównanie to pokazuje, że wspomniane podzbiory zadań (może z wyjątkiem jednego zadania ze świadomości językowej) w założeniu mierzyły analogiczne konstrukty. W 2014 roku za tę część zadań można było uzyskać 10 punktów, a w 2015 roku 14 punktów (porównanie struktury znajduje się w tabeli 1).

Przejdźmy do zadań odwołujących się do umiejętności pisania (jak to nazwano w 2014 roku) lub umiejętności tworzenia wypowiedzi (termin stosowany w 2015 roku). W 2014 roku umiejętność ta była sprawdzana za pomocą dwóch zadań: jednego krótkiej odpowiedzi (KO) i jednego rozszerzonej odpowiedzi (RO), czyli wypracowania. Za wypracowanie można było uzyskać 8 punktów, a za zadanie KO 2 punkty. W 2015 roku umiejętność tworzenia wypowiedzi była sprawdzana jednym zadaniem RO, jakim było wypracowanie. Można było za nie uzyskać maksymalnie 7 punktów. Ponownie mierzone umiejętności są porównywalne. W 2015 roku zmniejszyła się jedynie „waga" wypracowania, rozumiana jako wkład punktowy do ostatecznego sumarycznego wyniku.

Pozostałe zadania w obu egzaminach mierzyły tak naprawdę umiejętności matematyczne. Zacznijmy od kategorii wykorzystywania informacji. Analiza zapisów kartotek sprawdzianów oraz samych zadań pokazuje, że zadania wchodzące w skład tej grupy łączy to, że uczeń ma przeanalizować informację zawartą w postaci tabelarycznej lub graficznej (rysunek, diagram) w celu udzielenia odpowiedzi na pytanie, często po odpowiednim przetworzeniu danych. W 2014 roku umiejętność tę mierzyły 4 zadania WW. Wszystkie one odwoływały się do tej samej informacji tekstowo-tabelarycznej. W 2015 roku umiejętność ta także była badania przez 4 zadania (3 WW i 1 RO), jednak każde zadanie wymagało przeanalizowania informacji z innego źródła i tym samym były one bardziej zróżnicowane (tabela, rysunki, diagram). Kolejną kategorią, której nazwa powtarza się w koncepcjach obu testów, jest umiejętność rozumowania. W obu latach była ona mierzona 4 zadaniami. W 2015 roku zdefiniowano ją jako umiejętność prowadzenia prostego rozumowania składającego się z niewielkiej liczby kroków, ustalenia kolejności czynności (w tym obliczeń) prowadzących do rozwiązania problemu, wyciągnięcia wniosków z kilku informacji podanych w różnej postaci (Centralna Komisja Egzaminacyjna, 2013). W 2014 roku została ona zdefiniowana przez odwołanie do wielu umiejętności bardziej szczegółowych, jak np.: rozpoznawanie charakterystycznych właściwości i cech liczb, figur i in., posługuje się kategoriami czasu i przestrzeni w celu porządkowania wydarzeń, dostrzeganie prawidłowości, opisanie ich i sprawdzenie na przykładach, opisanie sytuacji przedstawionej w zadaniu za pomocą m.in. wyrażenia arytmetycznego i prostego wyrażenia algebraicznego, a także ustalenie sposobu rozwiązania zadania oraz prezentacji tego rozwiązania (pełny opis można znaleźć w: *Standardy wymagań będące podstawą przeprowadzenia sprawdzianu*…). Kategoria ta z 2014 roku zdaje się pokrywać częściowo z obszarem modelowania matematycznego z 2015 roku, zdefiniowanym jako umiejętność dobrania odpowiedniego modelu matematycznego do prostej sytuacji, stosowania poznanych wzorów i zależności, przetwarzania tekstu zadania na działania arytmetyczne i proste równania.

W 2014 roku w strukturze egzaminu wyróżniona została wyraźnie kategoria wykorzystywania wiedzy w praktyce, która była reprezentowana przez 8 zadań. W 2015 roku ta umiejętność ogólna nie została wyszczególniona. Czy to oznacza, że egzamin w 2015 roku w mniejszym stopniu mierzył umiejętność wykorzystywania wiedzy w praktyce? Po pierwsze, należy zwrócić uwagę na to, że w przypadku co najmniej dwóch zadań z tej kategorii ze sprawdzianu z 2014 roku można mieć wątpliwości, czy odwołują się one do kontekstu praktycznego, jest to zadanie 13 o treści: „Jaka jest łączna długość wszystkich krawędzi prostopadłościanu o wymiarach podanych na rysunku obok?" oraz zadanie 21 o treści: „Rysunek przedstawia fragment siatki prostopadłościanu. Uzupełnij siatkę – dorysuj brakujące ściany". Oba zadania są osadzone w czysto szkolnym – matematycznym kontekście. Na egzaminie z 2015 roku pojawiają się natomiast zadania, których kontekst jest typowo praktyczny: zadanie 19, w którym trzeba ustalić proporcje przepisu na ciasto, zadanie 21 dotyczące opłat za ogrzewanie, zadanie 25 o opłatach za SMS-y, zadanie 26, w którym trzeba obliczyć długość trasy wycieczki, czy zadanie 27, w którym trzeba ustalić ilość potrzebnej ziemi. Można więc przyjąć, że sprawdzian przeprowadzony w 2015 roku badał więc również umiejętność wykorzystywania wiedzy w sytuacjach praktycznych.

W strukturze egzaminu z 2015 roku wyróżniono dodatkowo kategorię sprawności rachunkowej mierzonej przez 3 zadania WW. Kategoria ta nie jest wyszczególniona dla sprawdzianu do roku 2014, w związku z czym nie pojawiały się wtedy zadania wymagające tylko obliczenia wartości jakiegoś wyrażenia. Sprawność rachunkowa była badana tylko pośrednio o tyle, o ile była ona potrzebna do wybrania lub uzyskania prawidłowej odpowiedzi w innych zadaniach.

Przedstawione powyżej zestawienie pokazało, że struktura mierzonych umiejętności na sprawdzianie w nowej formule jest bardzo podobna do tej na egzaminie w 2014 roku. W przypadku zadań z języka polskiego różnice są drobne i polegają przede wszystkim na trochę mniejszym udziale zadań mierzących umiejętność tworzenia wypowiedzi a troszkę większym znaczeniu zadań odwołujących się do umiejętności czytania. Całościowo za zadania z języka polskiego w 2015 roku można było uzyskać o 1 punkt więcej niż rok wcześniej, więc można powiedzieć, że nieznacznie wydłużyła się skala umiejętności polonistycznych. W przypadku zadań matematycznych struktura mierzonych umiejętności została trochę inaczej zdefiniowana. Zmiany polegały w dużej mierze na zmianie definicji kategorii mierzonych umiejętności, przez co trudno bezpośrednio je porównać. Dostrzega się jednak duże podobieństwo mierzonych umiejętności szczegółowych. Stąd uzasadnione wydaje się stwierdzenie, że grupy zadań matematycznych w obu testach mierzą podobne konstrukty. Dodatkowo podskale matematyczne są tej samej długości (20 punktów). Warte podkreślenia jest to, że mimo iż do 2014 roku mówiło się, że sprawdzian jest testem wiedzy ogólnej i mierzy umiejętności ponadprzedmiotowe, to struktura testu z 2014 roku jest bardzo podobna do sprawdzianu w nowej formule, w którym wyraźnie wyróżnia się część z języka polskiego i z matematyki. Wydaje się więc, że zmiana polegająca na wyróżnieniu w teście szóstoklasisty części z języka polskiego i z matematyki zaszła już wcześniej.

Jedna czy dwie skale umiejętności?

Jedną z wyraźnie zauważalnych zmian w formule sprawdzianu jest wydzielenie części z języka polskiego i z matematyki. Zestawienie przedstawione powyżej pokazało jednak, że sprawdzian w poprzedniej formule miał podobną strukturę mierzonych umiejętności. Czy zatem coś się zmieniło? Do 2014 roku sprawdzian był uważany za test wiedzy ogólnej, mierzący ponadprzedmiotowe umiejętności na jednej skali. Czy sprawdzian w nowej formule możemy traktować podobnie? A może istnieje konieczność uwzględnienia wymiaru umiejętności polonistycznych i matematycznych?

Pytania te sprowadzają się do problemu jednowymiarowości testu. Czy mierzone przez pierwszą część sprawdzianu w nowej formule umiejętności możemy sensownie opisać jednym wymiarem umiejętności ogólnych? Założenie o jednowymiarowości skali jest jednym z podstawowych założeń modelu Rascha (Smith, 1996). Złamanie tego założenia jest równoznaczne ze złamaniem założenia o lokalnej niezależności odpowiedzi na zadania (Modzelewski, 2015). Zadania mierzące dodatkowy wymiar oprócz związku z głównym wymiarem mierzą w takim przypadku jeszcze inny konstrukt, przez co nie są od siebie niezależne.

W celu odpowiedzi na pytanie o jednowymiarowość testu wykonane zostały analizy weryfikujące, czy po uwzględnieniu wpływu umiejętności ogólnej na odpowiedzi uczniów na zadania, obserwujemy jeszcze systematyczne związki między zadaniami, co do których podejrzewamy, że mogłyby być wskaźnikami dodatkowych wymiarów (umiejętności z zakresu języka polskiego i matematyki). Wyniki sprawdzianu zostały w pierwszej kolejności wyskalowane jednowymiarowym modelem Rascha. Następnie zapisano reszty z modelu (różnice między prawdopodobieństwem udzielenia danej odpowiedzi wynikającym z modelu a rzeczywistą odpowiedzią), a ich wartości zostały wystandaryzowane w obrębie zadań[2] (dzięki czemu wariancja reszt dla różnych zadań jest wyrównana). Reszty z modelu poddano analizie głównych składowych, wykorzystując macierz kowariancji. Podejście takie jest stosowane w teorii IRT do wykrywania wielowymiarowości, a badania symulacyjne pokazały, że tego typu metody oparte na analizie dopasowania do modelu Rascha są skuteczniejsze niż analiza czynnikowa lub analiza głównych składowych przeprowadzona bezpośrednio na odpowiedziach na zadania (Linacre, 1998; Smith, 1996). Opisane analizy zostały przeprowadzone dla sprawdzianu z 2014 roku oraz pierwszej części sprawdzianu z 2015 roku. Wyniki przedstawiono w tabeli 2.

[2] Reszty standaryzowane dają lepsze wyniki w zakresie wykrywania wielowymiarowości niż surowe czy wyrażone na skali logitowej (Linacre, 1998).

Tabela 2. Wyniki analizy głównych składowych na resztach dla zadań

2014					2015				
Składowa	Wartość własna	% wyjaśnionej wariancji	Zadanie	Ładunki dla 1 składowej	Składowa	Wartość własna	% wyjaśnionej wariancji	Zadanie	Ładunki dla 1 składowej
1	1,855	7,73%	**z1**	**0,048**	1	1,573	6,57%	**z1**	**0,059**
2	1,427	5,95%	**z2**	**0,071**	2	1,361	5,69%	**z2**	**0,068**
3	1,270	5,30%	**z3**	**0,069**	3	1,221	5,10%	**z3**	**0,083**
4	1,180	4,92%	**z4**	**0,052**	4	1,160	4,85%	**z4**	**0,116**
5	1,125	4,69%	**z5**	**0,036**	5	1,084	4,53%	**z5**	**0,096**
6	1,053	4,39%	**z6**	**-0,045**	6	1,081	4,52%	**z6**	**0,082**
7	1,014	4,23%	**z7**	**0,001**	7	1,034	4,32%	**z7**	**0,061**
8	0,981	4,09%	**z8**	**-0,014**	8	1,020	4,26%	**z8**	**0,091**
9	0,970	4,05%	**z9**	**-0,030**	9	0,987	4,12%	**z9**	**0,227**
10	0,952	3,97%	**z10**	**-0,045**	10	0,972	4,06%	**z10**	**0,061**
11	0,941	3,92%	z11	0,131	11	0,952	3,98%	**z11**	**0,107**
12	0,934	3,90%	z12	0,160	12	0,931	3,89%	**z12**	**0,390**
13	0,906	3,78%	z13	0,183	13	0,925	3,86%	**z13**	**0,560**
14	0,879	3,67%	z14	0,179	14	0,912	3,81%	z14	-0,030
15	0,864	3,60%	z15	0,181	15	0,883	3,69%	z15	-0,150
16	0,838	3,50%	z16	0,148	16	0,854	3,57%	z16	-0,118
17	0,837	3,49%	z17	0,188	17	0,811	3,39%	z17	-0,153
18	0,820	3,42%	z18	0,113	18	0,772	3,23%	z18	-0,144
19	0,806	3,36%	z19	0,072	19	0,746	3,12%	z19	-0,192
20	0,782	3,26%	z20	0,050	20	0,744	3,11%	z20	-0,152
21	0,763	3,18%	z21	0,200	21	0,716	2,99%	z21	-0,226
22	0,752	3,13%	z22	0,215	22	0,714	2,98%	z22	-0,052
23	0,696	2,90%	z23	0,140	23	0,685	2,86%	z23	-0,107
24	0,658	2,74%	z24	0,179	24	0,665	2,78%	z24	-0,172
25	0,644	2,69%	**z25**	**-0,332**	25	0,564	2,36%	z25	-0,158
26	0,034	0,14%	**z26**	**-0,715**	26	0,505	2,11%	z26	-0,257
					27	0,057	0,24%	z27	-0,288

W tabeli wyróżniono pogrubieniem zadania odwołujące się do umiejętności z języka polskiego.

Jeśli zadania ze sprawdzianów mierzyły poza głównym wymiarem umiejętności ogólnych wymiary umiejętności z języka polskiego i umiejętności matematyczne, reszty dla zadań powinny być ze sobą powiązane w ramach tych dwóch podwymiarów. Wtedy ładunki na pierwszym komponencie dla zadań należących do tego samego podwymiaru powinny być sobie bliskie, a dalekie dla par zadań należących do różnych podwymiarów. Jeśli ładunki na pierwszej

składowej odzwierciedlają założone podwymiary testu, to ważne jest też to, ile wariancji reszt wyjaśnia pierwsza składowa. Jeśli odsetek ten jest znaczący, możemy przyjąć, że test faktycznie ma wielowymiarową strukturę.

W przypadku sprawdzianu z 2014 roku reszty dla zadań 25 i 26 okazały się ze sobą powiązane (dla obu ujemne ładunki o względnie dużych wartościach). Oba zadania mierzyły umiejętność pisania (zadanie 25 było to zadanie krótkiej odpowiedzi, a zadanie 26 to wypracowanie). Powiązanie reszt może być wynikiem podobieństwa mierzonych umiejętności, ale także może być związane z efektem egzaminatora (Kulon, 2015), a reszty mogą odzwierciedlać to, co nie jest wskaźnikiem umiejętności ogólnych, ale jest związane z surowością oceniającego lub wpływem poprzednio ocenianego zadania na ocenę kolejnego. Relatywnie wyższe pozytywne ładunki obserwujemy dla reszt dla zadań matematycznych, jednak nie dla wszystkich. Pierwsza składowa wyjaśnia ponadto 7,73% wariancji reszt, czyli niewiele.

Ładunki na pierwszej składowej dla reszt dla zadań ze sprawdzianu z 2015 roku układają się w bardziej jednoznaczny wzorzec (pozytywne ładunki dla zadań z polskiego i negatywne dla zadań z matematyki), jednak ich wartości nie są duże. Największe co do wartości bezwzględnej obserwujemy dla wypracowania (zadanie 13) i zadania otwartego krótkiej odpowiedzi (zadanie 12), mierzącego tym razem jednak umiejętność czytania. Pierwsza składowa wyjaśnia jednak tylko 6,57% wariancji reszt, czyli mniej niż na sprawdzianie rok wcześniej. Podobnie wielkość wartości własnej pokazuje, że pierwsza składowa, która odzwierciedla powiązanie ze sobą zadań ze względu na związek z podwymiarami, wyjaśnia niewiele ponad 1,5 więcej wariancji reszt niż średnio jedna zmienna (reszty dla jednego zadania).

Analizy te pokazały, że w sprawdzianach z 2014 i 2015 roku, choć daje się zauważyć niewielką zależność reszt dla zadań pochodzących z tych samych podskal, to nie są to zależności silne. Daje to podstawę do przyjęcia, że analizowane testy są jednowymiarowe w wystarczającym stopniu, by posługiwać się jednym wynikiem reprezentującym umiejętności ogólne. Dodatkowo należy podkreślić, że wyniki sprawdzianu w nowej formule i sprawdzianu z 2014 roku są podobne, co pokazuje, że sprawdzian z 2015 roku możemy w co najmniej porównywalnym stopniu traktować jako test wiedzy ogólnej, jak egzamin realizowany rok wcześniej.

Jakość skal pomiarowych

Wyskalowany wynik z testu będzie dobrym wskaźnikiem mierzonych umiejętności, a skala będzie miała pożądane właściwości wynikające z zastosowania przyjętego modelu tylko wtedy, gdy dane będą dobrze dopasowane do modelu. Wystarczająco dobre dopasowanie będzie także świadczyło o tym, że mierzone umiejętności mogą być z powodzeniem opisane jednym wymiarem umiejętności (Modzelewski, 2015). W tabeli 3 podano statystyki opisowe dla miar dopasowania wyliczonych dla zadań ze sprawdzianów z 2014 i 2015 roku (a dokładniej dla każdej kategorii punktowej zadania wyznaczono niezależne miary). Wykorzystano statystyki infit i outfit. *Outfit* jest równie wrażliwa na odstępstwa od modelu dla całego zakresu skali, natomiast *infit* jest bardziej

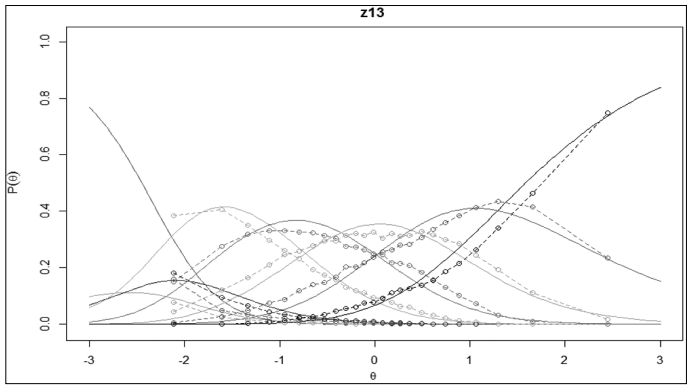
wrażliwa na odstępstwa dla tych przypadków na skali, które znajdują się w okolicy parametru trudności zadania. Im wartość bliższa 1, tym zadanie jest lepiej dopasowane do modelu. Dopuszcza się jednak pewne rozchwianie wartości tych statystyk (De Ayala, 2009, s. 55–57), przyjmując różne wartości graniczne.

Tabela 3. Statystyki opisowe dla miar dopasowania

Statystyka	2014		2015	
	Outfit	Infit	Outfit	Infit
Minimum	0,76	0,88	0,62	0,85
10 percentyl	0,85	0,92	0,77	0,91
25 percentyl	0,93	0,96	0,84	0,94
Mediana	1,00	1,01	0,99	1,00
75 percentyl	1,31	1,07	1,30	1,09
90 percentyl	1,81	1,17	2,05	1,17
Maksimum	3,50	1,20	217,10	1,29
Liczba zad. z wart. statystyki poza przedziałem (0,8;1,2)	5	1	8	1
Liczba zad. z wart. statystyki poza przedziałem (0,5;1,5)	2	0	1	0

Dopasowanie zadań do modelu jest umiarkowanie dobre zarówno dla egzaminu z 2014, jak i 2015 roku. Miary *infit* i *outfit* dla zadań na dłuższych skalach obliczono dla każdej wartości punktowej (np. za wypracowanie w 2015 roku można było uzyskać maksymalnie 7 punktów, więc miary dopasowania zostały policzone dla każdej z 7 kategorii). Zadania te wykazywały zwykle gorsze dopasowanie do modelu, co w połączeniu z analizą statystyk opisowych może prowadzić do mylnie pesymistycznych wniosków na temat odsetka zadań źle dopasowanych (są to statystyki dla kategorii punktowych, a nie dla zadań). Z tego powodu podano także liczbę zadań, dla których choć dla jednej kategorii punktowych wielkość miary dopasowania wykracza poza granice powszechnie przyjętych przedziałów. Biorąc pod uwagę bardziej restrykcyjne kryterium dopasowania, należałoby stwierdzić, że 8 zadań ze sprawdzianu z 2015 roku było niedopasowanych do jednowymiarowego modelu Rascha. Przy kryterium mniej rygorystycznym, jedno zadanie – wypracowanie – wykazuje niedopasowanie, a miary *outfit* są wysokie dla każdej wartości punktowej, choć wyraźnie wyższe dla kategorii 1 i 2 punktów[3]. Przyjrzyjmy się jednak rysunkowi z krzywymi charakterystycznymi dla tego zadania jako innemu kryterium oceny dopasowania. Na rysunku 1 liniami ciągłymi przedstawiono krzywe charakterystyczne z modelu, a odpowiadającymi im liniami przerywanymi krzywe empiryczne. Widoczne są zakresy skali, w których krzywe empiryczne nie pokrywają się dokładnie z krzywymi wynikającymi z modeli, jednak nie są to odstępstwa drastyczne. Warto też podkreślić to, że 1 punkt za zadanie 13 uzyskało mniej niż 1% uczniów, co może uzasadniać tak duże raportowane niedopasowanie dla tej kategorii punktowej.

[3] Warto dodać, że dopasowanie tego zadania nie poprawia się w modelu uwzględniającym tylko zadania z języka polskiego, więc nie jest to związane z potraktowaniem całego testu jako jednowymiarowego.



Rysunek 1. Krzywe charakterystyczne dla zadania 13 (wypracowanie) ze sprawdzianu w 2015 roku

Podsumowując, należy stwierdzić, że dopasowanie danych ze sprawdzianu w nowej formule do modelu Rascha jest porównywalnej jakości jak na egzaminie rok wcześniej, choć wypracowanie z 2015 roku w mniejszym stopniu spełnia założenia modelu.

Kolejnym kryterium jakości skali jest jej rzetelność. Do analizy rzetelności wykorzystano współczynnik rzetelności EAP/PV – jedną z miar rzetelności wykorzystywaną w przypadku modeli IRT (Adams, 2005; Jasińska i Modzelewski, 2014). Jest to stosunek wariancji oszacowania punktowego (EAP – *expected a posteriori*) do całkowitej wariancji zmiennej ukrytej (obliczanej jako wariancja PV – *plausible values*). Przyjmuje ona wartości od 0 do 1. Rzetelność skal policzono zarówno dla całych testów, jak i osobno dla skali z języka polskiego i osobno dla matematyki. Wyniki zaprezentowano w tabeli 4.

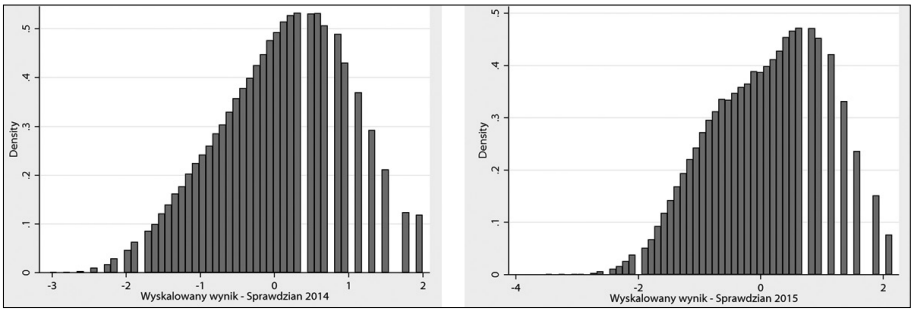
Tabela 4. Rzetelność skal

Sprawdzian	2014	2015
Cały test	0,859	0,863
Język polski	0,755	0,716
Matematyka	0,804	0,820

Rzetelność sprawdzianu w nowej formule jest porównywalna do rzetelności testu z 2014 roku. Jest ona umiarkowanie wysoka, wystarczająca do budowania wskaźników zagregowanych. Analiza w rozbiciu na dwa testy pokazała, że skala z języka polskiego jest znacznie mniej rzetelna niż skala matematyczna. Różnica w rzetelnościach jest większa w 2015 roku. Rzetelność testu z języka polskiego w 2015 roku jest dość niska, nawet na potrzeby konstrukcji zagregowanych wskaźników mogących nieść z sobą znaczące konsekwencje.

Dopełnieniem obrazu jakości skal będą rozkłady wyskalowanych wyników. Przedstawiono je na rysunku 2. Rozkład wyników sprawdzianu z 2014 roku jest bliższy rozkładowi normalnemu. W przypadku sprawdzianu z 2015 roku widać większe odstępstwa. Jego kształt jest konsekwencją dwumodalnego rozkładu wyników z testu matematycznego z tego roku (rozkład dostępny

w raporcie: CKE, 2015). Oba testy, jako że były testami dość łatwymi, słabiej różnicują osoby o wyższym poziomie umiejętności, co ujawnia się w skokowym charakterze skali w zakresie wyników wysokich.



Rysunek 2. Rozkłady wyskalowanych wyników: sprawdzian 2014 i sprawdzian 2015

Precyzja szacowania wyników uczniów o różnym poziomie umiejętności

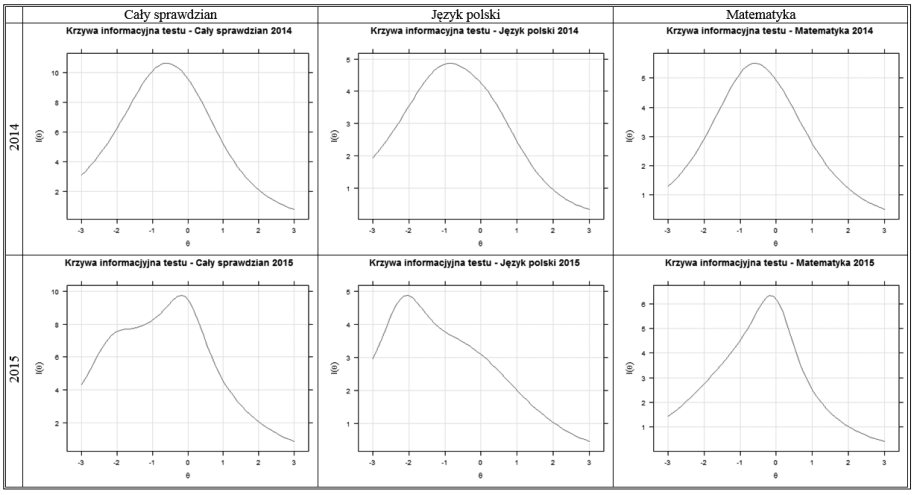
Na jakość skal pomiarowych możemy spojrzeć też przez pryzmat tego, z jaką precyzją pozwalają oszacować wyniki uczniów o różnym poziomie umiejętności. Umożliwia to narzędzie, jakim jest krzywa informacyjna testu (De Ayala, 2009), bazująca na informacji Fishera (Wright, 1990). Im większa wartość funkcji informacyjnej, tym większa precyzja pomiaru. W modelu Rascha ilość informacji w jakimś zakresie skali zależy od liczby zadań o określonej trudności. Krzywa będzie więc przybierała różne kształty w zależności od rozkładu trudności zadań uwzględnionych w teście. Porównując kształt krzywej informacyjnej z rozkładem umiejętności uczniów, można ocenić, na ile test pozwala precyzyjnie szacować wyniki uczniów o przeciętnych, niskich czy wysokich umiejętnościach. Dzięki temu możemy ocenić, na ile jest on dopasowany do pomiaru umiejętności określonej populacji i do celu, do którego został stworzony.

Sprawdzian w 2014 roku pozwalał na szacowanie z największą precyzją wyników uczniów o trochę poniżej średnim poziomie umiejętności (0 na skali na wykresie jest równe średniej). Precyzja pomiaru zmniejszała się podobnie w miarę oddalania się w obie strony od tego maksimum. W konsekwencji test trochę dokładniej mierzył umiejętności uczniów słabszych niż zdolnych. Dla skali zbudowanej z zadań z języka polskiego oraz skali umiejętności matematycznych krzywe te przyjmowały podobne kształty jak ta dla całego testu (choć oczywiście wartość funkcji informacyjnej jest niższa).

W przypadku sprawdzianu z 2015 roku maksimum funkcji informacyjnej także wypada niewiele poniżej średniego wyniku, jednak spadek informacji (czyli wzrost błędu pomiaru) wraz ze wzrostem poziomu umiejętności jest znacznie większy niż w zakresie niższych wyników, gdzie obserwujemy także pewne wypłaszczenie. Sprawdzian z 2015 roku pozwala na szacowanie z dość dobrą precyzją wyników uczniów o poniżej przeciętnych umiejętnościach. Precyzja pomiaru w zakresie umiejętności powyżej przeciętnych jest wyraźnie niższa. Oznacza to, że test składał się przede wszystkim z zadań łatwych i średnio trudnych. Testy z języka polskiego i matematyki mają jednak inne

profile krzywych informacyjnych. Test z języka polskiego pozwala na szacowanie z największą precyzją wyników uczniów o bardzo niskim poziomie umiejętności. Wraz ze wzrostem poziomu umiejętności informacja dostarczana przez test wyraźnie spada, przez co nawet wyniki uczniów o przeciętnym poziomie umiejętności szacowane są już ze sporym błędem. Test z matematyki najdokładniej mierzy poziom umiejętności w okolicach średniej, a informacja znacząco spada wraz z oddalaniem się od średniego wyniku, przy czym spadek w zakresie wyższych wyników jest trochę bardziej gwałtowny. Co ważniejsze, różnych kształtów krzywych informacyjnych testów z języka polskiego i matematyki nie daje się wytłumaczyć celem pomiaru. Testy z poszczególnych przedmiotów na sprawdzianie służą bowiem tym samym celom. Wydaje się, że taki kształt krzywych jest efektem braku wystarczającej kontroli trudności zadań wchodzących w skład testów na etapie ich konstrukcji. W 2014 roku, mimo formalnego niewyodrębniania dwóch części testu, z języka polskiego i matematyki udało się stworzyć skale o profilu precyzji pomiaru dającym się łatwiej uzasadnić. Może to szczęśliwy przypadek? Jeśli tak, to w 2015 roku tego szczęścia zabrakło. Konsekwencją jest dość niska precyzja szacowania wyników uczniów o wyższym poziomie umiejętności, w szczególności w zakresie języka polskiego. To stawia pod znakiem zapytania wiarygodność wskaźników liczonych dla szkół, których absolwenci uzyskują bardzo wysokie wyniki na sprawdzianie, a w szczególności jego części z języka polskiego.

Tabela 5. Krzywe informacyjne testu



Podsumowanie

Przedstawione w artykule wyniki pokazały, że sprawdzian przeprowadzony w 2015 roku (jego pierwsza część), mimo zmiany formuły egzaminu, zachował podobną strukturę mierzonych umiejętności oraz ma podobne właściwości psychometryczne do sprawdzianu z 2014 roku. Mimo wydzielenia części z języka polskiego i z matematyki wyniki sprawdzianu w nowej formule można przedstawić na jednowymiarowej skali ogólnych umiejętności. Większość

zadań jest dobrze dopasowana do jednowymiarowego modelu Rascha; największe niedopasowanie stwierdzono dla wypracowania, które jest oceniane na dość długiej skali. Być może jej skrócenie poprawiłoby jakość dopasowania zadania do modelu, jednak z punktu widzenia przygotowania skali osiągnięć na potrzeby modeli EWD dla szkół podstawowych jest to niemożliwe[4]. Pozostaje zachowanie w pamięci faktu, że informacja dostarczana na podstawie wyników za to zadanie jest obarczona większym błędem. Skala utworzona z wykorzystaniem jednowymiarowego modelu Rascha dla całego sprawdzianu ma wystarczająco dobrą rzetelność na potrzeby budowania miar zagregowanych, takich jak wskaźniki EWD. Sprawdzian z 2015 roku pozwolił na szacowanie z największą precyzją wyników uczniów o przeciętnym i poniżej przeciętnym poziomie umiejętności. Wyniki uczniów o wyższym poziomie umiejętności są obarczone większym błędem.

Wątpliwości wzbudzają natomiast skale zbudowane dla przedmiotów, w szczególności skala umiejętności z języka polskiego. Ma ona dość niską rzetelność, a także niską precyzję szacowania wyników uczniów o przeciętnym i powyżej przeciętnym poziomie umiejętności. Skala umiejętności matematycznych ma lepsze właściwości psychometryczne pod tym względem.

Z punktu widzenia budowy skal na potrzeby modeli EWD dla II etapu edukacyjnego uzasadnione wydaje się wykorzystanie jednego wyniku opisującego ogólne umiejętności. Skala taka ma wystarczająco dobre właściwości psychometryczne i pozwala na zachowanie ciągłości opracowanych modeli i interpretacji wskaźników. Wykorzystanie wyników dla skal przedmiotowych jest rekomendowane tylko wtedy, gdy zyski ze stosowania wskaźników w podziale na przedmioty przewyższają możliwe ryzyko związane z posługiwaniem się miarami o niewystarczająco dobrych właściwościach. W szczególności należy pamiętać, że oszacowania wyników z języka polskiego dla szkół, których absolwenci mają bardzo wysoki poziom tych umiejętności, obarczone będą dużym błędem. Wskaźniki z matematyki nie budzą tak wielu wątpliwości, więc można również rozważyć wykorzystanie tylko tej skali przedmiotowej.

Bibliografia

1. Adams R. J., *Reliability as a Measurement Design Effect*, "Studies in Educational Evaluation" 2005, nr 31(2), s. 162–172.
2. Ballou, D., *Test Scaling and Value-Added Measuremen*t. "Education Finance and Policy" 2009, nr 4(4), s. 351–383.
3. CKE, *Informator o sprawdzianie od roku szkolnego 2014/201*5, Centralna Komisja Egzaminacyjna, Warszawa 2013.

[4] Zmieniłoby to maksymalną liczbę punktów możliwych do zdobycia i uniemożliwiło przeliczenie wyników uczniów (przechowywanych zwykle tylko jako wyniki za cały test i poszczególne części, bez pełnego rekordu odpowiedzi) na wyniki wyskalowane.

4. CKE, *Rozkłady wyników i parametry statystyczne rozkładu wyników sprawdzianu przeprowadzonego w kwietniu 2015 r.*, CKE 2015, http://www.cke.edu.pl/images/_SPRAWDZIAN/OD_2014/Informacje_o_wynikach/20150529_SPRAWDZIAN_2015_Wstepne_informacje_o_wynikach_Parametry_i_rozklady.pdf [dostęp: 15.06.2015].
5. De Ayala, R. J, *The theory and practice of item response theory*, Guilford Press, New York 2009.
6. Dolata, R., Hawrot, A., Humenny, G., Jasińska, A., Koniewski, M., Majkut, P. i Żółtak, T., *Trafność metody edukacyjnej wartości dodanej dla gimnazjów*, Instytut Badań Edukacyjnych, Warszawa 2013.
7. Dolata, R., Hawrot, A., Humenny, G., Jasińska-Maciążek, A., Rappe, A., Stożek, E. i Żółtak, T., *Metoda edukacyjnej wartości dodanej w Polsce*, [w:] Liczą się egzaminy. Raport o stanie edukacji, Instytut Badań Edukacyjnych, Warszawa.
8. *Informator z aneksem dla uczniów ze specyficznymi trudnościami w uczeniu się. Sprawdzian w klasie szóstej szkoły podstawowej przeprowadzony od roku szkolnego 2009/2010*, Centralna Komisja Egzaminacyjna, Warszawa 2009.
9. Jakubowski, M. i Pokropek, A., *Badając egzaminy: podejście ilościowe w badaniach edukacyjnych*, Centralna Komisja Egzaminacyjna, Warszawa 2009.
10. Jasińska, A. i Modzelewski, M., *Testy osiągnięć szkolnych TOS3: przykład narzędzia skonstruowanego z wykorzystaniem modelu Rascha*. „Edukacja" 2014, nr 2(127), s. 85–107.
11. Kiefer, T., Robitzsch, A. i Wu, M. TAM: *Test Analysis Modules*, 2015, http://CRAN.R-project.org/package=TAM.
12. Kondratek, B. i Pokropek, A., *IRT i pomiar edukacyjny*. „Edukacja" 2013, nr 4(124), s. 42–66.
13. Kulon, F., *Modele analizy efektu oceniającego*, [w:] A. Pokropek (red.), Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania, s.185–203, Instytut Badań Edukacyjnych, Warszawa 2015.
14. Linacre, J. M., *Detecting multidimensionality: which residual data-type works best?,* "Journal of Outcome Measurement" 1998, nr 2, s. 266–283.
15. Modzelewski, M., *Specyficzne własności modelu Rascha*, [w:] A. Pokropek (red.), Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania, Instytut Badań Edukacyjnych, Warszawa 2015.
16. Pokropek, A., *Trafność testów egzaminacyjnych*, [w:] M. Karwowski (red.), *Ścieżki rozwoju edukacyjnego młodzieży – szkoły pogimnazjalne,* Wydawnictwo Instytutu Filozofii i Socjologii PAN, Warszawa 2013.
17. Rasch, G., *Probabilistic models for some intelligence and attainment tests*, University of Chicago Press, Chicago 1980.
18. Reardon, S. F. i Raudenbush, S. W., *Assumptions of value-added models for estimating school effects,* "Education Finance and Policy" 2009, nr 4(4), s. 492–519.
19. Smith, R. M., *A comparison of methods for determining dimensionality in Rasch measurement*, "Structural Equation Modeling: A Multidisciplinary Journal" 1996, nr 3(1), s. 25–40.
20. *Standardy wymagań będące podstawą przeprowadzenia sprawdzianu w ostatnim roku nauki w szkole podstawowej*, www.cke.edu.pl/images/stories/Standardy/masowe_spr.pdf, [dostęp 15.06.2015].
21. Szaleniec, H. (red.), *Teoria wyniku zadania IRT: zastosowania w polskim systemie egzaminów zewnętrznych: praca zbiorowa*. Centralna Komisja Egzaminacyjna, Warszawa 2009.
22. Wright, B. D., What is information? „Rasch Measurement Transactions" 1990, nr 4(2), s. 109.