Paulina Skórska
Uniwersytet Jagielloński

Maciej Koniewski
Instytut Badań Edukacyjnych, Uniwersytet Jagielloński

Przemysław Majkut
Okręgowa Komisja Egzaminacyjna w Krakowie, Uniwersytet Jagielloński,
Instytut Badań Edukacyjnych

Wpływ wersji arkusza egzaminacyjnego na zróżnicowane funkcjonowanie zadań na przykładzie egzaminu gimnazjalnego

Wprowadzenie

Jednym z istotnych problemów, z którym borykają się systemy egzaminacyjne na całym świecie, są oszustwa egzaminacyjne. Przybierają one różne formy, których klasyfikację w odniesieniu do polskich warunków edukacyjnych przedstawił Bolesław Niemierko (2006). Na najbardziej ogólnym poziomie oszustwa egzaminacyjne można podzielić na oszustwa dokonywane przez uczniów („ściąganie") i nauczycieli (np. pomoc w trakcie egzaminu, korekta odpowiedzi na arkuszu). Gdy dochodzi do oszustwa egzaminacyjnego, wyniki egzaminu przestają odzwierciedlać poziom umiejętności i wiedzy uczniów. Oszustwa egzaminacyjne stanowią więc zagrożenie dla trafnej interpretacji wyników testów jako miary poziomu wiedzy i umiejętności uczniów.

W Polsce zagadnieniem oszustwa egzaminacyjnego zajmował się także Henryk Szaleniec (2006). W swojej pracy wskazywał na „ściąganie" jako jedną z zasadniczych form oszustwa egzaminacyjnego. Problem „ściągania" jest związany z wykorzystaniem pytań zamkniętych jako głównej formy pytań testowych. W ramach systemu egzaminów zewnętrznych podejmowane są działania mające na celu ograniczenie do minimum zjawiska „ściągania". Jednym z nich było wprowadzenie różnych form arkuszy testowych, różniących się od siebie kolejnością możliwych do wyboru odpowiedzi na dane pytanie. Początkowo w arkuszach egzaminacyjnych były wskazane wersje testu. Obecnie są one zakodowane w kodzie kreskowym arkusza, co uniemożliwia łatwe rozpoznanie przez ucznia, którą wersję rozwiązuje on sam i uczniowie siedzący obok.

Wypowiadając się o wiedzy i umiejętnościach uczniów na podstawie wyników egzaminów, zakładamy, że są one trafnym wskaźnikiem ich wiedzy i umiejętności. Istotne jest również to, że powinny być one niezależne od innych cech ucznia (np. statusu społeczno-ekonomicznego czy płci) oraz od samej sytuacji testowej (Hornowska, 1999). Fakt istnienia różnych wersji testu zmienia sytuację egzaminacyjną uczniów. Pojawia się zatem pytanie, czy istnieje związek

między rozwiązywaną przez uczniów wersją testu a funkcjonowaniem poszczególnych pytań testowych. Problem ten zostanie przeanalizowany na przykładzie arkuszy egzaminacyjnych z przedmiotów humanistycznych (historia i wiedza o społeczeństwie) w wersji standardowej, rozwiązywanych na egzaminach gimnazjalnych w 2012 i 2013 r. W analizach zostaną wykorzystane wyniki egzaminów gimnazjalnych uczniów ze szkół działających na terenie OKE w Krakowie.

Problematyka związana ze zróżnicowaniem funkcjonowania zadań ze względu na cechy ucznia lub sytuacji testowej w literaturze znana jest pod nazwą *differential item functioning* (DIF). Można powiedzieć, że pozycja (pytanie) testowa jest obciążona, jeśli uczniowie o takim samym poziomie wiedzy i umiejętności, ale różniący się jakąś cechą (np. płcią, przynależnością do mniejszości narodowej, faktem rozwiązywania różnych form arkusza egzaminacyjnego) mają nierówne prawdopodobieństwo udzielenia prawidłowej odpowiedzi na dane pytanie testowe (Ironson, 1982; Linn i in., 1981). W polskich badaniach edukacyjnych diagnoza różnego funkcjonowania pozycji testowych jest stosunkowo rzadko przedstawiana. Przykładem jej zastosowania jest praca Magdaleny Grudniewskiej i Bartosza Kondratka (2012), w której autorzy zaprezentowali wyniki analizy różnego funkcjonowania zadań egzaminu gimnazjalnego w części matematyczno-przyrodniczej ze względu na płeć.

Analiza właściwości psychometrycznych zadań arkusza z historii i WOS z 2012 r.

Badanie zależności funkcjonowania zadań z historii i WOS rozpoczęto od prześledzenia zmian w poziomie wykonania zadań na egzaminie w 2012 i 2013 roku. Poziom wykonania zadania to odsetek uczniów, którzy odpowiedzieli na dane pytanie poprawnie. Jest to jedna z podstawowych miar używana w ramach Klasycznej Teorii Testów (KTT) do oceny jakości zadania. Pozwoliła ona na ocenę, które zadania różnią się od siebie między różnymi arkuszami testu. Zadania o największych różnicach zostały następnie poddane analizie testem Mantela-Haenszela (M-H) (Mantel i Haenszel, 1959). Test polega na porównaniu wyników egzaminacyjnych dwóch grup – jednej będącej przedmiotem zainteresowania (*focal group*) i drugiej stanowiącej grupę odniesienia (*reference group*), np. dziewczynki *vs* chłopcy. Możliwe jest zastosowanie testu M-H także wtedy, gdy dwie grupy są wyróżnione ze względu na charakterystyczne cechy testu. Tak jak ma to miejsce w przypadku różnych wersji arkusza (A i B). Hipoteza zerowa testu M-H zakłada, że szanse poprawnej odpowiedzi na dane pytanie testowe są równe w obu porównywanych grupach. Jako zmiennej kontrolnej użyto wyniku ucznia na teście. Test M-H jest najczęściej wykorzystywaną metodą w analizie DIF (Holland i Weiner, 1993). W analizach testem M-H wykorzystano program Stata 12[1] oraz IBM SPSS Statistics 21.

Od 2012 roku zadania z historii i WOS są zbierane w osobnym arkuszu, który stanowi jedną z części egzaminu gimnazjalnego w części humanistycznej. Wszystkie pytania są zamknięte, tzn. uczeń opowiada na nie, zaznaczając

[1] Specjalne podziękowania kierujemy do Bartosza Kondratka, który udostępnił nam przygotowane przez siebie procedury do liczenia wielkości efektu napisane w programie Stata.

odpowiednią odpowiedź na karcie odpowiedzi. Zadania, za które uczeń mógł otrzymać więcej niż 1 punkt, zostały rozbite na elementy, tak więc ostatecznie wszystkie analizowane pozycje testowe przybrały formę dychotomiczną (0-1).

Tabela 1. Porównanie poziomów wykonania zadań z historii i WOS z 2012 r.

	Poziom wykonania zadania (w nawiasie poprawna odpowiedź)		Różnica w poziomach wykonania zadania między wersjami A i B arkusza
	wersja A	wersja B	
z1	0,36 (B)	0,35 (B)	0,01
z2	0,57 (C)	0,58 (B)	0,01
z3	0,66 (E)	0,67 (D)	0,00
z4_1	0,34 (C)	0,31 (D)	0,02
z4_2	0,52 (D)	0,43 (C)	0,08
z5	0,57 (C)	0,49 (D)	0,08
z6	0,68 (A)	0,63 (C)	0,05
z7	0,66 (B)	0,64 (B)	0,02
z8	0,74 (A)	0,69 (C)	0,05
z9	0,79 (D)	0,80 (A)	0,01
z10	0,63 (B)	0,62 (B)	0,01
z11	0,52 (D)	0,54 (D)	0,02
z12_1	0,66 (B)	0,63 (B)	0,03
z12_2	0,81 (C)	0,76 (B)	0,05
z12_3	0,52 (B)	0,57 (A)	0,04
z13	0,79 (D)	0,78 (D)	0,01
z14	0,69 (C)	0,66 (D)	0,03
z15_1	0,72 (C)	0,72 (B)	0,00
z15_2	0,63 (B)	0,61 (D)	0,02
z15_3	0,75 (D)	0,72 (C)	0,03
z16_1	0,97 (A)	0,97 (B)	0,00
z16_2	0,86 (B)	0,84 (C)	0,02
z16_3	0,58 (C)	0,56 (C)	0,02
z17	0,50 (D)	0,51 (C)	0,01
z18	0,39 (A)	0,39 (A)	0,00
z19	0,55 (A)	0,55 (A)	0,00
z20	0,53 (B)	0,53 (B)	0,01
z21	0,70 (B)	0,63 (C)	0,07
z22	0,56 (C)	0,60 (B)	0,04
z23	0,42 (D)	0,43 (D)	0,01
z24_1	0,71 (D)	0,77 (C)	0,06
z24_2	0,90 (C)	0,88 (B)	0,02
z24_3	0,48 (B)	0,54 (D)	0,06

Wskaźnik wykonania zadania możemy w przypadku kodowania dychotomicznego interpretować jako procent osób, które prawidłowo odpowiedziały na dane zadanie. Analizując tabelę 1, widzimy 10 zadań, dla których różnica wskaźnika wykonania zadania między wersjami jest równa lub większa niż 0,04 (czyli 4%). Na szczególną uwagę zasługują zadania z4_2 oraz z5, gdzie różnica w poprawnych odpowiedziach między wersjami jest szczególnie duża

i wynosi 8%. W przypadku tych zadań lepiej wypadli uczniowie piszący wersję A, jednak nie jest to regułą w przypadku pozostałych wyróżnionych zadań. Nie ma większych różnic między średnimi wynikami w arkuszach w wersji A (wyniosła ona 20,8 punktu) oraz w wersji B (20,4 punktu)[2].

Wyniki testu M-H okazały się nieistotne statystycznie (na poziomie $p<0,05$) tylko w przypadku siedmiu pozycji testowych. Wskazywałoby to na obciążenie DIF większości pozycji testowych arkusza. Istotność różnic między grupami niekoniecznie oznacza, że efekt (oddziaływanie obciążenia DIF) jest duży i znaczący w kontekście praktycznym. Zwłaszcza w przypadku dużych prób osiągnięcie istotności statystycznej nie jest problemem. Z tego względu nie należy poprzestawać na analizie istotności wyników testu, ale porównać statystyk wielkości efektu (*effect size)* (ES) (Cohen, 1988).

Tabela 2. Wyniki testu MH oraz wielkości efektów obciążenia pozycji testowych z arkusza z historii i WOS za 2012 r.

	chi-kwadrat MH	istotność	cOR	ln(cOR)	MH P-DIF	STD P-DIF
z4_2	518,40	0,000	0,71	-0,34	0,09	0,07
z5	510,22	0,000	0,71	-0,34	0,08	0,07
z12_3	337,23	0,000	1,34	0,29	-0,07	-0,06
z21	430,34	0,000	0,72	-0,33	0,07	0,06
z24_1	724,66	0,000	1,61	0,48	-0,09	-0,07
z24_3	391,9	0,000	1,34	0,29	-0,07	-0,06

***Ograniczono się do pokazania w tabeli jedynie tych pozycji, które mają największą wielkość efektu ln(cOR) na poziomie 0,29 i więcej.**

Mantel i Haenszel (1959) wprowadzili wielkość efektu w postaci tzw. stałego ilorazu szans (*constant odds ratio*) (cOR). Ponieważ iloraz szans, przyjmując wartości od 0 do $+\infty$, gdzie wartość 1 wskazuje na brak obciążenia pozycji testowej (zerowy DIF), jest mało intuicyjny w interpretacji, częstą praktyką jest jego logarytmowanie (Holland i Weiner, 1993). Rozkład takiej wielkości efektu jest symetryczny wokół wartości 0, co ułatwia interpretację i jest bardziej intuicyjne (zerowy DIF w punkcie symetrii). Dodatkowo zastosowano dwie inne miary efektu, które należą do miar odnoszących wielkości efektu DIF do poziomu wykonania zadania (skali łatwości zadania), MH P-DIF oraz STD P-DIF (Kondratek i Grudniewska, 2013; Holland i Weiner,1993). Pozwalają one na interpretację, o ile dana pozycja testowa (zadanie) byłaby łatwiejsza/ trudniejsza w grupie będącej przedmiotem zainteresowania, gdyby funkcjonowała w niej tak, jak funkcjonuje w grupie odniesienia. Innymi słowy, jak zwiększyłoby/ zmniejszyłoby się prawdopodobieństwo poprawnej odpowiedzi na dane pytanie w grupie będącej przedmiotem zainteresowania (w tym przypadku wersja B testu), gdyby funkcjonowało tak samo jak w grupie odniesienia (w wykonywanych analizach wersja A testu)[3].

[2] Maksymalna liczba punktów do uzyskania wynosiła 33 punkty.

[3] Efekty MH P-DIF oraz STD P-DIF opierają się na warunkowej różnicy w łatwości/ trudności zadania, różnią się natomiast procedurą obliczeniową (Holland i Weiner,1993; Kondratek i Grudniewska, 2013). Wyjaśniając interpretację tych wielkości, można powiedzieć, że oznaczają różnicę między łatwością/ trudnością zadania w grupie będącej przedmiotem zainteresowania, a łatwością/ trudnością, jaką pozycja testowa miałaby w tej grupie, gdyby zadanie funkcjonowało tak jak w grupie odniesienia.

Tabela 2 zawiera (poza analizą istotności testu H-M) wartości wielkości efektu, tj. surowe (cOR) oraz zlogarytmowane (ln(cOR)). Dodatkowo zawarto w niej efekty wyrażone na skali łatwości zadania. Analiza potwierdza wnioski z tabeli 1. Pozycje testowe zidentyfikowane w niej jako problematyczne (duża różnica w poziomach wykonania zadania między wersjami A i B arkusza), charakteryzują się najwyższymi na tle pozostałych efektami DIF. Przyjęto, że efekty ln(cOR) odchylające się od 0 o 0,29 lub więcej mogą być najbardziej problematyczne.

Trzy zadania w wersji A dają uczniom szansę na uzyskanie nieco wyższych wyników (z4_2, z5, z21), trzy w wersji B (z12_3, z24_1, z24_3). Wielkość efektu we wskazanych zadaniach nie jest jednak bardzo duża. Jedyną cechą różniącą od siebie te pytania między arkuszami jest kolejność możliwych odpowiedzi. Ten wątek zostanie rozwinięty dokładniej w późniejszej części artykułu, gdzie zostaną podsumowane analizy DIF arkuszy z 2012 i 2013 r.

Analiza właściwości psychometrycznych zadań z historii i WOS z 2012 r.

Dla arkuszy z zakresu historii i wiedzy o społeczeństwie, rozwiązywanych przez uczniów w 2013 r. przeprowadzono analizy analogiczne jak dla arkuszy z 2012 r. Informacje dotyczące poziomu wykonania zadań w grupach wydzielonych ze względu na wersje testu zawiera tabela 3. W arkuszach z 2013 r. zidentyfikowano zadania, które w dużym stopniu różnią się poziomem wykonania między wersjami. Stosując kryterium przyjęte dla arkuszy z 2012 r. (różnica między wersjami na poziomie 0,04 i więcej), do dokładniejszej analizy wskazano osiem zadań lub czynności. Dla czynności 18_2 różnica ta wyniosła aż 13%, natomiast dla czynności 11_1 i 11_3 było to 12%. Są to stosunkowo duże wartości.

Tabela 3. Porównanie poziomów wykonania zadań z historii i WOS z 2013 r.

	Poziom wykonania zadania (w nawiasie prawidłowa odpowiedź)		Różnica w poziomach wykonania zadania między wersjami A i B arkusza
	wersja A	wersja B	
z1	0,71 (B)	0,72 (B)	0,00
z2	0,56 (A)	0,54 (A)	0,01
z3	0,26 (B)	0,26 (B)	0,00
z4	0,83 (D)	0,83 (D)	0,00
z5	0,90 (B)	0,91 (C)	0,01
z6	0,33 (D)	0,33 (D)	0,00
z7	0,63 (A)	0,61 (B)	0,01
z8	0,71 (C)	0,72 (C)	0,01
z9	0,66 (C)	0,66 (C)	0,01
z10	0,49 (D)	0,49 (D)	0,01
z11_1	0,55 (C)	0,43 (C)	0,12
z11_2	0,91 (B)	0,91 (C)	0,00
z11_3	0,66 (C)	0,54 (C)	0,12
z12	0,72 (B)	0,71 (C)	0,01
z13	0,86 (D)	0,87 (A)	0,01
z14	0,62 (B)	0,64 (B)	0,03
z15	0,49 (C)	0,53 (B)	0,03
z16	0,22 (C)	0,22 (C)	0,01

z17	0,49 (B)	0,48 (C)	0,01
z18_1	0,98 (A)	0,98 (A)	0,00
z18_2	0,43 (A)	0,57 (B)	0,13
z18_3	0,64 (A)	0,70 (C)	0,06
z19_1	0,45 (A)	0,44 (A)	0,01
z19_2	0,51 (C)	0,50 (C)	0,01
z19_3	0,32 (B)	0,36 (D)	0,04
z20_1	0,51 (B)	0,51 (A)	0,00
z20_2	0,31 (C)	0,31 (C)	0,00
z21	0,63 (C)	0,68 (B)	0,04
z22	0,52 (C)	0,53 (C)	0,01
z23	0,25 (D)	0,33 (A)	0,08
z24_1	0,94 (A)	0,91 (B)	0,02
z24_2	0,45 (C)	0,39 (A)	0,06
z24_3	0,93 (B)	0,94 (C)	0,01

Wyróżnione zadania z tabeli 3 zostały sprawdzone testem M-H. Wyniki analiz znajdują się w tabeli 4. Podobnie jak w przypadku wyników z 2012 r., można zauważyć, że większość pozycji testowych uznanych za problematyczne w analizie różnic między poziomem wykonania zadania między wersjami A i B arkusza wykazuje duże obciążenie w teście M-H. Statystyka wielkości efektu cOR wskazuje, że pozycji obciążonych w 2013 r. jest mniej. Natomiast należy mieć na uwadze, że jest więcej problematycznych pozycji testowych w teście z 2013 r. o wysokich efektach (wyższe obciążenie DIF) niż w 2012 r. Dotyczy to zwłaszcza pozycji z11_1, z11_3, z18_2 (w 2012 r. do podobnych wielkości efektów zbliżyła się tylko jedna pozycja testowa z_24_1). Również wielkości efektów opartych o MH P-DIF oraz STD P-DIF potwierdzają te wnioski.

Podobnie jak w 2012 r., nie można wskazać, iż w jednej z wersji arkuszy jest więcej zadań dających szansę na uzyskanie lepszych wyników. Cztery pozycje testowe (z11_1, z_11_3, z24_1, z24_2) wypadają lepiej w wersji A arkusza, pozostałe pozycje testowe (z18_2, z18_3, z23) w wersji B. Należy tutaj wskazać na relatywnie duże efekty w przypadku zadań z 2013 r. Ten wątek zostanie rozwinięty w końcowej części artykułu, gdzie podczas dyskusji wyników zostaną podsumowane analizy arkuszy z lat 2012 i 2013.

Tabela 4. Wyniki testu MH oraz wielkości efektów obciążenia pozycji testowych z arkusza z historii i WOS za 2013 r.

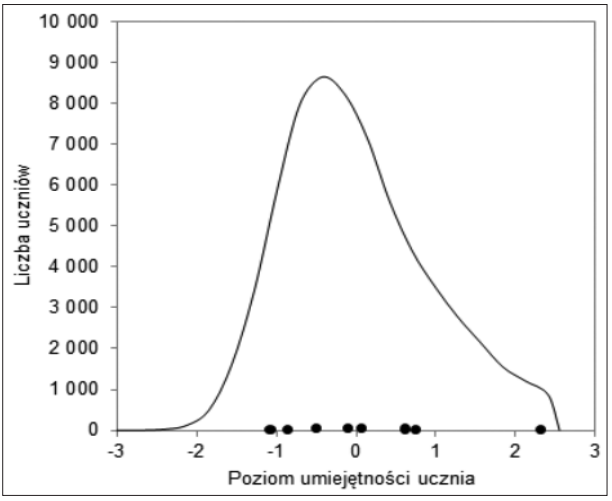
	chi-kwadrat MH	istotność	cOR	ln(cOR)	MH P-DIF	STD P-DIF
z11_1	1390,44	0,000	0,55	-0,59	0,15	0,12
z11_3	1529,21	0,000	0,55	-0,60	0,14	0,13
z18_2	1622,67	0,000	1,83	0,60	-0,15	-0,13
z18_3	330,42	0,000	1,33	0,29	-0,06	-0,06
z23	672,47	0,000	1,52	0,42	-0,09	-0,08
z24_1	229,11	0,000	0,65	-0,43	0,03	0,03
z24_2	377,38	0,000	0,75	-0,29	0,07	0,06

***Ograniczono się do pokazania w tabeli jedynie tych pozycji, które mają największą wielkość efektu (ln(cOR) na poziomie 0,29 bądź więcej.**

Ocena skali konsekwencji różnic między formami testu

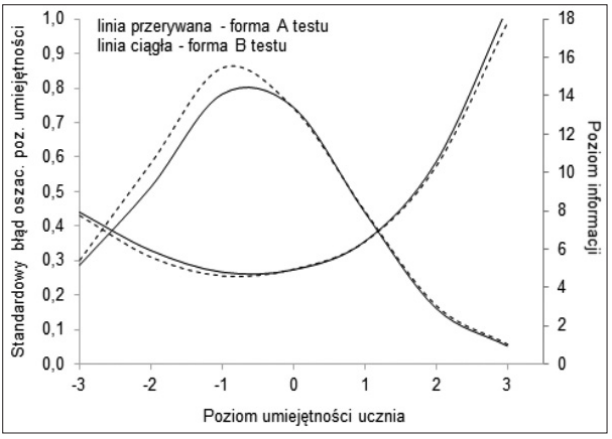
Porównanie proporcji prawidłowych odpowiedzi w grupach rozwiązujących różne arkusze testu oraz analiza testem M-H dzielą ze sobą ograniczenia wszystkich analiz wykonywanych w paradygmacie KTT. KTT rodzi co najmniej cztery główne problemy (Hambleton i in., 1991: 2-5). Po pierwsze, interpretacja wyników uczniów nie może wybiegać w KTT poza ramy konkretnego testu. Po drugie, właściwości pozycji testowych są w KTT zależne od próby. Oznacza to, że właściwości pozycji testowych będą inne, jeśli test będzie rozwiązywała inna grupa uczniów. Po trzecie, w KTT nie jest możliwe szacowanie indywidualnego prawdopodobieństwa prawidłowej odpowiedzi na daną pozycję testową. Dysponuje się jedynie proporcją prawidłowych odpowiedzi w grupie uczniów rozwiązujących test. Po czwarte, KTT nie daje możliwości oszacowania indywidualnych błędów pomiarowych. Zakłada natomiast równy błąd pomiarowy między wszystkimi uczniami rozwiązującymi test. Założenie to oczywiście nie jest do utrzymania. Oprócz systematycznych błędów pomiarowych, które w równym stopniu oddziałują na uczniów, istnieją jeszcze indywidualne źródła błędu, np. dyspozycja w dniu testu, stres itp. IRT przezwycięża problemy KTT. IRT daje możliwość wyrażenia na wspólnej skali zarówno umiejętności uczniów, jak i poziomu trudności pytań testowych. Zarówno więc oszacowania poziomu wiedzy uczniów są niezależne od testu, jak i właściwości psychometryczne pozycji testowych są niezależne od próby. Prawdopodobieństwo prawidłowej odpowiedzi na dane pytanie, jak i błąd pomiarowy szacowane są indywidualnie dla każdego ucznia.

Prezentowane poniżej analizy zostały zawężone do danych za 2013 r. Ocena skali konsekwencji różnic między formami testu została przeprowadzona głównie w ramach IRT. Odpowiedzi uczniów na pytania z części humanistycznej egzaminu w 2013 r. zostały wyskalowane dwuparametrycznym modelem logistycznym (2PLM) (Birnbaum, 1968). Wykres 1 to histogram poziomu umiejętności uczniów. Na osi *x*, która reprezentuje oprócz poziomu wiedzy uczniów także poziom trudności pytań testowych, zostały zaznaczone problematyczne pytania, tj. z11_1, z11_3, z18_2, z18_3, z19_3, z21, z23, z24_2. Absolutna różnica dla tych pytań między formą A i B testu w odsetku uczniów, którzy udzielili poprawnej odpowiedzi, wynosi nie mniej niż 4 p.p. Problematyczne pytania obciążają wyniki uczniów o średnich umiejętnościach, przez co dla tych uczniów mogą rodzić negatywne konsekwencje, gdy wyniki egzaminu brane są pod uwagę w rekrutacji do szkół średnich. Wykres pokazuje także, jak silny wpływ na wyniki może mieć obciążone pytanie w obszarach rozkładu umiejętności uczniów, które nie są dobrze pokryte przez test, tzn. obarczone dużym błędem oszacowania indywidualnych poziomów umiejętności. Pytanie z23 w znacznym stopniu zmienia kształt krzywej rozkładu umiejętności uczniów.



Wykres 1. Rozkład umiejętności uczniów w odniesieniu do obciążonych pytań z arkusza z historii i WOS za 2013 r.

Odpowiedzi uczniów na pytania z części humanistycznej egzaminu w 2013 r. zostały powtórnie wyskalowane w modelu 2PLM, jednak osobno dla subpopulacji rozwiązującej formę A testu i subpopulacji rozwiązującej formę B testu. Wykres 2 obrazuje różnice w krzywych informacyjnych i poziomie błędów dla formy A i formy B testu. Przy założeniu, że każdy uczeń miał równe prawdopodobieństwo otrzymania formy A lub B, konsekwencje, jakie rodzą problematyczne pytania dla właściwości testu, są znaczne.



Wykres 2. Krzywe informacyjne i błędów pomiaru dla historii i WOS za 2013 r.

Tabela 5 prezentuje ranking pytań w arkuszu z historii i WOS za 2013 r. Różne wersje rankingu sporządzono w oparciu o wskaźnik wykonalności oraz o parametr trudności oszacowany w modelu 2PLM. Bezwzględna różnica w rankingu trudności na podstawie procent prawidłowych odpowiedzi jest największa

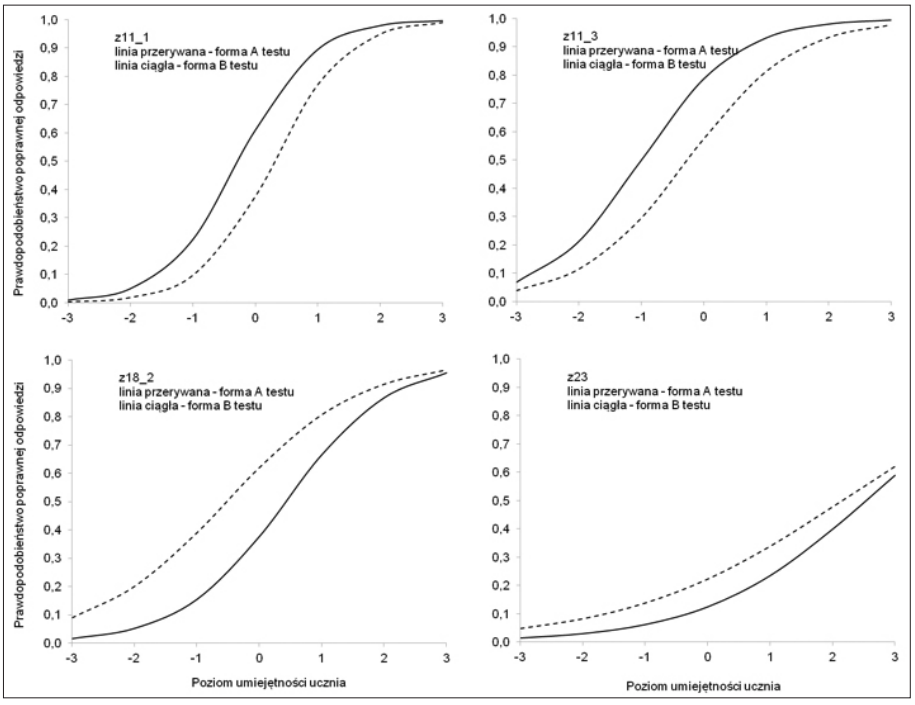
dla pytań z11_1 (8), z11_3 (6), z18_2 (11). Bezwzględna różnica w rankingu trudności na podstawie parametru trudności jest największa dla pytań z11_1 (8), z11_3 (7), z18_2 (12), z18_3 (4), z21 (5), z24_2 (4). Bezwzględna różnica trudności pytań wyrażonej w parametrze trudności jest największa dla z11_1 (0,559), z11_3 (0,74) i z18_2 (0,94).

Tabela 5. Ranking pytań w arkuszu z historii i WOS za 2013 r.

	Wersja A % prawd. odp.	Wersja A par. b	Wersja B % prawd. odp.	Wersja A par. b
z1	9	8	9	10
z2	17	17	17	18
z3	31	31	32	31
z4	7	6	7	6
z5	5	5	4	5
z6	28	29	30	29
z7	15	15	15	16
z8	10	11	8	12
z9	11	13	13	13
z10	23	22	23	23
z11_1	18	18	26	26
z11_2	4	4	5	4
z11_3	12	10	18	17
z12	8	9	10	11
z13	6	7	6	7
z14	16	16	14	14
z15	22	23	20	20
z16	33	32	33	33
z17	24	24	24	24
z18_1	1	2	1	2
z18_2	27	27	16	15
z18_3	13	12	11	8
z19_1	26	25	25	25
z19_2	20	20	22	22
z19_3	29	28	28	27
z20_1	21	21	21	21
z20_2	30	30	31	28
z21	14	14	12	9
z22	19	19	19	19
z23	32	33	29	32
z24_1	2	3	3	3
z24_2	25	26	27	30
z24_3	3	1	2	1

Ze względu na zdiagnozowany DIF rankingi pytań ze względu na trudność między wersjami A i B różnią się. Różnice jednak nie są duże. Rho Spearmana rankingów między wersjami A i B opartych o poziom realizacji zadań wynosi 0,954. Dla rankingów między wersjami A i B opartych o parametr trudności wynosi 0,942.

Różnice między problematycznymi pytaniami są bardzo dobrze widoczne na wykresach porównujących krzywe informacyjne pytań (wykres 3). Jeśli pytania byłby takie same w wersji A i B arkusza, to krzywe powinny na siebie nachodzić.



Wykres 3. Krzywe informacyjne wybranych pytań o wysokich DIF w wersji A i B arkusza z historii i WOS za 2013 r.

W niniejszym opracowaniu wykorzystano proste porównania parametrów dwuparametrycznego modelu IRT. Metod diagnozy DIF w ramach IRT jest więcej, ale zawarcie tych analiz przekracza możliwości niniejszego opracowania.

Dyskusja wyników

Analiza przedstawiona w tym artykule wykazała istnienie zróżnicowania funkcjonowania zadań (pozycji testowych) w arkuszach egzaminacyjnych z historii i WOS, używanych podczas egzaminu gimnazjalnego w 2012 i 2013 r. Pytania między arkuszami różnią się tylko co do kolejności możliwych odpowiedzi. W tabeli 5 znajduje się podsumowanie pozycji testowych z największym zróżnicowaniem funkcjonowania między wersjami.

Tabela 6. Typy zadań o największym zróżnicowaniu funkcjonowaniu zróżnicowania na w latach 2012 i 2013

Rok	Zadanie (pozycja testowa)	Typ zadania	ln(cOR)	MH P-DIF
2012	z4_2	Zadanie na dobieranie	-0,34	0,09
	z5	Zadanie wielokrotnego wyboru	-0,34	0,08
	z12_3	Zadanie z luką	0,29	-0,07
	z21	Zadanie prawda/fałsz	-0,33	0,07
	z24_1	Zadanie na dobieranie	0,48	-0,09
	z24_3	Zadanie na dobieranie	0,29	-0,07
2013	z11_1	Zadanie z luką	-0,59	0,15
	z11_3	Zadanie z luką	-0,60	0,14
	z18_2	Zadanie z luką	0,60	-0,15
	z18_3	Zadanie z luką	0,29	-0,06
	z23	Zadanie wielokrotnego wyboru	0,42	-0,09
	z24_1	Zadanie na dobieranie	-0,43	0,03
	z24_2	Zadanie na dobieranie	-0,29	0,07

Wśród obciążonych pytań są pytania zamknięte różnego rodzaju, tj. zadanie na dobieranie, wielokrotnego wyboru, zadanie z luką itp. Nie zdiagnozowano, że to forma pytania odpowiada za różne funkcjonowanie pytań między wersjami arkusza. W arkuszach egzaminacyjnych w 2012 i 2013 r. są inne zadania, o takiej samej konstrukcji, które nie powodują obciążenia.

Wydaje się jednak, że zadania, które konstrukcyjnie są skomplikowane, odwołują się do jednego zagadnienia w taki sposób, że odpowiedź na pytanie w pierwszej pozycji testowej danego zadania zmienia sposób odpowiedzi na pozostałe pozycje w tym zadaniu, mogą zwiększać szansę na to, że to zadanie będzie funkcjonowało w sposób zróżnicowany między wersjami arkusza.

Widać to zwłaszcza w przypadku zadań na dobieranie, które mają sprawdzać wiedzę uczniów z wiedzy o społeczeństwie. Są to zadania z24 zarówno w arkuszu z 2012, jak i 2013 r. Analiza DIF wykazała, że mają one zróżnicowane funkcjonowanie między wersjami arkusza. Cechą charakterystyczną tych zadań jest zmiana kolejności nie odpowiedzi, a pytań. Możemy domniemywać, że jest to jeden z powodów, który prowadzi do odmiennego funkcjonowania tych zadań. Tego rodzaju hipotezę należy jednak potwierdzić w dodatkowych badaniach.

Dane z 2013 r. dostarczają nam informacji, możliwych do postawienia nieco odmiennej hipotezy. Przypomnijmy w tym miejscu, że pozycje testowe o największym zróżnicowaniu należą do zadania z11 i z18. Przyglądając się prawidłowym odpowiedziom w tych zadaniach, możemy dostrzec interesujący wzór (tabela 6).

Tabela 7. Zadania o największym zróżnicowaniu funkcjonowania ze względu na treść w 2013 roku

	Poziom wykonania zadania (w nawiasie poprawna odpowiedź)	
	wersja A	wersja B
z11_1	0,55 (C)	0,43 (C)
z11_2	0,91 (B)	0,91 (C)
z11_3	0,66 (C)	0,54 (C)
z18_1	0,98 (A)	0,98 (A)
z18_2	0,43 (A)	0,57 (B)
z18_3	0,64 (A)	0,70 (C)

W przypadku zadania z11 prawidłowy wzór odpowiedzi w przypadku wersji B to trzy odpowiedzi C pod rząd. Natomiast w zadaniu z18 w arkuszu A wzór poprawnych odpowiedzi to trzy A pod rząd. W każdym przypadku, gdy mamy do czynienia z taką samą kolejnością prawidłowych odpowiedzi, uczniowie piszący tę wersję arkusza mają niższe wyniki w niektórych czynnościach składających się na dane zadanie niż uczniowie piszący drugą wersję arkusza. Możemy zatem postawić hipotezę, że tego rodzaju układ prawidłowych odpowiedzi w zadaniu może być jednym z powodów zwiększających zróżnicowanie funkcjonowania zadania między tymi dwoma grupami uczniów. Jest prawdopodobne, że uczniowie którzy zaznaczyli poprawnie np. A, A i A w trzech kolejnych pozycjach, zaczynają się wahać, czy taki wzorzec odpowiedzi nie jest błędny. Wydaje się bowiem uczniom mało prawdopodobny. Uczniowie ze względu na te obawy mogą zmieniać odpowiedzi, wybierając te złe. Ten efekt możemy nazwać „antywzorcem". Możliwe, że uczniowie spodziewają się, że prawidłowe odpowiedzi nie będą oznaczone tą samą literą w jednym zadaniu. Hipoteza ta wymaga potwierdzenia w dodatkowych badaniach, jednak na jej poprawność może wskazywać fakt relatywnie najsilniejszych efektów DIF w przypadku pozycji testowych w tych dwóch zadaniach oraz brak tego rodzaju wzorców poprawnych odpowiedzi w innych zadaniach. Jeśli okazałaby się ona prawdziwa, stanowiłaby ważną wskazówkę dla osób tworzących testy egzaminacyjne.

Podsumowanie

Przedstawione w artykule analizy wskazują na istnienie zróżnicowania funkcjonowania zadań i pozycji testowych w zależności od wersji arkusza z historii i WOS, używanych podczas egzaminu gimnazjalnego w 2012 i 2013 r. Szereg hipotez, które mogłyby tłumaczyć to zróżnicowanie, należy zweryfikować w dalszych badaniach. Należy przeprowadzić dodatkowe analizy, które będą mogły dostarczyć więcej informacji w zakresie tego tematu. Temat jest istotny także ze względu na fakt używania dwóch wersji arkuszy standardowych także na innych egzaminach zewnętrznych. Z tego względu rzetelne informacje dotyczące cech zadań, które mają zróżnicowane funkcjonowanie w zależności od wersji arkusza, mogą pomóc w doskonaleniu testów egzaminacyjnych.

Podobna liczba obciążonych pytań faworyzuje uczniów rozwiązujących wersję A, jak również wersję B arkusza, zarówno w 2012, jak i 2013 r. Dodatkowo trudność obciążonych pytań lokuje się wokół średnich wartości umiejętności

uczniów, gdzie test ma największą moc informacyjną. Obciążenie wynikające z pytań o dużym DIF zostało więc niejako „rozmyte". Różnice między średnimi wynikami w arkuszach w wersji A i B zarówno raportowanymi jako suma punktów, jak i oszacowane z modelu 2PLM, są nieistotne statystycznie dla testu w 2012 i 2013 r. Wydaje się jednak, że fakt występowania podobnej liczby obciążonych pytań między wersjami arkusza jest dziełem przypadku. Jeśli wszystkie lub przynajmniej większość obciążonych pytań faworyzowałaby formę A lub formę B to wyniki uczniów byłyby niesprawiedliwe. Należy więc z większą ostrożnością weryfikować wersje arkuszy, tak aby upewnić się, że są one faktycznie wersjami równoległymi.

Bibliografia

1. Birnbaum, A. (1968), *Some latent trait models and their use in inferring an examinees ability*, [w:] Lord, F. M. i Novick, M. R. (red.) *Statistical theories of mental test scores*, Addison-Wesley: 395-479.
2. Cohen, J. (1988), *Statistical power analysis for the behavior sciences*. Hillsdale, NJ: Erlbaum.
3. Grudniewska, M., Kondratek, B. (2012), *Zróżnicowane funkcjonowanie zadań w egzaminach zewnętrznych w zależności od płci na przykładzie części matematyczno-przyrodniczej egzaminu gimnazjalnego,* [w:] B. Niemierko i M.K. Szmigel (red.) *Regionalne i lokalne diagnozy edukacyjne*, Kraków: gRUPA TOMAMI, PTDE.
4. Kondratek, B., Grudniewska, M. (2013), *Test Mantel-Haenshel oraz modelowanie IRT jako narzędzia służące do wykrywania DIF i opisu jego wielkości na przykładzie zadań ocenianych dychotomicznie*, Edukacja 2013 2(122), s.34-55.
5. Hambleton, R. K., Swaminathan, H. i Rogers, H. J. (1991), *Fundamentals of Item Response Theory*, Sage Publications: Newbury Park London New Delhi.
6. Holland, P. W., Thayer, D. T. (1988), *Differential item performance and the Mantel-Haenszel procedure.* [w:] Wainer, H. i Braun, H. I. (red.) *Test validity*. Hillsdale, NJ: Erlbaum, 129-145.
7. Holland, P. W., Weiner, H. (1993), *Differential item functioning*. Hillsdale, NJ: Erlbaum.
8. Hornowska, E. (1999). *Stronniczość testów psychologicznych: problemy – kierunki – kontrowersje*, Poznań: Wydawnictwo Fundacji Humaniora.
9. *Ironson*, G. H. (*1982), Use of chi-square and latent trait approaches for detecting item bias,* [w:] R. Berk (red.). *Handbook of methods for detecting test bias.* Baltimore: Johns Hopkins University Press, 117-155.
10. Kondratek, B., Grudniewska, M. (2013), *Test Mantel-Haenshel oraz modelowanie IRT jako narzędzia służące do wykrywania DIF oraz opisu jego wielkości na przykładzie zadań ocenianych dychotomicznie*, Edukacja, 2.
11. Linn, R. L., Levine, M. V., Hastings, C. N. i Wardrop, J. L. (1981), *Item bias in a test of Reading comprehension*, Applied Psychological Measurement, 5: 159-173.
12. Mantel, N., Haenszel. W. (1959), *Statistical aspects of the analysis of data from retrospective studies of disease*, Journal of the National Cancer Institute, 22: 719-748.
13. Niemierko, B. (2006), *Oszustwo egzaminacyjne*, [w:] *O wyższą jakość egzaminów szkolnych*, Kraków: gRUPA TOMAMI, PTDE.
14. Szaleniec, H. (2006), *Oszukiwanie na egzaminie istotnym źródłem majowej porażki*. [w:] *O wyższą jakość egzaminów szkolnych.* Kraków: gRUPA TOMAMI, PTDE.