Edukacja 2014, 1(126), 79–94
ISSN 0239-6858

Zróżnicowane funkcjonowanie zadań testowych ze względu na wersję testu

Maciej Koniewski, Przemysław Majkut, Paulina Skórska

Instytut Badań Edukacyjnych*

Analizowano zróżnicowane funkcjonowanie zadań (DIF) ze względu na wersję arkusza testowego. Zadania wielokrotnego wyboru różniły się między arkuszami kolejnością odpowiedzi. Analizowano zadania z wersji A i B arkusza egzaminu gimnazjalnego z historii i wiedzy o społeczeństwie z 2013 r. Dane pochodziły od uczniów z województw lubelskiego, małopolskiego i podkarpackiego (n = 81 545). W celu detekcji DIF wykorzystano test Mantela–Haenshela, regresję logistyczną oraz standaryzację. Przedstawiono graficzne metody prezentacji DIF. Wyniki analiz wskazują na istotne różnice w funkcjonowaniu zadań między wersjami A i B testu w sytuacji, gdy w jednej wiązce zadań prawidłowa odpowiedź jest oznaczona zawsze tym samym symbolem, np. A, A, A. Taki wzór odpowiedzi nazwano antywzorcem, ponieważ może być uznawany przez uczniów za mało prawdopodobny i w konsekwencji prowadzić do udzielania błędnych odpowiedzi. Sformułowano rekomendacje ważne dla twórców testów.

Słowa kluczowe: psychometria, zróżnicowane funkcjonowanie zadań testowych, DIF, standaryzacja, test Mantela–Haenszela, regresja logistyczna.

Wnioskowanie na temat umiejętności ucznia na podstawie osiąganych wyników w teście wymaga upewnienia się, że trafnie odzwierciedlają one poziom jego wiedzy. Zadania testowe spełniają ten warunek, jeżeli prawdopodobieństwo poprawnej odpowiedzi zależy wyłącznie od poziomu wiedzy ucznia. Niekiedy prawdopodobieństwo udzielenia poprawnej odpowiedzi zależy od innych cech uczniów, takich jak płeć, pochodzenie etniczne, pochodzenie społeczne, lub od cech samego testu, np. wersji rozwiązywanego arkusza (Ironson, 1982; Linn, Levine, Hastings i Wardrop, 1981). W takich przypadkach zadanie testowe, a w konsekwencji cały test, nie jest sprawiedliwym i trafnym narzędziem pomiaru wiedzy uczniów. Ma to szczególne znaczenie w sytuacjach, gdy wyniki testu mają duże konsekwencje dla egzaminowanych. Jednym z takich testów wysokiej stawki[1] w polskim systemie oświaty jest egzamin gimnazjalny, który stanowi próg selekcyjny do szkół średnich.

Artykuł jest rozbudowaną wersją wystąpienia pt. „Wpływ wersji arkusza egzaminacyjnego na zróżnicowane funkcjonowanie zadań na przykładzie egzaminu gimnazjalnego", wygłoszonego przez autorów podczas XIX Konferencji Diagnostyki Edukacyjnej w Gnieźnie, 26–28 września 2013 r. Wystąpienie ukazało się drukiem w publikacji pokonferencyjnej: Niemierko, B. i Szmigiel, M. K. (red.). (2013). *Polska edukacja w świetle diagnoz prowadzonych z różnych perspektyw badawczych* (s. 212–224). Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

* Adres do korespondencji: Maciej Koniewski, Pracownia Edukacyjnej Wartości Dodanej, Instytut Badań Edukacyjnych, ul. Górczewska 8, 01-180 Warszawa. E-mail: m.koniewski@ibe.edu.pl

[1] Ang. *high-stakes testing*, w polskiej terminologii specjalistycznej nazywany również egzaminem doniosłym.

Sposobem identyfikacji zadań, które mogą dyskryminować niektóre grupy uczniów, jest analiza zróżnicowanego funkcjonowania zadań testowych (*differential item functioning,* DIF). Z DIF mamy do czynienia, gdy na prawdopodobieństwo udzielenia prawidłowej odpowiedzi na zadanie testowe (Y_i) wpływa nie tylko poziom umiejętności ucznia (θ), ale także to, do jakiej grupy (zdefiniowanej przez określoną cechę) należy uczeń (G_j). Formalną definicję można sformułować jako:

$$P(Y_i|\theta_i,G_j);\ \text{gdzie}\ i = 1, 2, 3, ..., n;\ j = 1, 2,$$

gdy porównujemy dwie grupy. Zróżnicowane funkcjonowanie zadań testowych może występować w dwóch formach: jednorodnej (*uniform* DIF) i niejednorodnej (*non-uniform* DIF). W pierwszym przypadku różnica w funkcjonowaniu zadań między dwoma analizowanymi grupami jest taka sama dla grup uczniów posiadających różny poziom umiejętności. Natomiast z przypadkiem niejednorodnego DIF mamy do czynienia w sytuacji, gdy różne funkcjonowanie zadań między grupami zmienia się w zależności od poziomu umiejętności uczniów.

Zagadnienie DIF jest rzadko przedstawiane w polskiej literaturze naukowej. Wyjątkiem są artykuły Bartosza Kondratka i Magdaleny Grudniewskiej (Grudniewska i Kondratek, 2012; Kondratek i Grudniewska, 2013), które pokazują wyniki analiz DIF ze względu na płeć uczniów. Celem tego artykułu jest zaprezentowanie wyników analiz DIF ze względu na rozwiązywaną przez ucznia wersję arkusza testowego. W polskim systemie egzaminów zewnętrznych stosuje się dwie wersje arkusza testów standaryzowanych, które różnią się kolejnością odpowiedzi w zadaniach wielokrotnego wyboru. Zadanie wielokrotnego wyboru jest zadaniem zamkniętym z jedną prawidłową odpowiedzią i kilkoma dystraktorami. Zamiana kolejności odpowiedzi między arkuszami służy zapobieganiu zjawisku „ściągania" (Szaleniec, 2006).

Wyjściowa hipoteza prezentowanych tu analiz zakłada, że różna kolejność odpowiedzi w różnych wersjach arkusza powoduje różne funkcjonowanie zadań testowych powiązanych w wiązki. Wiązkę tworzy treść trzonu wraz z odnoszącymi się do niego zadaniami. Do każdego zadania proponowanych jest kilka wariantów odpowiedzi, które prezentowane są w arkuszu testowym, jednen po drugim. Losowa manipulacja kolejnością tych wariantów może powodować sytuację, w której klucz prawidłowych odpowiedzi na trzy kolejne zadania w wiązce z jedną prawidłową odpowiedzią i dwoma dystraktorami tworzy wzór, czyli układ symboli określających prawidłowe odpowiedzi w ramach danej wiązki. W artykule stawiamy hipotezę, że niektórzy uczniowie, a zwłaszcza ci o przeciętnych umiejętnościach, wzór odpowiedzi, np. A, A, A lub B, B, B, lub C, C, C mogą uważać za „podejrzany" i mało prawdopodobny. Zwłaszcza wizualna bliskość na karcie udzielonych odpowiedzi na kolejne zadania z wiązki może budzić wątpliwości dotyczące poprawności wzorca. Wątpliwości te mogą sprawiać, że uczniowie celowo będą zmieniać jedną z odpowiedzi, aby otrzymany wzór bardziej „uprawdopodobnić". Hipotezę tę nazywamy w artykule hipotezą antywzorca, ponieważ prawidłowy (ale specyficzny) wzór odpowiedzi w ramach jednej wiązki zadań może prowadzić uczniów do popełniania błędów. W Tabelach 1 i 2 znajdują się dwa przykłady zadań powiązanych w wiązki, które pojawiły się w arkuszach części humanistycznej egzaminu gimnazjalnego z historii i wiedzy o społeczeństwie (WOS), zdawanego przez uczniów w 2013 r. Przykłady te będą szczegółowo

Tabela 1
Treść zadania 11 z arkusza historia i WOS egzaminu gimnazjalnego z 2013 r. i układ odpowiedzi w wersjach A i B arkusza

Tekst do zadania 11

My rady koronne, duchowne i świeckie i rycerstwo wszystko obiecujemy zjechać się i wspólnie akt elekcji podług woli Bożej do skutku słusznego przywieść. A i w Rzeczypospolitej naszej jest niezgoda niema a w sprawie religii chrześcijańskiej, zapobiegając temu, aby się z tej przyczyny między ludźmi rozterka jaka szkodliwa nie wszczęła, którą po innych królestwach jaśnie widzimy, obiecujemy to sobie wspólnie, za nas i potomków naszych, i którzy jesteśmy równi w wierze, pokój między sobą zachować, a dla różnej wiary i odmiany w kościołach krwi nie przelewać.

Na podstawie: *Historia. Teksty rodowe,* pod red. S. Sierpowskiego, Warszawa 1998.

Zadanie 11 (0–3 pkt.)

Uzupełnij poniższy tekst, przyporządkowując w każdym zdaniu właściwą odpowiedź spośród oznaczonych literami A–C.

Przedstawione w tekście porozumienie zawarto w 11.1. ____. W cytowanym fragmencie dokumentu szlachta gwarantowała sobie 11.2. ____. Wspomniana w dokumencie forma wyboru króla została wprowadzona w dobie 11.3. ____.

Układ odpowiedzi na zadanie 11 w wersji A arkusza:

11.1. A. XIV w.	B. XV w.	C. XVI w.
11.2. A. przywileje ekonomiczne	B. tolerancję religijną	C. wzajemną pomoc
11.3. A. monarchii patrymonialnej	B. rządów absolutnych	C. demokracji szlacheckiej

Układ odpowiedzi na zadanie 11 w wersji B arkusza:

11.1. A. XIV w.	B. XV w.	C. XVI w.
11.2. A. przywileje ekonomiczne	B. wzajemną pomoc	C. tolerancję religijną
11.3. A. monarchii patrymonialnej	B. rządów absolutnych	C. demokracji szlacheckiej

analizowane pod kątem DIF w dalszej części artykułu.

Prawidłowe odpowiedzi zaznaczono podkreśleniem. Wiązka zadań 11 w wersji B arkusza różni się tylko kolejnością odpowiedzi w zadaniu z11.2. Zgodnie z hipotezą antywzorca wiązka zadań 11 faworyzuje uczniów rozwiązujących wersję A arkusza, gdzie nie występuje antywzorzec.

Wiązka zadań 18 w wersji B arkusza różni się tylko kolejnością odpowiedzi w zadaniach z18.2 i z18.3. Hipoteza antywzorca sugeruje, że wiązka zadań 18 faworyzuje uczniów rozwiązujących wersję B arkusza, gdzie nie występuje antywzorzec.

Metodyka

Zgodnie z przyjętą metodyką postępowania, najpierw obliczone zostały proporcje prawidłowych odpowiedzi (*p*) na zadania w analizowanej części egzaminu w podziale na wersje arkusza. Tabela 3 prezentuje wartości *p* wybranych zadań testowych, w tym wszystkie, które zidentyfikowano jako potencjalnie obciążone przez zróżnicowane funkcjonowanie między arkuszami. Zadania te zaznaczono gwiazdką. Wartości *p* obliczono na podstawie wyników uczniów gimnazjów w województwach: lubelskim, małopolskim i podkarpackim. Wszystkie analizy w tym artykule prowadzono na tak

Tabela 2
Treść zadania 18 z arkusza historia i WOS egzaminu gimnazjalnego z 2013 r. i układ odpowiedzi w wersjach A i B arkusza

Zadanie 18 (0–3 pkt.)

Uzupełnij poniższy tekst, przyporządkowując do każdego zadania właściwą odpowiedź spośród oznaczonych literami A–C.

Zastosowanie maszyny parowej w XIX w. doprowadziło w Europie Zachodniej do gwałtownych przemian gospodarczych i społecznych nazywanych rewolucją 18.1. _____. Jednym ze skutków tych przemian był wyraźny spadek zatrudnienia w 18.2. _____. „Warsztatem świata", czyli krajem wówczas przodującym w zakresie uprzemysłowienia, nazywano 18.3. _____.

Układ odpowiedzi na zadanie 18 w wersji A arkusza:

18.1.	A. przemysłową	B. kulturalną	C. informatyczną
18.2.	A. rolnictwie	B. przemyśle	C. handlu
18.3.	A. Wielką Brytanię	B. Francję	C. Hiszpanię

Układ odpowiedzi na zadanie 18 w wersji B arkusza:

18.1.	A. przemysłową	B. kulturalną	C. informatyczną
18.2.	A. przemyśle	B. rolnictwie	C. handlu
18.3.	A. Francję	B. Hiszpanię	C. Wielką Brytanię

zdefiniowanej populacji uczniów. Egzamin był prowadzony w formie papierowej (*paper and pencil based test,* PPT). Wszystkie zadania egzaminu gimnazjalnego z historii i WOS były zadaniami wielokrotnego wyboru z jedną prawidłową odpowiedzią. Odpowiedzi były kodowane dychotomicznie: 1 punkt za prawidłową odpowiedź, 0 punktów za błędną lub brak odpowiedzi. Poziom wykonania zadania jest więc równoważny z proporcją jedynek lub średnią odpowiedzi na dane zadanie testowe.

Zadania w wiązkach 11 oraz 18 na tle pozostałych zadań, o stosunkowo dużych różnicach odsetka prawidłowych odpowiedzi między arkuszami, należy uznać za potencjalnie najbardziej obciążone. Specyficzny charakter zadań powiązanych w wiązki, ich układ w arkuszach oraz obecność antywzorca to prawdopodobne przyczyny największych różnic w poziomach wykonania między arkuszami. Należy zauważyć, że najbardziej obciążone w wiązce 11 są zadania z11.1 i z11.3. Różnica w kolejności odpowiedzi między arkuszami wystąpiła jednak tylko w z11.2. Ponieważ z11.2 było zadaniem bardzo prostym (poprawnie odpowiedziało 91% uczniów), to antywzorzec obecny w wiązce mógł skłaniać uczniów do zmiany z dobrej odpowiedzi na złą w z11.1 i z11.3, które były zadaniami trudniejszymi niż z11.2. Podobną sytuację zaobserwować można dla wiązki 18. Kolejność odpowiedzi różni się między arkuszami tylko dla z18.2 i z18.3, które w wiązce są akurat zadaniami trudnymi. Zaobserwowano potencjalnie duże obciążenie tych zadań.

Można więc sformułować kolejną hipotezę, że uczniowie szukają punktu zaczepienia, czyli takiego zadania w wiązce, które jest najprostsze, co do którego są przekonani, że znają prawidłową odpowiedź. Natomiast, gdy zaobserwują antywzorzec, wracają do zadań trudniejszych w wiązce i w nich zmieniają odpowiedzi. Zmiana kolejności wariantów odpowiedzi w danym zadaniu w wiązce, nie musi więc automatycznie

Tabela 3
Porównanie poziomów wykonania wybranych zadań egzaminu gimnazjalnego z arkusza z historii i WOS z 2013 r.

Zadanie	*p* [(a)]								Różnica w *p* między wersjami A i B arkusza[(b)]
	wersja A				wersja B				
	A	B	C	D	A	B	C	D	
z11.1	0,22	0,24	0,55*	-	0,21	0,36	0,43*	-	0,12
z11.2	0,02	0,91*	0,07	-	0,02	0,08	0,91*	-	0,00
z11.3	0,25	0,09	0,66*	-	0,30	0,16	0,54*	-	0,12
z14	0,12	0,62	0,18	0,09	0,09	0,64*	0,16	0,11	-0,03
z18.1	0,98*	0,02	0,00	-	0,98*	0,02	0,00	-	0,00
z18.2	0,43*	0,42	0,15	-	0,31	0,57*	0,12	-	-0,13
z18.3	0,64*	0,30	0,06	-	0,24	0,06	0,70*	-	-0,06
z19.1	0,45*	0,28	0,17	0,10	0,44*	0,12	0,17	0,26	0,01
z19.2	0,16	0,11	0,51*	0,22	0,16	0,24	0,50*	0,10	0,01
z19.3	0,30	0,32*	0,15	0,23	0,31	0,20	0,14	0,36*	-0,04
z21	0,16	0,21	0,63*	-	0,18	0,68*	0,15	-	-0,04
z23	0,26	0,43	0,06	0,25*	0,33*	0,07	0,38	0,21	-0,08
	n = 41 030				*n* = 40 515				

[(a)] Gwiazdką zaznaczono prawidłową odpowiedź.

[(b)] Podane różnice obliczone zostały z danych zaokrąglonych do dwóch miejsc po przecinku.

powodować obciążenia tego pytania. Może natomiast powodować obciążenie innych pytań w wiązce. Niestety, weryfikacja tej hipotezy i rekonstrukcja mechanizmów działania uczniów podczas rozwiązywania testu nie jest możliwa na podstawie dostępnych danych. Byłaby możliwa w kontrolowanym badaniu eksperymentalnym.

Analiza różnic w poziomach wykonania zadań może więc mieć jedynie charakter eksploracyjny. Wartości *p* nie biorą pod uwagę różnic w poziomach umiejętności uczniów. W celu obliczenia różnic w funkcjonowaniu zadań między arkuszami, przy kontroli poziomu umiejętności uczniów, należy zastosować metody wykrywania DIF.

Istnieje wiele takich metod. W najbardziej ogólnym wymiarze możemy podzielić je na dwie grupy (Haladyna, 2004; Wainer, 1993):

1. Oparte na empirycznych wynikach testów, do których należą:
   - test Mantela–Haenszela (MH; Holland i Thayer, 1988; Mantel i Haenszel, 1959);
   - metoda regresji logistycznej (Swaminathan i Rogers, 1990);
   - procedura standaryzacji (Dorans i Kulick, 1986).
2. Oparte na szacowanych wynikach, wykorzystujące modele teorii prawdopodobieństwa odpowiedzi na pozycje testowe *(item response theory*, IRT); należą do nich:
   - współczynnik wiarygodności *(general item response theory likelihood ratio, general IRT-LR*; Thissen, Steinberg i Wainer, 1988);
   - współczynnik wiarygodności oparty na ograniczonej informacji *(limited-information IRT-LR*; Muthén i Lehman, 1985);
   - podejście IRT-D2 oparte na pełnej informacji (*full-information IRT-D2*; Bock, Muraki i Pfeiffenberger, 1988).

Test MH oraz standaryzacja są metodami nieparametrycznymi. Regresja logistyczna oraz metody oparte na modelach IRT są metodami parametrycznymi. Przywołany powyżej podział jest arbitralny, ponieważ regresja logistyczna jest metodą pośrednią między testem MH a modelami 2PL i 3PL IRT. Z kolei standaryzacja może być rozumiana jako empiryczna odmiana metod opartych na porównaniu krzywych charakterystycznych (Wainer, 1993). Podział ten wskazuje jednak na pochodzenie metod z różnych tradycji i teorii pomiaru.

W niniejszym artykule do analiz zróżnicowanego funkcjonowania zadań z11.1, z11.3 oraz z18.2 między arkuszami zastosowano wyłącznie metody oparte na empirycznych wynikach testów. Za pomocą testu MH testowano istotność statystyczną zróżnicowanego funkcjonowania zadań. Regresja logistyczna posłużyła do wykazania, czy zaobserwowany DIF ma charakter jednorodny, czy niejednorodny. Oprócz tego, że standaryzacja jest metodą wykrywania DIF, przede wszystkim stanowi efektywne i intuicyjne narzędzie graficznego raportowania. Ma to szczególne znaczenie, gdy wyniki analiz DIF prezentowane są odbiorcom niezaznajomionym z wiedzą statystyczną, np. decydentom politycznym.

Test MH do wykrywania i standaryzacja do prezentacji są podstawowymi metodami w ocenie zadań testowych pod kątem DIF, stosowanymi przez Educational Testing Service (ETS; Dorans i Holland, 2009). Dodatkowo, w celu sprawdzenia stopnia wpływu antywzorca na prawdopodobieństwo poprawnej odpowiedzi na obciążone zadania, wykonano analizy regresji logistycznej. W analizach korzystano z oprogramowania STATA 12[2] oraz IBM SPSS Statistics 21.

[2] Autorzy składają serdeczne podziękowania Bartoszowi Kondratkowi za udostępnienie do analiz swojego autorskiego dodatku do programu STATA, przeznaczonego do modelowania IRT.

Metody wykrywania DIF oparte na empirycznych wynikach testów

Test MH polega na porównaniu wyników egzaminacyjnych dwóch grup: ogniskowej, która znajduje się w centrum naszego zainteresowania (*focal group*) i odniesienia (*reference group*), np. dziewcząt z chłopcami. Analogicznie można porównać wyniki uczniów rozwiązujących wersję A arkusza z wynikami uczniów rozwiązujących wersję B. Hipoteza zerowa testu MH zakłada, że szanse poprawnej odpowiedzi w określonym zadaniu testowym są równe w obu grupach uczniów: rozwiązujących wersję A i B testu. Sumę punktów uzyskanych przez ucznia na teście wykorzystuje się jako zmienną warstwującą (*matching variable*), aby kontrolować efekt obciążenia w poszczególnych przedziałach poziomu umiejętności ucznia. W praktyce test MH jest najczęściej wykorzystywaną metodą detekcji DIF (Holland i Wainer, 1993).

Alternatywną metodą jest regresja logistyczna (Swaminathan i Rogers, 1990). Model ten umożliwia oszacowanie prawdopodobieństwa udzielenia przez ucznia prawidłowej odpowiedzi na podstawie informacji o wyniku uzyskanym w teście i przynależności ucznia do grupy ogniskowej lub grupy odniesienia. Można powiedzieć, że test MH wprost bazuje na modelu logistycznym, w którym poziom umiejętności ucznia jest traktowany jako zmienna dyskretna i nie wchodzi (zgodnie z założeniem) w interakcję z przynależnością do grupy ogniskowej lub odniesienia. Jednak w przypadku wykrywania DIF wykorzystującej regresję logistyczną, zakładamy że na odpowiedź ucznia mogą wpływać: wynik końcowy ze wszystkich zadań w teście (lub inny wskaźnik umiejętności ucznia), przynależność ucznia do danej grupy oraz interakcja między nimi. Ogólny model logistyczny opisujący tę sytuację został przedstawiony poniżej:

$$\ln\left(\frac{P(u=1)}{P\,(u=0)}\right) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG \quad (1)$$

gdzie:

X – wynik końcowy ucznia w teście;

G – grupa, do której uczeń przynależy (ogniskowa lub odniesienia);

XG – interakcja pomiędzy wynikiem końcowym i przynależnością do grupy.

Występowanie obciążenia zadania testowego sprawdza się, porównując testem ilorazu wiarygodności, czy model z interakcją jest modelem lepiej dopasowanym do danych niż model bez interakcji. Jeśli lepszym modelem jest model bez interakcji, a β_2 istotnie różni się od zera, to mamy do czynienia z jednorodnym DIF. Jeśli lepszym modelem jest model z interakcją, a β_3 istotnie różni się od zera, to mamy do czynienia z niejednorodnym DIF.

Regresja logistyczna i test MH są alternatywnymi sposobami detekcji DIF. Regresja logistyczna ma jednak przewagę nad testem MH. Pozwala na wykrycie obu rodzajów DIF, podczas gdy test MH tylko jednorodnego. Z kolei częściej, niż test MH, regresja logistyczna prowadzi do popełnienia błędu pierwszego rodzaju, tzn. wskazuje na występowanie obciążenia pozycji testowych, które tak na prawdę obciążone nie są (Narayanan i Swaminathan, 1996).

Standaryzacja jest nieparametryczną alternatywą IRT służącą do opisywania funkcjonowania zadania testowego między grupami, wśród uczniów o podobnym poziomie umiejętności (*item-ability regression*; Kingston i Dorans, 1985). Dla zadań kodowanych dychotomicznie standaryzacja definiuje obciążenie DIF jako różnicę w wartościach *p* między analizowanymi grupami (wyodrębnionymi ze względu na jakąś cechę) na wszystkich poziomach umiejętności (Dorans, 1989; Dorans i Holland, 2009).

Graficznej prezentacji różnic w funkcjonowaniu zadania testowego między grupami dokonuje się za pomocą wykresów regresji odpowiedzi na zadanie testowe na skalę umiejętności. Skala umiejętności dzielona jest zwyczajowo na 15 interwałów co 0,4 odchylenia standardowego. Następnie oblicza się wartości *p* dla analizowanego zadania w interwałach, osobno dla dwóch grup i rzutuje na wykresie. Zgodnie z podejściem zaprezentowanym przez Neila Kingstona i Nell Dorans (1985), skalę umiejętności stanowić powinny wartości θ oszacowane w modelu IRT. Skalą umiejętności może być także suma punktów uzyskanych w teście.

Inną popularną metodą graficznej prezentacji DIF jest metoda *delta-plot*, nazywana także *transformed item-difficulty* (TID; Angoff, 1972; Thurstone, 1925). Metoda TID bazuje na wartościach *p* obliczonych dla każdej z analizowanych grup. Przyjęło się, że wartości *p* następnie transformuje się na skalę o średniej 13 i odchyleniu standardowym 4[3]. Następnie na wykresie rozrzutu rzutuje się pary wartości *p* dla wszystkich zadań testu. Każdy punkt na wykresie reprezentuje jedną parę wartości *p* dwóch grup dla danego zadania testowego. W sytuacji braku zadań obciążonych, należy się spodziewać korelacji 0,98 lub wyższej.

Logika metody *delta-plot* odwołuje się do szerokiego spektrum metod diagnostyki w analizie regresji pod kątem występowania przypadków odstających (Pedhazur, 1997). Mogą one z powodzeniem być wykorzystywane do prezentacji graficznej DIF. W prostej regresji wartości *p* dla wszystkich zadań testu w analizowanych grupach uczniów, pytania obciążone można traktować jako przypadki odstające. Ich wskazanie jest możliwe dzięki analizie reszt standaryzowanych (ZRESID), studentyzowanych (SRESID) lub usuniętych studentyzowanych (SDRESID). Do prezentacji DIF można także wykorzystać odległość Cooka

[3] Transformacji dokonuje się wg wzoru: $z_i = \frac{(x_i - \mu)}{0{,}25 \cdot \sigma} + 13$.

(1977; 1979). Umożliwia ona wskazanie wpływowych przypadków odstających ze względu na ich status na zmiennej zależnej lub niezależnej, jak również na obu jednocześnie. Z tego względu jest syntetyczną miarą wpływu poszczególnej obserwacji na linię regresji.

Wyniki

Wyniki testu MH są istotne statystycznie (na poziomie $p < 0{,}05$) dla wszystkich analizowanych zadań testowych. Wskazuje to na obciążenie większości zadań w arkuszach. Jednak istotność statystyczna jest łatwa do uzyskania w przypadku dużych prób. Nie należy poprzestawać na analizie istotności wyników, ale porównać statystyki wielkości efektu (*effect size*, ES; Cohen, 1988).

Nathan Mantel i William Haenszel (1959) definiują wielkość efektu za pomocą tzw. łącznego ilorazu szans (*common odds ratio,* cOR). Formuła obliczania tej wartości znajduje się poniżej.

$$\hat{\alpha}_{MH} = \frac{\sum_i a_i d_i / N_i}{\sum_i b_i c_i / N_i} \qquad (2)$$

gdzie:

a_i – liczba osób z grupy odniesienia, które odpowiedziały poprawnie;

b_i – liczba osób z grupy odniesienia, które odpowiedziały niepoprawnie;

c_i – liczba osób z grupy ogniskowej, które odpowiedziały poprawnie;

d_i – liczba osób z grupy ogniskowej, które odpowiedziały niepoprawnie;

N_i – łączna liczba osób ze wszystkich czterech grupach.

Ponieważ iloraz szans znajduje się w przedziale wartości od 0 do +∞ (1 wskazuje na brak obciążenia zadania testowego, czyli brak DIF), podlega logarytmowaniu, w celu ułatwienia interpretacji (Holland i Wainer, 1993). Rozkład efektu po logarytmowaniu jest symetryczny wokół wartości 0 (0 oznacza brak DIF). Wartości łącznego ilorazu szans i jego zlogarytmowanych wartości zawiera Tabela 4.

W celu ułatwienia interpretacji uzyskanych wielkości skorzystano z przekształcenia ilorazu szans w statystykę Δ_{MH}, zgodnie z poniższym wzorem:

$$\Delta_{MH} = -(2{,}35)\ln(\hat{\alpha}_{MH}) \qquad (3)$$

Rebecca Zwick i Kadriye Ercikan (1989) na bazie tej statystyki zaproponowały klasyfikację efektów, która na podstawie arbitralnych reguł pozwala ocenić uzyskany efekt obciążenia pozycji testowej DIF jako:

- mały, gdy: $|\Delta_{MH}| < 1$ lub test MH jest nieistotny statystycznie;
- średni, gdy: $1 \leq |\Delta_{MH}| < 1{,}5$ i test MH jest istotny statystycznie;
- duży, gdy: $|\Delta_{MH}| \geq 1{,}5$ i test MH jest istotny statystycznie.

W Tabeli 4 pokazano, że uczniowie wypełniający wersję B arkusza mają mniejszą szansę udzielenia poprawnej odpowiedzi na zadania z11.1 oraz z11.3, niż uczniowie

Tabela 4
Wynik analiz testu Mantela–Haenszela

Zadanie	χ^2_{MH}	$\hat{\alpha}_{MH}$	$\ln(\hat{\alpha}_{MH})$	Δ_{MH}	$\lvert\Delta_{MH}\rvert$	Interpretacja wielkości efektu
z11.1	1 390,44**	0,55	-0,59	1,40	1,40	średnia
z11.3	1 529,21**	0,55	-0,60	1,40	1,40	średnia
z18.2	1 622,67**	1,83	0,60	-1,42	1,42	średnia

** Istotne na poziomie $p < 0{,}01$

rozwiązujący arkusz w wersji A. Wielkość obciążenia jest średnia, choć zbliża się do granicy efektów dużych. W zadaniu z18.2 to uczniowie rozwiązujący wersję B arkusza mieli większą szansę na udzielenie poprawnej odpowiedzi i jest to największy efekt obciążenia we wszystkich analizowanych zadaniach testowych.

Wykrywanie DIF za pomocą regresji logistycznej potwierdza wyniki uzyskane za pomocą testu MH. Prawdopodobieństwo uzyskania prawidłowej odpowiedzi na pozycje testowe w zadaniach z11.1, z11.3 i z18.2 zmienia się pod wpływem wersji arkusza, którą wypełniał uczeń.

W Tabeli 5 zaprezentowano wyniki porównania wartości logarytmu wiarygodności (*log likelihood*) dla trzech modeli regresji logistycznej (różniących się liczbą predyktorów). Dzięki nim można testować występowanie oraz rodzaj DIF (jednorodny lub niejednorodny). W tym celu wykorzystuje się test *D* (test ilorazu wiarygodności).

Wykonane analizy wskazują na zróżnicowane funkcjonowanie zadań w zależności od wersji arkusza w każdym z analizowanych zadań testowych. Zastosowanie regresji logistycznej pozwala także na stwierdzenie, z jakiego rodzaju obciążeniem mamy do czynienia. W analizowanej sytuacji w przypadku z11.1 mamy do czynienia z jednorodnym DIF. Dla innych analizowanych pytań test ilorazu wiarygodności jest istotny w przypadku wszystkich porównywanych modeli. W takiej sytuacji mamy do czynienia z niejednorodnym DIF.

Standaryzacja pozwala na wizualną ocenę występowania DIF, dzięki czemu jest popularnym narzędziem raportowania wyników analiz tego zjawiska. Na Rysunku 1 zaprezentowano wykresy poziomu wykonania (*p*) analizowanych w artykule zadań w 15 interwałach skali umiejętności, będącej sumą punktów uzyskanych z historii i WOS na egzaminie.

Rysunek 2 przedstawia wykresy prawdopodobieństwa (oszacowanego w modelu regresji logistycznej) udzielenia poprawnej odpowiedzi na analizowane zadanie, przy kontroli poziomu umiejętności, wyrażonego jako suma punktów z testu.

Tabela 5
Wyniki analiz DIF wykonanych za pomocą regresji logistycznej

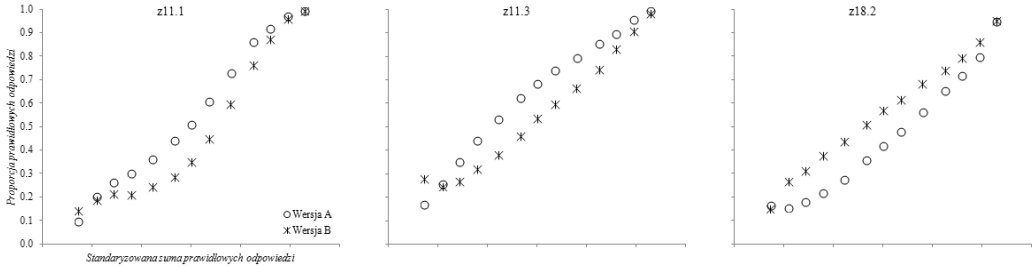
	Iloraz wiarygodności			2(LL Model 3-LL Model 2)	2(LL Model 2-LL Model 1)	Interpretacja	
Zadanie	Model 1[(a)]	Model 2[(b)]	Model 3[(c)]	Niejednorodny DIF	Jednorodny DIF	Niejednorodny DIF	Jednorodny DIF
z11.1	-48 017,2	-47 320,9	-47 320,0	1,8	1392,6**	Nie	Tak
z11.3	-49 461,3	-48 687,4	-48 681,2	12,4**	1547,8**	Tak	Nie
z18.2	-51 757,5	-50 950,1	-50 922,7	54,8**	1614,8**	Tak	Nie

** Istotny na poziomie $p < 0{,}01$

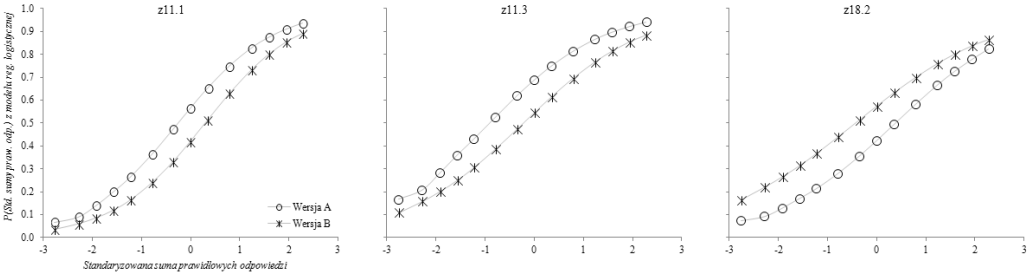
(a) $\ln\left(\frac{P(u=1)}{P(u=0)}\right) = \beta_0 + \beta_1 X$, gdzie X to wynik końcowy ucznia na teście.

(b) $\ln\left(\frac{P(u=1)}{P(u=0)}\right) = \beta_0 + \beta_1 X + \beta_2 G$, gdzie G to wersja testu (A lub B).

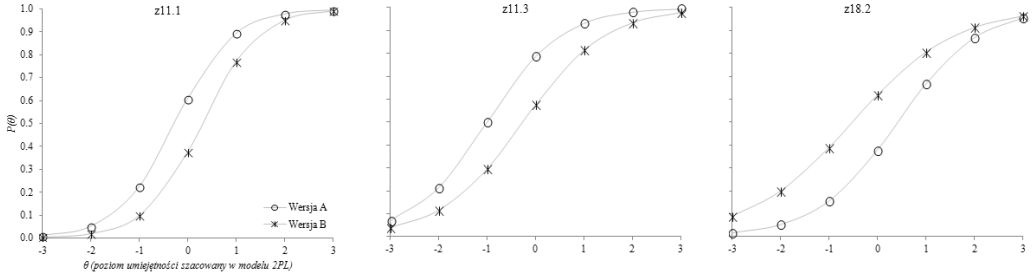
(c) $\ln\left(\frac{P(u=1)}{P(u=0)}\right) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG$, gdzie XG to interakcja między wynikiem końcowym a wersją arkusza.



Rysunek 1. Nieparametryczne wykresy regresji odpowiedzi na zadania testowe na skalę umiejętności.



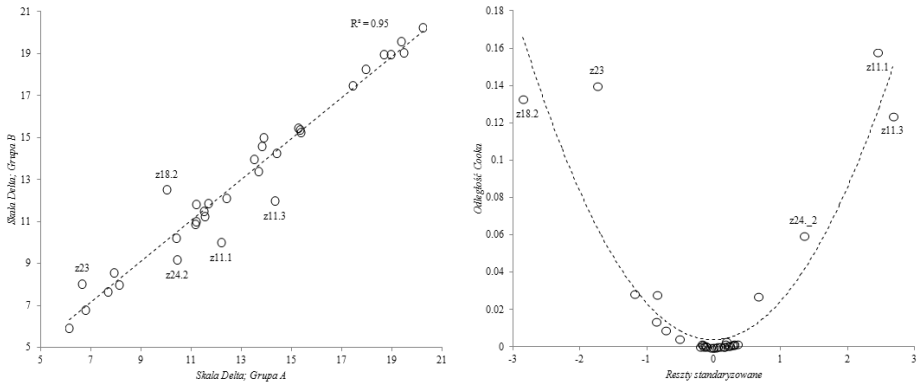
Rysunek 2. Parametryczne (regresja logistyczna) wykresy regresji odpowiedzi na zadania testowe na skalę umiejętności.



Rysunek 3. Parametryczne (2PLM) wykresy regresji odpowiedzi na zadania testowe na skalę umiejętności.

Rysunek 3 przedstawia wykresy prawdopodobieństwa (oszacowanego w modelu 2PL IRT) udzielenia poprawnej odpowiedzi na analizowane zadanie przy kontroli poziomu umiejętności, wyrażonego jako suma punktów z testu. Każdy wykres prezentuje krzywe charakterystyczne dla danego zadania w dwóch grupach uczniów wyodrębnionych ze względu na rozwiązywanie wersji A lub B testu.

Podejście nieparametryczne, oparte na regresji wartości *p* na skali umiejętności, oraz dwa podejścia parametryczne, oparte na wartościach θ oszacowanych w modelu regresji logistycznej i w modelu 2PL IRT, dają bardzo zbliżone rozwiązania. Wykresy



Rysunek 4. Wykresy oparte na metodzie *delta-plot*.

zaprezentowano dla trzech kodowanych dychotomicznie zadań obciążonych DIF z arkusza z historii i WOS z 2013 r. Zadania z11.1 oraz z11.3 faworyzują uczniów rozwiązujących wersję A arkusza. Z kolei zadanie z18.2 faworyzuje uczniów rozwiązujących wersję B arkusza. Są to wersje, w których nie występuje antywzorzec.

W celu graficznej prezentacji DIF wykonano także analizy z wykorzystaniem metody *delta-plot*, przedstawiono również przykład wykorzystania odległości Cooka. Przypomnijmy, że metody te wywodzą się z diagnostyki modeli regresji w celu identyfikacji przypadków odstających. Wyniki zastosowania tych procedur znajdują się na Rysunku 4.

W przypadku obu wykresów *delta-plot* wśród zadań, które należy określić jako przypadki odstające, znajdują się zadania w wiązkach, zidentyfikowane jako obciążone DIF. Oprócz nich możemy wskazać dodatkowo na inne zadania testowe, jednak w tych przypadkach nie można obciążenia DIF tłumaczyć hipotezą antywzorca.

Podsumowując wyniki analiz, należy stwierdzić, że potwierdzona została hipoteza o zróżnicowanym funkcjonowaniu zadań w wiązkach (z11.1, z11.3 oraz z18.2) ze względu na wersję arkusza testu. Wyniki testów MH wskazują na średnie efekty obciążenia. Analizy wykonane z użyciem regresji logistycznej wskazują, że zadanie z11.1 obciążone jest jednorodnym, z kolei zadania z11.3 i z18.2 niejednorodnym DIF. Graficzna detekcja DIF z użyciem standaryzacji oraz metody *delta-plot* potwierdza te wnioski, dając także możliwość prezentacji różnic między dwiema grupami uczniów oraz siłę obciążenia analizowanych pytań na tle pozostałych zadań w teście.

Testowanie hipotezy antywzorca

Analizy zadań z11.1, z11.2 oraz z18.2 wykazały ich zróżnicowane funkcjonowanie ze względu na wersję A i B arkusza. Jako prawdopodobną przyczynę wskazano specyficzny wzór poprawnych odpowiedzi na zadania w ramach wiązek 11 i 18. Określono to jako hipotezę antywzorca, czyli sytuację, w której prawidłowy (ale specyficzny) układ wariantów odpowiedzi w ramach jednej wiązki zadań prowadzi do popełniania błędów przez uczniów.

Hipotezę antywzorca testowano, wykorzystując serię regresji logistycznych, w których zmiennymi zależnymi były zadania o największym DIF w wiązkach 11 i 18. Jako zmienną wyjaśniającą w modelach stosowano sumę poprawnych odpowiedzi na wszystkie pozycje w teście (dzięki temu kontrolowano poziom umiejętności

uczniów) oraz interakcję odpowiedzi na pozostałe zadania w danej wiązce. W ten sposób sprawdzano, w jakim stopniu na prawdopodobieństwo udzielenia prawidłowej odpowiedzi na daną pozycję w wiązce wpływają odpowiedzi na pozostałe pozycje testowe, przy kontroli poziomu umiejętności uczniów. Formalnie, używany model regresji możemy zapisać następująco:

$$\ln\left(\frac{P(u=1)}{P(u=0)}\right) = \beta_0 + \beta_1 X + \beta_2 Z_i Z_j \qquad (4)$$

gdzie:

X – wynik końcowy ucznia w teście;

Z_iZ_j – interakcja pomiędzy wynikami za i-te oraz j-te zadanie w wiązce.

Każdy z modeli był analizowany osobno dla wersji arkusza A i B. Jest to zasadne, jeśli przyjmiemy, że to właśnie antywzorzec, jako cecha wersji arkusza, wpłynął na osiągane przez uczniów rezultaty w ramach badanych wiązek zadań. W przypadku wersji arkusza A przeprowadzono analizę dla zadania z18.2, natomiast dla wersji B policzono modele dla zadań z11.1 oraz z11.3.

Wyniki analiz przedstawiono w Tabeli 6. W przypadku zadań należących do wiązki 11 wykonane obliczenia potwierdziły istnienie zjawiska zmiany odpowiedzi pod wpływem antywzorca. Udzielenie prawidłowej odpowiedzi na pozostałe pytania w wiązce obniża prawdopodobieństwo poprawnej odpowiedzi na zadanie z11.1 o 45%, przy kontroli wpływu poziomu umiejętności uczniów (mierzonych jako sumaryczny wynik końcowy). W przypadku zadania z11.3 uzyskano bardzo podobny rezultat. Prawidłowe odpowiedzi na pytania z11.1 oraz z11.2 obniżają prawdopodobieństwo udzielenia dobrej odpowiedzi na zadanie z11.3 o 44%. W przypadku zadania z18.1 także uzyskano rezultaty zgodne z założoną hipotezą antywzorca. Udzielenie prawidłowej odpowiedzi na pozostałe zadania w tej wiązce zmniejsza prawdopodobieństwo udzielenia prawidłowej odpowiedzi na zadanie z18.2 o 60%.

Dyskusja wyników

Przedstawione w artykule wyniki analiz dowiodły istnienia zróżnicowanego funkcjonowania zadań w arkuszu egzaminacyjnym

Tabela 6
Wyniki regresji logistycznej dla zadań z11.1, z11.3 i z18.2

Model	β	*SE*	Exp(β)
Zadanie z11.1			
Wynik testu	0,216**	0,003	1,241
z11.2*z11.3	-0,595**	0,025	0,551
Stała	-4,247**	0,049	0,014
Zadanie z11.3			
Wynik testu	0,170**	0,002	1,186
z11.1*z11.2	-0,575**	0,026	0,563
Stała	-2,918**	0,044	0,054
Zadanie z18.2			
Wynik testu	0,175**	0,002	1,191
z18.1*z18.3	-0,905**	0,025	0,404
Stała	-3,159**	0,043	0,042

** Istotny na poziomie $p < 0,01$

z historii i WOS z 2013 r. ze względu na wersję arkusza egzaminacyjnego. Największe zróżnicowanie wykazano w przypadku zadań z11.1, z11.3 oraz z18.2. Na dwa pierwsze zadania poprawnie odpowiedziało zdecydowanie mniej egzaminowanych, rozwiązujących wersję B arkusza, natomiast dla zadania z18.2 odnotowano niższy odsetek prawidłowych odpowiedzi wśród rozwiązujących wersję A arkusza. Analizy potwierdziły, że są to zadania o istotnym efekcie DIF, którego siłę należy określić jako średnią. Sformułowano hipotezę, która obserwowane zależności tłumaczy specyficznym układem odpowiedzi w ramach danych wiązek zadań. W przypadku wersji A w wiązce zadań 18 wszystkie prawidłowe odpowiedzi miały symbol A, natomiast w wersji B w wiązce 11 prawidłowe odpowiedzi oznaczono jako C. Hipoteza antywzorca, jak nazwano tego rodzaju oddziaływanie, polega na uznaniu przez uczniów takiej sytuacji za mało prawdopodobną (trzy takie same odpowiedzi poprawne A lub C) i na celowej zmianie odpowiedzi na najtrudniejsze dla siebie pytanie, tak aby wzorzec odpowiedzi uprawdopodobnić. W literaturze podobny mechanizm istnieje pod nazwą tzw. paradoksu hazardzisty (*gambler's fallacy*; Sundali i Croson, 2006). Polega on na błędnym traktowaniu losowych zdarzeń niezależnych jako zależnych. Przejawia się w przekonaniu, że zdarzenie będące elementem serii ocenianej jako mało prawdopodobna powinno być przerwane bardziej prawdopodobnym zdarzeniem. Omawiane zagadnienie można także rozpatrywać w ramach psychologicznej teorii decyzji (Kozielecki, 1975).

Hipoteza antywzorca została zweryfikowana za pomocą serii regresji logistycznych, gdzie zmienną zależną było pytanie obciążone DIF, a kluczową zmienną wyjaśniającą była interakcja między pozostałymi zadaniami w wiązce przy kontroli sumy punktów uzyskanych w teście. Otrzymane rezultaty skłaniają do przyjęcia zakładanej hipotezy. Uwzględnienie odpowiedzi uczniów na pozostałe zadania w wiązce, zmniejsza prawdopodobieństwo udzielenia prawidłowej odpowiedzi na obciążone DIF zadania w dużym stopniu, tj. od 44 do 60%. Jest to mocny dowód potwierdzający istnienie efektu, który został w artykule nazwany hipotezą antywzorca. Weryfikacja tej hipotezy możliwa jest jednak jedynie w toku badań eksperymentalnych, które jednoznacznie wskazywałyby na przyczynę tego obciążenia. Badania te należy prowadzić w formie testów administrowanych na komputerach (*computer based testing*, CBT). Pozwoliłoby to kontrolować sekwencję pytań oraz odpowiedzi, jak również czas ich rozwiązywania (*test speededness effect*).

Na sam koniec należy rozważyć konsekwencje przedstawionych wniosków. Zróżnicowane funkcjonowanie zadań ze względu na wersję arkusza jest obciążeniem, które może być kontrolowane na etapie tworzenia testu. Wykonane analizy wskazują, że w przypadku analizowanego egzaminu kontrola ta nie była wystarczająca. Przedstawione analizy jasno wskazują, że obciążone zadania generują różne wyniki w zależności od wersji testu, co może być powodowane unikaniem wzorca odpowiedzi uznanego przez ucznia za mało prawdopodobny. Jako rozwiązanie praktyczne należy zalecić, aby unikać w arkuszach testowych wzorów prawidłowych odpowiedzi zadań w wiązce typu A, A, A itp. Tym bardziej należy się wystrzegać tego typu układu prawidłowych odpowiedzi, gdy są one wydrukowane w arkuszu w bezpośrednim sąsiedztwie, np. jedna po drugiej.

Rekomenduje się prowadzenie badań pilotażowych arkuszy egzaminacyjnych dokładnie w takiej formie, w jakiej są stosowane na egzaminach (z uwzględnieniem wersji arkuszy różniących się kolejnością odpowiedzi). Tylko analizując wyniki tak przeprowadzonego pilotażu, można

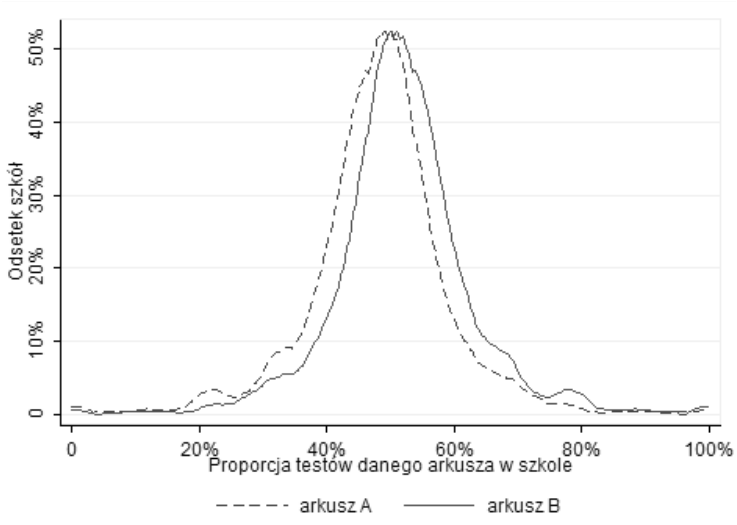
wykonać analizy mające na celu sprawdzenie zróżnicowanego funkcjonowania zadań testowych ze względu na wersje arkusza.

Następną kwestią, którą należy rozważyć w tym miejscu, są konsekwencje istnienia DIF w analizowanych zadaniach. W przypadku danych z egzaminu gimnazjalnego z historii i WOS z 2013 r. obciążenie całego testu wynikające z zadań o zróżnicowanym funkcjonowaniu nie jest widoczne w końcowym wyniku, zarówno surowym, jak i oszacowanym w modelu 2PL. Różnica średnich wyników między grupami jest nieistotna statystycznie. W tym konkretnym przypadku zatem nie możemy mówić o niepokojących konsekwencjach dla trafności interpretacji średnich wyników uczniów w szkołach czy jednostkach administracyjnych.

Dowiedziono jednak, że w przypadku tego konkretnego egzaminu jedna z wiązek zadań faworyzuje uczniów piszących wersję A, natomiast druga uczniów piszących wersję B. Taka sytuacja obniża trafność interpretacji indywidualnych wyników uczniów, co na egzaminie gimnazjalnym, który jest testem selekcyjnym do szkół średnich, jest bardzo niepokojące.

Jeśli wszystkie obciążone zadania znalazłyby się w jednej wersji arkusza, wyniki uczniów byłyby nietrafną miarą umiejętności także na zagregowanym poziomie (klasy, szkoły, jednostki administracyjne), co mogłoby mieć wpływ na nieadekwatne decyzje administracyjne, gdy egzaminy zewnętrzne pełnią funkcję diagnostyczną i rozliczeniową. Jest to powód, ze względu na który należy zawsze badać zróżnicowane funkcjonowanie zadań między wersjami arkusza już na poziomie pilotażu testów.

Dwie wersje arkuszy powinny być losowo dystrybuowane w szkołach. Niestety, można zaobserwować odstępstwa od tej procedury. Jeśli arkusz zawierający obciążone zadania dodatkowo trafi do większej liczby uczniów niż połowa (co wynikałoby z losowej dystrybucji), efekt obciążenia byłby dodatkowo wzmocniony faktem nie w pełni losowej dystrybucji. W przypadku analizowanych gimnazjów z województw: lubelskiego, małopolskiego i podkarpackiego, średnio



Rysunek 5. Funkcja gęstości Kernela dla proporcji arkuszy wersji A i B w analizowanych szkołach.

wersja A trafiła do 48% uczniów, z kolei wersja B do 52%, co jest dowodem na brak w pełni losowej dystrybucji. Istnieją jednak szkoły, w których proporcja wersji A testu do wersji B była znacznie bardziej zachwiana. Przy pełnej losowości rozkład prezentowany na Rysunku 5 powinien stanowić linię prostą przecinającą oś odciętych w punkcie 50%. Widać, że wykres gęstości przesunięty jest w prawo, ponieważ rozdystrybuowano więcej arkuszy w wersji B.

Literatura

Angoff, W. H. (1972) *The development of statistical indices for detecting cheaters*. Princeton, NJ: Educational Testing Service.

Bock, R., Muraki, E., i Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*(4), 275–285.

Cohen, J. (1988). *Statistical power analysis for the behavior sciences*. Hillsdale, NJ: Erlbaum.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*, 15–18.

Cook, R. D. (1979). Influential observation in linear regression. *Journal of the American Statistical Association*, *74*(365), 169–174.

Dorans, N. J. i Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*(4), 355–368.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: standardization and the Mantel–Haenszel method. *Applied Measurement in Education*, *2*(3), 217–233.

Dorans, N. J. i Holland, P. W. (2009). DIF detection and description: Mantel–Haenszel and standardization. W: P. W. Holland i H. Wainer (red.), *Differential item functioning* (s. 35–66). New York: Routledge.

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: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. London: Erlbaum.

Holland, P. W. i Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. W: H. Wainer i H. Braun (red.), *Test validity* (s. 129–145). Hillsdale, NJ: Erlbaum.

Holland, P. W. i Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

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* (s. 117–155). Baltimore: Johns Hopkins University Press.

Kingston, N. M. i Dorans, N. J. (1985). The analysis of item-ability regressions: an exploratory IRT model fit tool. *Applied Psychological Measurement*, 9, 281–288.

Kondratek, B. i 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*, *122*(2), 34–55.

Kozielecki, J. (1975). *Psychologiczna teoria decyzji*. Warszawa: PWN.

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*(2), 159–173.

Mantel, N. i Haenszel. W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748.

Muthén, B. i Lehman, J. (1985). Multiple-group IRT modeling: applications to item bias analysis. *Journal of Educational Statistics*, *10*(2), 133–142.

Narayanan, P. i Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257–274.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research. Explanation and prediction* (wyd. 3). Belmont, CA: Wadsworth Thomson Learning.

Sundali, J. i Croson, R. (2006). Biases in casino betting: the hot hand and the gambler's fallacy. *Judgment and Decision Making*, *1*(1), 1–12.

Swaminathan, H. i Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

Szaleniec, H. (2006). Oszukiwanie na egzaminie istotnym źródłem majowej porażki. W: B. Niemierko i M. K. Szmigel (red.), *O wyższą jakość egzaminów szkolnych*. Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Thissen, D., Steinberg, L. i Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. W: H. Wainer i H. Braun (red.), *Test validity* (s. 147–169). Hillsdale, NJ: Erlbaum.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433– 451.

Wainer, H. (1993). Model-based standardized measurement of an item's differential impact. W: P. W. Holland i H. Wainer (red.), *Differential item functioning* (s. 123–135). Hillsdale, NJ: Erlbaum.

Zwick, R. i Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*(1), 55–66.