Magdalena Grudniewska
Bartosz Kondratek
Instytut Badań Edukacyjnych

Zróżnicowane funkcjonowanie zadań w egzaminach zewnętrznych w zależności od płci na przykładzie części matematyczno-przyrodniczej egzaminu gimnazjalnego

Wstęp

U podstaw budowania zestawów testowych leży założenie, iż test mierzy jedną, ukrytą cechę, czyli umiejętności ucznia w zakresie objętym danych testem. Zakładamy zatem, iż wszelkie różnice w wynikach uczniów spowodowane są wyłącznie różnym rozkładem umiejętności i są niezależne od cech ucznia. Jeżeli porównamy dwóch uczniów o tym samych umiejętnościach, prawdopodobieństwo udzielenia poprawnej odpowiedzi na poszczególne zadania testowe powinno być dla tych uczniów porównywalne, niezależnie od ich charakterystyk społeczno-demograficznych.

Zagadnienie zróżnicowanego funkcjonowania zadań testowych jest ściśle powiązane z trafnością teoretyczną testu. Trafność teoretyczna zadania testowego oznacza, iż osoby o takich samych umiejętnościach powinny mieć taką samą szansę udzielenia prawidłowej odpowiedzi w tym zadaniu (Hornowska, 134). Jeżeli szansa udzielenia prawidłowej odpowiedzi jest różna w zależności od przynależności grupowej (przy kontroli umiejętności uczniów), wówczas trafność testu jest zaburzona i test mierzy dodatkowy aspekt, niezwiązany z umiejętnościami ucznia.

Analiza wyników egzaminu gimnazjalnego oraz uczestnictwa w olimpiadach przedmiotowych ze względu na płeć

Tabela 1. pozwala porównać średnie wyniki dziewcząt i chłopców w części matematyczno-przyrodniczej egzaminu gimnazjalnego na przestrzeni lat 2002-2011. Należy pamiętać, iż w poszczególnych latach zmieniała się trudność arkusza egzaminacyjnego. W związku z brakiem w polskim systemie egzaminów zewnętrznych procedur zrównujących, nieuprawnione jest porównywanie wyników uczniów pomiędzy poszczególnymi latami, gdyż wyniki są ściśle związane z cechami testu egzaminacyjnego w danym roku. Nie to jest jednak celem tego zestawienia. Tabela pozwala porównać różnice między średnimi wynikami chłopców i dziewcząt w poszczególnych latach. Zauważmy, iż w średnia chłopców w latach 2002-2007 jest nieznacznie wyższa aniżeli średnia dziewcząt, natomiast od roku 2008 ta tendencja się odwraca i aż do egzaminu z roku 2011 to dziewczęta uzyskują średnio wyższe wyniki niż chłopcy.

Tabela 1. Średni wynik uczniów w części matematyczno-przyrodniczej egzaminu gimnazjalnego w latach 2002-2011

Średni wynik			
rok	ogółem	chłopcy	dziewczęta
2002	28,16	28,32*	28,00*
2003	25,75	25,78*	25,71*
2004	24,49	24,63*	24,35*
2005	24,26	24,40*	24,11*
2006	23,90	23,90	23,89
2007	25,31	25,77*	24,83*
2008	27,07	26,89*	27,27*
2009	26,03	25,59*	26,47*
2010	23,90	23,89*	23,91*
2011	23,46	23,30*	23,62*

*różnica średnich między chłopcami a dziewczętami istotna statystycznie na poziomie 0,05

Analizując umiejętności chłopców i dziewcząt z zakresu przedmiotów humanistycznych i matematyczno-przyrodniczych, warto zwrócić również uwagę na udział chłopców i dziewcząt wśród laureatów olimpiad przedmiotowych. Analiza liczby laureatów zostanie ograniczona do roku 2011, ponieważ tylko dla tego roku zbierane były dane na temat liczby laureatów[1]. Tabela 2. zawiera informację na temat liczby laureatów olimpiad matematyczno-przyrodniczych i humanistycznych wśród uczniów klas trzecich gimnazjum ogółem oraz w podziale na chłopców i dziewczęta. Zwróćmy uwagę, iż wśród gimnazjalistów więcej jest laureatów olimpiad matematyczno-przyrodniczych: w 2011 roku takich uczniów było 2364, czyli o około 1000 więcej w porównaniu z liczbą laureatów olimpiad humanistycznych. Zauważmy, iż wśród laureatów olimpiad humanistycznych więcej jest dziewcząt, natomiast w przypadku olimpiad z zakresu przedmiotów matematyczno-przyrodniczych przewagę mają chłopcy. Co ciekawe, różnica między procentowym udziałem chłopców i dziewcząt w obu typach olimpiad jest niemal taka sama. Co jest powodem takiego zróżnicowania? Najprostsze wyjaśnienie to wyższe umiejętności dziewcząt z zakresu przedmiotów humanistycznych, a chłopców z przedmiotów matematyczno-przyrodniczych, jednak takie wyjaśnienie może być tylko pozornie trafne. Taki wynik może być pochodną sposobu nauczania przedmiotów matematyczno-przyrodniczych i humanistycznych, podejścia nauczycieli, sposobu

[1] W bazach danych z CKE z wynikami uczniów z egzaminów zewnętrznych laureatom olimpiad przedmiotowych, którzy z tego tytułu zwolnieni są z pisania egzaminu, przypisana zostaje maksymalna liczba punktów. Zidentyfikowanie tych uczniów możliwe jest przez dodatkową zmienną, która zawiera informację, czy dany uczeń był laureatem olimpiady z danego zakresu. W przypadku danych z egzaminu gimnazjalnego z pozostałych lat, taka informacja nie była odnotowana. W związku z niemożnością rozróżnienia w przypadku lat 2002-2010 laureatów od pozostałych uczniów, którzy otrzymali maksymalną liczbę punktów, wszystkie analizy funkcjonowania zadań zostały wykonane po odfiltrowaniu uczniów, którzy uzyskali maksymalną liczbę punktów.

motywowania uczniów oraz zainteresowań i decyzji uczniów, na które z kolei wpływa wiele obciążeń kulturowych i oczekiwań otoczenia. Powodów takiego stanu może być wiele i z pewnością ta kwestia wymaga głębszego wyjaśnienia i refleksji, nie to jest jednak przedmiotem tego artykułu. Ukazanie zróżnicowania wyników chłopców i dziewcząt w zakresie przedmiotów humanistycznych i matematyczno-przyrodniczych było jedynie tłem, a głównym celem artykułu jest zbadanie, jak zadania z egzaminu gimnazjalnego funkcjonują w zależności od płci ucznia.

Tabela 2. Laureaci olimpiad z zakresu przedmiotów humanistycznych i matematyczno-przyrodniczych wśród uczniów trzeciej klasy gimnazjum w roku 2011

	Liczba			% laureatów		% uczniów ogółem		
	chłopcy	dziewczęta	ogółem	chłopcy	dziewczęta	razem	chłopcy	dziewczęta
Część matematyczno-przyrodnicza	1299	1065	2364	55%	45%	0,58%	0,32%	0,26%
Część humanistyczna	585	782	1367	43%	57%	0,33%	0,14%	0,19%

Analiza funkcjonowania zadań egzaminacyjnych w grupie chłopców i dziewcząt

Analiza funkcjonowania zadań została przeprowadzona z wykorzystaniem programu jMetrik, który zawiera moduł umożliwiający obliczanie efektu DIF zarówno dla zadań dychotomicznych (punktowanych 0-1), jak również dla zadań wielopunktowych. Procedura służąca do obliczania efektu zróżnicowanego funkcjonowania zadań zbudowana jest w oparciu o test Mantel-Heanshel, który zostanie szerzej opisany w dalszej części artykułu. Niewątpliwą zaletą programu jMetrik jest fakt, iż zaimplementowana procedura pozwala nie tylko potwierdzić lub odrzucić hipotezę o zróżnicowanym funkcjonowaniu zadań, umożliwia również ocenę siły i kierunku tego efektu. Na podstawie wielkości efektu DIF oraz istotności statystycznej testu program przypisuje zadania do trzech klas („A", „B", „C" lub : „AA", „BB" i „CC"), wraz ze wskazaniem, dla której z grup dane zadanie jest łatwiejsze. Co oznaczają poszczególne klasy oraz na jakiej podstawie owa klasyfikacja została skonstruowana, zostanie również szerzej opisane w dalszej części artykułu.

Opis metody. Analiza DIF w oparciu o test Mantel-Haenshel[2]

W podejściu Mantel-Haenshel (MH) odpowiedzi na dychotomiczne zadanie uczniów z dwóch grup są stratyfikowane ze względu na liczbę punktów zdobytych w całym teście, w wyniku czego powstaje tablica kontyngencji o wymiarach 2x2x, gdzie M jest liczbą kategorii punktowych wyniku sumarycznego. Prawdopodobieństwa zaobserwowania danej odpowiedzi na rozpatrywane

[2] Część artykułu poświęcona metodzie Mantel-Haenshel oraz klasyfikacji efektu DIF została napisana w oparciu o artykuł: Grudniewska M., Kondratek B., *Test Mantel-Haenshel oraz modelowanie IRT jako narzędzia służące do wykrywania DIF oraz opisu jego wielkości na przykładzie zadań ocenianych dychotomicznie* [w druku].

zadanie w zależności od przynależności grupowej oraz od kategorii punktowej, oznaczymy w następujący sposób:

Tabela 3. Podział na grupy wyodrębnione ze względu na przynależność grupową oraz udzieloną odpowiedź w zadaniu 0-1 - prawdopodobieństwa

		Odpowiedź na zadanie (zadanie punktowane 0-1)		Łącznie
		1 (odpowiedź poprawna)	0 (odpowiedź błędna)	
Grupa	f	p_{1fm}	p_{0fm}	p_{fm}
	r	p_{1rm}	p_{0rm}	p_{rm}
Łącznie		p_{1m}	p_{0m}	p_m

Test MH klasycznie jest opisywany w języku ilorazu szans. Szansą (ang. *odds*) udzielenia odpowiedzi poprawnej określa się stosunek prawdopodobieństwa udzielenia odpowiedzi poprawnej do prawdopodobieństwa udzielenia odpowiedzi błędnej, iloraz takich szans dla uczniów z grup *r* i *f* w kategorii punktowej *m* jest zatem:

$$\alpha_m = \frac{p_{1rm}/p_{0rm}}{p_{1fm}/p_{0fm}} \tag{1}$$

Przy powyższych oznaczeniach hipotezę zerową i alternatywną testu Mantel-Haenshel zapisuje się w następujący sposób (por. Dorrans & Holland, 1993):

$$\begin{aligned} H_0&: \alpha_m = \alpha \qquad ,m \in \{1, \dots, M\}, \qquad \alpha = 1 \\ H_1&: \alpha_m = \alpha \qquad ,m \in \{1, \dots, M\}, \qquad \alpha \neq 1 \end{aligned} \tag{2}$$

Hipoteza zerowa zatem stanowi, że szanse udzielenia odpowiedzi poprawnej na zadanie w dwóch grupach są takie same w każdej kategorii punktowej *m*. Można by ją równoważnie zapisać w konwencji, która podkreśla rozumienie DIF jako zróżnicowane między grupami prawdopodobieństwo udzielenia określonej odpowiedzi, gdy kontrolowany jest umiejętności:

$$H_0: P(U_i = 1|m, f) = P(U_i = 1|m, r) \qquad ,m \in \{1, \dots, M\} \tag{3}$$

Hipoteza zerowa zapisana wzorem (3) stanowi, że prawdopodobieństwo udzielenia odpowiedzi poprawnej na zadanie nie zależy od przynależności grupowej, jeżeli uwzględnimy wynik w całym teście. Bardzo specyficzna dla testu MH jest hipoteza alternatywna, względem której jest testowana H_0. Wedle H_1 we wzorze (2) różnica tych prawdopodobieństw będzie niezerowa z tym samym znakiem dla każdej kategorii punktowej *m*, a co więcej - wszystkie ilorazy szans α_m będą równe wspólnemu ilorazowi szans α (ang. *common odds ratio*).

Jeżeli liczebności będziemy indeksować analogicznie do prawdopodobieństw (tabela 3.), uzyskamy tablicę kontyngencji przedstawioną w tabeli 4.

Tabela 4. Podział na grupy wyodrębnione ze względu na przynależność grupową oraz udzieloną odpowiedź w zadaniu 0-1 - liczebności

		Odpowiedź na zadanie (zadanie punktowane 0-1)		Łącznie
		0 (odpowiedź błędna)	1 (odpowiedź poprawna)	
Grupa	r	N_{0rm}	N_{1rm}	N_{rm}
	f	N_{0fm}	N_{1fm}	N_{fm}
Łącznie		N_{0m}	N_{1m}	N_m

Statystykę dla testu MH z poprawką na ciągłość przedstawiamy jako:

$$MH_{\chi^2} = \frac{\left(\sum_{m=1}^{M}\left(N_{1fm} - E(N_{1fm})\right) - 0{,}5\right)^2}{\sum_{m=1}^{M} D^2(N_{1fm})} \tag{4}$$

gdzie $E(N_{1fm})$ oraz $D^2(N_{1fm})$ są wartością oczekiwaną i wariancją liczebności N_{1fm} przy prawdziwości H_0. Przy prawdziwości hipotezy zerowej statystyka MH_{X^2} ma rozkład asymptotycznie $_{X^2}$ z jednym stopniem swobody (Dorrans & Holland, 1993).

Dla testu MH wykazano (Radhakrishn, 1965), że jest on jednorodnie najmocniejszym testem dla hipotezy zerowej o warunkowej niezależności proporcji między grupami, przy prawdziwości hipotezy o stałym ilorazie szans. W przypadku gdy hipoteza o stałym ilorazie szans nie jest prawdziwa, test MH traci na mocy. Oznacza to, że test MH będzie sobie gorzej radził z wykrywaniem niejednorodnego DIF w porównaniu do procedur, które dopuszczają interakcję między wielkością DIF mierzoną jako iloraz szans, a poziomem umiejętności, co zostało zauważone np. u Swaminathan & Rogers (1990). Ze wzoru (3) widać, że w skrajnych przypadkach, gdy ilorazy szans α_m będą się zmieniały w zależności od m tak, że dla części m będą powyżej 1, a dla części poniżej 1, to odpowiednie wkłady odchyleń liczebności N_{1fm} od ich wartości oczekiwanych będą się wzajemnie znosiły. Ze względu na opisaną zależność właściwości testu MH od spełnienia założenia o stałości ilorazu szans, przeprowadzaniu tego testu często towarzyszy dodatkowa procedura weryfikująca spełnienie tego założenia, np. test Wolfa (1955).

Mantel i Haenshel w swojej oryginalnej pracy (1959) zaproponowali również estymator wspólnego ilorazu szans w postaci:

$$\alpha_{MH} = \frac{\sum_{m=1}^{M} p_{1rm} p_{0fm} N_m}{\sum_{m=1}^{M} p_{1fm}\, p_{0fm} N_m} \tag{5}$$

Jak widać większą wagę przy obliczaniu α_{MH} mają komórki z większą brzegową liczebnością N_m. Dla zadania, które przy kontroli poziomu umiejętności jest łatwiejsze dla grupy r uzyskamy $\alpha_{MH} > 1$, dla sytuacji odwrotnej będzie $\alpha_{MH} < 1$.

Test Mantel-Haenshel, w oryginalnej wersji stosowany w analizie zmiennych kodowanych 0 - 1, można uogólnić na zmienną zależną przyjmującą postać 0 - k, co pozwala zaimplementować procedurę MH do oceny zadań wielokategorialnych. Wspomniane uogólnienie testu Mantel-Haenshel dla tabel przyjmujących postać *IxJxK* zostało opisane przez Agresti (2002).

Miary wielkości efektu DIF i klasyfikacja zadań ze względu na DIF

Wspólny iloraz szans statystyki MH (3) stanowi dość trudną w interpretacji miarę wielkości efektu DIF. W celu ułatwienia interpretacji wartości α_{MH} poddawana jest ona różnym przekształceniom. Jednym z nich jest wskaźnik *MH D – DIF* uzyskiwany w następujący sposób:

$$MH\ D - DIF = -2.35 \ln[\alpha_{MH}] \quad (6)$$

Takie przekształcenie pozwala uzyskać rozkład symetryczny, z wartościami z zakresu od minus do plus nieskończoności. Wartość 0 oznacza brak efektu DIF.

Educational Testing Service opracował system klasyfikacji efektu DIF, która opiera się na istotności statystyki MH_X^2 (przyjmuje się standardowy próg istotności statystycznej, tj. α=0,05) oraz wielkości miary *MH D – DIF*. Na podstawie tych parametrów zadania przypisywane są do trzech kategorii „A", „B" oraz „C" (Dorans & Holland 1993; Zeiky, 2003) w następujący sposób:

- Kategoria A, gdy test MH dał wynik negatywny albo gdy wynik testu był pozytywny, ale absolutna wartość *MH D – DIF* jest mniejsza od 1. Przypisanie zadania do kategorii A oznacza, iż nie została potwierdzona hipoteza o zróżnicowanym funkcjonowaniu zadania w analizowanych grupach.
- Kategoria B, gdy test MH dał wynik pozytywny oraz absolutna wartość *MH D – DIF* jest w przedziale od 1 do 1,5 lub gdy test MH dał wynik pozytywny oraz 95% przedział ufności wokół *MH D – DIF* nie znajduje się poza przedziałem od -1 do +1. Przypisanie zadania do tej kategorii oznacza mały lub średni efekt DIF.
- Kategoria C, gdy 95% przedział ufności wokół *MH D – DIF* znajduje się poza przedziałem od -1 do +1 oraz absolutna wartość *MH D – DIF* jest większa od 1,5 (w szczególności oznacza to pozytywny wynik testu MH). Przypisanie zadania do klasy C jest równoważne ze stwierdzeniem wysokiego efektu DIF.

Przedstawiona powyżej klasyfikacja dotyczy zadań dychotomicznych, za rozwiązanie których uczeń może otrzymać jeden punkt. Jak już wcześniej wspomniano, statystyka Mantel-Haenshel może zostać również zostać uogólniona na zmienne wielokategorialne przyjmujące postać 1-k. Generalizacja klasyfikacji efektu DIF do oceny zadań wielokategorialnych została zastosowana w badaniu NAEP (National Assessment of Educational Progress). Dla zadań wielokategorialnych zasady klasyfikacji efektu DIF (Michaelides, 2008) są następujące:

- Kategoria „AA" - wynik testu Mantel-Haenshel x^2 jest negatywny (p≥0,05) lub wartość absolutna efektu DIF (*effect size*[3]) jest mniejsza lub równa 0,17. Przypisanie zadania do kategorii AA oznacza, iż nie ma statystycznych podstaw do odrzucenia hipotezy o braku zróżnicowanego funkcjonowania zadania w analizowanych grupach lub że efekt DIF jest niewielki co do wartości.
- Kategoria „BB" - wynik testu Mantel-Haenshel jest pozytywny i wartość absolutna efektu DIF jest większa niż 0,17 oraz mniejsza lub równa 0,25. Przypisanie zadania do tej kategorii oznacza istotny statystycznie mały lub średni efekt DIF.

[3] Wielkość efektu (*effect size*) obliczamy w tym przypadku, dzieląc współczynnik SDM (Standardized Mean Difference) przez odchylanie standardowe wewnątrz grup dla analizowago itemu, zsumowane dla dwóch grup. Procedurę obliczania SDM opisuje Micheelides (2008).

- Kategoria „CC" - wynik testu Mantel-Haenshel jest pozytywny oraz wartość absolutna miary wielkości efektu DIF (effect size) jest większa niż 0,25. Przypisanie zadania do klasy CC jest równoważne ze stwierdzeniem istotnego statystycznie wysokiego efektu DIF.

Zadania oznaczone „C" (lub w przypadku zadań wielopunktowych „CC") oraz „B" (lub „BB") wymagają od konstruktorów zwrócenia szczególnej uwagi pod kątem stronniczości. Dodatkowo wraz z informacją o kategorii DIF, do której należy dane zadanie, znajduje się informacja, czy zadanie jest trudniejsze dla *focal group* (zadania oznaczone „-") czy też dla *refference group* (zadania oznaczone „+").

Analiza funkcjonowania zadań ze względu na płeć

Zdania funkcjonujące w odmienny sposób w grupie dziewcząt i chłopców - przegląd zadań egzaminacyjnych z części matematyczno-przyrodniczej egzaminu gimnazjalnego w latach 2002-2011

Tabela 5. stanowi podsumowanie wyników analizy zróżnicowanego funkcjonowania zadań z części matematyczno-przyrodniczej egzaminu gimnazjalnego 2002-2011 w zależności od płci ucznia. W tabeli możemy zobaczyć, jaka była w poszczególnych latach liczba zadań faworyzujących jedną z grup oraz ile zadań okazało się łatwiejsze dla chłopców, a ile dla dziewcząt. Zauważmy, iż na przestrzeni lat 2002-2011 liczba zadań z efektem DIF podlegała znacznym fluktuacjom - w roku 2002, gdy pierwsi absolwenci gimnazjum zdawali egzamin, było to tylko jedno zadanie, natomiast w zestawie testowym z 2003 roku znalazło się już 7 zadań, które potencjalnie mogą być stronnicze ze względu na płeć. Najwięcej zadań funkcjonujących różnie w zależności od płci ucznia - aż 9 - znalazło się e egzaminie gimnazjalnym z roku 2006. Spójrzmy na liczbę zadań faworyzujących chłopców i dziewczęta w poszczególnych edycjach egzaminu: tylko w przypadku roku 2010 liczba zadań łatwiejszych dla dziewcząt i chłopców równoważy się, natomiast w latach 2002 i 2003 jest więcej zadań faworyzujących chłopców, a w edycjach egzaminu z lat 2004-2009 - przewaga zadań faworyzujących dziewczęta. Należy jednak podkreślić, iż o stronniczości testu jako narzędzia nie świadczy liczba zadań faworyzujących jedną lub drugą grupę, lecz wielkość efektu DIF. Możemy sobie wyobrazić sytuację, gdy test będzie zawierał mniej zadań z DIF na korzyść chłopców niż zadań z DIF na korzyść dziewcząt, jednak wielkość efektu DIF w zadaniach faworyzujących chłopców będzie większy niż w zadaniach faworyzujących dziewczęta i w konsekwencji narzędzie jako całość będzie stronnicze na korzyść chłopców. Aby zatem ocenić, czy test jest stronniczy, nie wystarczy opisanie liczby zadań z efektem DIF na korzyść jednej i drugiej grupy, musimy również dysponować miarą wielkości efektu DIF. Sposób obliczania wielkości efektu DIF nie jest przedmiotem tego artykułu, procedura ta została opisana w artykule Grudniewskiej i Kondratka.

Tabela 5. Liczba zadań z efektem DIF w części matematyczno-przyrodniczej egzaminu gimnazjalnego na przestrzeni lat 2002-2011

Rok	Liczba zadań z DIF przypisanych do klasy „B" („BB") lub „C" („CC")	Liczba zadań faworyzujących chłopców	Liczba zadań faworyzujących dziewczęta	Liczba zadań ogółem
2002	1	0	1	36
2003	7	3	4	34
2004	6	4	2	34
2005	5	4	1	35
2006	9	6	3	34
2007	5	3	2	34
2008	6	4	2	33
2009	7	5	2	36
2010	4	2	2	36
2011	3	3	0	36

Przykład zadania funkcjonującego w różny sposób w grupie dziewcząt i chłopców

Obok analizy funkcjonowania zadań jako całości, w przypadku zadań wielopunktowych została przeprowadzona także analiza zdań w podziale na kryteria. Taka analiza pozwala stwierdzić, z jakimi umiejętnościami lepiej radzą sobie dziewczęta, a z jakimi chłopcy. Wyniki analizy zadań w podziale na kryteria okazały się bardzo interesujące, szczególnie w przypadku zadań z matematyki.

Przeanalizujmy przykład zadania z roku 2002:

Zadanie 29. (0-3)

Marcin przebywa autobusem drogi do jeziora, a pozostałą część piechotą. Oblicz odległość między domem Marcina a jeziorem, jeżeli trasa, którą przebywa pieszo jest o 8 km krótsza niż trasa, którą przebywa autobusem. Zapisz obliczenia.

Za poprawne rozwiązanie zadania uczeń może otrzymać odpowiednio 1, 2 lub 3 punkty.

Punktacja przedstawia się następująco:

- Ustalenie zależności między poszczególnymi odcinkami szukanej drogi - 1 pkt
- Ułożenie równania - 1 pkt
- Rozwiązanie równania (zapisanie poprawnego wyniku) - 1 pkt.

Z kryterium pierwszym i drugim lepiej radzą sobie dziewczęta niż chłopcy, natomiast łatwość trzeciego kryterium jest taka sama w obu grupach.

Podobna zależność występuje w wielu innych zadaniach z matematyki czy też fizyki, gdzie uczeń powinien najpierw poprawnie rozpisać zadanie i ułożyć równanie, a następnie wykonać obliczenia i podać odpowiedź. Dziewczęta lepiej sobie radzą z poprawnym rozpisaniem zadania, natomiast na etapie rozwiązywania zadania nie ma różnic między grupami. Jest to prawidłowość, która systematycznie powtarza się w wielu zadaniach. Prawdopodobnie

przyczyną może być inny sposób wychowywania dziewczynek i chłopców oraz inne wymagania rodziców i wychowawców w zależności od płci dziecka: od dziewcząt oczekuje się, iż będą wzorowe, sumienne, posłuszne, chłopcom natomiast pozwala się na więcej spontaniczności. Taki sposób socjalizacji chłopców i dziewcząt do ról związanych z płcią można odnieść do sposobu rozwiązywania zadań przez te dwie grupy uczniów: chłopcy skupiają się na poprawnym rozwiązaniu zadania, dziewczynki przywiązują większą wagę do wymogów formalnych: poprawnego rozpisania zadania oraz ukazania toku rozumowania, który doprowadził do uzyskania poprawnego rozwiązania.

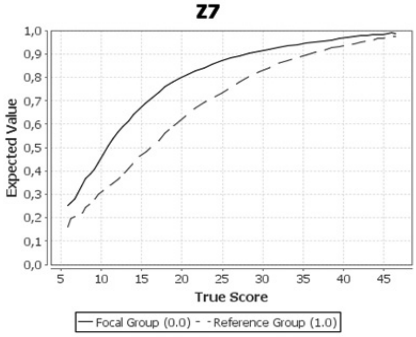
Warto odwołać się tutaj do wyników międzynarodowego badania osiągnięć uczniów PISA (*Programme for International Student Assessment)*. Interesujące są wyniki z roku 2009 dla Islandii: zaobserwowano przewagę dziewcząt w zakresie matematyki, przedmiotów ścisłych oraz czytania. Badania wykazują dużą stabilność wzorca nierówności płci na korzyść dziewcząt na przestrzeni lat (*Różnice w wynikach nauczania a płeć: obecna sytuacja i działania podejmowane w Europie*, 2010). Hallardórsen i Ólafsson wskazują natomiast, iż przewaga dziewcząt spowodowana jest czynnikami natury psychologicznej takimi jak większa przychylność dla „kultury uczenia się" w przypadku dziewcząt oraz słabszej dyscypliny i gorszego zachowania chłopców. Także w badaniu Międzynarodowej Oceny Rozwoju Edukacyjnego (*International Assessment of Educational Progress* - IAEP) stwierdzono, iż różnice we wzorach socjalizacji chłopców i dziewcząt mogą być czynnikiem najsilniej wpływającym na powstawanie różnic płci w zakresie umiejętności i osiągnięć (*Różnice w wynikach nauczania a płeć: obecna sytuacja i działania podejmowane w Europie*, 2010). Kwestia wpływu różnych wzorów socjalizacji chłopców i dziewcząt oraz ich wpływu na wyniki, umiejętności oraz motywację uczniów jest niezwykle interesująca i wymaga pogłębionych badań w tym zakresie.

Analiza graficzna funkcjonowania zadań z DIF

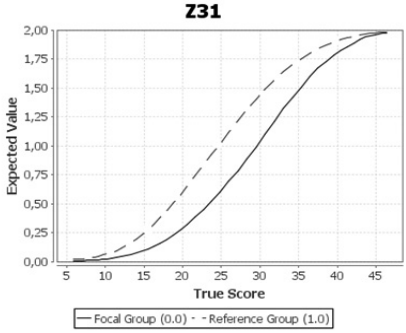
Na rysunku 1. przedstawione zostały przykładowe wykresy dla zadań, dla których odrzucona została hipoteza o braku zróżnicowanego funkcjonowania z zależności od płci. Wykresy są dobrym narzędziem oceny, w jaki sposób funkcjonuje zadanie w zależności od przynależności grupowej ucznia oraz ogólnego wyniku w teście. Oś pozioma symbolizuje wynik ucznia w teście, natomiast oś pionowa - prawdopodobieństwo udzielenia poprawnej odpowiedzi w danym zadaniu. Linia czarna przedstawia prawdopodobieństwo udzielenia prawidłowej odpowiedzi w grupie ogniskowej (*focal group*), natomiast linia czerwona - prawdopodobieństwo udzielenia prawidłowej odpowiedzi w grupie referencyjnej (*reference group*). We wszystkich wykonanych analizach przyjęto, iż grupa ogniskowa to chłopcy.

Na rysunku 1. widzimy, iż prawdopodobieństwo udzielenia poprawnej odpowiedzi w wybranych zadaniach różni się znacznie w grupie dziewcząt i chłopców. Dwa spośród wybranych zadań są łatwiejsze dla chłopców, dwa dla dziewcząt. Obserwując, jak zmienia się odległość między liniami w zależności od wyniku uczniów w całym teście, możemy ocenić, czy ta dysproporcja jest taka sama na wszystkich poziomach, czy też zmienia się w zależności od

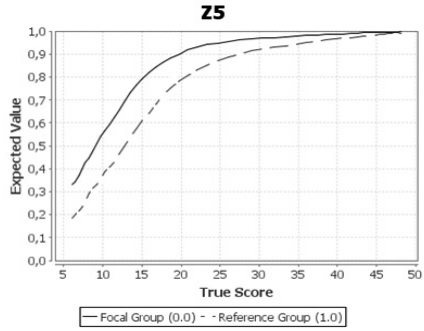
wyniku ucznia. Na przykład w przypadku zadania 5. z egzaminu gimnazjalnego w roku 2009 różnica między odpowiedziami chłopców i dziewcząt jest do pewnego wyniku punktowego taka sama, po czym dla uczniów, którzy uzyskali w całym teście powyżej 25 punktów, różnica między odpowiedziami chłopców i dziewcząt systematycznie maleje. W przypadku pozostałych zadań różnica między odpowiedziami chłopców jest największa dla uczniów, którzy uzyskali w teście wyniki z przedziału punktowego 15-30 punktów. Dla uczniów o najniższym i najwyższym wyniku różnica między odpowiedziami dziewcząt i chłopców systematycznie maleje.



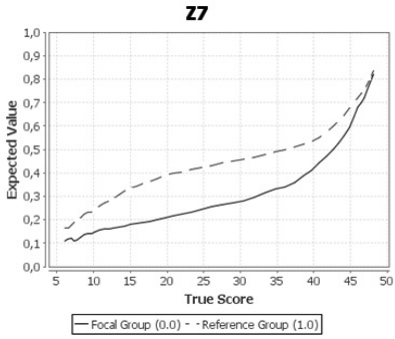
Zad. 7. część matematyczno-przyrodnicza egzaminu gimnazjalnego, rok 2010



Zad. 31. część matematyczno-przyrodnicza egzaminu gimnazjalnego, rok 2010



Zad. 5. część matematyczno-przyrodnicza egzaminu gimnazjalnego, rok 2009



Zad. 7. część matematyczno-przyrodnicza egzaminu gimnazjalnego, rok 2009

Rysunek 1. Wykresy z krzywymi nieparametrycznymi pokazujące funkcjonowanie zadania w grupie dziewcząt i chłopców w zależności od ogólnego wyniku w teście

Podsumowanie

Celem artykułu był przegląd zadań z części matematyczno-przyrodniczej w latach 2002-2011 pod kątem funkcjonowania w zależności od płci ucznia. Opisana została metoda analizy DIF, opierająca się na teście Mantel-Haenshel oraz klasyfikacja efektu DIF, która daje czytelną informację dla konstruktorów testów o sposobie funkcjonowania zadania. Należy pamiętać jednak, iż

analiza DIF jest jedynie narzędziem pozwalającym zidentyfikować zadania, które funkcjonują odmiennie w analizowanych grupach. Wykrycie takich zadań nie jest jednak jednoznaczne ze stwierdzeniem stronniczości zadania lub - w przypadku wykrycia w teście wielu takich zadań - stronniczości całego testu. Kolejnym etapem jest analiza jakościowa takich zadań i tu wielką rolę odgrywają autorzy zadań, którzy mogą je ocenić pod kątem merytorycznym. Z tego powodu autorzy artykułu nie podjęli się próby klasyfikacji zadań z efektem DIF ani wyjaśnienia przyczyn takiego zróżnicowania, ograniczając się do zaprezentowania przykładów takich zadań. Należy pamiętać, iż przyczyna zróżnicowanego funkcjonowania zadania może tkwić w jego treści czy konstrukcji lub też może świadczyć o wielowymiarowości testu. W drugim przypadku oznacza to, iż zadanie mierzy jakąś inną umiejętność, z którą radzi sobie lepiej jedna z grup. Niezależnie od przyczyny kwestia ta wymaga głębszej refleksji i jest bardzo istotna zarówno dla systemu egzaminów zewnętrznych, jak również uczniów, nauczycieli i wszystkich osób związanych z edukacją.

Bibliografia:

1. Agresti A. (2002), *Categoirical Data Analysis*, John Wiley & Sons, New Jersey.
2. Dorans N.J., Holland P.W (1993), *DIF detection and description: Mantel-Haenszel and standardization* **[w:]** *Differential Item Functioning*. Edited by Holland PW, Wainer H. Hillsdale, New Jersey: Lawrence Erlbaum Associates:35-66.
3. Grudniewska M., Kondratek B. (2012), *Test Mantel-Haenshel oraz modelowanie IRT jako narzędzia służące do wykrywania DIF oraz opisu jego wielkości na przykładzie zadań ocenianych dychotomicznie* [w druku].
4. Holland P.W., Weiner H. (1993), *Differential item functioning* New Jersey: Lawrence Erlbaum Associates, Hillsdale.
5. Hornowska E. (1999), *Stronniczość testów psychologicznych: problemy - kierunki - kontrowersje*, Poznań: Wydawnictwo Fundacji Humaniora.
6. Mantel N., Haenshel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*. 22 (4), s.719-48.
7. Michaelides P.M. (2008) An Illustration of a Mantel - Haenszel Procedure to Flag Misbehaving Common Items in Test Equating, *Practical Assessment, Research & Evaluation* 13 (7).
8. Radhakrishna S. (1965). Combination of Results from Several 2 x 2 Contingency Tables. *Biometrics*. 21(1), s. 86-98.
9. *Różnice w wynikach nauczania a płeć uczniów. Obecna sytuacja i działania podejmowane w Europie,* Eurydice, Agencja Wykonawcza ds. Edukacji, Kultury i Sektora Audiowizualnego 2010.
10. Swaminathan H, Rogers J.H., (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*. 27(4), s. 361-370.
11. Zieky M. (2003), *A DIF Primer*, Educational Testing Service.