Edukacja 2013, 2(122), 34-55
ISSN 0239-6858

Test Mantel–Haenshel oraz modelowanie IRT jako narzędzia wykrywania DIF i opisu jego wielkości na przykładzie zadań ocenianych dychotomicznie

Bartosz Kondratek*, Magdalena Grudniewska*

Artykuł porównuje dwie metody wykorzystywane do identyfikacji zróżnicowanego funkcjonowania zadań (DIF) ocenianych dychotomicznie: nieparametryczne rozwiązanie opierające się na statystyce Mantel–Haenshel (MH) oraz podejście bazujące na teście ilorazu funkcji wiarygodności. Porównanie przeprowadzono na gruncie teoretycznym i za pomocą symulacji. Wyniki symulacji potwierdziły przypuszczenie, że podejście opierające się na statystyce MH jest bardziej czułe na jednorodne efekty DIF, jednak traci moc, gdy wielkość DIF zmienia się w zależności od poziomu zmiennej ukrytej mierzonej testem. Oprócz mocy statystycznej analizowano również specyficzne miary wielkości efektu DIF stosowane w obu metodach: miarę *MH D – DIF*, wykorzystywaną standardowo przez *Educational Testing Service* do klasyfikacji wielkości DIF, oraz różne miary *P – DIF* określone na metryce łatwości zadania.

Słowa kluczowe: zróżnicowane funkcjonowanie zadań, DIF, test Mantel–Haenshel, IRT

Zróżnicowane funkcjonowanie zadania, a ogólniej – pozycji testowej (*Diffentential Item Functioning, DIF*), jest terminem statystycznym określającym zależność wykonania zadania nie tylko od poziomu umiejętności mierzonej danym testem, ale także od przynależności grupowej wykonujących je osób. Weryfikacja pozycji testowych pod kątem występowania DIF stanowi ważny element psychometrycznej analizy testu, ściśle związany z jego trafnością.

Jeżeli przez U_i oznaczymy odpowiedź na zadanie *i*, przez θ – poziom umiejętności mierzonej przez test, a przez G – przynależność grupową, to w najogólniejszej postaci DIF ze względu na przynależność grupową G występuje, gdy (por. Penfield i Camilli, 2007):

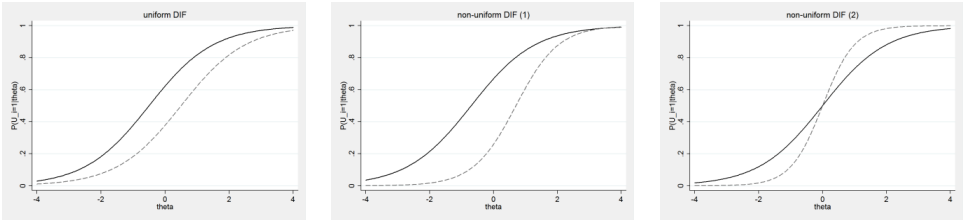
$$U_i|\theta, G \neq U_i|\theta,$$

czyli gdy warunkowy rozkład odpowiedzi na dane zadanie testowe nie zależy wyłącznie od poziomu umiejętności ucznia (θ), ale także od tego, do jakiej grupy (G) on należy. W wypadku zadań ocenianych 0–1 powyższe można zapisać w postaci:

$$P(U_i = 1|\theta, G) \neq P(U_i = 1|\theta),$$

Artykuł powstał w ramach projektu „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego" prowadzonego w Instytucie Badań Edukacyjnych. Projekt jest współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego.

* Pracownia Analiz Osiągnięć Uczniów, Instytut Badań Edukacyjnych. E-mail: m.grudniewska @ibe.edu.pl



Rysunek 1. Przykłady DIF (linią ciągłą zaznaczono $P(U_i = 1|\theta)$ dla G = *r*, przerywaną dla G = *f*).

co znaczy, że prawdopodobieństwo poprawnej odpowiedzi na zadanie U_i zależy nie tylko od θ, ale także od G. Jeżeli G przyjmuje dwie wartości G ∈ {*f*,*r*}, to zróżnicowane funkcjonowanie zadania *i* można również zapisać jako:

$$P(U_i = 1|\theta, G = f) \neq P(U_i = 1|\theta, G = r), \quad (1)$$

co znaczy, że prawdopodobieństwo udzielenia poprawnej odpowiedzi przez ucznia o poziomie umiejętności θ z grupy *f* różni się od prawdopodobieństwa udzielenia poprawnej odpowiedzi przez ucznia o takim samym poziomie umiejętności z grupy *r*.

Na Rysunku 1 przestawiono przykłady zróżnicowanego funkcjonowania zadania zdefiniowanego przez wzór (1). Lewy wykres pokazuje tzw. jednorodny DIF (*uniform* DIF) – krzywa określająca prawdopodobieństwo prawidłowej odpowiedzi dla jednej grupy powstaje przez równoległe przesunięcie krzywej dla drugiej grupy. W innych wypadkach mówi się o niejednorodnym (*non-uniform*) DIF. Na środkowym wykresie zadanie U_i jest łatwiejsze dla grupy *r* na wszystkich poziomach umiejętności (podobnie jak na wykresie z lewej), jednak wielkość DIF zależy od poziomu umiejętności. Interesujący przypadek niejednorodnego DIF przedstawiono z prawej strony – dla uczniów o poziomie umiejętności θ < 0 zadanie U_i jest łatwiejsze w grupie *r*, natomiast dla uczniów o θ > 0 zadanie jest łatwiejsze w grupie *f*.

Pionierskie prace dotyczące analizy DIF pochodzą z początku lat 60. ubiegłego wieku, gdy w Stanach Zjednoczonych uznano potrzebę identyfikowania zadań stronniczych względem grup mniejszościowych. Stąd w analizie DIF klasycznie występuje niesymetryczny podział na dwie grupy – grupę ogniskową (*focal*), na której koncentruje się badanie, oraz grupę odniesienia (*reference*) – odpowiadający podziałowi na grupę mniejszościową i większościową.

Przez stronniczość zadania rozumie się faworyzowanie jednej z grup wskutek odwołania się do czynników treściowo niezależnych od badanej umiejętności. Stronniczość zadania stanowi zatem specyficzne zaburzenie trafności testu i nie jest pojęciem tożsamym z występowaniem DIF. Występowanie DIF świadczy o zależności odpowiedzi na zadanie *i* od dodatkowego czynnika, ponad wspólną dla wszystkich zadań testu umiejętnością θ, którego poziom jest zróżnicowany między grupami G, co jest warunkiem koniecznym, ale niewystarczającym do stwierdzenia stronniczości. Uznanie zadania za stronnicze wymaga eksperckiej analizy treści zadania pod kątem możliwych przyczyn DIF. Może się okazać, że specyficzny dla zadania *i* czynnik powodujący DIF stanowi istotny element uniwersum treści badanej umiejętności, który nie jest reprezentowany w innych zadaniach, nie będąc tym samym zaburzeniem trafności testu niesprawiedliwie faworyzującym jedną z grup (zob. Zieky, 1993).

Warto również odróżniać DIF od międzygrupowych różnic w poziomie umiejętności. Pojęcie DIF w samej istocie ma na celu rozdzielenie faktycznych różnic w poziomie umiejętności uczniów między grupami i różnic w funkcjonowaniu zadania wynikających z innych czynników niż mierzona całym testem umiejętność. Pojawiające się w definicji warunkowanie ze względu na θ wskazuje, że analiza DIF odbywa się przy kontroli międzygrupowych różnic w poziomie umiejętności.

Zgodnie z tym, co napisano powyżej, wnioskujemy, że detekcja DIF dla zadań ocenianych dychotomicznie będzie wymagała analizy łatwości zadania w zależności od przynależności grupowej uczniów przy kontroli ich poziomu umiejętności. Operacyjnie poziom umiejętności jest zazwyczaj określany „wewnętrznie" jako jakaś forma wyniku uzyskiwanego w całym teście. Naturalnym i historycznie pierwszym rozwiązaniem tak postawionego problemu DIF było zastosowanie podejścia opierającego się na popularnym w badaniach klinicznych teście Mantel–Haenshel (MH), pozwalającym na statystyczną analizę różnic w rozkładzie dwuwartościowej zmiennej zależnej między dwoma grupami ustratyfikowanymi ze względu na istotną dla zmiennej zależnej zmienną uboczną. Test MH nazywany bywa również testem Cochran–Mantel–Haenshel, w celu podkreślenia zasług Williama Cochrana, który wcześniej zaproponował bardzo podobne rozwiązanie (Agresti, 2002). Alternatywne podejście do analizy DIF, jakie zostanie przedstawione w niniejszym artykule, pojawiło się wraz z gwałtownym rozwojem w ostatnich dekadach ubiegłego wieku modeli IRT (*Item Response Theory*), w których zależność między poziomem umiejętności a odpowiedzią na zadanie jest modelowana explicité.

Artykuł rozpocznie przedstawienie obu metod analizy DIF wraz z miarami wielkości efektu DIF, jakie można za ich pomocą skonstruować. Ponieważ określenie praktycznego znaczenia różnic w funkcjonowaniu zadania jest nie mniej istotne od wykrycia statystycznie znaczących różnic, wywód w dużej mierze będzie podporządkowany właśnie określeniu wielkości efektu DIF. Następnie zostaną opisane wyniki badań symulacyjnych ilustrujących działanie dwóch metod w różnych warunkach. Na zakończenie zostanie przeprowadzona dyskusja dotycząca wyników.

Analiza DIF na podstawie testu Mantel–Haenshel

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 2 × 2 × M, gdzie M jest liczbą kategorii punktowych wyniku sumarycznego. Prawdopodobieństwo zaobserwowania danej odpowiedzi na rozpatrywane zadanie w zależności od przynależności grupowej oraz od kategorii punktowej m oznaczymy w następujący sposób:

		Odpowiedź na zadanie		Razem
		1	0	
Grupa	f	p_{1fm}	p_{0fm}	p_{fm}
	r	p_{1rm}	p_{0rm}	p_{rm}
Razem		p_{1m}	p_{0m}	p_m

Test MH klasycznie jest opisywany w języku ilorazu szans. Szansą (*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:

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

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

$$H_0: \alpha_m = 1 \qquad m \in \{1, ..., M\},$$
$$H_1: \alpha_m = \alpha \neq 1 \qquad m \in \{1, ..., M\}.$$

Hipoteza zerowa stanowi zatem, 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, w jakiej został zdefiniowany DIF we wzorze (1):

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

Oznacza to, że prawdopodobieństwo udzielenia odpowiedzi poprawnej na zadanie nie zależy od przynależności grupowej, jeżeli uwzględnimy wynik w całym teście. Specyficzna dla testu MH jest hipoteza alternatywna, względem której jest testowana H_0. Wedle H_1 testu MH – 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 (*common odds ratio*).

Analogicznie do prawdopodobieństw indeksujemy w tablicy kontyngencji liczebności:

		Odpowiedź na zadanie		Razem
		0	1	
Grupa	*r*	N_{0rm}	N_{1rm}	N_{rm}
	f	N_{0fm}	N_{1fm}	N_{fm}
Razem		N_{0m}	N_{1m}	N_m

Statystykę dla testu MH z poprawką na ciągłość przedstawiamy jako:

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

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 rozkład statystyki MH_{χ^2} jest zbieżny do rozkładu χ^2 z jednym stopniem swobody (Dorans i Holland, 1993).

Wykazano (Radhakrishna, 1965), że test MH jest jednostajnie najmocniejszym testem dla hipotezy zerowej o warunkowej niezależności proporcji między grupami, przy prawdziwości hipotezy o stałym ilorazie szans. Jeśli hipoteza o stałym ilorazie szans nie jest prawdziwa, test MH traci moc. 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 (Swaminathan i Rogers, 1990). Ze wzoru (2) 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, 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).

Nathan Mantel i William Haenshel (1959) zaproponowali również estymator wspólnego ilorazu szans w postaci:

$$\alpha_{MH} = \frac{\sum_{m=1}^{M} p_{1rm} p_{0fm} N_m}{\sum_{m=1}^{M} p_{1fm}\, p_{0fm} N_m}, \qquad (3)$$

w którym 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 $\alpha_{MH} < 1$.

Analiza DIF na podstawie testu ilorazu wiarygodności w podejściu IRT

Analizę DIF na podstawie IRT przeprowadzimy na przykładzie dwuparametrycznego modelu logistycznego (2PLM), jednak wnioski łatwo można uogólnić również na inne modele, w tym dla zadań ocenianych na skali wielopunktowej. Zespół Davida Thissena (Thissen, Steinberg i Wainer, 1993) ogólnie przedstawił problem testowania DIF w modelowaniu IRT. Inne niż omawiane w niniejszym artykule metody wykorzystywane do analizy DIF przedstawiają Randall D. Penfield i Gregory Camilli (2007).

W podejściu IRT zależność między prawdopodobieństwem udzielenia poprawnej odpowiedzi na zadanie U_n a poziomem umiejętności ucznia θ, jaka pojawia się w przyjętej definicji DIF (1), jest modelowana w sposób bezpośredni. Prawdopodobieństwo udzielenia poprawnej odpowiedzi na zadanie w modelu 2PLM jest określone przez funkcję logistyczną, która zależy od dwóch parametrów b_n oraz a_n:

$$p_n(\theta) = P(U_n = 1|\theta, a_n, b_n) = \frac{1}{1 + e^{-a_n(\theta - b_n)}} \quad (4)$$

Parametr b_n (zwany parametrem trudności) odpowiada za przesunięcie krzywej logistycznej równolegle do osi θ, natomiast parametr a_n (zwany parametrem dyskryminacji) określa nachylenie tej krzywej. Dzięki tym dwu parametrom model pozwala uchwycić przypadki zarówno jednorodnego, jak i niejednorodnego DIF przedstawione na Rysunku 1.

Pełen model IRT opisuje rozkład prawdopodobieństwa całego wektora odpowiedzi na wszystkie zadania testu $u = (U_1, \ldots, U_n, \ldots, U_N)$, a nie tylko to analizowane ze względu na DIF zadanie *i*. Przyjmijmy zatem skrócony zapis $p_n(\theta)$ dla krzywych charakterystycznych poszczególnych zadań i załóżmy, że wszystkie są postaci (4) z parametrami (a_n, b_n) oraz że $\psi_G(\theta)$ oznacza rozkład umiejętności w grupie $G \in \{f,r\}$. Sytuacja braku DIF przy takich oznaczeniach będzie opisana modelem IRT, w którym prawdopodobieństwo zaobserwowania konkretnego wektora odpowiedzi $u = u$ jest dane całką:

$$P(U = u|G) = \\ = \int \left[\prod_{n \in \{1,\ldots,N\}} p_n(\theta)^{u_n} \left(1 - p_n(\theta)\right)^{1-u_n} \right] \psi_G(\theta) d\theta. \quad (5)$$

Wzięty w nawias kwadratowy iloczyn krzywych charakterystycznych i ich dopełnień do jedynki jest (przy założeniu, że odpowiedzi na zadania są warunkowo niezależne[2]) warunkową funkcją wiarygodności przedstawiającą prawdopodobieństwo zaobserwowania danego wektora odpowiedzi w zależności od poziomu umiejętności θ oraz od parametrów zadań określających funkcje p_n. Jak widać, iloczyn ten nie zależy od przynależności grupowej, a jedyną rzeczą zróżnicowaną międzygrupowo w modelu (5) jest rozkład umiejętności ψ_G, po którym odbywa się całkowanie.

We wzorze (5) przyjmujemy, że parametry krzywych są dla wszystkich zadań takie same w obu grupach. Natomiast model zakładający występowanie DIF dla zadania *i* powstaje przez wprowadzenie dla tego zadania innej pary parametrów dla uczniów z grupy *f* niż dla uczniów z grupy *r* – odpowiednio (a_i^f, b_i^f) oraz (a_i^r, b_i^r). Model zakładający DIF dla zadania *i* ma zatem postać:

[2] Założenie o warunkowej niezależności odpowiedzi na zadania testowe, zwanej również lokalną niezależnością (*local independence*), oznacza, że gdy znany jest poziom umiejętności θ, to odpowiedzi na zadania testowe są względem siebie statystycznie niezależne. Założenie to ma nie tylko bardzo ważne techniczne znaczenie przy estymacji parametrów modelu metodą największej wiarygodności, ale także istotną interpretację teoretyczną. Mianowicie oznacza, że poziom umiejętności *θ* wyjaśnia wszystkie obserwowane współzależności między zadaniami, czyli że test jest jednowymiarowy (Lord i Novick, 1968).

$$P(U=u|G)=\int\left[\prod_{n\in\{1,\ldots,N\}\setminus\{i\}}p_n(\theta)^{u_n}\left(1-p_n(\theta)\right)^{1-u_n}\right]p_i^G(\theta)^{u_i}\left(1-p_i^G(\theta)\right)^{1-u_i}\psi_G(\theta)d\theta. \quad (6)$$

Hipotezę zerową i alternatywną poddawane testowaniu w tym podejściu możemy opisać jako parę:

$$H_0\colon a_i^f=a_i^r\wedge b_i^f=b_i^r$$
$$H_1\colon a_i^f\neq a_i^r\vee b_i^f\neq b_i^r.$$

Do testowania prawdziwości hipotezy zerowej wykorzystuje się standardowy test ilorazu wiarygodności (*likelihood ratio test, LR test*), korzystając z faktu, że model (5) jest zagnieżdżony w modelu (6). Statystyka testowa ma postać:

$$LR=-2ln\left(\frac{L_0}{L_1}\right), \quad (7)$$

gdzie L_0 jest funkcją wiarygodności obliczoną na podstawie oszacowań parametrów modelu (5), a L_1 jest analogiczną funkcją wiarygodności dla modelu (6). Statystyka *LR* ma liczbę stopni swobody równą różnicy liczby parametrów szacowanych w dwóch modelach, co w rozpatrywanym przypadku wynosi 2 (jeden dodatkowy parametr trudności oraz jeden dodatkowy parametr dyskryminacji).

Co można zauważyć, w podejściu IRT testowanie występowania DIF w zadaniach wymaga wykorzystania oprogramowania, które pozwala bezpośrednio modelować różne rozkłady umiejętności dla grupy ogniskowej i dla grupy odniesienia. W przeciwnym wypadku model nie byłby w stanie poprawnie oddzielić różnic w poziomie umiejętności obu grup od różnic w funkcjonowaniu zadania w obu grupach, a to stanowi kwintesencję analizy zróżnicowanego funkcjonowania pozycji testowych.

Miary wielkości efektu DIF i klasyfikacja zadań ze względu na DIF

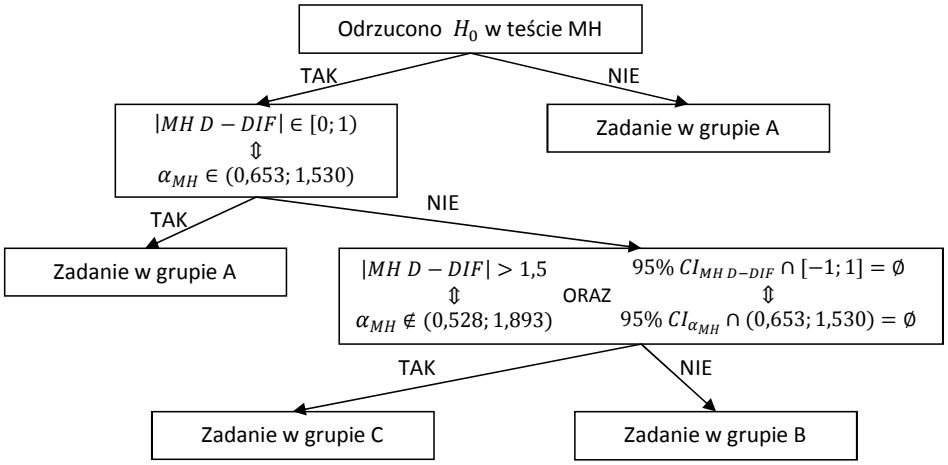
Wspólny iloraz szans statystyki MH (3) stanowi dość trudną w interpretacji miarę wielkości efektu DIF. W celu ułatwienia interpretacji wartości α_{MH} poddaje sie ją różnym przekształceniom. Jednym z nich jest wskaźnik *MH D – DIF* uzyskiwany w następujący sposób:

$$MH\,D-DIF=-2.35\ln[\alpha_{MH}]. \quad (8)$$

Takie przekształcenie pozwala uzyskać rozkład symetryczny, z wartościami z zakresu od -∞ do +∞. Wartość 0 oznacza brak efektu DIF.

Educational Testing Service (ETS) opracował system klasyfikacji efektu DIF, który opiera się na istotności statystyki MH_{χ^2} (przyjmuje się standardowy próg istotności statystycznej α = 0,05) oraz wielkości miary *MH D – DIF*. Na podstawie tych parametrów zadania są przypisywane do trzech kategorii: A, B i C (Dorans i Holland, 1993; Zieky, 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;
- 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-procentowy przedział ufności wokół *MH D– DIF* nie znajduje się poza przedziałem od -1 do +1;
- kategoria C – gdy 95-procentowy przedział ufności wokół *MH D – DIF* znajduje się poza przedziałem od -1 do +1



Rysunek 2. Drzewo decyzyjne klasyfikacji zadań ze względu na DIF na podstawie miary *MH D – DIF*.

oraz absolutna wartość *MH D – DIF* jest większa od 1,5 (w szczególności oznacza to pozytywny wynik testu MH).

Opisane reguły podziału na klasy A, B oraz C zestawiono schematycznie na Rysunku 2, w którym również zamieszczono wartości α_{MH} odpowiadające wartościom *MH D – DIF*, gdyż powszechnie dostępne programy statystyczne raportują wyniki testu MH właśnie na skali „surowego" ilorazu szans α_{MH}.

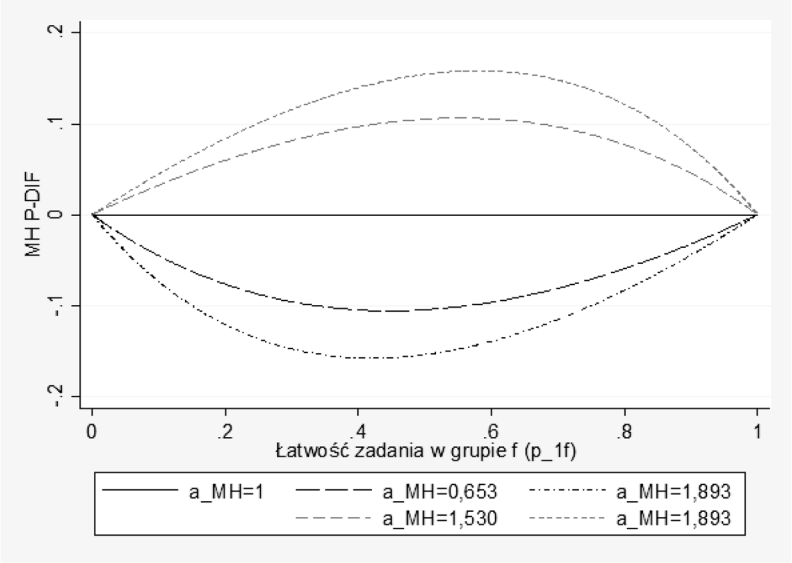
Zadania z kategorii C wymagają od konstruktorów zwrócenia szczególnej uwagi na stronniczość. Informacji o kategorii DIF, do której należy dane zadania, towarzyszy informacja, czy zadanie jest trudniejsze dla grupy ogniskowej (zadania oznaczone „–"), czy dla grupy odniesienia (zadania oznaczone „+").

Należy zauważyć, że współczynnik *MH D – DIF* przekształca rozkład α_{MH} do postaci bardziej symetrycznej i umożliwia stworzenie reguł decyzyjnych przy analizie wielkości DIF, ale sam przez się nadal nie dostarcza jasnej ilościowej interpretacji faktycznej wielkości DIF. Naturalną miarą DIF wydaje się skala łatwości zadania – o ile zadanie *i* byłoby łatwiejsze (trudniejsze) w grupie *f*, gdyby funkcjonowało w niej tak jak funkcjonuje w grupie *r*. Grupę miar wyrażających wielkości efektu DIF na skali łatwości zadania będziemy oznaczać w artykule poprzez symbol *P – DIF*, poprzedzony dodatkowym przedrostkiem.

Żeby zanalizować zależność między opisanymi kategoriami DIF opierającymi się na wyniku testu MH a różnicą w łatwości zadania między grupami przy kontroli poziomu umiejętności zauważmy, że dla każdej kategorii punktowej *m* (zatem przy kontroli poziomu umiejętności) prawdopodobieństwo udzielenia odpowiedzi poprawnej p_{1rm} można wyrazić za pomocą α_m oraz p_{1fm} w następujący sposób:

$$p_{1rm} = \frac{\alpha_m p_{1fm}}{1 - p_{1fm} + \alpha_m}.$$

Przy prawdziwości hipotezy o stałości ilorazów szans α_m, możemy oszacować, jakie byłoby prawdopodobieństwo udzielenia odpowiedzi poprawnej na rozpatrywane zadanie przez uczniów z grupy *f*, gdyby funkcjonowało ono w tej grupie tak samo jak w *r*:



Rysunek 3. Wartość *MH P – DIF* w zależności od łatwości zadania w grupie *f*.

$$p_{1r}^{\dagger} = \frac{\alpha_{MH} p_{1f}}{1 - p_{1f} + \alpha_{MH}}$$

i ostatecznie szukana różnica w łatwości zadania w kontekście grupy f na podstawie wartości statystyki MH przyjmuje postać (por. Dorans i Holland, 1993):

$$MH\ P - DIF = p_{1f} - p_{1r}^{\dagger}. \qquad (9)$$

Na Rysunku 3 przestawiono, jak wynikająca ze zróżnicowanego funkcjonowania różnica w łatwości zadania opisana przez miarę *MH P – DIF* (9) zależy od łatwości zadania w grupie ogniskowej p_{1f} oraz od granicznych wartości współczynnika α_{MH} pojawiających się przy klasyfikacji wielkości DIF przestawionej na Rysunku 2. Należy zwrócić uwagę na dwie prawidłowości. Po pierwsze, graniczne wartości α_{MH} oraz 95-procentowych przedziałów ufności wokół α_{MH} określające przejście między kategoriami A, B oraz C zależą od łatwości zadania w grupie *f* – dla zadań o przeciętnej łatwości konieczna jest większa absolutna różnica w łatwości zadania wynikająca z DIF, niż dla zadań o bardziej skrajnych poziomach łatwości. Po drugie, występuje niesymetryczne traktowanie DIF na korzyść grupy *f* (dodatnie wartości *MH P – DIF*) i na korzyść grupy *r* (ujemne wartości *MH P – DIF*). Ten brak symetrii wynika z przyjęcia symetrycznego kryterium ±1 lub ±1,5 wokół miary *MH D – DIF* (Rysunek 2) przy podejmowaniu decyzji o przynależności zadania do kategorii A–C, a zgodnie ze wzorem (8) *MH D – DIF* jest nieliniowym przekształceniem α_{MH}. Przyjmując, że adekwatną miarą efektu DIF jest wynikająca z DIF oczekiwana różnica w wyniku w zadaniu (i w konsekwencji w całym teście), należy obie obserwacje uznać za wady przedstawionej klasyfikacji ETS. Należy jednak zauważyć, że w przedziale 0,25–0,75, w którym znajdzie się większość zadań prawidłowo skonstruowanego testu, progi wyznaczone przez graniczne wartości α_{MH} są na zbliżonym poziomie.

Alternatywnie do wyrażonej wzorem (5) miary *MH P – DIF* różnicę między ła-

twością zadania w grupie *f* a łatwością, jakie zadanie miałoby w grupie *f*, gdyby funkcjonowało tak jak w grupie *r*, można oszacować bez odwoływania się do wspólnego ilorazu szans w następujący sposób:

$$STD\ P-DIF=\frac{\sum_{m=1}^{M}N_{fm}\left(p_{1fm}-p_{1rm}\right)}{\sum_{m=1}^{M}N_{fm}}. \quad (10)$$

Miara (10) stanowi zatem średnią z różnic łatwości zadania w każdej kategorii punktowej *m*, ważoną przez liczbę uczniów z grupy *f* wpadających do kategorii punktowej *m*. Neil Dorans i Paul Holland (1993), analizując zależności między *MH P – DIF* oraz *STD P – DIF* zauważają, że są one oszacowaniem tej samej wielkości, tj. warunkowej różnicy w łatwości zadania, przy czym różnią się sposobem, w jaki jest ona obliczana. W *MH P – DIF* różnica na metryce *p* jest obliczana przez odwołanie do warunkowego wspólnego ilorazu szans α_{MH}, natomiast w *STD P – DIF* – przez uśrednienie warunkowych różnic w łatwości. W konsekwencji wagi przypisywane każdej z kategorii punktowych *m* różnią się w obu podejściach (w teście *MH* są one dobrane optymalnie ze względu na statystyczną moc testu) i wartości podawane przez obie miary będą się nieznacznie różnić (Dorans i Holland, 1993).

Od wzoru (10) już tylko jeden krok do wprowadzenia miary efektu DIF na skali łatwości, która opierałaby się na podejściu IRT. Załóżmy, że w grupie *f* prawdopodobieństwo udzielenia poprawnej odpowiedzi na zadanie *i* jest określone funkcją p_i^f, a w grupie *r* funkcją p_i^r, co dla modelu 2PLM oznacza wzór (4) odpowiednio z parametrami (a_i^f, b_i^f) oraz (a_i^r, b_i^r). Naturalna miara efektu dla IRT, którą oznaczymy jako *IRT P – DIF* (por. wzór *T*(1) u Wainera, 1993), ma postać:

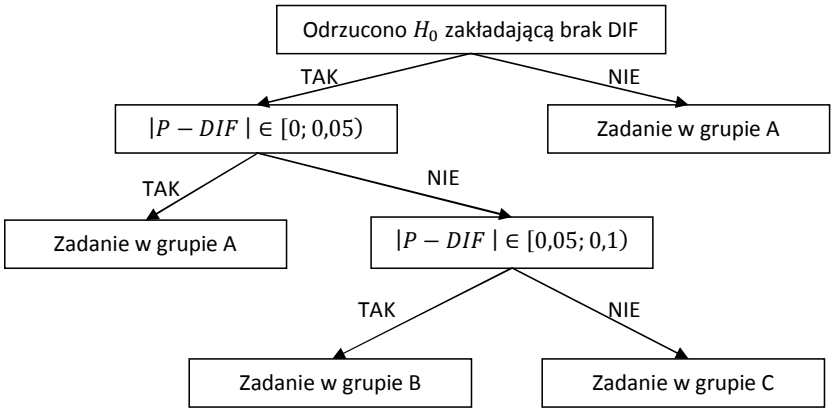
$$IRT\ P-DIF=\int\left[p_i^f(\theta)-p_i^r(\theta)\right]\psi_f(\theta)d\theta. \quad (11)$$

Wzór (11) w sposób jednoznaczny wyraża różnicę między łatwością zadania *i* w populacji *f* a łatwością, jaką miałoby to zadanie w populacji *f*, gdyby funkcjonowało w niej zgodnie z tymi parametrami, z jakimi funkcjonuje w populacji *r*. Należy zauważyć, że miarę *STD P – DIF* określoną wzorem (10) można uznać za nieparametryczną wersję *IRT P – DIF* (11) – w pierwszym wypadku całkowanie odbywa się po rozkładzie dyskretnego sumarycznego wyniku w teście podzielonego na *m* kategorii, w drugim po ciągłym rozkładzie ukrytej zmiennej umiejętności θ.

Mając na względzie krytyczne uwagi na temat klasyfikacji DIF opierającej się na wielkości *MH P – DIF* lub α_{MH}, którą schematycznie przedstawiono na Rysunku 2, można zaproponować alternatywną klasyfikację na podstawie miary *P – DIF*, przyjmując kryteria zespołu Patricka Monahana (Monahan, McHorney, Stump i Perkins, 2007):

- kategoria A – gdy test weryfikujący statystyczną istotność DIF dał wynik negatywny albo gdy wynik testu jest pozytywny, ale absolutna wartość *P – DIF* jest mniejsza od 0,05;
- kategoria B – gdy DIF jest statystycznie istotny oraz absolutna wartość *P – DIF* znajduje się w przedziale od 0,05 do 0,1;
- kategoria C – gdy DIF jest statystycznie istotny oraz absolutna wartość *P – DIF* wykracza poza przedział 0,1.

Na Rysunku 4 przedstawiono odpowiedni schemat dla tej kategoryzacji, analogiczny jak dla *MH D – DIF* na Rysunku 2. Pierwszą rzeczą, jaką można zauważyć jest ogólne, tj. nieodwołujące się do testu wykrywającego istotność statystyczną DIF, sformułowanie tej klasyfikacji. Klasyfikacja ta zatem mogłaby być stosowana zarówno po przeprowadzeniu testu MH i wykorzystania miary *MH P – DIF* (9) lub *STD P – DIF* (10), jak



Rysunek 4. Drzewo decyzyjne klasyfikacji zadań ze względu na DIF na podstawie miary *P – DIF*.

i po przeprowadzeniu testu LR i odwołania się do *IRT P – DIF* (11). Drugą istotną właściwością jest nieodwoływanie się do precyzji oszacowania *P – DIF*, w porównaniu z rozpatrywaniem 95-procentowych przedziałów ufności wokół *MH D – DIF* w poprzedniej klasyfikacji.

Nierozpatrywanie przedziałów ufności należy uznać za wadę tego podejścia, którą można by naprawić przez uwzględnienie błędu standardowego dla oszacowań *P – DIF*. Wyrażenie na błąd standardowy *STD P – DIF* można znaleźć u Doransa i Hollanda (1993), natomiast problem oszacowania błędu standardowego *IRT P – DIF* wydaje się zagadnieniem trudniejszym, wymagającym zapewne odwołania się do technik symulacyjnych.

Badanie symulacyjne

W celu porównania metody wykrywania DIF na podstawie testu MH z metodą opierającą się na IRT, przeprowadzono eksperyment Monte Carlo. Zgodnie z modelem IRT danym wzorem (6) generowano dane dla testu składającego się z $N = 20$ zadań, których krzywe charakterystyczne były zgodne z 2PLM (4). Zadania o numerach od 1 do 19 miały w obu populacjach *f* oraz *r* takie same parametry, tj. były zadaniami bez DIF. Zadania bez DIF miały parametry trudności b_n symetrycznie rozłożone wokół 0 i dobrane tak, że odpowiadały w przybliżeniu centylom: 5., 10., … 95. standardowego rozkładu normalnego $N(0;1)$, a wartości parametrów dyskryminacji przyjmowały naprzemiennie wartości 1 oraz 1,5. W ten sposób zestaw 19 zadań bez DIF tworzył „test", którego informatywność była optymalnie dopasowana do pomiaru umiejętności uczniów z rozkładu $N(0;1)$. Parametry wspomnianych zadań zebrano w Tabeli 1.

Dla grupy ogniskowej przyjęto standardowy rozkład normalny umiejętności $\psi_f = N(0;1)$. Dla grupy odniesienia przyjęto rozkład o takim samym kształcie, ale przesunięty o 0,253 $\psi_r = N(0{,}253;1)$, co odpowiada sytuacji, w której średni poziom umiejętności w grupie *r* przypada na 60. centyl poziomu umiejętności grupy *f*. Zadanie wykrywania DIF w symulacji było zatem przeprowadzane w sytuacji istotnej różnicy w poziomie umiejętności między grupami, na korzyść grupy odniesienia.

Manipulacji w przeprowadzonym eksperymencie Monte Carlo poddano parametry

Tabela 1
Zastosowane w symulacjach parametry 19 zadań bez DIF

n	b_n	a_n	n	b_n	a_n	n	b_n	a_n
1	-1,65	1	10	0	1,5	11	1,65	1
2	-1,28	1,5				12	1,28	1,5
3	-1,04	1				13	1,04	1
4	-0,84	1,5				14	0,84	1,5
5	-0,68	1				15	0,68	1
6	-0,52	1,5				16	0,52	1,5
7	-0,39	1				17	0,39	1
8	-0,25	1,5				18	0,25	1,5
9	-0,13	1				19	0,13	1

zadania o numerze 20 – było to zadanie, dla którego testowano występowanie DIF w obu podejściach. W grupie *f* rozpatrzono tylko dwa zestawy parametrów (a_{20}^f, b_{20}^f):

- a_{20}^f= 1,5 oraz b_{20}^f= 0, czyli sytuację, w której łatwość zadania 20 w populacji *f* wynosi 0,50;
- a_{20}^f= 1,5 oraz b_{20}^f = -0,79163, czyli sytuację, w której zadanie 20 jest łatwiejsze – odpowiada temu łatwość 0,70 w *f*.

Parametry zadania 20 w grupie ogniskowej zróżnicowano w większym stopniu, aby krzyżując je z dwoma powyższymi sytuacjami, uzyskać możliwe szerokie spektrum efektów DIF. Parametr dyskryminacji a_{20}^r przyjmował trzy wartości:

- a_{20}^r = 1, co oznacza niejednorodny DIF, ze względu na mniejsze nachylenie krzywej charakterystycznej w grupie *r* niż w *f*;
- a_{20}^r = 1,5, co oznacza jednorodny DIF;
- a_{20}^r = 2, co oznacza niejednorodny DIF, ze względu na większe nachylenie krzywej charakterystycznej w grupie *r* niż w *f*.

Parametry trudności b_{20}^r natomiast dobrano w taki sposób, żeby uzyskać określone łatwości zadania na parametrach (a_{20}^r, b_{20}^r) w populacji *f*, i tym samym określone wielkości efektu *IRT P – DIF* (11). Konkretnie, wykorzystując metodę bisekcji połączoną z całkowaniem Monte Carlo, parametry b_{20}^r dobrano tak, aby całka (por. wzór (11)):

$$\int [p_{20}^f(\theta) - p_{20}^r(\theta)]\, \psi_f(\theta) d\theta$$

przyjmowała w równych odstępach dziewięć wartości od -0,150 do 0,050.

Ostatecznie analizie poddano 2 × 3 × 9 warunków eksperymentalnych, które sumarycznie zebrano w Tabeli 2. Dla warunków w każdym polu Tabeli 2 przeprowadzono 10 000 niezależnych replikacji, losując w każdej replikacji po 1000 wektorów odpowiedzi dla uczniów grupy *f* i tyle samo dla uczniów z grupy *r*. Przy każdej replikacji:

- przeprowadzono test MH sprawdzający DIF dla zadania 20, stratyfikując wyniki ze względu na sumaryczny wynik w całym teście. Oszacowano $\widehat{\alpha_{MH}}$ oraz granice 95-procentowego przedziału ufności wokół tej statystyki. Wykorzystano do tego celu procedurę *cc* (*case-control*) dostępną w programie STATA;

Tabela 2
Zestawienie wartości poddanych manipulacji w badaniach symulacyjnych; „nu" – non uniform (niejednorodny) DIF, „u" – uniform (jednorodny) DIF, (-) – zadanie trudniejsze dla grupy f, (+) – zadanie łatwiejsze dla grupy f

IRT P – DIF	Łatwość zadania 20 w *f:* 0,5			Łatwość zadania 20 w *f:* 0,7		
	a^r_{20} = 1	a^r_{20} = 1,5	a^r_{20} = 2	a^r_{20} = 1	a^r_{20} = 1,5	a^r_{20} = 2
-0,15	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
-0,125	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
-0,1	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
-0,075	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
-0,05	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
-0,025	nu(-)	u(-)	nu(-)	nu(-)	u(-)	nu(-)
0	nu(0)	brak DIF	nu(0)	nu(0)	brak DIF	nu(0)
0,025	nu(+)	u(+)	nu(+)	nu(+)	u(+)	nu(+)
0,05	nu(+)	u(+)	nu(+)	nu(+)	u(+)	nu(+)

- przeprowadzono test LR sprawdzający DIF dla zadania 20. Dopasowanie modelu IRT bez DIF (5) oraz z DIF (6) przeprowadzono z wykorzystaniem oprogramowania MIRT (Glas, 2010);
- oszacowano $\widehat{MH\ P-DIF}$ (9), $\widehat{STD\ P-DIF}$ (10) oraz $\widehat{IRT\ P-DIF}$ (11);
- dokonano klasyfikacji wielkości DIF na trzy kategorie zgodnie ze schematem na Rysunku 2 oraz ze schematem na Rysunku 3, przy czym w drugim wypadku oparto się na wartościach $\widehat{IRT\ P-DIF}$ oraz wyniku testu LR.

Zasadniczym celem badania było:
- porównanie czułości testów MH oraz LR w warunkach różnej wielkości DIF oraz różnego typu DIF (jednorodny vs. niejednorodny);
- porównanie zdolności trzech estymatorów *P – DIF* (*MH*, *STD*, *IRT*) wielkości efektu DIF na skali łatwości zadania, przy założeniu, że prawdziwa wielkość efektu jest dana przez IRT *P – DIF*, zgodnie z którym wygenerowano dane (Tabela 2);
- porównanie dwóch klasyfikacji wielkości efektu DIF na kategorie A, B , C.

Wyniki symulacji

Pierwszym celem badania było porównanie czułości testów MH oraz LR w warunkach różnej wielkości DIF oraz różnego typu DIF (jednorodny vs. niejednorodny). W Tabeli 3 przedstawiono procent przypadków, w których metoda Mantel–Haenshel oraz metoda opierająca się na IRT dała wynik istotny statystycznie w zależności do warunków eksperymentalnych, wartości te również przedstawiono na wykresie na Rysunku 5.

Zauważmy, że w wypadku znacznej różnicy w efekcie DIF na metryce łatwości (IRT *P – DIF*) z zakresu -0,15 do -0,10 obie metody prawie we wszystkich przypadkach wskazały zgodnie istotny statystycznie DIF.

Zwróćmy uwagę, że test Mantel–Haenshel wykrywa więcej przypadków DIF, gdy a^r_{20} = 1,5 (jednorodny DIF), natomiast a^r_{20} test LR jest bardziej wrażliwy na wykrywanie efektu DIF, gdy a^r_{20} = 1 oraz a^r_{20} = 2 (niejednorodny DIF). Wyniki badania potwierdzają zatem wspomnianą wcześniej prawidłowość, że test Mantel–Haenshel jest najmocniejszy w wypadku jednorodnego

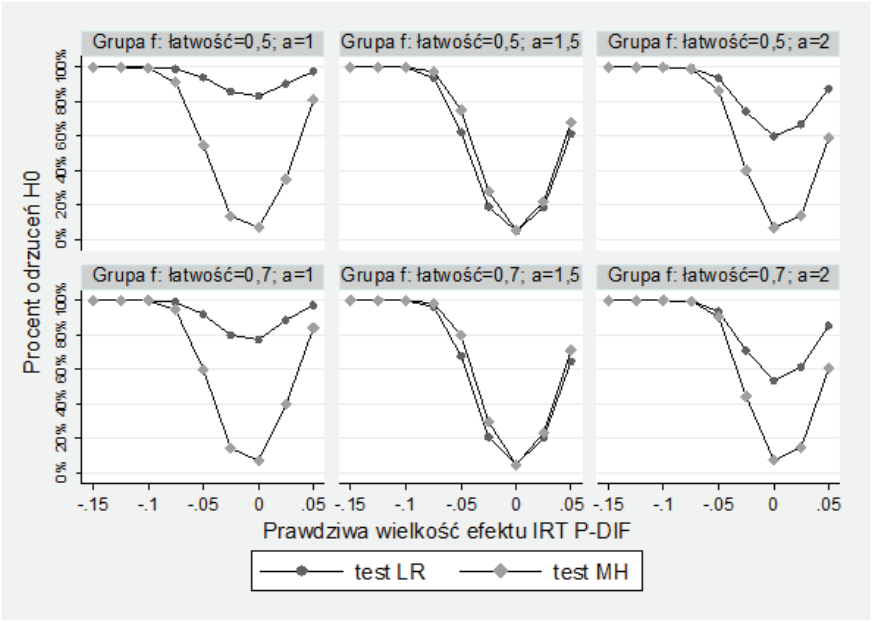
Tabela 3
Porównanie procentowe liczby przypadków, w których wynik testu Mantel–Haenshel oraz Likelihood Ratio jest istotny statystycznie

IRT *P – DIF*	Statystyka	Łatwość zadania 20. w *f*: 0,5			Łatwość zadania 20. w *f*: 0,7		
		a^r_{20} = 1	a^r_{20} = 1,5	a^r_{20} = 2	a^r_{20} = 1	a^r_{20} = 1,5	a^r_{20} = 2
-0,15	MH	100	100	100	100	100	100
	LR	100	100	100	100	100	100
-0,125	MH	100	100	100	100	100	100
	LR	100	100	100	100	100	100
-0,1	MH	99	100	100	100	100	100
	LR	100	100	100	100	100	100
-0,075	MH	91	97	99	95	99	100
	LR	99	94	100	99	96	100
-0,05	MH	54	75	86	60	80	90
	LR	94	62	93	92	68	94
-0,025	MH	14	28	40	14	30	44
	LR	86	19	74	80	21	71
0	MH	7	5	7	7	5	8
	LR	83	5	60	77	5	54
0,025	MH	35	22	14	40	23	15
	LR	90	18	67	89	20	61
0,05	MH	81	68	59	84	71	61
	LR	98	61	87	97	65	86

DIF oraz potwierdzają hipotezę, że test LR jest bardziej czuły na wykrywanie niejednorodnych różnic między grupami.

W wypadku niejednorodnego efektu DIF wrażliwość metod jest zależna od mocy dyskryminacyjnej oraz kierunku efektu DIF. Test MH jest mocniejszy, gdy moc dyskryminacyjna jest wyższa i efekt mniejszy bądź równy -0,05 oraz gdy moc dyskryminacyjna jest niższa i efekt większy lub równy -0,025. Test LR natomiast w większości analizowanych przypadków był mocniejszy, gdy dyskryminacja była niższa. Nie zmienia to jednak wcześniej wspomnianej prawidłowości, że LR jest bardziej czuły niż MH dla niejednorodnego DIF.

Jeszcze jedną ciekawą zmienną różnicującą czułość dwóch metod jest łatwość zadania w grupie ogniskowej. W przypadku jednorodnego DIF oba testy są bardziej wrażliwe na wykrywanie międzygrupowych różnic, gdy zadanie jest łatwiejsze w grupie *f* (tj. gdy łatwość wynosi 0,7). Natomiast gdy mamy do czynienia z niejednorodnym efektem DIF, prawidłowość ta się zaciera; jest prawdopodobne, że czułość metod zależy tu dodatkowo od wielkości efektu DIF. Bliższe zbadanie tego problemu wymagałoby zwiększenia liczby poziomów wielkości efektu DIF oraz zwiększenia liczby replikacji w eksperymencie.

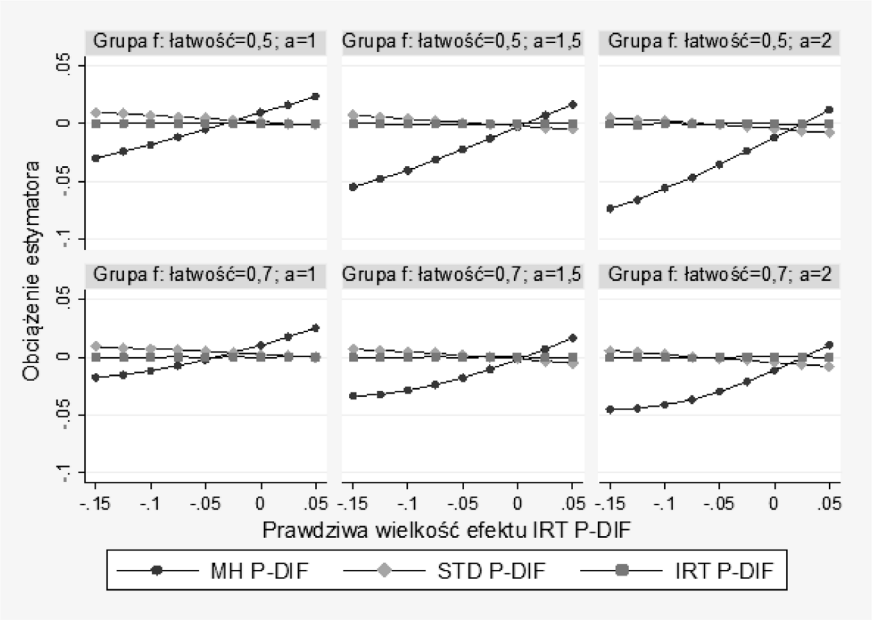


Rysunek 5. Porównanie procentowe liczby przypadków, w których wynik testu Mantel–Haenshel oraz *Likelihood Ratio* jest istotny statystycznie w zależności od warunków eksperymentalnych.

Kolejnym problemem, jaki poddano analizie, była jakość trzech różnych estymatorów szacujących wielkość DIF na skali łatwości zadania, tj.: $MH\widehat{P-DIF}$ (9), $STD\widehat{P-DIF}$ (10) oraz $IRT\widehat{P-DIF}$ (11). Kryterium do oceny jakości tych estymatorów stanowiła prawdziwa wartość *IRT P – DIF*, znana na podstawie parametrów modelu IRT użytych przy generowaniu danych. Na podstawie danych z 10 000 replikacji obliczono obciążenie oraz odchylenie standardowe wspomnianych estymatorów.

Na Rysunku 6 przedstawiono obciążenie trzech analizowanych estymatorów efektu DIF w zależności od warunków eksperymentalnych. Widzimy, że estymator opierający się na wspólnym ilorazie szans α_{MH} wykazuje się relatywnie najsilniejszym obciążeniem i jest ono tym większe, im większa jest dyskryminacja zadania w grupie odniesienia, i tym większe, im większa jest prawdziwa wielkość efektu DIF. Dodatkowo kierunek obciążenia jest taki, że $MH\widehat{P-DIF}$ zawyża prawdziwą wielkość efektu co do jej bezwzględnej wartości. O wiele mniejsze obciążenie obserwujemy dla $STD\widehat{P-DIF}$, nie wykazuje ono istotnego zróżnicowania w zależności od parametru a^r_{20}, a jedynie ze względu na wielkość prawdziwego efektu DIF, przy czym odwrotnie jak poprzednio – oszacowanie $STD\widehat{P-DIF}$ jest obciążone w stronę zera (zaniża bezwzględną wartość estymowanego parametru). Trzeci z analizowanych estymatorów, $IRT\widehat{P-DIF}$ praktycznie nie wykazuje obciążenia[2] dla wszystkich analizowanych warunków eksperymentalnych.

[2] Obciążenie $IRT\widehat{P-DIF}$ jest tak niewielkie, że nie ma praktycznego znaczenia. Ponieważ estymatory parametrów modelu IRT są obciążone (Lord, 1983), należy się spodziewać, że również obciążone będą estymatory parametrów obserwowanych wyników zbudowane przez ich przekształcenie, jak powstały poprzez podstawienie estymatorów IRT do wzoru (11) estymator $IRT\widehat{P-DIF}$. Obciążenie rozkładu obserwowanych wyników oszacowanego na podstawie estymatorów IRT zostało zbadane w kontekście zrównywania wyników (Kondratek, 2012).



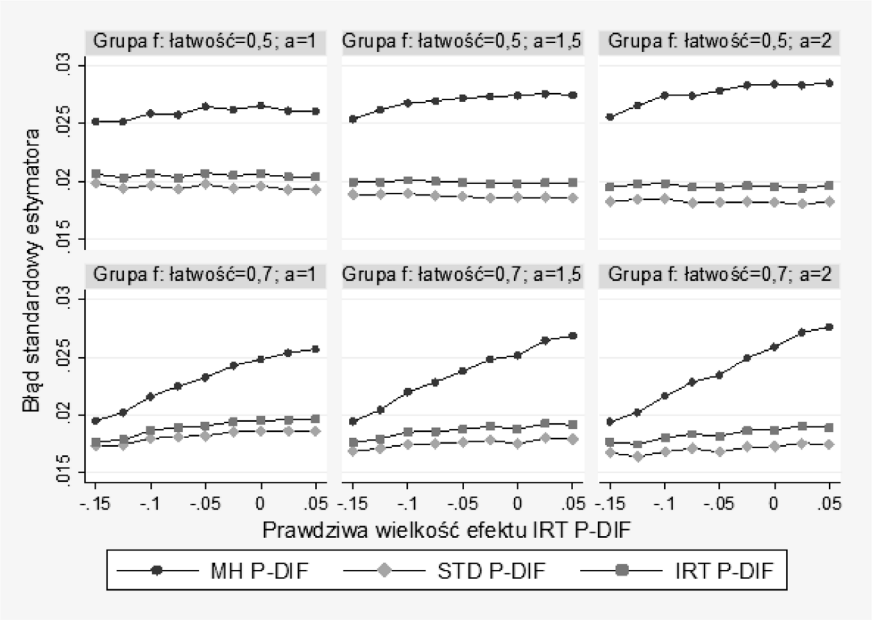
Rysunek 6. Obciążenie estymatorów efektu DIF na skali łatwości zadania w zależności od warunków eksperymentalnych (oś pozioma reprezentuje prawdziwą wartość parametru, oś pionowa – wartość obciążenia).

Na Rysunku 7 przedstawiono wartości odchyleń standardowych trzech estymatorów wielkości efektu DIF, które są jednocześnie oszacowaniem Monte Carlo błędu standardowego tych estymatorów. Widzimy, że estymator $\widehat{MH\,P-DIF}$ ma znacznie większy błąd standardowy niż dwa pozostałe. Błędy standardowe $\widehat{STD\,P-DIF}$ oraz $\widehat{IRT\,P-DIF}$ są do siebie zbliżone, przy czym dla pierwszego estymatora są systematycznie mniejsze. Zaobserwowana zależność między wielkością błędu a typem estymatora jest zapewne konsekwencją różnic w obciążeniu tych estymatorów (Rysunek 5) – im bardziej estymator „ściąga" oszacowania w stronę zera, tym mniejsza jest jego wariancja.

Zestawienie błędów standardowych estymatorów efektu DIF dla wszystkich analizowanych warunków przedstawiono w Tabeli 4. Obserwujemy mniejsze błędy standardowe, gdy łatwość zadania w grupie ogniskowej jest większa. Zauważalne jest natomiast obniżenie błędów standardowych dla estymatorów *STD P – DIF* oraz *IRT P – DIF* wraz ze wzrostem parametru dyskryminacji, przy braku jasnej zależności dla *MH P – DIF*.

Ostatnim celem eksperymentu było porównanie przedstawionych w artykule dwóch klasyfikacji efektu DIF. Zaczniemy od zanalizowania właściwości klasyfikacji opierającej się na *MH D – DIF,* ze względu na jej popularność. Następnie dwie metody zostaną ze sobą zestawione.

W Tabeli 5 pokazano wyniki klasyfikacji w zależności od łatwości zadania w grupie ogniskowej oraz od wartości mocy dyskryminacyjnej w grupie odniesienia dla klasyfikacji opierającej się na *MH D – DIF* (Rysunek 2). Możemy zobaczyć, w jaki sposób zmienia się procent przypadków



Rysunek 7. Błędy standardowe estymatorów efektu DIF na skali łatwości zadania w zależności od warunków eksperymentalnych (oś pozioma reprezentuje prawdziwą wartość parametru, oś pionowa – wartość błędu standardowego).

zaliczonych do poszczególnych kategorii efektu DIF w zależności od zmieniających się parametrów. Należy zwrócić szczególną uwagę na kategorię C, która, jak wcześniej wspomniano, wskazuje na istnienie dużych różnic między grupą ogniskową i grupą odniesienia i stanowi ostrzeżenie dla konstruktorów testów. Zauważmy, że im wyższa moc dyskryminacyjna w grupie ogniskowej, tym więcej przypadków zostaje przypisanych do kategorii C. Podobną zależność obserwujemy w wypadku zmian łatwości zadania z DIF w grupie ogniskowej: odsetek przypadków zaklasyfikowanych do kategorii C jest wyższy, gdy zadanie jest łatwiejsze. Widzimy zatem, że wyniki klasyfikacji są ściśle zależne od mocy dyskryminacyjnej oraz łatwości zadania w grupie ogniskowej, co jest zapewne konsekwencją wcześniej opisanej różnicy w mocy testu MH dla różnych warunków eksperymentalnych.

W Tabeli 6 pokazano szczegółowy rozkład przypadków według klasyfikacji efektu DIF w zależności od warunków eksperymentalnych. Analogiczne informacje przedstawiono na Rysunku 8. Co można zaobserwować na podstawie wykresów, do kategorii C wpada więcej przypadków, gdy łatwość w grupie ogniskowej równa się 0,7, niż przy łatwości równej 0,5 (przy takiej samej różnicy w łatwości zadania między grupą ogniskową i odniesienia). Zależność klasyfikacji ETS od łatwości zadania w grupie ogniskowej była sygnalizowana we wcześniejszej części artykułu.

Na poniższych wykresach widać również, że liczba przypadków zaklasyfikowanych do poszczególnych kategorii zależy od parametru dyskryminacji w grupie odniesienia – im wyższe a_{20}^r, tym większa szansa, że przy określonej wielkości różnicy przypadek zostanie zaliczony do wyższej

Tabela 4
Porównanie błędów standardowych estymatorów efektu DIF na skali łatwości zadania (wartości w tabeli należy pomnożyć przez 0,01)

IRT *P – DIF*	Statystyka	Łatwość zadania 20. w *f*: 0,5			Łatwość zadania 20. w *f*: 0,7		
		a^r_{20} = 1	a^r_{20} = 1,5	a^r_{20} = 2	a^r_{20} = 1	a^r_{20} = 1	a^r_{20} = 1,5
	MH	2,52	2,54	2,56	1,95	1,94	1,94
-0,15	STD	1,98	1,88	1,82	1,72	1,68	1,67
	IRT	2,06	1,99	1,95	1,76	1,75	1,76
	MH	2,52	2,62	2,66	2,02	2,04	2,02
-0,125	STD	1,94	1,88	1,84	1,74	1,70	1,63
	IRT	2,03	1,99	1,97	1,78	1,79	1,74
	MH	2,59	2,68	2,75	2,16	2,20	2,16
-0,1	STD	1,96	1,89	1,85	1,79	1,74	1,68
	IRT	2,06	2,01	1,98	1,86	1,85	1,80
	MH	2,58	2,70	2,74	2,25	2,29	2,28
-0,075	STD	1,93	1,87	1,81	1,81	1,75	1,71
	IRT	2,03	2,00	1,94	1,89	1,85	1,83
	MH	2,65	2,72	2,79	2,32	2,38	2,35
-0,05	STD	1,97	1,86	1,81	1,82	1,76	1,67
	IRT	2,07	1,99	1,95	1,90	1,87	1,81
	MH	2,62	2,73	2,84	2,43	2,48	2,49
-0,025	STD	1,94	1,85	1,82	1,85	1,78	1,72
	IRT	2,05	1,98	1,96	1,94	1,90	1,86
	MH	2,66	2,75	2,85	2,48	2,52	2,59
0	STD	1,96	1,86	1,81	1,86	1,75	1,72
	IRT	2,06	1,98	1,95	1,95	1,87	1,86
	MH	2,61	2,76	2,83	2,54	2,65	2,71
0,025	STD	1,93	1,86	1,80	1,86	1,80	1,75
	IRT	2,04	1,99	1,94	1,96	1,92	1,90
	MH	2,61	2,75	2,85	2,57	2,69	2,76
0,05	STD	1,93	1,86	1,82	1,86	1,79	1,74
	IRT	2,03	1,98	1,96	1,96	1,91	1,89

kategorii w klasyfikacji ETS. Spójrzmy przykładowo na wykres przedstawiający przypadki zaklasyfikowane do kategorii B, gdy łatwość grupie ogniskowej jest równa 0,5. Jak już zauważyliśmy, gdy różnica łatwości zadania między grupą ogniskową i odniesienia jest równa -0,025, większość przypadków zostaje zaklasyfikowana do kategorii A, co znaczy, że metoda nie wykrywa efektu DIF. Zauważmy jednak, że wykrycie efektu DIF jest zależne od wielkości parametru *a*: przy mocy dyskryminacyjnej równej 1 do kategorii B, czyli z efektem DIF, zostaje przypisanych zaledwie 10% przypadków, natomiast gdy moc dyskryminacyjna jest równa 2, w kategorii B znajdzie się około 40% przypadków. Podobną sytuację mamy również w wypadku różnicy łatwości w grupie ogniskowej i odniesienia równej -0,125 – gdy a^r_{20} = 1,

Tabela 5
Liczba przypadków zaliczonych do kategorii A, B, C w zależności od łatwości zadania w grupie ogniskowej oraz od wielkości parametru a^r_{20} w grupie odniesienia [w %]

Łatwość zadania 20. w grupie *f*	Wartość parametru *a*	Klasyfikacja *MH D – DIF*		
		A	B	C
0,5	a^r_{20} = 1	66	19	15
	a^r_{20} = 1,5	59	18	23
	a^r_{20} = 2	55	18	27
0,7	a^r_{20} = 1	58	17	25
	a^r_{20} = 1,5	53	17	31
	a^r_{20} = 2	49	16	35

do kategorii C zostaje zaklasyfikowane 40% przypadków, podczas gdy przy $a = 2$ jest to już 90%.

Na wykresach (Rysunek 8) możemy również zaobserwować wspomnianą we wcześniejszej części niesymetryczność metody MH w wykrywaniu efektu DIF (przedstawioną na Rysunku 3). Gdy prawdziwy efekt DIF jest równy -0,05, do kategorii B zostaje zaklasyfikowane więcej przypadków niż wówczas, gdy różnica jest równa 0,05. Metoda Mantel–Haenshel, na której opiera się klasyfikacja ETS, jest zatem bardziej wrażliwa w sytuacji, gdy zadanie jest łatwiejsze dla grupy odniesienia – wówczas nieznacznie więcej przypadków zostaje przypisanych do kategorii B. Spodziewalibyśmy się natomiast, że metoda będzie również skuteczna w wykrywaniu efektu DIF, zarówno

Tabela 6
Klasyfikacja efektu DIF na podstawie statystyki MH D – DIF w różnych warunkach eksperymentalnych [w %]

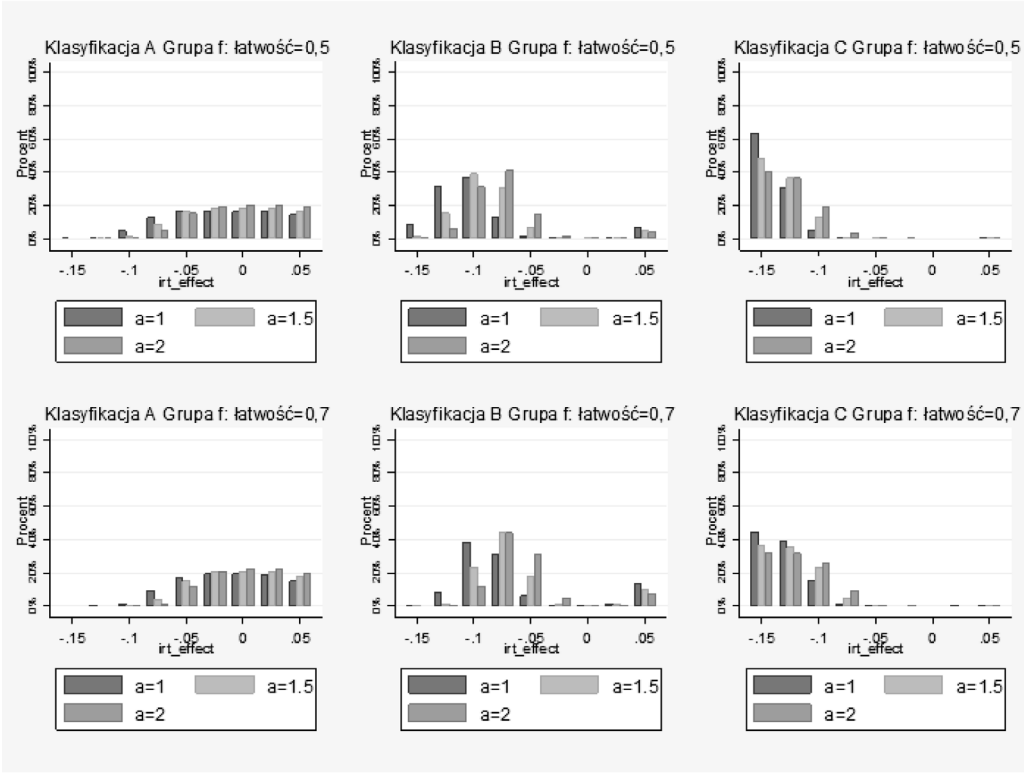
MH D – DIF	Łatwość zadania 20. w *f*: 0,5									Łatwość zadania 20. w *f*: 0,7								
	a^r_{20} = 1			a^r_{20} = 1,5			a^r_{20} = 1			a^r_{20} = 1,5			a^r_{20} = 1			a^r_{20} = 1,5		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
-0,15	0	14	85	0	3	98	0	0	100	0	0	100	0	0	100	0	0	100
-0,125	4	54	42	0	25	75	0	10	90	0	13	87	0	3	97	0	0	100
-0,1	30	63	7	9	63	28	3	49	48	6	60	33	1	35	64	0	17	83
-0,075	77	23	0	47	50	3	26	65	9	48	49	3	20	67	13	7	63	29
-0,05	97	3	0	89	11	0	76	24	0	90	10	0	72	27	1	53	45	2
-0,025	100	0	0	99	1	0	97	3	0	100	0	0	98	2	0	93	7	0
0	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0	100	0	0
0,025	99	1	0	100	0	0	100	0	0	98	2	0	99	1	0	99	1	0
0,05	88	12	0	92	8	0	93	7	0	78	22	0	85	15	0	89	11	0

gdy zadanie będzie faworyzowało grupę odniesienia, jak i grupę ogniskową.

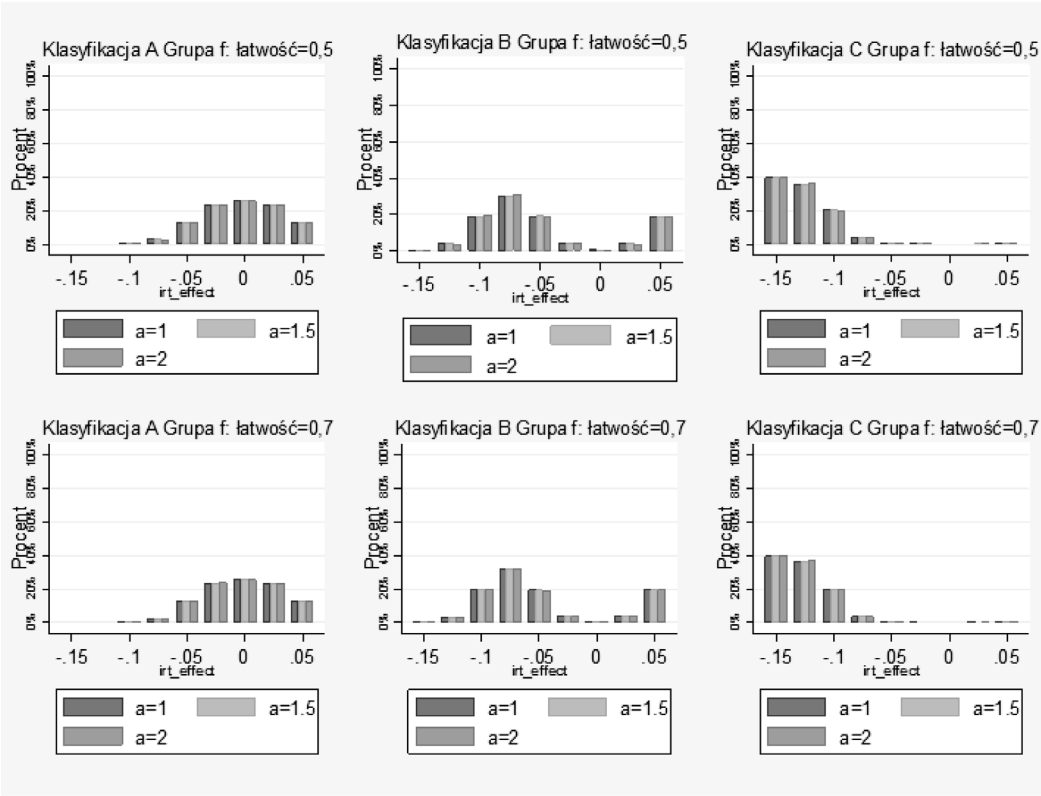
Podsumowując, skuteczność metody opierającej się na teście Mantel–Haenshel w wykrywaniu efektu DIF oraz klasyfikacja *MH D – DIF* są zależne od parametrów zadania. Metoda jest bardziej wrażliwa na różnice między grupami przy większej łatwości zadania oraz przy wyższej mocy dyskryminacyjnej. Należy również podkreślić wspomnianą wcześniej niesymetryczność: metoda Mantel–Haenshel jest bardziej skuteczna w wykrywaniu efektu DIF w wypadku różnic, które faworyzują grupę odniesienia.

Na Rysunkach 8 i 9 przedstawiono rozkład przypadków wpadających do poszczególnych kategorii w zależności od zmieniających się warunków eksperymentalnych. Rozkłady przypadków w klasyfikacji *D – DIF* i *P – DIF* są bardzo zbliżone, co potwierdzają również dane zamieszczone w Tabeli 8, porównującej obie klasyfikacje oraz procent przypadków zaklasyfikowanych zgodnie i niezgodnie według *D – DIF* i *P – DIF*. Należy jednak zwrócić uwagę, że w wypadku klasyfikacji *D – DIF* występują fluktuacje ze względu na łatwość zadania oraz wielkości współczynnika dyskryminacji, natomiast klasyfikacja *P – DIF* jest taka sama niezależnie od warunków eksperymentalnych.

W Tabeli 7 przedstawiono, analogicznie do Tabeli 6, klasyfikację efektu DIF, z tym, że zbudowaną na podstawie oszacowanych war-



Rysunek 8. Klasyfikacja efektu DIF na podstawie statystyki *MH D – DIF* w różnych warunkach eksperymentalnych.



Rysunek 9. Klasyfikacja efektu DIF na podstawie statystyki *IRT P – DIF* w różnych warunkach eksperymentalnych.

tości *IRT P – DIF*. Tak jak w wypadku poprzedniej tabeli, możemy zobaczyć, jak zmienia się procent przypadków przypisanych do poszczególnych kategorii, zależnie od zmieniających się warunków eksperymentalnych. W kolejnym akapicie obie klasyfikacje zostaną porównane, wówczas będzie możliwa ocena, w jakim stopniu są one zgodne.

W Tabeli 8 zestawiono porównanie klasyfikacji efektu DIF zbudowanej na podstawie oszacowań *MH D – DIF* oraz współczynnika *IRT P – DIF*, zależnie od warunków eksperymentalnych. Zauważmy, że obie metody przypisują do tej samej kategorii około 80% przypadków i że procent zgodnych przypisań tylko nieznacznie zależy od warunków symulacji. W odniesieniu do klasyfikacji niezgodnych w tabeli zostały rozróżnione dwie sytuacje: gdy tylko jedna z klasyfikacji wskazuje na istnienie efektu DIF oraz gdy obie klasyfikacje wskazują na istnienie efektu DIF, jednak różnią się co do wielkości tego efektu. Zwróćmy uwagę, że w wypadku klasyfikacji niezgodnych najczęściej mamy do czynienia z sytuacją, gdy na podstawie *D – DIF* zadanie zostaje przypisane do niższej klasy niż na podstawie klasyfikacji *P – DIF*.

Wnioski

Celem artykułu było porównanie dwóch narzędzi służących do wykrywania zróżnicowanego funkcjonowania zadań testo-

Tabela 7
Klasyfikacja efektu DIF na podstawie statystyki IRT P – DIF w różnych warunkach eksperymentalnych [w %]

P – DIF	Łatwość zadania 20. w *f*: 0,5						Łatwość zadania 20. w *f*: 0,7											
	a_{20}^r = 1			a_{20}^r = 1,5			a_{20}^r = 1			a_{20}^r = 1,5			a_{20}^r = 1			a_{20}^r = 1,5		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
-0,15	0	1	99	0	1	99	0	1	99	0	0	100	0	0	100	0	0	100
-0,125	0	11	89	0	11	89	0	10	90	0	8	92	0	8	92	0	8	92
-0,1	1	49	51	0	49	51	1	50	50	0	50	49	0	50	50	0	50	49
-0,075	11	78	11	11	79	10	9	80	10	9	82	9	9	82	9	8	83	9
-0,05	50	50	1	49	50	1	50	49	1	50	50	0	50	50	0	51	49	0
-0,025	89	11	0	89	11	0	90	10	0	90	10	0	91	9	0	91	9	0
0	98	2	0	99	1	0	99	1	0	99	1	0	99	1	0	99	1	0
0,025	89	11	0	90	10	0	90	10	0	90	10	0	90	10	0	90	10	0
0,05	49	50	1	50	49	0	50	49	1	50	50	1	50	49	0	50	50	0

Tabela 8
Porównanie procentowe ilości zgodnych i niezgodnych klasyfikacji efektu DIF według metod MH D – DIF oraz IRT P – DIF [w %]

Warunki symulacji		Klasyfikacje zgodne	Klasyfikacje niezgodne	Klasyfikacje niezgodne					
				P-DIF class=A		D-DIF-class=A,		D-DIF-class=B, P-DIF class=C	D-DIF-class=C, P-DIF class=B
				D-DIF-class=C	D-DIF-class=B	P-DIF class=C	P-DIF class=B		
Łatwość zadania 20 w *f*: 0,5	a_{20}^r =1	64	36	0	0	0	23	13	0
	a_{20}^r =1,5	78	22	0	0	0	16	5	0
	a_{20}^r =2	87	13	0	0	0	12	1	1
Łatwość zadania 20 w *f*: 0,7	a_{20}^r =1	82	18	0	0	0	15	3	0
	a_{20}^r =1,5	88	12	0	0	0	9	0	3
	a_{20}^r =2	86	14	0	1	0	6	0	7

wych: testu Mantel–Haenshel oraz metody wykorzystującej modelowanie IRT. Analizy przeprowadzone na danych symulacyjnych pozwoliły przetestować funkcjonowanie obu metod w różnych warunkach eksperymentalnych.

Badanie potwierdziło, że test Mantel–Haenshel jest testem mocniejszym w wypadku występowania systematycznych i jednorodnych różnic między grupą ogniskową i odniesienia, gorzej radzi sobie natomiast w przypadku niejednorodnego efektu DIF,

gdy różnica między łatwością zadania w obu grupach jest różna w zależności od poziomu mierzonej cechy ukrytej. W takich sytuacjach z test LR jest testem mocniejszym, który lepiej radzi sobie z wykrywaniem takich niuansów w funkcjonowaniu zadania.

Zauważono, że oszacowanie wielkości efektu DIF na skali łatwości zadania za pomocą modelowania IRT daje obciążenie, które jest zupełnie nieistotne w kontekście praktycznym. Ponadto wielkość błędu standardowego dla analogicznego oszacowania na podstawie *STD P – DIF* jest bardzo zbliżona do błędu *IRT P – DIF*, co rodzi możliwość wykorzystania błędów *STD P – DIF*, dla których wyznaczenia wyprowadzono odpowiednie wzory (Dorans i Holland, 1993) jako przybliżenie błędów *IRT P – DIF*. To z kolei umożliwiałoby wzbogacenie klasyfikacji na podstawie *IRT P – DIF* o informacje na temat precyzji oszacowania statystyki, analogicznie do klasyfikacji na podstawie *MH D – DIF*.

Literatura

Agresti, A. (2002). *Categorical data a nalysis*. New Jersey: John Wiley & Sons.

Dorans, N. J. i Holland, P. W. (1993). DIF detection and description: Mantel-Haenshel and standardization. W: P. W. Holland i H. Wainer (red.), *Differential item functioning* (s. 35–66). Hillsdale, NJ: Lawrence Earlbaum.

Glas, C. A. (2010). *Preliminary manual of the software program Multidimensional Item Response Theory (MIRT)*. Enschede: University of Twente.

Kondratek B. (2012). *Bias of IRT observed score equating under NEAT design*. Plakat naukowy zaprezentowany na konferencji Modern Modelling Methods, Storrs, Connecticut.

Lord, F. M. (1983). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika*, *48*(3), 425–435.

Lord, F. M. i Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley.

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

Monahan, P. O., McHorney, C. A., Stump, T. E. i Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*(1), 92–109.

Penfield, R. D. i Camilli, G. (2007). Differential Item Functioning and item bias. W: C. R. Rao i S. Sinharay (red.), *Handbook of statistics, Vol. 26. Psychometrics* (s. 125–167). New York, NY: Elsevier.

Radhakrishna, S. (1965). Combination of results from several 2 × 2 contingency tables. *Biometrics*, *21*(1), 86–98.

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

Thissen, D., Steinberg, L. i Wainer, H. (1993). Detection of Differential Item Functioning using the parameters of item response models. W: P. W. Holland i H. Wainer (red.), *Differential Item Functioning* (s. 67–113). Hillsdale, NJ: Lawrence Earlbaum.

Wainer, H. (1993). Model-based standardized measurement of an items differential impact. W: P. W. Holland i H. Wainer (red.), *Differential Item Functioning* (s. 255–276). Hillsdale, NJ: Lawrence Earlbaum.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics*, *19*(4), 251–253.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. W: P. W. Holland i H. Wainer (red.), *Differential Item Functioning* (s. 337–348). Hillsdale, NJ: Lawrence Earlbaum.

Zieky, M. (2003). *A DIF primer*. Princeton, NJ: ETS.