Edukacja 2017, 3(142), 52–69
doi: 10.24131/3724.170304
issn: 0239-6858

Czy nauczyciele edukacji wczesnoszkolnej potrafią bezstronnie ocenić osiągnięcia dziewcząt i chłopców z języka polskiego?

Paulina Skórska, Karolina Świst, Paweł Grygiel, Grzegorz Humenny

Instytut Badań Edukacyjnych*

Michał Modzelewski, Roman Dolata

Wydział Pedagogiczny, Uniwersytet Warszawski

Dotychczasowe badania sugerują, że w okresie wczesnoszkolnym przy tym samym poziomie umiejętności z zakresu języka ojczystego nauczyciele wyżej oceniają osiągnięcia szkolne dziewczynek niż chłopców. Celem artykułu jest weryfikacja tej hipotezy. Wykorzystano (a) oceny osiągnięć uczniów wystawione przez nauczycieli poza procesem nauczania (dla celów badawczych) i (b) wyniki standaryzowanych testów osiągnięć z języka polskiego. Analizy uwzględniające potencjalną stronniczość pozycji testowych ze względu na płeć ucznia przeprowadzono metodą modelowania wielu wskaźników i wielu przyczyn (MIMIC). Wykorzystano dane pochodzące z ogólnopolskiego badania 4144 uczniów trzeciej klasy szkoły podstawowej. Zgodnie z przewidywaniami, nauczyciele wyżej oceniali osiągnięcia dziewczynek niż chłopców, ale ta różnica zanikła, gdy do modelu analizy włączono wyniki standaryzowanych testów. Okazuje się więc, że nauczyciele potrafią bezstronnie ocenić osiągnięcia szkolne z języka polskiego dziewczynek i chłopców.

Słowa kluczowe: język polski; model wielu wskaźników i wielu przyczyn; oceny szkolne; płeć; stronniczość; zróżnicowane funkcjonowanie pozycji testowej.

Oceny nauczycielskie osiągnięć uczniów są i z pewnością długo jeszcze pozostaną ważnym elementem współczesnej szkoły, mimo że są krytykowane za subiektywizm i intuicyjność, a także za powiązania z funkcją selekcyjną szkoły i szkolnym konserwatyzmem (Niemierko, 1997; 2009). Zdaniem krytyków subiektywizm oceny prowadzi do różnego rodzaju stronniczości, w tym do ulegania stereotypom odnoszącym się do kategorii społecznych, do których należy uczeń. W tym artykule podejmujemy problem wpływu płci ucznia na oceny szkolne.

Nauczycielskie oceny osiągnięć szkolnych dziewcząt są przeciętnie wyższe niż chłopców. To dobrze znany fakt, szczególnie w zakresie umiejętności o silnym komponencie werbalnym. Średnia ocen z języka polskiego uczennic jest wyższa niż uczniów. Zobiektywizowane pomiary osiągnięć w zakresie języka polskiego potwierdzają faktyczną przewagę dziewcząt. Wystarczy spojrzeć na wyniki polskiego egzaminu gimnazjalnego, by zobaczyć, że mierzone

* Adres: ul. Górczewska 8, 01-180 Warszawa.
E-mail: k.swist@ibe.edu.pl

za pomocą standaryzowanych testów osiągnięcia dziewcząt z języka polskiego są wyższe niż osiągnięcia chłopców. W latach 2002–2012 różnica w części humanistycznej wynosiła ok. 0,3 odchylenia standardowego (Dolata i Sitek, 2015), a od 2013 r. przewaga uczennic w wyodrębnionym teście egzaminacyjnym z języka polskiego wynosiła ok. 0,4 odchylenia (CKE, 2012; 2016). Liczne analizy wskazują jednak, że różnice ocen z języka polskiego uczennic i uczniów nie dają się w pełni wyjaśnić wynikami standaryzowanych testów. Tę „resztową", prodziewczęcą tendencję w ocenianiu wyjaśnia się wielowymiarowym charakterem nauczycielskich ocen i ich wielofunkcyjnością.

Ocenianie jest zanurzone w szkolnej codzienności i podlega najróżniejszym wpływom. W artykule dostarczamy odpowiedzi na pytanie: Czy prodziewczęcość nauczycielskich ocen zaniknie, jeżeli poprosimy nauczycieli o ocenę umiejętności uczniów na potrzeby badania naukowego, czyli poza normalnym kontekstem życia szkolnego? Określenie resztowej, czyli nieuzasadnionej wynikami testowania prodziewczęcości, wymaga upewnienia się, że wyniki testowania nie są zaburzone przez zmienną płci (na podstawie analizy DIF, czyli zróżnicowanego funkcjonowania zadań testowych) oraz zastosowania odpowiedniego modelu analizy pozwalającego kontrolować ewentualną stronniczość testów (metoda wielu wskaźników i wielu przyczyn, MIMIC).

Międzypłciowe różnice nauczycielskich ocen osiągnięć uczniów

Metaanalizy dotyczące międzypłciowego zróżnicowania stopni szkolnych (Fischer, Schult i Hell, 2013; Richardson, Abraham i Bond, 2012; Voyer i Voyer, 2014) wskazują, że dziewczynki otrzymują istotnie lepsze oceny niż chłopcy. Przewaga uczennic utrzymuje się także przy kontroli faktycznego poziomu umiejętności uczniów mierzonych standaryzowanymi testami osiągnięć szkolnych. Różnica na korzyść uczennic utrzymuje się mimo braku znaczących różnic międzygrupowych w wynikach testów lub nawet przewagi chłopców – jak bywa w matematyce lub naukach przyrodniczych (Duckworth i Seligman, 2006; Ekstrom, 1994; Kling, Noftle i Robins, 2013). Potwierdza to wiele badań i metaanaliz (np. Else-Quest, Hyde i Linn, 2010; Hyde, Fennema i Lamon, 1990; Konarzewski, 1995; Lindberg, Hyde, Petersen i Linn, 2010). Również analizy polskiego systemu oświaty wykorzystujące wyniki egzaminacyjne potwierdzają, że nauczycielskie oceny osiągnięć szkolnych dziewcząt są wyższe, niż wynikałoby to z wyników standaryzowanych testów osiągnięć (Konarzewski, 2003). Zjawisko to wyjaśnia się rozmaicie (por.: Burusic, Babarovic i Seric, 2012; Hadjar, Krolak-Schwerdt, Priem i Glock, 2014; Voyer i Voyer, 2014), ale wszystkie koncepcje zakładają, że oceny nauczycielskie, oprócz ocenianej umiejętności, mogą odzwierciedlać także inne cechy (poznawcze i pozapoznawcze) oraz fakt wielofunkcyjności ocen.

Możliwe, że stopnie stawiane przez nauczycieli uwzględniają zdolności poznawcze, których nie mierzą standaryzowane testy osiągnięć (Conger i Long, 2010). Wyjaśnienie takie byłoby zgodne z teorią przeszacowania predykcji wyników zobiektywizowanych egzaminów dla uczennic (Duckworth i Seligman, 2006). Zgodnie z tą teorią standaryzowane testy nie doszacowują osiągnięć szkolnych dziewcząt, ponieważ nie mierzą ważnych, a dostrzeganych przez nauczycieli zdolności poznawczych (Shibley Hyde i Kling, 2001).

Jest możliwe, że stopnie odzwierciedlają także istotne z punktu widzenia skuteczności nauczania czynniki o charakterze pozapoznawczym (Ekstrom, 1994; Brookhart, 1997; McMillan, 2001; 2003; Rakoczy, Klieme, Bürgermeister i Harks, 2008; Randall i Engelhard, 2010), takie jak:

- wsparcie rodziny (Herbert i Stipek, 2005; Serbin, Stack i Kingdon, 2013),
- czynniki związane z osobowością i stosunkiem do uczenia się (Hicks, Johnson, Iacono i McGue, 2008; Noftle i Robins, 2007; Richardson i in., 2012; Spinath, Eckert i Steinmayr, 2014; Spinath, Harald Freudenthaler i Neubauer, 2010; Steinmayr i Spinath, 2008), w tym sumienność (Kling i in., 2013; Mattern, Sanchez i Ndum, 2017), motywacja (Preckel, Holling i Vock, 2006; Vecchione, Alessandri i Marsicano, 2014), samodyscyplina (Duckworth, Quinn i Tsukayama, 2012; Duckworth i Seligman, 2006; Weis, Heikamp i Trommsdorff, 2013), ugodowość (*agreeableness;* Laidra, Pullmann i Allik, 2007),
- kompetencje społeczne,
- pozytywne zachowania w klasie (Buchmann, DiPrete i McDaniel, 2008; Cornwell, Mustard i Parys, 2013; DiPrete i Jennings, 2012).

Zdaniem niektórych badaczy stopnie szkolne są raczej miarą spełniania wielowymiarowego standardu „dobrego ucznia" niż miarą jego osiągnięć w zakresie danego przedmiotu szkolnego (Kimball, 1989; Allen, 2005; Mullola i in., 2012; Spilt, Koomen i Jak, 2012). Ponieważ bliżej tego standardu są uczennice, to tłumaczy ich wyższe oceny. Niektórzy badacze dowodzą nawet, że feminizacja szkoły prowadzi do dominacji wartości związanych przez trening kulturowy z kobiecością. Podejmowano wiele badań nad wpływem feminizacji szkół na funkcjonowanie chłopców, ale wyniki nie są jednoznaczne (np. Driessen, 2007).

Inni badacze zwracają uwagę, że wystawiane przez nauczycieli stopnie nie są wynikiem szacowania osiągnięć intelektualnych ucznia, lecz mają charakter wychowawczy, czyli służą przede wszystkim kontrolowaniu motywacji i zachowań ucznia (Guskey, 2011; Remesal, 2011), kształceniu nawyków pracy oraz zarządzaniu klasą (Brookhart, 1997). Nauczyciele mogą też wykorzystywać oceny szkolne do kształtowania samooceny uczniów oraz oceny rówieśniczej (Grygiel, Modzelewski i Pisarek, 2016; Trautwein, Lüdtke, Marsh, Köller i Baumert, 2006).

Wielowymiarowość i wielofunkcyjność ocen szkolnych i związana z tym synkretyczność prowadzi niektórych analityków do kwestionowania ich wartości i odmawiania im jakiejkolwiek użyteczności (Bowers, 2011). Wyniki badań (Bacon i Bean, 2006; Guskey, 2011) wskazują jednak, że pomimo względnej nieokreśloności są one stosunkowo stabilne w całym okresie szkolnym, a oceny otrzymane na wcześniejszych etapach nauki są dobrymi predyktorami ocen późniejszych. Badania amerykańskie dowodzą, że stopnie uzyskiwane w szkołach podstawowych stosunkowo dobrze przewidują oceny otrzymywane w szkołach wyższego szczebla (Byrnes i Miller, 2007; Casillas i in., 2012). Co ciekawe, sukces w szkole wyższej może być trafniej przewidywany na podstawie ocen uzyskanych w szkole średniej niż wyników testów predyspozycji do studiowania (np. SAT, ACT; Richardson i in., 2012; Trapmann, Hell, Weigand i Schuler, 2007). Jednak trzeba pamiętać, że w realiach szkoły amerykańskiej testy znacznie silniej wpływają na wystawianie nauczycielskich ocen niż w Polsce. Mimo to w polskich badaniach również zaobserwowano wysoką moc prognostyczną ocen szkolnych, w szczególności oceny z matematyki (Konarzewski, 2003; Skórska i Świst, 2014; Świst i Skórska, 2016). Okazuje się, że na podstawie ocen gimnazjalnych można w dużej mierze przewidywać zarówno wynik uzyskany na egzaminie gimnazjalnym, jak i na egzaminie maturalnym odbywającym się trzy lata później. Zazwyczaj korelacje między stopniami szkolnymi a wynikami testów zewnętrznych są dość wysokie, wahają się między 0,4 a 0,8 (Brennan, Kim, Wenz-Gross i Siperstein, 2001; Martínez, Stecher i Borko, 2009; Ross i Gray, 2008; Ross i Kostuch, 2011; Zhu i Urhahne,

2014), a znana metaanaliza zespołu Anny Südkamp (Südkamp, Kaiser i Möller, 2012), obejmująca 75 badań, wykazała korelację na poziomie 0,63. Wydaje się więc, że wielowymiarowość i wielofunkcyjność nauczycielskich ocen nie przekreśla ich wartości jako wskaźnika poziomu osiągnięć uczniów. Tym bardziej warto pytać: Czy nauczyciele potrafią bezstronnie ocenić osiągnięcia szkolne dziewcząt i chłopców?

Międzypłciowe różnice wyników standaryzowanych testów osiągnięć szkolnych

Płeć ucznia różnicuje wyniki standaryzowanych testów osiągnięć. W testach umiejętności matematycznych, takich jak NAEP (*National Assessment of Educational Progress*) czy TIMSS (*Trends in International Mathematics and Science Study*), chłopcy mają przewagę nad dziewczynkami (Dalton, Ingels, Downing i Bozick, 2007) wielkości 0,16 odchylenia standardowego (Hyde, Fennema, Ryan, Frost i Hopp, 1990). Metaanaliza Sary Lindberg i współpracowników (Lindberg i in., 2010), obejmująca wyniki z matematyki z lat 1990–2007, wskazuje, że średni efekt płci ucznia wynosi 0,05 odchylenia standardowego, a więc nie ma praktycznego znaczenia. Nie jest to jednak prawidłowość uniwersalna. Na przykład w najnowszej edycji badania czwartoklasistów TIMSS 2015 w 8 krajach zaobserwowano istotną statystycznie różnicę na korzyść uczennic (w tym w Finlandii), w 23 nie zanotowano istotnej różnicy (w tym w Polsce), a w 18 krajach stwierdzono przewagę chłopców (Konarzewski i Bulkowski, 2016).

Inaczej jest w wypadku testów osiągnięć z zakresu języka ojczystego, tj. głównie czytania i pisania. Tu badania wskazują na silną i stabilną przewagę dziewczynek nad chłopcami. Badania PIRLS 2001, 2006 i 2011 (*Progress in International Reading Literacy Study*) pokazały, że dziewczynki po czterech latach nauki uzyskują wyższe wyniki w testach czytania (Konarzewski, 2012; Mullis, Martin, Kennedy i Foy, 2007). W edycji PIRLS 2011 w 5 krajach nie zaobserwowano istotnej statystycznie różnicy, a w pozostałych 40 wyniki dziewczynek były znacząco wyższe. Globalny efekt osiągnął wielkość 0,20 odchylenia standardowego (Robinson i Lubienski, 2011), w Polsce – 0,19 odchylenia. Inne badania przynosiły różne oszacowania wielkości efektu. W jednych (Logan i Johnston, 2009) dziewczynki przewyższały chłopców w testach czytania o ok. 2/3 odchylenia, w innych (Lietz, 2006) – tylko o 0,19 odchylenia. Badanie PISA (Programme for International Student Assessment) wskazuje na różnicę na poziomie 0,44 odchylenia (w Polsce 0,36; Reilly, 2012).

Drugim wymiarem osiągnięć szkolnych z języka polskiego jest umiejętność pisania. Niestety w tym zakresie znacznie trudniej o wyniki badań międzynarodowych. Naukowcy amerykańscy zwracają uwagę na mniejsze różnice płciowe w pisaniu (Logan i Johnston, 2009), które mogą wynosić ok. 0,04 odchylenia standardowego (Lietz, 2006). Wyniki polskich badań wykazują jednak silniejszy efekt. Na przykład w badaniu osiągnięć uczniów czwartej klasy szkoły podstawowej z 2015 r. efekt płci związany ze świadomością językową (ważnym aspektem umiejętności pisania) był znacząco silniejszy niż w wypadku czytania (Dolata, Hawrot, i in., 2015).

W budowaniu testów osiągnięć ważne jest sprawdzenie, czy poszczególne zadania mają takie same właściwości pomiarowe w różnych grupach respondentów o tym samym poziomie cechy ukrytej. Gdy parametry zadań w analizowanych grupach znacząco się różnią, wówczas mówimy o zróżnicowanym funkcjonowaniu zadań (pozycji) testowych (*differential item functioning*, DIF). Innymi słowy: DIF ujawnia się, gdy pojawiają się międzygrupowe różnice w sposobie odpowiadania na pozycje testowe

mimo kontrolowania poziomu umiejętności (cechy ukrytej). Na przykład jeżeli uczennice częściej udzielają poprawnych odpowiedzi na jedno lub więcej pytań, niż to wynika z ich poziomu umiejętności, to ogólny wynik uczennic będzie zawyżony. Wykrycie DIF może świadczyć o „stronniczości" określonej pozycji testowej, ale żeby uznać ją za rzeczywiście stronniczą, trzeba przeprowadzić analizę ekspercką, która wykaże, że wyższe prawdopodobieństwo poradzenia sobie z zadaniem wynika z cech zadania niezwiązanych z badaną umiejętnością (np. płci ucznia). Więcej o zjawisku DIF można przeczytać w innych publikacjach (np. Grygiel, Świtaj i Humenny, 2015; Kondratek, Skórska i Świst, 2015).

Metoda

Próba

W analizach wykorzystano dane zebrane w ramach ogólnopolskiego badania *Szkolne uwarunkowania efektywności kształcenia* (SUEK). Próba obejmowała ponad 5000 uczniów z losowo dobranych 274 oddziałów trzeciej klasy szkoły podstawowej. W analizach wykorzystano dane zebrane od 4144 uczniów (49,6% dziewczynek) o średniej wieku 9,6 lat (z wariancją 0,1), których osiągnięcia w języku polskim zostały ocenione przez nauczycieli, i którzy wykonali testy osiągnięć w tym przedmiocie. Szczegółowy opis metodologii badania można znaleźć w publikacjach książkowych (Dolata, 2014; Dolata, Grygiel i in., 2015).

Zmienne

Oceny nauczycielskie. Zebrano je pod koniec trzeciej klasy w roku szkolnym 2010/2011, prosząc nauczycieli o zaklasyfikowanie każdego ucznia do jednej z czterech kategorii:

- uczniowie słabi – słabo opanowali materiał, popełniają liczne błędy i wymagają systematycznej pomocy,
- uczniowie przeciętni – radzą sobie z wymaganiami, ale są niesamodzielni, popełniają błędy i potrzebują pomocy,
- uczniowie, którzy dobrze sobie radzą, bardzo rzadko popełniają błędy, wymagają niewielkiej pomocy,
- uczniowie wyróżniający się pod względem wszystkich wymaganych umiejętności i samodzielności wykonania.

Instrukcja podkreślała, że oceny powinny odzwierciedlać wyłącznie osiągnięcia ucznia i że będą wykorzystane jedynie w badaniu naukowym, czyli że nie zostaną ujawnione uczniom.

Trzeba podkreślić specyfikę ocen wykorzystanych w badaniu nauczycielskich. Dane będące podstawą przeprowadzonych analiz pochodzą z badania przeprowadzonego w trzeciej klasie szkoły podstawowej. Na tym etapie nauki nauczyciele nie wystawiają ocen końcowych w postaci tradycyjnych stopni szkolnych. Fakt, że oceny zostały sformułowane wyłącznie na potrzeby badania, ma istotne znaczenie. Po pierwsze, instrukcja badawcza zalecała ocenę ucznia jedynie ze względu na jego osiągnięcia szkolne. Po drugie, stopnie te z definicji nie miały zostać ujawnione uczniom – nauczyciel nie powinien zakładać więc, że mogą one pełnić funkcję inną niż diagnostyczna (np. motywacyjną). Stwierdzenie stronniczości płciowej w przypadku ocen wystawionych jedynie na potrzeby badania będzie świadczyć o tym, że „pozapoznawczy" element tradycyjnych stopni szkolnych jest tak głęboko wbudowany w świadomość nauczycieli, że stał się integralnym („przedświadomym") elementem nauczycielskiej percepcji uczniów. Innymi słowy, że nauczyciele nawet, gdy się ich o to prosi, nie potrafią oceniać uczniów w oderwaniu od ich przynależności kategorialnej.

Testy umiejętności szkolnych. Przeprowadzono je na początku czwartej klasy w roku szkolnym 2011/2012. Testy osiągnięć szkolnych opracowane w projekcie SUEK to testy papierowe, dostosowane do badania

audytoryjnego i skalowane w modelu Rascha (opis testów w: Jasińska-Maciążek i Modzelewski, 2014). Testy mają dwie równoległe wersje z pulą 15–16 zadań wspólnych (kotwiczących) dla obu wersji. W analizach wykorzystano wyniki testów osiągnięć w obszarach: (a) umiejętności czytania oraz (b) świadomości językowej.

Test umiejętności czytania składa się z 51 zadań (w tym 15 zadań kotwiczących), mierzy stopień zrozumienia czytanych przez ucznia poleceń, zadań i tekstów różnego typu (nie obejmuje techniki czytania) i obejmuje trzy aspekty czytania: (a) umiejętność wyszukiwania informacji, (b) interpretację, (c) refleksję i ocenę. Test świadomości językowej składa się z 43 zadań (w tym 16 kotwiczących) i mierzy wiadomości i umiejętności umożliwiające refleksję nad językiem jako narzędziem komunikowania się: umiejętności tworzenia tekstu pisanego zgodnie z zasadami ortografii i gramatyki języka polskiego, o czytelnej strukturze, w zgodzie z zasadą jasnego przekazywania myśli i zasadami logiki, a także umiejętności w zakresie argumentowania oraz zasób słownikowy dziecka. Test obejmuje trzy obszary: (a) umiejętności związanych z pisaniem tekstów, (b) zasoby słownikowe oraz (c) elementy wiedzy o języku.

Stworzenie ostatecznej wersji testu poprzedzono badaniem pilotażowym w roku szkolnym 2010/2011 zrealizowanym w trzecich i piątych klasach ogólnopolskiej losowej próby 80 szkół podstawowych (łącznie zbadano 5454 uczniów z 281 oddziałów). Z punktu widzenia problemu podjętego w tym artykule ważne jest, że w badaniu pilotażowym oceniano zadania na podstawie statystyki DIF (ze względu na płeć).

Hipoteza

Na podstawie wyników wcześniejszych badań można oczekiwać, że:

- nauczycielskie oceny uczennic będą wyższe niż oceny uczniów,
- wystąpi silny związek między ocenami nauczycielskimi a wynikami standaryzowanych testów osiągnięć,
- wyższość ocen uczennic utrzyma się mimo kontroli poziomu umiejętności za pomocą standaryzowanych testów osiągnięć.

Kluczowe dla badania jest ostatnie oczekiwanie, które traktujemy jako hipotezę badawczą.

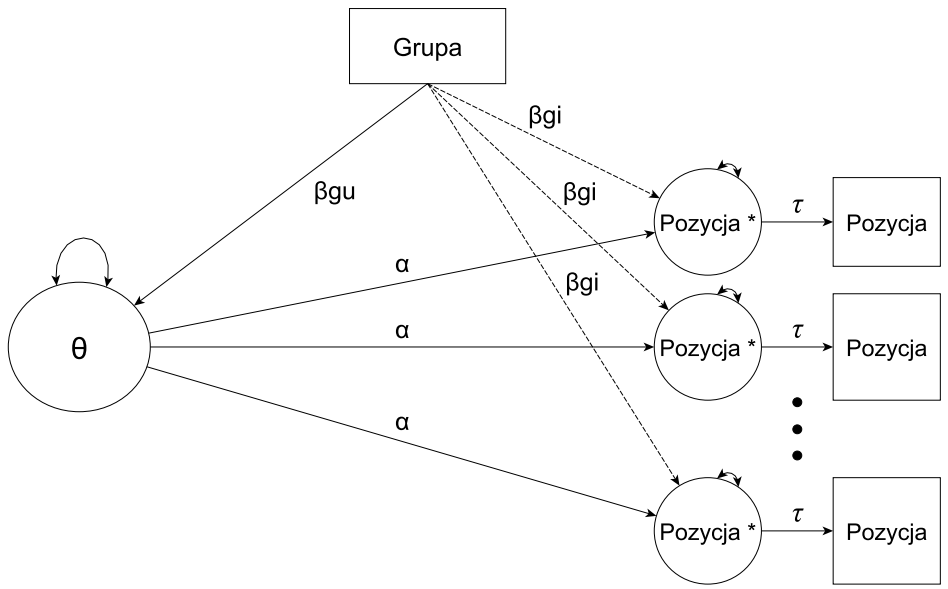
Analiza efektów DIF za pomocą modelu MIMIC

Współczesna statystyka dysponuje wieloma metodami wykrywania DIF. Należą do nich: metoda Mantela-Haenszela, symultaniczny test stronniczości pozycji (*simultaneous item bias test,* SIBTEST), regresja logistyczna, estymatory wariancji DIF, χ^2 Lorda, modele log-liniowe, hierarchiczne uogólnione modele liniowe DIF, metoda Raju zróżnicowanego funkcjonowania pozycji i testów (*differential functioning of items and tests,* DFIT), miary obszaru między krzywymi charakterystycznymi (*item characteristic curve*, ICC) itd. W niniejszej analizie do testowania DIF wykorzystano metodę wielu wskaźników i wielu przyczyn (*multiple indicators, multiple causes,* MIMIC), zaproponowaną na początku lat 70. ubiegłego wieku przez Roberta Hausera i Arthura Golbergera (1971). Metoda MIMIC ma kilka zalet (Jones, 2006). Umożliwia analizę DIF z uwzględnieniem wielu zmiennych niezależnych (predyktorów), które mogą być źródłem efektu zróżnicowanego funkcjonowania pozycji testowej. Zmienne te nie muszą być dychotomiczne, mogą mieć charakter ilościowy. Zastosowanie modelu nie ogranicza się do jednowymiarowej cechy ukrytej, można go łatwo rozszerzyć na badanie DIF w modelach wielowymiarowych czy w ich specyficznej odmianie: modelach podwójnego czynnika. Metoda MIMIC umożliwia również określenie relatywnej ważności zidentyfikowanych przypadków DIF przez porównanie wpływu różnic międzygrupowych (szerzej:

zmiennych niezależnych) na cechę ukrytą przed i po kontroli DIF. Co równie ważne, model wielu wskaźników i wielu przyczyn efektywnie wykrywa DIF przy relatywnie niewielkich próbach o liczebności ok. 200.

W sensie statystycznym modele MIMIC stanowią połączenie dwóch rodzajów technik analitycznych: konfirmacyjnej analizy czynnikowej (*confirmatory factor analysis,* CFA) oraz analizy ścieżek (*path analysis,* PA). Składa się więc z komponentu pomiarowego (CFA) oraz regresyjnego (PA), zwanego także komponentem strukturalnym (Bye, Gallicchio i Dykacz, 1985). Komponent pomiarowy służy do estymacji niedającego się bezpośrednio zaobserwować poziomu interesującej nas cechy (np. umiejętności) szacowanego na podstawie odpowiedzi udzielonych na pytania składające się na narzędzie badawcze – stąd pochodzi część nazwy „wiele wskaźników" (*multiple indicators*). Komponent regresyjny umożliwia poznanie wpływu zmiennych niezależnych na poziom cechy ukrytej uchwycony w ramach komponentu pomiarowego – stąd część nazwy „wiele przyczyn" (*multiple causes*). Formalny (statystyczny) opis modelu MIMIC przedstawił Roman Konarski (2009).

Metoda MIMIC użyta do wykrycia DIF dodatkowo różnicuje komponent regresyjny na dwie składowe: efekty bezpośrednie i pośrednie (Rysunek 1). Efekty pośrednie odnoszą się do ścieżek regresji od zmiennych niezależnych do cechy ukrytej θ. Efekty bezpośrednie są związane ze ścieżkami regresji od zmiennych niezależnych do obserwowalnych wskaźników tworzących skalę. Zaobserwowanie statystycznie istotnego efektu bezpośredniego oznacza jednocześnie występowanie znaczących różnic grupowych (wyznaczanych przez zmienne niezależne) w odpowiadaniu na pozycje testowe przy tym samym natężeniu cechy ukrytej i kontrolowanym wpływie zmiennych zależnych na poziom θ.



Rysunek 1. Model MIMIC w analizie DIF.

β_{gu} = średnia różnica poziomu zmiennej latentnej (np. umiejętności) między grupą ogniskową a grupą odniesienia (współczynnik regresji); β_{gi} = międzygrupowe zróżnicowanie progów (współczynnik regresji) poszczególnych pozycji testowych; α = współczynniki dyskryminacji (ładunki czynnikowe); ε= błąd pomiarowy dla danej pozycji; τ = progi (trudność) pozycji.

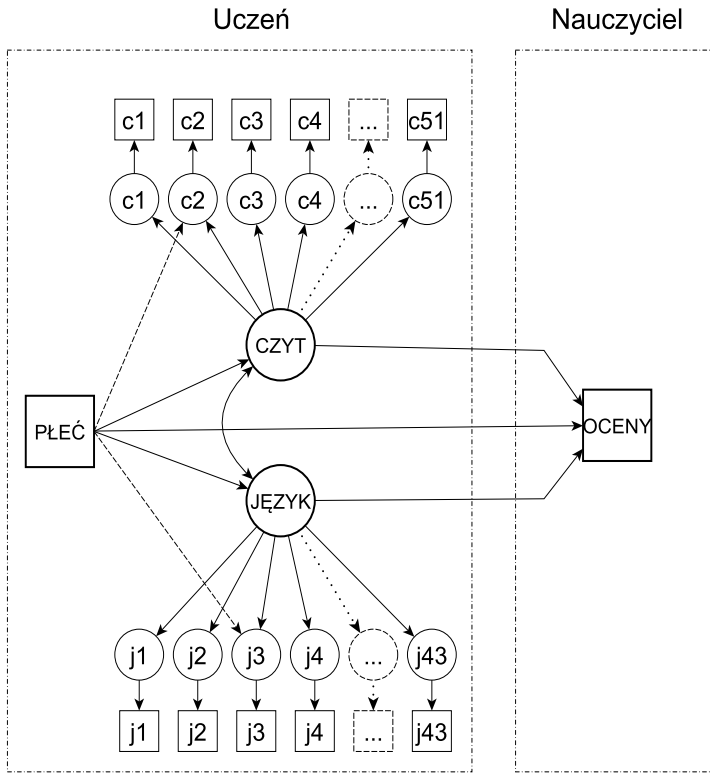
Odpowiedź na postawione pytania badawcze zakłada sprawdzenie, czy narzędzia badawcze, jakie zostaną wykorzystane w analizach właściwych, są porównywalne ze względu na płeć ucznia. Innymi słowy: czy sposób konstrukcji, wykorzystane pozycje testowe i ich „funkcjonowanie" sprzyja wyższym lub niższym wynikom dziewczynek niż chłopców. Identyfikacja tego typu przypadków (DIF) umożliwi zastosowanie odpowiednich procedur korygujących w ramach właściwych analiz.

Procedura testowania zróżnicowanego funkcjonowania pozycji testowej (DIF) – zgodnie z propozycją Carol M. Woods (2008; 2009) – została przeprowadzona w trzech kolejnych krokach. W pierwszym sprawdzono, czy przy kontroli poziomu cechy latentnej zmienna płci wpływa w istotny sposób na parametr trudności danej pozycji. Taką analizę przeprowadzono oddzielnie dla każdej pozycji, łącznie estymowano więc 51 modeli dla umiejętności czytania oraz 43 dla świadomości językowej. Często do grupy potencjalnie wolnej od DIF (grupy zadań kotwiczących) włącza się pozycje, dla których istotność współczynnika regresji (płeć → pozycja testowa) jest większa niż 0,05. Jednak gdy część z tych pozycji jest obciążona DIF, cały zestaw zadań kotwiczących może być obciążony, co może prowadzić m.in. do niedokładności w oszacowaniu parametrów oraz przeszacowania liczby pozycji wykazujących DIF. Woods (2009) rekomenduje więc wybór *g* pozycji o najmniejszym stosunku wartości logarytmu wiarygodności oraz liczby wolnych parametrów, gdzie *g* wynosi zazwyczaj 10–20% całkowitej liczby zadań w teście. Wszystkie pozostałe zadania zostają zaliczone do grupy „ryzyka", współczynnik regresji różny od 0 wskazuje bowiem, że przy tym samym poziomie cechy latentnej pozycja wykazuje różną „trudność" w różnych grupach respondentów (DIF). Ta grupa pozycji stanie się przedmiotem szczególnej uwagi w dalszej części analiz.

W ramach drugiego kroku sprawdzono, czy wprowadzenie do modelu założenia, że określona pozycja z grupy „ryzyka" nie jest obarczona DIF (tzn. współczynnik regresji grupa → pozycja testowa jest równy 0) powoduje znaczące pogorszenie dopasowania modelu do danych. Punktem odniesienia dla modelu z tak nałożonymi ograniczeniami jest model pełny, w którym – w przypadku wszystkich pozycji zaliczonych do grupy narażonej na DIF, łącznie z pozycją, która w modelu z ograniczeniami miała ustalony współczynnik regresji na 0 – nie nakłada się żadnych ograniczeń na współczynniki regresji grupa → pozycja. Znaczące różnice między nimi wskazują, że założenie braku DIF prowadzi do pogorszenia dopasowania modelu, oznaczając tym samym, że DIF daną pozycję cechuje.

W ostatnim kroku estymowano model, w którym – w odniesieniu do wszystkich pozycji o potwierdzonym wcześniej DIF – uwolniono współczynniki regresji (cecha → pozycja), zaś w wypadku pozostałych ustalono je na 0. W modelu tym sprawdzano, czy w przypadku wszystkich pozycji, w stosunku do których zakładano wystąpienie DIF, poziom istotności współczynnika regresji β nadal pozostaje mniejszy niż 0,05. Jeżeli okazywał się on większy, to ustalano go na 0, a całą procedurę powtarzano do momentu, w którym wszystkie pozycje z grupy ryzyka DIF wykazywały $p < 0{,}05$ dla β.

W konsekwencji, estymując zmienną latentną w modelu ostatecznym, uwzględnia się informacje o występujących DIF, a tym samym zwiększa jej odporność na potencjalne zniekształcenia związane z oddziaływaniem DIF. W ramach oszacowania ostatniego modelu otrzymujemy więc nie tylko informacje o poziomie zmiennej latentnej niezależnie od DIF, lecz także o: (a) różnicy w jej natężeniu ze względu na płeć uwzględniającej DIF, (b) współczynnikach dyskryminacji, (c) trudności zadań oraz (d) oszacowanej wielkości DIF.



Rysunek 2. Model analizy oddziaływania płci ucznia na oceny nauczycielskie przy kontroli poziomu umiejętności, uwzględniający występowanie DIF (ze względu na płeć).

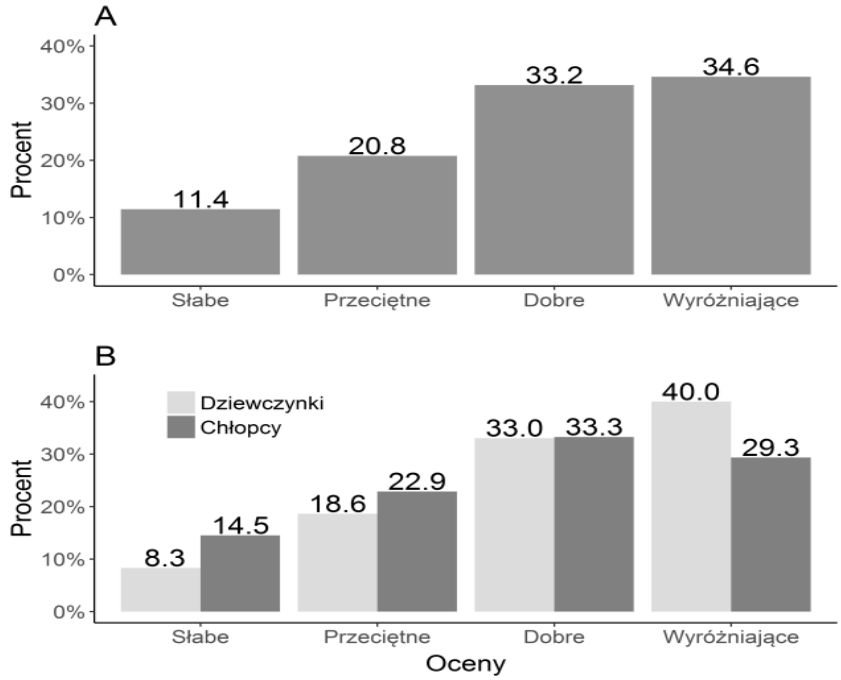
OCENY – poziom umiejętności z języka polskiego przypisany uczniowi przez nauczyciela (od 1 „uczeń słaby" do 4 „uczeń wyróżniający się"); PŁEĆ – płeć ucznia, w której grupą odniesienia są dziewczynki; CZYT – zmienna latentna powstała na podstawie wyników testu osiągnięć w obszarze umiejętności czytania; JĘZYK – zmienna latentna powstała na podstawie wyników testu osiągnięć w obszarze świadomości językowej; c1…c51 – zadania w teście osiągnięć w obszarze umiejętność czytania; j1…j43 – zadania w teście osiągnięć w obszarze świadomości językowej; Uczeń – dane zebrane za pomocą kwestionariusza wypełnianego przez ucznia; Nauczyciel – dane zebrane za pomocą kwestionariusza wypełnianego przez nauczyciela ucznia. Kwadratami oznaczono nazwy zmiennych jawnych; kołami oznaczono nazwy zmiennych ukrytych (latentnych); strzałki ciągłe o jednym grocie oznaczają zależności regresyjne; strzałki ciągłe o dwóch grotach oznaczają korelacje latentne; strzałki przerywane od zmiennej Płeć to znaczniki DIF.

Wyniki

Model analizy stronniczości płciowej nauczycielskich ocen osiągnięć z języka polskiego

Testowano dwa główne modele. Po pierwsze model, w którym na wystawione przez nauczycieli oceny wpływa jedynie płeć uczniów. Dzięki temu możliwa jest odpowiedź na pytanie: Czy nauczyciele inaczej oceniają osiągnięcia szkolne chłopców niż dziewczynek? Po drugie, model, w którym dodatkowo kontrolowano poziom uczniowskich umiejętności w zakresie czytania ze zrozumieniem i świadomości językowej – po wyłączeniu możliwych efektów DIF (Rysunek 2). Ten model daje odpowiedź na pytanie: Czy nauczycielskie oceny uczniów wynikają z grupowych różnic w poziomie ich wiedzy, czy też dodatkowo są wynikiem działania jakichś innych czynników?

Wszystkie analizy – o ile nie zaznaczono inaczej – zostały przeprowadzone w pakiecie



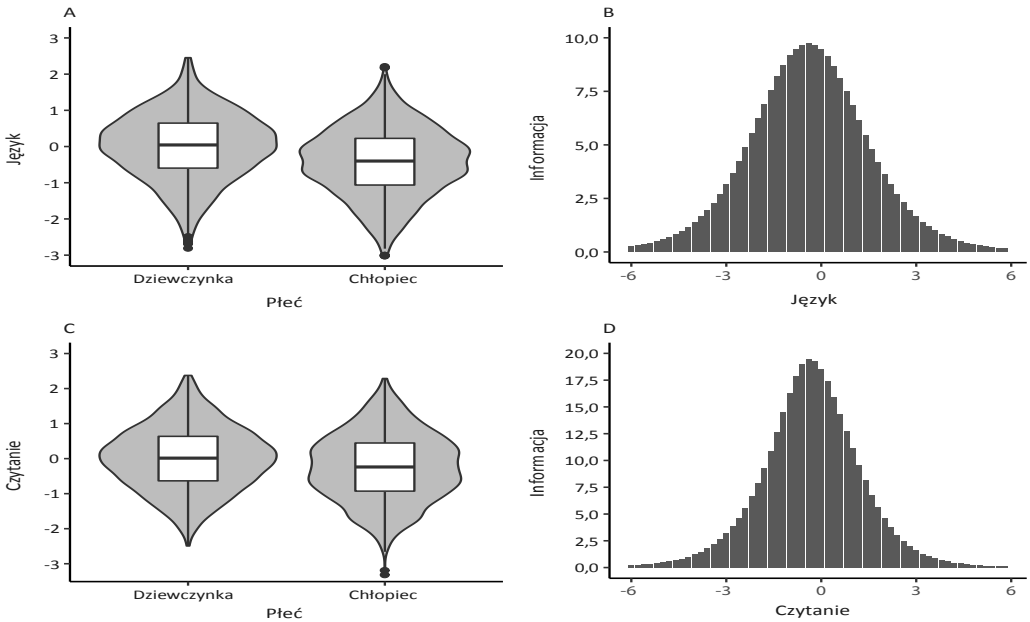
Rysunek 3. Rozkład nauczycielskich ocen osiągnięć szkolnych uczniów z języka polskiego (w %).
Część A – rozkład ocen ogółu dzieci; Część B – rozkład ocen według płci dzieci.

statystycznym Mplus 7.4 (Muthén i Muthén, 2012). W obliczeniach posłużono się pełnoinformacyjnym (*full-information*), odpornym estymatorem największej wiarygodności (*maximum likelihood estimation with robust standard errors,* MLR)[1]. Ze względu na specyfikę konstrukcji testów osiągnięć, zakładającej, że jedynie część pozycji testowych jest wypełniana przez wszystkich badanych (tzw. pozycje kotwiczące), nie istniała możliwość wykorzystania szybszego, niepełnoinformacyjnego (*limited-information*), opartego na macierzy korelacji polichorycznej estymatora ważonych najmniejszych kwadratów ze skorygowaną średnią i wariancją (*weighted least squares means and variance adjusted,* WLSMV).

[1] Estymator MLR maksymalizuje funkcję dopasowania modelu na podstawie wzorów odpowiedzi, a nie statystyk podsumowujących dane, np. wariancji czy kowariancji lub korelacji. Procedury wykorzystujące pełnoinformacyjny estymator MLR co do zasady działają podobnie jak modele IRT (Bovaird i Koziol, 2012). Różnice między nimi wynikają przede wszystkim z celów, którym służą: jak każda analiza czynnikowa, analiza z wykorzystaniem estymatora MLR służy wyjaśnieniu struktury korelacji pomiędzy zmiennymi obserwowalnymi, a w centrum zainteresowania IRT znajdują się relacje pomiędzy charakterystykami pozycji skali a charakterystykami respondentów udzielających na nie odpowiedzi (Brown, 2006; (Brown, 2006; Humenny i Grygiel, 2015; Kondratek i Pokropek, 2015; Kulon, 2015).

Rozkład ocen nauczycielskich i wyników testów osiągnięć z języka polskiego

Na Rysunku 3 przedstawiono rozkłady ocen nauczycielskich oraz wyników testów osiągnięć. Rozkład ocen jest zdecydowanie skośny – z przewagą wyników dobrych i wyróżniających. Oceny są zróżnicowane ze względu na płeć – w kategoriach uczniów „słabych" i „przeciętnych" występuje przewaga chłopców, natomiast w kategorii



Rysunek 4. Wartość informacyjna testów świadomości językowej i czytania oraz rozkłady umiejętności ze względu na płeć.

Część A – rozkład umiejętności językowych ze względu na płeć dzieci; część B – informacyjna krzywa testu umiejętności językowych; Część C – rozkład umiejętności z zakresu czytania ze względu na płeć dzieci; część D – informacyjna krzywa testu umiejętności z zakresu czytania.

uczniów „wyróżniających się" – przewaga dziewczynek (40% spośród dziewczynek i niespełna 30% chłopców). Kategoria uczniów „dobrych" w najmniejszym stopniu jest zróżnicowana ze względu na płeć – jednak zaliczono do niej trochę więcej chłopców.

Prawa strona Rysunku 4 obrazuje wartość informacyjną (*test information curve*) analizowanych testów – świadomości językowej oraz czytania. Oba testy mają największą wartość informacyjną dla osób o przeciętnym poziomie umiejętności, należy jednak zauważyć, że test z czytania jest bardziej informacyjny niż test świadomości językowej. Lewa strona Rysunku 4 prezentuje wykresy skrzypcowe (*violin plots*) rozkładów umiejętności z języka polskiego ze względu na płeć. Zwracają uwagę niższe mediany wyników osiąganych przez chłopców w obu testach oraz większa proporcja wyników niskich w tej grupie.

Efekty DIF w testach świadomości językowej i czytania

W teście świadomości językowej (Tabela 1) DIF ze względu na płeć ucznia ujawnił się w 17 (w 10 „dodatnio" i w 7 „ujemnie") spośród 43 pozycji testowych. Uwzględnienie DIF nie wpływa na oszacowanie różnic w poziomie umiejętności między chłopcami a dziewczynkami (standaryzowany współczynnik regresji β_{STDY} = =-0,41). Dziewczynki uzyskują wyższe wyniki w teście niż chłopcy.

W teście czytania DIF (Tabela 2) ujawnił się w 18 (w 9 „dodatnio" i w 9 „ujemnie") spośród 51 pytań testowych. Uwzględnienie DIF praktycznie nie wpływa na związek

Tabela 1
Standaryzowane współczynniki regresji płci na trudność pozycji (efekt DIF) – świadomość językowa

Pozycja	β[(a)]	(*se*)	Pozycja	β[(a)]	(*se*)	Pozycja	β[(a)]	(*se*)
J1	– –		J16	– –		J31	– –	
J2	0,14**	(0,04)	J17	– –		J32	-0,20**	(0,05)
J3	– –		J18	– –		J33	– –	
J4	– –		J19	– –		J34	– –	
J5	– –		J20	– –		J35	0,16*	(0,05)
J6	0,18**	(0,04)	J21	0,28*	(0,13)	J36	– –	
J7	0,11*	(0,04)	J22	0,16*	(0,05)	J37	0,16**	(0,05)
J8	0,09*	(0,04)	J23	0,33**	(0,10)	J38	– –	
J9	0,26**	(0,04)	J24	-0,25*	(0,11)	J39	– –	
J10	-0,11*	(0,04)	J25	-0,42**	(0,11)	J40	-0,14*	(0,05)
J11	-0,15**	(0,04)	J26	– –		J41	-0,19**	(0,05)
J12	-0,10*	(0,04)	J27	– –		J42	0,11*	(0,05)
J13	– –		J28	– –		J43	– –	
J14	-0,22**	(0,03)	J29	0,44**	(0,05)			
J15	0,19**	(0,04)	J30	– –				

[(a)] β (*se*) – standaryzowany współczynnik regresji (STDY) wraz z błędem standardowym.
„– –" oznacza brak DIF; * $p < 0,05$; ** $p < 0,01$.

Tabela 2
Standaryzowane współczynniki regresji płci na trudność pozycji (efekt DIF) – czytanie

Pozycja	β[(a)]	(*se*)	Pozycja	β[(a)]	(*se*)	Pozycja	β[(a)]	(*se*)
C1	0,12**	(0,04)	C18	0,10*	(0,05)	C35	– –	
C2	–		C19	– –		C36	– –	
C3	0,20**	(0,03)	C20	– –		C37	– –	
C4	-0,08*	(0,04)	C21	0,14*	(0,05)	C38	– –	
C5	– –		C22	-0,23**	(0,04)	C39	– –	
C6	– –		C23	– –		C40	0,30**	(0,05)
C7	– –		C24	– –		C41	– –	
C8	– –		C25	– –		C42	0,21**	(0,05)
C9	– –		C26	– –		C43	– –	
C10	-0,15**	(0,04)	C27	– –		C44	-0,16**	(0,05)
C11	-0,14**	(0,04)	C28	0,24**	(0,05)	C45	– –	
C12	–		C29	-0,11*	(0,04)	C46	– –	
C13	0,14**	(0,04)	C30	– –		C47	-0,16*	(0,05)
C14	– –		C31	– –		C48	-0,12*	(0,05)
C15	– –		C32	– –		C49	– –	
C16	0,20**	(0,05)	C33	– –		C50	-0,20*	(0,07)
C17	– –		C34	– –		C51	– –	

[(a)] β (*se*) – standaryzowany współczynnik regresji (STDY) wraz z błędem standardowym.
„– –" oznacza brak DIF; * $p < 0,05$; ** $p < 0,01$.

płci z badaną umiejętnością. O ile w modelu bez DIF współczynnik regresji β_{STDY} wyniósł -0,23 (*se* = 0,04; $p < 0,01$), o tyle w modelu z DIF wyniósł -0,24 (*se* = 0,04; $p < 0,01$). W obu wypadkach dziewczynki uzyskują średnio wyższe wyniki niż chłopcy.

Analiza stronniczości płciowej nauczycielskich ocen w zakresie języka polskiego

Dane zawarte w Tabeli 3 (Model 1) wskazują, że nauczycielskie oceny umiejętności językowych wystawione chłopcom są wyraźnie i istotnie statystycznie niższe niż te, które otrzymały dziewczynki (β_{STDY} = = -0,27). Nauczyciele oceniają, że w badanym zakresie przeciętne umiejętności chłopców są wyraźnie niższe niż dziewczynek. Uwzględnienie w modelu umiejętności oszacowanych za pomocą standaryzowanych testów osiągnięć powoduje, że związek pomiędzy płcią a oceną osiągnięć szkolnych z języka polskiego zanika. Innymi słowy: przekonania nauczycieli są trafne przynajmniej w zakresie różnic międzygrupowych, odpowiadają faktycznemu przeciętnemu poziomowi umiejętności dziewczynek i chłopców. Potwierdza się, że poziom umiejętności językowych uczniów jest znacznie niższy niż poziom umiejętności uczennic (w czytaniu współczynnik regresji dla płci wynosi β_{STDY} = -0,24; w świadomości językowej – β_{STDY} = -0,42).

Tabela 3
Standaryzowane współczynniki regresji w obu testowanych modelach

Współczynniki regresji	Model 1	Model 2
OCENY ← PŁEĆ	-0,27 (0,03)**	-0,007
OCENY ← CZYT	–	0,19 (0,05)**
OCENY ← JĘZYK	–	0,56 (0,05)**
CZYT ← PŁEĆ	–	-0,24 (0,04)**
JĘZYK ← PŁEĆ	–	-0,42 (0,03)**
CZYT ←→ JĘZYK	–	0,91 (0,01)**

„–" oznacza ścieżkę nieestymowaną w modelu; ← oznacza standaryzowany współczynnik regresji; ←→ oznacza korelację; OCENY – przypisany przez nauczyciela uczniowi poziom umiejętności polonistycznych (od 1 „uczeń słaby" do 4 „uczeń wyróżniający się"); PŁEĆ – płeć ucznia, w której grupą odniesienia są dziewczynki; CZYT – zmienna latentna powstała na podstawie wyników testu osiągnięć w obszarze umiejętności czytania; JĘZYK – zmienna latentna powstała na podstawie wyników testu osiągnięć w obszarze świadomości językowej; * $p < 0,05$; ** $p < 0,01$.

Dyskusja

Celem artykułu było sprawdzenie na dużej, ogólnopolskiej próbie, czy nauczyciele edukacji początkowej, oceniając osiągnięcia szkolne dziewczynek i chłopców z języka polskiego na potrzeby badawcze, a nie w naturalnym kontekście procesów nauczania–uczenia się, potrafią zrobić to bezstronnie. Dotychczasowe analizy prowadzone z wykorzystaniem ocen szkolnych wskazywały, że nauczyciele stawiają dziewczynkom wyższe stopnie, niż wynikałoby to z wyników standaryzowanych testów osiągnięć szkolnych (Skórska i Świst, 2014). Podkreślić należy, że oceny nauczycielskie wykorzystane w referowanym badaniu miały inny charakter niż zwykłe stopnie szkolne – nie były komunikowane uczniom i miały – zgodnie z instrukcją przedstawioną nauczycielom – zawierać wyłącznie informację o poziomie ich umiejętności, w związku z czym były wolne od funkcji formującej. Niewątpliwym walorem przedstawionych analiz jest także to, że poziom osiągnięć z języka polskiego dzieci był mierzony za pomocą starannie opracowanych testów, które były analizowane w sposób uwzględniający ich potencjalną stronniczość ze względu na płeć.

Wyniki przeprowadzonych analiz wykazują, że nauczyciele niżej oceniają umiejętności językowe chłopców niż dziewczynek – co

potwierdza pierwsze z oczekiwań badawczych. Zgodnie z przewidywaniem wystąpił też silny związek między nauczycielskimi ocenami umiejętności z języka polskiego a wynikami standaryzowanych testów osiągnięć w tym zakresie. Najważniejsze jest to, że gdy uwzględniono obiektywny pomiar poziomu umiejętności językowych, różnica ocen między chłopcami a dziewczynkami zanikła. Świadczy to o tym, że niższe oceny w zakresie umiejętności językowych wystawione chłopcom przez nauczycieli trafnie oddają ich faktycznie niższy poziom umiejętności w tym zakresie. Trzecia z postawionych hipotez nie została więc potwierdzona. Wbrew oczekiwaniom wyniki przeprowadzonych analiz sugerują, że nauczyciele potrafią formułować oceny nieobciążone efektem płci.

Przeprowadzone analizy mają jednak pewne ograniczenia, które wyznaczają równocześnie kierunki przyszłych badań. Słabością jest brak uwzględnienia, oprócz ocen uwolnionych od wpływu realiów procesu nauczania, ocen wystawianych w kontekście codziennego życia szkolnego. Następny problem to brak pewności, czy bezstronność oceniania dziewczynek i chłopców występuje w wypadku innych przedmiotów szkolnych, np. matematyki. Należy także pamiętać, że przedmiotem naszego badania były oceny wystawiane przez nauczycieli nauczania początkowego. Nie wiemy, czy nauczyciele nauczający na kolejnych etapach nauki w szkole podstawowej potrafią ocenić osiągnięcia bezstronnie ze względu na płeć dziecka. Możliwość uogólnienia wyników na inne przedmioty nauczania i inne klasy powinna być przedmiotem dalszych analiz. Ich celem mogłoby też być sprawdzenie, w jakim stopniu związek ocen i płci ucznia różni się między poszczególnymi klasami i między oddziałami (np. w zależności od średniego poziomu i zróżnicowania umiejętności przedmiotowych uczniów), a także jaki wpływ na siłę tego związku mają cechy nauczyciela (np. płeć i staż pracy).

Literatura

Allen, J. D. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *78*(5), 218–223. doi: 10.3200/TCHS.78.5.218-223

Bacon, D. R. i Bean, B. (2006). GPA in research studies: an invaluable but neglected opportunity. *Journal of Marketing Education*, *28*(1), 35–42. doi: 10.1177/0273475305284638

Bovaird, J. A. i Koziol, N. A. (2012). Measurement models for ordered-categorical indicators. W: R. H. Hoyle (red.), *Handbook of structural equation modeling* (s. 495–531). New York: Guilford Press.

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation*, *17*(3), 141–159. doi: 10.1080/13803611.2011.597112

Brennan, R. T., Kim, J. S., Wenz-Gross, M. i Siperstein, G. N. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: an analysis of the massachusetts comprehensive assessment system (MCAS). *Harvard Educational Review*, *71*(2), 173–216.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education*, *10*(2), 161–180. doi: 10.1207/s15324818ame1002_4

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Buchmann, C., DiPrete, T. A. i McDaniel, A. (2008). Gender inequalities in education. *Annual Review of Sociology*, *34*(1), 319–337. doi: 10.1146/annurev.soc.34.040507.134719

Burusic, J., Babarovic, T. i Seric, M. (2012). Differences in elementary school achievement between girls and boys: does the teacher gender play a role? *European Journal of Psychology of Education*, *27*(4), 523–538. doi: 10.1007/s10212-011-0093-2

Bye, B. V., Gallicchio, S. J. i Dykacz, J. M. (1985). Multiple-indicator, multiple-cause models for a single latent variable with ordinal indicators. *Sociological Methods & Research*, *13*(4), 487–509. doi: 10.1177/0049124185013004003

Byrnes, J. P. i Miller, D. C. (2007). The relative importance of predictors of math and science achievement: an opportunity–propensity analysis. *Contemporary Educational Psychology*, *32*(4), 599–629. doi: 10.1016/j.cedpsych.2006.09.002

Casillas, A., Robbins, S., Allen, J., Kuo, Y.-L., Hanson, M. A. i Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. *Journal of Educational Psychology*, *104*(2), 407–420. doi: 10.1037/a0027180

Centralna Komisja Edukacyjna (2012). *Sprawozdanie z egzaminu gimnazjalnego*. Warszawa: Centralna Komisja Edukacyjna.

Centralna Komisja Edukacyjna (2016). *Sprawozdanie z egzaminu gimnazjalnego*. Warszawa: Centralna Komisja Edukacyjna.

Conger, D. i Long, M. C. (2010). Why are men falling behind? Gender gaps in college performance and persistence. *The ANNALS of the American Academy of Political and Social Science*, *627*(1), 184–214. doi: 10.1177/0002716209348751

Cornwell, C., Mustard, D. B. i Parys, J. V. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: evidence from primary school. *Journal of Human Resources*, *48*(1), 236–264.

Dalton, B., Ingels, S. J., Downing, J. i Bozick, R. (2007). *Advanced mathematics and science coursetaking in the spring high school senior classes of 1982, 1992, and 2004*. Statistical Analysis Report. NCES 2007-312. Washington: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

DiPrete, T. A. i Jennings, J. L. (2012). Social and behavioral skills and the gender gap in early educational achievement. *Social Science Research*, *41*(1), 1–15. doi: 10.1016/j.ssresearch.2011.09.001

Dolata, R. (red.). (2014). *Czy szkoła ma znaczenie? Analiza zróżnicowania efektywności nauczania na pierwszym etapie edukacyjnym*. (t. 1). Warszawa: Instytut Badań Edukacyjnych.

Dolata, R., Grygiel, P., Jankowska, D. M., Jarnutowska, E., Jasińska-Maciążek, A., Karwowski, M., … Pisarek, J. (2015). *Szkolne pytania. Wyniki badań nad efektywnością nauczania w klasach IV–VI*. Warszawa: Instytut Badań Edukacyjnych.

Dolata, R., Hawrot, A., Humenny, G., Jasińska-Maciążek, A., Koniewski, M., Majkut, P., … Otręba-Szklarczyk, A. (2015). *(Ko)warianty efektywności nauczania. Wyniki badania w klasach IV–VI*. Warszawa: Instytut Badań Edukacyjnych.

Dolata, R. i Sitek, M. (2015). *Raport o stanie edukacji 2014. Egzaminy zewnętrzne w polityce i praktyce edukacyjnej*. Warszawa: Instytut Badań Edukacyjnych.

Driessen, G. (2007). The feminization of primary education: effects of teachers' sex on pupil achievement, attitudes and behaviour. *International Review of Education*, *53*(2), 183–203. doi: 10.1007/s11159-007-9039-y

Duckworth, A. L., Quinn, P. D. i Tsukayama, E. (2012). What No Child Left Behind leaves behind: the roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of Educational Psychology*, *104*(2), 439–451. doi: 10.1037/a0026280

Duckworth, A. L. i Seligman, M. E. P. (2006). Self-discipline gives girls the edge: gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, *98*(1), 198–208. doi: 10.1037/0022-0663.98.1.198

Ekstrom, R. B. (1994). Gender differences in high school grades: an exploratory study. *College Board Report*, *94*(3), 1–30.

Else-Quest, N. M., Hyde, J. S. i Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychological Bulletin*, *136*(1), 103–127. doi: 10.1037/a0018053

Fischer, F. T., Schult, J. i Hell, B. (2013). Sex-specific differential prediction of college admission tests: a meta-analysis. *Journal of Educational Psychology*, *105*(2), 478–488. doi: 10.1037/a0031956

Grygiel, P., Świtaj, P. i Humenny, G. (2015). Zróżnicowane funkcjonowanie pozycji testowych skali stygmatyzacji z Kwestionariusza piętna i dyskryminacji. W: A. Pokropek (red.), *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania* (s. 351–366). Warszawa: Instytut Badań Edukacyjnych.

Grygiel, P., Modzelewski, M. i Pisarek, J. (2016). Academic self-concept and achievement in Polish primary schools: cross-lagged modelling and gender-specific effects. *European Journal of Psychology of Education*. doi: 10.1007/s10212-016-0300-2

Guskey, T. R. (2011). Stability and change in high school grades. *NASSP Bulletin*, *95*(2), 85–98. doi: 10.1177/0192636511409924

Hadjar, A., Krolak-Schwerdt, S., Priem, K. i Glock, S. (2014). Gender and educational achievement. *Educational Research*, *56*(2), 117–125. doi: 10.1080/00131881.2014.898908

Hauser, R. M. i Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. *Sociological Methodology*, 3, 81–117.

Herbert, J. i Stipek, D. (2005). The emergence of gender differences in children's perceptions of their academic competence. *Journal of Applied*

Developmental Psychology, *26*(3), 276–295. doi: 10.1016/j.appdev.2005.02.007

Hicks, B. M., Johnson, W., Iacono, W. G. i McGue, M. (2008). Moderating effects of personality on the genetic and environmental influences of school grades helps to explain sex differences in scholastic achievement. *European Journal of Personality*, *22*(3), 247–268. doi: 10.1002/per.671

Humenny, G. i Grygiel, P. (2015). Wielowymiarowa struktura latentna w perspektywie analizy czynnikowej. W: A. Pokropek (red.), *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania* (s. 130–165). Warszawa: Instytut Badań Edukacyjnych.

Hyde, J. S., Fennema, E. i Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological Bulletin*, *107*(2), 139–155. doi: 10.1037/0033-2909.107.2.139

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A. i Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect: a meta-analysis. *Psychology of Women Quarterly*, *14*(3), 299–324. doi: 10.1111/j.1471-6402.1990.tb00022.x

Jasińska-Maciążek, A. i Modzelewski, M. (2014). Testy osiągnięć szkolnych TOS3: przykład narzędzia skonstruowanego z wykorzystaniem modelu Rascha. *Edukacja*, *127*(2), 85–107

Jones, R. N. (2006). Identification of measurement differences between english and spanish language versions of the mini-mental state examination: detecting differential item functioning using MIMIC modeling. *Medical Care*, *44*(Suppl. 3), S124–S133. doi: 10.1097/01.mlr.0000245250.50114.0f

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin*, *105*(2), 198–214. doi: 10.1037/0033-2909.105.2.198

Kling, K. C., Noftle, E. E. i Robins, R. W. (2013). Why do standardized tests underpredict women's academic performance? The role of conscientiousnes. *Social Psychological and Personality Science*, *4*(5), 600–606. doi: 10.1177/1948550612469038

Konarski, R. (2009). *Modele równań strukturalnych: teoria i praktyka*. Warszawa: Wydawnictwo Naukowe PWN.

Konarzewski, K. (1995). *Problemy i schematy: pierwszy rok nauki szkolnej dziecka*. Warszawa: Żak.

Konarzewski, K. (2003). *Reforma oświaty: podstawa programowa i warunki kształcenia*. Warszawa: Instytut Spraw Publicznych.

Konarzewski, K. (2012). *TIMSS i PIRLS 2011: osiągnięcia szkolne polskich trzecioklasistów w perspektywie międzynarodowej*. Warszawa: Centralna Komisja Egzaminacyjna.

Konarzewski, K. i Bulkowski, K. (red). (2016). *TIMSS 2015. Wyniki międzynarodowego badania osiągnięć czwartoklasistów w matematyce i przyrodzie*. Warszawa: Instytut Badań Edukacyjnych.

Kondratek, B., Skórska, P. i Świst, K. (2015). Wprowadzenie do zróżnicowanego funkcjonowania pozycji testowej. W: A. Pokropek (red.), *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania* (s. 62–90). Warszawa: Instytut Badań Edukacyjnych.

Kondratek, B. i Pokropek, A. (2015). Teoria odpowiedzi na pozycje testowe: jednowymiarowe modele dla cech ukrytych o charakterze ciągłym. W: A. Pokropek (red.), *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania* (s. 15–32). Warszawa: Instytut Badań Edukacyjnych.

Kulon, F. (2015). Wyjaśniające modele IRT. W: A. Pokropek (red.), *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania* (s. 91–105). Warszawa: Instytut Badań Edukacyjnych.

Laidra, K., Pullmann, H. i Allik, J. (2007). Personality and intelligence as predictors of academic achievement: a cross-sectional study from elementary to secondary school. *Personality and Individual Differences*, *42*(3), 441–451. doi: 10.1016/j.paid.2006.08.001

Lietz, P. (2006). A meta-analysis of gender differences in reading achievement at the secondary school level. *Studies in Educational Evaluation*, *32*(4), 317–344. doi: 10.1016/j.stueduc.2006.10.002

Lindberg, S. M., Hyde, J. S., Petersen, J. L. i Linn, M. C. (2010). New trends in gender and mathematics performance: a meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135. doi: 10.1037/a0021276

Logan, S. i Johnston, R. (2009). Gender differences in reading ability and attitudes: examining where these differences lie. *Journal of Research in Reading*, *32*(2), 199–214. doi: 10.1111/j.1467-9817.2008.01389.x

Martínez, J. F., Stecher, B. i Borko, H. (2009). Classroom assessment practices, teacher judgments, and student achievement in mathematics: evidence from the ECLS. *Educational Assessment*, *14*(2), 78–102. doi: 10.1080/10627190903039429

Mattern, K., Sanchez, E. i Ndum, E. (2017). Why do achievement measures underpredict female acade-

mic performance? *Educational Measurement: Issues and Practice*, *36*(1), 47–57. doi: 10.1111/emip.12138

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, *20*(1), 20–32. doi: 10.1111/j.1745-3992.2001.tb00055.x

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: implications for theory and practice. *Educational Measurement: Issues and Practice*, *22*(4), 34–43. doi: 10.1111/j.1745-3992.2003.tb00142.x

Mullola, S., Ravaja, N., Lipsanen, J., Alatupa, S., Hintsanen, M., Jokela, M. i Keltikangas-Järvinen, L. (2012). Gender differences in teachers' perceptions of students' temperament, educational competence, and teachability. *British Journal of Educational Psychology*, *82*(2), 185–206. doi: 10.1111/j.2044-8279.2010.02017.x

Mullis, I. V. S., Martin, M. O., Kennedy, A. M. i Foy, P. (red.). (2007). *PIRLS 2006 international report: IEA's progress in International Reading Literacy Study in primary schools in 40 countries*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Muthén, L. K. i Muthén, B. O. (2012). *Mplus user's guide* (wyd. 7). Los Angles: Muthén&Muthén.

Niemierko, B. (1997). *Między oceną szkolną a dydaktyką: bliżej dydaktyki*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.

Niemierko, B. (2009). *Diagnostyka edukacyjna: podręcznik akademicki*. Warszawa: Wydawnictwo Naukowe PWN.

Noftle, E. E. i Robins, R. W. (2007). Personality predictors of academic outcomes: big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, *93*(1), 116–130. doi: 10.1037/0022-3514.93.1.116

Preckel, F., Holling, H. i Vock, M. (2006). Academic underachievement: relationship with cognitive motivation, achievement motivation, and conscientiousnes. *Psychology in the Schools*, *43*(3), 401–411. doi: 10.1002/pits.20154

Rakoczy, K., Klieme, E., Bürgermeister, A. i Harks, B. (2008). The interplay between student evaluation and instruction: grading and feedback in mathematics classrooms. *Zeitschrift für Psychologie / Journal of Psychology*, *216*(2), 111–124. doi: 10.1027/0044-3409.216.2.111

Randall, J. i Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education*, *26*(7), 1372–1380. doi: 10.1016/j.tate.2010.03.008

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PLoS ONE*, *7*(7), e39904. doi: 10.1371/journal.pone.0039904

Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: a qualitative study. *Teaching and Teacher Education*, *27*(2), 472–482. doi: 10.1016/j.tate.2010.09.017

Richardson, M., Abraham, C. i Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353–387. doi: 10.1037/a0026838

Robinson, J. P. i Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *48*(2), 268–302. doi: 10.3102/0002831210372249

Ross, J. A. i Gray, P. (2008). Alignment of scores on large-scale assessments and report-card grades. *The Alberta Journal of Educational Research*, *54*(3), 327–341.

Ross, J. A. i Kostuch, L. (2011). Consistency of report card grades and external assessments in a Canadian province. *Educational Assessment, Evaluation and Accountability*, *23*(2), 159–180. doi: 10.1007/s11092-011-9117-3

Serbin, L. A., Stack, D. M. i Kingdon, D. (2013). Academic success across the transition from primary to secondary schooling among lower-income adolescents: understanding the effects of family resources and gender. *Journal of Youth and Adolescence*, *42*(9), 1331–1347. doi: 10.1007/s10964-013-9987-4

Shibley Hyde, J. i Kling, K. C. (2001). Women, motivation, and achievement. *Psychology of Women Quarterly*, *25*(4), 364–378. doi: 10.1111/1471-6402.00035

Skórska, P. i Świst, K. (2014). Wielkość efektu płci w wewnątrzszkolnych i zewnątrzszkolnych wskaźnikach osiągnięć ucznia. W: B. Niemierko i M. K. Szmigel (red.), *Diagnozy edukacyjne. Dorobek i nowe zadania* (s. 89–103). Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Spilt, J. L., Koomen, H. M. Y. i Jak, S. (2012). Are boys better off with male and girls with female teachers? A multilevel investigation of measurement invariance and gender match in teacher–student relationship quality. *Journal of School Psychology*, *50*(3), 363–378. doi: 10.1016/j.jsp.2011.12.002

Spinath, B., Eckert, C. i Steinmayr, R. (2014). Gender differences in school success: what are the roles of students' intelligence, personality and motiva-

tion? *Educational Research*, *56*(2), 230–243. doi: 10.1080/00131881.2014.898917

Spinath, B., Harald Freudenthaler, H. i Neubauer, A. C. (2010). Domain-specific school achievement in boys and girls as predicted by intelligence, personality and motivation. *Personality and Individual Differences*, *48*(4), 481–486. doi: 10.1016/j.paid.2009.11.028

Steinmayr, R. i Spinath, B. (2008). Sex differences in school achievement: what are the roles of personality and achievement motivation? *European Journal of Personality*, *22*(3), 185–209. doi: 10.1002/per.676

Südkamp, A., Kaiser, J. i Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. doi: 10.1037/a0027627

Świst, K. i Skórska, P. (2016). Trafność prognostyczna wskaźników osiągnięć gimnazjalnych względem wyników maturalnych dziewcząt i chłopców. *Edukacja*, *139*(4), 42–60.

Trapmann, S., Hell, B., Weigand, S. i Schuler, H. (2007). The validity of school grades for academic achievement. A meta-analysis. *Zeitschrift Fur Padagogische Psychologie*, *21*(1), 11–27. doi: 10.1024/1010-0652.21.1.11

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O. i Baumert, J. (2006). Tracking, grading, and student motivation: using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, *98*(4), 788–806. doi: 10.1037/0022-0663.98.4.788

Vecchione, M., Alessandri, G. i Marsicano, G. (2014). Academic motivation predicts educational attainment: does gender make a difference? *Learning and Individual Differences*, 32, 124–131. doi: 10.1016/j.lindif.2014.01.003

Voyer, D. i Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. doi: 10.1037/a0036620

Weis, M., Heikamp, T. i Trommsdorff, G. (2013). Gender differences in school achievement: the role of self-regulation. *Educational Psychology*, 4, 442. doi: 10.3389/fpsyg.2013.00442

Woods, C. M. (2008). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. doi: 10.1177/0146621607314044

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1–27. doi: 10.1080/00273170802620121

Zhu, M. i Urhahne, D. (2014). Teachers' judgements of students' foreign-language achievement. *European Journal of Psychology of Education*. doi: 10.1007/s10212-014-0225-6

Artykuł powstał na podstawie danych pochodzących z badania *Szkolne uwarunkowania efektywności kształcenia*, przeprowadzonego w Instytucie Badań Edukacyjnych w ramach projektu systemowego „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego" współfinansowanego ze środków Europejskiego Funduszu Społecznego w ramach Programu Operacyjnego Kapitał Ludzki 2007–2013. Priorytet III: Wysoka jakość systemu oświaty.

Tekst złożony 11 kwietnia 2017 r., zrecenzowany 8 maja 2017 r., przyjęty do druku 17 maja 2017 r.

Can primary school teachers grade the literacy level of girls and boys objectively?

Research suggests that primary school teachers grade the reading and writing skills of girls higher than of boys, even when they have the same level of ability. In this article, we try to verify this hypothesis. We analysed (a) teachers' grades (prepared particularly for the purpose of this research), and (b) achievement test results, which controlled for the differential item functioning due to gender. We used multiple indicators, the multiple causes (MIMIC) model on a representative sample of 4144 Polish third-grade students. Teachers graded girls' ability levels higher. However, the difference disappeared when we controlled for the test-based ability level. Therefore, we conclude that teachers can grade literacy levels without gender bias.

Keywords: literacy; multiple indicators multiple causes; MIMIC model; grading; gender; bias; differential item functioning.