Edukacja 2016, 4(139), 42–60
ISSN 0239-6858

Trafność prognostyczna wskaźników osiągnięć gimnazjalnych względem wyników maturalnych dziewcząt i chłopców

Karolina Świst, Paulina Skórska

Instytut Badań Edukacyjnych*

Artykuł został poświęcony różnicom w trafności prognostycznej wskaźników osiągnięć gimnazjalnych (oceny, średnia ocen, wyniki egzaminu) w analizach przewidywania wyników maturalnych wśród dziewcząt i chłopców. Na różnice w wynikach może wpływać wiele czynników – psychologicznych, społecznych oraz związanych z właściwościami arkuszy testowych. Można więc przyjąć hipotezę o różnej mocy prognostycznej tych wskaźników wśród dziewcząt i chłopców. Przeanalizowane zostały dwie kohorty: osób zdających egzamin gimnazjalny w latach 2011 i 2012 oraz maturę w latach 2014 i 2015. Analizy przeprowadzono przy pomocy hierarchicznych modeli liniowych oraz modelowania IRT. Wyniki wskazują na różnice w funkcjonowaniu wskaźników osiągnięć w zależności od płci oraz dziedziny egzaminu (język polski, matematyka). Wyniki egzaminów i oceny szkolne pozwalają przewidywać sukces ucznia, jednak różnice w trafności prognostycznej wśród chłopców i dziewcząt są niewielkie.

Słowa kluczowe: badania edukacyjne, trafność prognostyczna, płeć, hierarchiczne modelowanie liniowe, ocenianie.

Wskaźniki osiągnięć edukacyjnych uczniów powinny pozwalać na przewidywanie późniejszych sukcesów. Jednak jakość predykcji może być uwarunkowana płcią ucznia. Aby pomiar był sprawiedliwy (*fair*), oba wskaźniki powinny działać jednakowo dobrze w grupie dziewcząt i chłopców. Jeśli wskaźniki osiągnięć uczniów są obciążone płciowo, to zaprojektowanie skutecznych oddziaływań edukacyjnych na podstawie analiz trafności prognostycznych może być obciążone.

W literaturze wskazuje się, że istnieje rozbieżność osiągnięć szkolnych wśród dziewcząt i chłopców: zarówno ocen szkolnych, jak i wyników standaryzowanych testów (w tym egzaminów zewnętrznych). Dziewczęta mają zazwyczaj wyższe oceny szkolne (oraz ich średnią), natomiast ich wyniki w testach osiągnięć są zwykle niższe niż wyniki chłopców (Hyde i Kling, 2001; Konarzewski, 1996; Mau i Lynn, 2001; Noftle i Robins, 2007; Perkins, Kleiner, Roey i Brown, 2004). Jednocześnie nie stwierdzono znaczących różnic w ilorazie inteligencji pomiędzy dziewczętami a chłopcami (Cole, 1997; Pomerantz, Altermatt i Saxon, 2002). Trafność ocen szkolnych i wyników testów jest w różnym stopniu zagrożona podatnością tych wskaźników na problem wariancji niezwiązanej z mierzoną

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

© Instytut Badań Edukacyjnych

cechą (*construct irrelevant variance*, CIV; Haladyna i Downing, 2004). Ten efekt rozbieżności ocen oraz wyników egzaminów zewnętrznych uzyskiwanych przez dziewczęta i chłopców próbuje się tłumaczyć wieloma czynnikami: psychologicznymi, społeczno-kulturowymi oraz stanowiącymi ich interakcję z właściwościami narzędzi pomiarowych. Ponieważ wpływ tych czynników na obniżenie trafności wskaźników osiągnięć został dobrze udokumentowany w literaturze naukowej, można spodziewać się różnic międzypłciowych w ich trafności prognostycznej. Hipoteza ta nie była częstym przedmiotem badań, podobnie jak samo zagadnienie trafności prognostycznej ocen i wyników testów na niższych niż kształcenie wyższe etapach edukacji. Problem został podjęty w badaniu opisanym w tym artykule.

Źródła stronniczości ocen szkolnych

Wśród źródeł potencjalnej stronniczości ocen szkolnych ze względu na płeć można wskazać czynniki: psychologiczne (indywidualne), społeczno-kulturowe oraz specyficzne dla dziedziny nauczania. Wyższe oceny dziewcząt w porównaniu do ocen chłopców mogą wynikać z większego natężenia takich cech psychologicznych, jak sumienność, samokontrola i samodyscyplina (Donnellan i Lucas, 2008; Schmitt, Realo, Voracek i Allik, 2008; Srivastava, John, Gosling i Potter, 2003). Tego typu cechy mogą się przekładać na systematyczne odrabianie pracy domowej i lepsze przygotowywanie się do lekcji przez dziewczęta (Mac an Ghaill, 1994). Badania wskazują też na wyższą motywację wewnętrzną u dziewcząt i niższą motywację zewnętrzną w porównaniu do chłopców (Vecchione, Alessandri i Marsicano, 2014). Czynnikiem wyjaśniającym niższe oceny chłopców mogą być zaburzenia i trudności rozwojowe, w tym problemy z koncentracją (Zill i West, 2001).

Jeśli chodzi o czynniki społeczno-kulturowe, to w wielu aspektach funkcjonowania polskiego systemu szkolnego pomija się znaczenie równości płci i powiela role oraz stereotypy z nimi związane (Pankowska, 2005). Przykładowo Lucyna Kopciewicz (2008) zwróciła uwagę, że stereotypowo sukcesy dziewcząt są tłumaczone ich pracowitością i wytrwałością, natomiast chłopców – inteligencją i błyskotliwością. Nauczyciele przydzielają dziewczętom i chłopcom inne typy zadań – zadania będące wyzwaniami są stawiane chłopcom a rutynowe, proste zadania – dziewczętom. Inna jest też gotowość do pracy z chłopcami i dziewczętami oraz przekazywanie im informacji zwrotnych na temat poziomu wykonania (Gromkowska-Melosik, 2011; Suchocka, 2011). Ponieważ oceny w większym stopniu niż wyniki testów są wrażliwe na kryteria pozadydaktyczne, decydującą rolę mają też przekonania nauczycieli dotyczące zachowania dzieci (Drost-Rudnicka, 2012; Konarzewski, 1996; Niemierko, 2001). Dziewczęta postrzegane są jako grzeczniejsze, a przyznawane im oceny mają być „nagrodą" za dobre zachowanie. Dotyczy to przede wszystkim dziewcząt z grupy dzieci o niskim statusie społeczno-ekonomicznym (Konarzewski, 1996). Wskazane zjawiska mogą wchodzić w interakcję z dziedziną nauczania, gdzie różnice ocen na korzyść dziewcząt rysują się bardzo wyraźnie w przypadku języka polskiego i kursów językowych, a słabiej w przedmiotach ścisłych, zwłaszcza matematyce (Skórska i Świst, 2014; Voyer i Voyer, 2014).

Źródła stronniczości wyników testów

W przypadku wyników testów (egzaminów zewnętrznych) możemy wskazać na podobne potencjalne źródła stronniczości jak w przypadku oceniania wewnątrzszkolnego i wydzielić czynniki: psychologiczne

(indywidualne), społeczno-kulturowe i wynikające z interakcji czynników psychologicznych oraz właściwości psychometrycznych arkuszy testowych. Badania dowodzą, że chłopcy mogą uzyskiwać wyższe wyniki testów ze względu na wyższą pewność siebie (Preckel, Goetz, Pekrun i Kleine, 2008). Podkreśla się, że dziewczęta mogą być w mniejszym stopniu socjalizowane i przygotowywane do osiągania sukcesów (Fried-Buchalter, 1997; Steinmayr i Spinath, 2008). Metaanaliza Sabine Severiens i Geerta ten Dama (1998) wskazała, że dziewczęta mają wyższy stopień strachu przed porażką, choć nie wykazano różnic w potrzebie osiągnięć między płciami (Costa, Terracciano i McCrae, 2001).

Innym czynnikiem może być zróżnicowany wśród chłopców i dziewcząt styl odpowiadania na pytania testowe, wynikający z uwarunkowań psychologicznych. W badaniach wskazuje się na większą tendencję do zgadywania wśród chłopców (Slakter, Koehler, Hampton i Grennell, 1971) oraz większą tendencję do omijania (pozostawiania zadań bez odpowiedzi) wśród dziewcząt (Ben-Shakhar i Sinai, 1991; Pekkarinen, 2014) ze względu na aktywizację lęku przed porażką, zwłaszcza w testach matematycznych (Schrader i Ansley, 2006). Klimat społeczno-kulturowy w danym kraju może przekładać się na wyniki testów. Jak wykazał Luigi Guiso ze współpracownikami (Guiso, Monte, Sapienza i Zingales, 2008), w krajach o wysokim wskaźniku równości płci różnica między wynikami testów (wykorzystywanych w badaniach PISA) między dziewczętami a chłopcami jest niewielka. Wreszcie, wiele czynników związanych z właściwościami psychometrycznymi testów wpływa na wyniki uzyskiwane przez dziewczęta i chłopców. Różnice mogą być uwarunkowane formatem zadania egzaminacyjnego oraz stopniem skomplikowania testu. Chłopcy osiągają wyższe wyniki w zadaniach zamkniętych wielokrotnego wyboru, dziewczęta natomiast w zadaniach otwartych (Willingham i Cole, 1997). Efekt ten pojawia się zwłaszcza w grupie uczniów najzdolniejszych (DeMars, 1998), co może wynikać ze zróżnicowanej motywacji, natężenia lęku oraz skłonności do podejmowania ryzyka wśród dziewcząt i chłopców.

Trafność prognostyczna

Trafność prognostyczna (Carmines i Zeller, 1979; Cronbach i Meehl, 1955) pozwala odpowiedzieć na pytanie, w jakim stopniu punkty uzyskane na danej skali lub w teście umożliwiają przewidywanie wyników uzyskanych za pomocą innego narzędzia pomiarowego. Innymi słowy, wartość tego wskaźnika mówi nam o tym, w jakim stopniu wynik osoby w teście może być przewidywany na podstawie wyników innego, wykonanego wcześniej testu. W najnowszym wydaniu *The standards for educational and psychological testing* (AERA, APA i NCME, 2014), oprócz trafności prognostycznej zwraca się uwagę również na pojęcie trafności konsekwencyjnej (Messick, 1989), która określa, do jakiego stopnia decyzje podjęte na podstawie pomiaru (oraz idące za nimi konsekwencje) można uznać za trafne (więcej w: Niemierko, 1999).

Klasyczne podejścia sugerują pomiar trafności prognostycznej przy pomocy wielkości współczynnika korelacji testu wraz z wybranym kryterium zewnętrznym (Carmines i Zeller, 1979). Obok metod opartych na współczynnikach korelacji w literaturze wymienia się również metody oparte na analizie dyskryminacji (przewidywania przynależności do grupy na podstawie zestawu predyktorów), modelowaniu równań strukturalnych oraz analizie regresji (Mitchell, 1990). Elazar Pedhazur i Liora Schmelkin (1991) oraz Linda Crocker i James Algina (1986) zarekomendowali stosowanie metod opartych na regresji. Wskazali, że metody

korelacyjne dają jedynie ogólne wskazania co do wielkości i natury relacji pomiędzy wskaźnikami a ich zewnętrznymi kryteriami[1].

Badania nad trafnością prognostyczną wskaźników osiągnięć uczniów prowadzone w Stanach Zjednoczonych wykorzystują wyniki standaryzowanych testów zewnętrznych (np. SAT lub ACT) oraz średnią ocen szkolnych ze szkoły średniej. Na tej podstawie przewiduje się osiągnięcia edukacyjne na studiach, mierzone przy pomocy średniej ocen z pierwszego roku studiów lub z całego toku studiów. Dowody na trafność prognostyczną wskaźników osiągnięć szkolnych na niższych etapach edukacji (szkoła podstawowa i gimnazjum) przedstawiono w nielicznych pracach (np. Byrnes i Miller, 2007; Casillas i in., 2012).

Wyniki badań prowadzonych w najpopularniejszym nurcie sugerują, że uzyskana w szkole średniej średnia ocen w połączeniu z wynikami standaryzowanych testów pozwala w największym stopniu na przewidywanie wyników uzyskiwanych na studiach (Hezlett i in., 2001; Kobrin, Patterson, Shaw, Mattern i Barbuti, 2008; Zahner, Ramsaran i Steedle, 2012). Zastosowanie dwóch predyktorów pozwala na uchwycenie różnych aspektów osiągnięć uczniów, a pojedyncze wskaźniki osiągnięć uczniów mogą być nadmiernie obciążone. Ocenianie służy nie tylko do informowania o poziomie umiejętności, lecz także do motywowania uczniów (jako kara lub nagroda odzwierciedlająca np. złe lub dobre zachowanie), dlatego jego funkcja informacyjna może być pomniejszona. Standaryzowane testy, w tym te wykorzystywane do egzaminowania zewnętrznego są pozbawione tego obciążenia, jednak ich wyniki są wrażliwe na status społeczno-ekonomiczny ucznia (Atkinson i Geiser, 2009; Geiser i Santelices, 2007; Rothstein, 2004; Willingham, Pollack i Lewis, 2002).

W wielu wypadkach średnia ocen niesie więcej informacji niż pojedyncza ocena, a co za tym idzie – powinna być bardziej rzetelnym predyktorem osiągnięć edukacyjnych. Jednakże zarówno w zagranicznych (Schuler, Funke i Baron-Boldt, 1990), jak i polskich (Skórska, Świst i Szaleniec, 2014) badaniach wskazano na to, że do przewidywania późniejszych wyników z matematyki, pojedyncze oceny z tego przedmiotu mogą być bardziej prognostyczne niż średnia ocen.

W artykule przeanalizujemy różnice w trafności prognostycznej wskaźników osiągnięć gimnazjalnych (oceny, średnia ocen, wyniki egzaminu gimnazjalnego) względem osiągnięć maturalnych wśród dziewcząt i chłopców. Dane zostały zebrane dla dwóch kohort: osób zdających egzamin gimnazjalny w 2011 r. i maturę w 2014 r., oraz zdających egzamin gimnazjalny w 2012 r. i maturę w 2015 r. Na ich podstawie będą przewidywane wyniki z matury z języka polskiego i matematyki. Do analiz zostaną wykorzystane hierarchiczne modele liniowe oraz modelowanie IRT (*item response theory*).

Testowaniu podlegały następujące hipotezy postawione na podstawie stanu dotychczasowej wiedzy:

- Hipoteza 1: Wskaźniki osiągnięć uczniów wykazują się zróżnicowaną trafnością prognostyczną na etapie kształcenia gimnazjalnego.
- Hipoteza 2: Trafność prognostyczna wskaźników osiągnięć szkolnych zależy od płci oraz dziedziny nauczania (w tym wypadku języka polskiego i matematyki).

Metodologia

Dane

W analizie oparto się na podłużnych wynikach uczniów zdających egzamin gimnazjalny, a trzy lata później egzamin maturalny. Wykorzystano połączone

[1] Przy pomocy metody korelacyjnej nie można np. sprawdzić relacji pomiędzy wieloma predyktorami a zewnętrznymi kryteriami na raz, co możliwe jest dzięki wykorzystaniu regresji wielorakiej.

wyniki egzaminu maturalnego (lata 2014 i 2015[2]) oraz egzaminu gimnazjalnego (lata 2011 oraz 2012[3]) pochodzące z Centralnej Komisji Egzaminacyjnej (CKE) oraz dane dotyczące ocen szkolnych wystawionych za pierwszy semestr trzeciej klasy gimnazjum (lata 2011 oraz 2012). Dane dotyczące ocen pochodzą z drugiego i trzeciego etapu badań zrównujących przeprowadzonych przez Zespół Analiz Osiągnięć Uczniów Instytutu Badań Edukacyjnych (zob. Szaleniec i in., 2013; 2015). Zebrano dane dotyczące następujących przedmiotów szkolnych: języka polskiego, matematyki, historii, wiedzy o społeczeństwie, biologii, chemii, fizyki oraz geografii.

Próbę w badaniach zrównujących dobierano w sposób losowy, operatem była lista szkół udostępniona przez CKE w latach 2011 i 2012. Przeprowadzono losowanie, które miało charakter warstwowy (proporcjonalne do liczebności szkoły na poziomie oddziału szkolnego oraz wielopoziomowe; więcej w: Szaleniec i in., 2013). Wielkości prób przedstawiono w Tabeli 1. Znajdują się w niej informacje o liczbie uczniów, dla których możliwe było połączenie danych o wynikach egzaminu gimnazjalnego, egzaminu maturalnego oraz informacji o ocenach z badań zrównujących. Już na tym etapie może zwrócić uwagę znacznie niższa wielkość analizowanej próby w kohorcie 2012–2015. Należy zauważyć, że w 2015 r. wprowadzono dwie formuły egzaminu maturalnego, co prawdopodobnie mogło wpłynąć na trudności w łączeniu danych. Niższa wielkość próby może przekładać się na większą niepewność oszacowań w późniejszych analizach (np. szersze przedziały ufności dla współczynników regresji) w porównaniu do wyników dla starszej kohorty.

Procedura

Zastosowano modelowanie wielopoziomowe (*hierarchical linear modeling*) oraz oszacowania poziomu umiejętności uzyskane przy pomocy modeli IRT. W badaniach edukacyjnych rekomenduje się zastosowanie modelowania wielopoziomowego (np. Bryk i Raudenbush, 1992). Za zastosowaniem tej techniki przemawia zgrupowanie uczniów w oddziałach: uczniowie znajdujący się w tych samych oddziałach mogą osiągać bardziej zbliżone wyniki niż uczniowie w wielu oddziałach. Wyniki mogą się różnić w poszczególnych oddziałach np. pod względem nauczyciela danego przedmiotu lub ogólnym poziomem umiejętności danego oddziału oraz dostosowanego do niego sposobu oceniania (zob. teoria relatywnego oceniania szkolnego; Goldman i Hewitt, 1975). Wymienione czynniki mogą łamać założenie o niezależności obserwacji wykorzystanych w analizie. Zignorowanie hierarchicznej

[2] Wyłącznie w nowej formule.

[3] Należy zauważyć, że w analizowanych rocznikach zmieniła się podstawa programowa. Dlatego w 2011 r. przeanalizowano wyłącznie zadania matematyczne z części matematyczno-przyrodniczej oraz zadania z języka polskiego z części humanistycznej.

Tabela 1
Wielkości prób dla poszczególnych kohort zdających poszczególne przedmioty

Kohorta	Przedmiot	Wielkość próby	Liczba oddziałów[(a)]
2011–2014	Język polski	4 378	431
	Matematyka	4 368	443
2012–2015	Język polski	718	74
	Matematyka	718	74

[(a)] Losowano jedną klasę z każdej szkoły, dlatego liczba klas jest równa liczbie szkół.

struktury danych (czyli wykorzystanie zwykłej regresji liniowej do analizy danych) prowadziłoby do złamania założenia o niezależności składnika błędu na poziomie oddziału szkolnego (Hox, 2010).

W regresji za pierwszy poziom przyjęto poziom ucznia, a za drugi – poziom oddziału szkolnego. Wykorzystano modele z losową stałą (*random intercept*) oraz losowym nachyleniem (*random slope*; zob. Hox, 2010; Snijders i Bosker, 1999). Do raportowania wyników wykorzystano standaryzowane współczynniki regresji (ze względu na możliwość porównywania ze sobą siły predyktorów w modelach łączonych) oraz brzegowe i warunkowe R^2 do raportowania poziomu wyjaśnionej wariancji.

W klasycznych pracach (np. Snijders i Bosker, 1999) poświęconych modelowaniu wielopoziomowemu zwrócono uwagę na problematyczność oszacowania miar wyjaśnionej wariancji (R^2) na pierwszym i drugim poziomie analizy. Miary były trudne do przeliczenia (wariancja nie jest stała dla wszystkich poziomów kowariantów) szczególnie w przypadku modeli z losowym nachyleniem i losową stałą. Roel Snijders i Tom Bosker (1999) zaproponowali przybliżenie poziomu wyjaśnionej wariancji dla modeli z losowym nachyleniem oraz losową stałą poprzez przeliczenie R^2 dla modeli z losowym nachyleniem (czyli z tą samą specyfikacją efektów stałych oraz pominięciem efektów losowych). Wykazano jednak, że podejście zaprezentowane przez Snijdersa i Boskera może być problematyczne (Nakagawa i Schielzeth, 2013). Oszacowania wyjaśnionej wariancji na pierwszym i drugim poziomie analizy mogą spadać po dodaniu kolejnych predyktorów. Poza tym niejasne jest, czy rozwiązanie Snijdersa i Boskera można uogólnić na więcej niż dwa poziomy analizy. Shinichi Nakagawa i Holger Schielzeth (2013) przedstawili własne rozwiązanie polegające na wyliczeniu dwóch rodzajów R^2 – brzegowego (*marginal*) oraz warunkowego (*conditional*). Brzegowe R^2 określa poziom wariancji wyjaśnionej przez efekty stałe, warunkowe R^2 – poziom wariancji wyjaśnionej zarówno przez efekty stałe (równanie 1) jak i efekty losowe (równanie 2). Oba rodzaje R^2 mogą zostać wyliczone przy pomocy pakietu MuMIn i środowiska R (Bartoń, 2015; Johnson, 2014).

$$R^2_{GLMM(\mathrm{m})} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{l=1}^{u}\sigma_l^2 + \sigma_e^2 + \sigma_d^2} \qquad (1)$$

$$R^2_{GLMM(c)} = \frac{\sigma_f^2 + \sum_{l=1}^{u}\sigma_l^2}{\sigma_f^2 + \sum_{l=1}^{u}\sigma_l^2 + \sigma_e^2 + \sigma_d^2}, \qquad (2)$$

gdzie: σ_f^2 wariancja efektów stałych; σ_l^2 – wariancja związana z *l-tym* losowym czynnikiem (*variance component of* the *l-th random factor*) σ^2 – wariancja błędu (*residual variance*); σ_d^2 – wariancja specyficzna dla rozkładu (*distribution-specific variance*).

W analizie wykorzystano oszacowania poziomu umiejętności (θ) przy pomocy modeli IRT, zamiast prostej sumy punktów. Tego rodzaju oszacowania pozwalają na kontrolę różnych właściwości psychometrycznych testów, szczególnie ich zróżnicowanej trudności. Wykorzystano model dwuparametryczny (2PLM; Birnbaum, 1968) dla zadań zamkniętych oraz model GRM (*graded response model*; Samejima, 1968) dla zadań wielokategorialnych. Przy pomocy obu modeli wyliczono punktowe oszacowania poziomu umiejętności poprzez estymator EAP (*expected a posteriori*)[4] dla:

[4] Czytelnik może zapoznać się z dokładnym omówieniem sposobu szacowania poziomu umiejętności w paradygmacie IRT w pracach Frederica Lorda (1986) oraz Roberta Mislevy'ego (1986).

Tabela 2
Specyfikacja modeli wykorzystanych w analizie

Model	Zmienna zależna	Zmienne niezależne (predyktory)
0 (bazowy)	Poziom umiejętności oszacowany na podstawie wyniku egzaminu maturalnego	Brak
1		Pojedyncza ocena
2		Średnia ocen
3		Poziom umiejętności (oszacowany na podstawie egzaminu gimnazjalnego)
4		Pojedyncza ocena i poziom umiejętności
5		Średnia ocen i poziom umiejętności

- zadań matematycznych z części matematyczno-przyrodniczej egzaminu gimnazjalnego w 2011 r.,
- zadań językowych z części humanistycznej egzaminu gimnazjalnego w 2011 r.,
- egzaminu gimnazjalnego z matematyki w 2012 r.,
- egzaminu gimnazjalnego z języka polskiego w 2012 r.,
- wyników matury z języka polskiego (lata 2014 i 2015),
- wyników matury z matematyki (lata 2014 i 2015).

Przetestowano sześć modeli wielopoziomowych, których specyfikacja znajduje się w Tabeli 2. W pierwszych trzech modelach wykorzystano pojedyncze predyktory: ocenę z języka polskiego lub z matematyki wystawioną na pierwszy semestr gimnazjum, średnią ocen z gimnazjum[5] oraz poziom umiejętności oszacowany na podstawie wyniku egzaminu gimnazjalnego. Dwa kolejne modele wykorzystują kombinację predyktorów – jednego opierającego się na ocenianiu wewnątrzszkolnym (ocena lub średnia) oraz oszacowanego poziomu umiejętności.

[5] Język polski, matematyka, historia, wiedza o społeczeństwie (WOS), biologia, geografia, fizyka, chemia. W 2012 r. wykorzystano średnią z siedmiu przedmiotów, ponieważ WOS w znaczący sposób obniżał rzetelność średniej (jego usunięcie spowodowało zwiększenie się rzetelności skali mierzonej przy pomocy α-Cronbacha z 0,26 do 0,91). Problem ten nie pojawił się w 2011 r. – rzetelność średniej ocen w 2011 r. wyniosła ok. 0,9.

Wyniki

Statystyki opisowe

Analizę rozpoczynamy od przedstawienia rzetelności poszczególnych wskaźników osiągnięć wyliczonych przy pomocy α-Cronbacha oraz podstawowych różnic (i ich istotności statystycznej) we wskaźnikach osiągnięć uczniów wśród dziewcząt i chłopców. Istotność statystyczną dla egzaminów wyliczono przy pomocy testu *t* z poprawką Welcha na nierówność wariancji. Należy zwrócić również uwagę na to, że w maturze z języka polskiego uczeń pisze rozprawkę na jeden z dwóch wybranych przez siebie tematów, które nie muszą być treściowo ekwiwalentne oraz mierzyć tych samych umiejętności (zob. Szaleniec i in., 2015). Dlatego rzetelność egzaminu maturalnego z języka polskiego wyliczono osobno dla poszczególnych tematów wypracowań. Porównanie podstawowych statystyk opisowych dla poszczególnych wskaźników osiągnięć uczniów (zarówno zmiennych niezależnych – wskaźników oceniania wewnętrznego, jak i zewnętrznego na poziomie gimnazjum oraz zależnych, czyli wyników matury) przedstawiono w Tabeli 3.

W obu kohortach egzaminy z języka polskiego mają znacznie mniejszą rzetelność niż egzaminy z matematyki. Dla kohorty 2012–2015 rzetelność egzaminu maturalnego z języka polskiego (w nowej formule) jest jeszcze niższa niż dla kohorty

Tabela 3
Statystyki opisowe dla wykorzystanych w analizie wskaźników osiągnięć

Kohorta	Przedmiot	Zmienna	α-Cronbacha	*M*		*SD*	
				Chł.	Dz.	Chł.	Dz.
2011–2014	Matematyka	Ocena	nd.	3,61	3,60	1,05	0,99
		Wynik egzaminu gimnazjalnego (tylko matematyka)***	0,86	9,91	8,61	2,95	3,22
		Wynik matury***	0,90	31,59	28,06	11,54	11,25
	Język polski	Ocena***	nd.	3,57	4,00	0,88	0,89
		Wynik egzaminu gimnazjalnego (tylko język polski)***	0,72	22,97	24,22	5,84	5,81
		Wynik matury***	Temat I – 0,74 Temat II – 0,74	38,85	40,21	10,53	11,0
	Średnia 8 ocen z gimnazjum	Średnia***	0,90	3,71	3,90	0,80	0,76
2012–2015	Matematyka	Ocena	nd.	3,60	3,65	1,10	1,03
		Wynik egzaminu gimnazjalnego**	0,83	19,70	18,11	6,91	7,07
		Wynik matury	0,89	29,89	28,21	12,91	11,76
	Język polski	Ocena***	nd.	3,43	3,94	0,92	0,90
		Wynik egzaminu gimnazjalnego***	0,71	23,91	25,31	4,20	3,90
		Wynik matury***	Temat I – 0,63 Temat II – 0,67	28,80	30,50	6,10	5,54
	Średnia 7 ocen z gimnazjum	Średnia**	0,91	3,61	3,80	0,84	0,78

Różnice istotne statystycznie na poziomach: * $p < 0,05$; ** $p < 0,005$; *** $p < 0,001$.

2011–2014. Co ciekawe, średnia ocen wydaje się znacznie bardziej rzetelnym predyktorem (α-Cronbacha w obu kohortach wyniosła ok. 0,9) niż egzaminy zewnętrzne.

Międzypłciowe różnice we wskaźnikach osiągnięć są w zdecydowanej większości wypadków spójne między analizowanymi latami. Dla języka polskiego ocena jest w sposób istotny statystycznie wyższa dla dziewcząt, natomiast zróżnicowanie ocen jest na podobnym poziomie dla obu płci. Wynik egzaminu gimnazjalnego jest w sposób istotny statystycznie wyższy dla dziewcząt, natomiast odchylenie standardowe wyższe wśród chłopców. Dziewczęta uzyskują też istotnie statystyczne, wyższe wyniki z matury z języka polskiego. Odchylenie standardowe dla matury z języka polskiego jest zróżnicowane między latami – wyższe dla dziewcząt w 2014 r. i dla chłopców w 2015 r.

W przypadku matematyki nie ma istotnej statystycznie różnicy w średnich ocenach dziewcząt i chłopców, natomiast odchylenie standardowe jest wyższe wśród chłopców. Wyniki egzaminu gimnazjalnego są w sposób istotny statystycznie wyższe dla chłopców, jednak odchylenie standardowe jest wyższe dla dziewcząt. Podobna sytuacja (statystycznie istotny, wyższy średni wynik dla chłopców oraz wyższe odchylenie

standardowe dla dziewcząt) pojawia się dla matury z matematyki.

Średnia ocen jest w sposób istotny statystycznie wyższa dla dziewcząt niż dla chłopców, jednak odchylenie standardowe jest wyższe dla chłopców.

Analiza współczynników nachylenia

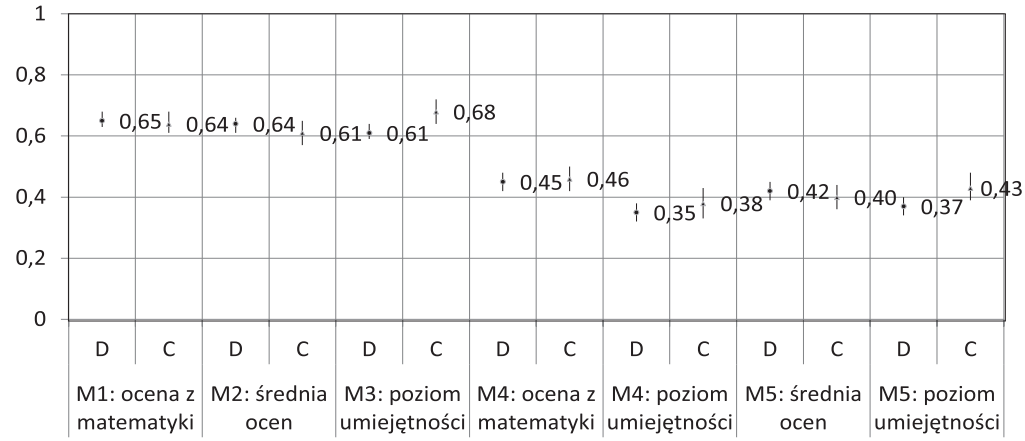
W kolejnych częściach opiszemy analizę standaryzowanych współczynników regresji hierarchicznej dla sześciu analizowanych modeli dla poszczególnych przedmiotów oraz kohort. Uzasadnienie wyboru standaryzowanych współczynników regresji jako wskaźników trafności prognostycznej zostało opisane w sekcji „Trafność prognostyczna". Wykorzystanie metody regresji zamiast korelacji umożliwia analizowanie związku wielu predyktorów i zmiennej wynikowej na raz. Współczynniki nachylenia zostały wystandaryzowane, by móc porównywać ze sobą siłę związku pomiędzy predyktorami mierzonymi na różnych skalach (np. średnia ocen oraz poziom umiejętności uczniów oszacowany na podstawie egzaminu gimnazjalnego). Analizę trafności prognostycznej rozpoczęto od porównania dobroci dopasowania modeli hierarchicznych oraz modeli regresji liniowej przy pomocy testu LR (*likelihood ratio*) dostępnego w programie Stata. Dla wszystkich analizowanych modeli opisanych w Tabeli 2, w obu kohortach i w przypadku obu przedmiotów, modele hierarchiczne okazały się lepiej dopasowane do danych niż modele regresji liniowej.

Matematyka

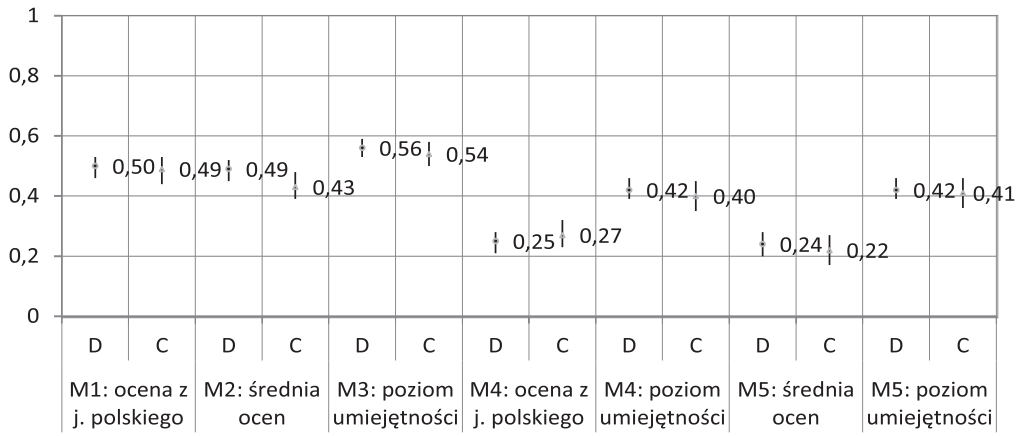
Wskaźniki oparte na ocenach szkolnych (pojedyncza ocena z matematyki lub średnia ocen szkolnych) dla kohorty uczniów zdających egzamin gimnazjalny w 2011 r. i maturę w 2014 r. są porównywalnie prognostyczne dla obu płci. Jedyną różnicą jest to, że średnia ocen jest nieco mniej prognostyczna niż ocena z matematyki w grupie chłopców – jednak różnica ta nie jest istotna statystycznie, (co obrazują pokrywające się przedziały ufności). Wśród predyktorów wyniku chłopców na maturze z matematyki największą moc prognostyczną ma wynik egzaminu gimnazjalnego z matematyki. U chłopców wzrost wyniku egzaminu gimnazjalnego o jedno odchylenie standardowe pozwala przewidywać wzrost wyniku maturalnego o 0,68 odchylenia standardowego. Podobne wnioski można zauważyć dla kohorty uczniów młodszych o rok (Rysunek 2). Co ciekawe, moc prognostyczna ocen szkolnych i wyników egzaminu gimnazjalnego w kolejnej kohorcie uczniów jest jeszcze wyraźniejsza. Podobnie jak w kohorcie uczniów zdających egzaminy w 2011–2014 r. wynik egzaminu gimnazjalnego z matematyki jest bardziej prognostyczny dla wyniku na maturze w grupie chłopców niż dziewcząt (różnica efektów na granicy istotności statystycznej – prawdopodobnie ze względu na niewielką liczebność próby).

Dla modeli z dwoma predyktorami (kohorta 2012–2015) wynik egzaminu gimnazjalnego jest znacząco lepszym predyktorem wyniku na maturze niż wskaźniki oceniania wewnętrznego. Efekt ten można zaobserwować zarówno w grupie dziewcząt, jak i w grupie chłopców (Rysunek 2). Moc prognostyczna oceny z matematyki i średniej ocen jest porównywalna w grupie dziewcząt, a w grupie chłopców nieco lepiej przewiduje ocena ($\beta = 0{,}34$) niż średnia ($\beta = 0{,}29$), choć różnica ta, prawdopodobnie ze względu na niewielką liczebność próby, nie osiąga istotności statystycznej. Natomiast w obu modelach z dwoma predyktorami (modele 4 i 5) pojedyncza ocena i średnia ocen są tylko w połowie tak prognostyczne, jak wyniki egzaminu gimnazjalnego.

Dla kohorty uczniów starszych o rok (2011–2014) tendencje te się rozmywają (Rysunek 1). Analiza modelu 5 ujawnia brak istotnie statystycznej różnicy w mocy prognostycznej średniej ocen i wyników



Rysunek 1. Wielkość standaryzowanych współczynników regresji (β) i odpowiadające im przedziały ufności jako wskaźniki efektu trafności prognostycznej dla wyników egzaminu maturalnego z matematyki (kohorta zdających egzamin gimnazjalny w 2011 r. i maturalny w 2014 r.).



Rysunek 2. Wielkość standaryzowanych współczynników regresji (β) i odpowiadające im przedziały ufności jako wskaźniki efektu trafności prognostycznej dla wyników egzaminu maturalnego z matematyki (kohorta zdających egzamin gimnazjalny w 2012 r. i maturalny w 2015 r.).

egzaminu gimnazjalnego. Wielkości efektów są porównywalne w grupie chłopców i dziewcząt – mieszczą się w przedziale od 0,37 do 0,43. Porównanie wyników egzaminu gimnazjalnego do oceny z matematyki wskazuje istotnie niższą moc prognostyczną tego pierwszego wskaźnika w grupie dziewcząt o ok. 10% odchylenia standardowego (w grupie chłopców różnica nie jest istotna statystycznie).

Język polski

Dla kohorty uczniów 2011–2014 zdających maturę z języka polskiego można zaobserwować wyższą moc prognostyczną wyników egzaminu gimnazjalnego w porównaniu

do dwóch wskaźników oceniania wewnętrznego (Rysunek 3). Moc prognostyczna egzaminu gimnazjalnego w tym zakresie jest porównywalna między dziewczętami a chłopcami – wynosi odpowiednio 0,56 i 0,54 odchylenia standardowego. Różnica jest widoczna zwłaszcza w grupie chłopców, gdzie średnia ocen w istotny statystycznie sposób gorzej przewiduje sukces na maturze w porównaniu do wyniku egzaminu gimnazjalnego (różnica ok. 10% odchylenia standardowego). Ocena z języka polskiego pozwala w grupie chłopców na lepsze przewidywanie niż średnia ocen i jako taka nie różni się jakością predykcji od wyników egzaminacyjnych. W przypadku młodszej kohorty różnice w trafności prognostycznej poszczególnych wskaźników są nieistotne statystycznie (Rysunek 4).

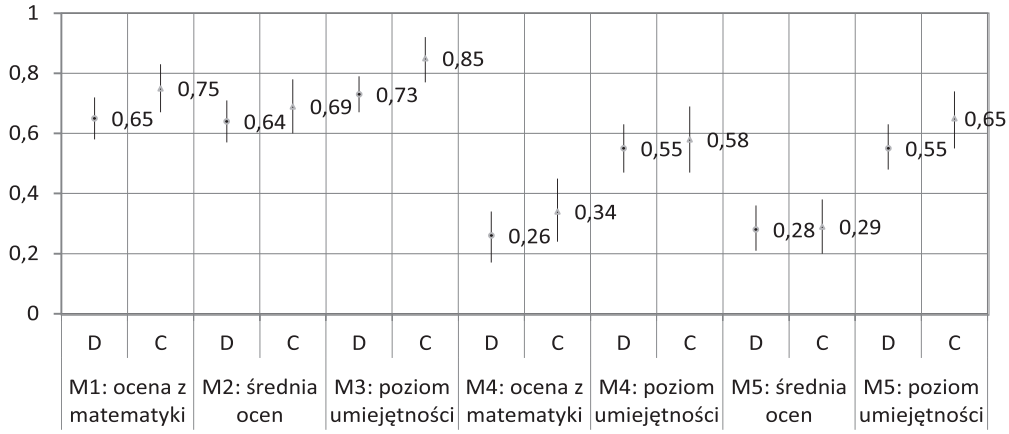
W modelach łączących dwa predyktory (Rysunek 3), zarówno w grupie chłopców, jak i dziewcząt, moc predykcyjna wyników egzaminu gimnazjalnego jest istotnie i znacząco wyższa (β od 0,4 do 0,42) niż oceny z języka polskiego (0,25 w grupie dziewcząt i 0,27 w grupie chłopców) oraz średniej ocen szkolnych (0,24 i 0,22 odpowiednio w grupie dziewcząt i chłopców). Różnice te zacierają się w przypadku kohorty 2012–2015, na co wskazuje brak istotności statystycznej w mocy prognostycznej poszczególnych wskaźników (Rysunek 4).

Wskaźniki osiągnięć z języka polskiego w obydwu analizowanych kohortach są zdecydowanie mniej prognostyczne niż w przypadku matematyki (Rysunki 3 i 4). Efekt ten jest szczególnie widoczny dla kohorty zdających egzaminy w 2012 i 2015 r. Różnica efektu przewidywania wyników egzaminu gimnazjalnego u chłopców wynosi niemal 1/3 odchylenia standardowego na korzyść matematyki w porównaniu do języka polskiego i efekt ten jest wysoce istotny statystycznie. Podobnie duża różnica zarysowuje się w grupie dziewcząt. Jeśli przeanalizujemy jakość wskaźników opartych na ocenianiu wewnątrzszkolnym, to okaże się, że oceny i średnia są istotnie słabszym predyktorem późniejszych osiągnięć na egzaminie z języka polskiego (różnica ok. 15% odchylenia standardowego w porównaniu do matematyki). Różnica jest widoczna w grupach chłopców, w obydwu porównywanych kohortach. Przykładowo w kohorcie 2012–2015 ocena z matematyki jest o niemal 1/3 odchylenia standardowego lepszym predyktorem sukcesu na maturze niż ocena z języka polskiego. U dziewcząt ta różnica jest widocznie mniejsza (osiąga ok. 20% różnicy odchylenia standardowego).

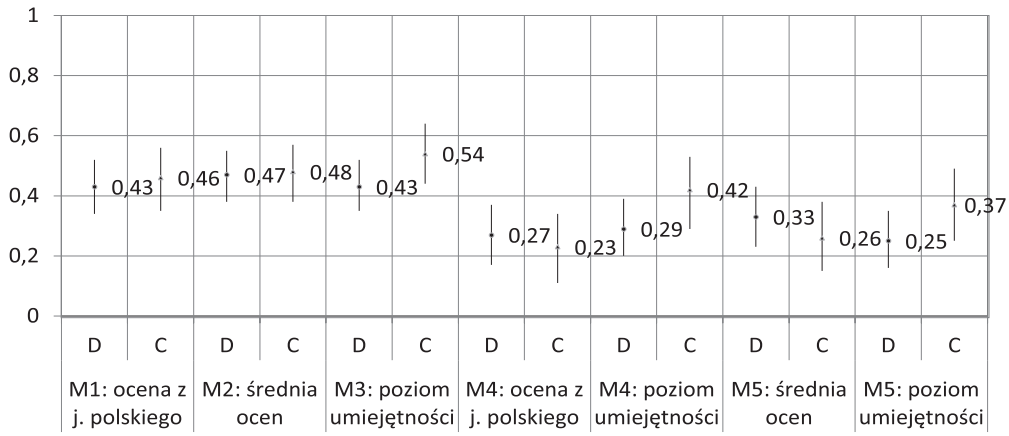
Analiza poziomu wyjaśnionej wariancji

Wyniki przedstawione w Tabeli 4 wskazują, że różnice w poziomie wyjaśnionej wariancji dla dziewcząt i chłopców są niewielkie (rzędu 0,01–0,04). W przypadku efektów stałych różnice układają się podobnie zarówno w przypadku matematyki, jak i języka polskiego. Więcej wariancji dla chłopców wyjaśniają modele: 1 (tylko ocena) i 4 (ocena i poziom umiejętności), a dla dziewcząt: 2 (średnia), 3 (poziom umiejętności) i 5 (średnia i poziom umiejętności). Dla wariancji wyjaśnionej przez efekty stałe i losowe pojawia się zróżnicowanie związane z typem egzaminu. Wszystkie modele przewidujące wynik maturalny z matematyki wyjaśniają więcej wariancji związanej z efektami stałymi i losowymi w przypadku chłopców, natomiast wszystkie modele przewidujące wynik z matury z języka polskiego wyjaśniają więcej warunkowego R^2 dla dziewcząt.

W 2015 r. sytuacja wygląda inaczej (Tabela 5). Podobnie, różnice pomiędzy chłopcami a dziewczętami są raczej niewielkie. Zarówno w przypadku języka polskiego, jak i matematyki, wszystkie z proponowanych modeli wyjaśniają więcej wariancji dla chłopców niż dla dziewcząt, zarówno poprzez efekty stałe, jak i mieszane. Najwięcej wariancji dla obu przedmiotów i obu płci wyjaśnia model 5,



Rysunek 3. Wielkość standaryzowanych współczynników regresji (β) i odpowiadające im przedziały ufności jako wskaźniki efektu trafności prognostycznej dla wyników egzaminu maturalnego z języka polskiego (kohorta zdających egzamin gimnazjalny w 2011 r. i maturalny w 2014 r.).



Rysunek 4. Wielkość standaryzowanych współczynników regresji (β) i odpowiadające im przedziały ufności jako wskaźniki efektu trafności prognostycznej dla wyników egzaminu maturalnego z języka polskiego (kohorta zdających egzamin gimnazjalny w 2012 r. i maturalny w 2015 r.).

w którym wykorzystuje się jako predyktory średnią i poziom umiejętności uczniów.

Dyskusja

Niniejsze wyniki sugerują, że nie ma znaczących różnic we wskaźnikach trafności prognostycznej dla obu płci. Informacja o wskaźnikach oceniania wewnętrznego oraz zewnętrznego pozwala na wyjaśnienie podobnego zakresu zmienności wskaźników maturalnych dziewcząt i chłopców, zdających oba analizowane przedmioty. Wyniki wskazują więc, że projektowanie interwencji edukacyjnych na podstawie informacji o wcześniejszych wynikach

Tabela 4
Poziom wyjaśnionej wariancji dla matury z 2014 r. dla dziewcząt i chłopców

		Procent wyjaśnionej wariancji			
		Efekty stałe		Efekty losowe i efekty stałe	
Matura 2014	Model	Dziewczęta	Chłopcy	Dziewczęta	Chłopcy
Matematyka	0	0,00	0,00	0,19	0,26
	1	0,43	0,45	0,58	0,61
	2	0,41	0,39	0,55	0,56
	3	0,42	0,41	0,46	0,49
	4	0,54	0,55	0,61	0,65
	5	0,53	0,52	0,59	0,62
Język polski	0	0,00	0,00	0,16	0,14
	1	0,22	0,23	0,38	0,35
	2	0,22	0,21	0,34	0,31
	3	0,31	0,30	0,38	0,34
	4	0,34	0,35	0,42	0,40
	5	0,34	0,34	0,41	0,38

uczniów, nie powinno być obciążone w znaczący sposób płcią.

Możemy stwierdzić na podstawie otrzymanych wyników zgodność obu typów oceniania, szczególnie w przypadku języka polskiego. Oceny szkolne z matematyki dziewcząt i chłopców nie różnią się w sposób istotny statystycznie. Jednocześnie chłopcy osiągają wyższe wyniki egzaminacyjne i jest to różnica istotna statystycznie oraz stabilna w czasie – ujawnia się zarówno na egzaminie gimnazjalnym, jak i trzy lata później na maturze z matematyki. Znając oceny z matematyki w grupie chłopców, można byłoby spodziewać się także niższych wyników egzaminacyjnych z matematyki. Wyniki jednak temu przeczą. Mamy więc do czynienia z niedoszacowaniem wyników egzaminacyjnych chłopców w obrębie matematyki w stosunku do ocen z matematyki. Biorąc pod uwagę niższe wyniki egzaminacyjne dziewcząt z matematyki, możemy też mówić o przeszacowaniu ich ocen w stosunku do umiejętności mierzonych testowo. Różnica w niedoszacowaniu lub przeszacowaniu wyników z matematyki nie jest jednak duża. W zakresie języka polskiego występuje spójność oceniania i wyników egzaminacyjnych – dziewczęta uzyskują zarówno wyższe oceny, jak i wyniki na egzaminie gimnazjalnym i maturalnym.

Może to sugerować, że mimo krytyki, której poddawane są oceny szkolne (ze względu na ich niską rzetelność oraz potencjalne obciążenie czynnikami niezwiązanymi z mierzoną cechą), mają większą niż przypuszczamy wartość informacyjną pozwalającą na wyjaśnienie zróżnicowania wyników maturalnych oraz są zgodne z wynikami egzaminów gimnazjalnych. Istnienie znacznych związków (współczynniki korelacji rzędu 0,4–0,8) pomiędzy stopniami szkolnymi zaobserwowano również w badaniach zagranicznych (np. Brennan, Kim, Wenz-Gross i Siperstein, 2001; Ross i Kostuch, 2011).

Zróżnicowanie ze względu na płeć związku pomiędzy wynikami oceniania wewnętrznego oraz egzaminami zewnętrznymi może być spowodowane relacją pomiędzy systematycznością dziewcząt i chłopców oraz jej związkiem z odrabianiem prac

Tabela 5
Poziom wyjaśnionej wariancji dla matury z 2015 r. dla dziewcząt i chłopców

		Procent wyjaśnionej wariancji			
		Efekty stałe		Efekty losowe i efekty stałe	
Matura 2015	Model	Dziewczęta	Chłopcy	Dziewczęta	Chłopcy
Matematyka	0	0,00	0,00	0,09	0,15
	1	0,44	0,53	0,51	0,62
	2	0,42	0,44	0,48	0,54
	3	0,58	0,62	0,62	0,65
	4	0,61	0,66	0,64	0,70
	5	0,62	0,66	0,65	0,70
Język polski	0	0,00	0,00	0,14	0,19
	1	0,18	0,19	0,32	0,39
	2	0,22	0,23	0,34	0,37
	3	0,20	0,27	0,28	0,38
	4	0,24	0,30	0,35	0,41
	5	0,27	0,32	0,37	0,42

domowych. Mniejsza pilność chłopców (Grygiel, 2016) może przekładać się na mniej systematyczne odrabianie prac domowych, co może być jednym z kryteriów branych pod uwagę podczas wystawiania ocen, nie musi natomiast przekładać się na wyniki egzaminów zewnętrznych. Hipoteza o związkach systematyczności z wynikami egzaminów i ocenami szkolnymi (oraz sile tych związków) wymaga jednak przetestowania.

Analizując trafność prognostyczną wskaźników osiągnięć szkolnych – łączne wykorzystanie dwóch predyktorów (ocenianie wewnątrzszkolne i wyniki egzaminacyjne) wyjaśnia najwięcej wariancji maturalnych wskaźników osiągnięć, szczególnie w grupie chłopców. Znajomość tych dwóch zmiennych niezależnych w znacznym stopniu (55–66%) pozwala przewidywać sukces ucznia na maturze z matematyki. Co ciekawe, w przypadku modeli, w których predyktorami były wskaźniki oceniania wewnątrzszkolnego (ocena lub średnia ocen), silne są efekty losowe. Oznacza to, że uwzględnienie poziomu klasy znacznie podwyższa poziom wyjaśnionej wariancji, a cechy nauczyciela oraz ogólny poziom danej klasy szkolnej może mieć wpływ na wyniki oceniania wewnętrznego.

Warto zauważyć, że w przypadku języka polskiego osoba nauczyciela ma nie tylko większe znaczenie dla chłopców podczas wystawiania ocen szkolnych, lecz także dla ich wyników na egzaminie gimnazjalnym. Może to świadczyć o tym, że chłopcy nie tylko otrzymują niższe oceny niż dziewczęta, lecz także mają niższe umiejętności, a być może sam sposób uczenia ich w klasie jest inny niż ten sposób uczenia dziewcząt. Możemy mieć tu do czynienia z tym, co badacze nazwali „zagrożeniem stereotypem" (Spencer, Steele i Quinn, 1999; Steele i Aronson, 1995). Zarówno nauczyciele, jak i rodzice podzielają stereotypy płciowe na temat dziewcząt i chłopców, zakładając wyższe, niejako „wrodzone" zdolności dziewcząt w przedmiotach humanistycznych, a chłopców w ścisłych. Przekłada się to wprost na wystawiane oceny i pośrednio (przez inny sposób nauczania) na wyniki egzaminów przedmiotowych. Hipoteza ta wymaga jednak osobnego sprawdzenia w kolejnych

badaniach empirycznych. Niemniej niższe umiejętności czytania ze zrozumieniem oraz pisania w grupie chłopców wymagają monitorowania.

Możliwość przewidywania wyników jest zróżnicowana ze względu na dziedzinę nauczania – w przypadku matematyki jest ok. dwukrotnie wyższa niż w przypadku języka polskiego. Ocenianie umiejętności z języka polskiego – zarówno mierzone ocenami szkolnymi, jak i egzaminami – może być w większym stopniu podatne na wariancję niezwiązaną z mierzoną cechą (CIV) niż w przypadku umiejętności matematycznych. Na egzaminie maturalnym może na nią wpływać efekt egzaminatora – znaczną część punktów z egzaminu determinuje wynik eseju na jeden z dwóch tematów. W idealnych warunkach oceniający powinni stosować te same kryteria w ten sam, spójny sposób podczas sesji oceniania. Jednakże w przypadku języka polskiego utrzymanie tego założenia jest bardzo trudne. W badaniach nad efektem oceniającego wykazano, że różnica pomiędzy 25% najbardziej łagodnych oraz 25% najbardziej surowych egzaminatorów oceniających maturę z języka polskiego, w zależności od analizowanego roku oraz tematu eseju, może wahać się w okolicach 3,1–3,7 pkt proc. całkowitego wyniku egzaminu[6] (Szaleniec i in., 2015). Dodatkowo egzaminy maturalne z języka polskiego (w przeciwieństwie do matematyki[7]) najczęściej składają się z pytań otwartych (67/70 pkt za zadania otwarte w 2014 r., 68/70 pkt w 2015 r.) oraz są punktowane na długich (np. 0–12, 0–18), nie zawsze precyzyjnych skalach, więc w większym stopniu mogą być podatne na oddziaływanie wariancji niezwiązanej z mierzoną cechą. Co więcej, rzetelność egzaminów z języka polskiego (szczególnie maturalnego z 2015 r.) jest niższa niż rzetelność egzaminów z matematyki, co również może wpływać na jakość predykcji.

Ocena z matematyki ma interesujący związek z wynikami maturalnymi, szczególnie w grupie chłopców. Choć dla modeli z pojedynczymi predyktorami najsilniejszy związek z wynikami egzaminu maturalnego ma poziom umiejętności oszacowany na podstawie egzaminu gimnazjalnego, to nie ma różnicy w mocy prognostycznej egzaminu gimnazjalnego i pojedynczej oceny z matematyki. Można więc zastanowić się, dlaczego pojedyncza ocena z matematyki (szczególnie w grupie chłopców) jest bardziej prognostyczna dla późniejszych sukcesów matematycznych ucznia niż średnia ocen, oraz dlaczego efekt ten nie występuje w przypadku języka polskiego.

Odpowiedzi na pytanie należy również poszukiwać w analizie źródeł wariancji niezwiązanej z mierzona cechą. W przypadku ocen można śmiało mówić o tym, że poza umiejętnościami poznawczymi ucznia mierzą, także czynniki pozapoznawcze. Ocena z matematyki może jednak być mniej podatna na zniekształcenia niż innego typu oceny. Wynika to z faktu, że podstawą oceniania uczniów w klasie są zadania, które mają bardziej przejrzyste kryteria oceny – są to zadania ze skończoną, niewielką liczbą rozwiązań (niezależnie od tego, czy są to sprawdziany czy odpowiedzi ustne). Prawdopodobnie oceny z innych przedmiotów, zwłaszcza humanistycznych, pozostawiają nauczycielom większe pole do manewru w ocenianiu – kryteria oceniania są bardziej subiektywne, więc w większym stopniu są obciążone CIV.

Wszystkie powyższe wnioski wskazują na to, że używanie łącznie wyników egzaminów i ocen szkolnych w znacznym stopniu pozwala przewidywać późniejszy sukces ucznia. Daje to podstawę projektowania skutecznych, wczesnych oddziaływań nauczycielskich na poziomie kształcenia gimnazjalnego. Jednocześnie zarówno słaba moc prognostyczna oceniania szkolnego (poza

[6] Ta różnica podczas oceniania zadań otwartych z matematyki wynosiła od 0,87 do 1,36 pkt proc.

[7] 50% pkt za zadania otwarte, maksymalna długość skali.

oceną z matematyki), jak i problemy z trafnością tych wskaźników (podatność na czynniki pozadydaktyczne) sugerowałaby niską użyteczność ocen jako wskaźników.

Ograniczeniem analiz przedstawionych w tym artykule jest przyjęcie założenia, że we wszystkich przypadkach przyrost umiejętności uczniów jest taki sam, tj. cechują się podobną edukacyjną wartością dodaną (EWD; więcej o wskaźniku można przeczytać w: Dolata i in., 2014; Żółtak, 2015). Założenie o tym, że wszystkie szkoły ponadgimnazjalne cechują się jednakowym wkładem w przyrost umiejętności uczniów jest pewnym uproszczeniem. Dlatego przyszłe analizy powinny kontrolować maturalny wskaźnik EWD (odpowiednio z języka polskiego i matematyki) poprzez włączenie go do modeli. W przyszłych analizach warto skupić się na dokładnej analizie różnic w trafności prognostycznej dla różnych wartości wskaźników osiągnięć gimnazjalnych wśród chłopców i dziewcząt. W języku polskim istnieją znaczne różnice we wskaźnikach osiągnięć wśród dziewcząt i chłopców w zależności od ich poziomu osiągnięć (Skórska, 2015). W przypadku egzaminu z języka polskiego wyniki chłopców są znacznie bardziej zmienne niż wyniki dziewcząt (o ok. 18–23% w zależności od analizowanego roku). W grupie uczniów osiągających wyniki powyżej 90. centyla stosunek liczby dziewcząt do chłopców wynosi od 1,91 do 2,5 (w zależności od analizowanego roku). W grupie uczniów osiągających najniższe wyniki nadreprezentacja chłopców jest ewidentna – ok. 4,5–5 razy więcej chłopców niż dziewcząt znajduje się w 1% najniższych wyników z języka polskiego (Skórska, 2015). W przypadku matematyki (Zawistowska, 2013) zaobserwowano względnie stały udział dziewcząt niezależnie od osiągniętych wyników na poziomie egzaminu gimnazjalnego (rok szkolny 2008/2009) oraz spadek odsetka kobiet w grupie osób zdających egzamin maturalny (w 2011 r.). Te wyniki wskazują, że nierówności mogą pojawić się dopiero na etapie szkoły ponadgimnazjalnej. Przytoczone badania sugerują, że szacowanie trafności prognostycznej ze względu na płeć oraz różny poziom osiągnięć mogłyby wnieść wartościowy wkład do przedstawionych analiz.

Literatura

American Educational Research Association, American Psychological Association i National Council for Measurement in Education (2014). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

Atkinson, R. C. i Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher*, *38*(9), 665–676.

Bartoń, K. (2015). Package 'MuMIn'. Pobrano z ftp://155.232.191.229/cran/web/packages/MuMIn/MuMIn.pdf

Ben-Shakhar, G. i Sinai, Y. (1991). Gender differences in multiple-choice tests: the role of differential guessing tendencies. *Journal of Educational Measurement*, *28*(1), 23–35.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. W: F. M. Lord i M. R. Novick (red.), *Statistical theories of mental test scores* (s. 17–20). Reading: Addison-Wesley.

Brennan, R., Kim, J., Wenz-Gross, M. i Siperstein, G. (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), 173–217.

Bryk, A. S. i Raudenbush, S. W. (1992). *Hierarchical linear models: applications and data analysis*. Newbury Park: Sage.

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.

Carmines, E. G. i Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks: Sage.

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.

Cole, N. S. (1997). *The ETS gender study: how males and females perform in educational settings*. Princeton: Educational Testing Service.

Costa Jr, P., Terracciano, A. i McCrae, R. R. (2001). Gender differences in personality traits across cultures: robust and surprising findings. *Journal of personality and social psychology*, *81*(2), 322–331.

Crocker, L. i Algina, J. (1986). *Introduction to classical and modern test theory*. Mason: Cengage Learning.

Cronbach, L. J. i Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: the role of response format. *Applied Measurement in Education*, *11*(3), 279–299.

Dolata, R., Hawrot, A., Humenny, G., Jasińska-Maciążek, A., Koniewski, M. i Majkut, P. (2014). *Kontekstowy model oceny efektywności nauczania po pierwszym etapie edukacyjnym*. Warszawa: Instytut Badań Edukacyjnych.

Donnellan, M. B. i Lucas, R. E. (2008). Age differences in the Big Five across the life span: evidence from two national samples. *Psychology and Aging*, *23*(3), 558–566.

Drost-Rudnicka, M. (2012). Edukacja wczesnoszkolna a problem nierówności płci – uczniowskie stereotypy postrzegania płci. W: N. Majchrzak, N. Starik i A. Zduniak (red.), *Podmiotowość w edukacji wobec odmienności kulturowych oraz społecznych* zróżnicowań (447–456). Poznań: Wydawnictwo Wyższej Szkoły Bezpieczeństwa.

Fried-Buchalter, S. (1997). Fear of success, fear of failure, and the imposter phenomenon among male and female marketing managers. *Sex Roles*, *37*(11–12), 847–859.

Geiser, S. i Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year: high-school record vs standardized tests as indicators of four-year college outcomes. *Research & Occasional Paper Series*: CSHE. 6.07. Berkeley: Center for Studies in Higher Education, University of California.

Goldman, R. D. i Hewitt, B. N. (1975). Adaptation-level as an explanation for differential standards in college grading. *Journal of Educational Measurement*, *12*(3), 149–161.

Gromkowska-Melosik, A. (2011). *Edukacja i (nie) równość społeczna kobiet: studium dynamiki dostępu*. Oficyna Wydawnicza Impuls.

Grygiel, P. (2016). *Chłopięca „krzywda" – społeczny kontekst ocen szkolnych*. Referat wygłoszony podczas XVI Ogólnopolskiego Zjazdu Socjologicznego, Gdańsk.

Guiso, L., Monte, F., Sapienza, P. i Zingales, L. (2008). Culture, gender, and math. *Science-New York Then Washington*, *320*(5880), 1164–1165.

Haladyna, T. M. i Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27.

Hezlett, S., Kuncel, N., Vey, A., Ones, D., Campbell, J. i Camara, W. (2001*). The effectiveness of the SAT in predictive success early and late in college: a comprehensive meta-analysis.* Referat wygłoszony podczas The Annual Meeting of The National Council of Measurement in Education, Seattle.

Hox, J. J. (2010). *Multilevel analysis. Techniques and application*. New York: Routledge.

Hyde, J. S. i Kling, K. C. (2001). Women, motivation, and achievement. *Psychology of Women Quarterly*, *25*(4), 364–378.

Johnson, P. C. (2014). Extension of Nakagawa and Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution*, *9*(5), 944–946.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D. i Barbuti, S. M. (2008). Validity of the SAT® for predicting first-year college grade point average. *Research Report*, 2008-5. New York: College Board.

Konarzewski, K. (1996). *Problemy i schematy. Pierwszy rok nauki szkolnej dziecka.* Warszawa: Żak.

Kopciewicz L. (2008). Grzeczne dziewczynki, niegrzeczni chłopcy – wytwarzanie różnic rodzajowych w dydaktyczno-wychowawczej pracy szkoły. W: M. Dudzikowa i M. Czerepaniak-Walczak (red.), *Wychowanie. Pojęcia. Procesy. Konteksty. Interdyscyplinarne ujęcie* (t. 4, s. 349–392). Gdańsk: Gdańskie Wydawnictwo Psychologiczne.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*(2), 157–162.

Mac an Ghaill, M. (1994). *The making of men: masculinities, sexualities and schooling.* Buckingham: Open University Press.

Mau, W. C. i Lynn, R. (2001). Gender differences on the Scholastic Aptitude Test, the American College Test and college grades. *Educational Psychology*, *21*(2), 133–136.

Messick, S. (1989). Validity. W: R. L. Linn (red.), *Educational Measurement* (wyd. 3, s. 13–103). New York: American Council on Education.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177–195.

Mitchell, K. J. (1990). Traditional predictors of performance in medical school. *Academic Medicine*, *65*(3), 149–58.

Nakagawa, S. i Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

Niemierko, B. (1999). *Pomiar wyników kształcenia*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.

Niemierko, B. (2001). Chłodne oblicze egzaminu zewnętrznego. *Edukacja*, *75*(3), 11–22.

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.

Pankowska, D. (2005). *Wychowanie a role płciowe: program edukacyjny*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.

Pedhazur, E. J. i Schmelkin, L. P. (1991). *Measurement design and analysis: an integrated approach*. Hillsdale: Lawrence Erlbaum.

Pekkarinen, T. (2014). Gender differences in strategic behaviour under competitive pressure: evidence on omission patterns in university entrance examinations. *IZA Discussion Paper*, 8018. Pobrano z http://ftp.iza.org/dp8018.pdf

Perkins, R., Kleiner, B., Roey, S. i Brown, J. (2004). *The high school transcript study: a decade of change in curricula and achievement, 1990–2000*. Washington: National Center for Education Statistics.

Pomerantz, E. M., Altermatt, E. R. i Saxon, J. L. (2002). Making the grade but feeling distressed: gender differences in academic performance and internal distress. *Journal of Educational Psychology*, *94*(2), 396–404.

Preckel, F., Goetz, T., Pekrun, R. i Kleine, M. (2008). Gender differences in gifted and average-ability students comparing girls' and boys' achievement, self-concept, interest, and motivation in mathematics. *Gifted Child Quarterly*, *52*(2), 146–159.

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.

Rothstein, J. M. (2004). College performance predictions and the SAT. *Journal of Econometrics*, *121*(1), 297–317.

Samejima, F. (1968). Estimation of latent ability using a pattern of graded scores. *ETS Research Bulletin Series*. Princeton: Educational Testing Service.

Schmitt, D. P., Realo, A., Voracek, M. i Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, *94*(1), 168–182.

Schrader, S. von i Ansley, T. (2006). Sex differences in the tendency to omit items on multiple-choice tests: 1980–2000. *Applied Measurement in Education*, *19*(1), 41–65.

Schuler, H., Funke, U. i Baron-Boldt, J. (1990). Predictive validity of school grades – a meta-analysis. *Applied Psychology*, *39*(1), 89-103.

Severiens, S. i Dam, G. ten (1998). A multilevel meta-analysis of gender differences in learning orientations. *British Journal of Educational Psychology*, *68*(4), 595–608.

Skórska, P. (2015). Gender gap in reading and writing achievements. *Kwartalnik Pedagogiczny*, *238*(4), 139–152.

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.

Skórska, P., Świst, K. i Szaleniec, H. (2014). Szacowanie trafności predykcyjnej ocen szkolnych z wykorzystaniem hierarchicznego modelowania liniowego. *Edukacja*, *128*(3), 75–94.

Slakter, M. J., Koehler, R. A., Hampton, S. H. i Grennell, R. L. (1971). Sex, grade level, and risk taking on objective examinations. *The Journal of Experimental Education*, *39*(3), 65–68.

Snijders, T. A. B. i Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling.* London: Sage.

Spencer, S. J., Steele, C. M. i Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28.

Srivastava, S., John, O. P., Gosling, S. D. i Potter, J. (2003). Development of personality in early and middle adulthood: set like plaster or persistent change? *Journal of Personality and Social Psychology*, *84*(5), 1041–1053.

Steele, C. M. i Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811.

Steinmayr, R. i Spinath, B. (2008). Sex differences in school achievement: what are the roles of personal-

ity and achievement motivation? *European Journal of Personality*, *22*(3), 185–209.

Suchocka, A. (2011). Przemoc symboliczna jako element ukrytego programu kształcenia polskiej szkoły. *Zeszyty Naukowe Akademii Marynarki Wojennej*, 52, 293–302.

Szaleniec, H., Grudniewska, M., Kondratek, B., Kulon, F., Pokropek, A., Stożek, E. i Żółtak, M. (2013). *Analiza porównawcza wyników egzaminów zewnętrznych – sprawdzian w szóstej klasie szkoły podstawowej i egzamin gimnazjalny*. Warszawa: Instytut Badań Edukacyjnych.

Szaleniec, H., Kondratek, B., Kulon, F., Pokropek, A., Skórska, P., Świst, K., Wołodźko, T. i Żółtak, M. (2015). *Efekt egzaminatora w ocenianiu prac maturalnych z języka polskiego i matematyki*. Warszawa: Instytut Badań Edukacyjnych.

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.

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

Willingham, W. W. i Cole, N. S. (1997). *Gender and fair assessment*. Mahwah: Lawrence Erlbaum.

Willingham, W. W., Pollack, J. M. i Lewis, C. (2002). Grades and test scores: accounting for observed differences. *Journal of Educational Measurement*, *39*(1), 1–37.

Zahner, D., Ramsaran, L. M. i Steedle, J. T. (2012). *Comparing alternatives in the prediction of college success*. Referat wygłoszony podczas The Annual Meeting of the American Educational Research Association, Vancouver.

Zawistowska, A. (2013). „Płeć matematyki". Zróżnicowania osiągnięć ze względu na płeć wśród uzdolnionych uczniów. *Studia Socjologiczne*, *210*(3), 75–95.

Zill, N. i West, J. (2001). *Entering kindergarten: a portrait of American children when they begin school*. Washington: National Center for Education Statistics.

Żółtak, T. (2015). *Statystyczne modelowanie wskaźników edukacyjnej wartości dodanej –podsumowanie polskich doświadczeń 2005–2015*. Warszawa: Instytut Badań Edukacyjnych.

Artykuł powstał na podstawie konkursu wewnętrznego o finansowanie badań naukowych lub prac rozwojowych oraz zadań z nimi związanych, służących rozwojowi młodych naukowców w Instytucie Badań Edukacyjnych.

Tekst złożony 30 września 2016 r., zrecenzowany 7 listopada 2016 r., przyjęty do druku 7 grudnia 2016 r.

Gender differences in the predictive validity of student achievement indicators from lower secondary school

In this article, we discuss the gender differences in the predictive validity of students' Matura results. We use the achievement indicators (school grades, grade point average and standardized exam results) from lower secondary school as predictors of success on the Matura exams. Due to a number of psychological, socio-cultural and other factors, achievement indicators may function differentially according to gender. Thus, we hypothesize that the predictive validity itself may differ for girls and boys. We analyzed two cohorts of students – the first one took the exam at the end of lower secondary school in 2011 and the Matura in 2014, the second one took its exams in 2012 and 2015. We conducted the analysis using hierarchical linear modeling and ability level estimated within the IRT paradigm. The results show the differential functioning of achievement indicators according to gender and domain tested (mathematics and literacy). Combining exam results and school grades is the best strategy to predict the Matura's results; however, the differences in the predictive validity between girls and boys is negligible.

Keywords: educational research, predictive validity, gender, hierarchical linear modeling, grading.