Edukacja 2014, 3(128), 75–94
ISSN 0239-6858

Szacowanie trafności predykcyjnej ocen szkolnych z wykorzystaniem hierarchicznego modelowania liniowego

Paulina Skórska, Karolina Świst, Henryk Szaleniec
Instytut Badań Edukacyjnych*

Artykuł omawia zagadnienie trafności predykcyjnej ocen szkolnych dla wyników uzyskanych w egzaminach zewnętrznych. Opierając się na ocenach szkolnych i ich średniej, oszacowano stopień, w jakim można było przewidzieć uzyskany przez ucznia wynik na egzaminie gimnazjalnym w latach 2012 i 2013. Ze względu na hierarchiczną strukturę danych, w analizach wykorzystano dwupoziomowy model regresji. Do oszacowania stopnia selekcyjności oddziałów szkolnych w gimnazjum użyto modelowania IRT. Wyniki świadczą o wysokiej mocy predykcyjnej ocen szkolnych i ich średniej, zarówno na poziomie indywidualnym, jak i na poziomie oddziału szkolnego. Współczynniki korelacji wewnątrzklasowej świadczą o wysokim stopniu selekcyjności oddziałów szkolnych. Przedstawione wyniki mają znaczenie w kontekście procesów rekrutacyjnych do szkół ponadgimnazjalnych, zwłaszcza wyboru przez ucznia konkretnego oddziału szkolnego.

Słowa kluczowe: trafność predykcyjna, model hierarchiczny, ocenianie szkolne, selekcyjność oddziałów szkolnych, egzamin zewnętrzny.

Źródła różnic między ocenami szkolnymi a wynikami egzaminów zewnętrznych

Testowanie i ocenianie szkolne już od kilkudziesięciu lat wzbudzają poważne kontrowersje ze względu na wątpliwości dotyczące trafności (i rzetelności) testów, ocen szkolnych oraz związanej z tym idei sprawiedliwości społecznej i równości szans (Cronbach, 1975; Cureton, 1971). Zarówno ocenianie szkolne, jak i egzaminy zewnętrzne odnoszą się do podstawy programowej (mają sprawdzać osiągnięcie przez ucznia odpowiedniego poziomu wiedzy i umiejętności)[1], jednak korelacja ocen i wyników egzaminów jest umiarkowana (Willingham, Pollack i Lewis, 2002). Oceny szkolne i wyniki standaryzowanych testów[2] nie są ekwiwalentne, gdyż ich wariancja zależy od różnych czynników. Ponadto pełnią różne funkcje – „ocenianie wewnątrzszkolne wspomaga uczenie się, natomiast ocenianie na egzaminach dokumentuje poziom osiągnięć na progach

Artykuł powstał w ramach projektu systemowego „Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego" finansowanego ze środków Europejskiego Funduszu Społecznego w ramach Programu Operacyjnego Kapitał Ludzki (Priorytet III: Wysoka jakość systemu oświaty, Poddziałanie 3.1.1. Tworzenie warunków i narzędzi do monitorowania, ewaluacji i badań systemu oświaty).

[1] Ustawa z dnia 7 września 1991 r. o systemie oświaty. Dz.U. 1991 nr 95 poz. 425 z późn. zm.

[2] W artykule określenia „standaryzowane testy" i „egzaminy zewnętrzne" używamy zamiennie.

* Adres do korespondencji: ul. Górczewska 8, 01-180 Warszawa. E-mail: p.skorska@ibe.edu.pl

edukacyjnych i to tylko w zakresie, który poddaje się zewnętrznemu ocenianiu technikami stosowanymi na egzaminie (egzamin pisemny, egzamin ustny)" (Szaleniec, 2010, s. 99). Roman Dolata (2008) wskazuje także, że wskaźniki osiągnięć uczniów są bardzo często wykorzystywane w innych analizach, np. dotyczących nierówności edukacyjnych. W zależności od tego, który wskaźnik wybierzemy: wynik standaryzowanych testów osiągnięć czy też ocenę (oceny) szkolną, możemy dojść do odmiennych wniosków. W artykule przedstawiamy czynniki wpływające na zróżnicowanie ocen szkolnych i wyników egzaminów zewnętrznych.

Oceny szkolne i ich średnia nie są uważane za rzetelny wskaźnik osiągnięć ucznia ze względu na brak jednolitych standardów oceniania we wszystkich szkołach (wynikający z różnic w wewnątrzszkolnych regulaminach oceniania), a nawet w obrębie tych samych przedmiotów nauczanych w jednej szkole (Camara i Michaelides, 2005, Zwick i Himelfarb, 2011). Badania pokazują, że oceny z różnych przedmiotów nie są porównywalne, a korelacja między ocenami szkolnymi a wynikami testów jest różna dla różnych przedmiotów (Ramist, Lewis i McCamley, 1990; Young, 1990). Czynnikiem mającym wpływ na to zjawisko jest m.in. inflacja ocen szkolnych (*grade inflation*), czyli systematyczne polepszanie się ocen szkolnych bez odpowiedniego przyrostu umiejętności (Bejar i Blew, 1981). Robert Ziomek i Joseph Svec (1995) wskazują, że na przestrzeni lat 1989–1994 średnia ocen szkolnych w Stanach Zjednoczonych systematycznie wzrastała, m.in. z uwagi na to, że nauczyciele stawiają dobre oceny w nagrodę dla najlepszych uczniów. W tym samym czasie średni wynik testu SAT-V (*Scholastic Assessment Test* – część „Słownictwo") pozostał na tym samym poziomie, choć średnie wyniki SAT-M (*Scholastic Assessment Test* – część „Matematyka") wzrosły. Zespół Wayne'a Camary (Camara, Kimmel, Scheuneman i Sawtell, 2003) zauważył, że średnia ocen uczniów w ciągu 30 lat wzrosła z 2,97 (w 1973 r.) do 3,29 (w 2002 r.). Przyspieszenie wzrostu średniej nastąpiło zwłaszcza w latach 1998–2002, kiedy zaobserwowano połowę całego przyrostu z okresu 1973–2002. Susan Brookhart (1998) wskazuje, że zjawisko to może być spowodowane trzema przyczynami: presją ze strony rodziców i uczniów, niechęcią nauczycieli do wystawiania niskich ocen oraz konfliktem, jaki mogą odczuwać, odgrywając z jednej strony rolę sędziego oceniającego osiągnięcia uczniów, a z drugiej, adwokata zaangażowania i wysiłku włożonego przez uczniów w proces uczenia się (zob. Konarzewski, 1991). Poza tym, jak wskazuje Grażyna Szyling (2011), sam nauczyciel może osiągnąć jedynie niską lub – rzadziej – umiarkowaną rzetelność oceniania. Z tych powodów przewaga funkcji motywacyjnej oceny szkolnej nad jej funkcją informacyjną prowadzi do ograniczenia jej znaczenia treściowego (Niemierko, 2001).

Innym związanym z tym zjawiskiem jest stosowanie przez nauczycieli kryteriów pozadydaktycznych (Liszka, 2001; Szyling, 2011). Bolesław Niemierko (2001) zwrócił uwagę na to, że wyniki oceniania szkolnego (w ramach tzw. egzaminów powszednich, wewnętrznych) są wyższe niż wyniki egzaminów zewnętrznych, ponieważ na oceny szkolne wpływają nie tylko kryteria dydaktyczne, ale także społeczno-wychowawcze. Zarówno polskie (Liszka, 2001), jak i zagraniczne badania (Astin, 1971) dowodzą, że czynniki takie jak: spóźnianie się, opieszałość w wykonywaniu poleceń czy niesystematyczność, są związane z niższymi ocenami. Sami nauczyciele przyznają, że podczas wystawiania ocen biorą pod uwagę różne aspekty zachowania ucznia, nie tylko jego umiejętności (Frary, Cross i Weber, 1993; Szyling, 2011). Krzysztof Konarzewski (1991) wskazuje, że stopnie pełnią często

funkcję represyjną – uczniowie podkreślają, że ocena negatywna jest często stosowana przez nauczycieli jako kara za złe zachowanie. Duże znaczenie ma również zachowanie uczniów oraz systematyczność i terminowość wykonywania przez nich prac domowych (Ekstrom, Goertz i Rock, 1988). Kryteria pozadydaktyczne pośrednio mają także wpływ na motywację do uczenia się oraz samoocenę, która koreluje na poziomie 0,34 z ocenami szkolnymi i 0,22 z wynikami w standaryzowanych testach (Hansford i Hattie, 1982).

Ponadto ocenianie szkolne bardziej polega na ocenie holistycznej, wynikającej z ogólnych doświadczeń pracy z uczniem, w której ważną rolę odgrywają emocje nauczyciela oraz przekonania o poziomie opanowanych przez ucznia umiejętności (Wojciszke, 2001). Zauważalne jest zjawisko zawyżania ocen szkolnych uczniom przejawiającym trudności w nauce, gdyż większe znaczenie ma dla nauczyciela nakład pracy (staranie się), jaką uczeń wkłada w opanowanie materiału. Silnym komponentem oceniania staje się więc kryterium perspektywiczno-rozwojowe, nieobjęte programem nauczania (Black i William, 1998; Brookhart, 1993). O tym, jak trudne jest obiektywne ocenianie, przekonuje Konarzewski (1991), zauważając, że nauczyciele często stosują kategorialne schematy oceny uczniów, pozwalające klasyfikować ich jako członków danej grupy (wyróżnionej np. ze względu na płeć lub status społeczno-ekonomiczny), z pominięciem ich wewnętrznego zróżnicowania.

Inflacja stopni szkolnych oraz uwzględnianie w ocenianiu uczniów kryteriów pozaprogramowych, były argumentami przemawiającymi za wprowadzeniem systemu powszechnych egzaminów zewnętrznych w Polsce. Zakładano, że wprowadzenie systemu wzorowanego na brytyjskim, umożliwi Ministerstwu Edukacji Narodowej ujednolicenie kryteriów oceniania (Marquand, 1993) oraz sprawowanie kontroli parametrycznej nad polskimi szkołami (Niemierko, 2001).

Standaryzowane testy, takie jak SAT, również nie pozwalają w pełni trafnie szacować osiągnięć uczniów na późniejszych etapach kształcenia. Ich wyniki są pozytywnie skorelowane z miarami statusu społeczno-ekonomicznego ucznia (*socioeconomic status*, SES), takimi jak: dochód rodziny czy wykształcenie rodziców (Geiser i Santelices, 2007). Znacznie lepiej odzwierciedlają kapitał społeczny wynikający z wyższego SES niż oceny szkolne, które silniej wiążą się z osiągnięciami w zakresie specyficznych celów nauczania czy zachowaniem w klasie (Willingham i in., 2002). Na podobną tendencję wskazują analizy przeprowadzone przez *College Entrance Examination Board* (2005). Średnie wyniki SAT w części „Rozumowanie" (*Reasoning)* różnią się o ponad jedno odchylenie standardowe w teście słownym, a prawie jedno odchylenie standardowe w teście matematycznym dla uczniów pochodzących z rodzin o wysokich dochodach, w porównaniu do rodzin o niższych dochodach. Również w literaturze polskiej (Dolata, 2008) zaznacza się, że choć testy miały zapewnić trafność decyzji selekcyjnych na późniejszych etapach kształcenia, mogą być narzędziem stronniczym ze względu na korelację ze statusem społeczno-ekonomicznym ucznia.

Drugim czynnikiem wpływającym na wariancję wyników egzaminów zewnętrznych są charakterystyki demograficzne i charakterystyki szkoły, które wyjaśniają 23% ich zróżnicowania, natomiast odpowiadają jedynie za 5% wariancji poszczególnych ocen szkolnych[3] (Rothstein, 2004). Dodatkowo na obniżenie rzetelności wyników standaryzowanych testów wpływa to,

[3] W Stanach Zjednoczonych zazwyczaj wykorzystuje się następującą skalę ocen: A, B, C, D i F (*failing*), gdzie A jest oceną najlepszą, a F oznacza niezaliczenie danego kursu (np. Arizona Department of Education, 2013).

że opierają się na pojedynczym pomiarze, podczas gdy ocena i średnia ocen szkolnych (np. *high school grade point average*, HSGPA) opierają się na powtarzalnym pomiarze osiągnięć uczniów w różnych okolicznościach (Geiser i Santelices, 2007).

Podsumowując, oba wskaźniki osiągnięć ucznia są obarczone błędem pomiarowym. Do systematycznych błędów można zaliczyć różnice między skalami (kilkupunktowa skala ocen szkolnych i obejmująca kilkadziesiąt punktów skala na egzaminach zewnętrznych). Wariancja ocen szkolnych, poza poziomem wiedzy i umiejętności, może wynikać ze zróżnicowania między szkołami (w tym z różnych regulaminów oceniania wewnątrzszkolnego) oraz oddziałami szkolnymi (w tym różnic w programie i wybranych podręcznikach; Willingham i in., 2002). Zróżnicowanie wyników testów z kolei może być uwarunkowane charakterystykami społeczno-demograficznymi uczniów. Innym czynnikiem powodującym rozbieżności obu wskaźników osiągnięć gimnazjalistów jest różna skala nieetycznych zachowań egzaminacyjnych, takich jak ściąganie (Szaleniec, 2006).

Samodzielnie każdy z tych wskaźników (oceny szkolne i wyniki egzaminów zewnętrznych) ma ograniczoną moc predykcyjną. Przykładowo, William Kidder i Jay Rosner (2002) wskazują, że wynik SAT dodaje jedynie 5,4% do procentu wariancji wyjaśnionej przez średnią ocen z liceum (HSGPA). Na podstawie przytoczonej literatury można wnioskować, że chociaż HSGPA jest najlepszym pojedynczym predyktorem dla osiągnięć na studiach, to największą moc predykcyjną uzyskuje w połączeniu z wynikami standaryzowanych testów. Uwzględnienie w predykcji tego typu testów podnosi poziom wyjaśnionej wariancji osiągnięć akademickich (Geiser i Santelices 2007; Kobrin i in., 2008). Ze względu na wady i zalety obydwu wskaźników należy traktować je komplementarnie i używać jednocześnie w predykcji przyszłych osiągnięć edukacyjnych uczniów.

Trafność predykcyjna wskaźników osiągnięć szkolnych

Wskaźniki osiągnięć uczniów, niezależnie od tego, czy są nimi wyniki egzaminów zewnętrznych, czy też oceny szkolne, powinny podlegać ocenie pod kątem rzetelności i trafności (zwłaszcza, jeśli są wykorzystywane w procesach selekcji na kolejnych etapach kształcenia[4]). Nie będziemy w tym artykule analizować zagadnienia rzetelności (dokładności pomiaru), pamiętając jednak, że jest ona warunkiem koniecznym trafności wskaźników osiągnięć uczniów. Tę z kolei, nawiązując do tradycyjnych typologii (Gipps, 1994; Hornowska, 2007; Niemierko, 1999), można podzielić na: trafność teoretyczną (*construct validity*), trafność treściową/wewnętrzną (*content validity*) oraz trafność kryterialną (*criterion-related*), a w jej ramach można wyróżnić trafność diagnostyczną (*concurrent*) i predykcyjną (*predictive*).

Samuel Messick (1989, za: Hornowska, 2007, s. 81) proponuje holistyczną koncepcję trafności testu jako „zintegrowanego procesu oceny stopnia, w jakim dowody empiryczne i rozważania natury teoretycznej potwierdzają adekwatność i poprawność interpretacji oraz programów działania wyprowadzonych na podstawie wyników testowych czy innych narzędzi pomiaru".

[4] W Polsce w poszczególnych województwach kuratoria oświaty nadają różną wagę ocenom szkolnym na świadectwie i szczególnym osiągnięciom ucznia, co także może modyfikować proces rekrutacji. W poszczególnych województwach różnice w wagach nadawanych przez kuratorium (dla roku 2014/2015) ocenom szkolnym i szczególnym osiągnięciom ucznia są dosyć wysokie. Przykładowo, w województwie wielkopolskim maksymalna liczba punktów do uzyskania za szczególne osiągnięcia to 52 (48 punktów za oceny), natomiast w województwach lubelskim, dolnośląskim, małopolskim, podkarpackim i warmińsko-mazurskim to 20 (przy 80 możliwych do uzyskania punktach za oceny).

Podobne rozumienie trafności pojawia się w *Standardach dla testów stosowanych w psychologii i pedagogice* (AERA, APA i NCME, 2007, s. 31), gdzie trafność jest definiowana jako „stopień, w jakim dane empiryczne oraz teoria uzasadniają interpretację wyników pomiaru". Obie definicje kładą więc nacisk nie tyle na trafność samych narzędzi badawczych, lecz na interpretację wyników i wniosków wyprowadzonych na ich podstawie. W kontekście tematu artykułu możemy więc mówić o trafności ocen szkolnych (lub innych wskaźników osiągnięć szkolnych), biorąc pod uwagę konsekwencje ich wykorzystywania np. do celów rekrutacyjnych. Koncepcja Messicka jest nazywana modelem unitarnym (ze względu na jednorodność zjawiska trafności), w polskiej literaturze pedagogicznej i psychologicznej pojawiła się w drugiej połowie lat 90. ubiegłego wieku (Skorupiński, 2013).

Nie bez powodu badacze zajmujący się społecznym aspektem badań nad trafnością w największym stopniu skupiają się na mocy predykcyjnej różnych wskaźników osiągnięć uczniów (Findley, 1963). Badania nad trafnością predykcyjną mają silny komponent praktyczny – pozwalają ocenić możliwość przewidywania przyszłych osiągnięć uczniów na podstawie wyników uzyskiwanych w szkole i wyników standaryzowanych testów. To z kolei daje możliwość reagowania na poszczególnych etapach kształcenia, a także dostarcza wiedzy użytecznej dla rozwoju środowiska uczenia się oraz tworzenia lokalnej, regionalnej czy krajowej polityki oświatowej.

W Stanach Zjednoczonych najpopularniejszym nurtem badań nad trafnością predykcyjną wskaźników osiągnięć uczniów jest wykorzystanie wyników standaryzowanych testów, np. SAT lub ACT (*American College Testing*) i/lub średniej ocen szkolnych z liceum (HSGPA). Na ich podstawie przewiduje się osiągnięcia ucznia na studiach, mierzone za pomocą FGPA (*freshman grade point average*), czyli średniej ocen z I roku studiów oraz CGPA/ /CFYG (*cumulative grade point average/ /cumulative fourth-year grades*) – skumulowanej średniej ocen z kilku, zazwyczaj czterech, lat studiów.

Odnosząc się do pierwszego wskaźnika – wyników egzaminów zewnętrznych, badacze wskazują na ich trafność predykcyjną w stosunku do osiągnięć na etapie kształcenia wyższego (Camara i Echternacht, 2000). Przy czym Saul Geiser i Roger Studley (2004) podkreślają, że w przeciwieństwie do trafności testu SAT I (matematyka, pisanie, czytanie ze zrozumieniem), zagadnienie trafności testu SAT II (egzaminy z konkretnych przedmiotów) jest przedmiotem dyskusji (brakuje przekonujących dowodów dotyczących trafności tego narzędzia). Równocześnie, wskazuje się, że trafność predykcyjna większości standaryzowanych testów systematycznie spada (np. Willingham, Lewis, Morgan i Ramist, 1990).

Drugi wskaźnik osiągnięć ucznia, średnia ocen z przedmiotów w szkole średniej, jest uważany za najlepszy pojedynczy predyktor osiągnięć studentów zarówno na I roku studiów (*freshman grades in college*), jak i uzyskanych w toku całych studiów (CFYG), niezależnie od dyscypliny akademickiej (np. Atkinson i Geiser, 2009; Zahner, Ramsaran i Steedle, 2012). Nawet w przypadku wykorzystania jako zmiennej zależnej binarnego kryterium „ukończył(-a) studia/nie ukończył(-a) studiów", HSGPA jest jedynym istotnym statystycznie predyktorem dla wszystkich dyscyplin akademickich przy kontroli innych czynników (Geiser i Santalices, 2007). Średnia ocen z przedmiotów w szkole średniej wyjaśnia około 30% wariancji średniej ocen z I roku studiów (Atkinson, 2001, Kobrin i in., 2008, Zahner i in., 2012). Co więcej, HSGPA zyskuje na mocy predykcyjnej po I roku studiów (Geiser i Santelices, 2007). Zarówno ocenianie szkolne, jak i ocenianie

na studiach opiera się na podobnych narzędziach sprawdzania wiedzy i umiejętności (testy, eseje, aktywność na zajęciach), więc osiągnięcia w liceum mogą być trafnym predyktorem osiągnięć na studiach (Geiser i Santalices, 2007; Zahner i in., 2012). Choć HSGPA jest najlepszym pojedynczym predyktorem osiągnięć na studiach, to największą moc predykcyjną uzyskuje w połączeniu z wynikami standaryzowanych testów (Hezlett i in., 2001; Kobrin i in., 2008; Zahner i in., 2012). Richard Atkinson i Saul Geiser (2009) wskazują, że HSGPA i wyniki SAT w części „Pisanie" (*Writing*) są najlepszymi predyktorami średniej ocen na IV roku studiów i razem wyjaśniają około 30% jej wariancji.

Według Saula Geisera i Marii Santelices (2007) HSGPA wyjaśnia samodzielnie 20,4% wariancji całościowych, czteroletnich osiągnięć akademickich (CFYG). Wyniki standaryzowanych testów wyjaśniają 13,4% zmienności osiągnięć akademickich przy kontroli zmiennych mierzących status społeczno-ekonomiczny uczniów. Zmiana w średniej wynosząca jedno odchylenie standardowe prowadzi do zmiany w osiągnięciach akademickich o około jedną trzecią odchylenia standardowego. Wszystkie predyktory ujęte w jednym modelu: HSGPA, wyniki standaryzowanych testów oraz status społeczno-ekonomiczny, wyjaśniają łącznie 27% zmienności osiągnięć studentów (jednocześnie wariancja wewnątrz kohort wiekowych studentów, jak i w ramach dyscyplin akademickich, jest bardzo mała i wynosi setne części odchylenia standardowego czteroletniego GPA studenta). W analizie Jesse Rothsteina (2004) charakterystyki ucznia (rasa, płeć), charakterystyki szkoły (proporcja uczniów danej rasy, przeciętny poziom wykształcenia rodziców oraz proporcja uczniów otrzymująca dotowane obiady) razem z HSGPA wyjaśniły 45% wariancji średniej ocen z I roku studiów.

Wykorzystanie średniej ocen szkolnych jako predyktora późniejszych osiągnięć jest lepszym rozwiązaniem niż wykorzystanie pojedynczych ocen. Heinz Schuler, Uwe Funke i Jutta Baron-Boldt (1990) przeprowadzili metaanalizę, z której wynika, że trafność predykcyjna pojedynczych ocen jest niższa niż dla średniej ocen wykorzystanej jako zmienna niezależna. Średnią ocen można potraktować więc jako skuteczniejszy predyktor. Co więcej, średnia ocen szkolnych nie jest silnie związana z czynnikami takimi jak dochody rodziny czy wykształcenie rodziców (w przeciwieństwie do wyników standaryzowanych testów, co opisujemy w poprzednim podrozdziale).

Cel badania

Celem tego artykułu jest określenie trafności predykcyjnej ocen szkolnych uzyskanych przez uczniów na zakończenie pierwszego semestru w III klasie gimnazjum względem zewnętrznego kryterium, jakim jest wynik ucznia uzyskany na standaryzowanym teście, tj. egzaminie gimnazjalnym. Cel główny może być zoperacjonalizowany w postaci czterech celów szczegółowych:

- oszacowanie zróżnicowania wewnątrzklasowego umiejętności gimnazjalistów w latach 2012 i 2013,
- oszacowanie wielkości zmiany (współczynnika regresji) w wynikach egzaminu gimnazjalnego przy użyciu różnych predyktorów osiągnięć (oceny lub średniej ocen),
- rozstrzygnięcie, czy wyniki egzaminu gimnazjalnego skuteczniej można przewidywać na podstawie oceny z przedmiotu odpowiadającego rodzajowi egzaminu gimnazjalnego, czy też średniej ocen z kilku przedmiotów,
- uzasadnienie wyboru modelowania hierarchicznego do analizy problemu badawczego oraz przedstawienie podstawowych zasad posługiwania się tą techniką.

Opis wykorzystanych danych

Informacje o ocenach wystawionych na koniec pierwszego półrocza pochodzą z II i III etapu badań zrównujących dla egzaminu gimnazjalnego przeprowadzonych w Instytucie Badań Edukacyjnych. W analizach, których wyniki prezentujemy, szacowano moc predykcyjną ocen szkolnych z następujących przedmiotów[5]: język polski, historia, matematyka, geografia, biologia, fizyka i chemia.

W Tabeli 1 przedstawiono wielkość losowych prób uczniów, którzy przystąpili do egzaminu gimnazjalnego i uczestniczyli w badaniach zrównujących w poszczególnych latach, a także specyfikację, które oceny zostały wykorzystane jako predyktory w modelu. Populację docelową stanowili uczniowie klas III ze szkół dla młodzieży, z wyłączeniem szkół specjalnych i przyszpitalnych. Operatem pierwotnym w losowaniu była lista szkół udostępniona przez Centralną Komisję Egzaminacyjną w 2011 r. Próba losowa szkół miała charakter wielopoziomowy, warstwowy z prawdopodobieństwem wylosowania szkoły proporcjonalnym do liczby klas (Szaleniec i in., 2013).

Wykorzystane metody

W analizach wykorzystano model wielopoziomowej regresji (*multilevel regression model*), ponieważ struktura populacji uczniów jest hierarchiczna: uczniowie grupują się w oddziałach szkolnych, a te z kolei w szkołach (Domański i Pokropek, 2011; Geiser i Santelices, 2007). Jeśli jednostki obserwacji (uczniowie) są zgrupowane, to uczniowie z tego samego oddziału są do siebie bardziej podobni (pod względem różnych

[5] Oceny traktowano jako zmienne mierzone na ilościowej skali. Wynika to z tradycyjnego w edukacji podejścia do ocen, np. obliczania na ich podstawie średniej arytmetycznej. Autorzy powtórzyli analizy w modelu hierarchicznym, traktując oceny jako zmienne kategorialne, co wykazało brak znaczących różnic w wynikach (maksymalne różnice w odsetku wyjaśnionej wariancji wyników egzaminu gimnazjalnego sięgały 0,011). W literaturze (Konarzewski, 1991; Dolata, 2008) podkreśla się, że przyjęcie założenia o tym, że oceny szkolne są wskaźnikiem definicyjnym (Nowak, 1970) osiągnięć dopuszcza wykonywanie na nich operacji numerycznych zazwyczaj zarezerwowanych dla skal przedziałowych.

Tabela 1
Wielkości prób dla poszczególnych części egzaminu gimnazjalnego w latach 2012–2013

Rodzaj egzaminu	Rok egzaminu	Nazwa	Przedmioty, których oceny wykorzystano jako predyktory[(a)]	Wielkość próby
Matematyka	2012	MAT 2012	matematyka	1 645
	2013	MAT 2013	matematyka	1 682
Przyroda	2012	PRZYR 2012	geografia, fizyka, biologia, chemia	1 642
	2013	PRZYR 2013	geografia, fizyka, biologia, chemia	1 679
Język polski	2012	POL 2012	język polski	1 656
	2013	POL 2013	język polski	1 700
Historia[(b)]	2012	HIS 2012	historia	1 655
	2013	HIS 2013	historia	1 701

[(a)] Część przyrodnicza egzaminu gimnazjalnego ma charakter międzyprzedmiotowy, składa się z zadań odpowiadających więcej niż jednemu przedmiotowi szkolnemu. Dlatego zamiast pojedynczej oceny, używano w tym wypadku średniej ocen z przedmiotów przyrodniczych. Skuteczność takiego predyktora porównywano ze skutecznością średniej ze wszystkich wskazanych powyżej przedmiotów.

[(b)] Z uwagi na braki danych w ocenach z wiedzy o społeczeństwie (62 przypadki w 2012 r., 174 w 2013 r.), wynikające z faktu, że w niektórych szkołach nie uczono tego przedmiotu w pierwszym półroczu klasy III gimnazjum, zdecydowano się użyć jako predyktora wyłącznie oceny z historii.

charakterystyk) niż uczniowie z innych oddziałów. To z kolei narusza założenie o niezależności obserwacji wykorzystywanych w analizie. Zależności między zgrupowanymi jednostkami (uczniami) można wyrazić w postaci tzw. współczynnika korelacji wewnątrzklasowej (*intraclass correlation coefficient*, ICC). Wskazuje on na korelację między dwoma losowo wybranymi uczniami z tego samego (losowo dobranego) oddziału (np. Hox, 2010; Snijders i Bosker, 1999). Formułę obliczania tego wskaźnika można przedstawić w następujący sposób:

$$\mathrm{ICC} = \frac{U_{0j}^2}{U_{0j}^2 + R_{ij}^2}$$

gdzie:

U_{0j} – efekt losowy (*random effect*) na poziomie oddziału szkolnego; reprezentuje odchylenie stałej dla *j*-tego oddziału od średniej stałej γ_{00};

R_{ij} – efekt losowy (*random effect*) na poziomie ucznia.

Współczynnik korelacji wewnątrzklasowej reprezentuje zależności wewnątrz grupy (oddziału szkolnego) i jest pojęciem równoważnym zróżnicowaniu międzygrupowemu (*between-cluster heterogenity*). Jeśli nie ma wariancji między grupami (oddziałami), obydwa współczynniki osiągają zero, a wraz ze wzrostem wariancji międzyklasowej (względem wariancji wewnątrzklasowej) ich wartość rośnie. Innymi słowy, wysoka wartość współczynnika korelacji wewnątrzklasowej wskazuje na istnienie znaczących różnic w wynikach pomiędzy grupami (Stockford, 2009).

Wykorzystanie zwykłej regresji wielokrotnej jest niewskazane, ponieważ w przypadku zgrupowania uczniów w oddziałach szkolnych mamy do czynienia ze złamaniem założenia o homoskedastyczności, a więc niezależności składnika błędu (*random residual error term*) na poziomie oddziału szkolnego od predyktorów wykorzystanych w modelu regresji (wielkość błędu nie jest stała dla wszystkich predyktorów). Użycie zwykłej regresji wielokrotnej prowadziłoby do błędnego wnioskowania o zależnościach w populacji (Hox, 2010). O znaczeniu poziomu oddziału w analizach trafności predykcyjnej wskaźników osiągnięć uczniów świadczy też fakt, że jako predyktora w analizach używa się czasami pozycji rankingowych poszczególnych oddziałów szkolnych pod względem osiąganych przez nie średnich ocen szkolnych (Blacklow, Goepp i Hojat, 1991; 1993).

Specyfikacja modelu

W analizie zastosowano model dwupoziomowy (pierwszy poziom – ucznia, drugi – poziom oddziału szkolnego), ponieważ w ramach jednej szkoły badano jeden oddział szkolny. Gdyby przebadano więcej oddziałów w ramach szkół, lepszym rozwiązaniem byłoby zastosowanie modelu trzypoziomowego, gdzie poziom pierwszy reprezentowałby poziom ucznia, drugi – oddział, a poziom trzeci – szkołę. W celu określenia, czy model dwupoziomowy jest lepiej dopasowany do danych niż prosty model regresji (nieuwzględniający hierarchicznej struktury danych), skorzystano z testu opartego na ilorazie wiarygodności (*likelihood-ratio test*, LR). Dla wszystkich analizowanych danych (niezależnie, czy predyktorem była ocena z przedmiotu czy średnia z ocen), test ten jest istotny statystycznie, co wskazuje na lepsze dopasowanie do danych modelu dwupoziomowego w porównaniu do prostego modelu regresji liniowej. Specyfikacja zastosowanego modelu regresji dwupoziomowej wyglądała następująco:

$$Y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + U_{0j} + U_{1j}x_{ij} + R_{ij}$$

gdzie:

Y_{ij} – zmienna zależna, wynik egzaminu gimnazjalnego *i*-tego ucznia w *j*-tym oddziale szkolnym;

γ_{00} – średni współczynnik stałej regresji (*mean intercept*);

$\gamma_{10}x_{ij}$ – średni współczynnik regresji dla zmiennej niezależnej (ocen/średniej ocen) *i*-tego ucznia w *j*-tym oddziale;

$U_{1j}x_{ij}$ – losowa interakcja (*random interaction*) między grupą (oddział szkolny) a zmienną niezależną (oceną/średnią ocen);

U_{1j} – losowy współczynnik regresji (*random slope)* dla zmiennej niezależnej (ocena/średnia ocen); reprezentuje odchylenie współczynnika regresji *j*-tego oddziału od średniego współczynnika regresji γ_{10}.

Parametry modelu zostały oszacowane za pomocą metody maksymalnej wiarygodności (*maximum likelihood*, ML), z niezależnie szacowanymi wariancją i kowariancją efektów losowych. Drugą popularną metodą estymacji parametrów modelu wielopoziomowego jest metoda resztowej maksymalnej wiarygodności (*residual/restricted maximum likelihood*, REML). W przypadku dużej liczby grup na drugim poziomie (więcej niż 30), a tak jest w przypadku prezentowanych analiz, różnice między obiema metodami estymacji są nieznaczące (Snijders i Bosker, 1999). Macierz kowariancji efektów losowych została oszacowana bez nakładania ograniczeń na szacowane wartości (z uwagi na brak podstaw teoretycznych do zastosowania takiego rozwiązania).

Z uwagi na to, że oceny szkolne mogą zależeć od osoby nauczyciela uczącego w konkretnym oddziale, specyfikacja uwzględnia obecność losowego współczynnika regresji (*random slope*[6]). W modelu zakładamy, że związek między ocenami szkolnymi/średnią ocen a wynikiem egzaminu gimnazjalnego może się różnić między grupami (oddziałami szkolnymi). W terminologii statystycznej zjawisko to jest znane pod nazwą heterogeniczności regresji między grupami lub interakcją między grupą a innymi kowariantami (Snijders i Bosker, 1999). Za pomocą modelu regresji wielopoziomowej oszacowano także:

- współczynniki regresji dla pojedynczych ocen szkolnych i ich średniej (cel 2),
- poziom wariancji wyników egzaminu gimnazjalnego wyjaśnionej przez pojedyncze oceny oraz średnią ocen szkolnych (R^2) (cel 3).

Używając modeli regresji wielorakiej, wartość R^2 można zinterpretować jako proporcjonalną redukcję błędu przewidywania (*proportional reduction of prediction error*), czyli wartość, o jaką zmniejsza się poziom zmienności niewyjaśnionej przez model, dzięki zastosowaniu predyktorów $X_1 - X_q$. Natomiast w przypadku modeli dwupoziomowych pojęcie proporcjonalnej redukcji błędu przewidywania można zdefiniować w dwojaki sposób:

- jako proporcjonalną redukcję błędu przewidywania na poziomie indywidualnym,
- jako proporcjonalną redukcję błędu przewidywania średniej grupowej.

Tom Snijders i Roel Bosker (1999, s. 102) podkreślają, że jeśli wartości *X* są znane, najlepszym liniowym predyktorem dla zmiennej zależnej Y_{ij} jest współczynnik regresji $\sum_{h=0}^{q}\gamma_h X_{hij}$ (gdzie $X_{ho\,j}$ jest zdefiniowany jako 1 dla wszystkich *h* i *j*). Średni kwadrat błędu przewidywania wynosi więc:

$$\mathrm{var}(\bar{Y}_{.j} - \sum_{h}\gamma_h\bar{X}_{h.j}) = U_{0j}^2 + \frac{R_{ij}^2}{n}$$

Szacowanie wielkości wyjaśnionej wariancji opiera się na określeniu stosunku średniego kwadratu błędu przewidywania dla pełnego modelu do średniego kwadratu błędu przewidywania dla modelu pustego (*null*

[6] W celu sprawdzenia, czy spośród modeli dwupoziomowych lepiej dopasowany do danych jest model z losową stałą, czy też model z dodatkowo uwzględnionym losowym współczynnikiem regresji, wykorzystano ponownie test LR. Ujawnił on, że dla wszystkich typów analizowanych danych model z losowym współczynnikiem regresji jest lepiej dopasowany do danych, co oznacza, że relacja między oceną lub średnią ocen jest specyficzna dla grupy (oddziału).

model). Snijders i Bosker (1999) zauważają, że formuła ta jest wykorzystywana do szacowania wyjaśnionej wariancji w modelu z losową stałą, ale może ona być zaaplikowana także do liczenia poziomu wyjaśnionej wariancji w modelu uwzględniającym losowy współczynnik regresji (ze względu na niemal identyczne wyniki).

Proporcjonalna redukcja błędu przewidywania (R_1^2) na pierwszym poziomie będzie równa:

$$R_1^2 = 1 - \frac{\mathrm{var}(Y_{ij} - \sum_h \gamma_h X_{hij})}{var(Y_{ij})}$$

Średni kwadrat błędu przewidywania dla drugiego poziomu wynosi:

$$\mathrm{var}(\bar{Y}_{.j} - \sum_h \gamma_h \bar{X}_{h.j}) = U_{0j}^2 + \frac{R_{ij}^2}{n}$$

Aby uwzględnić nierówną liczebność oddziałów szkolnych (Snijders i Bosker, 1999) zastosowano poprawkę polegającą na użyciu średniej harmonicznej liczonej jako $N/\{\Sigma_j\left(\frac{1}{n_i}\right)\}$, gdzie n_j to liczebność *j*-tego oddziału szkolnego. Proporcjonalna redukcja błędu przewidywania (R_2^2) na drugim poziomie może być oszacowana za pomocą następującego wzoru:

$$R_2^2 = 1 - \frac{\mathrm{var}(\overline{Y_{.j}} - \sum_h \gamma_h \bar{X}_{h.j})}{\mathrm{var}(\bar{Y}_{.j})}$$

Należy jednak pamiętać, że zaproponowana miara jest jedynie analogiem miary R^2 z regresji liniowej, z uwagi na to, że wariancja danych pochodzi z dwóch poziomów analizy (w tym wypadku z poziomu ucznia i poziomu oddziału szkolnego). Każdy poziom danych wymaga wyliczenia osobnej macierzy kowariancji efektów losowych (*random effects covariance matrix*) oraz wariancji błędu, które mogą zachowywać się w sposób nieoczekiwany (Recchia, 2010).

Zainteresowany czytelnik może zapoznać się z 95-procentowymi przedziałami ufności dla U_{0j}^2 oraz R_{ij}^2 (dla analizowanych modeli pustych i z predyktorem – oceną i średnią ocen), które zamieszczono w Aneksie. W przypadku R^2 oszacowanie przedziału ufności może być obciążone błędem. Do oszacowania przedziału ufności potrzebne są oszacowania 95-procentowego przedziału ufności U_{0j}^2 oraz R_{ij}^2, zarówno dla modelu pustego, jak i modelu z predyktorem (ocena i średnia ocen). Kowariancja tych wartości oraz funkcja opisująca stosunek tych dwóch wartości jest nieznana. Rozkłady błędów dla tych elementów mogłyby być szacowane z wykorzystaniem metody *bootstrap* (Efron, 1982). W literaturze przedmiotu brakuje jednak informacji na temat konsekwencji takiego wyboru w przypadku szacowania przedziałów ufności w modelu wielopoziomowym z losowym nachyleniem. Określenie stabilności takiego podejścia wymagałoby osobnego opracowania.

Wyniki

W Tabeli 2 przedstawiono zestawienie współczynników korelacji wewnątrzklasowej (ICC) dla poszczególnych testów i poszczególnych lat, czyli oszacowanie podobieństwa w wynikach egzaminacyjnych między dwoma losowo wybranymi uczniami z danej klasy. ICC wyliczono za pomocą poziomu umiejętności uczniów (θ_{ij}) szacowanego na podstawie surowych wyników egzaminu. Do oszacowania wskaźnika umiejętności uczniów (θ_{ij}) wykorzystano trzyparametryczny model logistyczny dla zadań punktowanych 0–1 oraz model GRM dla zadań o maksymalnej liczbie punktów większej niż 1 (Samejima, 1969). W specyfikacji modelu pustego wykorzystano 15 losowych wartości z rozkładu a posteriori umiejętności ucznia (*plausible values*, PV; np. Zwick i Mislevy, 2011). Na podstawie

Tabela 2
Współczynniki korelacji wewnątrzklasowej dla wyników poszczególnych części egzaminów w latach 2012 i 2013 (z uwzględnieniem 95-procentowych przedziałów ufności)

Egzamin	MAT 2012	MAT 2013	PRZYR 2012	PRZYR 2013	POL 2012	POL 2013	HIS 2012	HIS 2013
ICC	0,28	0,35	0,27	0,34	0,31	0,26	0,24	0,25
95% przedział ufności	(0,21; 0,36)	(0,27; 0,43)	(0,20; 0,35)	(0,26; 0,41)	(0,24; 0,39)	(0,20; 0,34)	(0,18; 0,32)	(0,19; 0,33)

oszacowanych w modelu wielopoziomowym przedziałów ufności dla wariancji na poziomie ucznia i oddziału szkolnego, obliczono minimalną i maksymalną granicę przedziału dla ICC.

ICC obliczone na podstawie modeli pustych z piętnastoma PV[7] są najwyższe dla matematyki w 2013 r. oraz przyrody w 2013 r. Najniższe ICC można zaobserwować dla obydwu egzaminów z historii (w sesjach 2012 i 2013). Analiza 95-procentowych przedziałów ufności nie pozwala jednak stwierdzić w sposób uprawniony, że różnice w wartościach ICC pomiędzy różnymi przedmiotami są istotne statystycznie (Tabela 2).

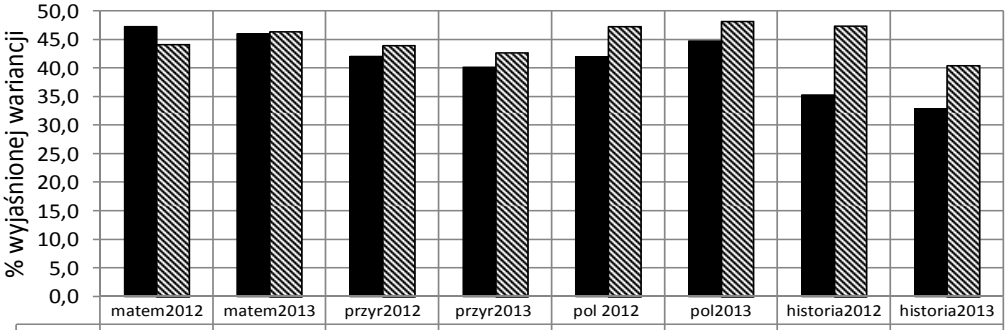
W Tabeli 3 przedstawiono współczynniki regresji (γ_{10}) dla poszczególnych testów[8]. Zmiana na skali ocen szkolnych (lub ich średniej) o jeden wiąże się ze znaczną zmianą liczby punktów uzyskanych na egzaminie gimnazjalnym. W przypadku matematyki wzrost oceny szkolnej o 1 odpowiada wynikowi egzaminu gimnazjalnego wyższemu o ok. 4 punkty (w odniesieniu do 29–30 punktów maksymalnie możliwych do uzyskania). W przypadku egzaminu z historii w 2013 r. związek ten jest najsłabszy wśród wszystkich analizowanych przedmiotów w obu sesjach egzaminacyjnych, biorąc

[7] Po oszacowaniu 15 PV dla każdej szkoły, obliczono 15 modeli pustych z poszczególnymi PV jako zmiennymi zależnymi, następnie oszacowano na ich podstawie 15 ICC, po czym uśredniono ich wartości dla poszczególnych testów.

[8] Należy zwrócić uwagę, że dokonywanie bezpośrednich porównań pomiędzy analizowanymi przedmiotami jest nieuprawnione, ze względu na różną maksymalną liczbę punktów egzaminacyjnych. Możliwe jest dokonywanie porównań współczynników regresji oceny i średniej ocen w ramach jednego egzaminu.

Tabela 3
Wielkości współczynników regresji (i ich 95% przedziały ufności) dla oceny szkolnej i średniej ocen szkolnych wraz z maksymalną punktacją poszczególnych egzaminów

Egzamin	Współczynnik regresji dla pojedynczej oceny		Współczynnik regresji dla średniej ocen		Maksymalna liczba punktów do uzyskania z danego egzaminu
	Estymator punktowy	95% przedział ufności	Estymator punktowy	95% przedział ufności	
Matematyka 2012	3,59	(3,37; 3,80)	4,16	(3,90; 4,42)	29
Matematyka 2013	4,07	(3,79; 4,34)	4,58	(4,29; 4,87)	30
Przyroda 2012	2,87	(2,64; 3,10)	3,02	(2,80; 3,25)	26
Przyroda 2013	2,96	(2,74; 3,18)	3,18	(2,96; 3,41)	28
Polski 2012	3,71	(3,40; 4,01)	4,26	(3,93; 4,61)	32
Polski 2013	3,90	(3,60; 4,19)	4,50	(4,20; 4;80)	32
Historia 2012	3,45	(3,16; 3,74)	4,41	(4,08; 4,75)	33
Historia 2013	2,92	(2,63; 3,21)	3,60	(3,31; 3,90)	33

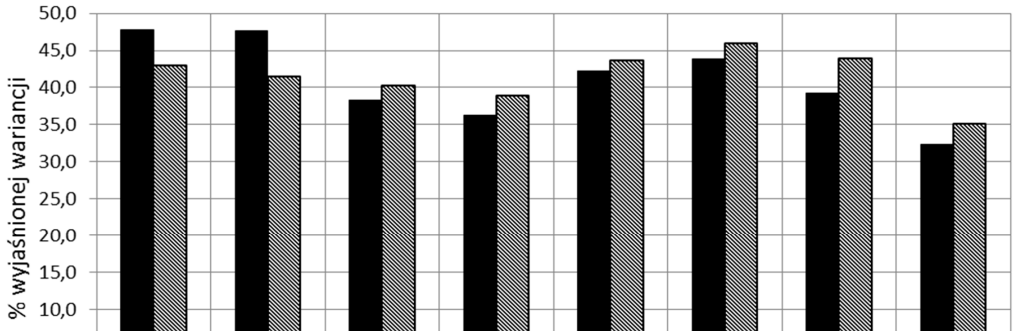


	matem2012	matem2013	przyr2012	przyr2013	pol 2012	pol2013	historia2012	historia2013
ocena	47,3	46,0	42,0	40,2	42,0	44,7	35,3	32,8
średnia	44,1	46,4	44,0	42,7	47,3	48,3	47,4	40,5

Rysunek 1. Procent wyjaśnionej wariancji wyników egzaminu gimnazjalnego na poziomie indywidualnym (ucznia).

pod uwagę najdłuższą skalę punktów dla tego egzaminu. Wzrost oceny (lub średniej ocen) o jeden prowadzi do wzrostu wyniku egzaminu gimnazjalnego z historii o około 3 punkty (przy nieco dłuższej skali, gdyż maksymalna liczba punktów do uzyskania w egzaminie z historii wynosiła 33).

Rysunki 1 i 2 prezentują procenty wyjaśnionej wariancji na obu poziomach analizy, obliczone zgodnie z procedurami zaprezentowanymi w części poświęconej specyfikacji modelu. Rysunek 1 wskazuje, że najwięcej wariancji wyniku egzaminacyjnego w zakresie matematyki na poziomie indywidualnym wyjaśnia w 2012 r. pojedyncza ocena z matematyki. Jednocześnie ma ona najwyższą moc predykcyjną dla wyników egzaminu gimnazjalnego w porównaniu z ocenami z innych przedmiotów nauczanych w gimnazjum. Najmniej wariancji na poziomie indywidualnym wyjaśniają pojedyncze oceny z historii z lat 2012 i 2013.



	matem2012	matem2013	przyr2012	przyr2013	pol 2012	pol2013	historia2012	historia2013
ocena	47,8	47,6	38,3	36,3	42,2	43,8	39,2	32,3
średnia	42,9	41,4	40,2	38,9	43,6	45,9	43,9	35,1

Rysunek 2. Procent wyjaśnionej wariancji wyników egzaminu gimnazjalnego na poziomie grupy (oddziału).

Dla tych egzaminów przyrost wyjaśnionej wariancji po zmianie predyktora z oceny na średnią ocen jest także największy (różnica odpowiednio: 12,1 punktu procentowego w 2012 r. i 7,6 p.p. w 2013 r.). Należy zwrócić uwagę, że zastąpienie pojedynczej oceny średnią zmniejsza poziom wyjaśnionej wariancji dla egzaminu z matematyki w 2012 r., natomiast w 2013 r. przyrost ten wynosi 0,3 punktu procentowego.

Biorąc pod uwagę oceny szkolne, poziom wyjaśnionej wariancji na poziomie oddziału szkolnego (Rysunek 2) jest najwyższy dla egzaminów z matematyki w sesjach 2012 i 2013. Najniższy procent wyjaśnionej wariancji na poziomie oddziału można zaobserwować dla historii i przyrody w 2013 r. Dzięki zastąpieniu pojedynczej oceny ich średnią, w największym stopniu polepsza się predykcja dla egzaminu z historii (o 4,7 p.p. dla 2012 r. i 2,8 p.p. dla 2013 r.), choć zmiana nie jest tak duża jak na poziomie indywidualnym. Podobnie, jak w przypadku predykcji wyniku egzaminacyjnego z matematyki na poziomie indywidualnym – zastąpienie pojedynczej oceny ich średnią, zmniejsza procent wyjaśnionej wariancji (odpowiednio o 4,9 i 6,2 p.p.).

Dyskusja wyników

Przeprowadzone analizy sugerują, że oceny szkolne (a także ich średnia) są bardzo dobrymi predyktorami wyników egzaminacyjnych. Pojedyncze oceny wyjaśniają od 32,8 do 47,3% zróżnicowania wyników egzaminacyjnych na poziomie indywidualnym, średnia ocen wyjaśnia od 40,4 do 48,2%. Wzrost oceny o jeden stopień przekłada się na przyrost wyniku egzaminacyjnego od 2,9 do 4,1 punktów (przy skali maksymalnej punktacji na egzaminie 26–33), wzrost średniej ocen o jeden – przyrost wyniku egzaminacyjnego od 3 do 4,6 punktów. Oznacza to, że zmiana wyniku egzaminu gimnazjalnego o jeden punkt może mieć decydującą rolę dla losów ucznia, dla którego znaczenie może mieć dostanie się do konkretnego oddziału szkolnego w danej szkole ponadgimnazjalnej. Z kolei jakość przewidywania późniejszych osiągnięć polskich uczniów na podstawie ocen szkolnych jest porównywalna z jakością przewidywania raportowaną przez przytoczone wcześniej badania zagraniczne. Wszystkie te wyniki świadczą o tym, że oceny szkolne są dobrym predyktorem późniejszych osiągnięć uczniów i można je wykorzystywać do wielu analiz implikujących rozwiązania praktyczne.

W przypadku wszystkich analizowanych testów można zauważyć dość wysokie zróżnicowanie międzygrupowe, o czym świadczą wielkości korelacji wewnątrzklasowej na poziomie od 0,24/0,25 (historia 2012 i 2013) do 0,34/0,35 (przyroda i matematyka 2013). Korelacja wyników egzaminacyjnych dwóch losowo wybranych uczniów z tego samego losowo dobranego oddziału szkolnego jest dość wysoka, co może świadczyć o znacznej selekcyjności oddziałów szkolnych. Nie można jednak w sposób uprawniony wnioskować o istotnych statystycznie różnicach między wartościami ICC ze względu na konkretne przedmioty szkolne. Pełne wyjaśnienie opisanych zjawisk wymaga dalszych badań, podobnie jak uwzględnienie jako dodatkowych predyktorów takich zmiennych, jak płeć uczniów i wielkość miejscowości zamieszkania. Przykładowo Ruth Ekstrom (1994) wskazuje, że dziewczęta uzyskują wyższe oceny niż chłopcy, natomiast chłopcy uzyskują lepsze wyniki w standaryzowanych testach – co oznacza, że nauczyciele mogą mieć inne oczekiwania wobec chłopców i dziewcząt podczas oceniania. Na podobny efekt wskazał Konarzewski (1991), mówiąc o tym, że dziewczynkom podczas pierwszych etapów nauczania częściej stawia się zadania-ćwiczenia (wymagające rutynowego wykonania), natomiast chłopcom

zadania-problemy (wymagające twórczych rozwiązań). Nie wiadomo jednak, czy ten efekt utrzymuje się na poziomie gimnazjum i dalszych etapach edukacyjnych. Z kolei po zrównaniu wyników egzaminów gimnazjalnych dziewczęta uzyskują wyższe wyniki niż chłopcy w części humanistycznej, natomiast takie różnice nie występują w części matematyczno-przyrodniczej. Innym czynnikiem wpływającym na zróżnicowanie międzygrupowe jest status społeczno-ekonomiczny rodzin uczniów (Dolata, 2008). Wskaźnik segregacji międzyszkolnej wyjaśnia 19% wariancji SES, natomiast wskaźnik segregacji międzyklasowej wyjaśnia 35% wariancji SES. Wyniki badań wskazują, że siła determinacji związanej ze statusem społeczno-ekonomicznym jest silniejsza w mieście niż na wsi. Co więcej, w dużych miastach przejście uczniów ze szkoły podstawowej do gimnazjum oznacza także nasilenie segregacji gimnazjalistów pod względem poziomu uprzednich osiągnięć szkolnych (mierzonych wynikiem osiąganym na sprawdzianie po klasie VI). W przypadku gimnazjów wiejskich efekt ten nie występuje – można powiedzieć nawet o spadku zróżnicowania międzygrupowego na poziomie szkoły (Dolata, 2008). Wszystko to oznacza, że zmienna określająca wielkość i rodzaj miejscowości zamieszkania może być również kluczowa w pogłębieniu tych analiz.

W kontekście trafności predykcyjnej ocen szkolnych interesujący wydaje się wysoki procent wyjaśnionej wariancji wyników egzaminacyjnych przez pojedynczą ocenę szkolną z matematyki oraz stosunkowo niski procent wyjaśnionej wariancji przez pojedynczą ocenę z historii, a także jego znaczne zwiększenie przez zmianę predyktora na średnią wszystkich ocen szkolnych. Wyniki dla przyrody (stosunkowo wysoki stopień wyjaśnionej wariancji dzięki mniej złożonemu predyktorowi i niewielki przyrost przez zastąpienie go średnią ze wszystkich ocen) przypominają w większym stopniu wyniki dla matematyki niż historii.

Wysoka trafność predykcyjna ocen szkolnych w stosunku do wyników egzaminu gimnazjalnego z matematyki (niezależnie czy przewiduje się wynik indywidualnego ucznia, czy też wynik na poziomie całego oddziału szkolnego) może być spowodowana znacznym pokrywaniem się treści nauczania na lekcjach matematyki w gimnazjum i umiejętności mierzonych na egzaminie gimnazjalnym. W literaturze od dawna znane jest zjawisko zawężania treści nauczanych w szkole (*narrowing the curriculum)* i dostosowywania ich do treści sprawdzanych w egzaminie (Au, 2007). Ponieważ arkusze egzaminów zewnętrznych zawierają zadania, które są tylko próbką całego spektrum poleceń mogących mierzyć wymagane umiejętności, z natury mierzą węższy zakres kompetencji, niż ma to miejsce w trakcie procesu nauczania szkolnego.

Jeżeli wynik egzaminacyjny jest wykorzystywany jako główna miara skuteczności szkoły, to nauczyciele stawiają sobie za cel przygotowanie ucznia do uzyskania jak najwyższego wyniku na egzaminie. Dlatego też mogą w znacznej mierze poświęcać czas na lekcji (który jest ograniczony) na nauczanie specyficznych umiejętności mających zwiększyć szansę ucznia na osiągnięcie tego celu. Często w literaturze efekt ten nazywa się nauczaniem pod test (*teaching to the test*) i wiąże się nie tylko z położeniem zbyt dużego nacisku na nauczanie specyficznych dla standaryzowanych testów umiejętności, ale także z wykluczeniem z procesu dydaktycznego rozwijania umiejętności społecznych oraz pominięciem czynników kulturowych i kontekstowych w nauczaniu (Linn, 2000; Duffy, Giordano, Farrell, Paneque i Crump, 2008).

Badania wskazują, że czas poświęcony wyłącznie na przygotowanie do egzaminu zewnętrznego wynosi około 100 godzin, co jest ekwiwalentem około czterech tygodni nauczania wszystkich przedmiotów w szkole

(Smith, Edelsky, Draper, Rottenberg i Cherland, 1989). Zawężenie treści nauczania może być związane z wykorzystywaniem przez nauczycieli opracowań publikujących zadania z matematyki, które zostały wykorzystane we wcześniejszych testach. Przygotowując sprawdziany czy zadania do rozwiązania przez ucznia przy tablicy (których efekty stają się składową oceny semestralnej z matematyki), nauczyciele mogą korzystać z publikowanych zadań egzaminacyjnych w oryginalnej wersji lub modyfikować ich treść (np. przez zmianę danych). W związku z tym zadania rozwiązywane na lekcjach matematyki mają podobny format i zakres treściowy, co zadania występujące w standaryzowanym teście. Wskutek wysokiej doniosłości egzaminu gimnazjalnego wśród nauczycieli może wzmacniać się przekonanie, że istnieje jeden, poprawny schemat rozwiązania określonego zadania, a w związku z tym jeden, poprawny algorytm nauczania przygotowującego do rozwiązania określonego zadania. Skutecznym sposobem nauczania (z punktu widzenia uzyskania jak najlepszych wyników w egzaminie zewnętrznym) staje się rozłożenie zadania na komponenty i powtarzające się ćwiczenie umiejętności odpowiadających poszczególnym komponentom do czasu, aż uczeń je opanuje. W związku z tym uczeń podczas egzaminu nie musi się zastanawiać nad rozwiązaniem danego zadania – może bezrefleksyjnie zastosować algorytm znany mu z lekcji (Smith, 1991).

Drugim możliwym czynnikiem wyjaśniającym wysoką moc predykcyjną ocen z matematyki może być podobieństwo treści nauczania i umiejętności sprawdzanych na podstawie testu (gdy zadania w teście w wysokim stopniu reprezentują treści nauczania wynikające z podstawy programowej). Rozstrzygnięcie, czy wysoka trafność predykcyjna ocen z matematyki jest spowodowana nauczaniem pod testy, czy też wysoką adekwatnością testu względem podstawy programowej, wymaga sprawdzenia w kolejnych badaniach.

Pewną wątpliwość może budzić stosunkowo niski stopień wyjaśnionej wariancji na poziomie indywidualnym przez pojedynczą ocenę z historii i znaczny jej przyrost przez zastąpienie oceny z historii średnią ocen ze wszystkich przedmiotów. To zjawisko można interpretować na dwa różne sposoby. Egzamin gimnazjalny z historii (zgodnie z celami nauczania wymienionymi w nowej podstawie programowej na tym etapie) sprawdza umiejętności czytania ze zrozumieniem oraz analizy źródeł (np. tekstów źródłowych, map, danych statystycznych), ich syntezy z posiadaną już wiedzą, czy umiejętności dostrzegania relacji przyczynowo-skutkowych (Podstawa programowa z komentarzami, 2012). Uwzględnienie jako predyktorów ocen z języka polskiego (czytanie ze zrozumieniem) czy matematyki (umiejętności analityczne, logiczne myślenie) może więc poprawić jakość predykcji. Co więcej, jak wskazują raporty z badania *Diagnoza kompetencji gimnazjalistów* 2011 i 2012 (IBE, 2012; 2013), uczniowie wciąż mają problemy z wyżej wymienionymi umiejętnościami. Syntetyzowanie czy interpretacja danych zawartych w tekście źródłowym sprawia gimnazjalistom problem – wydaje się, że podczas analizy tekstu poprzestają jedynie na wykorzystaniu słów kluczowych[9]. Raport DKG 2011 (IBE, 2012) wskazuje, że przyczyną może być tradycyjny sposób nauczania historii,

[9] Przykładowo w badaniu DKG 2012 (IBE, 2013) przedstawiono zadanie polegające na analizie tekstu źródłowego (deklaracji Katarzyny II uzasadniającej I rozbiór Polski), a następnie wskazaniu spośród trzech odpowiedzi (Stanisława Leszczyńskiego, Augusta III Sasa i Stanisława Augusta Poniatowskiego) osoby, do której odnosi się źródło w kontekście okoliczności jej wyboru na tron. 25% uczniów wybrało Augusta III Sasa, o którym napisano wprost w pierwszych słowach tekstu („Po śmierci Augusta III […] aby nie dopuścić do zgubnych skutków, które wywołać mogło bezkrólewie w Polsce…"), że nie żyje w momencie opublikowania źródła.

ograniczony do kształtowania specyficznych umiejętności, takich jak zapamiętywanie i odtwarzanie konkretnych wydarzeń, natomiast nie na holistyczne umiejętności wyższego rzędu, jak myślenie logiczne, integracja i interpretacja informacji oraz dostrzeganie sekwencyjności i związków przyczynowo-skutkowych między nimi (co jest sprawdzane w egzaminie gimnazjalnym). Jeśli pojedyncza ocena z historii nadal odzwierciedla głównie umiejętność zapamiętywania i odtwarzania konkretnych wydarzeń, może być to przyczyną jej stosunkowo niskiej predykcji wyniku na egzaminie gimnazjalnym z historii odwołującym się do zupełnie innych kompetencji. Jeśli wyniki badań prezentowane w tym artykule zostałyby zreplikowane dla innych typów egzaminów zewnętrznych, należałoby zastanowić się, czy ze względu na niski poziom wyjaśnionej wariancji wyników ocena z historii powinna być uwzględniania w modelach i wykorzystywana jako predyktor późniejszych osiągnięć uczniów.

Obydwa wskaźniki osiągnięć szkolnych mają znaczenie dla późniejszej ścieżki edukacyjnej uczniów. O rekrutacji do szkół ponadgimnazjalnych (*Ustawa z dnia 6 grudnia 2013 r. o zmianie ustawy o systemie oświaty oraz niektórych innych ustaw*, 2014) decyduje wynik egzaminu gimnazjalnego oraz osiągnięcia szkolne (a więc oceny szkolne z języka polskiego i trzech wybranych przedmiotów oraz szczególne osiągnięcia). Zarówno wyniki egzaminu gimnazjalnego, jak i osiągnięcia szkolne mają taką samą wagę dla wyniku rekrutacji (choć miary te różnią się rzetelnością). To zalicza je do sposobów sprawdzania wiedzy ucznia, które mają poważne konsekwencje dla jego ścieżki edukacyjnej.

Przedstawiony artykuł jest pierwszą w Polsce (na taką skalę i z wykorzystaniem danych zebranych w reprezentatywnych badaniach) próbą oszacowania trafności predykcyjnej ocen szkolnych dla wyniku egzaminu gimnazjalnego. Po uzyskaniu danych z najbliższej edycji badań, autorzy planują także oszacować trafność predykcyjną ocen i wyników egzaminów zewnętrznych dla kolejnego etapu kształcenia. Celem tych analiz będzie określenie, w jakim stopniu oceny szkolne i średnia ocen w gimnazjum, a także wyniki egzaminu gimnazjalnego są predykcyjne w stosunku do osiągnięć ucznia w szkole ponadgimnazjalnej (wyników egzaminu maturalnego). Wyniki analiz prezentowanych w niniejszym artykule, wskazujące na wysoką moc predykcyjną zarówno ocen szkolnych, jak i ich średniej, uzasadniają potrzebę prowadzenia dalszych badań w tym obszarze.

Literatura

AERA, APA i NCME (2007). *Standardy dla testów stosowanych w psychologii i pedagogice*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne.

Arizona Department of Education (2013). *A–F letter grade accountability system – 2013 technical manual*. Pobrano z http://www.azed.gov/research-evaluation/files/2013/11/2013-a-f-technical-manual.pdf

Astin, A. W. (1971). *Predicting academic performance in college*. New York, NY: The Free Press.

Atkinson, R. (2001). *Standardized tests and access to American universities*. Washington, D.C.: American Council on Education. Pobrano z http://works.bepress.com/cgi/viewcontent.cgi?article=1035&context=richard_atkinson

Atkinson, R. i Geiser, S. (red.). (2009). *Reflections on a century of college admissions test*. Berkeley, CA: University of California.

Au, W. (2007). High-stakes testing and curricular control: a qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267.

Bejar, I. I. i Blew, E. O. (1981). *Grade inflation and the validity of the Scholastic Aptitude Test (College Board Report. No. 81–3)*. New York, NY: College Entrance Examination Board.

Black, P. i Wiliam, D. (1998). *Inside the black box: raising standards through classroom assessment*. London: School of Education, King's College.

Blacklow, R. S., Goepp, C. E. i Hojat, M. (1991). Class ranking models for deans' letters of recommendation and their psychometric evaluation. *Academic Medicine*, *66*(9), 10–12.

Blacklow, R. S., Goepp, C. E., Hojat, M. (1993). Further psychometric evaluation of a class ranking model as a predictor of graduates' clinical competence in the first year of residency. *Academic Medicine*, *68*(4), 295–297.

Brookhart, S. M. (1993). Teachers' grading practices: meaning and values. *Journal of Educational Measurement*, *30*(2), 123–42.

Brookhart, S. M. (1998). Why 'grade inflation' is not a problem with a 'just say no' solution. *National Forum*, *78*(2), 3–5.

Camara, W. J. i Echternacht, G. (2000). *The SAT I and high school grades: utility in predicting success in college. Research Report RN-10*. New York, NY: College Entrance Examination Board.

Camara, W. J., Kimmel, E., Scheuneman, J., i Sawtell, E. (2003). *Whose grades are inflated? (College Board Research Report No. 2003–04*). New York, NY: College Entrance Examination Board.

Camara, W. i Michaelides, M. (2005). *AP use in admissions: a response to Geiser and Santelices*. College Board Research Note, May 11, 2005. Pobrano z http://www.collegeboard.com/research/pdf/051425Geiser_050406.pdf

College Entrance Examination Board (2005). *2005 college-bound seniors: total group profile report*. Pobrano z http://www.collegeboard.com/prod_downloads/about/news_info/cbsenior/yr2005/2005-college-bound-seniors.pdf

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist*, *30*(1), 1–14.

Cureton, L. W. (1971). The history of grading practices. *Measurement in Education*, *2*(4), 1–8.

Dolata, R. (2008). *Szkoła – segregacje – nierówności*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Domański, H. i Pokropek, A. (2011). *Podziały terytorialne globalizacja a nierówności społeczne. Wprowadzenie do modeli wielopoziomowych*. Warszawa: Wydawnictwo IFiS PAN.

Duffy, M., Giordano, V. A., Farrell, J. B., Paneque, O. M. i Crump, G. B. (2008). No Child Left Behind: values and research issues in high-stakes assessments. *Counseling and Values*, *53*(1), 53–66.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Ekstrom, R., Goertz, M., i Rock, D. (1988). *Education and American youth*. London: Falmer Press.

Ekstrom, R. (1994). *Gender differences in high school grades: an exploratory study*. New York, NY: College Entrance Examination Board. Pobrano z https://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1994-3-gender-differences-high-school-grades-study.pdf http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1994-3-gender-differences-high-school-grades-study.pdf

Findley, W. G. (1963). *The impact and improvement of school testing programs*. Chicago, IL: University of Chicago Press. Pobrano z http://www.questia.com/library/353545/the-impact-and-improvement-of-school-testing-programs

Frary, R. B., Cross, L. H., i Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice*, *12*(3), 23–30.

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.* Berkeley, CA: Center for Studies in Higher Education, University of California.

Geiser, S. i Studley, R. E. (2004). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. W: R. Zwick (red.), *Rethinking the SAT: the future of standardized testing in university admissions* (1–25). New York, NY: Routledge.

Gipps, C. V. (1994). *Beyond testing. Towards a theory of educational assessment*. London: The Falmer Press.

Hansford, B. C. i Hattie, J. A. (1982). The relationship between self and achievement/performance measures. *Review of Educational Research*, *52*(1), 123–142.

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 w trakcie *National Council of Measurement in Education*. Seattle, WA.

Hornowska, E. (2007*). Testy psychologiczne. Teoria i praktyka.* Warszawa: Scholar.

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

Instytut Badań Edukacyjnych (2012). *Diagnoza Kompetencji Gimnazjalistów: Historia i WOS. Raport z badania.* Pobrano z http://eduentuzjasci.pl/images/stories/badania/diagnizakg/dkghistoria.pdf

Instytut Badań Edukacyjnych (2013). *Diagnoza Kompetencji Gimnazjalistów: Historia*. *Raport z badania*. Pobrano z http://eduentuzjasci.pl/images/stories/badania/diagnizakg/Historia_raport_DKG_2012w.pdf

Kidder, W. C., i Rosner, J. (2002). How the SAT creates „built-in headwinds": an educational and legal analysis of disparate impact. *Santa Clara Law Review*, *43*(1), 131–211.

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*. New York, NY: The College Board.

Konarzewski, K. (1991). *Problemy i schematy: pierwszy rok nauki szkolnej dziecka*. Poznań: Akademos.

Linn, R. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4–16.

Liszka, K. (2001). *Nauczycielskie kryteria oceniania uczniów*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Marquand, J., (1993). *Studium wstępne krajowego systemu oceniania w polskim szkolnictwie ponadpodstawowym*. Warszawa: Fundusz Współpracy PHARE, BKKK.

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

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

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

Podstawa programowa z komentarzami. Tom 4 – Edukacja historyczna i obywatelska w szkole podstawowej, gimnazjum i liceum. (2012). Pobrano z https://archiwum.men.gov.pl/index.php?option=com_content&view=article&id=2060%3Atom-4-edukacja-historyczna-i-obywatelska-w-szkole-podstawowej-gimnazjum-i-liceum--&catid=230%3Aksztacenie-i-kadra-ksztacenie--ogolne-podstawa-programowa&Itemid=290

Nowak, S. (1970). *Metodologia badań socjologicznych*, Warszawa: PWN.

Ramist, L., Lewis, C., i McCamley, L. (1990). Implications of using freshman GPA as the criterion for the predictive validity of the SAT. W: W. W. Willingham, C. Lewis, R. Morgan, i L. Ramist (red.), *Predicting college grades: an analysis of institutional trends over two decades* (253–288). Princeton, NJ: Educational Testing Service.

Recchia, A. (2010). R-squared measures for two-level hierarchical linear models using SAS. *Journal of Statistical Software*, *32*(2), 1–9.

Rothstein, J. M. (2004). College performance predictions and the SAT. *Journal of Econometrics*, *121*(1–2), 297–317.

Samejima, F. (1969). *Estimation of latent ability using a pattern of graded scores* (Psychometric Monograph nr 17). Richmond, VA: Psychometric Society.

Schuler, H. Funke U. i Baron-Boldt, J. (1990). Predictive validity of school grades – a meta-analysis. *Applied Psychology: In International Review*, *39*(1), 83–103.

Skorupiński, P. M. (2013). Modele trafności pomiaru. W: M. Karwowski (red.), *Ścieżki rozwoju edukacyjnego młodzieży – szkoły pogimnazjalne. Trafność wskaźników edukacyjnej wartości dodanej dla szkół maturalnych* (13–26). Warszawa: Wydawnictwo IFiS PAN.

Smith, M. L. (1991). Put to the test: the effects of external testing on teachers. *Educational Researcher*, *20*(5), 8–11.

Smith, M. L., Edelsky, C., Draper, K., Rottenberg, C. i Cherland, M. (1989). *The role of testing in elementary schools*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, University of California.

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

Stockford, S. M. (2009). *Meta-analysis of intraclass correlation coefficients from multilevel models of educational achievement* [Rozprawa doktorska]. Phoenix, AZ: Arizona State University. Pobrano z http://search.proquest.com/docview/304845471

Szaleniec, H. (2006). Oszukiwanie na egzaminie istotnym źródłem majowej porażki. W: B. Niemierko i M. K. Szmigel (red.), *O wyższą jakość egzaminów szkolnych. Część I Etyka egzaminacyjna i zagadnienia ogólne*. Referat wygłoszony na XII Krajowej Konferencji Diagnostyki Edukacyjnej, Lublin, 9–11.10.2010 r. Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

Szaleniec, H. (2010). Czy egzaminy zewnętrzne wpływają na wewnątrzszkolne ocenianie i politykę edukacyjną? W: B. Niemierko i M. K. Szmigel (red.), *Teraźniejszość i przyszłość oceniania szkolnego*. Referat wygłoszony na XVI Krajowej Konferencji Diagnostyki Edukacyjnej, Toruń, 22–24.10.2010 r. Kraków: Polskie Towarzystwo Diagnostyki Edukacyjnej.

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.

Szyling, G. (2011). *Nauczycielskie praktyki oceniania poza standardami*. Kraków: Oficyna Wydawnicza Impuls.

Ustawa z dnia 7 września 1991 r. o systemie oświaty (Dz.U. 1991 nr 95, poz. 425 z późn. zm.). Pobrano z http://isap.sejm.gov.pl/DetailsServlet?id=WDU19910950425

Ustawa z dnia 6 grudnia 2013 r. o zmianie ustawy o systemie oświaty oraz niektórych innych ustaw (Dz.U. 2014, poz. 7). Pobrano z http://isap.sejm.gov.pl/DetailsServlet?id=WDU20140000007

Willingham, W. W., Lewis, C., Morgan, R. I. i Ramist, L. (1990). *Predicting College grades: an analysis of institutional trends over two decades*. Princeton, NJ: Educational Testing Service.

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

Wojciszke, B. (2001). Psychologia oceniania: mechanizmy, pułapki, środki zaradcze. W: A. Brzezińska i J. Brzeziński (red.), *Ewaluacja procesu kształcenia w szkole wyższej* (s. 211–237). Poznań: Wydawnictwo Fundacji Humaniora.

Young, J. W. (1990). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement*, *27*(2), 175–186.

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

Ziomek, R. L. i Svec, J. C. (1995). *High school grades and achievement: evidence of grade inflation.* (*ACT Research Report 1995-3*). Iowa City, IA: American College Testing.

Zwick, R., i Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement*, *48*(2), 101–121.

Zwick, R. i Mislevy, R. J. (2011). *Scaling and linking through-course summative assessments*. Princeton, NJ: Center for K-12 Assessment and Performance Management, Educational Testing Service.

Podziękowania

Autorzy artykułu składają podziękowania Bartoszowi Kondratkowi za udostępnienie autorskiego oprogramowania do analiz z wykorzystaniem modeli IRT.

Aneks

Tabela A1

Wielkości odchyleń standardowych dla poziomu ucznia (I) i oddziału szkolnego (II) (i ich 95-procentowe przedziały ufności) dla wykorzystywanych w analizie modeli: pustego, z oceną (o) i średnią ocen (sr) jako predyktorami

Rodzaj egzaminu	Rodzaj modelu	Odchylenie standardowe (poziom I)			Odchylenie standardowe (poziom II)		
		Wartość	95% przedział ufności		Wartość	95% przedział ufności	
MAT 2012	pusty	5,36	5,18	5,55	2,94	2,45	3,53
	z predyktorem (o)	3,86	3,72	3,99	2,20	1,83	2,63
	z predyktorem (sr)	4,03	3,89	4,17	2,15	1,79	2,59
MAT 2013	pusty	5,65	5,45	5,85	3,92	3,30	4,66
	z predyktorem (o)	4,06	3,92	4,20	3,01	2,54	3,57
	z predyktorem (sr)	4,28	4,14	4,43	2,65	2,22	3,16
PRZYR 2012	pusty	4,06	3,92	4,20	2,01	1,67	2,43
	z predyktorem (o)	3,18	3,07	3,29	1,33	1,09	1,63
	z predyktorem (sr)	3,13	3,02	3,24	1,30	1,06	1,59
PRZYR 2013	pusty	4,09	3,95	4,23	2,45	2,05	2,93
	z predyktorem (o)	3,25	3,14	3,37	1,74	1,45	2,08
	z predyktorem (sr)	3,18	3,08	3,30	1,70	1,41	2,04
POL 2012	pusty	5,36	5,17	5,55	3,01	2,51	3,61
	z predyktorem (o)	4,06	3,92	4,20	2,33	1,95	2,79
	z predyktorem (sr)	4,00	3,86	4,14	1,98	1,64	2,39
POL 2013	pusty	5,50	5,31	5,69	2,85	2,36	3,42
	z predyktorem (o)	4,10	3,97	4,25	2,09	1,73	2,52
	z predyktorem (sr)	4,02	3,89	4,17	1,91	1,58	2,31
HIS 2012	pusty	5,61	5,42	5,81	2,72	2,25	3,28
	z predyktorem (o)	4,37	4,22	4,52	2,46	2,05	2,95
	z predyktorem (sr)	4,18	4,04	4,33	1,72	1,40	2,11
HIS 2013	pusty	4,89	4,72	5,06	2,53	2,11	3,04
	z predyktorem (o)	4,01	3,87	4,15	2,06	1,72	2,48
	z predyktorem (sr)	3,92	3,79	4,06	1,63	1,34	1,99