Filip Kulon
Mateusz Żółtak
Instytut Badań Edukacyjnych

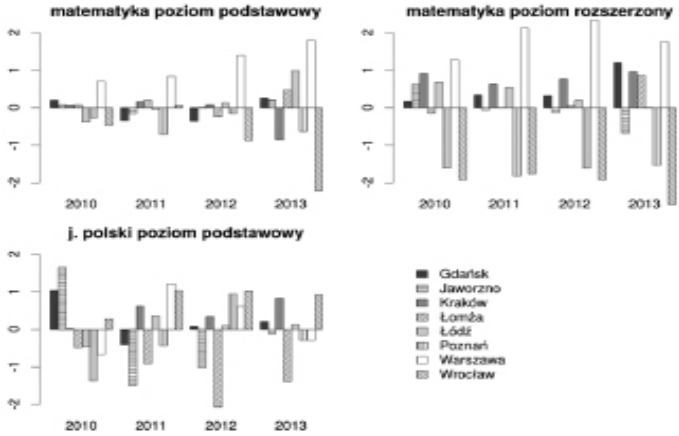
Zróżnicowanie łagodności egzaminatorów między okręgowymi komisjami egzaminacyjnymi

Wstęp

Łagodność (bądź surowość) oceniania można zdefiniować jako systematyczne przesunięcie oceny przyznawanej przez egzaminatora, niezwiązane z merytorycznym poziomem ocenianej odpowiedzi. W systemie egzaminacyjnym powinno dążyć się do tego, aby zróżnicowanie łagodności egzaminatorów było jak najmniejsze. Próbuje się to osiągnąć, przygotowując jak najbardziej precyzyjne i jednoznaczne schematy oceny oraz uniformizując proces szkolenia egzaminatorów. W praktyce zmuszeni jesteśmy jednak zaakceptować istnienie pewnego zróżnicowania łagodności pomiędzy egzaminatorami, choćby ze względu na zróżnicowanie osobowościowych cech egzaminatorów (Dubiecka, Szaleniec i Węziak, 2006). Wpływu osobowościowych, psychologicznych cech oceniających nie jesteśmy w stanie wyeliminować z procesu oceniania, sprawiedliwość oceny wymaga jednak, aby przynajmniej nie istniał związek pomiędzy cechami egzaminowanego, a łagodnością egzaminatora, który ocenia jego pracę. Ma to szczególne znaczenie w przypadku, gdy od wyników zależą dalsze losy edukacyjne zdających egzamin, zwłaszcza podczas matury, której wyniki decydują o przyjęciu na studia. Niespełnieniem tego wymagania mogłoby być np. istnienie związku pomiędzy łagodnością egzaminatora, do którego trafi praca, a miejscem pisania egzaminu przez ucznia (determinującym okręgową komisję egzaminacyjną, w której sprawdzana będzie jego praca). Biorąc pod uwagę występowanie dla niektórych OKE systematycznych różnic średnich wyników egzaminacyjnych (rys. 1), zasadne wydaje się pytanie, czy na gruncie dostępnych danych jesteśmy w stanie odrzucić hipotezę o istnieniu takiej zależności.

Odpowiedź na nie stała się możliwa dzięki przeprowadzonemu w 2013 r. przez Instytut Badań Edukacyjnych „Badaniu porównywalności oceniania i efektu egzaminatora dla egzaminu maturalnego z języka polskiego i matematyki", które objęło ogólnopolską losową próbę egzaminatorów. Poprzednie badania nad efektem egzaminatora prowadzone w Polsce albo stosowały modele statystyczne, które nie dawały informacji na temat łagodności egzaminatorów (Dolata, Putkiewicz i Wiłkomirska, 2004), albo przeprowadzane były w ramach tylko jednej OKE (Dubiecka, Szaleniec i Węziak, 2006). Warto jednak nadmienić, że w badaniu Dubieckiej, Szaleńca i Węziak analizowano występowanie różnic w łagodności egzaminatorów oceniających sprawdzian szóstoklasisty pomiędzy ośrodkami koordynacji oceniania w ramach OKE Kraków (w ośrodku takim pracowało od 4 do 7 zespołów oceniających, po kilkunastu egzaminatorów w każdym zespole; w OKE Kraków istniało kilkadziesiąt ośrodków

koordynacji oceniania[1]). Wyniki wskazywały na istotne na poziomie 0,012 zróżnicowanie łagodności egzaminatorów pomiędzy ośrodkami koordynacji oceniania, a przypisanie egzaminatora do ośrodka wyjaśniało 7,15% całkowitej wariancji łagodności egzaminatorów w badaniu. Choć nie sposób na tej podstawie przewidywać wyników na poziomie OKE i/lub dla innych poziomów egzaminu, widać jednak, że założenie o losowym ze względu na łagodność egzaminatora przydziale prac do oceniających może nie być spełniane.



Rysunek 1. Różnice w średnich wynikach egzaminu maturalnego (j. polski i matematyka) pomiędzy OKE w latach 2010-2013 względem średniego wyniku ogólnopolskiego (wyrażone w punktach na egzaminie)

Dane i model efektu egzaminatora

Dane

W analizach oparto się na danych z przeprowadzonego w 2013 r. przez Instytut Badań Edukacyjnych badania „Porównywalności oceniania i efektu egzaminatora". Objęło ono losową próbę egzaminatorów oceniających prace maturalne z języka polskiego i matematyki w sesji egzaminacyjnej w 2013 roku. W każdej OKE wylosowanych zostało 29 egzaminatorów oceniających prace maturalne z języka polskiego oraz 29 oceniających prace z matematyki. Oceniali oni, w warunkach symulujących normalną sesję oceniania, prace maturalne z terminu majowego z lat 2010 i 2011 (por. tab. 1). W przypadku matematyki oceniane były wszystkie zadania otwarte, a w przypadku języka polskiego ocenie poddano jedynie wypracowanie. Prace dobrane zostały w sposób celowy, tak by możliwie równomiernie były reprezentowane prace z zakresu całej skali wyników uzyskanych na egzaminie, pominięto jednak prace puste (w których uczeń nie udzielił odpowiedzi na żadne z zadań otwartych w przypadku matematyki oraz nie podjął próby napisania wypracowania w przypadku języka polskiego). Ponieważ przy tak dużej liczbie prac i egzaminatorów niemożliwe było dokonanie oceny każdej pracy przez wszystkich egzaminatorów, zastosowano dość złożony schemat przydziału prac do oceniających, który pozwolił

[1] W chwili obecnej koordynacja oceniania jest jednostopniowa na poziomie OKE.

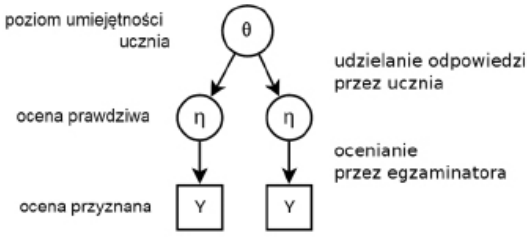
na powiązanie całej puli prac wykorzystywanych w badaniach z wszystkimi egzaminatorami (Szaleniec i inni, 2014). Z punktu widzenia opisanych analiz istotne jest, że każda praca z języka polskiego oceniana była przez jednego egzaminatora z każdej z OKE, natomiast każda praca z matematyki oceniana była przez co najmniej czterech egzaminatorów z różnych OKE oraz że każdy z egzaminatorów oceniał od 29 do 31 prac. Taki schemat badawczy pozwolił na oszacowanie każdego z parametrów zastosowanego modelu analizy na skali wspólnej dla wszystkich egzaminatorów.

Tabela 1. Liczba prac wykorzystanych w badaniu efektu egzaminatora

	2011		2012	
	p. podstawowy	p. rozszerzony	p. podstawowy	p. rozszerzony
j. polski	442	0	455	0
matematyka	442	450	455	447

Model i estymacja

Do oszacowania parametrów egzaminatorów użyto modelu *Hierarchical Rater Model with Signal Detection Theory* (HRM-SDT) (DeCarlo, Kim, & Johnson, 2011). Jest to model dwupoziomowy – pierwszy poziom odnosi się do wyniku oceny egzaminatora, zaś drugi do odpowiedzi na zadanie przez ucznia (por. rys. 2).



Rysunek 2. Model efektu egzaminatora HRM-SDT

Na poziomie oceniania przez egzaminatora HRM-SDT wykorzystuje model odpowiedzi stopniowej IRT – Graded Response Model (GRM), oparty o rozkłady logistyczne lub normalne. W opisywanych tu analizach zdecydowano się na wykorzystanie rozkładów logistycznych, a więc pierwszy poziom modelu opisać można wzorem:

$$P(Y_{nkr} \leq j|\eta_{nk} = \eta) = \frac{1}{1+\exp\left(-d_{kr}(\eta_{nk}-1-c_{kjr})\right)} \quad (1)$$

gdzie:

$Y_{nkr} \leq j$ – ocena przyznana za zadanie k w pracy ucznia n przez egzaminatora r mniejsza lub równa j (j – poziom wykonania zadania),

$\eta_{nk} = \eta$ – ocena prawdziwa[2] za rozwiązanie w zadaniu k pracy ucznia n równa η (η – poziom wykonania zadania),

c_{kjr} – subiektywny próg trudności poziomu wykonania j w zadaniu k dla egzaminatora r,

[2] Ocena pozbawiona wpływu efektu egzaminatora.

d_{kr} – parametr dyskryminacji egzaminatora *r* w zadaniu *k*.

Na drugim poziomie modelu oceny prawdziwe η poszczególnych zadań w danej pracy zależą od poziomu umiejętności ucznia piszącego pracę. W HRM-SDT także ta zależność modelowana jest za pomocą IRT – w opisywanych analizach na tym poziomie modelu również zastosowano model GRM:

$$P(\eta_{nk} \leq j|\theta_n = \theta) = \frac{1}{1+\exp\left(-a_k(\theta_n - b_{kj})\right)} \quad (2)$$

gdzie:

$\eta_{nk} \leq j$ – ocena prawdziwa za zadanie *k* w pracy ucznia *n* mniejsza lub równa *j* (*j* – poziom wykonania zadania),

$\theta_n = \theta$ – poziom umiejętności ucznia *n* równy θ,

a_k – parametr dyskryminacji zadania *k*,

b_{kj} – parametr trudności poziomu wykonania *j* w zadaniu *k*.

Na wykorzystanie modelu HRM-SDT zdecydowano się z uwagi na to, że pozwala on w analizie na:

- odseparowanie poziomu udzielania przez ucznia odpowiedzi na zadanie od przydzielania oceny przez egzaminatora,
- oddzielenie wpływu rzetelności oceniania od efektu łagodności,
- szacowanie parametrów dla interakcji zadania i egzaminatora.

Ostatnia z ww. cech ma szczególne znaczenie z uwagi na duże zróżnicowanie długości skal, na których oceniane są zadania maturalne (od 3 do 7 poziomów wykonania dla matematyki i od 4 do aż 26 poziomów wykonania dla języka polskiego). Sugeruje ono występowanie zróżnicowanych, w zależności od zadania, wartości parametrów dyskryminacji egzaminatorów. Zastosowanie prostszych modeli, w których parametr ten szacowany był jedynie na poziomie egzaminatora, a nie w podziale na zadania (por. Dubiecka, Szaleniec i Węziak, 2006), byłoby nadmiernym uproszczeniem.

Z uwagi na stopień komplikacji modelu został on oszacowany bayesowsko, z użyciem metody MCMC (ang. *Markov Chain Monte Carl*o) w programie JAGS. Zastosowanie tej metody oszacowania umożliwiło łatwe, a jednocześnie eleganckie formalnie obliczanie błędów standardowych uzyskanych parametrów modelu oraz ich agregatów.

Zagregowany wskaźnik łagodności egzaminatorów

Przedstawiony model dostarcza informacji o łagodności egzaminatorów na bardzo szczegółowym poziomie – interakcji danego egzaminatora z danym poziomem wykonania danego zadania. Na potrzeby porównań pomiędzy OKE niezbędne jest zagregowanie tych informacji. Pewną niedogodnością jest nieintuicyjna skala estymowanych w modelu parametrów łagodności egzaminatorów. Jest to skala ciągła, nieposiadająca minimum ani maksimum, opisująca subiektywne położenie progów kolejnych poziomów wykonania dla danego egzaminatora. W szczególności nie da się jej w prosty sposób odnieść

do punktacji zadania używanej na egzaminie. Aby w wygodny sposób interpretować wyniki, korzystne byłoby przeliczenie jej na bardziej intuicyjną skalę, np. punktów egzaminacyjnych. Ostateczną formę wskaźników łagodności egzaminatorów na poziomie OKE osiągnięto poprzez:

1. wprowadzenie wskaźnika łagodności egzaminatora w odniesieniu do pojedynczego zadania wyrażonego w punktacji egzaminu;
2. zagregowanie łagodności egzaminatorów do poziomu arkuszy egzaminacyjnych przez proste zsumowanie dla każdego egzaminatora wartości wskaźników łagodności wszystkich zadań w danym arkuszu egzaminacyjnym;
3. obliczenie średniej z wyliczonych w punkcie 2. wartości w podziale na OKE.

Procedura ta była powtarzana niezależnie dla każdej iteracji łańcucha Markowa. Końcowe oszacowania punktowe wskaźników uzyskano jako średnią z wartości obliczonych dla wszystkich iteracji łańcucha. Z kolei odchylenie standardowe wartości pomiędzy iteracjami łańcucha posłużyło do obliczenia przedziałów ufności dla uzyskanych wskaźników.

Wskaźnik łagodności egzaminatora w odniesieniu do pojedynczego zadania wyrażony w punktacji egzaminu zdefiniowano jako różnicę wartości oczekiwanych rozkładu punktacji ocen prawdziwych oraz ocen przydzielonych przez danego egzaminatora:

$$l_{kr} = E(pkt\ (Y_{kr}\ )) - E(pkt(\eta_k)) \tag{3}$$

gdzie:

l_{kr} – wartość wskaźnika łagodności egzaminatora,

$pkt(x)$ – funkcja przekształcająca poziom wykonania zadania na liczbę punktów na egzaminie[3],

Y_{kr} – ocena przyznana przez egzaminatora r za rozwiązanie zadania k,

η_k – ocena prawdziwa za rozwiązanie zadania k (estymowana w modelu).

Po podstawieniu Y i η z zastosowanego modelu efektu egzaminatora wzór (3) przyjmuje postać:

$$l_{kr} = \left(\sum_{\eta} P(\eta_k = \eta)\left(pkt(\eta) - \left(\sum_j f(Y_{kr} = j|\eta_k = \eta)pkt(j)\right)\right)\right) \tag{4}$$

gdzie[4]:

$f(j \mid \eta)$ – funkcja charakterystyczna modelu efektu egzaminatora:

$$f(j|\eta) = \frac{1}{1+\exp\left(-d(\eta-1-c_j)\right)} \tag{5}$$

$P(\eta_k = \eta)$ – przyjęty rozkład ocen prawdziwych w populacji.

Obliczenie wartości wskaźnika wymaga przyjęcia rozkładu ocen prawdziwych w populacji. Ponieważ próba prac użyta w badaniu była próbą celową,

[3] W wypadku niektórych zadań z języka polskiego liczba poziomów wykonania zadania pomniejszona o jeden nie odpowiada liczbie punktów możliwych do zdobycia, np. kryterium styl posiada 4 poziomy wykonania, którym odpowiada 0, 1, 3 lub 5 punktów.

[4] Pominięto oznaczenia wprowadzone już we wzorze (2).

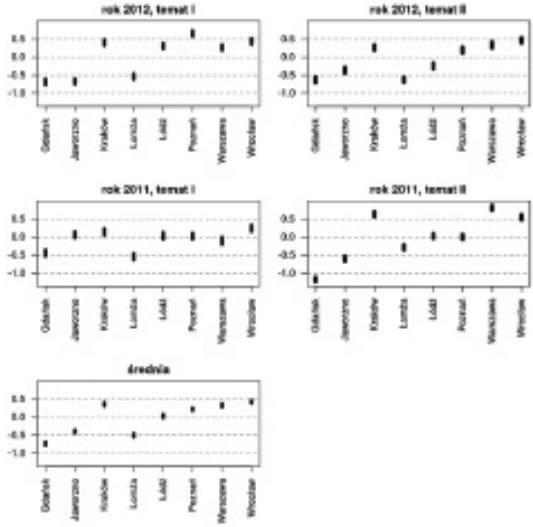
zdecydowano się na estymowanie rozkładu ocen prawdziwych w populacji poprzez przemnożenie rozkładu łącznego ocen obserwowanych oraz oszacowanych w modelu ocen prawdziwych przez rozkład ocen obserwowanych w populacji (wyliczony na podstawie ogólnopolskich wyników egzaminacyjnych):

$$P_{pop}(\eta_k = \eta, Y_k = y) = P_{próba}(\eta_k = \eta, Y_k = y)P_{pop}(Y_{pop} = y) \qquad (6)$$

Wyniki analiz dla języka polskiego

Uzyskane wyniki wskazują na istotne statystycznie[5] różnice średniej łagodności egzaminatorów pomiędzy niektórymi okręgowymi komisjami egzaminacyjnymi (por. rys. 3). Co więcej, dla niektórych OKE różnice te są stałe niezależnie od roku egzaminu i tematu wypracowania, np. OKE Gdańsk i OKE Łomża są najbardziej surowymi komisjami niezależnie od roku i tematu wypracowania, a OKE Kraków oraz OKE Wrocław plasują się niezmiennie wśród najbardziej łagodnych.

Wartości bezwzględne tych różnic mogą nie wydawać się duże. Dla średniej z wszystkich lat i tematów uwzględnionych w badaniu różnica pomiędzy skrajnie surową i skrajnie łagodną OKE wynosi 1,15 punktu na egzaminie. Należy jednak pamiętać, że dotyczą one różnic średniego wyniku w całej OKE. Różnica 1,15 punktu oznacza więc, że średnio każdy uczeń w bardziej łagodnej OKE dostał o 1,15 punktu więcej od ucznia w bardziej surowej OKE (za pracę o takiej samej ocenie prawdziwej). W wypadku wyniku egzaminacyjnego mieszczącego się w środkowej części skali[6] oznacza to przesunięcie ucznia w rozkładzie populacyjnym o ok. 3 centyle, co z punktu widzenia zdającego maturę należy uznać za istotną różnicę.



Rysunek 3. 99% przedziały ufności średniej łagodności egzaminatorów w podziale na OKE – język polski (wyrażone w punktach na egzaminie)

[5] Rozłączność 99% przedziału ufności.

[6] Od 31 do 45 punktów – w zakresie tym mieszczą się wyniki ok. połowy wszystkich uczniów.

Interesujące jest również przyjrzenie się zagadnieniu od strony OKE i postawienie pytania, na ile różnice średniej łagodności egzaminatorów pomiędzy OKE (por. rys.3) tłumaczą różnice średnich wyników (por. rys. 1). W tym celu użyto wskaźnika R^2 prostego modelu liniowego, gdzie zmienną zależną był średni wynik uczniów w OKE, a zmienną niezależną średnia wartość parametru łagodności egzaminatorów w OKE.

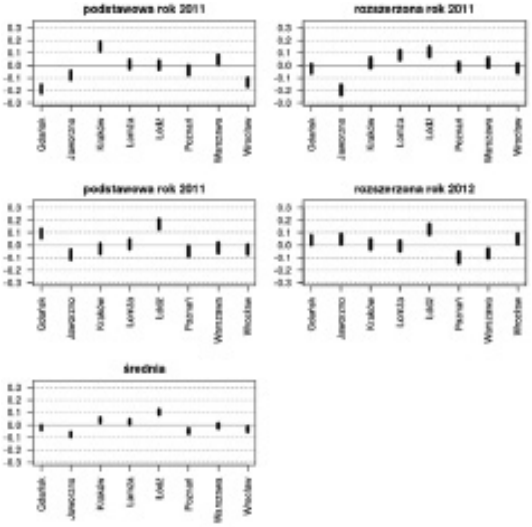
Jeśli analizę przeprowadzić łącznie dla obydwu uwzględnionych w badaniu tematów wypracowania w obydwu latach, różnice średniej łagodności egzaminatorów tłumaczą 10,8% wariancji średnich wyników egzaminacyjnych uczniów pomiędzy OKE, zależność jest więc wyraźna, ale jej siła ograniczona. Jeśli jednak powtórzyć obliczenia w rozbiciu na tematy wypracowań, okazuje się, że dla tematu I (odwołującego się do analizy poezji lub dramatu, wybieranego przez ok. ¼ uczniów) wyjaśniane jest jedynie 4,5% wariancji, natomiast dla tematu II (odwołującego się do prozy, wybieranego przez ok. ¾ uczniów, w obiegowej opinii uważanego za łatwiejszy) aż 44%. Wyniki te wskazują, że w wypadku wypracowania na temat II w analizowanych latach udział zróżnicowania średniej łagodności egzaminatorów pomiędzy OKE w różnicach średnich wyników egzaminu pomiędzy OKE jest bardzo duży. Z punktu widzenia systemu egzaminacyjnego jest to sytuacja wysoce niepożądana, istotnie zakłóca ona bowiem analizę wyników egzaminacyjnych na poziomie makroskopowym (np. porównania między regionami kraju), jak również oznacza istnienie systematycznych różnic w sposobie oceniania uczniów ze względu na miejsce, w którym podchodzą do egzaminu maturalnego z języka polskiego.

Wyniki analizy dla matematyki

Zarówno na poziomie poszczególnych arkuszy egzaminacyjnych, jak i na poziomie całego badania odnotowano istotne statystycznie różnice w łagodności oceniania pomiędzy OKE (por. rys. 4). Odmiennie jednak niż w języku polskim, gdzie różnice między OKE były systematyczne między tematami wypracowań i latami, uporządkowanie OKE ze względu na łagodność oceniania egzaminu z matematyki zmienia się w zależności od roku i poziomu egzaminu. O ile więc dla języka polskiego można było mówić o różnicach łagodności pomiędzy OKE, o tyle w wypadku matematyki ograniczyć się trzeba co najwyżej do stwierdzenia, że dana OKE ma tendencję do bycia bardziej łagodną lub surową niż inne komisje.

Brak systematycznych różnic łagodności pomiędzy OKE jest zjawiskiem pożądanym, ważne są jednak także wielkości występujących tendencji oraz wartości bezwzględne różnic dla poszczególnych arkuszy. Siła obserwowanych tendencji bez podziału na poszczególne lata i poziomy egzaminów waha się od -0,08 (OKE Jaworzno) do 0,10 punktu (OKE Łódź) z odchyleniem standardowym wynoszącym 0,05 punktu. Rozstęp skrajnych wartości wynoszący 0,18 punktu jest ponad sześciokrotnie mniejszy niż w wypadku języka polskiego. Zestawiając te wartości ze zróżnicowaniem łagodności między egzaminatorami w ramach OKE (rozstęp międzyćwiartkowy od 0,29 do 0,75 punktu na egzaminie, różnice pomiędzy skrajnymi egzaminatorami od 1,19 do 2,22 punktu, w zależności od OKE, roku i poziomu egzaminu) oraz uwzględniając,

że ocena każdej pracy obarczona jest jeszcze błędem losowym, wynikającym z niedoskonałej rzetelności oceniania, wydaje się, że wpływ tendencji łagodności egzaminatorów pomiędzy OKE na całkowity błąd egzaminatora, jakim obarczona jest ocena pojedynczej pracy z matematyki, można uznać za pozbawiony praktycznego znaczenia. Także bezwzględne wartości tych różnic (0,18 punktu między skrajnymi OKE, odchylenie standardowe średniej łagodności OKE równe 0,05 punktu) wydają się pomijalne z punktu widzenia pojedynczego ucznia. Podsumowując, wpływ różnic w tendencji łagodności egzaminatorów z matematyki pomiędzy OKE na ocenę pojedynczej pracy, choć istotny statystycznie, z praktycznego punktu widzenia uznać można za zaniedbywalny.



Rysunek 4. 99% przedziały ufności średniej łagodności egzaminatorów w podziale na OKE – matematyka (wyrażone w punktach na egzaminie)

Biorąc pod uwagę, że zróżnicowanie OKE ze względu na łagodność różni się w zależności od roku i poziomu egzaminu, warto przeanalizować go także w rozbiciu na poszczególne arkusze egzaminacyjne. Wyniki na tym poziomie wykazują większe wartości różnic – rozstęp łagodności pomiędzy najbardziej łagodną i surową OKE waha się w zależności od roku i poziomu egzaminu od 0,23 do 0,42 punktu, a odchylenie standardowe od 0,08 do 0,13 punktu. Wartości te, choć wyższe, nadal pozostają niewielkie.

Warto odnotować, że różnice średniej łagodności egzaminatorów pomiędzy OKE wyjaśniają jedynie 1,7% wariancji średnich wyników uczniów pomiędzy OKE dla poziomu podstawowego egzaminu[7] i 0,03% dla poziomu rozszerzonego egzaminu (porównaj różnice średnich wyników uczniów pomiędzy OKE na rys. 1). Wyniki te utwierdzają w przekonaniu, że wpływ systematycznych różnic łagodności egzaminatorów pomiędzy OKE na wyniki egzaminu z matematyki można uznać za pozbawiony praktycznego znaczenia.

[7] Wyniki egzaminacyjne z uwzględnieniem zadań zamkniętych.

Wnioski

Przeprowadzone analizy wskazują na:

- pozbawiony praktycznego znaczenia (zarówno ze względu na niewielką bezwzględną wielkość efektu, jak i brak jego stałości w czasie) związek pomiędzy średnią łagodnością egzaminatorów w poszczególnych OKE a średnimi wynikami egzaminu maturalnego z matematyki;
- istotny (zarówno na poziomie ucznia, jak i porównań w skali makro, np. między OKE lub województwami) związek pomiędzy średnią łagodnością egzaminatorów w poszczególnych OKE a średnimi wynikami części podstawowej egzaminu maturalnego z języka polskiego.

W szczególności za niepokojące należy uznać wyjaśnianie przez różnice w średniej łagodności egzaminatorów na poziomie OKE aż 44% wariancji średnich wyników uczniów między OKE dla wypracowania z języka polskiego w tematach odnoszących się do prozy (temat II w analizowanych latach). Jako sposoby na wyeliminowanie tego zjawiska wskazać można:

- Losowy przydział prac do egzaminatorów, nieuwzględniający podziału na OKE. Przy tradycyjnej organizacji oceniania, gdy egzaminator otrzymuje do sprawdzenia papierową pracę, byłoby to rozwiązanie niezwykle karkołomne logistycznie, jednak w wypadku e-oceniania jego wdrożenie byłoby bardzo łatwe (i praktycznie bezkosztowe).
- Położenie większego nacisku na uniformizację szkolenia egzaminatorów pomiędzy OKE oraz stała troska o przygotowywanie jak najbardziej jednoznacznych schematów oceniania.

Bibliografia

1. DeCarlo, L. T., Kim, Y., Johnson, M. S. (2011), A Hierarchical Rater Model for Constructed Responses, with a Signal Detection Rater Model. *Journal of Educational Measurement, 48*(3).
2. Dolata, R., Putkiewicz, E., Wiłkomirska, A. (2004), *Reforma egzaminu maturalnego – oceny i rekomendacje,* Instytut Spraw Publicznych, Warszawa..
3. Dubiecka, A., Szaleniec, H., Węziak, D. (2006), *Efekt egzaminatora* w *egzaminach zewnętrznych* [w:] Niemierko B., Szmigel M.K. (red.), O *wyższą jakość egzaminów szkolnych,* gRUPA TOMAMI, Lublin.
4. Szaleniec, H. i in. (2014), *Raport badania porównywalności oceniania* i *efektu egzaminatora*, Instytut Badań Edukacyjnych (w druku).