



KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

IBE



*entuzjaści
edukacji*

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY



Henryk Szaleniec
Bartosz Kondratek
Filip Kulon
Artur Pokropek
Paulina Skórska
Karolina Świst
Tymoteusz Wołodźko
Mateusz Żółtak



Efekt egzaminatora w ocenianiu prac maturalnych z języka polskiego i matematyki

Praca zbiorowa



KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

IBE



entuzjaści
edukacji

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY



Henryk Szaleniec
Bartosz Kondratak
Filip Kulon
Artur Pokropek
Paulina Skórska
Karolina Świst
Tymoteusz Wołodźko
Mateusz Żółtak

Efekt egzaminatora w ocenianiu prac maturalnych z języka polskiego i matematyki

Praca zbiorowa

Autorzy:

Henryk Szaleniec

Bartosz Kondratek

Filip Kulon

Artur Pokropek

Paulina Skórska

Karolina Świst

Tymoteusz Wołodźko

Mateusz Żółtak

Recenzenci:

prof. dr hab. Roman Konarski, Uniwersytet Gdański

prof. dr hab. Jerzy Biniewicz, Instytut Dziennikarstwa i Komunikacji Społecznej, Uniwersytet Wrocławski

Redakcja językowa:

Beata Dąbrowska

Korekta, skład, łamanie, druk:

Business Point Sp. z o.o.

ul. Erazma Ciołka 11A/302

01-402 Warszawa

Tel. +48 22 188 18 72

biuro@businesspoint.pl

www.businesspoint.pl

Wydawca:

Instytut Badań Edukacyjnych

ul. Górczewska 8

01-180 Warszawa

tel. (22) 241 71 00; www.ibe.edu.pl

© Copyright by: Instytut Badań Edukacyjnych, Warszawa 2015

ISBN: 978-83-61693-95-6

Wzór cytowania:

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

Publikacja współfinansowana przez Unię Europejską ze środków Europejskiego Funduszu Społecznego w ramach projektu:

Badanie jakości i efektywności edukacji oraz instytucjonalizacja zaplecza badawczego

Spis treści

Streszczenie.....	9
Summary.....	11
Rekomendacje	13
Podziękowania.....	17
1. Wprowadzenie	18
2. Charakterystyka egzaminu maturalnego z języka polskiego	21
3. Charakterystyka egzaminu maturalnego z matematyki.....	23
4. Organizacja oceniania w polskim systemie egzaminów zewnętrznych	25
4.1. Specyfika oceniania zadań otwartych rozszerzonej odpowiedzi z języka polskiego	26
4.2. Specyfika oceniania zadań otwartych rozszerzonej odpowiedzi z matematyki.....	26
4.3. Egzaminator podczas procesu oceniania	27
5. Efekt egzaminatora – podstawowe pojęcia	29
5.1. Badania prowadzone w Polsce.....	30
5.2. Monitorowanie oceniania w skali kraju przez CKE w latach 2006–2009.....	32
5.3. Monitorowanie efektu egzaminatora w procesie e-oceniania.....	34
6. Metodologia i sposób realizacji badania.....	35
6.1. Cel ogólny i cele szczegółowe.....	35
6.2. Problemy, pytania badawcze.....	35
6.3. Specyfikacja zmiennych zależnych i niezależnych.....	35
6.4. Schemat planu badawczego	36
6.5. Dobór próby egzaminatorów i prac do oceniania.....	37
6.5.1. Dobór próby egzaminatorów	37
6.5.2. Dobór prac do oceniania	37
6.6. Opis modelu linkowania prac i egzaminatorów wewnątrz OKE i pomiędzy OKE.....	39
6.6.1. Schemat rozdziału prac pomiędzy egzaminatorów uczestniczących w badaniach	39
7. Statystyczna koncepcja analizy efektu egzaminatora	45
7.1. Modelowanie kryterium oceny <i>treść</i> w języku polskim.....	45
7.2. Interpretacja parametrów modelu.....	46
7.3. Wskaźniki siły efektu egzaminatora	47
7.3.1. Standaryzowana dyskryminacja egzaminatora.....	48
7.3.2. Odsetek poprawnych klasyfikacji.....	48
7.3.3. Łagodność.....	48
7.3.4. Zniekształcenie skali.....	49
7.3.5. Rozkład ocen „prawdziwych” w populacji.....	49
7.4. Walidacja ocen „prawdziwych” z modelu – ocenianie ekspertów	50

8. Realizacja badań 53

8.1. Realizacja zadań wynikających z przyjętego schematu badania	53
8.1.1. Osoby oceniające (punktujące) zadania w pracach egzaminacyjnych	53
8.2. Opis koordynacji i monitorowania badania	54
8.3. Opis przetwarzania i weryfikowania danych	55

9. Efekt egzaminatora w ocenianiu maturalnych prac pisemnych z języka polskiego i matematyki 57

9.1. Język polski	57
9.1.1. Treść	58
9.1.2. Kompozycja, styl, język, zapis	59
9.1.3. Walory	60
9.1.4. Nauczyciele niebędący egzaminatorami oraz studenci	60
9.1.5. Różnice łagodności na poziomie testów	62
9.1.6. Różnice na poziomie okręgowych komisji egzaminacyjnych	63
9.1.7. Błędy kardynalne	64
9.2. Matematyka	66
9.2.1. Nauczyciele niebędący egzaminatorami	69
9.2.2. Różnice łagodności na poziomie testów	70
9.2.3. Różnice na poziomie okręgowych komisji egzaminacyjnych	71
9.3. Pismo odręczne vs zapis komputerowy a ocena za wypracowanie	72
9.3.1. Dobór prac do badań	73
9.3.2. Wyniki badań – efekt formy zapisu wypracowania (pismo odręczne vs zapis komputerowy)	73
9.4. Wynik oceniania a czytelność pisma	77
9.4.1. Oceny charakteru pisma	78
9.4.2. Porównanie ocen prac różniących się charakterem pisma	80
9.4.3. Podsumowanie	84
9.5. Kolejność i czas oceniania pracy egzaminacyjnej a wynik oceny	84
9.6. Próg zdania matury a rozkład wyników z egzaminu maturalnego	87
9.6.1. Miary efektów związanych z progiem zaliczenia	87
9.6.2. Język polski	88
9.6.3. Matematyka	90
9.7. Analiza wyników badań ankietowych	92
9.7.1. Język polski	92
9.7.2. Matematyka	97
9.7.3. Łagodność i precyzja egzaminatorów	101

Bibliografia 105

Aneks 1. Parametry statystyczne (KTT i IRT) zadań matury z matematyki i języka polskiego 2011 i 2012 r. 109

Tabela 1. Matura z matematyki 2011 – charakterystyka zadań arkusza na poziomie podstawowym	109
Tabela 2. Matura z matematyki 2011 – charakterystyka zadań arkusza na poziomie rozszerzonym	110
Tabela 3. Matura z matematyki 2012 – charakterystyka zadań arkusza na poziomie podstawowym	112
Tabela 4. Matura z matematyki 2012 – charakterystyka zadań arkusza na poziomie rozszerzonym	113
Tabela 5. Matura z języka polskiego 2011 – charakterystyka zadań arkusza na poziomie podstawowym	115
Tabela 6. Matura z języka polskiego 2012 – charakterystyka zadań arkusza na poziomie podstawowym	106

Aneks 2. Wykresy rozkładów parametrów rkr (zniekształceń skali).....	119
Rysunek 1. Wykresy rozkładów parametrów rkr dla języka polskiego w podziale na tematy i kryteria oceny	119
Rysunek 2. Wykresy rozkładów parametrów rkr dla matematyki w podziale na zadania	120
Aneks 3. – Kwestionariusz ankiety egzaminatora	121

Spis ilustracji

Rysunek 4.1. Schemat koordynacji i zapewniania jakości oceniania.....	25
Rysunek 5.1. Model analizy rzetelności systemów punktacji.....	30
Rysunek 5.2. Średnie poziomy surowości zespołów egzaminatorów sprawdzianu w jednostkach logit przy odchyleniu standardowym 0,117	31
Rysunek 5.3. Rozstęp poziomu surowości egzaminatorów w poszczególnych zespołach.....	32
Rysunek 6.1. Schemat planu badawczego efektu egzaminatora.....	36
Rysunek 7.1. Hierarchiczny model efektu egzaminatora z użyciem teorii detekcji sygnału (HRM-SDT).....	43
Rysunek 7.2. Zgodność między ekspertami w poszczególnych kryteriach jako procent takich samych ocen.....	50
Rysunek 7.3. Zgodność między ekspertami i oceną „prawdziwą” z modelu HRM-SDT w poszczególnych kryteriach jako procent takich samych ocen.....	51
Rysunek 8.1. Schemat organizacji sesji oceniania w trakcie badań.....	54
Rysunek 9.1. Rozkłady parametrów ρ_{kr} oraz b_{kr} w podziale na kryterium oceny z języka polskiego (tylko egzaminatorzy).....	58
Rysunek 9.2. Rozkłady parametrów ρ_{kr} oraz b_{kr} w podziale na kryterium oceny z języka polskiego – egzaminatorzy (czarne linie) oraz studenci i nauczyciele niebędący egzaminatorami (przerwane linie oraz parametry rozkładów).....	61
Rysunek 9.3. Rozkłady parametrów b_{tr} (łagodność egzaminatorów wyrażona w punktach procentowych wyniku egzaminacyjnego zagregowana do poziomu testu) w podziale na rok i temat wypracowania z języka polskiego: rzeczywiste (linia ciągła) i w sytuacji braku związku między łagodnością egzaminatorów w poszczególnych kryteriach oceny (linia przerywana).....	62
Rysunek 9.4. 99% przedziały ufności łagodności egzaminatorów na poziomie roku, tematu i OKE dla języka polskiego (w punktach procentowych na egzaminie)	63
Rysunek 9.5. 99% przedziały ufności łagodności egzaminatorów na poziomie roku, tematu i OKE (wraz z grupą kontrolną nauczycieli oczekujących na nadanie statusu egzaminatora) dla matematyki (w punktach procentowych na egzaminie).....	70
Rysunek 9.6. Rozkłady parametrów b_{tr} (łagodność egzaminatorów wyrażona w punktach procentowych wyniku egzaminacyjnego zagregowana do poziomu testu) w podziale na rok i poziom matury z matematyki: rzeczywiste (linia ciągła) i w sytuacji braku związku między łagodnością egzaminatorów w poszczególnych kryteriach oceny (linia przerywana).....	71
Rysunek 9.7. Porównanie rozkładów ocen ostatecznych (oszacowanych na podstawie oceny kilku egzaminatorów) dla różnych sposobów ich ustalania.....	74
Rysunek 9.8. Porównanie rozkładów wyników uzyskanych przy ocenie wypracowania przez egzaminatorów w jego oryginalnym zapisie pismem odręcznym (PO) i po przepisaniu wypracowania na komputerze (PK), w rozbiciu na poszczególne kryteria oceny.....	76
Rysunek 9.9. Reprezentacja w próbie pięciu kategorii wielkości pisma z uwzględnieniem czytelności	79
Rysunek 9.10. Reprezentacja w próbie trzech kategorii wyrobienia pisma z uwzględnieniem czytelności.....	79
Rysunek 9.11. Oceny przyznane za poszczególne kryteria oceniania dla prac z poszczególnych kategorii jakości pisma odręcznego.....	80
Rysunek 9.12. Rozkłady wyników dla egzaminu maturalnego z języka polskiego na poziomie podstawowym w latach 2010–2014.....	89
Rysunek 9.13. Rozkłady wyników dla egzaminu maturalnego z matematyki na poziomie podstawowym w latach 2010–2014.....	91
Rysunek 9.14. Wiek (w latach) egzaminatorów z języka polskiego ($n = 232$).....	93
Rysunek 9.15. Płeć (wykres po lewej) i staż pracy w latach (wykres po prawej) egzaminatorów z języka polskiego ($n = 232$)	93
Rysunek 9.16. Opinie egzaminatorów z języka polskiego na temat tego, jak stresująca jest praca w charakterze nauczyciela i egzaminatora	94
Rysunek 9.17. Liczba majowych sesji maturalnych, w których brali udział egzaminatorzy z języka polskiego (procent).....	95
Rysunek 9.18. Wiek (w latach) egzaminatorów z matematyki ($n = 232$)	97
Rysunek 9.21. Liczba majowych sesji maturalnych, w których brali udział egzaminatorzy z języka polskiego (procent).....	100

Streszczenie

Badania efektu egzaminatora i porównywalności oceniania prowadzone przez Instytut Badań Edukacyjnych są spójne z szeregiem działań prowadzonych przez Centralną Komisję Egzaminacyjną zmierzających do zapewnienia rzetelności oceniania egzaminów zawierających zadania otwarte rozszerzonej odpowiedzi. Ocenę za rozwiązanie takich zadań z natury obciążone są efektem egzaminatora wynikającym z faktu, że wartościowanie dokonuje człowiek – z całym bagażem swojej wiedzy, doświadczenia, postaw, empatii i innych cech indywidualnych. Dlatego też zróżnicowanie wyników egzaminu zależy nie tylko od zróżnicowania poziomu sprawdzanych na egzaminie umiejętności, ale także od egzaminatora. W egzaminach z przedmiotów humanistycznych, w których zastosowane są zadania wymagające napisania wypracowania, niebagatelna część zmienności wyników jest generowana przez efekty związane z cechami egzaminatora (nazywane ogólnie w niniejszym opracowaniu efektem egzaminatora). W prezentowanych badaniach efekt ten jest analizowany dla dwóch głównych maturalnych przedmiotów egzaminacyjnych – języka polskiego i matematyki, które są w diametralnie różny sposób podatne na wpływ tego czynnika.

W poszczególnych rozdziałach niniejszego opracowania przedstawiono: dotychczasowe badania w Polsce w tej dziedzinie, główne założenia metodologiczne, specyfikację wykorzystanego do analiz hierarchicznego modelu oceniającego oraz rezultaty badań. Szczególną wartość stanowią poszukiwania rozwiązań mających prowadzić do minimalizacji wpływu efektu egzaminatora na wyniki zdających egzamin.

Do badania efektu egzaminatora wybrano dwa kolejne lata: 2011 i 2012. Ze względów praktycznych prace egzaminacyjne zostały wylosowane z dwóch okręgowych komisji egzaminacyjnych: Jaworzna i Krakowa. Istotne było zgromadzenie zbioru prac posiadających następujące cechy:

1. zróżnicowanie poziomu umiejętności badanych w zakresie całej skali zastosowanej w ocenie zadań;
2. dla języka polskiego – zróżnicowanie tematów wypracowań do analizy interakcji egzaminator ↔ temat wypracowania.

Badania zostały przeprowadzone na reprezentatywnej próbie oceniających prace maturalne, warstwowanej ze względu na okręgowe komisje egzaminacyjne (rozszerzonej o nauczycieli niebędących egzaminatorami, a także o studentów w przypadku egzaminu maturalnego z języka polskiego). W celu powiązania wszystkich prac i egzaminatorów z ośmiu OKE zastosowano złożony model, w wyniku którego każda z prac z języka polskiego była oceniana przez jednego egzaminatora z każdej OKE. W przypadku egzaminu maturalnego z matematyki, dla którego badano ocenianie prac z egzaminu na poziomie podstawowym i rozszerzonym, każda z prac oceniana była przez czterech egzaminatorów (każdy z innej OKE).

Ocenianie zostało przeprowadzone w warunkach możliwie najbardziej zbliżonych do tych, które mają miejsce podczas głównej (majowej) sesji oceniania egzaminu maturalnego. Zastosowano analogiczną strukturę koordynacji oceniania, zespoły ulokowano w tych samych miastach oraz użyto niezmienionych kryteriów i schematów oceniania.

Wyniki analiz

Badania wykazały istnienie istotnego wpływu egzaminatorów na wyniki oceniania zarówno dla matury z języka polskiego jak i matematyki. Przy tym w wypadku języka polskiego siła tego wpływu jest wyraźnie większa. Różnica łagodności pomiędzy 25% najbardziej łagodnych a 25% najbardziej surowych egzaminatorów wynosi (w zależności od roku i tematu wypracowania) od 3,1 do 3,7 punktu procentowego na egzaminie¹. Dla matematyki różnice te przyjmują wartości od 0,87 do 1,36 punktu procentowego wyniku egzaminu. Wynik taki nie jest zaskoczeniem. Ocenianie wypracowań (esejów) zawsze będzie obarczone pewną dozą subiektywizmu na co wskazuje wiele badań na świecie. Osiągnięcie pełnej zgodności oceniania jest utopią. Usilne dążenie do 100 procentowej zgodności, które pociąga za sobą rozszerzanie i usztywnianie oceniania kryterialnego może mieć negatywny wpływ na uczenie tak ważnej umiejętności jak pisanie własnych tekstów.

¹ Wyniki dla całego egzaminu – łącznie zadania krótkiej odpowiedzi i wypracowanie dla j. polskiego oraz zamknięte i otwarte dla matematyki.

Różnice w łagodności oceniania występują nie tylko na poziomie indywidualnym, ale utrzymują się również na poziomie okręgowych komisji egzaminacyjnych. W przypadku języka polskiego w niektórych OKE różnice te są systematyczne i występują we wszystkich badanych tematach wypracowań. Średnia różnica pomiędzy najbardziej surową i najbardziej łagodną komisją wyniosła 1,6 punktu procentowego wyniku egzaminu. Jeśli spojrzeć na różnice w wynikach egzaminacyjnych uczniów na poziomie OKE, to okazuje się, że efekt łagodności egzaminatorów wyjaśnia dużą część tego zróżnicowania (średnio dla wszystkich tematów wyjaśniane jest 11% wariancji różnicy wyników). Dla matematyki różnice na poziomie OKE również występują, lecz są mniejsze i nie mają systematycznego charakteru. Zamiast o różnicach, w tym przypadku można by mówić raczej o pewnych tendencjach łagodności egzaminatorów, którymi cechują się poszczególne OKE.

Innym wnioskiem płynącym z analizy efektu egzaminatora jest zaobserwowane zaburzenie rozkładu wyników, które występuje na progu zdawalności egzaminu. Ponieważ próg zdawalności służy certyfikacji, pożądane jest zwiększenie rzetelności oceny w jego okolicy. W polskich egzaminach próg zdawalności jest niestety źródłem efektu egzaminatora polegającego na lokalnym zwiększeniu łagodności oceniania tuż poniżej progu zaliczenia. Warto zaznaczyć, że nawet jeśli efekt lokalnego zwiększenia rzetelności oceniania ma pozytywny wpływ na oceny zdających, których prace znajdują się w okolicy progu zdawalności, to lokalny wzrost łagodności jest czynnikiem negatywnym w kontekście egzaminu jako całości. Powoduje on, że część prac zdających oceniana jest w sposób systematycznie różny od pozostałych. Efekt ten jest znacznie większy dla egzaminu maturalnego z języka polskiego niż z matematyki. Istnieje też statystycznie istotne zróżnicowanie łagodności egzaminatora w okolicy progu zdawalności pomiędzy okręgowymi komisjami egzaminacyjnymi.

Kolejnym zagadnieniem, które było przedmiotem analiz w prezentowanych badaniach, jest porównywalność oceniania wypracowań pisanych odręcznie i przepisanych komputerowo. Na podstawie analizy danych z badań porównawczych rezultatów oceniania wypracowań maturalnych w odręcznym, oryginalnym zapisie i ich kopii w zapisie komputerowym zaobserwowano, że prace pisane ręcznie oceniane są wyżej niż ich kopie zapisane na komputerze. Różnica średniej ocen dla wypracowania z egzaminu maturalnego ocenianego w skali pięćdziesięciopunktowej przekracza 2 punkty i jest statystycznie istotna na poziomie istotności $p = 0,01$. Rezultat ten jest zaskoczeniem dla praktyków, chociaż podobny wynik uzyskano w wielu badaniach na świecie porównujących jakość oceniania tekstów pisanych odręcznie i ich kopii wykonanych z wykorzystaniem procesora tekstu.

W prezentowanych badaniach przedmiotem analiz był także związek oceny za wypracowanie z cechami pisma autora. Wyniki sugerują, że prace pisane przez autorów, których pismo cechuje się małym wyrobieniem, oceniane są niżej w stosunku do prac pisanych wyrobionym lub średnio wyrobionym charakterem pisma. Różnica ta jest znaczna i statystycznie istotna. Związki czytelności pisma z oceną otrzymaną przez zdającego za wypracowanie są umiarkowane. Należy tu jednak zaznaczyć, że czytelność i wyrobienie są silnie skorelowanymi cechami pisma. Większość wypracowań mało czytelnych to również prace pisane mało wyrobionym charakterem pisma. Wyniki te są zgodne z rezultatami badań w innych krajach.

W przypadku wpływu zmiennych opisujących indywidualne cechy egzaminatorów (zarejestrowanych podczas badania ankietowego) na stopień precyzji i łagodności egzaminatorów zarówno w przypadku języka polskiego, jak i matematyki zaobserwowano wyłącznie związki specyficzne dla poszczególnych przedmiotów. Jedynie w przypadku łagodności egzaminatorów z języka polskiego można zaobserwować związek z indeksem nastawienia do pracy w zespołach podczas oceniania. Zaobserwowano różnice w poziomie motywacji wewnętrznej i nastawienia do pracy w zespołach w konkretnych OKE.

Sprawdzono także wpływ czasu oceniania prac oraz kolejności ich oceniania na uzyskaną liczbę punktów w zadaniach otwartych. W przypadku zarówno kolejności, jak i czasu oceniania efekty te, choć istotne statystycznie, są bliskie zeru (zarówno w przypadku prac z języka polskiego, jak i matematyki). Wyniki te sugerują, że zmęczenie egzaminatora nie zaburza procesu oceniania.

Summary

Studies on comparability of assessment conducted by the Educational Research Institute are consistent with a range of activities undertaken by the Central Examination Board (*Centralna Komisja Egzaminacyjna – CKE*) that are designed to ensure reliability of assessment of exam responses to constructed responses. Assessment of responses to such open questions is inherently burdened with the rater effect which results from the fact that assessment is conducted by humans with all their knowledge, experience, beliefs, empathy and other individual characteristics. Therefore, differences in exam results depend not only on the varying level of skills examined, but also on the rater. At humanities exams, where writing of an essay is required, a significant portion of variation in the exam results is generated by the effects associated with the personal characteristics of the rater (in this report, in general referred to as the rater effect). In the presented studies, the effect was analysed for two major Matura exam subjects – Polish language and mathematics, which are susceptible in a completely different way to these factors.

Respective chapters of this report present studies in this area conducted in Poland to date, main methodological assumptions, specification of the hierarchical rater model used in the analyses and study results. The main value of the report is the search for solutions designed to minimise the rater effect on the performance of students sitting the exam.

2011 and 2012 were selected for the rater effect study. For practical reasons, exam papers were drawn from two Regional Examination Boards (*Okręgowa Komisja Egzaminacyjna – OKE*): Jaworzno and Kraków. It was important to collect a set of papers with the following characteristics:

1. diversity of the level of examined skills against the entire scale used in the study;
2. for the Polish language – diversity of essay topics for the analysis of the rater-essay topic interaction.

Studies were conducted on a representative sample of raters involved in the assessment of Matura exams, stratified for Regional Examination Boards (extended by teachers that are not raters and with students, in the case of the Polish language Matura exam). A complex model was used to link all the papers and raters from the eight OKE districts, as a result of which each essay from the Polish language exam was assessed by one rater from each OKE. In the case of Matura exam in mathematics, for which papers on the basic and extended level were selected, each paper was assessed by four raters (each from a different OKE).

Assessment was conducted in conditions that were possibly closest to those present at the main Matura exam session (held in May). An analogical assessment coordination structure was applied, the teams were located in the same cities, and unchanged assessment criteria and schemes were applied.

Results of the analyses

Studies have shown that raters have significant influence on assessment results, both for the Polish language and mathematics Matura exam. In addition, in the case of the Polish language, this influence is clearly stronger. Difference in leniency at the level of test between 25% of the most lenient raters and 25% of the most stringent raters for the Polish language is (depending on the year and essay topic) from 3.1 to 3.7 percentage points of the exam result. For mathematics, these differences are between 0.87 to 1.36 percentage points of the exam result². Such result is not surprising. As research show, rating essays will always be to some extent subjective. Reaching full rater agreement is utopian. Demanding 100% rater agreement leads to petrification of criterion-referenced assessment. As a consequence, it can have a negative influence on teaching important skill of writing essays.

Differences in the leniency of assessment occur not only at the individual level, but also at the level of Regional Examination Boards. In the case of the Polish language, in some OKE these differences are systematic and are present in all essay topics examined. The average difference between the most stringent and most lenient board was 1.6 percentage points of the exam result. Differences in the exam results of students at the OKE level show that the effect of raters' leniency explains a large portion of this diversity (on average for all topics 11% of results' variance is explained). In the case of mathematics differences at the OKE level also occur, but

² Result for the entire exam - including short-answer tasks and an essay for the Polish language as well as closed and open tasks for mathematics.

they are smaller and are not systematic. In this case, instead of differences, we may talk about trends in leniency of raters, which are typical of particular OKE.

Another conclusion from the analysis of the rater effect is the observed disruption of result distribution that occurs at the exam cut-off score. Since the exam cut-off score is used for certification, it is desirable to increase the reliability of assessment in its area. In the case of the Polish language exams, the cut-off score is unfortunately the source of the rater effect involving increased local leniency just below the pass grade threshold. It is worth noting that even if the effect of a local increase in reliability of assessment has a positive impact on the assessment of students whose papers are in the area of the exam cut-off score, a local increase in leniency is a negative factor in the context of the exam as a whole. Because of it, some papers are assessed in a systematically different manner than others. This effect is much stronger for the Polish language Matura exam than for mathematics. There is also a statistically significant difference in rater leniency around the exam cut-off score among the regional examination boards.

Another issue analysed in this study is the comparability of assessment of handwritten and computer-typed essays. Based on the analysis of data from comparative studies of Matura essays assessment results in the original handwritten copy and computer-typed copy, it was observed that handwritten essays are assessed higher than copies typed on a computer. The difference in the average grades at the Matura exam assessed on a five-point scale is more than 2 points and is statistically significant at the significance level of $p = 0.01$. This result is surprising for practitioners, although a similar result was obtained in a number of studies conducted throughout the world, comparing the quality of assessment of handwritten texts and their copies written using a word processor.

In the presented study, the link between assessment and the author's handwriting traits was also the subject of analysis. The results suggest that essays written by authors in poor handwriting are assessed lower compared to essays written in a sophisticated and moderately sophisticated handwriting. This difference is large and statistically significant. Links of handwriting legibility with a rater's essay assessment are moderate. It should be noted, however, that legibility and sophistication are highly correlated traits of handwriting. Most of the barely legible essays are written in poor handwriting. These results are consistent with the results of studies conducted in other countries.

In the case of the impact of variables describing individual characteristics of raters (recorded during the study) on the degree of precision and leniency of raters for both the Polish language and mathematics, only links specific for the particular subject were observed. A link with the index of attitude to work in teams during the assessment may be observed only in the case of leniency of the Polish language raters. Differences in the level of internal motivation and attitude to work in teams in specific REBs were observed.

The impact of time spent assessing essays and the order in which they were assessed on the number of scores obtained in the constructed responses was also examined. In the case of the order in which essays were assessed and time spent assessing them, these effects, although statistically significant, are close to zero (for both Polish language and mathematics papers). These results suggest that the fatigue does not impair the rater's evaluation process.

Rekomendacje

Przeprowadzone badania dostarczyły wniosków, które stanowią podstawę do sformułowania wielu rekomendacji, zorientowanych na zwiększenie rzetelności oceniania i zminimalizowania efektu egzaminatora. Rekomendacje te można podzielić na trzy kategorie:

1. Propozycje związane z wykorzystaniem e-oceniania.
2. Propozycje związane z konstrukcją schematów oceniania.
3. Propozycje związane z przeprowadzaniem oraz koordynacją oceniania prac egzaminacyjnych.

Pierwsza wymieniona grupa rekomendacji wiąże się z wykorzystaniem unikalnych możliwości, jakich dostarcza oprogramowanie umożliwiające przeprowadzanie e-oceniania. Znacznie ułatwia ono monitorowanie efektów egzaminatora na poziomie wybranych zbiorowości i na poziomie indywidualnego egzaminatora. Szczególnie cenną właściwością e-oceniania jest możliwość wykrycia prac (lub zadań) ocenionych nierzetelnie i ich powtórna ocena, co ma fundamentalne znaczenie z perspektywy zdającego egzamin.

Druga grupa rekomendacji zawiera postulaty odnośnie konstrukcji schematów oceniania, mając na względzie wykryte w badaniach zależności między liczbą kategorii oceny zadania a rzetelnością jego oceny.

Trzecia grupa rekomendacji odnosi się do samego procesu oraz oceniania prac w trakcie sesji i uwzględnia uwagi dotyczące takich kwestii jak szkolenia egzaminatorów, współpraca między okręgowymi komisjami egzaminacyjnymi (OKE) bądź bieżące monitorowanie jakości procesu oceniania przez egzaminatorów. Wszystkie trzy grupy rekomendacji są opisane szczegółowo poniżej.

Propozycje związane z wykorzystaniem e-oceniania

E-ocenianie, czyli ocenianie z wykorzystaniem sieci (*electronic marking*), jest rozwiązaniem istotnie różniącym się od oceniania tradycyjnego. Podstawowe różnice to, po pierwsze, inny nośnik, na którym dostarczana jest egzaminatorowi praca zdającego. Egzaminator otrzymuje obraz rozwiązania zadania egzaminacyjnego, który może przeglądać i punktować na ekranie komputera, korzystając z wielu funkcjonalności oprogramowania ułatwiających ocenianie. Po drugie, dzięki zastosowaniu Internetu zapewniona jest interakcja egzaminatora z przewodniczącym zespołu egzaminatorów (PZE) niezależnie od tego, czy ocenianie odbywa się w zespole, czy też w domu egzaminatora. Trzecią i chyba najważniejszą różnicą z punktu widzenia rzetelności oceniania jest monitorowanie na bieżąco całego procesu oceniania w kraju, a w szczególności monitorowanie pracy egzaminatora – prowadzone z wykorzystaniem zadań kalibrujących (zadań ocenionych przez zespół ekspertów) i przez PZE. Egzaminator oceniający nierzetelnie zadanie lub wiązkę zadań automatycznie zostaje niedopuszczony do dalszego oceniania aż do czasu kontaktu on-line z PZE. Przewodniczący zespołu po zidentyfikowaniu przyczyn i wyjaśnieniu zaistniałej sytuacji może dopuścić egzaminatora do dalszego oceniania, może przydzielić oceniającemu dodatkową pulę zadań ćwiczeniowych przed dalszym ocenianiem, a w skrajnych przypadkach nawet odsunąć od dalszego oceniania. Istotne jest to, że już ocenione prace (zadania) łatwo mogą być skierowane do powtórnej oceny przez innego egzaminatora. Ponadto w procesie oceniania zadania do oceny mogą być przydzielone losowo, niezależnie od tego, w jakiej części kraju zdający pisał egzamin. Zaletą e-oceniania jest także łatwość pobierania prób z całej populacji, co daje możliwość wczesnego identyfikowania opuszczeń czy innych anomalii w wynikach oceniania. Prace nad przygotowaniem warunków do wdrożenia e-oceniania w Polsce rozpoczęły się w CKE pod koniec 2007 roku. Obecnie system egzaminacyjny dysponuje w tym zakresie koniecznymi rozwiązaniami informatycznymi oraz przygotowaną kadrą ekspertów i egzaminatorów. W 2015 roku e-ocenianie wprowadzane jest dla egzaminu gimnazjalnego z matematyki na obszarze 4 okręgowych komisji egzaminacyjnych.

Zastosowanie systemu e-oceniania w kontekście efektu egzaminatora sprowadza się do trzech postulatów:

1. Wprowadzenie w pełni losowego przyporządkowania prac egzaminacyjnych z całego kraju do egzaminatorów z różnych okręgowych komisji egzaminacyjnych.
2. Wprowadzenie kompleksowych i wystandaryzowanych procedur umożliwiających szacowanie wskaźników rzetelności wszystkich egzaminatorów na podstawie wielokrotnych ocen pracy w trakcie trwania sesji egzaminacyjnej.
3. Zbudowanie systemu umożliwiającego wykrywanie, na podstawie przesłanek statystycznych, w czasie rzeczywistym prac, co do których są wątpliwości, że nie zostały rzetelnie ocenione oraz egzaminatorów, którzy mogą systematycznie oceniać prace w sposób nierzetelny.

Pierwsza rekomendacja wydaje się nie generować dodatkowych kosztów po okresie wdrożenia, gdyż jest funkcjonalnością wbudowaną bezpośrednio w gotowy już system e-oceniania. Natomiast korzyści płynące z losowego przydziału prac (zadań) z całego kraju pomiędzy egzaminatorów z różnych OKE wydają się nie do przecenienia w kontekście możliwości monitorowania efektów egzaminatora na poziomie większych zbiorowości egzaminatorów jak cała OKE, ale również na niższym poziomie, np. zespołów egzaminatorów (ZE). W sytuacji, w której prace pomiędzy wszystkimi egzaminatorami byłyby rozprowadzone w sposób losowy, jakiegokolwiek istotne statystycznie różnice w rozkładach wyników uzyskiwanych między OKE lub między zespołami egzaminatorów musiałyby być uznane za przejaw różnic w sposobie oceniania prac, a nie umiejętności uczniów. W szczególności:

1. różnice w średnich wynikach na poziomie zadań oraz całych arkuszy egzaminacyjnych świadczyłyby o zróżnicowaniu w zakresie łagodności/surowości;
2. różnice w zakresie rozproszenia wyników na poziomie zadań oraz całych arkuszy egzaminacyjnych świadczyłyby o poszerzeniu/zawężeniu skali;
3. różnice w zakresie rzetelności części egzaminów zawierającej zadania otwarte świadczyłyby o występowaniu zróżnicowania w zakresie błędów losowych przy ocenie prac, tj. o zróżnicowanej nierzetelności ocen.

Zebrane z takiej analizy wyniki stanowiłyby bardzo wartościową informację o występowaniu różnic w sposobie oceniania prac między OKE i ZE, które można by wykorzystać w celach monitorowania skali zjawisk związanych z efektem egzaminatora na poziomie wybranych zbiorowości. Należy również zauważyć, że pod względem metodologicznej trafności oraz mocy statystycznej wyciąganych wniosków o systematycznych efektach egzaminatora na poziomie całych OKE, analizy tak zebranych danych przewyższałyby znacznie te, które przedstawiono w niniejszym raporcie. Osiągnięcie większej mocy wnioskowania byłoby skutkiem objęcia analizą całej populacji zdających egzamin oraz wszystkich egzaminatorów. Trafność analiz nie byłaby w żaden sposób zagrożona, gdyż badanie dotyczyłoby faktycznych danych egzaminacyjnych i byłoby, ze względu na zastosowanie losowego przydziału prac (zadań), przeprowadzone de facto w zgodzie ze schematem eksperymentalnym.

Druga rekomendacja postuluje wykorzystanie informacji zbieranych z wielokrotnych ocen pracy egzaminacyjnej do automatycznego obliczania dla każdego egzaminatora wskaźników określających stopień rzetelności dokonywanych przez niego ocen, na podstawie zgodności z innymi ocenami danej pracy. Część prac egzaminacyjnych jest poddawana wielokrotnej ocenie, lecz nie stworzono spójnych dla całego systemu instrukcji ani procedur, które na podstawie zebranej w ten sposób informacji o zgodności ocen umożliwiałyby w sposób ilościowy oraz obiektywny profilowanie jakości pracy poszczególnych egzaminatorów względem ogólnokrajowych norm. Taka obiektywna ocena jakości pracy egzaminatora ułatwiałaby identyfikację egzaminatorów, którzy wymagałoby odpowiedniej informacji zwrotnej czy dodatkowego szkolenia mającego na celu poprawę sposobu, w jaki oceniają lub nawet wykluczenia z dalszego oceniania.

Trzecia rekomendacja jest powiązana z drugą w taki sposób, że ma na celu zoptymalizowanie decyzji o przekazywaniu pracy do powtórnej oceny. Wiadomo, że każda dodatkowa ocena pracy generuje dodatkowe jednostkowe koszty, które powodują, że ze względów ekonomicznych system egzaminów zewnętrznych może sobie pozwolić na powtórna, zwiększającą rzetelność, ocenę jedynie pewnego odsetka wszystkich prac. W związku z tym, zamiast rozsyłać do powtórnej oceny prace bez uwzględnienia jakichkolwiek kryteriów (losowo), racjonalne wydaje się rozważenie możliwości identyfikacji w czasie rzeczywistym prac, co do których istnieją pewne przesłanki, że mogą one być ocenione nierzetelnie, i wysłanie do powtórnej oceny takich właśnie prac.

W zakresie trzeciej rekomendacji postuluje się wprowadzenie rozwiązania, które, w ramach procedury e-oceniania, umożliwiłoby identyfikowanie potencjalnie nierzetelnych ocen w czasie rzeczywistym na podstawie przesłanek statystycznych. Podczas sesji egzaminacyjnej zbierane jest wiele danych o zdających egzamin, które spełniają dwie przesłanki: (I) są/mogą być dostępne przed przystąpieniem do oceny pracy przez egzaminatora oraz (II) są istotnie skorelowane z wynikami uzyskiwanymi w zadaniach podlegających ocenie przez egzaminatorów. Nazwijmy te dane „danymi warunkującymi”. Wśród takich potencjalnych danych warunkujących można wymienić wyniki z części danego egzaminu zawierającej zadania zamknięte, ale również wyniki zadań zamkniętych z arkuszy egzaminacyjnych z innych przedmiotów lub części egzaminu.

W momencie, gdy pewien odsetek zadań otwartych z danego egzaminu zostałby oceniony, możliwe byłoby wykalibrowanie modelu statystycznego, który przewidywałby oceny za zadania otwarte z danego egzaminu na podstawie danych warunkujących. Umożliwiłoby szacowanie na bieżąco odchylenia oceny danego zadania przez egzaminatora od oceny przewidywanej przez

model statystyczny. W przypadku odchyleń przekraczających określony próg zadanie powinno być automatycznie przekazane do powtórnej oceny przez innego egzaminatora. Głównym zadaniem takiego systemu byłoby zoptymalizowanie decyzji o przekazywaniu zadań do powtórnej oceny tak, aby nakłady ponoszone przez system na powtórna ocenę pewnego odsetka prac egzaminacyjnych przekładały się jak najbardziej efektywnie na zwiększenie rzetelności ocen.

Propozycje związane z konstrukcją schematów oceniania

Rekomenduje się unikanie wykorzystania w schematach punktowania nadmiernej liczby poziomów wykonania zadań – w praktyce powyżej 6. Przeprowadzone badania wskazują, że więcej poziomów wykonania (przede wszystkim kryterium *treści* w języku polskim, ale także zadania z matematyki na 7 poziomach wykonania) prowadzi do wyraźnie wyższego poziomu błędów oceny. Skutkuje to sytuacją, w której końcowy wynik egzaminu nie szereguje uczniów poprawnie ze względu na poziom umiejętności.

Jeśli niezbędne jest nadanie danemu zadaniu lub kryterium oceny stosownej wagi, rekomenduje się stosowne przeliczenie skali poziomów wykonania na punkty. Rozwiązanie takie jest już zresztą stosowane w kryteriach kompozycji, stylu i języka przy ocenie wypracowania z języka polskiego, np. kompozycja i styl posiadają cztery poziomy wykonania punktowane: 0, 1, 3, 5. Alternatywnie można rozbić kryterium na kilka części, ocenianych niezależnie od siebie na krótszych skalach. Takie rozbieżności zostało zastosowane dla rozwinięcia tematu w schematach oceniania w badanych egzaminach z języka polskiego, lecz na karcie odpowiedzi egzaminatorzy zaznaczali sumę za całe kryterium. Istotne w przypadku takiego rozbieżności jest to, aby do systemu informatycznego wprowadzane były poszczególne części, a nie suma za całe kryterium. Utrata informacji przy sumowaniu punktacji za poszczególne części skutkuje gorszą jakością przeprowadzanych analiz statystycznych.

Wskazane jest również przeprowadzanie z większą starannością panelu, którego celem jest uzyskanie konsensusu odnośnie kryteriów i schematu oceniania, głównie dla języka polskiego. Należy rozważyć zwiększenie puli prac, na podstawie których doprecyzowane są autorskie propozycje kryteriów oceniania pracy pisemnej z języka polskiego. Zaledwie 10 prac (w porównaniu z przeszło 200 w przypadku matematyki) wydaje się niewystarczających, aby doprecyzować autorskie propozycje kryteriów przygotowane w procesie tworzenia arkusza egzaminacyjnego.

Propozycje związane z przeprowadzeniem oraz koordynacją oceniania prac egzaminacyjnych

Koordinacja oceniania zaczyna się podczas panelu w CKE, którego celem jest uzyskanie konsensusu odnośnie kryteriów i schematu oceniania, Konsensus, który zwykle prowadzi do ustaleń sygnowanych podpisami wszystkich uczestników panelu, poprzedza bardzo ważny proces kaskady szkoleń – przewodniczących zespołów egzaminatorów, a następnie wszystkich egzaminatorów. Jakość tych szkoleń ma fundamentalne znaczenie dla profesjonalizmu PZE w przeprowadzeniu szkoleń przygotowujących bezpośrednio do oceniania w danej sesji. Jak wykazały wyniki prezentowanych w raporcie badań, dobrze przygotowani podczas szkolenia przed ocenianiem nauczyciele niebędący egzaminatorami oceniali w porównywalny sposób z doświadczonymi egzaminatorami.

Bardzo istotne dla poprawy rzetelności oceniania jest podjęcie działań prowadzących do likwidacji zaburzeń rozkładu ocen na progu zdawalności w egzaminach maturalnych na poziomie podstawowym z przedmiotów obowiązkowych. Zaburzenia te są konsekwencją celowego wyboru do podwójnego oceniania prac progowych. W przypadku matematyki zadania zamknięte oceniane są automatycznie po skanowaniu kart odpowiedzi w OKE. Aby rozpoznać prace z wynikiem progowym, okręgowe komisje egzaminacyjne stosują w tym zakresie różne rozwiązania. W niektórych przypadkach wymaga to dodatkowo od egzaminatorów dokonania ręcznej oceny zadań zamkniętych w celu identyfikacji takich prac już podczas oceniania, gdyż automatyczna ocena następuje po zeskanowaniu karty odpowiedzi po uzupełnieniu jej przez egzaminatorów o oceny zadań otwartych. Są też stosowane takie rozwiązania, w których prace progowe są identyfikowane dopiero po zakończeniu pracy zespołów egzaminacyjnych i po skanowaniu kart odpowiedzi. W takim przypadku prace progowe są dodatkowo oceniane przez zespół PZE. W przypadku języka polskiego egzaminator dokonuje oceny wszystkich zadań, zatem w momencie oceniania pracy pisemnej jest w stanie stwierdzić, czy dana praca znajduje się na progu zaliczenia. Aby uniknąć zaburzeń w ocenianiu prac progowych, należałoby zrezygnować z celowego ich doboru do podwójnego oceniania, a także niezależnie oceniać poszczególne zadania. Niezależna ocena zadań w egzaminie maturalnym z matematyki ma miejsce w przypadku, gdy egzaminatorzy nie oceniają całej pracy danego ucznia, lecz zaledwie jej część (przydzielone zadania). Wprowadzenie takiego sposobu oceniania w przypadku języka polskiego wymagałoby zastąpienia wypracowania dwoma zadaniami sprawdzającymi umiejętność pisanego własnego tekstu.

Powyższe działania wiążą się z pierwszą kategorią rekomendacji, gdyż można je bez dodatkowych kosztów wdrożyć w przypadku e-oceniania. Proponowane dobieranie prac przy użyciu modelu statystycznego pozwala uniknąć skupiania się na pracach progowych, a przekazywanie egzaminatorowi pojedynczych zadań do oceny zapewnia niezależną ich ocenę.

Wdrożenie e-oceniania jest procesem długofalowym, który zaczyna się w 2015 roku dla egzaminu gimnazjalnego z matematyki. Dlatego też istotne jest podjęcie działań, które podczas tradycyjnego oceniania prowadzić będą do poprawy rzetelności oceniania każdego egzaminu, a w szczególności matury, która jest egzaminem wstępnym na wyższe uczelnie. Działania takie mogą być wprowadzone niemal natychmiast, gdyż już są stosowane, ale w zbyt wąskim zakresie lub też miały miejsce w przeszłości ale ich zaniechano. W tym obszarze (w tradycyjnym ocenianiu) rekomenduje się:

1. Zwiększenie procentu podwójnie ocenianych prac podczas sesji.
2. Powrót do monitorowania jakości oceniania w skali kraju z wykorzystaniem kalibrujących prac egzaminacyjnych. W latach 2006–2009 stosowane były rozwiązania mające na celu monitorowanie porównywalności oceniania. Wyniki monitorowania miały istotne znaczenie jako informacja zwrotna na poziomie indywidualnym, zespołów egzaminatorów i OKE w trakcie danej sesji, ale nie były wykorzystywane systemowo do refleksji i doskonalenia oceniania w kolejnych sesjach egzaminacyjnych.
3. Prowadzenie systematycznych analiz rzetelności oceniania po każdej sesji.

Podziękowania

Autorzy dziękują serdecznie prof. dr hab. Wojciechowi Guzickiemu za pomoc w przygotowaniu modelu linkowania prac i egzaminatorów wewnątrz OKE i pomiędzy OKE oraz Wojciechowi Małeckiemu, dyrektorowi OKE we Wrocławiu, za cenne uwagi.

1. Wprowadzenie

Egzamin maturalny obejmuje przedmioty obowiązkowe – matematykę, język polski (część pisemna i ustna), język obcy nowożytny (część pisemna i ustna) oraz przedmioty dodatkowe do wyboru. Egzaminy pisemne mogą być zdawane na poziomie podstawowym lub rozszerzonym, przy czym matematyka, język polski i język obcy nowożytny pisane jako egzaminy obowiązkowe wymagane są na poziomie podstawowym. Część ustna zdawana jest na jednym poziomie i oceniana jest w szkole przez przedmiotowy zespół egzaminacyjny.

Dla egzaminu maturalnego z przedmiotów obowiązkowych (części pisemnej i ustnej) obowiązuje taki sam próg zdawalności egzaminu – 30% maksymalnej liczby punktów. Dla egzaminów zdawanych jako dodatkowe w polskim systemie egzaminacyjnym nie ma progu zaliczenia. Absolwent, który przystąpił do wszystkich egzaminów z przedmiotów obowiązkowych w części ustnej i pisemnej i nie zdał jednego egzaminu (ustnego lub pisemnego), może przystąpić ponownie do egzaminu z tego przedmiotu w sesji poprawkowej w sierpniu.

W polskim systemie egzaminów zewnętrznych prace z pisemnej części egzaminu z poszczególnych przedmiotów oceniane są przez egzaminatorów rekrutowanych spośród nauczycieli, którzy po przeszkoleniu i spełnieniu określonych wymagań zostali wpisani do ewidencji egzaminatorów jednej z okręgowych komisji egzaminacyjnych. Wymagania i ramowy program szkolenia określony jest przez rozporządzenie ministra edukacji narodowej³. Sposób rekrutacji spośród wpisanych do ewidencji egzaminatorów danej OKE jest zróżnicowany. Niektóre komisje stosują kryteria polegające na poprawnej ocenie on-line na platformie MOODLE kilku prac z zakresu danego egzaminu. Stawki za ocenę jednej pracy danego egzaminu ustalane są centralnie poprzez rozporządzenie Ministra Edukacji Narodowej Rozporządzenie⁴ w stosunku do minimalnej stawki wynagrodzenia zasadniczego nauczyciela dyplomowanego posiadającego tytuł zawodowy magistra oraz przygotowanie pedagogiczne.

Dla zapewnienia rzetelności oceniania, a tym samym egzaminu, ważne jest, aby poszczególni egzaminatorzy posługiwali się w taki sam sposób i stabilnie tymi samymi kryteriami oceny w trakcie całej sesji oceniania. W uproszczeniu można problem przedstawić następująco: czy praca danego ucznia zostałaby oceniona tak samo przez różnych egzaminatorów? Czy ten sam egzaminator oceniłby tę samą pracę tak samo na początku sesji i pod koniec, kiedy ma za sobą doświadczenie punktowania kilkudziesięciu prac? Opisywane zagadnienie zwane jest zbiorczo efektem egzaminatora. Efekt egzaminatora jest zjawiskiem charakterystycznym towarzyszącym nieodłącznie ocenom dokonywanym w trakcie punktowania zadań otwartych. Nie da się go całkowicie wyeliminować. Ważne jest jednak, aby je kontrolować i dążyć do jego minimalizacji.

Niekontrolowany efekt egzaminatora prowadzi do błędów pomiaru, które powinny być identyfikowane i minimalizowane (eliminowane), jeżeli chcemy zapewnić wysoki poziom rzetelności egzaminu. W zależności od typu egzaminu efekt egzaminatora niesie ze sobą konsekwencje, które z punktu widzenia egalitarności systemu edukacji powinny zostać wyeliminowane. Sprawdzian po szóstej klasie szkoły podstawowej czy egzamin gimnazjalny nie posiadają progów zaliczenia i ich funkcja selekcyjna nie jest zbyt duża, zatem w tych przypadkach ewentualna niesprawiedliwość (wynikająca z efektu egzaminatora) w ocenianiu nie ma tak dużych konsekwencji, jak w egzaminie maturalnym i egzaminie potwierdzającym kwalifikacje zawodowe. Należy jednak pamiętać, że wynik egzaminu gimnazjalnego stanowi wraz z ocenami szkolnymi podstawę rekrutacji do szkoły ponadgimnazjalnej i decyduje o dostępie do wybranej ścieżki edukacji. Natomiast jeśli chodzi o egzamin maturalny, to jest on jednocześnie egzaminem wstępnym na studia. W związku z tym jakikolwiek wpływ na otrzymaną ocenę niezależny od zdającego może spowodować, że absolwent (nie w związku ze zbyt niskim poziomem umiejętności) nie dostanie się na wybrany kierunek studiów. Z tego powodu bardzo istotne jest zapewnienie każdemu zdającemu szansy na uzyskanie nieobciążonego efektem egzaminatora wyniku z egzaminu.

Szczególnie podatne na wpływ efektu egzaminatora są zadania wymagające dłuższych wypowiedzi pisemnych. Takie zadania znajdują się w każdym z opisywanych egzaminów. Egzamin maturalny z języka polskiego na poziomie rozszerzonym to tylko jedno zadanie, do którego oceny egzaminatorzy wykorzystują szereg kryteriów. Na poziomie podstawowym za wypowiedź pisemną

³ Rozporządzenie Ministra Edukacji Narodowej z dnia 21 kwietnia 2009 r. w sprawie ramowego programu szkolenia kandydatów na egzaminatorów, sposobu prowadzenia ewidencji egzaminatorów oraz trybu wpisywania i skreślenia egzaminatorów z ewidencji (Dz. U. z 2009 r., Nr 70, poz. 600).

⁴ Rozporządzenie Ministra Edukacji Narodowej z dnia 8 czerwca 2009 r. w sprawie warunków wynagradzania egzaminatorów za udział w przeprowadzaniu sprawdzianu i egzaminów oraz nauczycieli akademickich za udział w przeprowadzaniu części ustnej egzaminu maturalnego (Dz. U. z 2009 r., Nr 94, poz. 773).

(na jeden z tematów do wyboru) można otrzymać około 70% całkowitej liczby punktów. Przy tak dużym wpływie oceny za tę wypowiedź na końcową punktację bardzo ważne jest to, aby kontrolować i minimalizować efekt egzaminatora. Należy dbać, aby zdający otrzymywali ocenę odpowiednią do prezentowanego poziomu umiejętności, a niewynikającą z tego, do jakiego egzaminatora praca została przydzielona.

W Polsce efekt egzaminatora jest słabo zbadany – brakuje badań prowadzonych na dużą skalę, w odniesieniu do egzaminów zewnętrznych. System egzaminacyjny powinien zawierać w sobie mechanizm kontroli i monitorowania efektu egzaminatora oraz dążyć do jego zminimalizowania. Niestety nie dysponujemy wiarygodnym oszacowaniem poziomu efektu egzaminatora dla poszczególnych egzaminów. Przeprowadzone badania pozwoliły na określenie efektu egzaminatora i porównywalności oceniania pomiędzy okręgowymi komisjami egzaminacyjnymi dla dwóch egzaminów maturalnych – z języka polskiego i matematyki.

2. Charakterystyka egzaminu maturalnego z języka polskiego

Prezentowane w niniejszym opracowaniu badania efektu egzaminatora przeprowadzone zostały w kontekście egzaminów funkcjonujących w odniesieniu do podstawy programowej obowiązującej w szkołach ponadgimnazjalnych do 2012 roku. W konsekwencji wprowadzania zmian podstawy programowej kształcenia ogólnego⁵ następują także zmiany dotyczące egzaminów na progach poszczególnych etapów edukacyjnych. W egzaminach gimnazjalnych taka zmiana miała miejsce w 2012 roku. Zmiany w egzaminie maturalnym będą miały miejsce w sesji 2015 roku. Szczegółowo egzamin maturalny z języka polskiego obowiązujący od 2015 roku przedstawiony jest w *Informatorze o egzaminie maturalnym z języka polskiego od roku szkolnego 2014/2015* (CKE i OKE, 2013a). W pisemnej części matury z języka polskiego struktura egzaminu nie ulega zmianie. Podobnie jak w poprzednich latach, pisemny egzamin maturalny z języka polskiego składa się z dwóch części: pisemnej i ustnej. Część pisemna zdawana jest na poziomie podstawowym obowiązkowo, a na poziomie rozszerzonym jako przedmiot dodatkowy wybrany przez zdającego. Począwszy od 2015 roku na poziomie podstawowym w części testowej sprawdzane będą: rozumienie tekstu nieliterackiego, świadomość językowa oraz umiejętność pracy z tekstem. W drugiej części egzaminu pisemnego, podobnie jak dotychczas, maturzyści będą pisać wypracowanie, mając do wyboru jeden z dwóch zaproponowanych w arkuszu egzaminacyjnym tematów wymagających napisania rozprawki w przypadku pierwszego tematu lub interpretacji tekstu poetyckiego – temat drugi.

Na poziomie rozszerzonym pisemny egzamin, podobnie jak dotychczas, będzie jednoczęściowy. Sprawdzaniu będą podlegały umiejętności interpretacji porównawczej utworów literackich lub tworzenia wypowiedzi argumentacyjnej wymagającej odniesienia się do tekstu historycznoliterackiego, teoretycznoliterackiego albo krytycznoliterackiego. Zdający będzie miał możliwość wyboru formy wypowiedzi np. rozprawka lub szkic. O ile w przypadku pisemnej części egzaminu zarówno na poziomie podstawowym, jak i rozszerzonym forma pozostaje bez zmian, to zmieniają się istotnie charakter zadań i sprawdzane umiejętności.

Jeżeli chodzi o ocenianie prac egzaminacyjnych, to zmianę stanowi przejście w kierunku oceniania holistycznego i modyfikacja kryteriów oceniania. Zmiany te mogą mieć wpływ na efekt egzaminatora. Także wprowadzenie stałych form wypowiedzi na egzaminie pisemnym, ujednolicenie formuł tematów, a także wprowadzenie stałych, niezmiennych z roku na rok kryteriów oceniania dla poszczególnych form gatunkowych może okazać się zmianą istotną z punktu widzenia jakości oceniania. Ocena tych zmian będzie możliwa dopiero po sesji egzaminacyjnej w 2015 roku.

Egzamin maturalny z języka polskiego obowiązujący do końca roku 2014 opisany jest informatorze maturalnym z języka polskiego z 2008 roku (CKE, 2007). Część pisemna, podobnie jak to będzie miało miejsce od sesji 2015, zdawana jest na poziomie podstawowym obowiązkowo a na poziomie rozszerzonym jako przedmiot dodatkowy wybrany przez zdającego. Oceniana jest przez egzaminatorów zewnętrznych okręgowej komisji egzaminacyjnej, w obrębie której zlokalizowana jest szkoła. Przykładowo, jeżeli szkoła zlokalizowana jest w Opolu, to prace z tej szkoły oceniane są przez egzaminatorów Okręgowej Komisji Egzaminacyjnej we Wrocławiu, ponieważ Opole jest w obrębie działania tej OKE. Pisemna część egzaminu na poziomie podstawowym składa się z dwóch części, z których pierwsza sprawdza rozumienie czytanego tekstu, a druga umiejętność sformułowania własnej wypowiedzi. Część sprawdzająca rozumienie czytanego tekstu odnosi się do zamieszczonego w arkuszu egzaminacyjnym tekstu lub jego fragmentu, z którym powiązane jest 10–16 zadań zamkniętych lub otwartych typu krótkiej odpowiedzi. Część sprawdzająca umiejętność pisania własnego tekstu to dwa tematy, którym również towarzyszy tekst lub jego fragment, z których zdający wybiera jeden. Na poziomie podstawowym za część sprawdzającą rozumienie czytanego tekstu zdający maksymalnie może uzyskać 20 punktów, natomiast w części sprawdzającej umiejętność pisania 50 punktów.

Prace egzaminacyjne oceniane są w całym kraju według takich samych kryteriów. Podczas oceniania wypracowania zdającego na poziomie podstawowym przyjęte są następujące wagi kryteriów:

- rozwinięcie tematu – 25 punktów (50%)
- kompozycja – 5 punktów (10%)

⁵ Rozporządzenie Ministra Edukacji Narodowej z dnia 23 grudnia 2008 r. w sprawie podstawy programowej wychowania przedszkolnego oraz kształcenia ogólnego w poszczególnych typach szkół (Dz. U. z dnia 15 stycznia 2009 r. Nr 4, poz. 17)

Efekt egzaminatora w ocenianiu prac maturalnych z języka polskiego i matematyki

- styl – 5 punktów (10%)
- język – 12 punktów (24%)
- zapis (ortografia i interpunkcja) – 3 punkty (6%).

Od 2010 roku pisemna część egzaminu maturalnego z języka polskiego na poziomie rozszerzonym polega na napisaniu własnego tekstu na podstawie tekstu literackiego lub jego fragmentu zamieszczonego w arkuszu egzaminacyjnym z wykorzystaniem wiedzy o całym utworze, znajomości kontekstu literackiego, historycznego itp. Podobnie jak w przypadku egzaminu na poziomie podstawowym zdający ma do wyboru jeden z dwóch tematów. Za pracę zdający może otrzymać maksymalnie 40 punktów. Zgodnie z przyjętymi przez CKE kryteriami oceny zdający może otrzymać maksymalnie za:

- rozwinięcie tematu – 26 punktów (65%)
- kompozycję – 2 punkty (5%)
- styl – 2 punkty (5%)
- język – 8 punktów (20%)
- zapis (ortografia i interpunkcja) – 2 punkty (5%).

Za szczególne walory pracy maturzysty mógł w latach 2005–2014 otrzymać dodatkowo do 4 punktów, które przyznawano za wypracowanie ocenione poniżej maksymalnej liczby punktów, jednak wyróżniające się oryginalnością w zakresie rozwinięcia tematu, kompozycji, stylu lub języka. Łączna liczba punktów za całą pracę egzaminacyjną nie mogła jednak przekroczyć maksymalnej liczby punktów do osiągnięcia.

Podczas egzaminu z języka polskiego zdający otrzymuje jeden arkusz egzaminacyjny. Na rozwiązanie zadań z arkusza z poziomu podstawowego zdający ma 170 minut, a z arkusza z poziomu rozszerzonego 180 minut.

Arkusze egzaminacyjne egzaminu maturalnego na poziomie podstawowym, którego wyniki zostały włączone do analiz są dostępne na stronie CKE pod adresem <http://www.cke.edu.pl/index.php/2011-em> oraz <http://www.cke.edu.pl/index.php/2012-em/maj-2012-r>. Parametry psychometryczne zadań z tych dwóch arkuszy przedstawione są w aneksie 1.

3. Charakterystyka egzaminu maturalnego z matematyki

Egzamin maturalny z matematyki jest (od roku 2010) egzaminem obowiązkowym na poziomie podstawowym. Podobnie jak w przypadku języka polskiego, w następstwie zmian podstawy programowej kształcenia ogólnego następują także zmiany dotyczące egzaminu maturalnego z matematyki poczynając od sesji wiosennej 2015 roku. Szczegółowo egzamin maturalny z matematyki obowiązujący od 2015 roku przedstawiony jest w *Informatorze o egzaminie maturalnym z matematyki od roku szkolnego 2014/2015* (CKE i OKE, 2013b). Jeżeli chodzi o matematykę, to w porównaniu z egzaminem maturalnym obowiązującym do 2015 roku struktura egzaminu na poziomie podstawowym nie ulega zmianie. Natomiast na poziomie rozszerzonym egzamin zmienia się w kierunku lepszego sprawdzania stopnia opanowania umiejętności opisanych przez wymagania ogólne podstawy programowej. W związku z tym w arkuszach egzaminacyjnych można się spodziewać mniej rozbudowanych zadań sprawdzających znajomość algorytmów i umiejętność posługiwania się nimi w typowych zastosowaniach. Akcent zostanie przesunięty w kierunku sprawdzania rozumienia pojęć matematycznych oraz umiejętności dobierania własnych strategii matematycznych do nietypowych sytuacji. Ponieważ prezentowane badania przeprowadzone zostały przed rokiem 2015, to wszystkie opisy w dalszej części tekstu odniesione są do egzaminów maturalnych obowiązujących do sesji wiosennej 2014 roku.

Egzamin na poziomie podstawowym trwa 170 minut i wymaga od zdającego rozwiązania zadań egzaminacyjnych sprawdzających rozumienie pojęć i umiejętność ich zastosowania w życiu codziennym oraz zadań o charakterze problemowym. Zadania w arkuszach egzaminacyjnych obowiązujących do 2015 roku obejmują zakres wymagań dla poziomu podstawowego i odnoszą się do umiejętności określonych w *Standardach wymagań egzaminacyjnych* (2007). Arkusz egzaminacyjny stosowany na poziomie podstawowym składa się z trzech części różniących się w zakresie konstrukcyjnym formą zadań i długością skali punktowania. Łącznie za poprawne rozwiązanie wszystkich zadań zdający może uzyskać 50 punktów. Pierwsza grupa zawiera od 20 do 30 zadań zamkniętych wielokrotnego wyboru z jedną poprawną odpowiedzią. Zadania te są punktowane na skali 0–1. Druga grupa składa się z 5 do 10 zadań otwartych krótkiej odpowiedzi punktowanych w skali 0–2. Trzecią grupę stanowią zadania otwarte rozszerzonej odpowiedzi (od 3 do 5 zadań). Za zadania z tej grupy zdający może maksymalnie uzyskać od 4 do 6 punktów.

Egzamin maturalny z matematyki na poziomie rozszerzonym (jako przedmiot do wyboru) trwa 180 minut. Zadania egzaminacyjne obejmują zakres wymagań przedstawiony w *Standardach wymagań egzaminacyjnych* (2007) dla poziomu rozszerzonego. Sprawdzają one między innymi umiejętność budowania modelu matematycznego dla sytuacji problemowej przedstawionej w zadaniu, tworzenia strategii rozwiązania problemu, posługiwania się sformalizowanym językiem matematycznym do opisu własnego rozumowania oraz umiejętność wnioskowania i dowodzenia matematycznego.

Arkusz egzaminacyjny dla poziomu rozszerzonego składa się z zadań otwartych rozszerzonej odpowiedzi. Zwykle zawiera on od 10 do 12 zadań, za które zdający może maksymalnie uzyskać od 3 do 6 punktów. Łącznie za poprawne rozwiązanie wszystkich zadań na egzaminie z matematyki na poziomie rozszerzonym zdający może uzyskać 50 punktów. Arkusze egzaminacyjne egzaminu maturalnego z matematyki, którego wyniki zostały włączone do badań, są dostępne na stronie CKE pod adresem <http://www.cke.edu.pl/index.php/2011-em> oraz <http://www.cke.edu.pl/index.php/2012-em/maj-2012-r>. Parametry psychometryczne zadań z tych dwóch arkuszy przedstawione są w aneksie 2.

4. Organizacja oceniania w polskim systemie egzaminów zewnętrznych

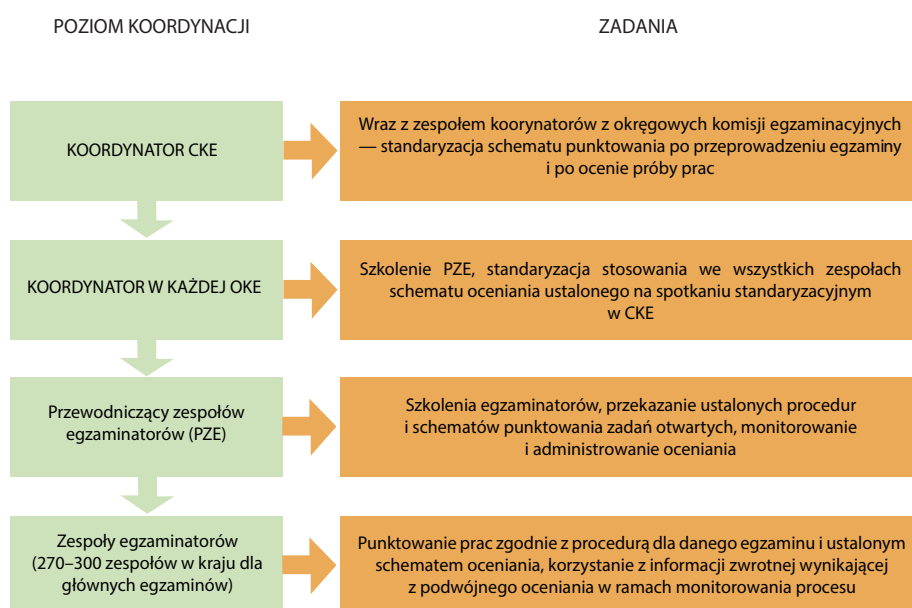
Po przeprowadzeniu egzaminu w ośrodku egzaminacyjnym, którym najczęściej jest szkoła macierzysta zdającego (z wyjątkiem praktycznej części egzaminu potwierdzającego kwalifikacje zawodowe), prace egzaminacyjne zdających przekazywane są bezpośrednio lub pośrednio ze szkół do ośrodków oceniania. W niektórych komisjach (np. OKE Wrocław) prace najpierw są przekazywane do OKE i dopiero po zarejestrowaniu rozdzielane są do ośrodków oceniania, w których zespoły egzaminatorów sprawdzają i oceniają zadania otwarte zgodnie z ustalonymi centralnie zasadami. Zadania zamknięte są punktowane automatycznie poprzez wprowadzenie do systemu danych egzaminacyjnych zeskanowanych kart odpowiedzi oraz „maski” określającej wzorzec poprawnych odpowiedzi.

Każdy egzaminator pracuje w zespole egzaminatorów i po przeszkoleniu przez przewodniczącego zespołu egzaminatorów (PZE) otrzymuje ustaloną liczbę prac egzaminacyjnych (50–80 zależnie od egzaminu) i ocenia je, posługując się kryteriami i schematem punktowania zadań. Przez szereg lat dominowała strategia oceniania kryterialnego. W ostatnich latach w niektórych egzaminach praktykowane jest ocenianie holistyczne⁶ (np. egzamin maturalny z matematyki, egzamin gimnazjalny z języka polskiego). Jakość punktowania zadań monitorowana jest na podstawie podwójnego oceniania 10–15% prac, którego zgodność jest weryfikowana przez przewodniczącego zespołu egzaminatorów (PZE).

W celu zapewnienia porównywalności oceniania w skali kraju proces oceniania jest koordynowany centralnie przez CKE, a na terenie właściwej OKE przez koordynatora oceniania danego egzaminu w tej komisji. Bezpośrednio po zakończeniu egzaminu odbywa się w CKE spotkanie standaryzacyjne schematu punktowania z wykorzystaniem autorskiego schematu zaproponowanego przez autorów arkusza egzaminacyjnego oraz wyników oceny próby prac egzaminacyjnych dostarczonej przez koordynatorów z poszczególnych OKE. Rysunek 4.1 przedstawia schematycznie strukturę koordynacji oceniania.

Dla zapewnienia szybkiego przepływu informacji pomiędzy poziomami koordynacji, a szczególnie pomiędzy koordynatorem OKE i przewodniczącym zespołu egzaminatorów (PZE), w trakcie procesu oceniania wykorzystywane są platforma MOODLE i połączenia telefoniczne.

Rysunek 4.1. Schemat koordynacji i zapewniania jakości oceniania



⁶ Pomimo że to ocenianie nazywane jest przez CKE holistycznym, to właściwie ma ono charakter mieszany – hybrydowy.

4.1. Specyfika oceniania zadań otwartych rozszerzonej odpowiedzi z języka polskiego

W przypadku egzaminu z języka polskiego wszystkie zadania, zarówno sprawdzające rozumienie czytanego tekstu, jak i umiejętność pisania na podstawie tekstu literackiego zamieszczonego w arkuszu egzaminacyjnym, oceniane są przez zewnętrznych egzaminatorów zatrudnianych przez okręgowe komisje egzaminacyjne. Ocenianie odbywa się w zespołach egzaminatorów (ZE) pod kierownictwem przewodniczących zespołów egzaminatorów (PZE). Przed przystąpieniem do oceniania przewodniczący przeprowadzają szkolenie egzaminatorów do oceniania arkuszy egzaminacyjnych z danej sesji, podczas którego oceniają przykładowe prace z wypracowaniem dla pierwszego i drugiego tematu. Wcześniej PZE są szkoleni przez koordynatora oceniania w każdej OKE. Podczas tego szkolenia przekazywane są PZE ustalenia i materiały przygotowane podczas spotkania koordynatorów (w poszerzonym składzie) w CKE, mającego na celu osiągnięcie zgodności kryteriów oceniania. W spotkaniu koordynatorów trwającym trzy dni w 2014 roku uczestniczyli przedstawiciele CKE, IBE i przedstawiciele okręgowych komisji egzaminacyjnych (koordynatorzy i ich zastępcy z OKE o zasięgu większym niż jedno województwo). Na spotkanie do CKE koordynatorzy przywożą po 5 prac (zarówno na poziomie podstawowym, jak i rozszerzonym). Po wstępnej analizie z 40 prac wybierane jest 10 do dalszych prac w zespole. Podczas pracy nad uzgadnianiem kryteriów oceniania wykorzystywane są autorskie propozycje kryteriów przygotowane w procesie tworzenia arkusza egzaminacyjnego oraz wyniki oceniania próby 10 prac ocenionych przez wszystkich koordynatorów z OKE uczestniczących w spotkaniu. Rezultatem spotkania jest protokół podpisany przez wszystkich uczestników spotkania zawierający:

- przykładowe wypracowania i ich oceny
- kryteria oceniania wraz z punktacją dla testu i wypracowania
- modele odpowiedzi
- materiały dla egzaminatora:
 - informacje dla egzaminatora oceniającego prace uczniów ze stwierdzonymi dysfunkcjami
 - wypracowania do oceny przez egzaminatora podczas szkolenia
 - materiał ćwiczeniowy dla egzaminatorów.

Ocenianie odbywa się w weekendy (od 1 do 2). W trakcie sesji oceniania dziesięć procent wybranych losowo prac ocenianych jest przez drugiego egzaminatora. W przypadku różnicy ocen pierwszego i drugiego egzaminatora decyzję o ostatecznej ocenie podejmuje PZE po dodatkowej ocenie pracy. W niektórych OKE PZE ocenia także prace progowe (na granicy zdawalności) i prace z maksymalną liczbą punktów, które nie mieszczą się w ustalonym schemacie oceniania. W tym zakresie OKE stosują różne rozwiązania – nie wypracowano wspólnych szczegółowych uregulowań.

Każdy koordynator nadzoruje ocenianie w swojej OKE, kontaktując się z przewodniczącymi zespołów egzaminatorów telefonicznie oraz poprzez platformę MOODLE. Platforma MOODLE służy także do komunikacji koordynatorów z wszystkich OKE oraz koordynatora krajowego.

4.2. Specyfika oceniania zadań otwartych rozszerzonej odpowiedzi z matematyki

Arkusz egzaminacyjny stosowany w egzaminie na poziomie podstawowym zawiera zarówno zadania zamknięte, jak i zadania otwarte (krótkiej i rozszerzonej odpowiedzi). Zadania zamknięte są oceniane automatycznie po skanowaniu, natomiast zadania otwarte są oceniane przez egzaminatorów. Egzamin w części rozszerzonej składa się tylko z zadań otwartych ocenianych przez egzaminatorów. Podobnie jak w przypadku języka polskiego, ocenianie odbywa się w zespołach koordynowanych przez PZE, a koordynacja całej sesji oceniania w każdej OKE prowadzona jest przez koordynatora z wykorzystaniem platformy MOODLE. Podobnie jak w przypadku egzaminu z języka polskiego, koordynator krajowy prowadzi forum dla koordynatorów z OKE. Podstawowe różnice pomiędzy ocenianiem prac z języka polskiego i matematyki polegają na sposobie organizacji pracy w zespołach. Obecnie większość OKE stosuje specjalizację egzaminatorów w ocenianiu zadań. Oznacza to, że egzaminator w danej sesji nie ocenia całej pracy, tylko przydzieloną grupę zadań. Stosuje się także inne rozwiązania podczas procedury podwójnego oceniania. Funkcjonują trzy modele. Pierwszy model (podobnie jak w innych egzaminach) polega na ocenianiu losowo wybranych prac. Drugie rozwiązanie polega na ocenianiu prac z wynikiem znajdującym się na progu zdawalności. Trzecie rozwiązanie polega na podwójnym ocenianiu

prac z wynikami maksymalnymi. Trzeci model stosowany jest np. w OKE Kraków. Modele te (z wyjątkiem drugiego) są stosowane zarówno dla egzaminu na poziomie podstawowym, jak i na poziomie rozszerzonym. Jeżeli dla egzaminu na poziomie podstawowym podwójne ocenianie realizowane jest dla prac z wynikiem progowym lub maksymalnym, to takie metody w niektórych rozwiązaniach organizacyjnych wymagają od egzaminatorów także oceny zadań zamkniętych, w celu selekcji prac z progowym lub maksymalnym wynikiem.

Jeżeli chodzi o ustalanie kryteriów i schematu oceniania na poziomie krajowym, zastosowane jest podobne rozwiązanie jak w języku polskim. W panelu koordynacyjnym w CKE spotykają się koordynatorzy OKE, koordynator CKE oraz dwóch ekspertów, którzy reprezentują środowisko akademickie (często są to recenzenci arkuszy egzaminacyjnych). W przypadku matematyki w trakcie panelu oceniane jest przeszło 200 prac z obszaru wszystkich OKE. W tym ocenianiu biorą udział wszyscy uczestnicy panelu. W ciągu trzech dni przygotowywane są finalne kryteria oraz materiały pomocnicze dla PZE i egzaminatorów z wykorzystaniem autorskich kryteriów sformułowanych na etapie tworzenia arkuszy oraz danych empirycznych z oceniania próby prac. Organizacja szkolenia PZE przed rozpoczęciem prac zespołów jest także zróżnicowana pomiędzy okręgowymi komisjami egzaminacyjnymi. Stosuje się następujące rozwiązania: bezpośrednie spotkania z koordynatorem, wykorzystanie e-learningu, platformy MOODLE oraz z zastosowaniem szkoleń z wykorzystaniem interaktywnej telewizji.

4.3. Egzaminator podczas procesu oceniania

Egzaminator zatrudniany podczas sesji oceniania prac egzaminacyjnych lub prac wykorzystywanych w badaniach edukacyjnych to osoba oceniająca i kodująca odpowiedzi na zadania otwarte. Egzaminator podczas pracy posługuje się schematem oceniania (zawierającym kryteria oceniania), którego nieodłączną częścią jest skala (określona w modelu oceniania). Na tej podstawie przypisuje on egzaminowanemu pozycje na skali w zależności od ich poziomu cechy (umiejętności), która jest przedmiotem pomiaru. Pozycje te, jak podkreśla Popham (1990), mogą być obciążone trzema potencjalnymi źródłami błędów, które powinny być monitorowane podczas procesu kodowania wyników:

1. schemat oceniania
2. procedura oceniania
3. osoba kodującego odpowiedzi w zadaniach otwartych (czyli egzaminatora).

Podczas trwającej często wiele godzin procedury oceniania egzaminator wykonuje bardzo złożoną pracę, podatną na różnorakie źródła błędów (potencjalnie związanych ze schematem oceniania i jego procedurą). Thorndike i Hagen (1977) podkreślają, że proces przypisania badanemu (egzaminowanemu) odpowiedniej pozycji na skali oceniania za rozwiązanie danego zadania przypomina model „czarnej skrzynki”. Znana jest sytuacja wejściowa – rezultat interakcji badanego z treścią zadania i wyjściowa – pozycja na skali za dane zadanie, jaką badanemu przypisał egzaminator. Procesy poznawcze egzaminatora towarzyszące kodowaniu odpowiedzi za zadania otwarte rozszerzonej odpowiedzi w złożony sposób zależne są od wielu czynników, wśród których możemy między innymi wymienić:

1. przeszłe doświadczenia nauczycielskie
2. przeszłe doświadczenie jako egzaminatora
3. przygotowanie merytoryczne w zakresie dziedziny, z której kompetencje poddane są egzaminowaniu (badaniu)
4. motywacja
5. poziom zmęczenia
6. cechy osobowościowe
7. interakcja z przewodniczącym zespołu egzaminatorów i innymi członkami zespołu.

Egzaminator zwykle ocenia w określonych warunkach, które także mogą mieć wpływ na rezultat jego pracy. Ma za sobą określone szkolenie (uprawniające go do wykonywania tej pracy) oraz trening przed samym przystąpieniem do oceniania odpowiedzi na zadania otwarte. Pracuje w określonym zespole zapewniającym mu wsparcie. Ponadto musi przestrzegać dyscypliny czasowej podczas oceniania przydzielonej puli prac. Można przypuszczać, że czynniki związane bezpośrednio z procedurami i organizacją oceniania mają bezpośredni lub pośredni (poprzez efekt egzaminatora) wpływ na rezultaty oceniania.

5. Efekt egzaminatora – podstawowe pojęcia

Efekt egzaminatora jest terminem określającym szeroką gamę efektów generujących wariancję ocen niezwiązanych z rzeczywistym poziomem umiejętności ocenianego, ale z cechami egzaminatora (Scullen, Mount i Goff, 2000). Wśród tych efektów można wyróżnić kilka najczęściej badanych i opisywanych (Saal, Downey i Lahey, 1980):

1. efekt halo
2. łagodność i surowość
3. tendencja centralna (ograniczenia skali)
4. zgodność.

Zagadnienia związane z ocenianiem można wyprowadzić z początków XX wieku, z badań Thorndike'a nad efektem halo. Koncepcje wywodzące się z psychologii ewoluowały i znalazły swoje zastosowanie w pomiarze edukacyjnym. Niewątpliwie punktem wyjścia i najistotniejszym elementem prac nad efektem egzaminatora jest założenie, że każdy pomiar obarczony jest błędem, wynikającym także z czynników związanych z osobą eksperta dokonującego oceny pracy.

Przyjrzyjmy się teraz krótko wymienionym powyżej efektom. Z efektem halo mamy do czynienia, gdy oceniający przypisuje ocenianemu ten sam poziom umiejętności w różnych wymiarach (kryteriach) na podstawie ogólnego wrażenia, zamiast oceniać poszczególne wymiary niezależnie. Łagodność i surowość w ocenianiu polegają na systematycznym przypisywaniu wszystkim ocenianym ocen niższych lub wyższych niż odpowiednie do ich poziomu umiejętności. Tendencja centralna, opisywana także jako ograniczenie skali, polega na przypisywaniu oceniaczom ocen blisko środka skali niezależnie od poziomu ich umiejętności. Ze zgodnością mamy z kolei do czynienia, gdy dwóch lub więcej oceniaczy przypisuje tym samym ocenianym takie same oceny. Warto zwrócić uwagę na to, że zgodność pomiędzy różnymi oceniaczami niekoniecznie oznacza przyznawanie ocen odpowiednich do poziomu umiejętności, choć bywa nazywana rzetelnością oceniania.

Kulon (2014) proponuje nieco inną typologię efektów egzaminatora niż zaproponowana wyżej, odnosząc ją do błędów systematycznych i losowych. Interpretacja efektu halo i łagodności/surowości nie uległa zmianie, lecz definicja dwóch pozostałych efektów została poszerzona. Ograniczenie skali jest traktowane jako zawężanie przyznawanych przez egzaminatora ocen w stosunku do całej skali. Zjawisko tendencji centralnej jest natomiast szczególnym przypadkiem ograniczenia skali – wyłącznie do jej środka. Innym rodzajem podobnego typu efektu jest tendencja do używania ocen skrajnych, czyli ekstremizm. Zjawiska te zbiorczo nazywane są zniekształceniami skali. O ile rzetelność oceniania jest miarą błędu losowego, o tyle zgodność łączy w sobie błąd losowy i systematyczny. W celu uniknięcia nieporozumień dotyczących tych terminów proponowane jest używanie terminu *precyzja* w odniesieniu do efektu związanego z błędem losowym egzaminatorów.

Większość badań nad efektem egzaminatora wpisuje się w jeden z trzech nurtów (Wolfe, 2004). Część badań koncentruje się na percepcji oceniaczego oraz związkach pomiędzy jego procesami poznawczymi oraz doświadczeniem. (np. Wolfe, Kao i Ranney, 1998). Badania koncentrują się na różnicach pomiędzy egzaminatorami nowicjuszami oraz egzaminatorami ekspertami, a uzyskane wyniki odzwierciedlają różnice pomiędzy ekspertami i nowicjuszami z innych dziedzin (Glaser, Chi i Farr, 1988). Ten nurt badań wskazuje na to, że egzaminatorzy z dużym doświadczeniem koncentrują się na specyficznych charakterystykach eseju oraz podchodzą holistycznie do całości procesu oceniania. W niniejszym badaniu doświadczenie egzaminatora było również kontrolowane. Kolejny nurt badań koncentruje się na charakterystykach oceniaczych, specyfice ocenianego zadania oraz warunkach oceniania w kontekście występowania efektu egzaminatora (np. Hoyt, 2000). Badania te sugerują, że sposób prezentacji informacji egzaminatorom oraz sposób percepcji informacji mogą indukować specyficzne efekty egzaminatora, związane np. z efektem halo. Inne badania sugerują, że specyficzne rodzaje efektów egzaminatora mają różny wpływ na trafność i rzetelność ocen. W niniejszej analizie kontrolowano zarówno szereg charakterystyk oceniaczego (poprzez badania ankietowe), specyfikę ocenianych zadań (do badania wybrano zadania egzaminacyjne z dwóch różnych przedmiotów, o różnym poziomie trudności) oraz warunki oceniania (np. kolejność oceniania prac oraz czas ich oceniania). Ostatni typ badań koncentruje się na wpływie efektu egzaminatora na wystawiane oceny oraz na rozwijaniu statystycznych procedur w celu modelowania i korekty efektu egzaminatora (np. Vance, Winne i Wright, 1983).

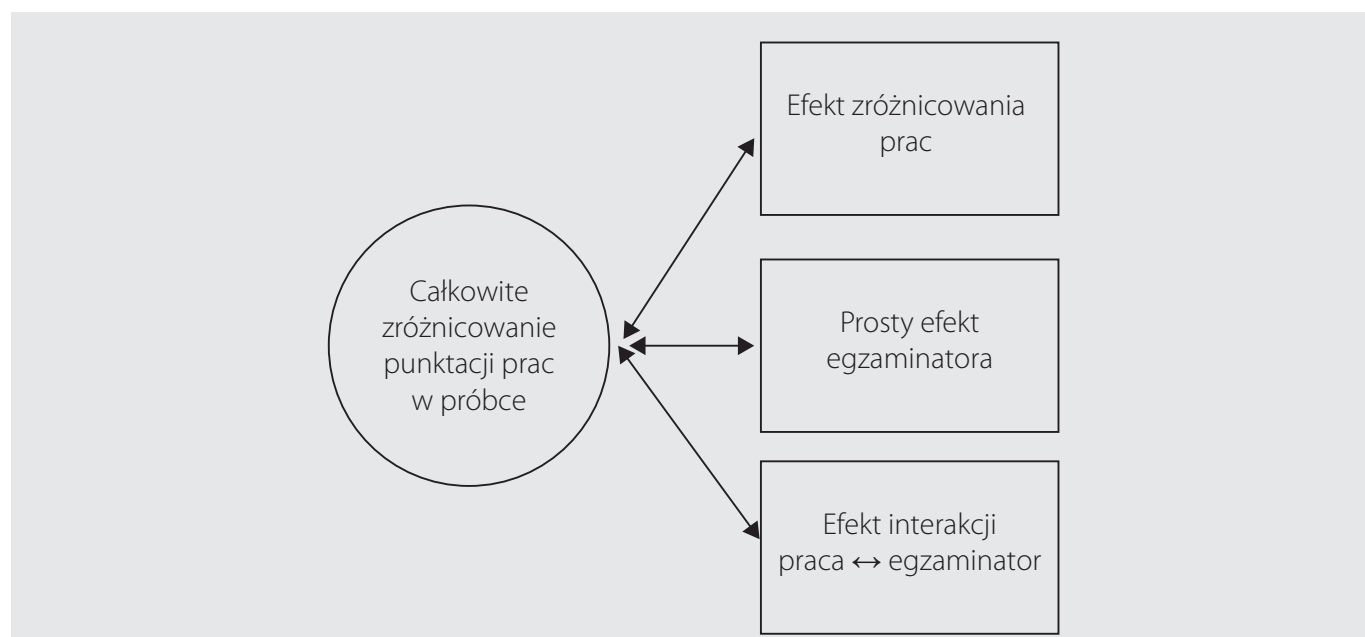
Efekt egzaminatora badano w kontekście tak znanych testów jak *Advanced Placement* (Wolfe, 2004), *TOEFL (Test of English as a Foreign Language)* (Xi i Mollaun, 2009) czy *International English Language Testing System (IELTS)* (Brown, 2007). Efekt egzaminatora bada się także bardzo często w kontekście egzaminów ustnych (np. Kang, 2008; Kim, 2009).

5.1. Badania prowadzone w Polsce

W przypadku polskiego systemu egzaminacyjnego w obrębie jednego egzaminu wszyscy zdający w całym kraju rozwiązują te same zadania. Stosowane są też ten sam schemat oceniania i taka sama procedura organizacyjna. Prace poszczególnych uczniów trafiają jednak do różnych egzaminatorów, co może mieć wpływ na końcową ocenę ucznia. W ocenianiu egzaminów zewnętrznych i w prowadzeniu badań często stosowana jest procedura tzw. podwójnego oceniania, gdzie pewną liczbę prac uczniów ocenia dwóch lub więcej egzaminatorów. Ma to przeważnie na celu bieżące monitorowanie zgodności (rzetelności) oceniania. Kontrolowanie efektu egzaminatora jest niewątpliwie zagadnieniem bardzo istotnym, szczególnie dla egzaminu maturalnego. Dlatego też stało się ono przedmiotem projektu opisywanego w niniejszym opracowaniu.

W Polsce problem efektu egzaminatora podejmowany był kilkakrotnie. Egzamin maturalny badali pod tym kątem Dolata, Putkiewicz i Wiłkomirska (2004), a sprawdzian po szóstej klasie szkoły podstawowej Dubiecka, Szaleniec i Węziak (2006). W latach 2006–2008 Centralna Komisja Egzaminacyjna prowadziła badania monitorujące porównywalność oceniania w trakcie sesji egzaminacyjnej, które obejmowały w pierwszym roku próbę losową zespołów egzaminatorów, a w dwóch kolejnych całą populację ocenających dla wybranych egzaminów (Węziak, 2007). Efekt egzaminatora w egzaminie gimnazjalnym był częściowo badany podczas powołanej przez IBE *Diagnozy Kompetencji Gimnazjalistów* (DKG, 2011; DKG, 2012). W przypadku pierwszego z wymienionych badań (Dolata i in., 2004) głównym celem było określenie jakości systemów punktacji egzaminów maturalnych wykorzystanych w 2002 roku. Badanie objęło 350 prac maturalnych, w tym 150 z języka polskiego oraz po 100 z historii i 100 z matematyki (po 50 dla każdego arkusza egzaminacyjnego) oraz próbę losową 24 egzaminatorów (wybrano po jednym egzaminatorze z każdej okręgowej komisji egzaminacyjnej z trzech przedmiotów egzaminacyjnych – język polski, historia i matematyka). Przyjęty model analizy rzetelności systemów punktacji zakładał sprawdzenie, jaka część całkowitej wariancji punktacji prac w próbie przypada na efekty składowe.

Rysunek 5.1. Model analizy rzetelności systemów punktacji



Źródło: Dolata, Putkiewicz i Wiłkomirska, 2004

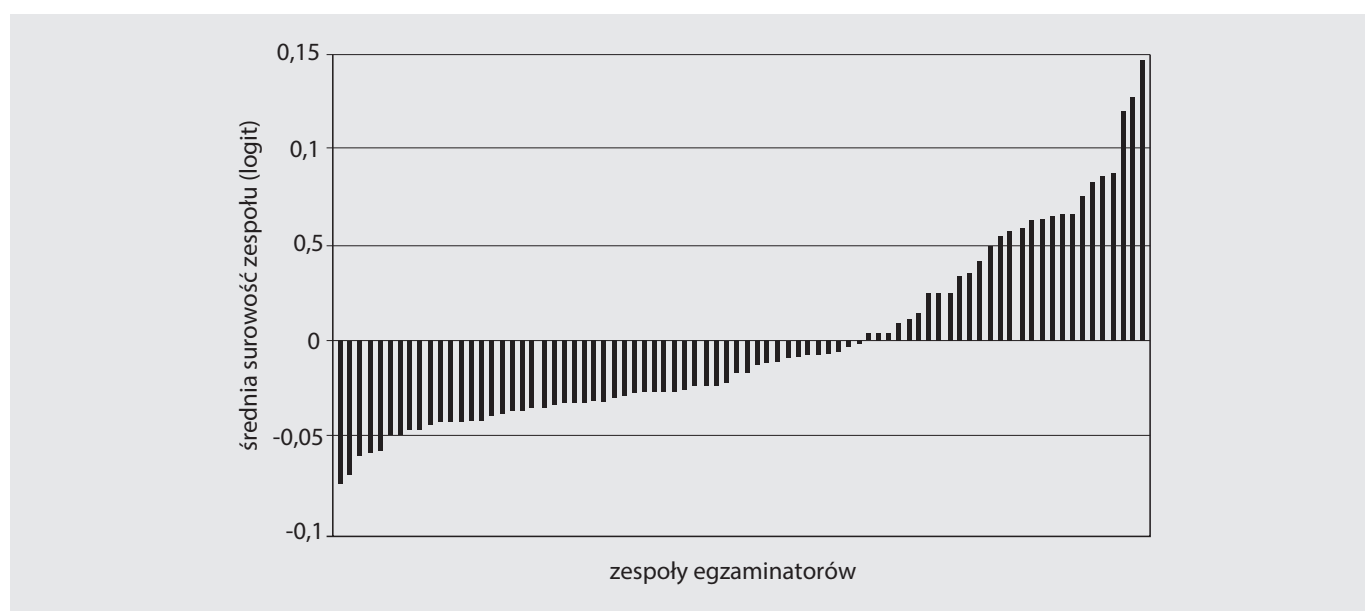
Jak pokazują wyniki badań, w przypadku matematyki (arkusz I poziom podstawowy) efekt zróżnicowania jakości prac wyjaśniał 99 procent wariancji, natomiast prosty efekt egzaminatora pozostały jeden procent. Nie zaobserwowano efektu interakcji praca ↔

egzaminator. W przypadku drugiego arkusza (poziom rozszerzony), efekt zróżnicowania jakości prac wyjaśniał 97 procent wariancji, prosty efekt egzaminatora 1 procent i efekt interakcji praca ↔ egzaminator 2 procent.

Diametralnie inaczej wyglądają wyniki badań odnośnie oceniania prac maturalnych z języka polskiego. Dla pierwszego arkusza na poziomie podstawowym (egzamin maturalny z języka polskiego obejmował 3 arkusze egzaminacyjne) efekt zróżnicowania jakości prac wyjaśniał 55 procent wariancji. Prostý efekt egzaminatora 16 procent i efekt interakcji praca ↔ egzaminator aż 29 procent. Dla arkusza II, w którym sprawdzano rozumienie czytanego tekstu (zadania krótkiej odpowiedzi), efekt zróżnicowania jakości prac wyjaśniał 80 procent wariancji, prostý efekt egzaminatora 7 procent i efekt interakcji 13 procent. Dla arkusza III zastosowanego na poziomie rozszerzonym dla uczniów piszących temat II, efekt zróżnicowania jakości prac wyjaśniał tylko 49 procent wariancji, efekt egzaminatora 22 i efekt interakcji egzaminatora z pracą zdającego 29 procent. Najważniejsze wnioski wynikające z przeprowadzonych prac to wysoka rzetelność oceniania egzaminu z matematyki i bardzo niska rzetelność oceniania egzaminu z języka polskiego – szczególnie wypracowania pisanego na jeden z dwóch wybranych tematów. Wyniki tych badań miały szczególne znaczenie na etapie planowania opisywanego w niniejszym opracowaniu projektu badawczego. Pamiętać jednak należy, że obydwa egzaminy (język polski i matematyka) obecnie różnią się istotnie od egzaminu, jaki zdawali absolwenci w 2002 roku, którzy ten egzamin wybrali pomimo zmiany rozporządzenia przesuującego wprowadzenie nowej matury na 2005 rok (Rozporządzenie MENiS z 2001 r. Dz. U. z 2001 r. Nr 128, poz. 1419).

Drugie ze wspomnianych badań polskich (Dubiecka i in., 2006) dotyczyło ponadprzedmiotowego testu diagnostycznego w szóstej klasie szkoły podstawowej (sprawdzian) i objęło niewielką liczbę prac (15), ale dużą grupę egzaminatorów – 1349 osób z Okręgowej Komisji Egzaminacyjnej w Krakowie (wszyscy egzaminatorzy oceniający w danym roku prace ze sprawdzianu w OKE Kraków). Badanie przeprowadzone było w 2005 roku, tydzień przed przystąpieniem egzaminatorów do oceniania w sesji egzaminacyjnej. Jednym z celów było sprawdzenie, czy istnieje zróżnicowanie efektu egzaminatora (a konkretnie surowości oceniania) pomiędzy zespołami egzaminatorów pracujących pod przewodnictwem PZE. W badaniu zastosowano analizę wariancji, a także wieloaspektowe skalowanie Rascha (*Many-Faceted Rasch Model*). Analizę przeprowadzono, narzucając średnie wartości oszacowań poziomu surowości egzaminatorów równe zero. Oszacowania trudności zadań egzaminacyjnych przyjęto za stałe i znane (oszacowane wcześniej na podstawie wyników egzaminacyjnych krajowej populacji szóstoklasistów). Natomiast oszacowanie poziomu umiejętności uczniów pozostawiono bez narzucania żadnych warunków ograniczających. Zróżnicowanie surowości oceniania w poszczególnych zespołach egzaminatorów okazało się bliskie dwóch odchyłeń standardowych dla efektu egzaminatora.

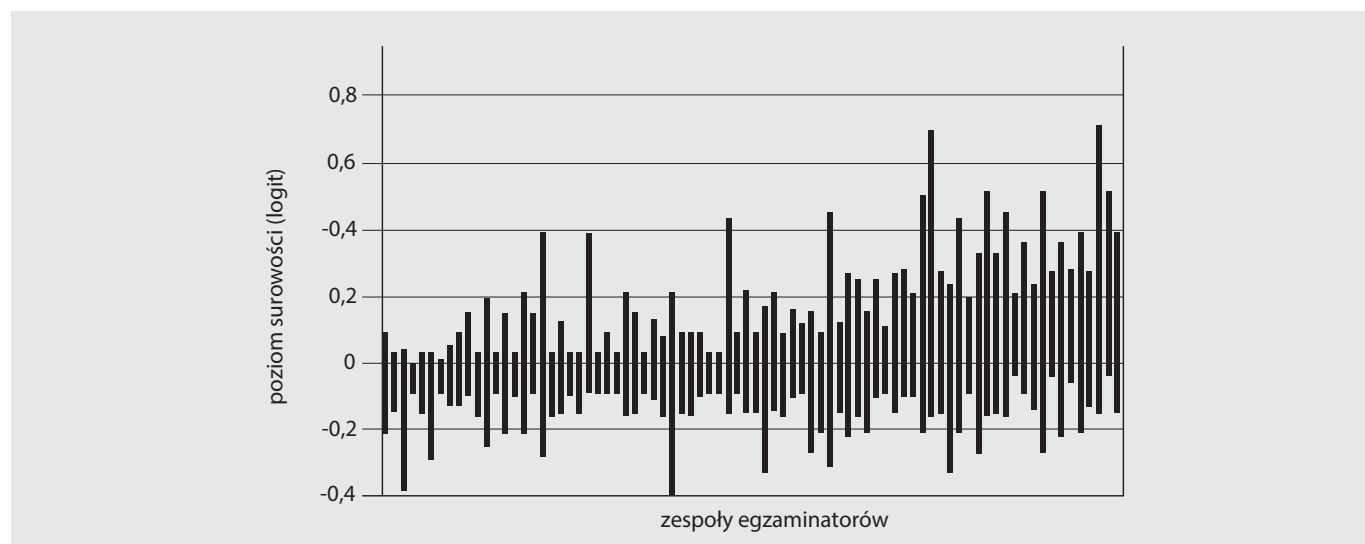
Rysunek 5.2. Średnie poziomy surowości zespołów egzaminatorów sprawdzianu w jednostkach logit przy odchyleniu standardowym 0,117



Źródło: Dubiecka, Szaleniec i Węziak, 2006

Na kolejnym rysunku przedstawione zostały rozstępy surowości w poszczególnych zespołach egzaminatorów.

Rysunek 5.3. Rozstęp poziomu surowości egzaminatorów w poszczególnych zespołach



Źródło: Dubiecka Szaleniec i Węziak D, 2006

Jako najważniejsze wnioski autorzy podają to, że istnieje zróżnicowanie poziomu surowości egzaminatorów nie tylko na poziomie indywidualnym, ale także na poziomie zespołów ocenających czy ośrodka koordynacji oceniania (w jednym ośrodku pracowało kilka zespołów egzaminatorów).

Największe obciążenie efektem egzaminatora zaobserwowano w przypadku zadań ocenianych przez niespecjalistów⁷ (matematyków ocenających zadania polonistyczne i polonistów ocenających zadania matematyczne). Wnioski z tych badań miały istotne znaczenie dla wprowadzenia specjalizacji egzaminatorów sprawdzianu w ocenianiu poszczególnych części prac uczniowskich, a także przyczyniły się do wprowadzenia monitorowania oceniania w skali kraju w latach 2006–2009.

5.2. Monitorowanie oceniania w skali kraju przez CKE w latach 2006–2009

Monitorowanie oceniania z zastosowaniem testu diagnostycznego CKE prowadziła w latach 2006, 2007, 2008 i 2009⁸. Rozwiązanie to miało charakter systemowy i stanowiło integralną część procesu oceniania. W roku 2006 monitorowanie zostało pilotażowo przeprowadzone na próbie 10 procent zespołów egzaminacyjnych dla wybranych 6 egzaminów (sprawdzianu, obydwu części egzaminu gimnazjalnego i egzaminu maturalnego z biologii, geografii i WOS-u). W 2007 monitorowaniem objęta została cała populacja zespołów egzaminacyjnych ocenających prace napisane z zastosowaniem standardowych arkuszy egzaminacyjnych sprawdzianu, egzaminu gimnazjalnego i pięciu egzaminów maturalnych. W 2008 roku monitorowaniem objęto sprawdzian, egzamin gimnazjalny i 12 egzaminów maturalnych. W 2009 roku monitorowanie zostało zawężone tylko do egzaminu gimnazjalnego. Proces monitorowania obejmował 5 grup działań:

1. przygotowanie narzędzi
2. przeprowadzenie testu diagnostycznego dla egzaminatorów – próba pracy polegająca na ocenianiu
3. wprowadzenie i przetworzenie danych
4. przygotowanie informacji zwrotnej
5. wykorzystanie informacji zwrotnej do zarządzania procesem oceniania.

⁷ Do oceniania sprawdzianu w szóstej klasie szkoły podstawowej uprawnienia mogli uzyskać (po ukończeniu kursu dla egzaminatorów) egzaminatorzy uczący różnych przedmiotów.

⁸ W opisie nie uwzględniono monitorowania oceniania w egzaminach potwierdzających kwalifikacje zawodowe.

Przygotowanie narzędzi poprzedzało sesję oceniania i przebiegało równoległe z panelem ustalania kryteriów i schematów oceniania przez koordynatorów OKE i CKE. Pozostałe 4 działania miały miejsce podczas sesji oceniania.

Głównym narzędziem zastosowanym w badaniu był test diagnostyczny złożony z rozwiązanych zadań wybranych z prac egzaminacyjnych z danego roku. Zadaniem egzaminatorów była ocena rozwiązań zgodnie z kryteriami obowiązującymi dla danego egzaminu.

Odpowiedzi i rozwiązania zadań (włączone do testu diagnostycznego) zostały ocenione przez zespół sędziów kompetentnych złożony z koordynatorów OKE i ekspertów CKE, a uzgodnione wyniki zostały przyjęte jako oceny prawdziwe. Oceny te stanowiły układ odniesienia do szacowania liczby błędów popełnianych przez egzaminatorów. Dodatkowym narzędziem była ankieta pozwalająca na zarejestrowanie charakterystyk egzaminatorów ze względu na nauczany przedmiot, stopień awansu zawodowego, liczbę lat pracy w zawodzie, liczbę sesji, w których egzaminator brał udział, pełnioną funkcję w zespole egzaminatorów i wielkość miejscowości, z której pochodził egzaminator.

W 2007 roku badania monitorujące miały charakter systemowego rozwiązania i przeprowadzone zostały dla całej populacji egzaminatorów sprawdzianu, egzaminu gimnazjalnego i dla wybranych egzaminów maturalnych na poziomie rozszerzonym (z biologii, chemii, geografii, historii, matematyki i WOS-u). Celem tych badań było zapewnienie jakości oceniania zadań egzaminacyjnych poprzez monitorowanie, czy prace uczniowskie są oceniane zgodnie z kryteriami i porównywalnie we wszystkich OKE, a także w różnych zespołach wewnątrz poszczególnych komisji egzaminacyjnych. Zidentyfikowanie ewentualnych różnic w czasie rzeczywistym po szkoleniu i przed przystąpieniem do oceniania dawało możliwość dostarczenia przewodniczącym zespołów egzaminacyjnych oraz koordynatorom, a także dyrektorom OKE i CKE informacji zwrotnej o przebiegu procesu sprawdzania, mającej umożliwić zwiększenie rzetelności oceniania.

Dla sprawdzianu i egzaminu gimnazjalnego badania obejmowały dwa etapy. Pierwszy test diagnostyczny zastosowano po zakończeniu szkolenia, a przed przystąpieniem do oceniania, natomiast drugi test diagnostyczny przeprowadzono w połowie czasu oceniania przez egzaminatorów. W przypadku matury w 2007 roku zastosowany był tylko jeden test – po zakończeniu szkolenia, ale przed przystąpieniem do oceniania.

Test diagnostyczny (kalibrujący) przeprowadzany był we wszystkich zespołach, a asystenci PZE wprowadzili oceny przyznane przez poszczególnych egzaminatorów do systemu informatycznego. Dzięki zastosowaniu aplikacji przygotowanej do monitorowania oceniania poszczególni egzaminatorzy, zespoły egzaminatorów, koordynatorzy oceniania na poziomie OKE i CKE otrzymywali oszacowania zgodności oceniania natychmiast po zakończeniu wprowadzania danych. Raporty zawierały zestawienia indywidualne i zagregowane, podsumowujące procent błędów w ocenianiu szacowany w stosunku do ocen prawdziwych za każde kryterium w zadaniu, za każde zadanie i za cały test diagnostyczny. Rezultaty te były wykorzystywane do podejmowania decyzji w czasie rzeczywistym na każdym z poziomów decyzyjnych oceniania. Najistotniejsza była informacja zwrotna dla egzaminatora i PZE dająca możliwość interwencji, np. szkolenie uzupełniające dotyczące kryteriów oceniania dla egzaminatorów popełniających błędy, podwójne ocenianie ich prac, aż do odsunięcia (niedopuszczenia) egzaminatora od dalszego oceniania.

W 2008 roku, podobnie jak w poprzednim, dla sprawdzianu i egzaminu gimnazjalnego monitorowano ocenianie w skali kraju. Zastosowano tylko jeden test diagnostyczny, który był przeprowadzony na zakończenie szkolenia, przed przystąpieniem egzaminatorów do oceniania swojej puli prac. Na poziomie maturalnym rozszerzono liczbę egzaminów do dwunastu. Oprócz przedmiotów monitorowanych w 2007 roku badania dotyczyły także egzaminów z języka polskiego, języków obcych (język angielski, język niemiecki, język francuski i język rosyjski) i fizyki na poziomie podstawowym, a na poziomie rozszerzonym włączono informatykę.

Ogółem na poziomie maturalnym w 2008 roku próba obejmująca całą populację oceniających w sesji majowej liczyła 22489 egzaminatorów pracujących w 1061 zespołach. W grupie tej znalazło się 1053 przewodniczących zespołu egzaminatorów (PZE) oraz 3007 weryfikatorów. Z badań monitorujących w każdym z trzech lat i dla każdego poziomu egzaminu (sprawdzian, egzamin gimnazjalny i maturalny) zostały opracowane raporty do użytku wewnętrznego, które były wykorzystywane w doskonaleniu procesu oceniania.

Wyniki monitorowania z 2007 roku zostały także wykorzystane do analizy efektu egzaminatora dla sprawdzianu, egzaminu gimnazjalnego i egzaminu maturalnego z matematyki na poziomie rozszerzonym. W badaniu zastosowano analizę wariancji, a także

wieloaspektowe skalowanie Rascha. Zgodnie z analizą autorstwa Doroty Węziak (2007) wpływ przynależności egzaminatora do OKE na wynik oceniania był znikomy (wyjaśniał 1,7% wariancji) i był statystycznie nieistotny. Natomiast wpływ przynależności do zespołu oceniania był większy (16,2% wariancji) i nie tylko istotny statystycznie, ale także znaczący z punktu widzenia jakości oceniania.

5.3. Monitorowanie efektu egzaminatora w procesie e-oceniania

Problem efektu egzaminatora był przedmiotem badań w projekcie *Monitorowanie efektu egzaminatora oraz przygotowanie wprowadzenia oceniania z wykorzystaniem narzędzi elektronicznych (ang. electronic marking)* realizowanego przez CKE w ramach grupy projektów *Badania dotyczące podnoszenia jakości narzędzi systemów egzaminów zewnętrznych* w latach 2007–2009 (Szaleniec i Węziak-Białowolska, 2008).

Podczas realizacji projektu przeprowadzono eksperyment dla sprawdzenia, czy przy ocenianiu indywidualnym poza zespołem, a więc w domu egzaminatora, technika oceniania (całymi pracami vs. zadaniami) ma wpływ na poziom surowości oceniania. W eksperymencie wzięło udział łącznie 328 egzaminatorów, reprezentujących wszystkie okręgowe komisje egzaminacyjne, dla czterech egzaminów (sprawdzian w szóstej klasie szkoły podstawowej, egzamin gimnazjalny w części humanistycznej, egzamin gimnazjalny w części matematyczno-przyrodniczej i matura z matematyki na poziomie rozszerzonym). Egzaminatorzy każdego z egzaminów zostali losowo podzieleni na dwie grupy. Egzaminatorzy pierwszej grupy otrzymali zestaw autentycznych prac z egzaminu przygotowanych w formie książki zawierającej kolejne, następujące po sobie prace. Natomiast druga grupa otrzymała zestawy zebrane w książce w ten sposób, że następowały po sobie kolejno rozwiązania tych samych zadań. Każdy z egzaminatorów oceniał tylko jedną techniką. Przeprowadzone analizy obejmowały trzy etapy. W pierwszym etapie oszacowano parametry statystyczne zadań dla każdego z egzaminów stosując model Rascha (1-PL) na dziesięcioprocentowej losowej próbie pobranej z całej populacji, piszącej wymienione egzaminy w wiosennej sesji 2007 r. Wskaźniki poziomu umiejętności zdających zostały oszacowane z wykorzystaniem wskaźników trudności zadań uzyskanych na pierwszym etapie. Do analizy wskaźników poziomu umiejętności uczniów zostały wykorzystane prace włączone do oceniania elektronicznego podczas pilotażu z zastosowaniem oprogramowania e-Marker angielskiej firmy DRS (*Data Services Limited*). W dalszej kolejności dokonano oszacowania surowości egzaminatorów oceniających prace dwoma technikami. Jak przedstawia w raporcie z badań Dorota Węziak-Białowolska (2008), we wszystkich czterech egzaminach zaobserwowano nieznacznie wyższą surowość egzaminatorów oceniających zadaniami w porównaniu z ocenianiem całymi pracami. Różnice te jednak okazały się statystycznie nieistotne. Rezultat ten miał istotne znaczenie dla organizacji oceniania w tradycyjny sposób, jak i dla dalszych prac w kierunku wdrożenia e-oceniania.

Pod koniec 2008 roku odbył się pilotaż e-oceniania w zespołach zorganizowanych w czterech ośrodkach oceniania (sprawdzian – 85 egzaminatorów, egzamin gimnazjalny w części humanistycznej – 81, egzamin gimnazjalny w części matematyczno-przyrodniczej – 82, matura z matematyki na poziomie rozszerzonym – 83) z wykorzystaniem narzędzi elektronicznych. Łącznie w pilotażu uczestniczyło 331 oceniających prace. Pilotaż odbył się z udziałem ekspertów z DRS (*Data Services Limited*) z Wielkiej Brytanii. Głównym celem pilotażu było zapoznanie egzaminatorów wytypowanych przez dyrektorów z ośmiu OKE z nową technologią i opanowanie umiejętności oceniania z wykorzystaniem oprogramowania e-Marker. Dodatkowym celem było zbadanie zróżnicowania efektu egzaminatora w zespołach w obrębie czterech egzaminów. Każdy zespół w obrębie danego egzaminu oceniał tę samą pulę prac egzaminacyjnych, natomiast zadania do oceny były przydzielane egzaminatorom losowo. Do weryfikacji hipotezy o równości średniego poziomu łagodności egzaminatorów w zespołach stosowano analizę wariancji (ANOVA), a w przypadku niespełnienia założenia o równości wariancji w zespołach – test Welcha. Jak raportuje Dorota Węziak-Białowolska (2009), pomimo występowania różnic w poziomach surowości oceniania między zespołami egzaminatorów, dla żadnego z egzaminów różnic tych nie można uznać za statystycznie istotne.

W drugim etapie projektu prowadzonego w latach 2010–2011 *Przygotowanie systemu informatycznego do e-oceniania* został zakupiony od angielskiej firmy RM Education system informatyczny Scoris, w polskiej wersji językowej. Przeprowadzono także pilotaż zastosowania tego oprogramowania dla 4 egzaminów w trakcie jego testowania. Jednocześnie eksperci zostali przygotowani w zakresie zarządzania sesjami e-oceniania. W ramach kolejnego etapu realizowanego w CKE projektu (2012–2013) zatytułowanego *Wdrożenie systemu informatycznego do e-oceniania* zastosowano system informatyczny Scoris, na hostowanych przez CKE serwerach, do powtórnego oceniania dwóch egzaminów dla całej populacji – poprawkowej matury z matematyki w 2012 roku i egzaminu gimnazjalnego z matematyki z sesji głównej 2013 roku (CKE, 2013a). Do e-oceniania zostało przeszkolonych 8926 egzaminatorów (egzamin maturalny z matematyki 3998 osób i egzamin gimnazjalny z matematyki 5066 osób). Zespół wdrożeniowy przygotował również studium wykonalności wdrożenia e-oceniania w latach 2015–2020 (CKE, 2013b).

6. Metodologia i sposób realizacji badania

6.1. Cel ogólny i cele szczegółowe

Ogólnym celem projektu było oszacowanie efektu egzaminatora i innych czynników mających wpływ na wynik w ocenianiu zadań otwartych rozszerzonej odpowiedzi w egzaminach maturalnych z języka polskiego i matematyki.

Na cel główny złożyły się następujące cele szczegółowe:

1. Oszacowanie efektu egzaminatora w ocenianiu prac maturalnych z języka polskiego na poziomie podstawowym (temat I i temat II).
2. Oszacowanie efektu egzaminatora w ocenianiu prac maturalnych z matematyki na poziomie podstawowym i rozszerzonym.
3. Oszacowanie porównywalności oceniania pomiędzy okręgowymi komisjami egzaminacyjnymi.
4. Oszacowanie związku wyniku oceniania z doświadczeniami w ocenianiu egzaminów zewnętrznych (porównanie efektu egzaminatora dla trzech grup o różnym doświadczeniu w tej dziedzinie: studenci ostatniego roku magisterskich studiów polonistycznych, nauczyciele bez uprawnień egzaminatora i egzaminatorzy wpisani do rejestru egzaminatorów i biorący czynny udział w ocenianiu).
5. Oszacowanie związku wyniku oceniania wypracowania z języka polskiego z cechami graficznymi zapisu wypracowania przez zdającego egzamin – charakterem i czytelnością pisma.

6.2. Problemy, pytania badawcze

W badaniu poszukiwano odpowiedzi na następujące pytania:

1. Czy i jak efekt egzaminatora jest zróżnicowany ze względu na okręgowe komisje egzaminacyjne?
2. Czy i jak efekt egzaminatora zależy od czynników związanych z obciążeniem w trakcie oceniania - kolejnością oceniania pracy podczas sesji?
3. Czy i jaki wpływ na efekt egzaminatora ma doświadczenie zawodowe (praca tylko w jednym typie szkoły, doświadczenie w różnych typach szkół)?
4. Czy i jak wynik oceniania zależy od profilu oceniającego (egzaminator wpisany do ewidencji OKE, nauczyciel danego przedmiotu niebędący egzaminatorem OKE, student polonistyki)?
5. Czy i jak wynik oceniania zależy od cech graficznych zapisu wypracowania przez zdającego egzamin – charakteru i czytelności pisma?

6.3. Specyfikacja zmiennych zależnych i niezależnych

Zmienną zależną w badaniach była łagodność/surowość oceniania zadań otwartych w pracach:

1. maturalnych z języka polskiego
2. maturalnych z matematyki.

Zmiennymi niezależnymi były:

1. Organizacja oceniania – przynależność do zespołu koordynowanego przez danego przewodniczącego zespołu egzaminatorów (OKE).
2. Kolejność oceniania danej pracy (sekwencja czasowa).
3. Czas oceniania każdej pracy.
4. Czytelność zapisu wypracowania:

- a. zdefiniowana przez grafologa skala czytelności wypracowania
- b. ocenianie pracy oryginalnej i zapisu komputerowego tej pracy.
5. Cechy deskryptywne egzaminatora:
 - a. doświadczenie w ocenianiu (egzaminator, nauczyciel niebędący egzaminatorem, student)
 - b. typ zatrudnienia (w przypadku nauczycieli): LO, technikum, gimnazjum, szkoła podstawowa, inne
 - c. płeć
 - d. staż w ocenianiu egzaminów zewnętrznych.

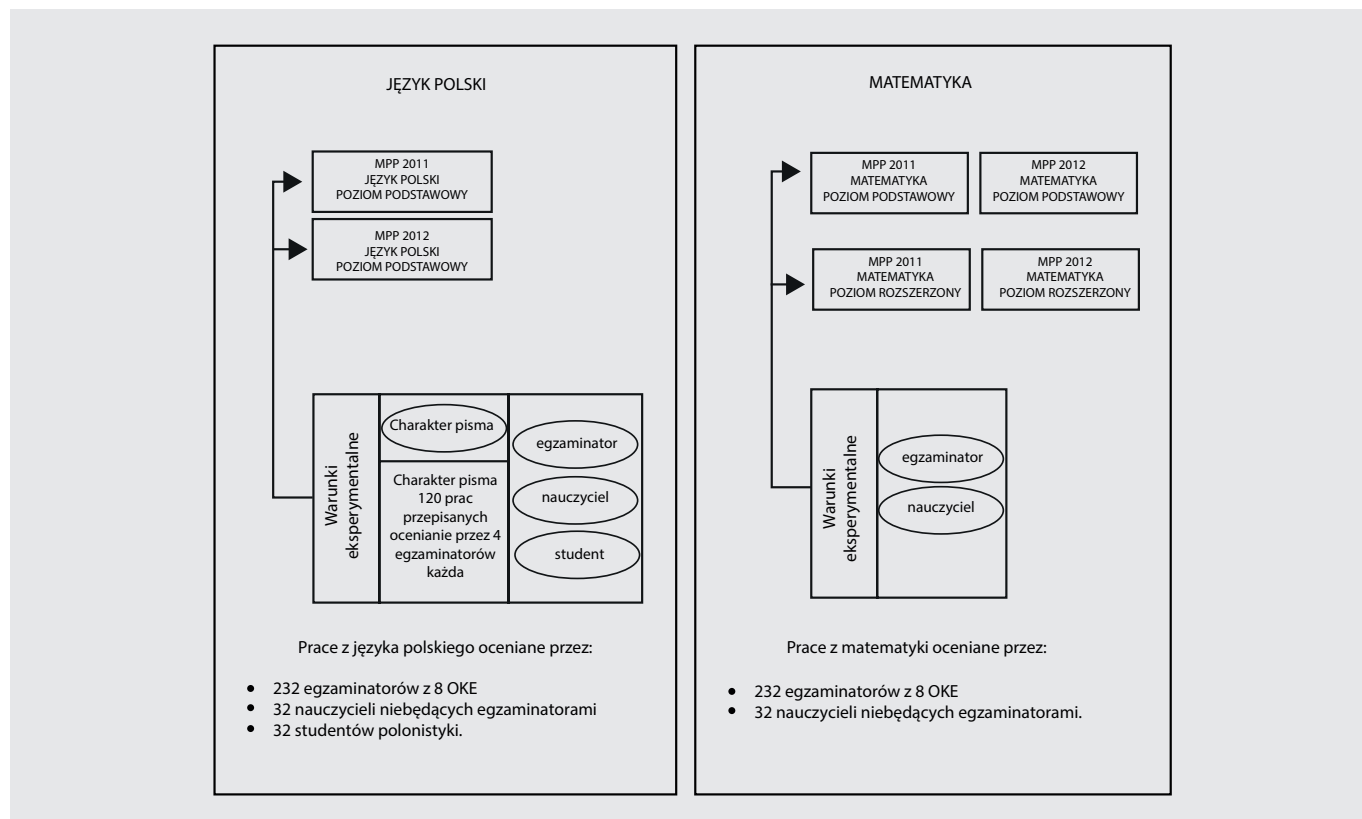
6.4. Schemat planu badawczego

Jak już wcześniej sygnalizowano, studium badawcze obejmowało dwa egzaminy obowiązkowe na maturze. Do analiz włączono prace z dwóch lat (2011 i 2012) z języka polskiego na poziomie podstawowym i z matematyki na poziomie podstawowym i rozszerzonym. Badanie efektu egzaminatora przeprowadzono dla prób egzaminatorów wylosowanych z każdej okręgowej komisji egzaminacyjnej. Ponadto do eksperymentu włączono grupę nauczycieli matematyki i języka polskiego niebędących egzaminatorami oraz studentów ostatniego roku filologii polskiej. W badaniach założono schemat ścisłego przydziału prac do oceny dla konkretnych egzaminatorów i odpowiedniego kotwiczenia pomiędzy różnymi grupami eksperymentalnymi w celu umożliwienia wnioskowania statystycznego z odpowiednią mocą na temat zależności pomiędzy badanymi zmiennymi niezależnymi a wynikami oceniania. Poniżej przedstawiono schematy przydziału prac w poszczególnych częściach badania.

Głównymi zmiennymi eksperymentalnymi były przynależność egzaminatora do OKE oraz kwalifikacje i doświadczenie egzaminatora przyjmujące 3 wartości:

1. egzaminator doświadczony w wyniku uczestniczenia w ocenianiu OKE (egzaminator wpisany do rejestru egzaminatorów)
2. nauczyciel specjalista nauczający danego przedmiotu, ale niebędący egzaminatorem
3. student V roku polonistyki.

Rysunek 6.1. Schemat planu badawczego efektu egzaminatora



6.5. Dobór próby egzaminatorów i prac do oceniania

6.5.1. Dobór próby egzaminatorów

Populacja egzaminatorów z matematyki i języka polskiego egzaminu maturalnego została zdefiniowana jako egzaminatorzy, którzy byli zatrudnieni w OKE do oceniania prac w głównej sesji 2013 roku. Do badania została wylosowana reprezentatywna próba egzaminatorów (29 egzaminatorów z każdej OKE). Podczas losowania egzaminatorzy zostali uporządkowani w każdej OKE alfabetycznie, a następnie z takiej listy w każdej OKE odbywało się losowanie systematyczne z odpowiednim dla każdej komisji interwałem. Łącznie w badaniu wzięło udział 232 egzaminatorów z języka polskiego i matematyki. Próba została rozszerzona o 64 nauczycieli niebędących egzaminatorami (po 32 z języka polskiego i matematyki). Nauczyciele ci zostali wybrani ze szkół wchodzących do próby badań zrównujących⁹ wyniki egzaminu maturalnego z matematyki w 2013 roku oraz z grupy oczekujących na szkolenie w OKE Kraków i OKE Warszawa. Grupa studentów (32 osoby) ostatniego roku polonistyki do badania obejmującego język polski została wybrana z trzech uniwersytetów (Uniwersytet Jagielloński, Uniwersytet Warszawski, Uniwersytet Marii Curie-Skłodowskiej w Lublinie).

Tabela 6.1. Populacja i próby egzaminatorów

OKE	01	02	03	04	05	06	07	08	
Nazwa OKE	Gdańsk	Jaworzno	Kraków	Łomża	Łódź	Poznań	Warszawa	Wrocław	Razem
Próba – język polski	29	29	29	29	29	29	29	29	232
Próba – matematyka	29	29	29	29	29	29	29	29	232
Liczba egzaminatorów z języka polskiego	778	905	1778	618	879	1003	1071	699	7731
Liczba egzaminatorów z matematyki	795	830	1862	620	623	1000	818	745	7293

6.5.2. Dobór prac do oceniania

Do badania efektu egzaminatora z języka polskiego istotne było zgromadzenie zbioru prac posiadających następujące cechy:

1. różnicowanie tematów wypracowań do analizy interakcji egzaminator ↔ temat wypracowania,
2. różnicowanie poziomu umiejętności badanych wypracowaniem w zakresie całej skali zastosowanej w ocenie zadań.

Tabela 6.2. Podział prac ze względu na temat i poziom egzaminu maturalnego z języka polskiego

Rok	Egzamin	Poziom	Liczba prac		
			temat I	temat II	Razem
2011	Język polski	podstawowy	275	167	442
2012		podstawowy	204	251	455
Razem	Język polski		479	418	897

Układ ten daje 4 rodzaje prac egzaminacyjnych: po 2 z każdego roku w podziale na temat.

⁹ Więcej o badaniach zrównujących można przeczytać w: Szaleniec i in. (2013).

Analogiczny zbiór prac dla egzaminu z matematyki musiał charakteryzować się zróżnicowaniem zestawów zadań. Było to konieczne do analizy interakcji egzaminator ↔ zestaw egzaminacyjny z danego roku.

Tabela 6.3. Podział prac ze względu na poziom egzaminu maturalnego z matematyki

Rok	Egzamin	Poziom	Liczba prac
2011	Matematyka	podstawowy	442
		rozszerzony	450
2012		podstawowy	455
		rozszerzony	447
Razem	Matematyka		1794

Do zapewnienia pracom egzaminacyjnym wymienionych cech konieczna była wcześniejsza analiza wyników i powarstwowanie prac ze względu na ich pożądane cechy.

Prace z 2011 roku (język polski i matematyka na poziomie maturalnym) zostały wybrane z OKE Jaworzno, ze względu na ułatwiony do nich dostęp. Prace do badania wylosowane zostały z baz Centralnej Komisji Egzaminacyjnej zawierających wyniki egzaminów maturalnych dla Okręgowej Komisji Egzaminacyjnej w Jaworznie. Ponieważ w bazach CKE wyniki są zanonimizowane i posiadają unikalny kod wygenerowany w OKE do archiwizacji danych w CKE, to do fizycznego wyboru prac konieczne było zidentyfikowanie tych prac zgodnie z kodami dostępnymi w archiwum OKE.

Przed losowaniem z baz danych usunięto wszystkie rekordy oznaczone jako „laureat”. Takie rekordy oznaczają laureatów olimpiad zwolnionych z egzaminu maturalnego. W bazie danych CKE laureaci otrzymują maksymalną liczbę punktów za każde z zadań.

Prace podzielono na warstwy ze względu na wynik sumaryczny w zadaniach otwartych. Przy czym przyjęto, iż jedna warstwa nie może liczyć mniej niż 50 prac. Takie założenie skutkowało łączeniem skrajnych kategorii punktowych, aby uzyskać wystarczająco liczne warstwy. Na przykład w języku polskim na poziomie podstawowym maksymalna liczba uzyskanych punktów za zadania otwarte wynosiła w OKE Jaworzno 52 pkt (włączając kryterium „szczególne walory”). Jednak 52 pkt uzyskały tylko 2 osoby, 51 pkt tylko 12 osób, a 50 pkt 54 osoby. Tym samym trzy kategorie punktowe: 52, 51 i 50, zostały połączone w jedną warstwę liczącą 68 prac.

Wewnątrz każdej warstwy prace dobierano, używając losowania prostego z równym prawdopodobieństwem wyboru. Liczebność wylosowanych prac ustalono na około 500. Końcowa liczba wylosowanych prac z każdego egzaminu zależała od liczby warstw i ustalonej liczby prac przypadających na warstwę. W tabeli 6.4 przedstawiono liczebność warstw, liczbę prac przypadającą na warstwę i finalną liczbę prac dla każdego przedmiotu z osobna.

Tabela 6.4. Liczebność warstw oraz wylosowanych prac z 2011 roku

Przedmiot	Matematyka poziom podstawowy	Matematyka poziom rozszerzony	Język polski poziom podstawowy	Język polski poziom rozszerzony
Liczba warstw	28	51	51	28
Liczba prac w warstwie	18	10	10	19
Liczba prac razem	504	510	510	532

Prace z 2012 roku zostały wylosowane z OKE Kraków w taki sam sposób jak zostały wylosowane prace do badań z 2011 roku z OKE Jaworzno.

6.6. Opis modelu linkowania prac i egzaminatorów

Projektując schemat przydziału prac do wielokrotnej oceny przez losowe próby egzaminatorów z 8 OKE, kierowano się postawionymi problemami badawczymi. W szczególności istotnymi przesłankami były: (I) konieczność możliwie rzetelnego oszacowania prawdziwej oceny danego zadania (rzetelność oszacowania rośnie wraz z krotnością powtórnych ocen) oraz (II) chęć porównania efektów egzaminatora zagregowanych na poziomie ośmiu OKE. Zestawienie tych dwóch przesłanek doprowadziło do poszukiwania schematu przydziału prac, który maksymalizowałby wariancję ocen tego samego zadania między OKE, w którym z każdego OKE dobrano by po około 30 egzaminatorów oraz w ramach którego poszczególne OKE funkcjonowałyby w jak najbardziej symetryczny sposób.

Opis podziału prac pomiędzy egzaminatorów OKE zastosowany w badaniach rozpoczniemy od schematu zastosowanego dla matury z języka polskiego (schemat dla matury z matematyki był jedynie nieznaczną modyfikacją tego rozwiązania). Następnie omówione zostanie rozproszanie prac do powtórnej oceny w grupie nauczycieli niebędących egzaminatorami i studentów. Przedstawienie każdego z wymienionych schematów rozdziału prac rozpocznie jego opisowa definicja, wyjaśniająca jego właściwości istotne w kontekście oceny jakości zebranych danych oraz interpretacji wyników późniejszych analiz. W dalszej kolejności będzie przedstawiony bardziej szczegółowy, techniczny opis konstrukcji każdego schematu rozdziału prac.

6.6.1. Schemat podziału prac dla matury z języka polskiego

Dla prac maturalnych z języka polskiego ocenianych przez egzaminatorów OKE przyjęto w podstawowym schemacie rozdziału prac, że:

- 1. każda praca będzie oceniana 8 razy, przy czym każda ocena będzie pochodziła od egzaminatora z innej OKE;
- 2. z każdej OKE do udziału w badaniach będzie wylosowanych 29 egzaminatorów, co daje łączną liczbę 232 egzaminatorów;
- 3. każdy egzaminator oceni 29 prac, przy czym pozostałe 7 ocen tych 29 prac zostanie przeprowadzone przez wszystkich 203 (7x29) egzaminatorów z pozostałych OKE. Innymi słowy, każdy egzaminator z dowolnej OKE poprzez ocenę 29 prac będzie miał jedną wspólną ocenę z każdym z 29 egzaminatorów z każdej innej OKE.

Opisany powyżej schemat rozdziału prac skutkuje łączną liczbą 841 (29x29) unikalnych prac ocenionych jednokrotnie w każdej OKE w taki sposób, aby każda para egzaminatorów z dwóch różnych OKE miała ze sobą dokładnie jedną wspólną pracę. Poniżej zostanie opisana konstrukcja schematu podziału prac między egzaminatorów spełniająca takie założenia.

Założmy, że przydział 841 prac między 29 egzaminatorów w pierwszej OKE (OKE_1) jest dany macierzą przedstawioną w tabeli 6.5.

Tabela 6.5. Macierz OKE_1 przedstawiająca rozproszanie 841 prac z języka polskiego w pierwszej OKE

indeks oceny (o) \ indeks egz. (e)								
		0	1	2	...	26	27	28
0		1	2	3	...	27	28	29
1		30	31	32	...	56	57	58
2		59	60	61	...	85	86	87
...		...	...	...	...	...	...	...
26		755	756	757	...	781	782	783
27		784	785	786	...	810	811	812
28		813	814	815	...	839	840	841

Ze względu na wykorzystanie operacji modulo w kolejnych krokach założmy dla ułatwienia, że wiersze i kolumny tej macierzy są indeksowane począwszy od 0 aż do 28. Wprowadźmy teraz następujące oznaczenia:

1. $OKE_1, OKE_2, \dots, OKE_8$ – analogiczne do OKE_1 macierze 29×29 przedstawiające przyporządkowanie prac egzaminatorom dla kolejnych siedmiu OKE;
2. $v = (0, 1, \dots, 27, 28)^T$ – kolumnowy wektor indeksów od 0 do 28;
3. $c(v, k, n)$ – przesunięcie cykliczne w lewo o k pozycji modulo n indeksów zebranych w kolumnowym wektorze v , np.:
 $c(v, k, n) = (\text{mod}(0+3, 29), \text{mod}(1+3, 29), \dots, \text{mod}(27+3, 29), \text{mod}(28+3, 29))^T =$
 $= (3, 4, 5, \dots, 28, 0, 1, 2)^T$.

Macierz OKE_2 powstaje poprzez transpozycję macierzy OKE_1 :

$$OKE_2 = (OKE_1)^T.$$

Macierze $OKE_3, \dots, OKE_8$ zostały natomiast skonstruowane poprzez następujące przesunięcia cykliczne kolumn macierzy:

$$OKE_x[v, o] = OKE_{x-1}[c(v, o, 29), o],$$

gdzie $x \in \{3, 4, \dots, 8\}$ jest indeksem OKE natomiast $o \in \{0, 1, 2, \dots, 27, 28\}$ jest indeksem wskazującym kolumnę ocenianej pracy w macierzy OKE_x . Powyższe można zapisać również nierekurencyjnie, poprzez odwołanie się do macierzy OKE_2 :

$$OKE_x[v, p] = OKE_2[c(v, o(x-2), 29), o].$$

Wartości $o(x-2)$ przesunięć cyklicznych na kolumnach macierzy OKE_2 w zależności od indeksu OKE oraz kolumny ocenianej pracy zebrano w tabeli 6.6.

Tabela 6.6. Wartości przesunięcia cyklicznego w kolumnie o w macierzach OKE_x względem macierzy OKE_2

$x \backslash o$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
3	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
4	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	1	3	5	7	9	11	13	15	17	19	21	23	25	27
5	0	3	6	9	12	15	18	21	24	27	1	4	7	10	13	16	19	22	25	28	2	5	8	11	14	17	20	23	26
6	0	4	8	12	16	20	24	28	3	7	11	15	19	23	27	2	6	10	14	18	22	26	1	5	9	13	17	21	25
7	0	5	10	15	20	25	1	6	11	16	21	26	2	7	12	17	22	27	3	8	13	18	23	28	4	9	14	19	24
8	0	6	12	18	24	1	7	13	19	25	2	8	14	20	26	3	9	15	21	27	4	10	16	22	28	5	11	17	23

Wzór na postać OKE_x można także sformułować dla pojedynczego elementu macierzy:

$$OKE_x[e, o] = OKE_{x-2}[\text{mod}(e + o(x-2), 29), o],$$

gdzie $e \in \{0, 1, 2, \dots, 27, 28\}$ jest indeksem wskazującym wiersz egzaminatora w macierzy OKE_x . Jeżeli uwzględnimy jeszcze numery prac nadane w OKE_1 (tabela A1) oraz że $OKE_2 = (OKE_1)^T$, to w jawny sposób możemy przypisać numer pracy dla dowolnego $OKE_x[e, o]$, gdzie $x \in \{3, 4, \dots, 8\}$:

$$OKE_x[e, o] = 29o + \text{mod}(e + o(x-2), 29).$$

6.6.1.1. Przypisanie dodatkowych linkujących egzaminatorów wewnątrz OKE oraz prac przepisanych komputerowo

Powyższy schemat rozdziału prac rozdziela optymalnie 8 ocen każdej pracy pomiędzy OKE. Aby zebrać również informację o zgodności ocen egzaminatorów wewnątrz tych samych OKE, ów schemat został uzupełniony o przypisanie każdemu egzaminatorowi dodatkowych dwóch prac, które były oceniane przez egzaminatorów wewnątrz tej samej OKE. Wewnątrz każdej OKE oceniono dodatkowe 5 prac, które były oceniane ośmiokrotnie oraz 2 prace oceniane dziewięciokrotnie w taki sposób, aby każdy egzaminator wewnątrz pewnej OKE miał jedną wspólnie ocenianą pracę z 14–16 innymi egzaminatorami z tej samej OKE. Oznaczało to dołączenie dodatkowych 56 (7x8) prac wielokrotnie sprawdzanych wewnątrz OKE do wcześniejszego schematu dla 841 prac wielokrotnie ocenianych pomiędzy OKE, zatem w sumie wielokrotnej ocenie poddano w badaniu 897 prac. Jeżeli 7 prac przeznaczonych do wielokrotnej oceny wewnątrz OKE oznaczmy kolejnymi literami alfabetu A, B, ..., F, G, to schemat przydziału tych prac egzaminatorom w danym OKE przedstawia tabela 6.7.

Tabela 6.7. Macierz ilustrująca rozprowadzenie ocen dodatkowych 7 prac z języka polskiego wewnątrz OKE

$\begin{matrix} & e \\ o \end{matrix}$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
29	A	B	C	D	E	F	G	A	C	E	G	B	D	F	A	D	G	C	F	B	E	A	E	B	F	C	G	D	F
30	B	C	D	E	F	A	A	C	E	G	B	D	F	G	D	G	C	F	B	E	A	E	B	F	C	G	D	A	G

Do badania efektu egzaminatora dla egzaminu maturalnego z języka polskiego włączono również dodatkowe oceny 120 przepisanych komputerowo prac wybranych przez grafologa spośród 841 prac znajdujących się w schemacie 29x29. Wspomnianych 120 prac oprócz ośmiu ocen formy oryginalnej (pismo odręczne) przeprowadzonych między OKE zostało dodatkowo 4 razy ocenione w formie zapisu komputerowego. Przyporządkowanie tych 4x120 dodatkowych ocen przepisanych prac 232 egzaminatorom uczestniczącym w badaniach zostało przeprowadzone losowo, przy nałożeniu ograniczenia, aby żaden egzaminator nie oceniał tej samej pracy w formie przepisanej komputerowo i nieprzepisanej. 216 egzaminatorów oceniało po 2 prace przepisane komputerowo (łącznie 32 prace), 16 egzaminatorów oceniało po 3 prace przepisane komputerowo (łącznie 33 prace).

6.6.2. Schemat podziału prac dla matury z matematyki

Schemat przyporządkowania prac maturalnych egzaminatorom matematyki z ośmiu OKE opierał się na tych samych 8 macierzach 29x29 z tą różnicą, że każda praca była oceniana czterokrotnie, a nie ośmiokrotnie. Sprawdzaniu poddano 841 unikalnych prac z poziomu podstawowego oraz 841 unikalnych prac z poziomu rozszerzonego. Zastosowanie opisanego wcześniej schematu zostało dokonane poprzez losowe rozdzielanie każdych 8 ocen pracy z języka polskiego na 4 oceny pracy z matematyki na poziomie rozszerzonym oraz 4 oceny pracy z matematyki na poziomie podstawowym.

Dodatkowe oceny dwóch prac wewnątrz OKE przeprowadzono zgodnie ze schematem przedstawionym w tabeli 6.8. Ze względu na mniejszą liczbę wielokrotnych ocen tej samej pracy w przypadku matematyki mamy mniejszą liczbę niż dla języka polskiego innych egzaminatorów wewnątrz tej samej OKE, z którymi dany egzaminator ma wspólną ocenę jakiejś pracy – od 6 do 8.

Tabela 6.8. Macierz ilustrująca rozprowadzenie ocen dodatkowych 14 prac z matematyki wewnątrz OKE; dużymi literami oznaczono prace z poziomu podstawowego, małymi z rozszerzonego

$\begin{matrix} & e \\ o \end{matrix}$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
29	G	A	B	C	D	E	F	G	A	B	C	D	E	F	G	A	B	C	D	E	F	G	A	B	C	D	E	F	G
30	a	b	c	d	e	f	g	b	c	d	e	f	g	a	c	d	e	f	g	A	b	d	e	f	g	a	b	c	g

6.6.3. Schemat podziału prac w grupie nauczycieli niebędących egzaminatorami oraz studentów

Dla prac maturalnych z języka polskiego oraz z matematyki ocenianych przez grupę nauczycieli niebędących egzaminatorami oraz przez grupę studentów przyjęto schemat podziału prac, w którym:

1. w każdej grupie znajdzie się 32 oceniających;
2. każdy oceniający oceni 31 unikalnych prac maturalnych;
3. każda praca maturalna zostanie oceniona czterokrotnie wewnątrz grupy;
4. każdy oceniający będzie miał dokładnie 3 wspólne oceny jakiejś pracy z każdym oceniającym wewnątrz tej samej grupy.

Liczba oceniających w grupie oraz liczba prac ocenianych przez jedną osobę (warunki 1–2) zostały tak dobrane, aby były możliwie zbliżone do wartości zastosowanych dla poszczególnych OKE. Dzięki temu później przeprowadzane porównania między grupami odbywały się w oparciu o zbliżone:

- I. liczby oceniających w grupie (tutaj – 32, w pojedynczej OKE – 29);
- II. obciążenie pojedynczego przedstawiciela grupy liczbą ocenianych prac (tutaj – 31, w pojedynczej OKE – $29 + 2/3$ prace oceniane dodatkowo);
- III. liczby wszystkich ocen w grupie (tutaj – $32 \cdot 31$, w pojedynczej OKE – $29 \cdot [29 + 2/3 \text{ prace oceniane dodatkowo}]$).

Warunki 1–3 określają łączną liczbę unikalnych prac ocenianych przez grupę nauczycieli niebędących egzaminatorami oraz studentów na $32 \times 31/4 = 248$ prac. Liczba 248 jest podzielna przez 8 co umożliwiło dobór równej liczby prac z różnych roczników oraz tematów i (w przypadku matematyki) poziomów egzaminu maturalnego. Dla grupy nauczycieli niebędących egzaminatorami oraz studentów wybrano ten sam zbiór 248 prac, który został dobrany poprzez wylosowanie z 841 prac z podstawowego schematu 29×29 rozdziału prac pomiędzy 8 OKE (w przypadku matematyki wylosowano 124 prace z 841 prac z poziomu rozszerzonego oraz 124 prace z 841 prac z poziomu podstawowego). Warunek 4 nałożył na schemat podziału prac ograniczenie, że średnia liczba wspólnych ocen między oceniającymi z tej samej grupy wynikająca z warunków 1–3, czyli $31 \times (4-1)/31 = 3$, będzie stałą wartością wspólnych ocen między oceniającymi z tej samej grupy, co daje pełną symetryczność powiązań między oceniającymi w tej samej grupie.

Nakreślone powyżej warunki definiują niekompletny zrównoważony schemat blokowy (*Balanced Incomplete Block Design, BIBD*) o parametrach (v, b, r, k, y) równych $(32, 248, 31, 4, 3)$, gdzie v jest liczbą unikalnych punktów w zbiorze V określającym schemat (liczbą nauczycieli), b jest liczbą bloków, r jest liczbą bloków zawierających określony punkt, k jest liczbą punktów w bloku, natomiast y jest liczbą zbiorów, w których występują dowolne dwa unikalne punkty. Ze względu na zależności funkcyjne w schemacie blokowym *BIBD* (v, b, r, k, y) schemat taki również bywa oznaczany jako schemat (v, k, y) lub schemat $2 - (v, k, y)$, zatem opisywany przypadek stanowi schemat $2 - (32, 4, 3)^{10}$.

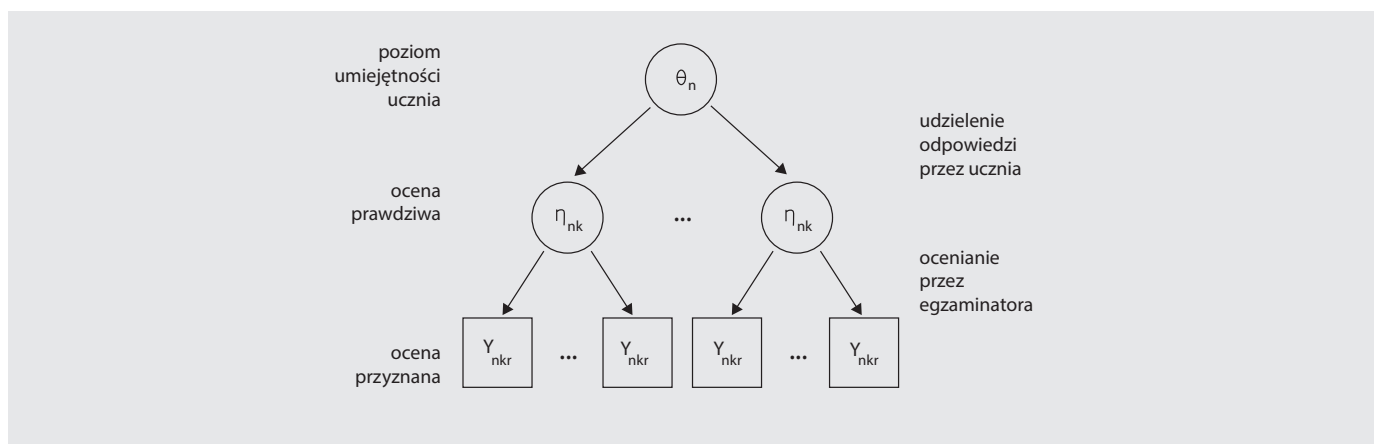
¹⁰ Schemat $2 - (32, 4, 3)$ jest podany w *Handbook of Combinatorial Designs* (Colbourn i Dinitz, 2007) w przykładzie VI.16.83, poprzez zadanie generujących go 8 bloków bazowych: $\{\infty; 17; 22; 23\}$, $\{0; 4; 7; 20\}$, $\{2; 14; 25; 28\}$, $\{1; 8; 10; 16\}$, $\{5; 9; 18; 19\}$, $\{6; 11; 13; 21\}$, $\{3; 12; 24; 30\}$, $\{15; 26; 27; 29\}$. Wymienione bloki bazowe są reprezentantami 8 orbit dla szukanych 248 bloków przy przekształceniu automorficznym $\sigma: V \rightarrow V, \sigma(x) = x + 1 \bmod 31$; $\sigma: V \rightarrow V, \sigma(x) = x + 1 \bmod 31$, gdzie $\text{mod}(\infty + n, 31) = 32$.

7. Statystyczna koncepcja analizy efektu egzaminatora

Zagadnienie efektu egzaminatora można modelować z użyciem różnych metod, w zależności od celu, któremu ma służyć analiza. Jednym z nich może być diagnostyka systemu egzaminacyjnego i dążenie do zwiększenia rzetelności ocen poprzez eliminację nierzetelnych oceniających czy udoskonalenie procedur. Innym celem może być oszacowanie wpływu egzaminatorów na oceny konkretnych uczniów i wykorzystanie tej wiedzy do uwzględnienia poprawki w ostatecznej ocenie ucznia. Oczywiście jeden model może być odpowiedni do różnych celów, choć niektóre z nich pozwalają na oszacowanie tylko jednego typu efektu egzaminatora. Najczęściej stosowane modele można podzielić na dwie grupy: modele analizy wariancji (Dolata i in., 2004; Glas, 2012; Scheerens, Glas i Thomas, 2003) oraz modele z wykorzystaniem *Item Response Theory* (IRT; DeCarlo, Kim i Johnson, 2011; Dubiecka i in., 2006; Patz i Junker, 1999; Patz, Junker, Johnson i Mariano, 2002). W kontekście polskiego systemu egzaminów zewnętrznych efekt egzaminatora nie był często analizowany. Wśród systematycznych badań na ten temat można wymienić prace Romana Dolaty, Elżbiety Putkiewicz i Anny Wiłkomirskiej (2004) w zakresie egzaminu maturalnego oraz Anny Dubieckiej, Henryka Szaleńca i Doroty Węziak (2006) w odniesieniu do sprawdzianu po szóstej klasie szkoły podstawowej. Prace te reprezentują wspomniane wyżej, odmienne podejścia do analizy.

Z reguły zainteresowanie badaczy pomiaru edukacyjnego skupia się na precyzji oceniających oraz na ich łagodności/surowości, a rzadziej na efektach zniekształcenia skali. Prawdopodobną przyczyną takiego stanu rzeczy jest trudność w dobrym rozróżnieniu między łagodnością/surowością i zjawiskami takimi jak ograniczenie skali, ekstremizm czy tendencja centralna. Bardzo rzadko wspomina się o efekcie halo, jednak w danych z polskiego systemu egzaminów zewnętrznych nie mamy do czynienia z sytuacją, kiedy jeden egzaminator dokonuje oceny znacząco różnych cech (umiejętności) ucznia. Nawet jeśli ocenie podlega kilka zadań, z reguły mierzą one te same lub bardzo podobne umiejętności, zatem efekt halo właściwie nie występuje. Optymalny model analizy powinien pozwalać na oszacowanie przynajmniej łagodności/surowości i precyzji oceniających. Pożądaną cechą byłaby także możliwość szacowania efektów zniekształceń skali. Krótkiego przeglądu modeli analizy efektu egzaminatora dokonał Kulon (2014).

Rysunek 7.1. Hierarchiczny model efektu egzaminatora z użyciem teorii detekcji sygnału (HRM-SDT)



Do analizy danych z badania zastosowano nieco zmodyfikowaną postać hierarchicznego modelu oceniającego z teorią detekcji sygnału (*Hierarchical Rater Model with Signal Detection Theory*; HRM-SDT; DeCarlo i in., 2011), którego graficzną reprezentację przedstawiono na rysunku 7.1. Jest to model dwupoziomowy – pierwszy poziom odpowiada procesowi udzielania odpowiedzi na poszczególne pytania (k) przez ucznia (n), drugi natomiast ocenianiu tej odpowiedzi przez egzaminatora (r). W modelu przyjmuje się, że każdej odpowiedzi udzielonej przez ucznia odpowiada pewna ocena prawdziwa (η_{nk}), która ma charakter obiektywny, ale jest nieobserwowalna. Zmienną latentną jest również poziom umiejętności ucznia (θ_n). Jedynymi zmiennymi obserwowalnymi są oceny przyznane przez egzaminatorów (Y_{kr}).

Z punktu widzenia interpretacji wyników modelu istotne są jedynie parametry dla poziomu przydzielania ocen przez egzaminatorów. Poziom udzielania odpowiedzi przez uczniów pełni jednak ważną rolę w procesie estymacji modelu. Sprawia, że przy szacowaniu oceny prawdziwej danego zadania w danej pracy uwzględniane są nie tylko informacje o tym, jak egzaminatorzy

oceniali daną pracę i dane zadanie w ogóle, ale również informacje o tym, jak dany uczeń odpowiadał na pozostałe zadania w tej samej pracy. W modelu HRM-SDT zaproponowanym przez DeCarla i współpracowników (2011) na poziomie tym użyto modelu *Generalized Partial Credit Model* (GPCM), wspominając jednocześnie, że rolę tę pełnić mogą także inne modele z rodziny *Item Response Theory* (IRT). W badaniu zdecydowano się na wykorzystanie modelu *Graded Response Model* (GRM), ponieważ jest on (z uwagi na rosnące uporządkowanie trudności poziomów wykonania) bardziej stabilny obliczeniowo od GPCM, co odgrywało ważną rolę w wypadku kryterium oceny **treść** dla języka polskiego.

Sam efekt egzaminatora opisywany jest przez drugi poziom modelu. Jego autorzy zakładają, że decyzja ocenającego odnośnie do oceny przyznawanej uczniowi za zadanie jest uzależniona od percepcji jakości odpowiedzi ucznia na zadanie, a percepcja ocenającego jest ukrytą, ciągłą zmienną losową. W przypadku konkretnego zadania percepcja jest realizacją z rodziny normalnych lub logistycznych rozkładów prawdopodobieństwa, z innym parametrem położenia dla każdej kategorii „prawdziwej” odpowiedzi ucznia. Odległość pomiędzy poszczególnymi rozkładami (ich parametr położenia) jest zależna od zdolności ocenającego do rozróżnienia ukrytych kategorii „prawdziwych”, pozwala zatem na oszacowanie precyzji ocenającego. Drugim założeniem poczynionym na potrzeby tego modelu jest to, że ocenający wyznaczają progi wykonania zadania tak, aby zaklasyfikować zadanie do odpowiedniej kategorii w zależności od tego, pomiędzy którymi progami znajduje się ich percepcja. Dla J kategorii, które mogą przydzielić ocenający, wyznaczanych jest $J-1$ progów – poniżej pierwszego progu przyznawana jest pierwsza kategoria, a powyżej ostatniego najwyższa kategoria. Liczba ukrytych, „prawdziwych” kategorii nie musi być równa liczbie kategorii, które mogą przyznać egzaminatorzy. Konstruowanie skal ocen opiera się jednak na przekonaniu, że można rozróżnić tyle poziomów umiejętności, ile kategorii używanej skali. Można zatem uznać, że liczba kategorii skali używanej przez ocenających jest równa liczbie ukrytych kategorii, do której należy badana umiejętność.

HRM-SDT na poziomie przyznawania oceny przez egzaminatora można przedstawić następująco¹¹:

$$P(Y_{nkr} \leq j | \eta_{nk} = \eta) = F(-d_{kr}(\eta - 1 - c_{kjr})),$$

gdzie:

$Y_{nkr} \leq j$ – ocena (kategoria) ucznia n w zadaniu k przypisana przez egzaminatora r mniejsza lub równa j ;

$\eta_{nk} = \eta$ – „prawdziwa” ocena (kategoria) ucznia n w zadaniu k równa η ;

F – dystrybuanta rozkładu normalnego lub logistycznego;

c_{kjr} – próg kategorii j w zadaniu k dla egzaminatora r ;

d_{kr} – precyzja egzaminatora r w zadaniu k .

Jeśli przyjmiemy, że rozkład prawdopodobieństwa percepcji egzaminatorów należy do rodziny logistycznych rozkładów prawdopodobieństwa, to przedstawione równanie opisuje kumulatywne prawdopodobieństwa dla poszczególnych kategorii zadania w modelu GRM. Jest to szeroko stosowany model IRT dla zmiennych wielokategorialnych. Różnica w stosunku do GRM polega na tym, że w przypadku HRM-SDT mamy do czynienia z dyskretną, a nie ciągłą, zmienną ukrytą. Dzięki użytej parametryzacji progi są wyznaczane na skali zmiennej η , a więc bezpośrednio odnoszą się do kategorii i są niezależne od precyzji. Wartość „idealnego” progu c_j w takiej sytuacji jest równa $j - 0,5$.

Z opisanego powyżej modelu HRM-SDT otrzymujemy dla każdego egzaminatora dwie grupy parametrów: d_{kr} , które oznaczają precyzję egzaminatora w danym zadaniu, a także c_{kjr} , które wyznaczają progi służące do przyznawania poszczególnych kategorii w danym zadaniu. Dzięki porównaniu położenia progów oszacowanych dla poszczególnym egzaminatorów do progów „idealnego” ocenającego możliwe jest uchwycenie kilku istotnych efektów egzaminatora.

¹¹ W stosunku do wzoru 2 w publikacji DeCarla i współpracowników (2011) zmieniono tu jedynie parametryzację na, równoważną jej, klasyczną parametryzację IRT (która pozwala na bardziej intuicyjną interpretację wartości współczynników d i c) oraz dodano korektę $n - 1$ wynikającą z przyczyn technicznych, związanych z zapisem modelu w języku BUGS i używanym w nim sposobem indeksacji tablic.

Łagodność oceniającego można definiować jako umieszczanie przez niego progów poszczególnych kategorii poniżej progów kategorii „idealnych”, natomiast surowość – powyżej. Oznacza to, iż łagodny oceniający przypisuje kolejną kategorię przy niższym poziomie umiejętności ucznia niż „idealny” egzaminator. Surowy oceniający natomiast przypisuje kolejną kategorię przy wyższym poziomie tej cechy.

Jeśli egzaminator swój pierwszy próg umieści dużo przed pierwszym progiem „idealnego” egzaminatora, a ostatni dużo za ostatnim „idealnym” progiem, będzie to oznaczało, iż w praktyce nie będzie on przyznawał pierwszej i ostatniej kategorii. Jest to przejaw tendencji centralnej oceniającego. Analogicznie, jeśli egzaminator skrajne progi umieści odpowiednio powyżej pierwszego „idealnego” progu i poniżej ostatniego „idealnego” progu i jednocześnie pozostałe progi przesunie w kierunku środkowych kategorii, to będziemy mieli do czynienia z efektem używania ocen skrajnych. W ten sposób można również identyfikować ograniczenia skali w dowolnym jej obszarze.

Model HRM-SDT pozwala na jednoczesne szacowanie większej gamy efektów egzaminatora niż pozostałe modele opisywane w literaturze. Jednocześnie przeprowadzone przez Kulona (2014) symulacje sugerują, że model ten odpowiada potrzebom analizy danych z *Badania porównywalności oceniania i efektu egzaminatora* i specyfice danych z maturalnych arkuszy egzaminacyjnych. Sposoby interpretacji parametrów modelu, a także konstrukcję użytych wskaźników opisano w rozdziale 7.3. Estymacji dokonano pełną metodą bayesowską z użyciem metody MCMC (*Markov chain Monte Carlo*) za pomocą pakietu statystycznego R (biblioteka rjags) i programu JAGS 3.4.0. Dla każdej kombinacji przedmiotu egzaminacyjnego, poziomu egzaminu i (w wypadku j. polskiego) tematu wypracowania przeprowadzono następującą procedurę estymacji modelu:

1. Obliczenie 3000 iteracji łańcucha na danych, z których wyłączono osoby nieposiadające statusu egzaminatora. Z tego pierwszych 1000 iteracji odrzucano, a na podstawie kolejnych 2000 szacowano oceny „prawdziwe” poszczególnych zadań w poszczególnych pracach η_{nk} .
2. Obliczenie 3000 iteracji łańcucha na pełnych danych, z ocenami „prawdziwymi” η_{nk} zafiksowanymi na wartościach obliczonych w poprzednim kroku. W tym wypadku wykorzystywano de facto model jednopoziomowy (tylko poziom przyznawania oceny przez egzaminatora). Również tutaj pierwszych 1000 iteracji było odrzucanych, a ostateczne wartości parametrów obliczano na podstawie kolejnych 2000.

Przyjęte rozkłady a priori dla poszczególnych parametrów modeli przedstawiono w tabeli 7.1. Dla progów (c_{kjr}) wybrano rozkład normalny ze średnią odpowiadającą położeniu „idealnego” progu dla danej kategorii i wariancją równą 1, przy czym zgodnie z wymogami GRM wartości dla kolejnych progów posortowano rosnąco. Dla precyzji (d_{kr}) wybrano rozkład lognormalny z wariancją $\ln(2)$ i średnią równą $\ln(4)$ dla wszystkich zadań z wyjątkiem kryterium *treść* z języka polskiego oraz $\ln(1,2)$ dla kryterium *treść*.

Tabela 7.1. Parametry rozkładów a priori dla parametrów modelu HRM-SDT

Parametr	Rozkład	M	VAR	Uwagi
θ_n	normalny	0	1	
a_k	normalny	1	1	Wartości ograniczono do $a_k > 0$.
b_{kj}	normalny	0	1	Wartości sortowano rosnąco dla kolejnych kategorii.
d_{kr}	lognormalny	$\ln(4)$	$\ln(2)$	Dla kryterium <i>treść</i> z języka polskiego $M = \ln(1,2)$.
c_{kjr}	normalny	$j - 0,5$	1	Wartości sortowano rosnąco dla kolejnych kategorii.

7.1. Modelowanie kryterium oceny *treść* w języku polskim

Kryterium oceny *treść* w języku polskim wyraźnie różni się od pozostałych długością skali. O ile najdłuższe skale zadań z matematyki obejmują 7 poziomów wykonania, a z języka polskiego, wyłączwszy *treść*, 6 poziomów wykonania, o tyle kryterium *treść* oceniane jest na skali o 26 poziomach wykonania. Zasadne wydaje się pytania, w jaki sposób tak długa skala wpłynie na stabilność obliczeniową modelu (konieczność oszacowania wielokrotnie większej liczby parametrów) oraz czy modelowane zjawiska przebiegają w ten sam sposób dla tak długich skal. Niestety żadna znana nam publikacja nie poruszała tego problemu. Z uwagi na to,

że kluczowym celem badania było opisanie efektu egzaminatora w odniesieniu do rzeczywistego egzaminu, w którym faktycznie stosowana jest dla **treści** skala o 26 poziomach wykonania, zdecydowano się ją modelować właśnie w takiej formie. Poprzedzone zostało to jednak analizami teoretycznymi oraz symulacyjnymi.

Rozważania teoretyczne nad zachowaniem modelu HRM-SDT na poziomie przyznawania oceny przez egzaminatora dla bardzo długich skal¹² wykazały, że spodziewać się należy spadku wartości parametru d wraz ze wzrostem długości skali (dokładnie $dJ = \text{const}$, gdzie J – długość skali). W praktyce oznacza to przyjęcie założenia, że percepcja egzaminatora nie jest doskonale rozciągliwa i od pewnego momentu zwiększanie długości skali nie będzie zwiększać precyzji oceny, a jedynie zwiększenie wariancji uzyskiwanych ocen. Założenie to wydaje się zgodne z tym, w jaki sposób intuicyjnie wyobrażać można sobie przebieg procesu oceniania dla bardzo długich skal, a więc jedna z wątpliwości została rozwiązana. Jednocześnie wyniki tych analiz wykazały, że nie jest poprawne metodologicznie porównywanie ze sobą wartości współczynników d dla skal o różnych długościach. W związku z tym przy omawianiu wyników użyty został wskaźnik standaryzowanej wartości d (patrz rozdział 7.3).

Oddzielny problem stanowi stabilność obliczeniowa modelu w momencie szacowania dla danej pracy tak dużej liczby parametrów. Z uwagi na zastosowanie estymacji bayesowskiej podstawowym ryzykiem jest w tym wypadku nadmierne ściągnięcie wyników modelu do wartości przyjętych dla danego parametru rozkładów a priori. W wypadku parametrów c_{kjr} oznaczałoby to niedoszacowanie efektów związanych z różnicami w łagodności egzaminatorów i zniekształceń skali (wartości oczekiwane rozkładów a priori parametrów c_{kjr} wskazują wartości idealne, dla których wspomniane efekty nie istnieją), natomiast w wypadku parametrów d_{kr} po prostu ściągnięcie do wartości opisanych rozkładem a priori. Aby zweryfikować, czy problemy te faktycznie występują, przeprowadzono badania symulacyjne, stosując zróżnicowane wartości rozkładów a priori parametrów c_{kjr} i d_{kr} , także zbadano wrażliwość uzyskiwanych wyników na liczbę przeprowadzonych iteracji MCMC. Analizy te wykazały wysoką stabilność szacowania przez model wartości parametrów d_{kr} niezależnie od zastosowanych rozkładów a priori i liczby iteracji MCMC. Co więcej, uzyskiwane oszacowania d_{kr} były zgodne z przedstawionymi w poprzednim akapicie rozważaniami teoretycznymi. Nieco problematyczne okazało się szacowanie wartości parametrów c_{kjr} . W tym wypadku zaobserwowano ściągnięcie wyników do przyjętych rozkładów a priori, co sugeruje, że siła efektów związanych ze zróżnicowaniem łagodności egzaminatorów w odniesieniu do kryterium oceny **treść** została w modelu nieco niedoszacowana. Uzyskaniu większej stabilności obliczeniowej modelu posłużyło również zastąpienie na poziomie udzielania odpowiedzi przez ucznia modelu GPCM zaproponowanego przez DeCarla i współpracowników (2011) modelem GRM.

Ostatecznie, ważąc wszystkie za i przeciw, zdecydowano się na to, by modelować kryterium **treść** z języka polskiego tak samo, jak wszystkie inne zadania w badaniu, jednocześnie (mając świadomość konsekwencji tej decyzji w odniesieniu do szacowania parametrów c_{kjr}) stosując większą powściągliwość przy omawianiu dla tego kryterium wyników odnoszących się do łagodności egzaminatorów i zniekształceń skali.

7.2. Interpretacja parametrów modelu

Parametry użytego modelu *Graded Response Model* – dyskryminacja egzaminatora oraz subiektywne progi trudności poszczególnych poziomów wykonania – można interpretować jako, odpowiednio, miarę błędu losowego (rzetelności) oraz miarę błędu systematycznego (surowości/łagodności oraz zniekształceń skali) danego egzaminatora w danym zadaniu.

Rzetelność oceny pracy odpowiada za zróżnicowanie ocen, które ten sam egzaminator przydzieliłby za dokładnie tak samo wykonane zadania. W praktyce chcielibyśmy, aby zróżnicowanie takie było jak najmniejsze, a najlepiej, aby nie było go w ogóle, jest ono bowiem niezwiązane z tym, jak uczeń rozwiązał zadanie. Im bliższe 0 wartości dyskryminacji egzaminatora, tym niższa rzetelność (większy błąd losowy). Egzaminator, którego parametr dyskryminacji miałby wartość 0 (na szczęście w praktyce nie obserwujemy takich egzaminatorów), przydzielałby oceny całkowicie losowo, nie biorąc w ogóle pod uwagę tego, jakiej odpowiedzi udzielił uczniów. Z kolei im wyższe wartości parametru dyskryminacji egzaminatora, tym mniejszy błąd losowy popełniany przez egzaminatora (przy czym skala parametru nie wyznacza niestety żadnego maksimum, gdzie błąd losowy już nie występuje¹³). Wartości

¹² Ze względu na ich charakter – stosunkowo zaawansowanej analizy matematycznej z wykorzystaniem rachunku całkowego – zdecydowano, że wykraczają one poza ramy niniejszej publikacji, w której wystarczy omówienie uzyskanych wyników.

¹³ Formalnie byłaby to sytuacja, w której parametr dyskryminacji egzaminatora miałby wartość plus nieskończoność.

parametrów dyskryminacji egzaminatorów możemy też utożsamiać ze stopniem pewności, z jaką egzaminator przydziela ocenę – im wyższa wartość parametru, tym egzaminator ma większą pewność co do tego, jaką ocenę przyznać, a im mniejsza, tym więcej różnych ocen rozważa w procesie przyznawania oceny. Warto przy tym zaznaczyć, że z użytego modelu explicite wynika, że im większa liczba poziomów wykonania zadania, tym mniejsze będą wartości dyskryminacji egzaminatorów dla tego zadania¹⁴. Odpowiada to intuicji, że proces oceny zadania komplikuje się wraz z wydłużaniem skali oceny, a więc dla dłuższych skal podjęcie właściwej decyzji jest trudniejsze, zatem odbywa się z mniejszą pewnością (jest mniej rzetelne). Powyższe omówienie ma charakter jakościowy. Niestety przejście na opis ilościowy, w którym umielibyśmy przeliczyć wartości parametrów dyskryminacji np. na wariancję liczby punktów przydzielanych przez egzaminatora za ocenę tak samo rozwiązanego zadania, nie jest łatwy. Wymaga on wprowadzenia specjalnych wskaźników, które zostały omówione w dalszej części rozdziału.

Surowość/łagodność (błąd systematyczny) egzaminatora sprowadza się do różnicy ocen przydzielonych za tak samo rozwiązane zadanie pomiędzy różnymi egzaminatorami (z pominięciem tej części różnicy ocen, która wynika z błędów losowych tych egzaminatorów¹⁵). W modelu różnica ta mierzona jest względem oceny „prawdziwej” zadania w danej pracy wyestymowanej w modelu, ale warto zauważyć, że z praktycznego punktu widzenia istotne są przede wszystkim różnice między egzaminatorami (jeśli wszyscy egzaminatorzy byłoby tak samo surowi lub łagodni, to pomijając oceny skrajne, nie wpływałoby to na wyniki). Stąd też chcielibyśmy, aby różnice ocen przydzielanych za tak samo rozwiązane zadanie przez różnych egzaminatorów były jak najmniejsze. Miarą surowości/łagodności (błędu systematycznego) w użytym modelu są występujące pomiędzy egzaminatorami różnice wartości progów trudności poszczególnych poziomów wykonania. Im mniej wartości te różnią się między egzaminatorami, tym mniejszy błąd systematyczny. Także w tym wypadku nie istnieje proste przejście od wartości różnic parametrów modelu do różnic wyrażonych w liczbie punktów na egzaminie, lecz niezbędne jest wprowadzenie odpowiednich wskaźników.

Zniekształcenia skali również wiążą się z różnicami w położeniu progów trudności względem „idealnego” egzaminatora. Poszczególne konfiguracje przesunięć progów mogą świadczyć o ograniczeniu skali (w tym tendencji centralnej) lub używaniu ocen skrajnych. Zjawiska te mogą przyjmować różny charakter i występować w różnych miejscach skali, zwłaszcza w przypadku długich skal. Różnice w położeniu progów wprost można interpretować tylko w odniesieniu do konkretnego ocenianego i poszczególnych progów konkretnego zadania. Już obliczenie miary dla całego zadania (wszystkich progów) sprawia pewne trudności. Opis konstrukcji wskaźnika zniekształcenia skali znajduje się poniżej.

7.3. Wskaźniki siły efektu egzaminatora

Aby móc wygodnie interpretować wyniki modelowania, musimy przedstawić je w łatwej do zrozumienia i interpretacji postaci. Parametrom modelu można co prawda przypisać intuicyjne interpretacje jakościowe, jednak nie pozwalają one na wygodne interpretacje ilościowe. Stąd potrzeba wprowadzenia wskaźników, które przekształcą parametry modelu na intuicyjne w interpretacji skale. W badaniu wprowadzono i zastosowano cztery takie parametry:

1. standaryzowany współczynnik dyskryminacji (ozn. d_j) – współczynnik dyskryminacji przeliczony na skalę dychotomiczną (o jedynie dwóch poziomach wykonania);
2. miara zniekształcenia skali (ozn. r) – rozstęp progów egzaminatora w zadaniu podzielony przez rozstęp progów „idealnych”;
3. odsetek poprawnych klasyfikacji (ozn. p) – odsetek ocen, dla których egzaminator przydzielił ocenę zgodną z oceną „prawdziwą” wyestymowaną w modelu;
4. łagodność (ozn. b) – systematyczna różnica pomiędzy ocenami przydzielanymi przez różnych egzaminatorów wyrażona na skali surowych wyników egzaminacyjnych (wyrażonych w procentach).

Wymienione wyżej parametry można agregować na różne poziomy, przy czym z punktu widzenia analiz interesujące będą:

1. poziom egzaminatora i zadania (ozn. p_{kr} , b_{kr} i r_{kr})
2. poziom egzaminatora i testu lub tematu z j. polskiego (ozn. p_{tr} i b_{tr})
3. poziom Okręgowej Komisji Egzaminacyjnej i testu lub tematu z j. polskiego (ozn. p_{to} i b_{to}).

¹⁴ Formalnie iloczyn liczby poziomów wykonania i wartości parametru dyskryminacji jest w przybliżeniu stały.

¹⁵ Formalnie jest to więc różnica wartości oczekiwanych ocen pomiędzy egzaminatorami.

Każdy ze wskaźników obliczany jest niezależnie dla każdej iteracji łańcucha modelu, po czym jako wartość parametru przyjmowana jest średnia spośród wartości we wszystkich iteracjach. Także agregaty wyższego poziomu (np. poziomu OKE) obliczane są najpierw oddzielnie dla każdej iteracji, a następnie uśredniane po iteracjach. Taki sposób wyliczania i agregacji parametrów pozwala na proste i eleganckie obliczenie przedziałów ufności wartości obliczanych parametrów w oparciu o odchylenie standardowe wartości parametru pomiędzy iteracjami łańcucha.

7.3.1. Standaryzowana dyskryminacja egzaminatora

Z przyjętego modelu explicite wynika, że jeśli pozostałe parametry modelu mają pozostać niezmiennicze, to wartość parametru dyskryminacji musi spadać wraz z wydłużaniem się skali, na której oceniane jest zadanie. Powoduje to, że nie da się porównywać ze sobą bezpośrednio wartości parametrów dyskryminacji dla pytań ocenianych na skalach różnej długości. Najprostszym sposobem porównania sobie z tym problemem jest standaryzacja wartości parametrów do tej samej długości skali oceny pytania. W wypadku badania wybrano standaryzację do skali o dwóch poziomach wykonania $d_s = J * d/2$, gdzie J to liczba poziomów wykonania zadania. Porównując ze sobą wartości współczynnika d_s dla różnych pytań, możemy stwierdzić, które z zadań byłoby bardziej rzetelnie oceniane, gdyby miały one tyle samo poziomów wykonania.

7.3.2. Odsetek poprawnych klasyfikacji

Odsetek poprawnych klasyfikacji jest najbardziej ogólną miarą przyjętą w badaniu. Kumuluje on w jednej wartości zarówno błąd losowy, jak i systematyczny egzaminatora. Formalnie zdefiniowany został jako średni odsetek poprawnych klasyfikacji (czyli takich, gdzie ocena przydzielona przez egzaminatora Y_{kr} jest równa ocenie „prawdziwej” η) dla poszczególnych ocen „prawdziwych” ważony prawdopodobieństwem wystąpienia każdej z ocen „prawdziwych”:

$$\rho_{kr} = \sum_{\eta} f(Y_{kr} = \eta | \eta_k = \eta) P(\eta_k = \eta),$$

gdzie f to funkcja charakterystyczna modelu efektu egzaminatora: $f(j|\eta) = \frac{1}{1 + \exp(-d(\eta - c_j))}$, której wartości możemy obliczyć na podstawie parametrów wyestymowanych w modelu.

Aby móc policzyć wartość wskaźnika, niezbędne jest jeszcze przyjęcie rozkładu ocen „prawdziwych” w populacji. Dzieje się tak dlatego, że w zastosowanym modelu prawdopodobieństwo poprawnej klasyfikacji może być różne w zależności od oceny „prawdziwej” danego zadania. Przyjęty w badaniu rozkład ocen „prawdziwych” opisany został w rozdziale poniżej.

7.3.3. Łagodność

Łagodność jest zagregowaną miarą błędu systematycznego egzaminatora. Formalnie zdefiniowana została w badaniu jako różnica wartości oczekiwanych punktacji odpowiadającej ocenie przyznanej przez egzaminatora (ozn. Y) oraz punktacji odpowiadającej ocenie „prawdziwej” (ozn. η) dla danego kryterium oceny k :

$$b_{kr} = \left(E(pkt(Y_{kr})) - E(pkt(\eta_k)) \right) 100\% / \max_{pkt}$$

Ze względu na różnice w maksymalnej liczbie punktów możliwych do zdobycia na egzaminie z języka polskiego (70 punktów) i matematyki (50 punktów) zdecydowano się wyrazić wskaźnik łagodności nie w surowych punktach wyniku egzaminacyjnego, lecz w punktach procentowych wyniku egzaminacyjnego. Pozwala to na łatwe porównywanie siły efektu pomiędzy językiem polskim i matematyką.

W wypadku zadań z matematyki przekształcenie z poziomów wykonania na punkty jest trywialne: $pkt(Y) = Y - 1$, jednak dla niektórych kryteriów oceny z polskiego jest ono bardziej skomplikowane (np. język oceniany jest na skali o 6 poziomach wykonania, którym odpowiada punktacja 0, 1, 3, 6, 9 lub 12 punktów), stąd konieczność zachowania we wzorze funkcji pkt .

Ponieważ dla zmiennych dyskretnych (a takimi są punkty odpowiadające ocenom przyznanym przez egzaminatorów) wartość oczekiwana to średnia z wartości zmiennej ważona prawdopodobieństwem jej wystąpienia, możemy zapisać powyższy wzór jako:

$$b_{kr} = \left[\left(\sum_{\eta} pkt(\eta) P(\eta_k = \eta) \right) - \left(\sum_{\eta} \left(\sum_j f(Y_{kr} = j | \eta_k = \eta) pkt(j) \right) P(\eta_k = \eta) \right) \right] 100\% / \max_{pkt}$$

co można uprościć do postaci:

$$b_{kr} = \left(\sum_{\eta} P(\eta_k = \eta) \left(pkt(\eta) - \left(\sum_j f(Y_{kr} = j | \eta_k = \eta) pkt(j) \right) \right) \right) 100\% / \max_{pkt}$$

gdzie f to funkcja charakterystyczna modelu efektu egzaminatora: $f(j|\eta) = \frac{1}{1 + \exp(-d(\eta - c_j))}$, której wartości możemy obliczyć na podstawie parametrów wyestymowanych w modelu.

Aby móc policzyć wartość wskaźnika, niezbędne jest jeszcze przyjęcie rozkładu ocen „prawdziwych” w populacji. Dzieje się tak dlatego, że w zastosowanym modelu błąd systematyczny egzaminatora może być różny w zależności od oceny „prawdziwej” danego zadania. Przyjęty w badaniu rozkład ocen „prawdziwych” opisany został w rozdziale poniżej.

Agregacja wskaźnika b_{kr} do poziomu testu przeprowadzona została jako proste zsumowanie wartości wskaźnika danego egzaminatora dla wszystkich zadań danego testu. Agregację do poziomu testu i OKE obliczono jako średnią z wartości wskaźnika zagregowanego do poziomu testu w grupach ze względu na OKE, z której pochodził egzaminator.

7.3.4. Zniekształcenie skali

Wskaźnik zniekształcenia skali jest konstruowany nieco inaczej niż dwa wskaźniki opisane powyżej. Opiera się on na prostej mierze rozstępu pomiędzy progami w danym zadaniu. W celu uzyskania porównywalności pomiędzy zadaniami rozstęp ten jest dodatkowo dzielony przez rozstęp kategorii „idealnych” dla danego zadania. Jego obliczenie jest niemożliwe dla zadań dychotomicznych – istnieje dla nich tylko jeden próg i nie może w nich występować zjawisko zniekształcenia skali inne niż łagodność lub surowość. W ten sposób otrzymana miara przyjmuje teoretyczne wartości od 0 do plus nieskończoności. Dla zadania k o liczbie kategorii J można ją zapisać następująco:

$$r_{kr} = \frac{\max(c_{jkr}) - \min(c_{jkr})}{J - 2}.$$

W sytuacji, gdy egzaminator swoje progi umieszcza tam, gdzie zrobiłby to idealny egzaminator, wskaźnik ten przyjmuje wartość 1 – rozstęp progów egzaminatora jest równy rozstępowi progów „idealnych”. Wartości mniejsze od 1 świadczą o tendencji do udzielania ocen skrajnych (ekstremizmu). W przypadku przyznawania przez egzaminatora tylko skrajnych kategorii miara ta przyjmie wartość 0, gdyż rozstęp progów będzie wynosił 0. Wartości większe od 1 oznaczają zawężanie skali – rozstęp progów egzaminatora jest większy niż rozstęp „idealnych” progów wskutek nieprzyznawania przez egzaminatora niektórych (skrajnych) kategorii oceny. Trudno ocenić wielkość danego efektu bezwzględnie, miara ta pozwala jedynie na stwierdzenie odstępstw od idealnych progów i względne porównania pomiędzy egzaminatorami. W połączeniu z informacją o łagodności lub surowości danego egzaminatora pozwala ona na stwierdzenie, w którym miejscu skali następuje jej ograniczenie.

7.3.5. Rozkład ocen „prawdziwych” w populacji

W zastosowanym modelu surowość/łagodność egzaminatora może być różna dla każdego poziomu wykonania zadania. Powoduje to, że aby policzyć jakikolwiek agregat biorący pod uwagę parametry surowości/łagodności egzaminatora, potrzebne jest przyjęcie rozkładu ocen „prawdziwych”. Najprostszym rozwiązaniem byłoby przyjęcie rozkładu ocen „prawdziwych” wyestymowanego w modelu, jednak należy pamiętać, że próba prac dobranych do badania miała charakter celowy ze względu na wartości sumy punktów przyznanej za daną pracę. Prace do badania dobierane były w ten sposób, aby rozkład sumy punktów prac był możliwie równomierny, podczas gdy rozkłady ocen obserwowanych w populacji są (mniej lub bardziej) zbliżone do normalnych. Stąd też chcąc wnioskować o sile efektu egzaminatora w populacji uczniów, należy odtworzyć rozkład ocen „prawdziwych” w populacji. Jako estymator rozkładu ocen „prawdziwych” w populacji przyjęto rozkład łączny ocen obserwowanych oraz wyestymowanych w modelu ocen „prawdziwych” przemnożony przez rozkład ocen obserwowanych w populacji:

$$\hat{P}_{pop}(\eta_k = \eta, Y_k = y) = P_{próba}(\eta_k = \eta, Y_k = y) P_{pop}(Y_{pop} = y)$$

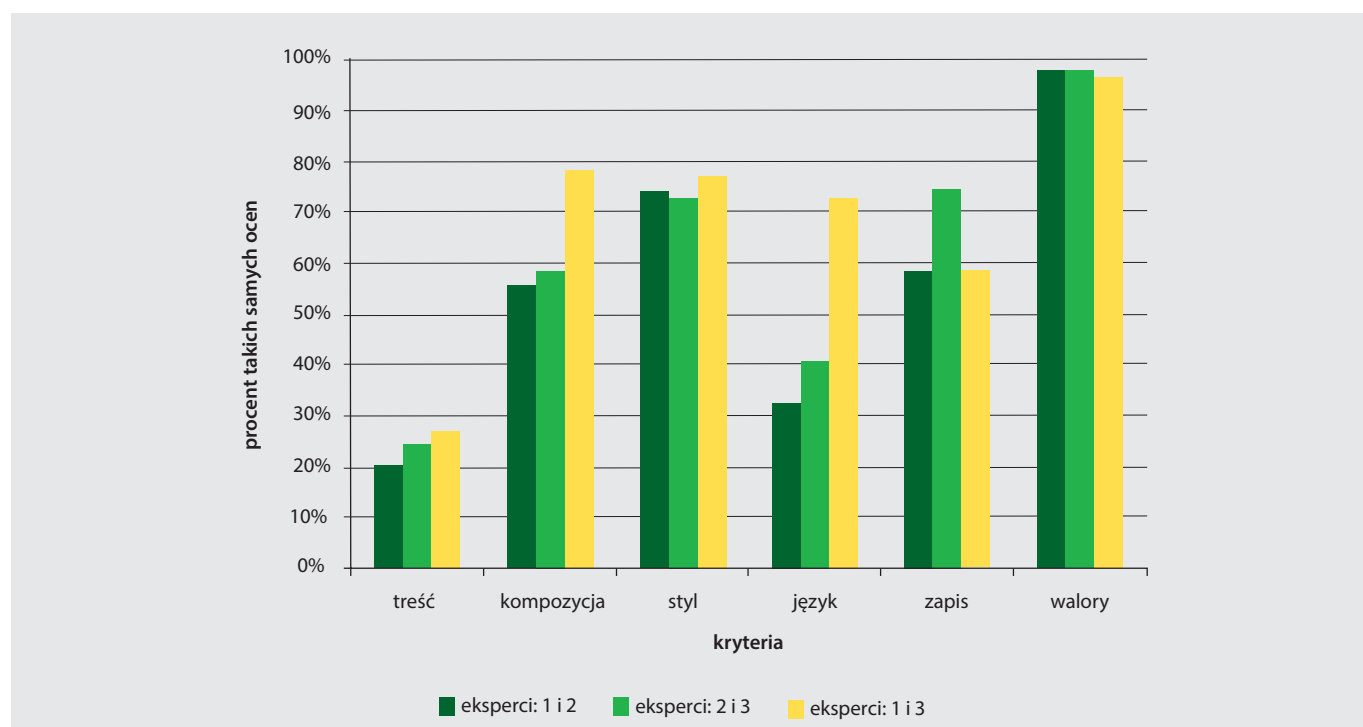
7.4. Walidacja ocen „prawdziwych” z modelu – ocenianie ekspertów

Wybrany model analizy bazuje na szacowaniu ocen „prawdziwych” na podstawie ocen poszczególnych egzaminatorów. Zasadne wydaje się pytanie, na ile te „prawdziwe” oceny oszacowane w modelu zgodne są z ocenami, które uczniowie rzeczywiście powinni otrzymać, gdyby nie istniał efekt egzaminatora. Weryfikacji tego aspektu wybranego modelu analizy postanowiono dokonać poprzez porównanie oszacowanych ocen „prawdziwych” z uzgodnionymi ocenami panelu ekspertów. W zasadzie jest to jedyny sposób określenia, czy ostateczne oceny uczniów oszacowane w modelu są adekwatne do ich poziomu umiejętności. Porównanie bezpośrednio poziomu umiejętności z modelu HRM-SDT z umiejętnością tych samych uczniów oszacowaną z danych egzaminacyjnych jest skazane na porażkę, gdyż istotą zastosowanego modelu jest pozbawienie ostatecznej oceny ucznia wpływu egzaminatora. Ocena ustalona przez panel ekspertów jest najlepszym przybliżeniem „prawdziwej” oceny ucznia, które nie wymaga stosowania modelu statystycznego¹⁶.

Do porównania wybrano losowo 121 prac z II tematu z 2011 roku z języka polskiego spośród tych, które zostały zakwalifikowane do badania. Panel ekspertów składał się z trójki koordynatorów oceniania na poziomie OKE. Podczas pierwszego etapu procedury każdy z ekspertów miał samodzielnie ocenić wszystkie prace zgodnie ze schematem oceniania. W drugim etapie eksperci podczas wspólnego spotkania w siedzibie IBE wspólnie omawiali poszczególne prace i uzgadniali ostateczną ocenę. Zebrane dane składały się z indywidualnych ocen ekspertów oraz oceny ostatecznej w podziale na poszczególne kryteria oceniania. Dzięki temu możliwe było porównanie zarówno zgodności pomiędzy ekspertami, jak i zgodności „prawdziwych” ocen z modelu z ostateczną oceną ekspertów.

Jako pierwszy krok analizy zbadano zgodność pomiędzy parami ekspertów rozumianą jako procent takich samych ocen przyznanych za poszczególne kryteria. Otrzymane wyniki przedstawia rysunek 7.2. Zgodnie z oczekiwaniami najniższy procent takich samych odpowiedzi eksperci osiągnęli w kryterium **treść**. Wynika to z dużej liczby kategorii punktów możliwych do przyznania wynoszącej 26 (punktacja 0-25). Największą zgodność eksperci osiągnęli w kryterium **szczególne walory pracy**, lecz jedynie w przypadku sześciu prac którykolwiek z ekspertów przyznał 1 punkt za to kryterium, zatem bardzo duża zgodność wynika z minimalnej wariancji ocen. Zgodność pomiędzy ekspertami bez uwzględnienia podziału na kryteria wynosiła 56% dla ekspertów 1 i 2, 61% dla ekspertów 2 i 3 oraz 68% dla ekspertów 1 i 3. Natomiast zgodność poszczególnych ekspertów z uzgodnionymi przez nich ocenami była równa 71%, 79% i 75% (w kolejności eksperci: 1, 2 i 3).

Rysunek 7.2. Zgodność między ekspertami w poszczególnych kryteriach jako procent takich samych ocen

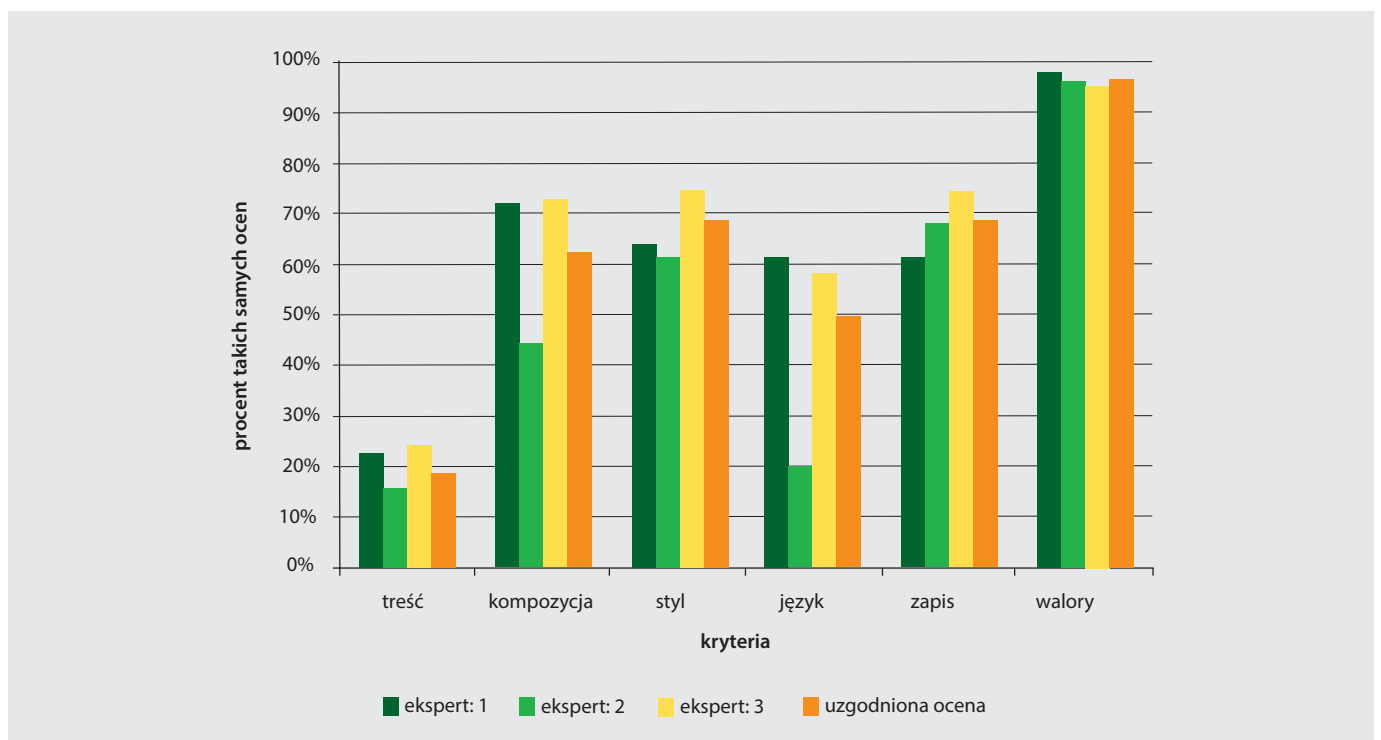


¹⁶

W przypadku obliczenia średniej dla kilku egzaminatorów również korzystamy z modelu statystycznego, choć jest on bardzo uproszczony.

W drugim kroku sprawdzono zgodność ocen pochodzących od ekspertów (zarówno indywidualnych, jak i uzgodnionych) z ocenami „prawdziwymi” z zastosowanego modelu HRM-SDT. Można to interpretować jako sytuację, kiedy model statystyczny byłby traktowany jako jeden z ekspertów. Otrzymane wyniki przedstawiono na rysunku 7.3. Znow z najmniejszą zgodnością mamy do czynienia w przypadku kryterium *treść*, a największą w przypadku *szczególnych walorów pracy*, co nie powinno dziwić. Otrzymano wyraźnie niższą zgodność pomiędzy ocenami „prawdziwymi” z modelu, a jednym z ekspertów (nr 2). W przypadku dwóch pozostałych ekspertów poziom zgodności w poszczególnych kryteriach nie odbiega znacząco od zgodności pomiędzy ekspertami (por. rysunek 7.2). Średnio zgodność pomiędzy ekspertami i ocenami „prawdziwymi” z modelu wynosiła 63%, 51% i 66% (kolejno dla ekspertów: 1, 2 i 3). W związku z tym, że ekspert nr 2 miał mniejszą zgodność z dwójką pozostałych ekspertów, można wnioskować, że zastosowany model nie uzyskuje znacząco niższej zgodności w ocenach niż eksperci. Zgodność dla ocen „prawdziwych” z modelu i ostatecznej oceny uzgodnionej przez ekspertów bez uwzględnienia podziału na kryteria wynosiła 61%.

Rysunek 7.3. Zgodność między ekspertami i oceną „prawdziwą” z modelu HRM-SDT w poszczególnych kryteriach jako procent takich samych ocen



Do oszacowania zgodności pomiędzy oceniającymi można się posłużyć specjalnie do tego celu zbudowanymi współczynnikami zgodności kappa. Podstawowa wersja, zwana kappą Cohena, pozwala na zbadanie zgodności pomiędzy dwoma oceniającymi. Istnieje też druga wersja tego współczynnika, zwana kappą Fleissa, która pozwala na jednoczesne badanie zgodności pomiędzy wieloma oceniającymi. Bazują one na zgodności w przyznanych ocenach, co zostało przedstawione powyżej. Zaletami tych współczynników jest to, że uwzględniają one możliwość pojawienia się przypadkowej zgodności w ocenianiu. Współczynniki te mają zakres 0–1, gdzie 0 oznacza brak zgodności, a 1 pełną zgodność. Współczynniki te nie posiadają szeroko akceptowanej interpretacji co do wielkości, a więc poziomu zgodności i często są z tego powodu krytykowane. W przypadku opisywanych danych można ich jednak użyć bez większych przeszkód, gdyż będą one służyły jedynie do porównania bardzo podobnych pod kątem pochodzenia zestawów ocen. Dla wszystkich 121 prac analizowane były indywidualne oceny pochodzące od ekspertów, przy czym w trzech sytuacjach oceny jednego z ekspertów zastąpiono ocenami „prawdziwymi” z modelu. Możemy zatem porównać współczynniki bez wypowiadania się co do poziomu zgodności, a jedynie sprawdzić, czy istnieją różnice pomiędzy tymi zestawami ocen.

Dla analizowanych danych obliczono wskaźniki kappa Fleissa, których wartości dla poszczególnych kryteriów oceny przedstawiono w tabeli 7.2 (wszystkie współczynniki są istotne statystycznie na poziomie $p < 0,01$). Najmniejszą zgodność mierzoną współczynnikiem kappa można ponownie zaobserwować dla kryterium *treść*. Z trzech sytuacji, kiedy oceny jednego z ekspertów zastąpiono ocenami „prawdziwymi” z modelu, w dwóch współczynniki kappa są niższe niż te dla samych ekspertów, a w jednej wyższe

(oprócz *stylu* i *szczególnych walorów pracy*). Oznacza to, iż jeden z ekspertów osiągał z dwoma pozostałymi mniejszą zgodność niż oceny oszacowane przez model statystyczny. Różnice pomiędzy wartościami opisywanych współczynników dla samych ekspertów i przypadków wymiany ocen jednego z ekspertów na oceny „prawdziwe” z modelu nie są duże. W przypadkach niższej zgodności modelu niż jednego z ekspertów maksymalnie sięgają około 0,15, a w przypadku wyższej zgodności około 0,18. Można w związku z tym uznać, że użyty model statystyczny nie szacuje ostatecznych ocen uczniów gorzej, niż czynią to eksperci.

Tabela 7.2. Współczynniki kappa dla indywidualnych ocen ekspertów i ocen „prawdziwych” z modelu (kappa Fleissa)

	Treść	Kompozycja	Styl	Język	Zapis	Walory	Bez podziału
Eksperci między sobą	0,183	0,474	0,604	0,324	0,490	0,486	0,526
Eksperci 1 i 2 i oceny „prawdziwe”	0,133	0,396	0,482	0,191	0,477	0,375	0,466
Eksperci 2 i 3 i oceny „prawdziwe”	0,156	0,410	0,538	0,209	0,612	0,443	0,502
Eksperci 1 i 3 i oceny „prawdziwe”	0,191	0,612	0,565	0,506	0,510	0,333	0,589

8. Realizacja badań

8.1. Realizacja zadań wynikających z przyjętego schematu badania

Przyjęty schemat badania warunkował zarówno dobór prac do oceniania, jak i sposób losowania próby osób oceniających prace. Dlatego też badania musiały być przeprowadzone w tym samym czasie przy obecności wszystkich egzaminatorów włączonych do próby i przy rygorystycznym przestrzeganiu schematu przydziału prac (por. rozdział 6.7). Nieobecność egzaminatora lub błąd w przydziale prac stanowiłby istotne zagrożenie dla jakości prowadzonych w badaniach analiz.

Tak jak zostało przedstawione w części metodologicznej, w badaniach oprócz 29 egzaminatorów z każdej komisji uczestniczył losowo wybrany przewodniczący zespołu egzaminatorów (PZE). Podobnie jak to ma miejsce podczas sesji egzaminacyjnej, pracami wszystkich zespołów kierowali koordynator oceniania prac z języka polskiego i koordynator oceniania prac maturalnych z matematyki. Sesja oceniania została też poprzedzona szkoleniem PZE zorganizowanym przez koordynatorów tydzień przed sesją oceniającą. W czasie sesji oceniającej koordynacja odbywała się z wykorzystaniem platformy MOODLE. W każdym z ośmiu ośrodków oceniania prace były monitorowane przez przedstawicieli zespołu badawczego (ZAOU).

Do badania związku wyniku oceniania wypracowania z języka polskiego z cechami graficznymi pisma zdającego zostały podjęte przez grafologa następujące działania:

1. Przegląd 1000 wypracowań maturalnych z języka polskiego i analiza wyrobienia, czytelności i wielkości pisma zdających.
2. Stworzenie katalogu form graficznych poszczególnych znaków alfabetu, które ulegają psychomotorycznym modyfikacjom w stosunku do elementarnego wzorca.
3. Opracowanie trzystopniowej skali czytelności tekstu wypracowań z egzaminu maturalnego z języka polskiego na podstawie stworzonego katalogu.
4. Wybór 120 wypracowań reprezentujących wszystkie stopnie ze stworzonej skali.

Wybrane przez grafologa 120 prac różniących się skalą czytelności zostało przepisanych na komputerze z zachowaniem struktury tekstu, skreśleń, nadpisań itp. Każda z prac przepisanych na komputerze była oceniana przez 4 egzaminatorów.

8.1.1. Osoby oceniające (punktujące) zadania w pracach egzaminacyjnych

W badaniach poddano analizie ocenianie prac na poziomie maturalnym z języka polskiego i matematyki. Z języka polskiego badano ocenianie wypracowania z arkuszy egzaminacyjnych na poziomie podstawowym (MPP) z 2011 i 2012 roku z uwzględnieniem dwóch tematów, które mogli wybierać zdający. Z matematyki poddano analizie ocenianie prac z arkuszy egzaminacyjnych na poziomie podstawowym (MMP) i na poziomie rozszerzonym (MPR) z 2011 i 2012 roku.

W badaniu efektu egzaminatora podczas oceniania prac egzaminu maturalnego z języka polskiego próba oceniających (punktujących odpowiedzi na zadania) obejmowała:

1. egzaminatorów posiadających uprawnienia do oceniania i doświadczenie w ocenianiu (egzaminatorzy zatrudniani do oceniania przez okręgowe komisje egzaminacyjne),
2. nauczycieli języka polskiego bez uprawnień egzaminatora,
3. studentów polonistyki.

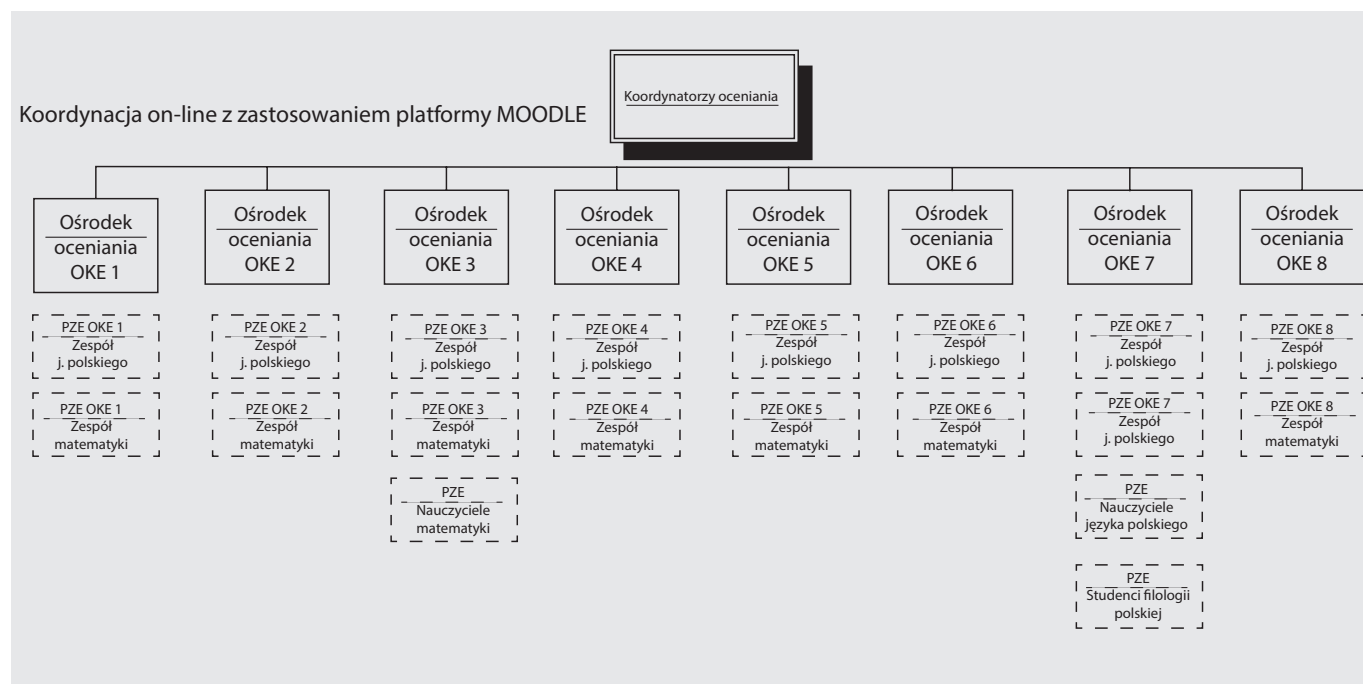
W przypadku oceniania prac maturalnych z matematyki oceniającymi byli egzaminatorzy i nauczyciele szkół ponadgimnazjalnych nieposiadający uprawnień egzaminatora. Charakterystyka próby oceniających uczestniczących w badaniach przedstawiona jest w tabeli 6.1. w rozdziale 6 zatytułowanym *Metodologia i sposób realizacji badania*.

8.2. Opis koordynacji i monitorowania badania

Ocenianie prac zorganizowane zostało w taki sam sposób, jak przebiega ocenianie w trakcie sesji egzaminacyjnej. Koordynowane było przez dwóch koordynatorów mających doświadczenie w koordynowaniu oceniania w trakcie sesji egzaminacyjnej (koordynator oceniania z języka polskiego i koordynator z matematyki). Tydzień przed badaniami głównymi koordynatorzy przeprowadzili szkolenie przewodniczących zespołów egzaminatorów (PZE), podczas którego przypomnieli zasady oceniania prac z 2011 i 2012 roku i dostarczyli materiały opracowane do szkolenia przewodniczących i egzaminatorów. Członkowie zespołu badawczego przedstawili w trakcie spotkania cele badania i obowiązujące procedury organizacyjne wynikające z koncepcji badania. Dla zespołów egzaminatorów, reprezentujących poszczególne okręgowe komisje egzaminacyjne, ośrodki oceniania, w których pracowali egzaminatorzy, zlokalizowane były w miastach, w których jest siedziba danej OKE (por. rysunek 8.1). Nauczyciele matematyki niebędący egzaminatorami oceniali swoje prace w ośrodku zlokalizowanym w Krakowie (ośrodek 3 na rysunku 8), a nauczyciele i studenci oceniający prace z języka polskiego w ośrodku 7 w Warszawie.

Koordynacja oceniania w ośmiu ośrodkach obejmujących 19 zespołów (9 zespołów z matematyki i 10 z języka polskiego) prowadzona była z wykorzystaniem platformy MOODLE. Dla przewodniczących zespołów egzaminatorów (PZE) dostępny był serwis on-line podczas całego czasu oceniania. Serwis ten zapewniał bezpośredni kontakt PZE z koordynatorem oceniania prac z języka polskiego i koordynatorem oceniania prac z matematyki. Ocenianie odbywało się w weekend 4–5 października 2013 roku. W piątek przewodniczący spotkali się z przedstawicielami wykonawcy odpowiedzialnego za organizację oceniania oraz przedstawicielami zespołu badawczego i przygotowali prace do dystrybucji zgodnie z planem zapewniającym właściwe powiązanie prac ocenianych przez egzaminatorów w poszczególnych zespołach. W sobotę po szkoleniu przeprowadzonym przez przewodniczących zespołów egzaminatorów odbyło się ocenianie zgodnie z obowiązującymi zasadami podczas sesji. Przewodniczący byli w ciągłym kontakcie on-line z koordynatorem danego egzaminu i w razie potrzeby uzgadniali konieczne kwestie. Zgodnie z założeniami wynikającymi z koncepcji badawczej nie stosowano obowiązującej w sesji egzaminacyjnej procedury podwójnego oceniania (10% prac). W zespołach oceniających prace z matematyki zastosowano ocenianie „całymi pracami” niezależnie od tego czy w danej OKE stosowane jest takie rozwiązanie, czy też egzaminatorzy specjalizują się w ocenianiu poszczególnych przydzielonych zadań.

Rysunek 8.1. Schemat organizacji sesji oceniania w trakcie badań



We wszystkich zespołach, oprócz przedstawicieli wykonawcy odpowiedzialnych za prawidłową organizację warunków oceniania, byli obecni przedstawiciele zespołu badawczego (ZAOU), zapewniając właściwą realizację zadania ze względu na przyjęte procedury badawcze. Po zakończeniu procesu oceniania prace z każdego ośrodka zostały sprawdzone pod względem właściwego opisu

i przetransportowane do ośrodka skanowania i przetwarzania danych. W trakcie przetwarzania danych zespół badawczy zapewnił procedury ułatwiające weryfikację danych, a w szczególności sprawdzanie błędów, które mogły powstać przy kodowaniu prac przez egzaminatorów.

8.3. Opis przetwarzania i weryfikowania danych

Złożoność logistyczna fazy realizacyjnej badania wymagała wdrożenia odpowiednich procedur w celu zapewnienia spójności otrzymanych w wyniku oceniania danych. Procedury te miały na celu:

1. Zapewnienie jednoznacznego i wiarygodnego przypisania każdej z kart oceny wypełnianej przez egzaminatora do ocenianej pracy oraz oceniającego ją egzaminatora, zgodnego z założonym schematem badania.
2. Uniknięcie opuszczeń w przenoszeniu przyznanych ocen na karty oceny.
3. Maksymalną automatyzację procesu digitalizacji kart oceny.
4. Weryfikację podstawowych niezmienników wynikowego zbioru danych (kontrola dozwolonych zakresów wartości i filtrów poszczególnych zmiennych).

W celu zapewnienia jednoznaczności i wiarygodności przypisania kart ocen do prac i egzaminatorów zarówno numer pracy, jak i identyfikator egzaminatora kodowane były na karcie oceny podwójnie. Numer pracy kodowany i odnotowywany był przez ręczne dopisanie przez egzaminatora ostatnich 4 cyfr numeru pracy oraz przez naklejenie nalepki z kodem kreskowym i numerem pracy. Niezgodność tych wartości powodowała uruchomienie procedury wyjaśniania właściwego dla danej karty oceny numeru pracy. Identyfikator egzaminatora odnotowywany był poprzez ręczne wpisanie oraz zakreszenie odpowiednich krutek wyboru. Dodatkowo w trakcie zdawania materiałów przez egzaminatora osoba koordynująca realizację badania w danym ośrodku sprawdzała spójność tych zapisów na wszystkich kartach oceny danego egzaminatora. Niezgodność obydwu wpisanych wartości była podstawą do uruchomienia procedury wyjaśniania właściwego dla danej karty oceny identyfikatora egzaminatora. Karty oceny od każdego egzaminatora pakowane były w oddzielne koperty, a skanowanie kart oceny prowadzone było w ten sposób, aby ten podział zachować (pliki wynikowe skanowania prac z każdej z kopert zapisywane były w oddzielnym katalogu). Kolejnym poziomem zabezpieczeń była weryfikacja wartości zmiennych z kodem pracy oraz identyfikatorem egzaminatora w powstałym po digitalizacji kart oceny zbiorze danych z założonym w badaniu schematem przydziału prac do egzaminatorów. Wykrycie par „praca–egzaminator” niepasujących do schematu oraz par „praca–egzaminator”, które znajdowały się w schemacie, lecz nie było ich w wynikowym zbiorze danych, powodowało uruchomienie procedur wyjaśniających. Ostatni stopień zabezpieczeń stanowiła weryfikacja ocen odstających, gdzie ocena przyznana za dane zadanie (lub kryterium oceny z języka polskiego) w danej pracy przez jednego z egzaminatorów różniła się wyraźnie od ocen przyznanych przez pozostałych egzaminatorów.

Procedury wyjaśniające w wypadku wykrycia nieprawidłowości w przypisaniu karty oceny do pracy bądź egzaminatora opierały się przede wszystkim na tym, że każdy z egzaminatorów dysponował przygotowaną z góry listą naklejek z kodami prac obejmującą jedynie oceniane przez niego prace. O ile więc podczas ręcznego dopisywania 4 ostatnich cyfr numeru pracy egzaminator mógł popełnić dowolne błędy, o tyle przyklejając nalepki mógł jedynie zamienić ze sobą prace z puli tych, które sam ocenia. Stąd pojawienie się nieistniejącej w założonym schemacie pary „praca–egzaminator”, gdzie numer pracy odczytywany był z naklejki, wskazywało na błąd w kodzie egzaminatora. Błędy tego rodzaju były łatwe do weryfikacji, gdyż skanowanie kart oceny odbywało się paczkami poszczególnych egzaminatorów i wystarczyło porównać kod egzaminatora odnotowany na problematycznej karcie oceny z kodami na pozostałych kartach oceny tego samego egzaminatora. W ten sam sposób rozwiązywano wątpliwości związane z różnicą w kodzie egzaminatora pomiędzy zapisem ręcznym i przez zaczernianie pól. Łącznie zweryfikowano 8 błędów tego typu. Weryfikacja niezgodności pomiędzy kodami prac odnotowanymi ręcznie oraz sczytanymi z naklejki, jak również obserwacji odstających była bardziej złożona. Ułatwiała ją wspomniana już właściwość, że numery prac odczytane z naklejek mogły być co najwyżej pozamieniane w obrębie tego samego egzaminatora. Wybór właściwego kodu pracy sprowadzał się więc do analizy wszystkich prac danego egzaminatora, w których występowały niezgodności kodów prac, i odwołanie się do dostępnych informacji kontekstowych, np.:

1. tematu pracy odnotowanego na karcie i wynikającego z pierwszej części numeru pracy, wydrukowanego na karcie odpowiedzi;

2. notatek naniesionych na ocenianych pracach przez egzaminatora (kryterium informacyjne zwłaszcza w wypadku prac z polskiego, gdzie egzaminator najczęściej najpierw odnotowuje oceny na pracy, a dopiero na koniec przenosi je na kartę oceny);
3. podobieństwa rozważanych numerów prac (egzaminatorom zdarzało się przyklejać nalepkę o takich samych ostatnich 4 cyfrach kodu pracy, lecz różniącej się na 4 pierwszych cyfrach kodu, których nie musieli wpisywać ręcznie na karcie odpowiedzi).

Łącznie zweryfikowano 8 błędów tego typu. Wykryto także jeden błąd, w którym kod pracy na nalepce nie odpowiadał żadnej z użytych w badaniu prac. Ponieważ była to sytuacja jednostkowa, odpowiedni kod pracy znaleziono, łatwo odwołując się do założonego przydziału prac do egzaminatorów.

Aby uniknąć problemu przypadkowych opuszczeń w przenoszeniu przyznanej oceny na kartę oceny przez egzaminatora, wszystkie karty oceny danego egzaminatora sprawdzane były przed ich przyjęciem najpierw przez przewodniczącego zespołu, a następnie przez osobę koordynującą realizację badania w danym ośrodku.

Automatyzację digitalizacji kart oceny osiągnięto przez zastosowanie profesjonalnych skanerów i oprogramowania do rozpoznawania tekstu (OCR). Wykorzystano te same profesjonalne urządzenia i oprogramowanie, które służą do digitalizacji kart oceny w sesjach egzaminacyjnych w OKE Łomża. Aby uczynić proces rozpoznawania tekstu ze skanów jak najbardziej rzetelnym, do identyfikacji prac zastosowano kody kreskowe, natomiast wszelkie wartości wpisywane odręcznie były jedynie wartościami liczbowymi, wpisywanymi w odpowiednio oznaczone pola – po jednej cyfrze w każdym polu.

Założone niezmienniki wynikowego zbioru danych (kontrola dopuszczalnych zakresów wartości zmiennych oraz spełnienia warunków filtrów decydujących o tym, czy dla danej obserwacji dana zmienna powinna posiadać wartość) zweryfikowano za pomocą internetowego weryfikatora zbiorów danych IBE (<https://zai.ibe.edu.pl/zbiory>).

Odnosząc liczbę wykrytych podczas przetwarzania i weryfikacji danych błędów do skali i stopnia złożoności logistycznej realizacji badania, skuteczność zastosowanych procedur należy ocenić wysoko. Wydają się one godne polecenia w kolejnych badaniach tego typu.

9. Efekt egzaminatora w ocenianiu maturalnych prac pisemnych z języka polskiego i matematyki

Efekt egzaminatora jest naturalnym zjawiskiem, którego nie da się całkowicie wyeliminować z procesu oceniania, można jedynie dążyć do jego minimalizacji. W rozdziale 5 opisano szereg aspektów efektu egzaminatora, a także przytoczono wyniki dotychczasowych, polskich badań w tym zakresie. Wyniki badania na poziomie maturalnym (Dolata, Putkiewicz i Wiłkomirska, 2004) wyraźnie wskazują, że efekt egzaminatora występuje z różną intensywnością w różnych egzaminach maturalnych. Długie wypowiedzi pisemne są bardziej nim obciążone niż mocno ustrukturyzowane zadania z matematyki. Podobnie jak w cytowanych badaniach, w wynikach badania opisanego w niniejszym opracowaniu można zauważyć różną siłę efektu egzaminatora w analizowanych egzaminach – z języka polskiego i matematyki. Możliwe jest zatem przygotowywanie zadań egzaminacyjnych i organizacji oceniania w sposób zmniejszający efekt egzaminatora i zwiększający rzetelność oceniania.

Poniższy rozdział zawiera szczegółowy opis wyników przeprowadzonego badania dla egzaminu maturalnego z języka polskiego i matematyki. Prezentowane są wyniki nie tylko odnoszące się do siły efektu egzaminatora na poziomie indywidualnym. Szczególną uwagę poświęcono wynikom na poziomie okręgowych komisji egzaminacyjnych. Podyktowane to jest próbą oceny, na ile spójny jest system egzaminacyjny. Regionalne zróżnicowanie siły efektu egzaminatora może budzić pewien niepokój, gdyż w spójnym systemie egzaminacyjnym oczekiwalibyśmy braku systematycznych różnic. Dodatkową uwagę poświęcono aspektowi pisma w pracach z języka polskiego. Interesującym zagadnieniem jest to, na ile rodzaj pisma (odręczne lub komputerowe) jest związany z ocenianiem (por. rozdział 9.3). W rozdziale 9.4 podjęto także próbę określenia, czy takie cechy jak czytelność pisma są znaczące dla siły efektu egzaminatora. W związku z tym, że analizowane egzaminy maturalne posiadały próg zaliczenia, w rozdziale 9.6 opisano problemy związane z tym faktem. Niepokojące jest zjawisko zniekształcenia rozkładów wyników w okolicach progu zdawalności, co prowadzi do obniżenia rzetelności egzaminów. Cały rozdział zamyka analiza badań ankietowych, które zostały przeprowadzone na egzaminatorach w trakcie oceniania prac. Brak wyraźnego związku pomiędzy wskaźnikami efektu egzaminatora i cechami społeczno-demograficznymi czy doświadczeniem egzaminatorów wskazuje na brak systematycznych różnic w sile efektu egzaminatora na poziomie indywidualnym.

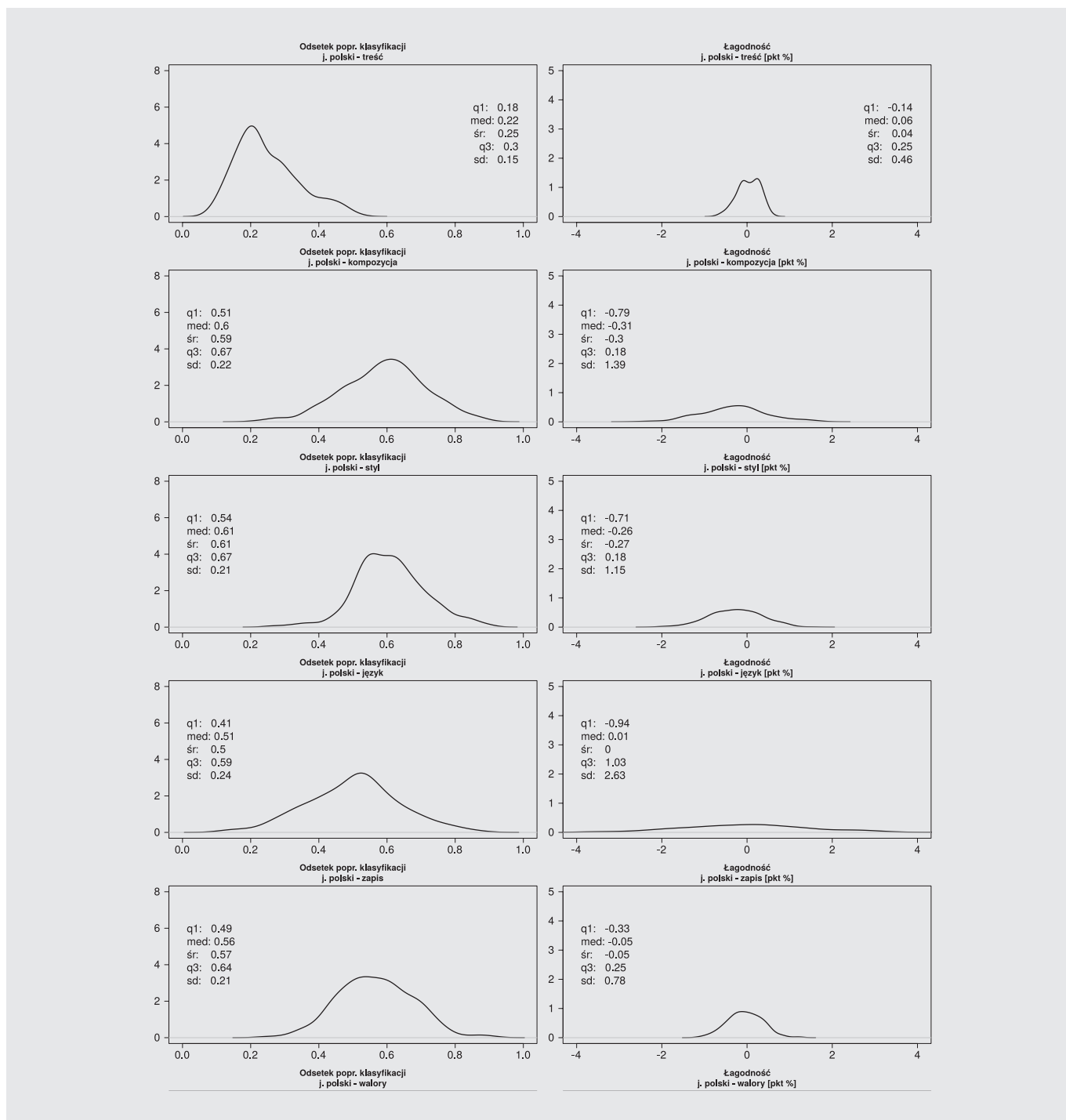
9.1. Język polski

Te same kryteria oceny wypracowania z języka polskiego charakteryzują się bardzo zbliżonymi wartościami parametrów i wskaźników pomiędzy tematami i latami. Jest to właściwość pożądana, oznacza bowiem, że wpływ efektu egzaminatora na ocenę wypracowania z języka polskiego nie zależy w sposób istotny od wyboru tematu ani roku pisania egzaminu. Zaobserwowano natomiast bardzo duże zróżnicowanie wartości wskaźników pomiędzy poszczególnymi kryteriami oceny, dzielące je na trzy wyraźnie od siebie odmienne grupy (rysunek 9.1):

1. **treść** – oceniana z bardzo dużym błędem losowym (nierzetelnie) i niewielkim błędem systematycznym (małe różnice łagodności egzaminatorów);
2. **kompozycja, styl, język, zapis** – oceniane z umiarkowanym błędem losowym i dużym błędem systematycznym (duże różnice łagodności egzaminatorów);
3. **dodatkowe walory** – oceniane z małym błędem losowym (rzetelnie) i małym błędem systematycznym (małe różnice łagodności egzaminatorów).

Dalsze analizy przeprowadzone zostały według powyższego podziału, przy czym uwzględniono w nich jedynie egzaminatorów. Wyniki dla grup kontrolnych, nauczycieli niebędących egzaminatorami oraz studentów omówione zostały w oddzielnym rozdziale, oddzielnie przeanalizowano także błędy kardynalne.

Rysunek 9.1. Rozkłady parametrów ρ_{kr} oraz b_{kr} w podziale na kryterium oceny z języka polskiego (tylko egzaminatorzy)



*q1, q3 – pierwszy i trzeci kwartyl, sd – odchylenie standardowe

Im bardziej skupione rozkłady, tym mniejsze zróżnicowanie egzaminatorów ze względu na dane kryterium. W wypadku parametru ρ_{kr} efekt egzaminatora jest tym mniejszy, im bardziej rozkład przesunięty jest w prawo (im więcej egzaminatorów osiąga wysokie odsetki poprawnych klasyfikacji). W wypadku parametru b_{kr} efekt egzaminatora jest tym mniejszy, im bardziej skupiony jest rozkład (im mniejsze zróżnicowanie łagodności egzaminatorów).

9.1.1. Treść

Treść jest bardzo nietypowym kryterium oceny, gdyż oceniana jest na skali o aż 26 poziomach wykonania. W praktyce układania testów bardzo rzadko stosuje się tak długie skale, zaś w systemach egzaminacyjnych, w których egzaminy poddaje się ścisłej

kontroli psychometrycznej, tak długich skal w ogóle się nie spotyka. Duża liczba poziomów wykonania prowadzi do obniżenia rzetelności oceniania, co znajduje potwierdzenie w badaniach. Co prawda wartości standaryzowanego współczynnika dyskryminacji d_s nie odbiegają dla **treści** od tych dla kryteriów **kompozycja**, **styl**, **język** czy **zapis** (tabela 9.1). Jednak pozwala to przewidywać, że skrócenie skali kryterium oceny za **treść** zaowocowałoby osiągnięciem rzetelności zbliżonej do wymienionych kryteriów oceny. Tymczasem skala w kryterium **treść** jest wielokrotnie dłuższa, co wyraźnie negatywnie rzutuje na obserwowaną rzetelność oceniania. Mediana odsetka poprawnych klasyfikacji wynosi 22%. Oznacza to, że przeciętny egzaminator poprawnie klasyfikował tylko dwie na dziewięć sprawdzanych prac. Żaden z egzaminatorów biorących udział w badaniu nie przekroczył progu 60% poprawnych klasyfikacji. Niskie wartości odsetka poprawnych klasyfikacji mogą być również spowodowane dużymi różnicami w łagodności egzaminatorów. Niewielkie różnice wartości wskaźnika b_{kr} pomiędzy egzaminatorami (rozstęp międzykwartkowy 0,39 punktu procentowego na egzaminie, odchylenie standardowe 0,46 punktu procentowego na egzaminie) świadczą jednak przeciwko tej hipotezie i potwierdzają decydujący wpływ niskiej rzetelności na niski odsetek poprawnych klasyfikacji. Warto podkreślić, że nie można argumentować, że przy tak długiej skali błędne zakwalifikowanie z różnicą o jeden czy dwa punkty jest dopuszczalne, tego rodzaju błędy zmieniają bowiem uszeregowanie uczniów według liczby punktów, którą zdobyli na egzaminie maturalnym, co jest potem np. podstawą przyjęcia na studia wyższe. W wypadku, gdy nie jesteśmy w stanie zapewnić rzetelnego oceniania na długich skalach – a z badania jasno wynika, że w wypadku kryterium **treść** z języka polskiego nie jest to możliwe – należy zmniejszyć liczbę poziomów wykonania zadania. Nie odbiera to możliwości nadania kryterium oceny stosownego wkładu w punktację całego egzaminu – wystarczy taką skróconą skalę przemnożyć przez odpowiednią wagę. Rozwiązanie to jest już zresztą stosowane w wypadku kryteriów oceny **kompozycja**, **styl**, **język** czy **zapis**.

Tabela 9.1. Parametry rozkładu współczynnika dyskryminacji egzaminatora (d) oraz jego standaryzowanej wartości d_s dla poszczególnych kryteriów oceny z języka polskiego (tylko egzaminatorzy)

	Treść		Kompozycja		Styl		Język		Zapis		Walory	
	d	d_s	d	d_s	d	d_s	d	d_s	d	d_s	d	d_s
mnożnik		13		2		2		3		2		2,5
1 kwartył	0,7	9,5	4,0	8,0	4,1	8,2	3,1	9,4	2,8	5,5	5,5	13,8
mediana	1,0	13,3	5,4	10,8	5,5	10,9	4,9	14,8	4,7	9,4	7,0	17,6
3 kwartył	1,6	20,3	7,1	14,2	7,2	14,3	6,5	19,6	6,5	13,0	7,9	19,8

W przypadku zniekształceń skali wartości wskaźnika r_{kr} są minimalnie zróżnicowane ($sd = 0,006$), a ich mediana oraz 1 i 3 kwartył wynoszą 1,02. Oznacza to, że średnio egzaminatorzy nie ograniczają skali podczas oceniania treści prac uczniów ani nie wykazują ekstremizmu. Warto jednak mieć na uwadze, że długość tej skali może wpływać na takie wyniki.

9.1.2. Kompozycja, styl, język, zapis

Parametry egzaminatorów dla kryteriów **kompozycja**, **styl**, **język** i **zapis** wypadają w podobny sposób. Kryteria te charakteryzują się przeciętną rzetelnością i dużym zróżnicowaniem łagodności egzaminatorów. Standaryzowany parametr dyskryminacji osiąga wartości zbliżone do tych dla kryterium **treść**, porównywalne z najmniej rzetelnie ocenianymi zadaniami z matematyki i wyraźnie niższe od wartości dla kryterium **walory** (por. tabela 9.1 i tabela 9.2). Pewne różnice zaobserwować można dla kryterium **zapis**, dla którego rozkład parametrów d i d_s jest bardziej lewoskośny (więcej jest egzaminatorów o wyraźnie niższej dyskryminacji) oraz **język**, dla którego standaryzowana dyskryminacja jest nieco wyższa od pozostałych kryteriów (jej 3 kwartył prawie zrównuje się z tym dla **walorów**). Rozstęp międzykwartkowy parametru b_{kr} (łagodności egzaminatorów) waha się od 0,58 (dla **zapisu**) do 1,97 (dla **języka**) punktu procentowego na egzaminie. Oznacza to, że 25% najbardziej łagodnych egzaminatorów przyznaje za te kryteria oceny (średnio) co najmniej o od 0,58 do 1,97 punktu procentowego więcej niż 25% najbardziej surowych egzaminatorów. Najbardziej jasnym problemem różnic łagodności egzaminatorów ujawnia się jednak przy porównaniu najmniej i najbardziej surowego ocenianego – w zależności od kryterium oceny wyniosła ona od 3,11 (dla **zapisu**) do 9,34 (dla **języka**) punktu procentowego na egzaminie, co odpowiada ponad połowie punktów, które można zdobyć za każde z tych kryteriów. Dla porównania, różnica między skrajnymi egzaminatorami w wypadku kryterium **treść** to 1,44 punktu procentowego na egzaminie, co w dodatku stanowi jedynie niecałe

4% skali tego kryterium. Omówione czynniki przekładają się na dość niskie odsetki poprawnych klasyfikacji – średnia wartość parametru ρ_{kr} waha się zależnie od kryterium oceny od 0,5 (*język*) do 0,61 (*styl*).

Zniekształcenia skali w przypadku tych kryteriów mają zbliżone rozkłady, dla których wartości mediany oraz 1 i 3 kwartyla przedstawia tabela 9.2. Mediany wszystkich czterech kryteriów są większe od 1, przy czym dla *kompozycji* i *stylu* wynoszą one odpowiednio 1,39 i 1,47, a dla *języka* i *zapisu*: 1,15 i 1,17. Oznacza to, że w przypadku tych czterech kryteriów przeciętnie egzaminatorzy mają tendencję do ograniczania skali (trudno jednak określić wielkość tego efektu). Jednakże w przypadku kryterium *zapisu* 1 kwartyl ma wartość 1, zatem jedna czwarta ocenających ma tendencję do przyznawania ocen skrajnych. Podobnie w przypadku kryterium *języka* 1 kwartyl jest dość niski (1,08), co świadczy o istnieniu grupy egzaminatorów przejawiających tendencję do przyznawania ocen skrajnych w tym kryterium. Dla kryterium *języka* rozstęp międzykwartkowy jest przy tym najmniejszy z analizowanych kryteriów i wynosi 0,14 (dla porównania: 0,26 dla *kompozycji*, 0,28 dla *stylu* i 0,35 dla *zapisu*), co oznacza nieco mniejszą tendencję do ograniczania skali niż w pozostałych kryteriach. Wykresy rozkładów opisywanych parametrów umieszczono w aneksie 2.

Tabela 9.2. Parametry rozkładów współczynnika r_{kr} dla poszczególnych kryteriów oceny z języka polskiego (tylko egzaminatorzy)

	Treść	Kompozycja	Styl	Język	Zapis	Walory
1 kwartyl	1,02	1,25	1,32	1,08	1,00	0,88
mediana	1,02	1,39	1,47	1,15	1,17	0,93
3 kwartyl	1,02	1,51	1,60	1,22	1,35	0,98

9.1.3. Walory

Niezwykle interesującym kryterium oceny w języku polskim są *dodatkowe walory pracy*. Po kategorii oceny, która w założeniach uwzględniać ma te cechy pracy, których nie udało się uchwycić w pozostałych, dobrze określonych kryteriach oceny, spodziewaliśmy się niskiej rzetelności, a także dużego błędu systematycznego. Tymczasem wyniki estymacji wskazują jednoznacznie, że rzetelność oceny *walorów* jest najwyższa spośród wszystkich kryteriów oceny wypracowania z języka polskiego i porównywalna z tymi dla zadań z matematyki (por. tabela 9.1 i tabela 9.5). Niewielki okazuje się również błąd systematyczny – rozstęp międzykwartkowy oraz odchylenie standardowe parametru b_{kr} (łagodności egzaminatorów) wynoszą odpowiednio 0,21 oraz 0,37 punktu procentowego na egzaminie. Przekłada się to na wysoki odsetek poprawnych klasyfikacji b_{kr} wynoszący średnio 84%. Warto zauważyć, że fakt, że zdecydowana większość prac w populacji (ok. 96% prac) otrzymuje 0 punktów za to kryterium oceny, nie oznacza automatycznie, że będzie ono oceniane rzetelnie i z jednakową surowością przez różnych egzaminatorów (mogliby oni bowiem przydzielać punkty za walory różnym pracom lub przyznawać różną ich liczbę). Wyniki badania wskazują, że takie zjawisko nie występuje i można stwierdzić, że egzaminatorzy mają nie tylko wspólne wyczucie tego, czym są *dodatkowe walory pracy*, jak również tego, jak wiele tych walorów posiada dana praca. Wskaźniki r_{kr} dla kryterium *walorów*, dla których 3 kwartyl wynosi 0,98, 1 kwartyl 0,88, a mediana 0,93 sugerują, że egzaminatorzy w przypadku tego kryterium wykazują pewną tendencję do przydzielania ocen skrajnych. Dokładniejsza analiza pokazuje, że odnotowana tendencja jest praktycznie w całości związana ze skłonnością niektórych egzaminatorów do przyznawania 0 punktów (brak dodatkowych walorów) za kryterium *walory*, podczas gdy zniekształcenia skali dla ostatniego, czwartego progu prawie nie występują. Bardzo niskie były również zniekształcenia dla trzeciego progu.

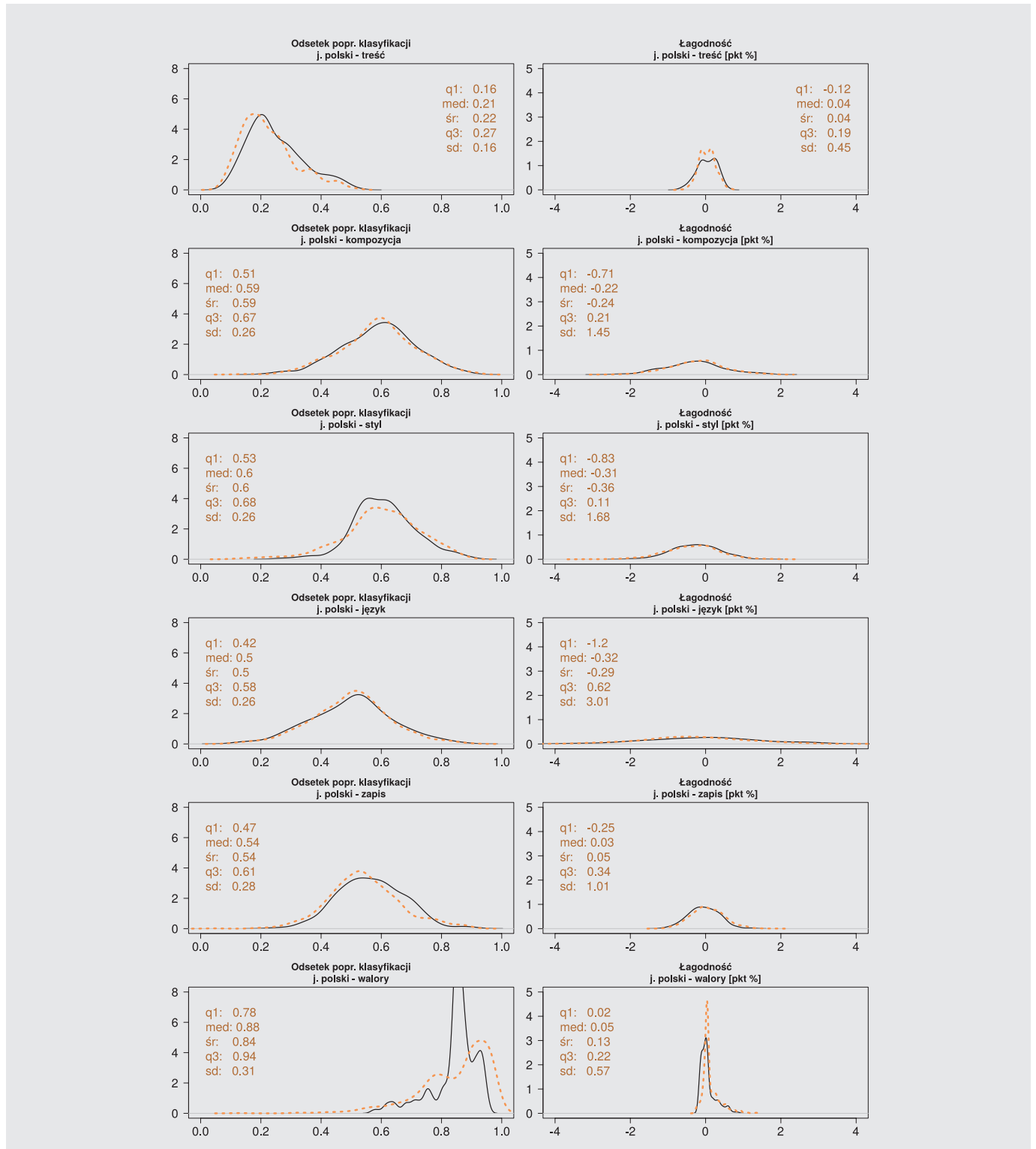
9.1.4. Nauczyciele niebędący egzaminatorami oraz studenci

Wyestymowane parametry egzaminatorów wykazały bardzo niewielkie różnice pomiędzy grupą egzaminatorów a grupą nauczycieli oczekujących na nadanie statusu egzaminatora oraz studentów (patrz rysunek 9.2). Różnice średniego odsetka poprawnych klasyfikacji między grupami nie były istotne statystycznie¹⁷ dla żadnego spośród 24 kryteriów oceny (rozpatrywanych w tym wypadku osobno dla każdego tematu i roku), natomiast średnie wartości parametru b_{kr} w sposób istotny statystycznie różniły się między grupami w 14 spośród 24 kryteriów oceny. Tam, gdzie różnice były istotne statystycznie, nieegzaminatorzy wykazywali się nieco większą łagodnością – różnice wyniosły od 0,39 do 0,03 punktu procentowego na egzaminie. Przy zagregowaniu wyników

¹⁷ Test t Welcha na poziomie istotności 0,01.

do poziomu całego badania (rysunek 9.2, ostatni wykres) nieegzaminatorzy okazali się co prawda istotnie statystycznie¹⁸ bardziej łagodni od egzaminatorów, różnica ta była jednak bardzo niewielka (0,19 punktu procentowego na egzaminie) i, co warto podkreślić, wyraźnie mniejsza od zróżnicowania pomiędzy OKE. W świetle tych wyników można stwierdzić, że siła efektu egzaminatora nie zależy w sposób istotny od tego, czy oceniający posiada status egzaminatora.

Rysunek 9.2. Rozkłady parametrów ρ_{kr} oraz b_{kr} w podziale na kryterium oceny z języka polskiego – egzaminatorzy (czarne linie) oraz studenci i nauczyciele niebędący egzaminatorami (przerywane linie oraz parametry rozkładów)



¹⁸

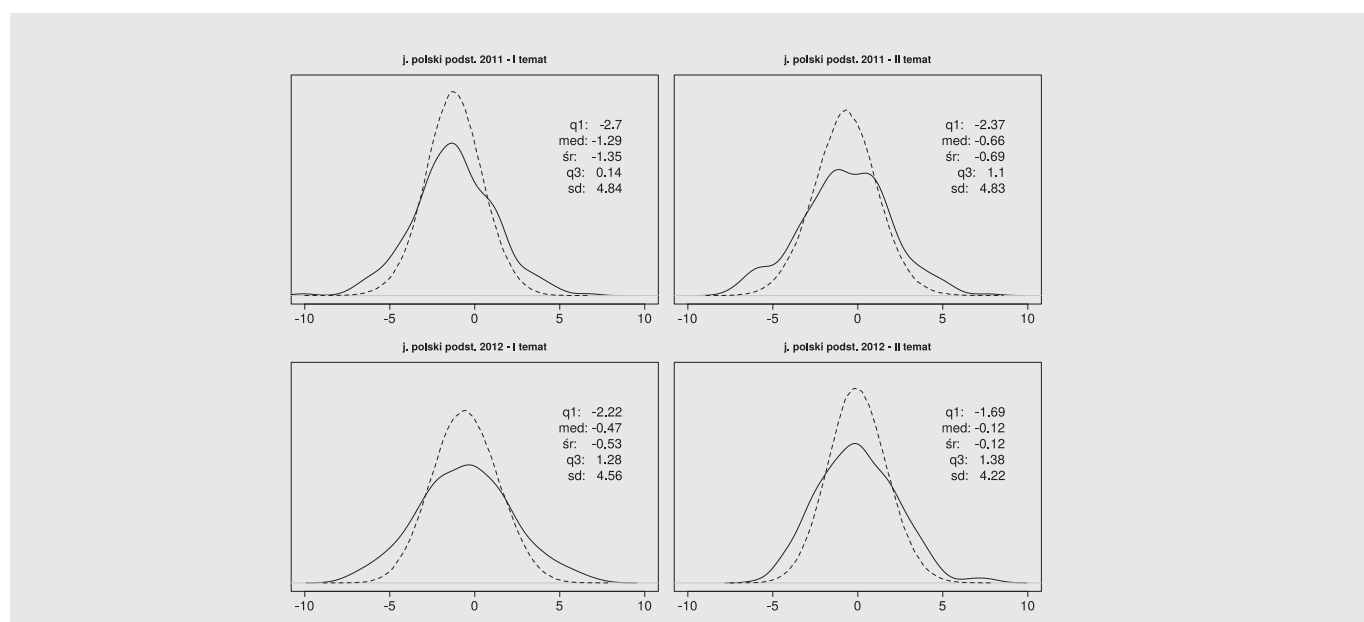
Rozłączność 99% przedziału ufności średniej ze średnią egzaminatorów.

9.1.5. Różnice łagodności na poziomie testów

Ważne wydaje się pytanie, w jaki sposób różnice łagodności egzaminatorów w poszczególnych kryteriach oceny przekładają się na sumaryczny wynik dla całego wypracowania. Wyniki tych analiz przedstawia rysunek 9.3. Linią ciągłą oznaczono na nim rozkłady faktycznych łagodności egzaminatorów zagregowanych do poziomu testu. Uzyskane wyniki są zbliżone dla obydwu lat i tematów egzaminu. Różnica łagodności pomiędzy 25% najbardziej łagodnych a 25% najbardziej surowych egzaminatorów wynosi od co najmniej 3,1 do 3,7 punktu procentowego na egzaminie. Dla ucznia o przeciętnym wyniku egzaminacyjnym z języka polskiego¹⁹ oznacza to przesunięcie w rozkładzie populacyjnym o co najmniej 5 centyli. Zarówno z punktu widzenia ucznia, jak i systemu egzaminacyjnego różnicę tę należy uznać za znaczną. Różnica pomiędzy skrajnie łagodnym a skrajnie surowym na poziomie testu egzaminatorem wyniosła, w zależności od roku i tematu wypracowania, od 13,7 do 16,8 punktu procentowego na egzaminie, co dla ucznia o przeciętnym wyniku egzaminacyjnym oznaczałoby przesunięcie w rozkładzie populacyjnym o ponad 20 centyli. Jest to oczywiście przykład skrajny, pokazuje jednak, jak ważne jest dążenie do zmniejszania zróżnicowania łagodności wśród egzaminatorów.

Ciekawe wydaje się również pytanie, na ile łagodności egzaminatorów w poszczególnych kryteriach oceny są ze sobą powiązane. Innymi słowy, czy dany egzaminator jest raczej konsekwentnie łagodny bądź surowy, czy może jego łagodność w jednym kryterium oceny pozostaje bez związku z łagodnością w pozostałych kryteriach i jaki ma to wpływ na łagodność zagregowaną do poziomu całego wypracowania. Aby odpowiedzieć na to pytanie, możemy porównać rozkłady łagodności egzaminatorów zagregowane do poziomu testu z rozkładami dla sytuacji, w której łagodność w poszczególnych kryteriach oceny byłaby przydzielana egzaminatorowi losowo (ale spośród łagodności faktycznie zaobserwowanych dla danego kryterium oceny w badaniu) – por. rysunek 9.3 – linie przerywane²⁰. Rozkłady faktycznych łagodności na poziomie testu mają nieco większe wariancje, ich średnie są natomiast bardzo zbliżone do rozkładów przy założeniu o losowym związku łagodności egzaminatorów pomiędzy poszczególnymi kryteriami oceny. Wskazuje to na istnienie jedynie niewielkiego związku pomiędzy łagodnością egzaminatorów w poszczególnych kryteriach oceny. Wynik ten można wiązać z zasadniczymi różnicami konstruktów poszczególnych kryteriów oceny z języka polskiego – *treść, kompozycja, język, styl, zapis* czy *walory* są konstruktami na tyle odrębnymi, że egzaminator surowy/łagodny w jednym z nich, wcale nie musi być surowy/łagodny także w pozostałych.

Rysunek 9.3. Rozkłady parametrów b_{tr} (łagodność egzaminatorów wyrażona w punktach procentowych wyniku egzaminacyjnego zagregowana do poziomu testu) w podziale na rok i temat wypracowania z języka polskiego: rzeczywiste (linia ciągła) i w sytuacji braku związku między łagodnością egzaminatorów w poszczególnych kryteriach oceny (linia przerywana)



*q1, q3 – pierwszy i trzeci kwartył, sd – odchylenie standardowe

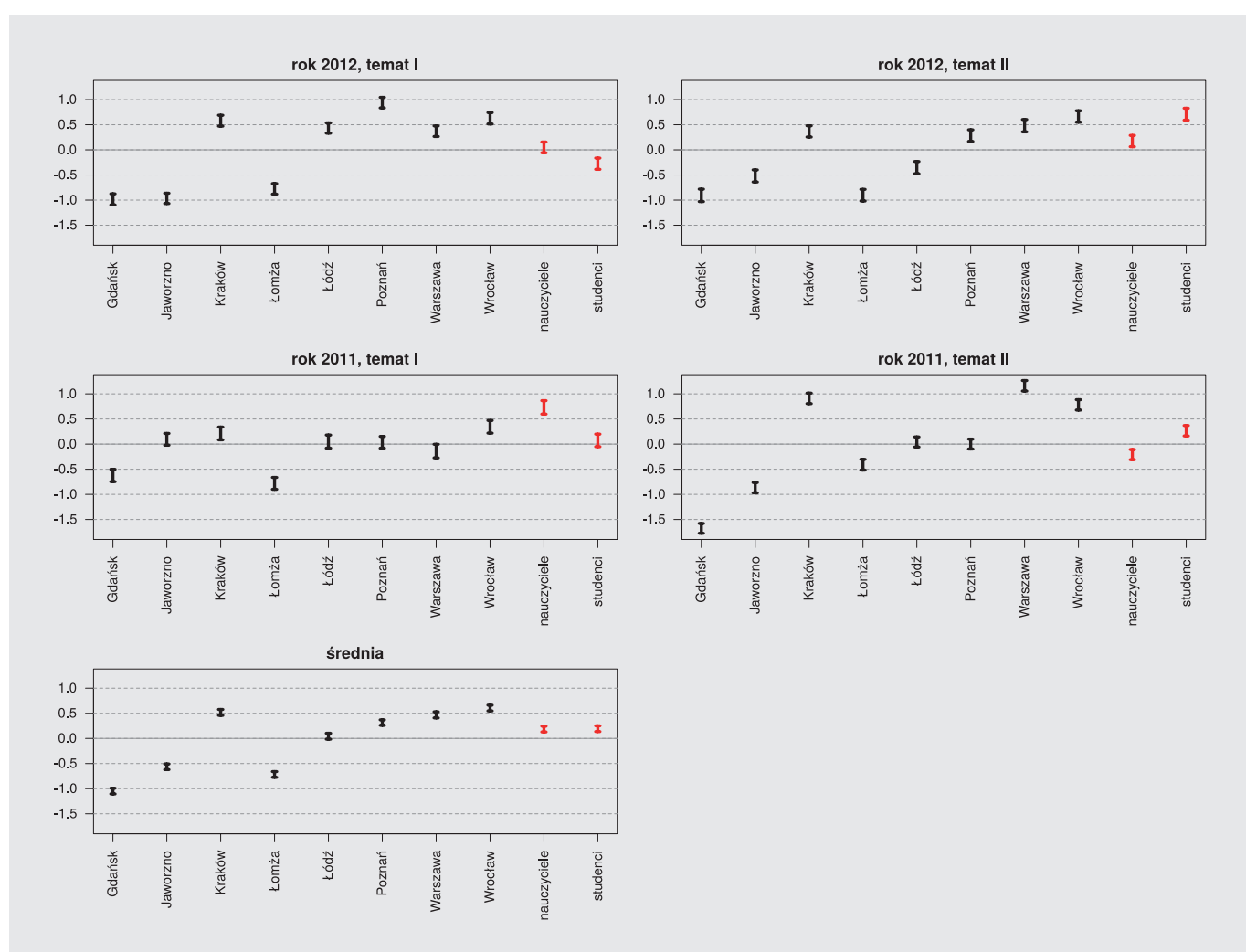
¹⁹ Od 44% do 64% punktów za cały egzamin – w zakresie tym mieszczą się wyniki ok. połowy wszystkich uczniów.

²⁰ Rozkłady uzyskano poprzez tysiąckrotne losowe permutowanie faktycznie uzyskanych w badaniu wartości łagodności egzaminatorów i sumowanie ich do poziomu testu, a następnie policzenie funkcji gęstości z wszystkich uzyskanych sum.

9.1.6. Różnice na poziomie okręgowych komisji egzaminacyjnych

Wyniki badania wykazują istotne statystycznie różnice w łagodności oceniania wypracowania pomiędzy niektórymi okręgowymi komisjami egzaminacyjnymi (rysunek 9.4). Co więcej, dla niektórych OKE różnice te są stałe niezależnie od roku egzaminu i tematu wypracowania, np. OKE Gdańsk i OKE Łomża są niezmiennie najbardziej surowymi komisjami, a OKE Kraków oraz OKE Wrocław plasują się 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 1,64 punktu procentowego na egzaminie. Jednak należy pamiętać, że dotyczą one różnic średniego wyniku w całej OKE. Różnica 1,64 punktu procentowego oznacza więc np., że każdy uczeń w bardziej łagodnej OKE dostał o 1,64 punktu procentowego więcej od ucznia w bardziej surowej OKE (za tak samo napisaną pracę). W wypadku wyniku egzaminacyjnego mieszczącego się w środkowej części skali²¹ oznacza to przesunięcie ucznia w rozkładzie populacyjnym o ok. 3 centyle, co z punktu widzenia zdającego maturę należy uznać za istotną różnicę.

Rysunek 9.4. 99% przedziały ufności łagodności egzaminatorów na poziomie roku, tematu i OKE dla języka polskiego (w punktach procentowych na egzaminie)



*Jako 0 przyjęto średnią łagodność spośród wszystkich egzaminatorów.

Interesujące jest również przyjrzenie się zagadnieniu z poziomu OKE i postawienie pytania, na ile różnice średniej łagodności egzaminatorów na poziomie prac (b_p) pomiędzy OKE tłumaczą różnice średnich wyników uczniów pomiędzy OKE. Jeśli analizę przeprowadzić dla obydwu tematów w obydwu rocznikach (2011, 2012), to różnice średniej łagodności egzaminatorów tłumaczą 10,8% wariancji średnich wyników egzaminacyjnych pomiędzy OKE²², zależność jest więc wyraźna, ale ograniczona. Jeśli jednak powtórzyć

²¹ Od 44% do 64% punktów za cały egzamin – w zakresie tym mieszczą się wyniki ok. połowy wszystkich uczniów.

²² Uwzględniono pełne wyniki egzaminu na poziomie podstawowym, włączając zadania zamknięte.

ją w rozbiciu na tematy wypracowań, okazuje się, że dla tematu I (odwołującego się zwykle do poezji, wybieranego przez ok. ¼ uczniów) wyjaśniane jest jedynie 4,5% wariacji, 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 II tematu wypracowania udział różnicowania 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. Po pierwsze, zaobserwowane różnicowanie oznacza systematyczne różnice w sposobie oceniania maturzystów ze względu na miejsce, w którym podchodzą do egzaminu maturalnego. Po drugie, zakłóca istotnie analizę wyników egzaminacyjnych na poziomie makro (np. porównania między regionami kraju). Na koniec warto odnotować, że opisany związek nie występuje w wypadku egzaminu z matematyki (por. rozdział 9.4.2).

9.1.7. Błędy kardynalne

Błąd kardynalny jest to poważny błąd rzeczowy, którego wystąpienie w wypracowaniu prowadzi do przyznania 0 punktów za wszystkie kryteria oceny, bez względu na pozostałą zawartość wypracowania. Błąd ten nie musi być przy tym bezpośrednio związany z tematem wypracowania czy prowadzoną w nim narracją. Z uwagi na bardzo poważne konsekwencje (niezaliczenie egzaminu) w systemie egzaminacyjnym wprowadzone zostały procedury mające gwarantować, że błędy kardynalne nie będą stwierdzane zbyt pochopnie. Egzaminator, który uważa, że praca zawiera błąd kardynalny, musi skonsultować się z przewodniczącym zespołu egzaminacyjnego, a w wypadku, gdyby nie był on pewny decyzji, dokonywane są dalsze konsultacje z koordynatorem na poziomie OKE. W badaniu każda spośród 897 prac sprawdzana była od 8 do 20 razy. W wypadku, gdy egzaminator stwierdzał błąd kardynalny, był on najpierw konsultowany z przewodniczącym zespołu, a następnie także z koordynatorem oceniania języka polskiego, który musiał potwierdzić istnienie w pracy błędu kardynalnego (i fakt ten odnotowywał). Informacja o stwierdzeniu w danej pracy błędu kardynalnego nie była jednak rozpowszechniana w innych zespołach egzaminacyjnych oraz wśród innych egzaminatorów w tym samym zespole. Pozwoliło to na sprawdzenie, czy egzaminatorzy stwierdzać będą błędy kardynalne w tych samych pracach. Dodatkowo 121 spośród 897 prac oceniona została przez trójkę ekspertów, którzy najpierw oceniali je niezależnie, a potem przyznawali wspólnie ocenę końcową, co pozwoliło zweryfikować, czy egzaminatorzy nie pomijali błędów kardynalnych pomimo tego, że prace faktycznie je zawierały. Do prawie wszystkich prac udało się również przyłączyć rzeczywiste wyniki z egzaminu maturalnego.

W wyniku sprawdzania prac przez egzaminatorów, 10 z nich zostało zakwalifikowanych jako zawierające błędy kardynalne. Okazuje się jednak, że opinie egzaminatorów były bardzo niespójne (tabela 9.3). Dla żadnego wypracowania błąd kardynalny nie został stwierdzony przez wszystkich oceniających, zaś dla większości prac stwierdzony został przez nie więcej niż połowę oceniających je egzaminatorów. Na rzeczywistym egzaminie tylko jedna z tych prac zaklasyfikowana została jako zawierająca błąd kardynalny, trzy natomiast przekroczyły próg zaliczenia (21 punktów). Biorąc pod uwagę oceny przydzielone rozważanym pracom w badaniu, na przekroczenie progu zdawalności mogłoby liczyć (w zależności od przydziału do egzaminatora) w najlepszym wypadku 5 z nich (z czego trzy nie przekroczyły progu zdawalności w rzeczywistym egzaminie). Co ciekawe, odnotowano także jedną pracę, która przekroczyła próg zdawalności w rzeczywistym egzaminie, choć nie pozwoliłaby jej na to żadna z ocen przydzielonych w badaniu (choć 7 egzaminatorów oceniło ją w badaniu normalnie, nie stwierdzając błędu kardynalnego). Warto zwrócić uwagę na bardzo silną, ujemną korelację (-0,79) pomiędzy średnią oceną pracy (w wypadku, gdy nie stwierdzono błędu kardynalnego) a odsetkiem egzaminatorów, którzy zaklasyfikowali daną pracę jako zawierającą błąd kardynalny. Sugeruje to istnienie tzw. efektu halo, gdzie egzaminatorzy są mniej skłonni do stwierdzenia błędów kardynalnych w pracach ogólnie dobrych i bardziej skłonni w pracach ogólnie słabych.

Sprawę dodatkowo komplikują wyniki oceniania wybranych 121 prac przez ekspertów (tabela 9.4), wykryli oni bowiem w tej puli 15 prac zawierających błędy kardynalne. Stanowi to blisko 12% ogółu prac sprawdzanych przez ekspertów, podczas gdy odsetek prac zaklasyfikowanych przez przynajmniej jednego egzaminatora jako zawierających błąd kardynalny wyniósł w zasadniczym badaniu jedynie 1,1%. Biorąc pod uwagę 121 prac, to wszystkie wypracowania zaklasyfikowane w badaniu przez co najmniej jednego egzaminatora jako zawierające błędy kardynalne zostały ocenione jako zawierające błędy kardynalne także przez ekspertów. Jednocześnie eksperci zaklasyfikowali jako zawierające błędy kardynalne dodatkowe 12 prac, dla których błędu kardynalnego nie wskazał żaden z oceniających je egzaminatorów. Przy tym połowa z prac wskazanych przez ekspertów (a niewskazanych przez żadnego z egzaminatorów) przekroczyła w badaniu próg zdawalności egzaminu dla wszystkich 8 ocen dokonanych przez egzaminatorów, nawet tych najbardziej surowych. Przy wzięciu pod uwagę średniej z ocen przyznanych przez egzaminatorów, próg przekroczyło 5/6 z takich prac. Widać więc, że największe różnice między egzaminatorami a ekspertami dotyczą prac ogólnie nienajgorszych, dla których stwierdzenie błędu kardynalnego jest w praktyce decyzją o tym, czy uczeń zaliczy, czy nie zaliczy egzaminu maturalnego. Warto przy tym odnotować, że

współwystępowanie to odnotowano pomimo tego, że egzaminatorzy w badaniu nie znali punktacji pracy za zadania krótkiej odpowiedzi, a więc na ogół nie posiadali wiedzy o tym, czy stwierdzenie błędu kardynalnego decyduje o zdaniu egzaminu przez ucznia²³. Gdyby informacje takie posiadali, spodziewać by się można jeszcze rzadszego wskazywania błędów kardynalnych dla tego typu prac (por. rozdział 9.8.2).

Powyższe wyniki wskazują na to, że w dotychczasowej praktyce procesu klasyfikowania wypracowania jako zawierającego błąd kardynalny nie można postrzegać jako zjawiska deterministycznego, lecz jako proces losowy obarczony dużą niepewnością. Prowadzi to do nierównego traktowania uczniów, przy czym mechanizm działa tu odwrotnie do intuicji i polega na tym, że większość prac, które powinny zostać sklasyfikowane jako zawierające błąd kardynalny, ocenia się normalnie. Jest to krzywdzące dla tych uczniów, których prace zostaną mimo wszystko sklasyfikowane jako zawierające błąd kardynalny. Z badania wynika bowiem, że ta sama praca oceniona przez innego egzaminatora lub nawet innym razem przez tego samego egzaminatora zostałyby z wysokim prawdopodobieństwem oceniona normalnie. Szczególnego znaczenia nierówność ta nabiera dla tych prac, dla których stwierdzenie bądź nie błędu kardynalnego decyduje o tym, czy uczeń zaliczy egzamin maturalny. Błędy kardynalne (a dokładniej brak konsekwencji w ich stwierdzaniu) prowadzą również do obniżenia (już i bez tego niezbyt wysokiej) rzetelności oceny prac z języka polskiego.

Tabela 9.3. Zestawienie prac, dla których choć jeden z egzaminatorów stwierdził wystąpienie błędu kardynalnego

ID pracy	Suma punktów na egzaminie maturalnym (razem/ tylko zad. krótkiej odp.)	Suma punktów, gdy nie stwierdzono błędu kardynalnego (maksimum/średnia)	Liczba ocen, gdzie nie stwierdzono/ stwierdzono błąd kardynalny	Odsetek ocen, gdzie stwierdzono błąd kardynalny
11120237	30 / 11	23 / 15,3	7 / 1	12%
11120470	28 / 10	19 / 7,5	13 / 3	19%
11120304	9 / 4	19 / 11,0	15 / 5	25%
11110506	13 / 0	25 / 14,5	8 / 4	33%
11110491	21 / 0	14 / 7,7	7 / 5	42%
11120331	12 / 7	13 / 11,3	4 / 4	50%
11120420	2 / 2	19 / 12,0	4 / 4	50%
11120280	16 / 6	12 / 9,3	3 / 5	62%
11110209	13 / 0	12 / 6,7	3 / 13	81%
11120129	8 / 6	3 / 3,0	2 / 14	88%

*Wytuszczeniem w drugiej i trzeciej kolumnie zaznaczono sytuacje, w których praca przekroczyłaby próg zdawalności egzaminu.

Tabela 9.4. Charakterystyka prac sklasyfikowanych jako zawierające błędy kardynalne w puli 121 prac ocenianych przez ekspertów

ID pracy	Czy błąd kardynalny wg egzaminatorów (Tak/Nie)	Ocena na egzaminie maturalnym (razem/ tylko zad. krótkiej odp.)	Maksymalna ocena	Średnia ocen	Minimalna ocena
11110073	Nie	39 / 19	46	23,0	14
11110101	Nie	24 / 15	33	20,9	10
11110367	Nie	21 / 10	26	18,2	11
11110306	Nie	10 / 10	22	16,0	13

²³

Z wyjątkiem pojedynczych sytuacji, gdy rozważali przyznanie pracy co najmniej 21 punktów.

11110220	Nie	13 / 13	21	15,4	10
11110233	Nie	23 / 11	28	15,0	2
11110506	Tak	13 / 13	25	14,5	9
11110252	Nie	25 / 11	21	14,3	10
11110388	Nie	24 / 10	24	13,1	10
11110035	Nie	21 / 12	21	13,1	1
11110177	Nie	24 / 11	20	12,1	3
11110393	Nie	19 / 9	15	10,8	7
11110214	Nie	16 / 9	19	8,9	4
11110491	Tak	21 / 9	14	7,7	3
11110209	Tak	13 / 10	12	6,7	2

*Wytłuszczeniem w kolumnach od trzeciej do piątej zaznaczono prace, które w wypadku niestwierdzenia błędu kardynalnego przekroczyłyby prób zdawalności egzaminu.

9.2. Matematyka

Obserwowany w badaniu wpływ efektu egzaminatora na końcową ocenę dla zadań z matematyki jest mniejszy niż dla zadań z polskiego. Wynika to zarówno z mniejszego błędu systematycznego (mniejszego zróżnicowania łagodności egzaminatorów), jak również z większej rzetelności oceniania.

Wspólną cechą wszystkich zadań z matematyki jest stosunkowo niewielkie zróżnicowanie łagodności egzaminatorów – rozstęp międzykwartkowy parametru b_{kr} dla 25% zadań jest mniejszy od 0,2 punktu procentowego na egzaminie, dla 50% zadań jest mniejszy od 0,3 punktu procentowego na egzaminie, dla 75% zadań jest mniejszy od 0,4 punktu procentowego na egzaminie, a dla żadnego zadania nie przekracza 1 punktu procentowego na egzaminie. Przy tym w wypadku zadań z części podstawowej egzaminu zróżnicowanie łagodności egzaminatorów jest co do zasady nieco niższe niż w wypadku części rozszerzonej. Są to wartości od 3 do ponad 10 razy mniejsze niż dla kryteriów *kompozycji, stylu, języka* czy *zapisu* w języku polskim. Z uwagi na większą liczbę zadań w arkuszu z matematyki warto przyjrzeć się, czy różnice łagodności egzaminatorów akumulują się, czy znoszą pomiędzy wszystkimi zadaniami danej części egzaminu. Wyniki badania wskazują na to, że egzaminatorzy z matematyki są konsekwentnie surowi bądź łagodni tylko w ograniczonym stopniu – rozstęp międzykwartkowy łagodności zagregowanej do poziomu testu (b_p) waha się w zależności od roku i poziomu egzaminu od 0,8 do 1,2 punktu procentowego na egzaminie, podczas gdy średnia wartość rozstępu międzykwartkowego łagodności na poziomie zadań wynosiła 0,32 punktu procentowego na egzaminie, a w arkuszu egzaminacyjnym znajduje się od 9 do 12 zadań (a więc przy egzaminatorach tak samo łagodnych bez względu na zadanie spodziewalibyśmy się wartości na poziomie testu równej ok. 2,88 – 3,84 punktu procentowego na egzaminie). Różnice na poziomie testu pomiędzy skrajnie surowym i łagodnym egzaminatorem wyniosły w zależności od poziomu i roku egzaminu od 4,32 do 5,02 punktu procentowego na egzaminie. Dla porównania, w języku polskim uśredniony pomiędzy tematami wypracowania i latami rozstęp międzykwartkowy łagodności egzaminatorów na poziomie testu wyniósł w badaniu 3,31 punktu procentowego na egzaminie, a różnica pomiędzy skrajnie surowym i łagodnym egzaminatorem aż 15,2 punktu procentowego na egzaminie, a więc są to wartości 3 do 4 razy większe. Choć wielokrotnie mniejsze niż w języku polskim różnice w łagodności egzaminatorów z matematyki trudno jeszcze uznać za pomijalne. Różnica ponad czterech punktów procentowych na egzaminie pomiędzy skrajnymi egzaminatorami odpowiada w wypadku matury z matematyki przesunięciu ucznia w rozkładzie populacyjnym o co najmniej 4 centyle, biorąc pod uwagę wynik za cały egzamin²⁴.

²⁴ Dla poziomu podstawowego w roku 2011 od 16% do 60% punktów, dla poziomu podstawowego w roku 2012 od 16% do 96% punktów, dla poziomu rozszerzonego w 2011 roku od 4% do 72% punktów, dla poziomu rozszerzonego w 2012 roku od 18% do 74% punktów. We wszystkich wymienionych przypadkach jest to punktacja za cały egzamin.

Drugim czynnikiem zmniejszającym obserwowany dla matematyki efekt egzaminatora jest większa rzetelność oceniania. Wpływ na to mają zarówno wyższe wartości standaryzowanego współczynnika dyskryminacji poszczególnych zadań (dla matematyki kwartyle rozkładu mediany d_s wynoszą 12, 16,9, 18,3, podczas gdy dla j. polskiego 10,4, 12,4 i 14,3), jak i znacznie krótsze skale stosowane w zadaniach z matematyki. Najdłuższe skale stosowane na egzaminie z matematyki to 6 poziomów wykonania (od 0 do 5 punktów) na poziomie podstawowym i 7 poziomów wykonania (od 0 do 6 punktów) na poziomie rozszerzonym przy średniej długości skali wynoszącej 3,7 poziomów wykonania dla części podstawowej oraz 5,3 poziomów wykonania dla części rozszerzonej. Dla języka polskiego najdłuższa skala to 26 poziomów wykonania, a średnia długość skali wynosi 8,2 poziomu wykonania. O istotnym związku długości skali oceny zadania i wielkości efektu egzaminatora świadczy korelacja pomiędzy długością skali a średnim odsetkiem poprawnych klasyfikacji danego zadania, która wyniosła w wypadku matematyki -0,44.

Opisane powyżej czynniki przekładają się na dość wysokie odsetki poprawnych klasyfikacji zadań z matematyki – średni odsetek poprawnych klasyfikacji waha się w zależności od zadania od 0,59 do 0,89 z medianą 0,75. Dla porównania, dla języka polskiego wartości te wynoszą od 0,24 do 0,94 z medianą 0,58. W badaniu nie odnotowano związku pomiędzy zakresem treściowym zadań a odsetkiem poprawnych klasyfikacji (tabela 9.7). Niższe odsetki klasyfikacji odnotowane w zadaniach z poziomu rozszerzonego wydają się związane głównie z dłuższymi skalami ocen zadań spotykanych na poziomie rozszerzonym. Na koniec warto wspomnieć, że wyniki analiz wskazują kilka zadań z matematyki (zadania 24 i 26 z roku 2011 oraz 26 z roku 2012, obydwa z poziomu podstawowego), dla których odsetek poprawnych klasyfikacji był niewspółmiernie niski w porównaniu do pozostałych pytań, a jednocześnie były one oceniane na skali o jedynie trzech poziomach wykonania. Wskazywałoby to na niedostateczną szczegółowość lub niejednoznaczność klucza oceny tych zadań.

Tabela 9.5. Parametry rozkładu (1 kwartyl, mediana, 3 kwartyl) współczynnika dyskryminacji egzaminatora (d) oraz jego standaryzowanej wartości d_s dla poszczególnych zadań z matematyki (tylko egzaminatorzy)

Zad.		Podstawowa 2011			Podstawowa 2012			Rozszerzona 2011			Rozszerzona 2012		
		Q1	med.	Q3	Q1	med.	Q3	Q1	med.	Q3	Q1	med.	Q3
1.	d	5,4	7,2	8,1	5,1	5,8	7,6	5,2	7,2	9,1	5,4	7,4	8,1
	d_s	8,1	10,8	12,1	7,6	8,7	11,4	13,1	17,9	22,6	13,6	18,5	20,2
2.	d	7,1	8	8,9	5,7	7,4	8,8	3,5	5,3	7,3	4,8	6,7	7,9
	d_s	10,6	12	13,4	8,6	11,1	13,2	8,7	13,2	18,4	12	16,7	19,8
3.	d	4,8	5,3	6,3	5,4	7,3	8,1	5,1	7,4	8,8	5,3	7,1	8,5
	d_s	7,1	7,9	9,4	8,1	10,9	12,1	17,9	25,8	30,7	13,1	17,6	21,3
4.	d	5	5,6	7,5	6,7	8	8,8	5,5	7,9	9,2	4,8	6,8	8,2
	d_s	7,6	8,4	11,3	10,1	12	13,2	13,9	19,7	23	16,8	23,8	28,6
5.	d	5,4	7,5	8,4	6,4	7,4	8	5,6	7,4	8,2	4,1	6,2	8
	d_s	8,1	11,3	12,6	9,7	11	12	14	18,4	20,4	14,5	21,5	27,9
6.	d	6,7	7,4	8,2	7,5	9,1	10,4	4,9	7,1	8,9	1,7	5	7,5
	d_s	10	11,1	12,4	11,3	13,7	15,5	12,3	17,7	22,2	6	17,5	26,2
7.	d	8,2	9,2	10,5	5,3	7,1	7,7	3,1	6,4	8,2	4,1	6,9	8,5
	d_s	12,4	13,8	15,7	13,3	17,9	19,4	7,8	16	20,4	8,3	13,9	17,1
8.	d	6,3	7,9	8,8	5,2	7	8,4	4,6	6,8	8,5	5,3	7,1	8,5
	d_s	15,8	19,7	22	12,9	17,6	21	11,6	17	21,4	13,2	17,7	21,3
9.	d	4,8	6,8	8	5,7	7,6	8,4	5,4	7,1	8,7	2,7	5,6	7,4
	d_s	14,3	20,3	24	17,2	22,8	25,2	13,5	17,7	21,6	8,2	16,7	22,1

Efekt egzaminatora w ocenianiu prac maturalnych z języka polskiego i matematyki

10.	d	5,6	7,4	8,6	6,4	7,5	8,1	5,6	7,4	8,5
	d _s	14	18,5	21,5	12,8	15	16,1	16,8	22,2	25,4
11.	d				2,1	5,1	7,5	2,8	6,2	7,4
	d _s				7,4	17,9	26,3	5,6	12,4	14,9
12.	d				2,8	5,1	7			
	d _s				5,5	10,3	14,1			

*W kolumnie „Zadania” umieszczono kolejność zadań otwartych w arkuszu.

Analiza wskaźników zniekształcenia skali (r_{kr}) wykazała, że jedynie w przypadku 5 zadań mediana dla wszystkich egzaminatorów jest większa od 1. Wg kolejności zadań otwartych w arkuszach są to zadania: 1 i 3 z poziomu podstawowego z 2011 roku, 4 i 5 z poziomu rozszerzonego z 2011 roku i 5 z poziomu rozszerzonego z 2012 roku, przy czym dla obydwu zadań 5 wartość ta jest wyższa od 1 o zaledwie 0,01–0,02. Oznacza to, że tylko w przypadku trzech pozostałych zadań egzaminatorzy średnio wykazują tendencję do ograniczania skali. W piętnastu zadaniach współczynnik r_{kr} mieści się w granicach 0,9–1, można zatem uznać, że przeciętnie nie następuje w nich zniekształcenie skali przez egzaminatorów (podobnie jak we wspomnianych zadaniach nr 5 z poziomów rozszerzonych). W pozostałych 22 zadaniach wartości współczynnika r_{kr} wahają się pomiędzy 0,52 a 0,89 (ze średnią równą 0,78), można zatem wnioskować, że zachodzi w nich z różną siłą zjawisko przydzielenia ocen skrajnych. Tabela 9.6 zawiera medianę oraz pierwszy i trzeci kwartył parametrów r_{kr} dla poszczególnych zadań, a wykresy ich rozkładów można znaleźć w aneksie 2.

Tabela 9.6. Parametry rozkładów współczynnika r_{kr} dla poszczególnych zadań z matematyki (tylko egzaminatorzy). Pogrubieniem i gwiazdką zaznaczono wartości mediany większe od 1

Zadania	Podstawowa 2011			Podstawowa 2012			Rozszerzona 2011			Rozszerzona 2012		
	Q1	med.	Q3	Q1	med.	Q3	Q1	med.	Q3	Q1	med.	Q3
1.	0,95	1,23*	1,65	0,81	0,88	0,95	0,62	0,66	0,76	0,76	0,84	0,94
2.	0,60	0,86	1,12	0,82	0,95	1,08	0,86	0,91	1,00	0,78	0,86	0,99
3.	0,99	1,19*	1,53	0,90	0,95	1,02	0,65	0,90	1,24	0,86	0,95	1,07
4.	0,65	0,92	1,13	0,88	0,97	1,07	1,06	1,36*	1,69	0,86	0,91	0,99
5.	0,52	0,57	0,88	0,73	0,78	0,87	0,77	1,01*	1,35	0,94	1,02*	1,09
6.	0,53	0,74	0,89	0,88	0,97	1,08	0,52	0,58	0,84	0,85	0,91	0,98
7.	0,87	0,99	1,13	0,74	0,84	0,92	0,52	0,62	0,9	0,62	0,77	0,9
8.	0,79	0,88	0,98	0,80	0,87	0,98	0,46	0,53	0,69	0,80	0,94	1,05
9.	0,82	0,87	0,98	0,75	0,89	1,01	0,89	1,00	1,16	0,80	0,88	0,98
10.	0,79	0,86	0,92				0,77	0,90	0,98	0,84	0,90	0,98
11.							0,78	0,86	0,96	0,51	0,67	0,84
12.							0,81	0,84	0,90			

*W kolumnie „Zadania” umieszczono kolejność zadań otwartych w arkuszu.

Tabela 9.7. Zróżnicowanie treściowe zadań otwartych w egzaminach maturalnych z matematyki wraz z oznaczeniem przedziału średniej wartości odsetka poprawnych odpowiedzi pomiędzy egzaminatorami

	2011	2012	2011	2012
	podstawa	podstawa	rozszerzenie	rozszerzenie
Równania i nierówności	rów. kwadratowe (24) [3]	rów. kwadratowe (26) [3]		rów. kwadratowe (1) [5]
	dowodzenie (25) [3]	dowodzenie (27) [3]	dowodzenie (1) [5]	dowodzenie (7) [4]
			dowodzenie (2) [5]	
	zad. z treścią (32) [6]	zad. z treścią (34) [6]		zad. z treścią (8) [5]
Geometria	Geometria (29) [3]	Geometria (30) [3]	Geometria (6) [5]	Geometria (9) [6]
			Geometria (10) [4]	
	Geometria analityczna (31) [5]	Geometria analityczna (29) [3]	Geometria analityczna (7) [5]	Geometria analityczna (6) [7]
	Geometria przestrzenna (33) [5]	Geometria przestrzenna (33) [5]	Geometria przestrzenna (8) [5]	Geometria przestrzenna (10) [6]
			Geometria przestrzenna (11) [7]	
Pozostałe	Ciągi (27) [3]	Ciągi (32) [5]	Ciągi (5) [5]	Ciągi (5) [7]
	Kombinatoryka (30) [3]	Kombinatoryka (31) [3]	Kombinatoryka (9) [5]	Kombinatoryka (11) [4]
			Kombinatoryka (12) [4]	
		Wielomiany (28) [3]	Wielomiany (3) [7]	Wielomiany (2) [5]
				Wielomiany (4) [7]
	Trygonometria (28) [3]		Trygonometria (4) [5]	Trygonometria (3) [5]
	Funkcje (26) [3]			

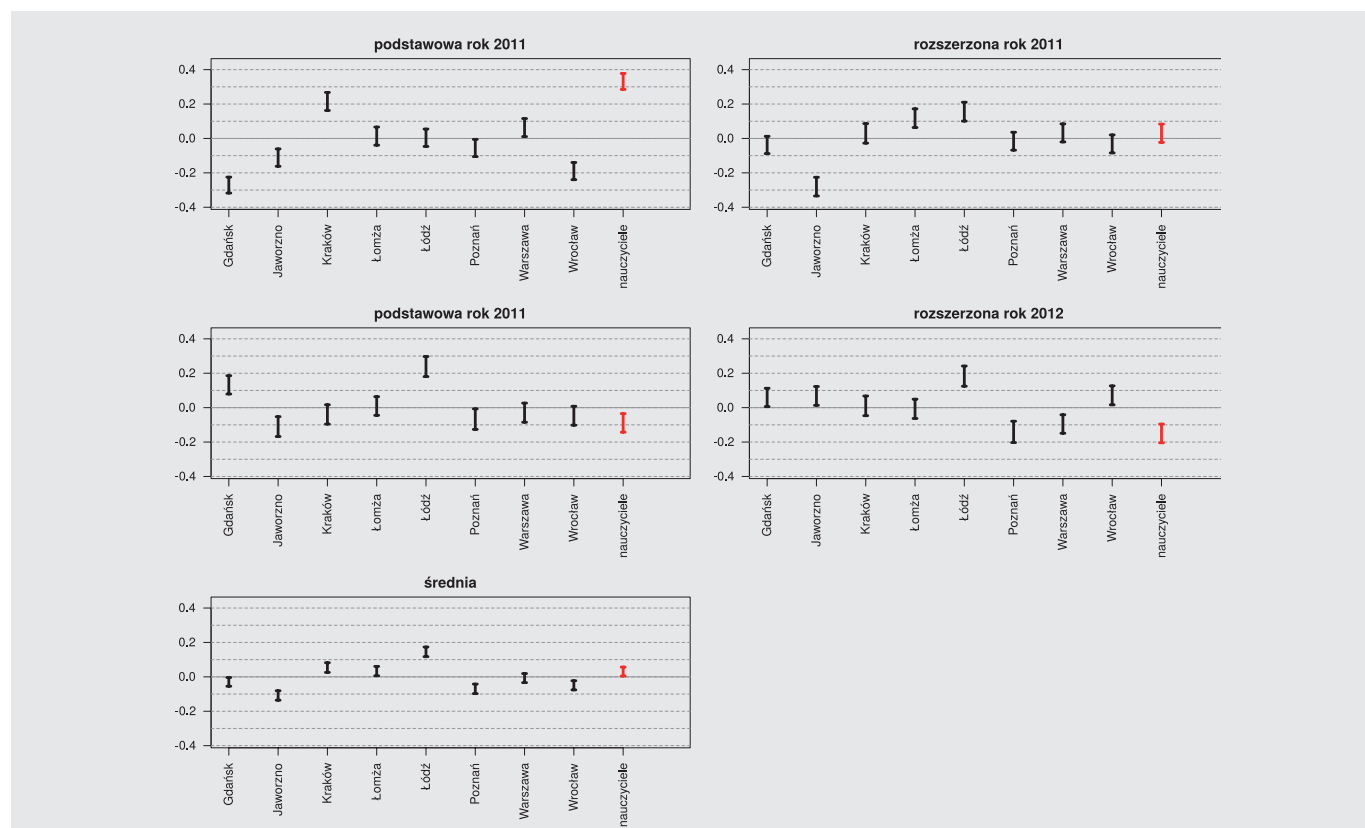
*(ρ_{kr} – odsetek poprawnych klasyfikacji egzaminatora): powyżej 0,8 – kolor zielony, pomiędzy 0,7 a 0,8 (włącznie) – kolor biały, poniżej 0,7 – kolor różowy. W nawiasach podano numery zadań w arkuszu egzaminacyjnym oraz liczbę poziomów wykonania zadania.

9.2.1. Nauczyciele niebędący egzaminatorami

Podobnie jak w wypadku języka polskiego różnice pomiędzy egzaminatorami a osobami nieposiadającymi statusu egzaminatora okazały się dla matematyki niewielkie. Odsetek poprawnych klasyfikacji był istotnie statystycznie różny²⁵ jedynie dla 3 spośród 42 zadań. Dla tych pytań średni odsetek poprawnych klasyfikacji był dla nieegzaminatorów wyższy o od 0,02 do 0,05. Średnia łagodność okazała się istotnie statystycznie różna dla 16 spośród 42 zadań (7 z poziomu podstawowego i 3 z rozszerzonego w 2011 r., 1 z poziomu podstawowego i 5 z rozszerzonego w 2012 r.). We wszystkich przypadkach, gdy różnica była istotna statystycznie, nieegzaminatorzy byli bardziej łagodni – różnice wyniosły od 0,02 do 0,24 punktu procentowego na egzaminie. Na poziomie całego badania różnica nie była istotna statystycznie, zaś dla poszczególnych arkuszy egzaminacyjnych wynik był zależny od roku i poziomu egzaminu – w dwóch wypadkach nieegzaminatorzy byli istotnie statystycznie bardziej surowi, w jednym łagodni, a w jednym różnica nie była istotna statystycznie (rysunek 9.5). Biorąc pod uwagę brak istotnych statystycznych różnic na poziomie całego badania oraz niekonsekwentne wyniki na poziomie poszczególnych arkuszy, uprawnione wydaje się stwierdzenie, że nie występują istotne różnice w sile efektu egzaminatora pomiędzy osobami posiadającymi i nieposiadającymi statusu egzaminatora.

²⁵ Test t Welcha na poziomie istotności 0,01.

Rysunek 9.5. 99% przedziały ufności łagodności egzaminatorów na poziomie roku, tematu i OKE (wraz z grupą kontrolną nauczycieli oczekujących na nadanie statusu egzaminatora) dla matematyki (w punktach procentowych na egzaminie)



*Jako 0 przyjęto średnią łagodność spośród wszystkich egzaminatorów.

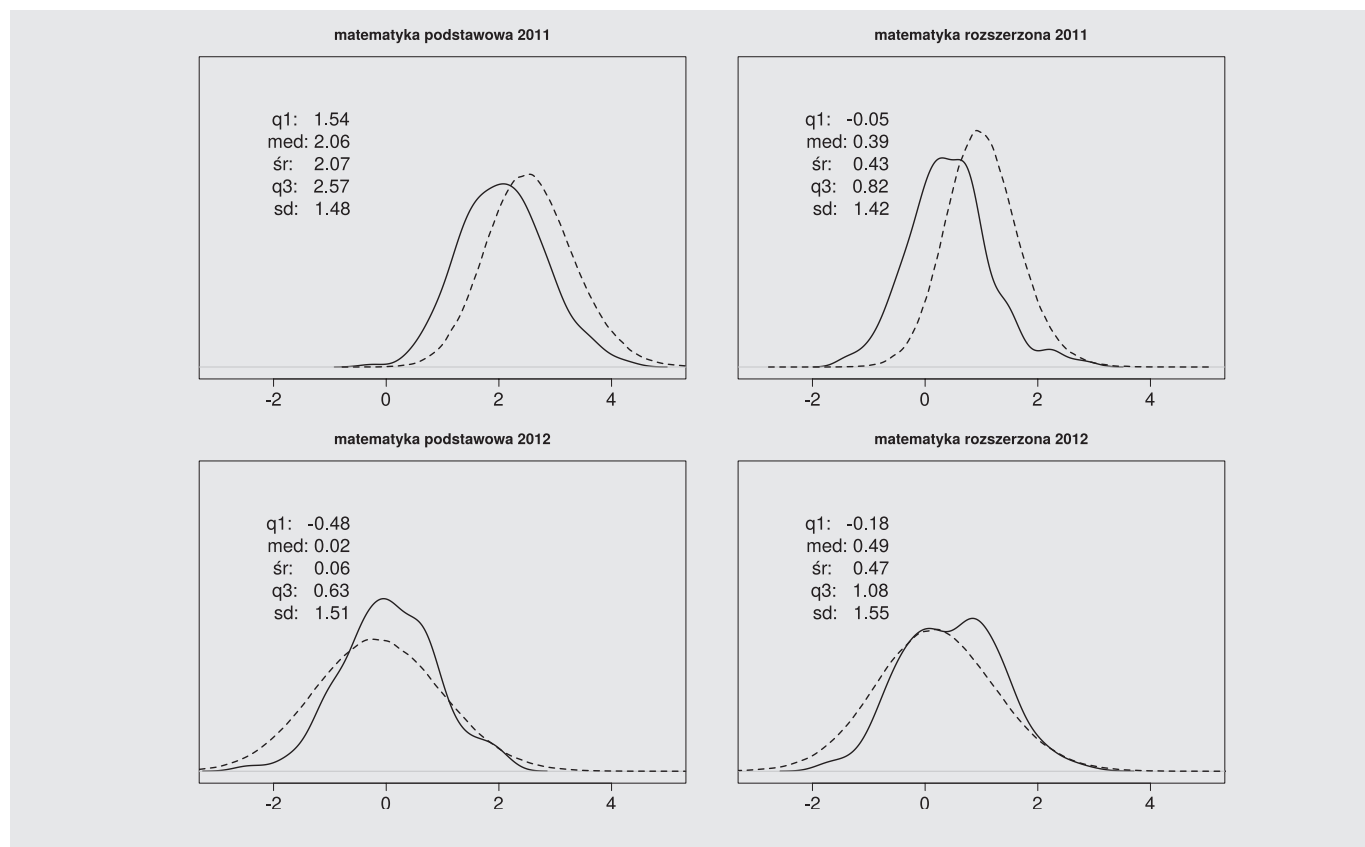
9.2.2. Różnice łagodności na poziomie testów

Skumulowany do poziomu całego testu wpływ różnic łagodności egzaminatorów w poszczególnych zadaniach przedstawia rysunek 9.6. Uzyskane wyniki są zbliżone dla obydwu lat i poziomów egzaminu (należy pamiętać, że łagodność egzaminatorów jest miarą względną, a więc przesunięcia całych rozkładów względem punktu 0 nie mają znaczenia), choć w wypadku poziomu rozszerzonego matury z 2011 roku zróżnicowanie jest nieco niższe niż w pozostałych wypadkach. Różnica łagodności pomiędzy 25% najbardziej łagodnych a 25% najbardziej surowych egzaminatorów wynosi od co najmniej 0,87 punktu procentowego wyniku egzaminu (poziom rozszerzony w roku 2011) do co najmniej 1,36 punktu procentowego wyniku egzaminu (poziom rozszerzony w roku 2012). Są to różnice od dwa do trzech razy niższe niż w wypadku egzaminu z języka polskiego, jednak cały czas mogą one powodować przesunięcie ucznia w rozkładzie populacyjnym o ponad 1 centyl. Różnice pomiędzy skrajnie surowym i łagodnym egzaminatorem wyniosły, w zależności od roku i poziomu egzaminu, od 4,4 do 4,7 punktu procentowego wyniku egzaminu, co odpowiada przesunięciu ucznia o przeciętny wynik w rozkładzie populacyjnym o ok. 5 centyli. Z punktu widzenia ucznia może to być znacząca różnica, pamiętać należy jednak o tym, że jest to przypadek skrajny. Różnice pomiędzy większością egzaminatorów są znacznie mniejsze. Jest to również czterokrotnie mniejsza wartość niż w wypadku egzaminu z języka polskiego.

Ciekawych wyników dostarcza porównanie uzyskanych rozkładów łagodności egzaminatorów na poziomie testu z wartościami hipotetycznymi przy założeniu, że łagodność egzaminatora w jednym zadaniu nie jest związana z łagodnością w pozostałych (rysunek 9.6 – linia przerywana). Okazuje się, że w 2011 roku, zarówno dla poziomu podstawowego, jak i rozszerzonego, liczba egzaminatorów systematycznie surowych wyraźnie przewyższała liczbę egzaminatorów systematycznie łagodnych (stąd przesunięcie rozkładów obserwowanych w lewo). Efektu tego nie zaobserwowano w roku 2012, a wręcz przeciwnie – w tym roczniku liczba egzaminatorów systematycznie łagodnych nieznacznie dominowała nad liczbą systematycznie surowych (nieznaczne przesunięcie rozkładów obserwowanych w prawo). Brak stałych w czasie wzorców w rozważanych wynikach sugeruje, że egzaminatorzy nie przejawiają silnych tendencji do bycia systematycznie surowymi lub łagodnymi we wszystkich ocenianych zadaniach. Jeśli związek taki istnieje, ma on

bardziej złożony charakter, związany np. z zakresem tematycznym zadań lub sposobem konstruowania kluczy oceny, co wymagałoby jednak przeprowadzenia pogłębionych badań.

Rysunek 9.6. Rozkłady parametrów b_{tr} (łagodność egzaminatorów wyrażona w punktach procentowych wyniku egzaminacyjnego zagregowana do poziomu testu) w podziale na rok i poziom matury z matematyki: rzeczywiste (linia ciągła) i w sytuacji braku związku między łagodnością egzaminatorów w poszczególnych kryteriach oceny²⁶ (linia przerywana)



*q1, q3 – pierwszy i trzeci kwartył, sd – odchylenie standardowe

9.2.3. Różnice na poziomie okręgowych komisji egzaminacyjnych

Zarówno na poziomie poszczególnych arkuszy egzaminacyjnych, jak i na poziomie całego badania w przypadku egzaminu maturalnego z matematyki odnotowano istotne statystycznie różnice w łagodności oceniania pomiędzy okręgowymi komisjami egzaminacyjnymi. 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 (rysunek 9.5). 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. Brak systematycznych różnic łagodności pomiędzy OKE jest cechą pożądaną, niemniej ważne są jednak wielkości występujących tendencji oraz różnic dla oceniania zadań z poszczególnych arkuszy egzaminacyjnych.

W wypadku matematyki siła obserwowanych tendencji waha się od -0,16 (OKE Jaworzno) do 0,20 punktu procentowego na egzaminie (OKE Łódź) z odchyleniem standardowym wynoszącym 0,1 punktu procentowego na egzaminie. Rozstęp skrajnych wartości wynoszący 0,36 punktu procentowego na egzaminie jest ponad sześciokrotnie mniejszy niż w wypadku języka polskiego. Dla porównania, rozstęp międzykwartkowy zróżnicowania łagodności egzaminatorów w ramach OKE wynosi od 0,58 do 1,5 punktu procentowego na egzaminie, a różnice pomiędzy skrajnymi egzaminatorami od 2,38 do 4,44 punktu procentowego, w zależności od

²⁶ Rozkłady uzyskano poprzez tysiąckrotne losowe permutowanie faktycznie uzyskanych w badaniu wartości łagodności egzaminatorów i sumowanie ich do poziomu testu, a następnie policzenie funkcji gęstości w wszystkich uzyskanych sum.

OKE, roku i poziomu egzaminu. Stąd, jeśli uwzględnić, że ocena każdej pracy obciążona jest jeszcze błędem losowym wynikającym z niedoskonałej rzetelności oceniania, wydaje się, że wpływ tendencji łagodności pomiędzy OKE na całkowity błąd egzaminatora, jakim obciążona jest ocena pojedynczej pracy, można uznać za pomijalny. Także bezwzględne wartości tych różnic (0,36 punktu procentowego na egzaminie między skrajnymi OKE, odchylenie standardowe łagodności OKE równe 0,1 punktu procentowego na egzaminie) wydają się pomijalne z punktu widzenia pojedynczego ucznia. Podsumowując, wpływ różnic w tendencji łagodności pomiędzy OKE na ocenę pojedynczej pracy, choć istotny statystycznie, z praktycznego punktu widzenia uznać można za zanedbywalnie mały.

Biorąc pod uwagę, że zróżnicowanie OKE ze względu na łagodność różni się w zależności od roku i poziomu egzaminu, warto przyjrzeć mu się w rozbiciu na poszczególne arkusze egzaminacyjne. Analiza na tym poziomie wykazuje większe różnice – rozstęp łagodności pomiędzy najbardziej łagodną i surową OKE waha się w zależności od roku i poziomu egzaminu od 0,46 do 0,84 punktu procentowego na egzaminie, a odchylenie standardowe od 0,16 do 0,26 punktu procentowego na egzaminie. 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 i 0,03% dla poziomu rozszerzonego egzaminu (dla porównania w języku polskim było to, zależnie od tematu wypracowania, 4,5% i 44%). Wyniki te utwierdzają w przekonaniu, że wpływ systematycznych różnic pomiędzy OKE na wyniki egzaminu z matematyki można uznać za mało znaczący z praktycznego punktu widzenia.

9.3. Pismo odręczne vs zapis komputerowy a ocena za wypracowanie

Jak już wspomniano w rozdziale 5, na wynik oceny wypracowania, oprócz jego jakości odzwierciedlającej poziom badanej umiejętności egzaminowanego, ma wpływ wiele czynników ogólnie nazywanych efektem egzaminatora (Scullen i in., 2000). Czynniki te mogą mieć różne źródło. Na ocenę jaką przyznaje egzaminator może mieć wpływ czytelność pisma (Wood, 1991), czy też forma zapisu – pismo odręczne lub zapis komputerowy (Mogey, Paterson, Burk i Purcell, 2010). Przeprowadzono stosunkowo niewiele badań dotyczących zróżnicowania ocen przyznawanych przez egzaminatorów za te same prace zapisane odręcznie i na komputerze. Jeżeli pojawiają się badania w tym zakresie, to stanowią one jeden z wątków podczas studiów dotyczących zróżnicowania wyników standaryzowanych testów w zakresie umiejętności pisania własnego tekstu z wykorzystaniem przez zdających komputerowej klawiatury i edytora tekstu oraz tekstów pisanych odręcznie. W takich przypadkach przedmiotem dociekań jest nie tylko problem efektu egzaminatora, ale przede wszystkim różnice w zakresie procesów kognitywnych podczas używania przez zdających klawiatury i funkcjonalności edytora tekstu w porównaniu z arkuszem papieru i długopisem (Lottridge, Nicewander, Schulz i Mitzel, 2008). Pisanie wypracowania podczas egzaminu odręcznie i na komputerze niekoniecznie dotyczy tego samego konstruktu i pomimo że zadania w swoim zapisie nie różnią się, to sprawdzane umiejętności niekoniecznie się pokrywają. Pojawiają się więc dwa problemy. Po pierwsze, problem ekwiwalentności wyników oceniania tych samych prac zapisanych odręcznie i na komputerze oraz po drugie, problem ekwiwalentności konstruktu – umiejętności sprawdzanych za pomocą tak samo sformułowanego zadania i tych samych kryteriów oceniania w przypadku, gdy zdający używają różnych narzędzi do tworzenia tekstu – długopisu i kartki papieru lub klawiatury i edytora tekstu.

Problem ekwiwalentności konstruktu w ocenianiu prac pisanych odręcznie i na komputerze nie był przedmiotem prezentowanych w tym rozdziale badań, gdyż eksperyment dotyczył tylko ekwiwalentności ocen tych samych wypracowań ocenianych w dwóch formach ich zapisu – pisma odręcznego i kopii sporządzonej z wykorzystaniem komputera i edytora tekstu. Celem przeprowadzonego eksperymentu było zbadanie, czy forma zapisu wypracowania maturalnego z języka polskiego na poziomie podstawowym (pismo odręczne – PO vs zapis komputerowy – PK) generuje efekt egzaminatora. Rozważano, czy forma zapisu (PK lub PO) różniuje wynik oceniania, a dokładniej poszukiwano odpowiedzi na dwa pytania:

1. Czy ocena przyznawana przez egzaminatorów za wypracowanie jest obciążona efektem formy zapisu (pismo odręczne vs zapis komputerowy)?
2. Czy któryś z kryteriów oceniania jest szczególnie narażony na obciążenie efektem egzaminatora ze względu na formę zapisu?

9.3.1. Dobór prac do badań

Do eksperymentu polegającego na analizie porównawczej wyników oceniania prac zapisanych oryginalnie pismem odręcznym i ich kopii przepisanych na komputerze wybrano 120 prac z próby 841 wypracowań włączonych do podstawowego schematu rozdziału prac maturalnych z języka polskiego pomiędzy OKE (por. rozdział 6.7). Wybrane prace zostały przepisane na komputerze z zachowaniem wszystkich poprawek, skreśleń i zakreśleń dokonanych przez zdającego. Wybór prac został przeprowadzony przez grafologa.

Prace wybrane do badań ($n = 120$) oceniane były przez egzaminatorów pochodzących z losowej próby warstwowanej ze względu na OKE w trakcie sesji oceniającej w ramach badań porównywalności oceniania i efektu egzaminatora, która odbyła się w październiku 2013 roku. Organizacja sesji oceniającej ściśle odpowiadała warunkom zapewnionym podczas regularnej sesji podczas egzaminów. Prace były oceniane w ośmiu zespołach zlokalizowanych w miastach będących siedzibami poszczególnych OKE. Każdy zespół składał się z 29 egzaminatorów pod kierownictwem przewodniczącego. Zarówno egzaminatorzy, jak i przewodniczący pochodzili z terenu OKE, gdzie ulokowano siedziby zespołów. Każda z prac w zapisie oryginalnym oceniana była ośmiokrotnie (przez jednego egzaminatora z próby z każdej OKE), co wynikało z celów badań porównywalności oceniania i efektu egzaminatora prowadzonych przez IBE (Kulon i Żółtak, 2014). Każda kopia wypracowania przepisane na komputerze oceniana była przez czterech egzaminatorów (każdy egzaminator z innej OKE).

9.3.2. Wyniki badań – efekt formy zapisu wypracowania (pismo odręczne vs zapis komputerowy)

W trakcie egzaminu maturalnego z języka polskiego oceniane prace w zapisie komputerowym stanowią tylko 2 promile wszystkich prac ocenianych podczas sesji. Jest to jednak liczba kilkuset prac w każdej sesji. Dlatego też wiedza, czy wynik tych prac jest obciążony efektem egzaminatora ze względu na formę zapisu, może mieć znaczenie dla procesu szkolenia egzaminatorów oceniających wypracowania w zapisie komputerowym.

W związku z tym, że każda praca była oceniana przez ośmiu (PO) lub czterech (PK) egzaminatorów, pierwszym krokiem do przeprowadzenia analiz jest ustalenie „ostatecznej” oceny danej pracy. Można to zrobić na kilka sposobów.

Jednym z podejść jest skorzystanie z surowych wyników, a więc ocen przydzielonych przez poszczególnych oceniających. Naturalnym sposobem obliczenia uzgodnionej oceny dla pracy jest obliczenie średniej ze wszystkich ocen. W związku z tym, że mamy do czynienia z dyskretną, a nie ciągłą skalą ocen, warto zastanowić się, czy inne rozwiązania nie byłyby lepsze. Jednym z nich może być użycie dominanty zamiast średniej. Średnia jest podatna na wartości skrajne, zatem jeśli daną pracę jakiś egzaminator ocenił znacząco niżej lub wyżej niż pozostali, będzie to miało wpływ na ostateczną ocenę. W przypadku użycia dominanty za ostateczną ocenę wypracowania uznalibyśmy ocenę przyznaną przez największą liczbę zgodnych egzaminatorów. O ile w przypadku prac pisanych ręcznie (PO) i ośmiu ocen mamy dość duże prawdopodobieństwo, że kilku (przynajmniej dwóch) egzaminatorów przyzna taką samą ocenę, to w przypadku prac przepisanych (PK) i czterech ocen może się zdarzyć, że nie będzie wartości dominującej. W takim przypadku należałoby obliczyć średnią z najczęściej występujących wartości, co w skrajnych przypadkach wszystkich ocen różniących się od siebie odpowiadałoby po prostu obliczeniu średniej. Wadą obliczania dominanty jest to, że przy dużych rozbieżnościach ocen pomiędzy egzaminatorami wystarczy, że dwóch z nich osiągnie (nawet przypadkową) zgodność i ich ocena zostanie uznana za ostateczną. Istotnym problemem przy obliczaniu ostatecznej oceny było to, że niektóre kryteria oceniania posiadają liczbę kategorii punktowych różną od liczby punktów, które można przyznać. Przykładowo, *kompozycja* czy *styl* są punktowane według schematu na 0, 1, 3 lub 5 punktów. Powoduje to, iż dla części kryteriów musi wystąpić między oceniającymi zgodność przynajmniej co do jednej kategorii. W pierwszej kolejności warto w związku z tym posłużyć się sumaryczną oceną za wypracowanie, gdzie nie powinna występować zgodność wymuszona przez małą liczbę kategorii.

Drugim, podejściem jest skorzystanie z ocen oszacowanych z wykorzystaniem *Hierarchical Rater Model with Signal Detection Theory* (HRM-SDT). Model ten został użyty na potrzeby badań porównywalności oceniania i efektu egzaminatora i szczegółowo opisany jest w rozdziale 7.

Model ten szacowany był bayesowsko, z użyciem metody *Markov chain Monte Carlo* (MCMC). W wyniku użycia tej metody uzyskuje się ciągi iteracji (zwane łańcuchami Markowa) dla każdego z szacowanych parametrów. Zwyczajowo wartość parametru oblicza się, wyciągając średnią z łańcucha. Użycie „prawdziwych” ocen z tego modelu ma tę zaletę, że powinny one być bliższe rzeczywistości

poziomowi umiejętności uczniów niż prosta średnia czy dominanta z surowych ocen. Zatem trzecią możliwością obliczenia ostatecznej oceny danej pracy jest obliczenie (zwyczajowej) średniej z łańcucha Markowa wspomnianego modelu. Podobnie jak w przypadku używania ocen surowych, z racji dyskretnej skali ocen, jako alternatywne rozwiązanie można użyć dominanty z łańcucha.

Mamy zatem cztery, nieco różniące się, sposoby obliczania ostatecznej oceny danej pracy (średnia wyników surowych, dominanta wyników surowych, średnia wyników „prawdziwych”, dominanta wyników „prawdziwych”). Warto sprawdzić, jak wyglądają rozkłady ocen w zależności od użytego sposobu. Rysunek 9.7 przedstawia porównanie rozkładów wyników ucznia szacowanych tymi czterema metodami bez podziału na prace pisane ręcznie i przepisane. Widać wyraźnie różnicę pomiędzy rozkładami reprezentującymi poszczególne podejścia – w przypadku średnich i dominant wyników surowych uzyskano nieco więcej wyników w środku skali w porównaniu z wynikami „prawdziwymi” z modelu HRM-SDT. Użycie dominanty wyników surowych prowadzi do uzyskania wartości nieco wyższych niż w przypadku użycia średniej – można zauważyć rozbieżność pomiędzy tymi rozkładami poniżej 20 punktów. Rozkłady ocen obliczonych na podstawie „prawdziwych” wyników z modelu HRM-SDT nie różnią się zbyt. Tabela 9.8 zawiera podsumowanie statystyk omawianych rozkładów.

Rysunek 9.7. Porównanie rozkładów ocen ostatecznych (oszacowanych na podstawie oceny kilku egzaminatorów) dla różnych sposobów ich ustalania

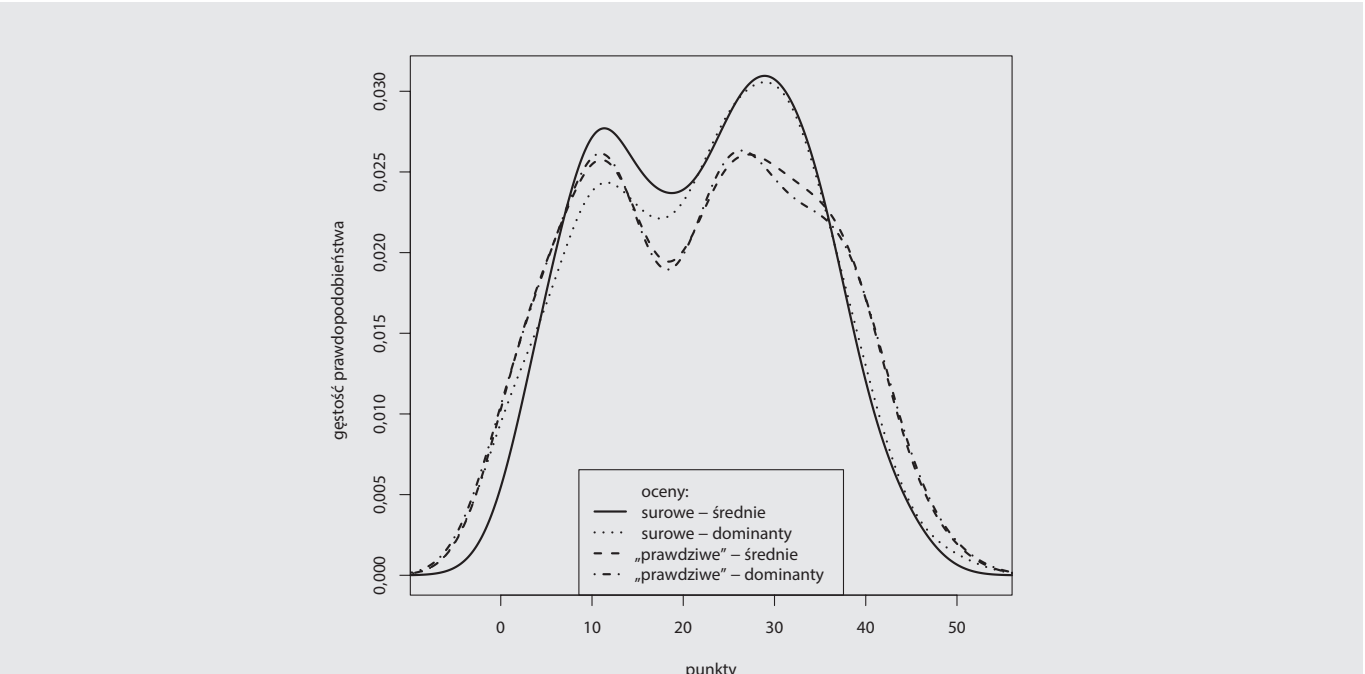


Tabela 9.8. Statystyki rozkładów ocen ostatecznych (oszacowanych na podstawie oceny kilku egzaminatorów) dla różnych sposobów ich ustalania

	Minimum	1 kwartył	Mediana	3 kwartył	Maksimum	Średnia	Odchylenie standardowe
surowe – średnie	0	12,16	23,12	30,75	46,12	21,96	10,99
surowe – dominanty	0	11,69	23,88	31,00	50,00	21,86	11,70
„prawdziwe” – średnie	0	11,38	23,66	32,04	48,58	22,01	12,50
„prawdziwe” – dominanty	0	11,00	23,00	32,00	49,00	21,92	12,62

Pierwszym pytaniem badawczym jest to, czy ocena przyznawana przez egzaminatorów za wypracowanie jest obciążona efektem formy zapisu (pismo odręczne vs zapis komputerowy). Możemy posłużyć się testem t, aby na nie odpowiedzieć. Dzięki niemu uzyskamy informację, czy istnieje statystycznie istotna różnica w średnich pomiędzy tymi dwoma grupami prac (PO vs PK). W związku z tym, że oceniana jest jedna praca w dwóch formach zapisu, należy użyć testu t dla prób zależnych. Oprócz opisanych powyżej czterech sposobów ustalenia ostatecznej oceny pracy, której można użyć w testach t, istnieje jeszcze jeden sposób na przeprowadzenie takiej analizy. Wspomniano wcześniej o łańcuchach Markowa zawierających wielokrotne oszacowania parametrów z modelu HRM-SDT. Mamy zatem do dyspozycji (pochodzące z użytego modelu) łańcuchy iteracji ocen „prawdziwych” każdej z prac. Dzięki temu możemy wykonać testy t dla każdej iteracji łańcuchów, a potem wyciągnąć średnią z otrzymanych wartości. Takie rozwiązanie umożliwia uwzględnienie podczas przeprowadzania testu t niepewności w oszacowaniu wyników „prawdziwych” przez model statystyczny. Tabela 9.9 przedstawia informacje o różnicach w średnich pomiędzy pracami pisanymi ręcznie i przepisnymi na komputerze w zależności od sposobu przeprowadzenia analizy. Prace przepisane na komputerze były oceniane niżej niż oryginalne o około 2–3 punktów. Wszystkie różnice okazały się istotne statystycznie przy poziomie istotności $p = 0,01$.

Tabela 9.9. Porównanie średnich wyników (test t) oceny wypracowań w zapisie oryginalnym (pismo odręczne – PO) i zapisie komputerowym (PK) w zależności od sposobu przeprowadzenia analizy

	Różnica wyników surowych (PO-PK)		Różnica wyników „prawdziwych” (PO-PK)		Średnia dla testów t wykonanych niezależnie dla każdej iteracji łańcuchów Markowa (PO-PK)
	średnia	dominanta	średnia	dominanta	
Różnica średnich (wraz z 95% przedziałem ufności)	2,52 (1,88–3,15)	2,30 (1,45–3,16)	3,09 (2,42–3,76)	3,21 (2,47–3,95)	3,09 (2,31–3,87)

Niezależnie od przyjętej metodologii obliczania testów t różnice oceny wypracowań w zapisie oryginalnym i ich kopii w zapisie komputerowym okazały się istotne statystycznie. Można więc wnioskować, iż taka zależność rzeczywiście występuje. Większe różnice w średnich przy użyciu wartości „prawdziwych” z użytego w badaniach IBE modelu HRM-SDT mogą wynikać z tego, iż wartości te powinny być pozbawione efektów egzaminatora, takich jak np. łagodność, o czym wspomniano wcześniej. Efekt egzaminatora może wpływać na oszacowania wyników surowych, poza tym zarówno obliczanie średniej, jak i dominanty z tych wyników posiada pewne wady, które opisano powyżej. Rozwiązaniem najbardziej poprawnym metodologicznie z punktu widzenia analizy bayesowskiej zastosowanej w badaniach porównywalności oceniania prowadzonych przez IBE jest przeprowadzenie testów t dla każdej iteracji łańcuchów i uśrednienie wyników. Warto zwrócić uwagę na duże podobieństwo otrzymanych w ten sposób wartości do wartości uzyskanych przy użyciu pozostałych dwóch metod bazujących na wynikach „prawdziwych” z modelu HRM-SDT.

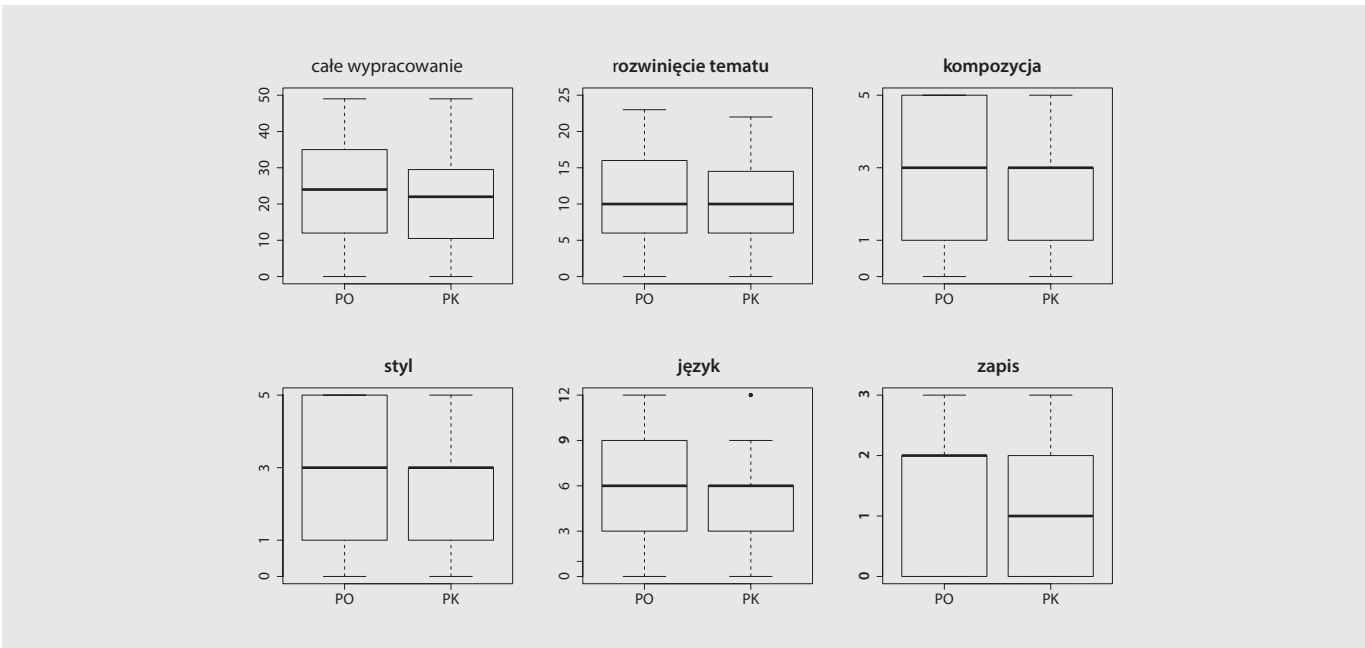
Zaprezentowane powyżej wyniki testów t użytych na sumarycznej ocenie za wypracowanie pozwalają odpowiedzieć na pierwsze z postawionych pytań badawczych. Istnieje obciążenie oceny za wypracowanie w zależności od formy jego zapisu. Można zatem szukać odpowiedzi na drugie z postawionych pytań – czy któreś z kryteriów oceniania jest szczególnie narażone na taki rodzaj obciążenia. W tym celu można również posłużyć się testami t, tym razem badając osobno poszczególne kryteria. Z poniższych analiz wyłączono kryterium *szczególne walory*, gdyż jest ono bardzo rzadko stosowane i w analizowanej próbie prac w przeważającej większości egzaminatorzy przyznawali za nie 0 punktów. Tabela 9.10. przedstawia różnice średnich pomiędzy pracami oryginalnymi (PO) i pracami przepisnymi (PK) w podziale na poszczególne kryteria. Podobnie jak w przypadku wyniku sumarycznego za całe wypracowanie, wyniki oceny za poszczególne kryteria były niższe dla prac przepisanych na komputerze w porównaniu z wynikami oceny w zapisie oryginalnym, odręcznym. Różnice średnich dla wszystkich kryteriów okazały się istotne statystycznie przy poziomie ufności $p = 0,01$. Analizowano głównie różnice obliczone jako średnia dla wyników testów t dla każdej iteracji łańcuchów, jako podejście uznane za najbardziej poprawne metodologicznie. Dla porównania użyto również ocen ustalonych na podstawie ocen „prawdziwych” z modelu HRM-SDT za pomocą średniej i dominanty.

Tabela 9.10. Porównanie średnich wyników oceny wypracowań w zapisie oryginalnym (pismo odręczne – PO) i zapisie komputerowym (PK) w rozbiciu na kryteria

Ocena za:	Skala	Średnia miar obliczonych niezależnie dla każdej iteracji łańcuchów Markowa			Różnica wyników „prawdziwych” PO-PK	
		różnica wyników PO-PK	odchylenie standardowe (SD)	efekt standardowy (PO-PK)/SD	średnia	dominanta
całe wypracowanie	0–50	3,09 (2,31–3,87)	12,82	24%	3,09 (2,42–3,76)	3,21 (2,47–3,95)
rozwińnięcie tematu	0–25	0,80 (0,41–1,20)	5,67	14%	0,80 (0,49–1,12)	0,85 (0,50–1,20)
kompozycję	0,1,3,5	0,33 (0,16–0,50)	1,78	19%	0,33 (0,20–0,46)	0,31 (0,16–0,46)
styl	0,1,3,5	0,54 (0,34–0,74)	1,7	32%	0,54 (0,38–0,70)	0,55 (0,35–0,75)
język	0,1,3,6,9,12	1,17 (0,83–1,52)	3,31	35%	1,17 (0,89–1,45)	1,23 (0,89–1,56)
zapis	0,1,2,3	0,24 (0,12–0,36)	1,10	22%	0,24 (0,14–0,33)	0,26 (0,15–0,36)

Analiza różnicy średnich ocen pomiędzy pismem odręcznym i pismem komputerowym w podziale na poszczególne kryteria oceny wyraźnie wskazuje, że obciążenie ocen efektem rodzaju pisma jest zróżnicowane ze względu na kryteria oceniania. Umieszczony w tabeli efekt standardowy (różnica pomiędzy grupami podzielona przez odchylenie standardowe) wskazuje, że kryterium **rozwińnięcia tematu** charakteryzuje się najmniejszym wpływem rodzaju zapisu. Z największym obciążeniem mamy natomiast do czynienia w kryteriach **styl** i **język** – efekt standardowy jest tu dwukrotnie wyższy niż dla **rozwińnięcia tematu**. Rysunek 9.8 przedstawia wykresy skrzynkowe ilustrujące różnicę w rozkładach wyników za realizację wypracowania w całości oraz w zakresie poszczególnych kryteriów.

Rysunek 9.8. Porównanie rozkładów wyników uzyskanych przy ocenie wypracowania przez egzaminatorów w jego oryginalnym zapisie pismem odręcznym (PO) i po przepisaniu wypracowania na komputerze (PK), w rozbiciu na poszczególne kryteria oceny



Wypracowanie w zakresie **języka** oceniane było na sześciostopniowej skali, gdzie wyniki mogły przybierać dyskretne wartości 0; 1; 3; 6; 9; 12. Zgodnie z kryteriami CKE oceniania egzaminu maturalnego na poziomie podstawowym w 2012 roku, jeżeli język w całej pracy, zdaniem egzaminatora, był komunikatywny, a zdający stosował poprawną, urozmaiconą składnię, poprawne słownictwo, frazeologię i fleksję, to powinien otrzymać 12 punktów. Praca charakteryzująca się komunikatywnym językiem, poprawną składnią, słownictwem, frazeologią i fleksją zasługiwała na 9 punktów. Jeżeli język w całej pracy był komunikatywny, poprawna fleksja, w większości poprawne składnia, słownictwo i frazeologia, egzaminator powinien przyznać 6 punktów. Tylko na 3 punkty w zakresie tego kryterium oceniane było wypracowanie, w którym język był komunikatywny mimo błędów składniowych, leksykalnych (słownictwo i frazeologia), fleksyjnych. Przy występowaniu błędów fleksyjnych, licznych błędów składniowych, leksykalnych wypracowanie oceniane było na 1 punkt (CKE, 2012).

Jeżeli chodzi o kryterium **styl**, dla którego także zaobserwowano znaczny efekt standardowy różnicy oceniania wypracowania (PO-PK), to maksymalną liczbę punktów (5) zdający mógł uzyskać, jeżeli egzaminator stwierdził, że styl pracy jest jasny, żywy, swobodny, zgodny z zastosowaną formą wypowiedzi, a leksyka jest urozmaicona. Jeżeli styl wypracowania był zgodny z zastosowaną formą wypowiedzi, na ogół jasny, a leksyka wystarczająca, to egzaminator mógł przyznać 3 punkty. Na 1 punkt pod względem stylu oceniane było wypracowanie, w którym styl był na ogół komunikatywny (dopuszczalne schematy językowe) (CKE, 2012).

Na podstawie przeprowadzonych analiz hipoteza zerowa o braku różnic w ocenach przyznawanych przez egzaminatorów pracom różniącym się jedynie formą zapisu musi być odrzucona. Dziwić może szczególne obciążenie kryteriów **stylu i języka**, gdyż przytoczone powyżej opisy nie zawierają cech, które mogłyby być bardzo zależne od formy zapisu. Pewnym wyjaśnieniem w przypadku kryterium **języka** mogłaby być dobra czytelność tekstu, a zatem pewność co do popełnionych błędów, które w formie pisma ręcznego mogą być niewyraźne. Nie potwierdzają się też intuicyjne odczucia egzaminatorów (które autorzy mogli obserwować, wielokrotnie uczestnicząc w sesjach egzaminacyjnych), że prace przepisane na komputerze uzyskują wyższe oceny niż przedstawione do oceny w zapisie pisma odręcznego. Obciążenie oceniania prac różniących się tylko formą zapisu (przyznawanie tym samym pracom w odręcznym zapisie wyższych ocen niż w zapisie komputerowym) zostało także zaobserwowane w innych badaniach (Powers i in., 1994).

Reasumując, na podstawie analizy danych z badań przeprowadzonych możemy wnioskować, że prace pisane ręcznie oceniane są wyżej niż ich kopie zapisane na komputerze. Różnica średniej ocen dla wypracowania z egzaminu maturalnego ocenianego w skali pięćdziesięciopunktowej przekracza 2 punkty i jest statystycznie istotna na poziomie istotności $p = 0,01$. Podobny wynik uzyskano w wielu badaniach porównujących jakość oceniania tekstów pisanych odręcznie i ich kopii wykonanych z wykorzystaniem procesora tekstu zarówno przez uczniów z dysfunkcjami w pisaniu, jak i bez dysfunkcji w tym zakresie (Arnold i in., 1990; Cahalan-Laitusis, 2003; Powers, Fowles, Farnum i Ramsey, 1994; Russell i Plati, 2001).

Egzaminatorzy oceniający w sesji egzaminacyjnej prace zarówno w formie odręcznego zapisu, jak i w zapisie komputerowym podkreślają, że ten efekt może być spowodowany dwoma czynnikami. Po pierwsze, podczas oceniania wypracowania napisanego odręcznie łatwiej sobie wyobrazić ucznia, który jest autorem wypracowania, co ma wpływ na efekt łagodności oceniania. Ponadto w zapisie odręcznym w przypadku wątpliwości egzaminatora co do jakości fragmentu tekstu (np. skutek obniżonej czytelności tekstu) egzaminatorzy podejmują decyzję na korzyść zdającego, zresztą z obowiązującymi zasadami oceniania w sesji egzaminacyjnej. Ponadto w komputerowym tekście wszystkie błędy są bardziej widoczne. Zjawisko to, polegające na tendencji zawyżania oceny w wyniku utożsamiania się oceniającego z autorem wypracowania napisanego odręcznie, możemy nazwać efektem empatii w ocenianiu (ang. *Reader Empathy Assessment Discrepancy [R-E-A-D effect]*) (Arnold i in., 1990). Zostało ono zaobserwowane zarówno w przypadku tekstów pisanych ręcznie, a potem skopiowanych z wykorzystaniem procesora tekstu, jak i w sytuacji odwrotnej, kiedy tekst oryginalnie pisany na komputerze został skopiowany przez odręczne przepisanie (Powers i in., 1994).

9.4. Wynik oceniania a czytelność pisma

W ramach badania podjęto także próbę weryfikacji w warunkach polskich wyników wcześniejszych badań, które sugerują, że na ocenę może mieć wpływ czytelność pisma (Wood, 1991). Zgodnie z powszechnym przekonaniem, prace pisane mniej czytelnie są oceniane niżej, niż prace pisane czytelnie (James, 1927; Klein i Taub, 2005; Markham, 1976). Zjawisko to można łączyć z szerzej rozumianym efektem egzaminatora, w tym efektem halo, czyli świadomego, bądź nie, uwzględniania przez egzaminatora przy

ocenie ogólnego wrażenia związanego z wypracowaniem (Dubiecka i in., 2006). W tym kontekście prace pisane mniej czytelnym charakterem pisma mogą zostać potraktowane jako napisane mniej starannie, a więc też potencjalnie jako dzieło ucznia, który się do pracy „mniej przyłożył”, a tym samym mogą zostać ocenione niżej.

9.4.1. Oceny charakteru pisma

Przy wyborze 120 prac wykorzystano skalę zaproponowaną przez Jerzego Frąszczaka (2013), wyróżniającą trzy kryteria oceny charakteru pisma, które w znacznym stopniu mają wpływ na komfort czytania i oceny wypracowań maturalnych z języka polskiego. Są to: czytelność pisma, wyrobienie pisma i wielkość pisma.

W zakresie czytelności wyróżniono trzy kategorie:

- A. Pismo czytelne w aspekcie wszystkich desygnatów brzmieniowych i znaczeniowych. Konstrukcje znaków graficznych zbliżone do wzorców elementarzystycznych i drukowych, z nielicznymi modyfikacjami modelunków znaków graficznych – wynikających z ekonomiki kreślenia.
- B. Pismo czytelne w aspekcie desygnatów brzmieniowych i znaczeniowych, ale zawierające uproszczenia i modyfikacje konstrukcji graficznych, które jednakże nie utrudniają ich prawidłowej interpretacji.
- C. Pismo częściowo nieczytelne – zawierające uproszczenia, modyfikacje i redukcje poszczególnych konstrukcji liter, części i całych wyrazów oraz ciągów wyrazowych, powodujące trudności w ich jednoznacznej interpretacji brzmieniowej i znaczeniowej.

Podobnie w zakresie wyrobienia pisma wyróżniono – (a) pismo wyrobione, (b) średnio wyrobione i (c) niewyrobione²⁷.

- a. Pismo wyrobione, charakteryzujące się między innymi dużą płynnością linii graficznej, subtelnym cieniowaniem pisma, swoistym rozwiązaniem budowy znaków oraz znaczną różnorodnością elementów powtarzalnych.
- b. Pismo średnio wyrobione, charakteryzujące się między innymi umiarkowaną płynnością linii graficznej, zauważalnym cieniowaniem pisma, pojawiającymi się niekiedy swoistymi rozwiązaniami budowy znaków, średnim bogactwem elementów powtarzalnych i średnim tempem pisania.
- c. Pismo niewyrobione, charakteryzujące mało zróżnicowanym cieniowaniem, mało płynnym przebiegiem linii i wolnym tempem kreślenia.

W zakresie wielkości pisma rozróżniono pięć kategorii:

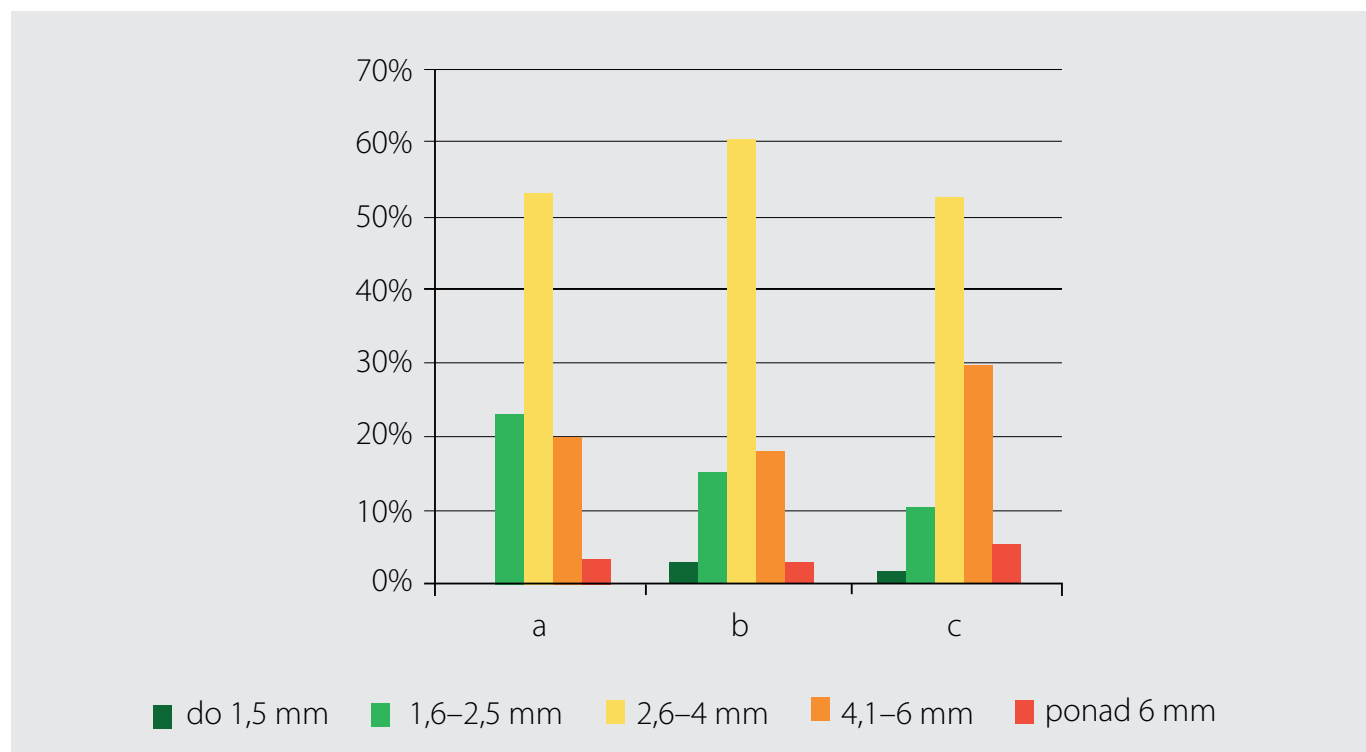
- I. pismo drobne – do 1,5 mm
- II. pismo małe – od 1,5 mm do 2,5 mm
- III. pismo średnie – od 2,6 mm do 4,0 mm
- IV. pismo duże – od 4,1 mm do 6,0 mm
- V. pismo bardzo duże – ponad 6 mm.

Próba 120 wypracowań składała się w 25 procentach z wypracowań czytelnych w kategorii A i 27 procentach czytelnych w kategorii B. Prace mogące sprawiać trudność w czytaniu (kategoria C) stanowiły 48 procent próby. Udział wypracowań charakteryzujących się różnymi klasami wyrobienia pisma i jego wielkości w wymienionych kategoriach czytelności przedstawiają rysunek 9.9 i rysunek 9.10. Prace nieczytelne (kategoria C) i równocześnie niewyrobione (kategoria c) stanowiły 41 procent próby.

²⁷

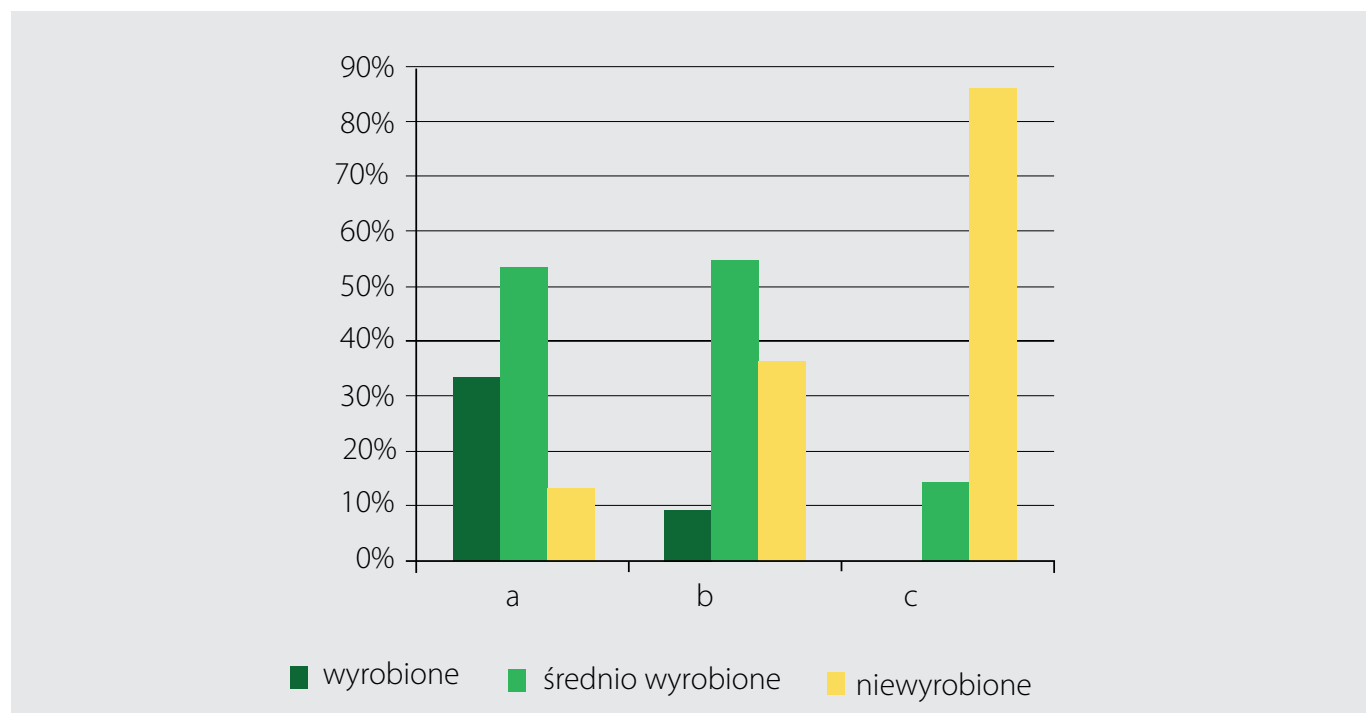
Na podstawie *Słownika Terminów Pismoznawczych*, <http://prawouam-stp.home.amu.edu.pl/p.htm>.

Rysunek 9.9. Reprezentacja w próbie pięciu kategorii wielkości pisma z uwzględnieniem czytelności



Uwaga: Procent dla każdej klasy wielkości obliczony jest względem liczebności wypracowań w poszczególnych kategoriach czytelności a, b, c

Rysunek 9.10. Reprezentacja w próbie trzech kategorii wyrobienia pisma z uwzględnieniem czytelności

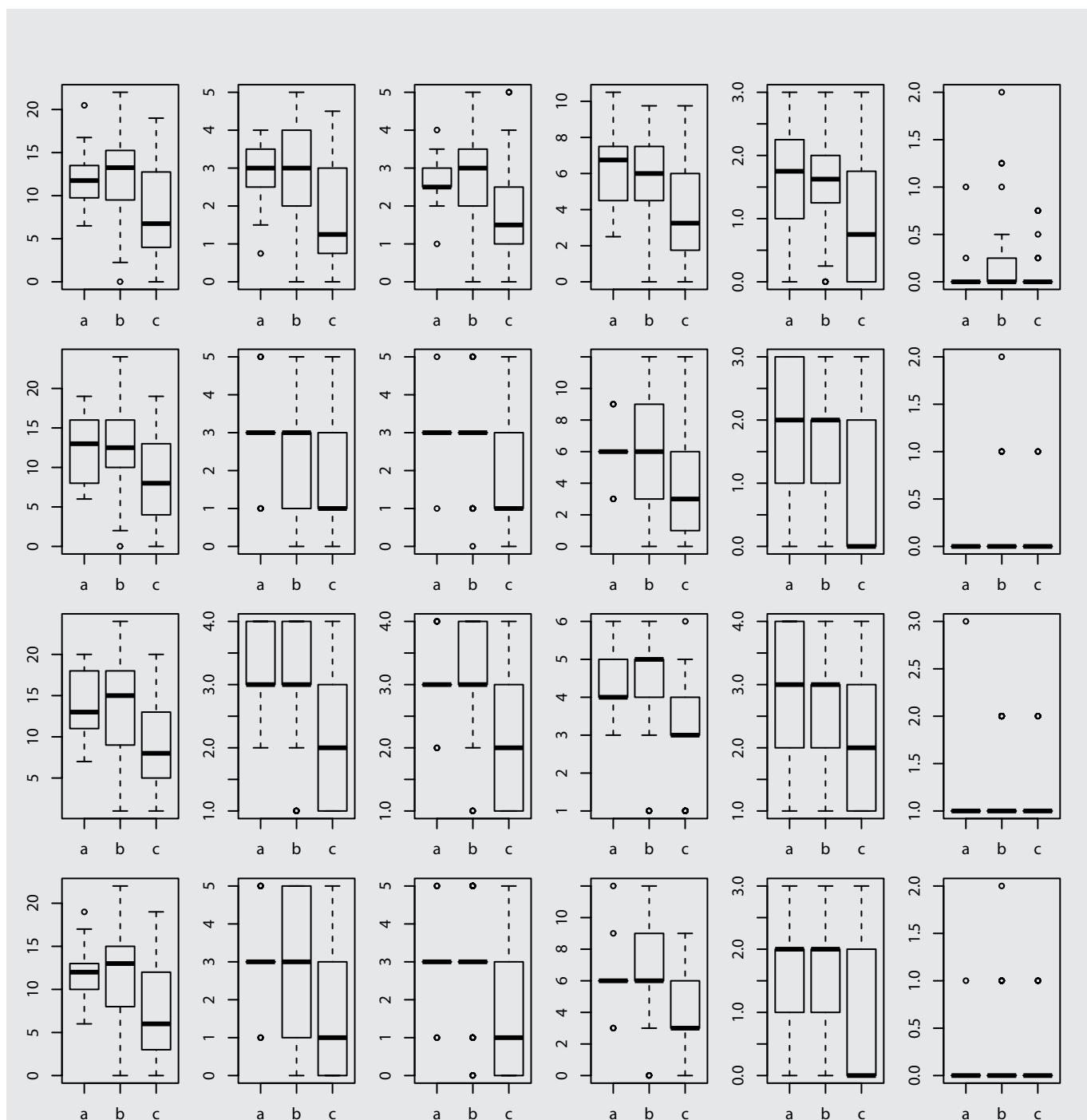


Uwaga: Procent dla każdej klasy wyrobienia obliczony jest względem liczebności wypracowań w poszczególnych kategoriach czytelności a, b, c.

9.4.2. Porównanie ocen prac różniących się charakterem pisma

W toku analiz porównano oceny przyznane pracom pisany różnym charakterem pisma. Rysunek 9.11 przedstawia zestawienie ocen przyznawanych przez egzaminatorów pracom pochodzącym z poszczególnych kategorii charakteru pisma ocenianego w oparciu o kryteria czytelności, rozmiaru i wyrobienia.

Rysunek 9.11. Oceny przyznane za poszczególne kryteria oceniania dla prac z poszczególnych kategorii jakości pisma odręznego



Uwaga: Na osi pionowej – liczba punktów przyznawanych za poszczególne kryteria, na osiach poziomych – kategorie czytelności, rozmiaru i wyrobienia pisma.

9.4.2.1. Porównanie ocen prac różniących się charakterem pisma przy użyciu jednoczynnikowej analizy wariancji

W związku z faktem, iż w większości przypadków poszczególne kategorie zmiennych nie były reprezentowane równolicznie oraz charakteryzowały się różnymi wariancjami, w porównaniach użyto testu ANOVA z poprawką Welcha. Wyniki porównań przedstawiono w tabelach 9.11, 9.12 i 9.13. W związku z małą liczebnością próby dane te nie zawsze pozwalały na porównania między poszczególnymi kategoriami, co w szczególności dotyczyło rozmiaru pisma, ponieważ ta zmienna przyjmuje aż pięć różnych kategorii. Podobnie jak w przypadku porównań dotyczących prac pisanych odręcznie i komputerowo, w analizie uwzględniono cztery wskaźniki ocen (por. podrozdział 9.5.2): średnią z ocen egzaminatorów, dominantę z ocen egzaminatorów, średni wynik „prawdziwy” oraz dominantę z łańcuchów wyników „prawdziwych”. Jak możemy się przekonać z zamieszczonych poniżej tabel, w przypadku czytelności zaobserwowano różnice w ocenianiu istotne na poziomie $p < 0,01$ dla *stylu i języka*. W przypadku rozmiaru liter, z powodu małej wielkości prób, co wynikało z porównań między pięcioma kategoriami wielkości liter, w większości przypadków nie udało się dokonać wiarygodnej oceny związków tej zmiennej z ocenianiem. Trzecie z wyróżnionych przez grafologa kryteriów, wyrobienie pisma, różnicowało na poziomie $p < 0,01$ wszystkie kategorie ocen z wyjątkiem ocen *szczególnych walorów pracy*, należy tu jednak pamiętać, że oceny za to kryterium przyznawane są rzadko, w związku z czym trudno jest stwierdzić czy ten brak istotności wynika z małej liczności próby, czy z faktycznego braku związku.

Tabela 9.11. Wpływ czytelności na przyznane oceny

	Ocena obserwowana				Ocena „prawdziwa”			
	średnia		dominanta		średnia		dominanta	
	<i>F</i>	<i>Pr(>F)</i>	<i>F</i>	<i>Pr(>F)</i>	<i>F</i>	<i>Pr(>F)</i>	<i>F</i>	<i>Pr(>F)</i>
Treść	2,20	0,12	2,04	0,14	3,22	0,05	2,43	0,10
Kompozycja	3,10	0,05	2,98	0,06	5,30	<0,01	4,41	0,02
Styl	5,29	<0,01	6,46	<0,01	5,56	<0,01	5,40	<0,01
Język	5,13	<0,01	4,50	0,10	7,42	<0,01	6,24	<0,01
Zapis	3,70	0,03	4,07	0,02	2,13	0,13	4,01	0,02
Walory	0,79	0,46	0,53	0,59	0,12	0,89	0,44	0,64

Tabela 9.12. Wpływ rozmiaru pisma na przyznane oceny

	Ocena obserwowana				Ocena „prawdziwa”			
	średnia		dominanta		średnia		dominanta	
	<i>F</i>	<i>Pr(>F)</i>	<i>F</i>	<i>Pr(>F)</i>	<i>F</i>	<i>Pr(>F)</i>	<i>F</i>	<i>Pr(>F)</i>
Treść	2,22	0,18	2,63	0,13	2,05	0,20	2,90	0,11
Kompozycja	35,52	<0,01	–	–	–	–	–	–
Styl	38,42	<0,01	–	–	–	–	–	–
Język	22,94	<0,01	–	–	2,37	0,16	–	–
Zapis	–	–	–	–	–	–	–	–
Walory	–	–	–	–	–	–	–	–

Uwaga: Puste pola (–) w tabeli wskazują na niemożność oszacowania wartości testu w związku z niewielką liczebnością próby.

Tabela 9.13. Wpływ wyrobienia pisma na przyznane oceny

	Ocena obserwowana				Ocena „prawdziwa”			
	średnia		dominanta		średnia		dominanta	
	F	Pr(>F)	F	Pr(>F)	F	Pr(>F)	F	Pr(>F)
Treść	9,00	<0,01	8,37	<0,01	11,06	<0,01	9,96	<0,01
Kompozycja	10,92	<0,01	9,49	<0,01	16,53	<0,01	12,01	<0,01
Styl	11,92	<0,01	11,39	<0,01	13,52	<0,01	13,97	<0,01
Język	12,14	<0,01	9,28	<0,01	14,15	<0,01	12,53	<0,01
Zapis	7,89	<0,01	8,71	<0,01	6,32	<0,01	10,54	<0,01
Walory	2,74	0,08	–	–	1,89	0,17	1,83	0,18

Uwaga: Puste pola w tabeli wskazują na niemożność oszacowania wartości testu w związku z niewielką liczebnością próby.

9.4.2.2. Porównanie ocen prac różniących się charakterem pisma przy użyciu alternatywnego modelu efektu egzaminatora

W celu oszacowania związku efektu egzaminatora z innymi zmiennymi wyjaśniającymi użyto dodatkowo mieszanego modelu liniowego. Metoda ta służy do estymacji modeli liniowych dla danych o charakterze hierarchicznym, gdzie obserwacje z niższego poziomu zagnieżdżone są wewnątrz grup na wyższym poziomie. Z taką sytuacją mamy do czynienia w przypadku zbioru danych pochodzących z tego badania, ponieważ każdy egzaminator oceniał kilka różnych wypracowań, a więc wypracowania zagnieżdżone były w egzaminatorach, a równocześnie każda praca oceniana była przez kilku egzaminatorów, więc egzaminatorzy byli zagnieżdżeni w wypracowaniach. Mieszane modele liniowe pozwalają uwzględnić te zależności w modelu statystycznym interesującego nas zjawiska. W najprostszej postaci model taki można zapisać jako:

$$y_m = \beta_{typ} + \beta_{pismo} + b_r + b_n + \varepsilon_{rn}$$

$$b_r \sim N(0, \sigma_{b_r}^2), \quad b_n \sim N(0, \sigma_{b_n}^2), \quad \varepsilon_{rn} \sim N(0, \sigma^2),$$

gdzie β_{typ} to efekt stały typu wypracowania (rok i temat), β_{pismo} to efekt stały jednego z kryteriów charakteru pisma, b_r to efekt losowy r -tego egzaminatora, b_n to efekt losowy n -tego wypracowania, a ε_{rn} to błąd. Oszacowania pochodzące z takiego modelu poddają się łatwej interpretacji, ponieważ efekt losowy wypracowania mówi nam o tym, w jaki sposób dane wypracowanie różni się od pozostałych, a efekt losowy egzaminatora mówi o tym, czy i w jakim stopniu dany egzaminator jest bardziej łagodny lub surowy od pozostałych egzaminatorów. W modelu możemy uwzględnić dodatkowo inne zmienne wyjaśniające, na przykład czytelność, rozmiar lub wyrobienie pisma. Tabela 9.14 poniżej zaprezentuje porównanie dopasowania modeli: (a) zerowego – uwzględniającego jedynie efekt losowy pracy, efekt losowy wypracowania i efekt stały typu wypracowania oraz uwzględniającego dodatkowo (b) czytelność, (c) wyrobienie i (d) rozmiar pisma dla wszystkich kategorii oceny wypracowania. Jak możemy się przekonać, najlepiej dopasowanym modelem jest ten uwzględniający wyrobienie pisma i to dla wszystkich kategorii oceny pracy. Nieco gorzej, choć wciąż istotnie lepiej w stosunku do modelu zerowego, dopasowany jest model uwzględniający czytelność pisma, w szczególności dla ocen **kompozycji, stylu i języka**. Skala tych różnic waha się między poszczególnymi kryteriami oceny, przedstawiając jednak podobny trend, gdzie prace pisane pismem niewyrobionym (kategoria c) oceniane są niżej, a prace o charakterze pisma średnio wyrobionym (kategoria b) oceniane są tak samo lub nieznacznie niżej w porównaniu do prac o wyrobionym charakterze pisma. Ze względu na podobieństwa między wynikami dla poszczególnych kryteriów oceny zrezygnowano z przedstawiania wyników szczegółowych, zamiast tego za przykład posłuży oszacowanie dla całościowej oceny wypracowania (tabela 9.15). Prace charakteryzujące się niewyrobionym pismem, w stosunku do pisanych najlepiej wyrobionym, oceniane były średnio o 9,84²⁸ punktów niżej przy kontroli innych zmiennych w opisanym powyżej modelu.

²⁸ Różnica surowych średnich nieuwzględniająca korekty ze względu na efekt egzaminatora, umiejętności ucznia i wybrany przez ucznia temat to 8,33 punktów.

Tabela 9.14. Porównanie dopasowania mieszanych modeli liniowych efektu egzaminatora

Zmienne wyjaśniające	Miary dopasowania modelu					
	Kryterium Informacyjne Aikake (AIC)	Bayesowskie Kryterium Informacyjne (BIC)	Logarytm Wiarygodności	Dewiancja	χ^2	$\Pr(>\chi^2)$
Treść						
(model zerowy)	22123	22171	-11055	22109	–	–
Czytelność	22119	22181	-11051	22101	8,06	0,02
Wyrobienie	22109	22171	-11046	22091	10,12	<0,01
Rozmiar	22122	22198	-11050	22100	0,00	1,00
Kompozycja						
(model zerowy)	9922,8	9970,6	-4954,4	9908,8	–	–
Czytelność	9915,9	9977,4	-4949,0	9897,9	10,87	<0,01
Wyrobienie	9902,8	9964,3	-4942,4	9884,8	13,14	<0,01
Rozmiar	9916,4	9991,6	-4947,2	9894,4	0,00	1,00
Styl						
(model zerowy)	9966,3	10014	-4976,2	9952,3	–	–
Czytelność	9953,7	10015	-4967,9	9935,7	16,62	<0,01
Wyrobienie	9945,7	10007	-4963,8	9927,7	8,03	<0,01
Rozmiar	9956,5	10032	-4967,3	9934,5	0,00	1,00
Język						
(model zerowy)	18696	18744	-9340,8	18682	–	–
Czytelność	18686	18748	-9334,0	18668	13,67	<0,01
Wyrobienie	18679	18741	-9330,6	18661	6,74	<0,01
Rozmiar	18691	18766	-9334,5	18669	0,00	1,00
Zapis						
(model zerowy)	5372,7	5420,5	-2679,3	5358,7	–	–
Czytelność	5372,1	5433,6	-2677,0	5354,1	4,61	0,10
Wyrobienie	5368,9	5430,4	-2675,4	5350,9	3,17	<0,01
Rozmiar	5373,4	5448,6	-2675,7	5351,4	0,00	1,00
Walory						
(model zerowy)	-2794,0	-2746,2	1404,0	-2808,0	–	–
Czytelność	-2795,8	-2734,3	1406,9	-2813,8	5,77	0,06
Wyrobienie	-2797,3	-2735,8	1407,7	-2815,3	1,49	<0,01
Rozmiar	-2792,9	-2717,7	1407,5	-2814,9	0,00	1,00

Uwaga: Niższe wartości AIC, BIC i dewiancji oraz wyższa wartość logarytmu wiarygodności sugerują lepsze dopasowanie modelu. Dwie ostatnie kolumny prezentują wyniki porównań parami kolejnych modeli na podstawie testu χ^2 , istotność $\Pr(>\chi^2)$ na poziomie $p < 0,01$ wskazuje, że dany model jest lepiej dopasowany do danych w stosunku do poprzedzającego.

Tabela 9.15. Oszacowania mieszanego modelu liniowego efektu egzaminatora dla całościowej oceny wypracowania

Efekty losowe			
	Wariancja	Odchylenie Standardowe	
Praca	106,76	10,33	
Egzaminator	28,98	5,38	
(reszta)	4,56	2,13	
Efekty stałe			
	Oszacowanie	Błąd standardowy	Wartość t
(stała) ²⁹	29,40	3,97	7,40
Rok 2011, temat II	-6,28	2,90	-2,16
Rok 2012, temat I	-1,75	3,06	-0,57
Rok 2012, temat II	-2,29	3,33	-0,69
kategoria b	-0,17	3,34	-0,05
kategoria c	-9,84	3,24	-3,04

9.4.3. Podsumowanie

W trakcie analiz porównano oceny maturalne przyznawane wypracowaniom z języka polskiego w oparciu o charakter pisma ucznia. Znaczny odsetek wypracowań pisanych jest nieczytelnym i/lub niewyrobionym charakterem pisma, co może sprawiać egzaminatorom trudności w odczytaniu treści prac. Wyniki analiz wskazują, że prace pisane charakterem pisma cechującym się małym wyrobieniem oceniane są niżej w stosunku do prac pisanych wyrobionym lub średnio wyrobionym charakterem pisma. Związki czytelności pisma z oceną otrzymaną przez ucznia są umiarkowane, jednak należy tu zaznaczyć, że czytelność i wyrobienie są silnie skorelowanymi cechami pisma, więc znaczna większość prac mało czytelnych to również prace pisane mało wyrobionym charakterem pisma. Nie zaobserwowano związków wielkości liter używanych przez uczniów z uzyskiwanymi ocenami. Wyniki te są zgodne z rezultatami pochodzącymi z wcześniejszych badań. Równocześnie warto zauważyć, iż fakt, że to raczej wyrobienie niż czytelność pisma związana jest z oceną pracy, wydaje się zgodny z hipotezą, że efekt ten wynika nie tyle z niemożności odczytania treści pracy przez egzaminatora, ale raczej z niskich walorów estetycznych liter, co wiąże się z wrażeniem, jakie praca wywiera na egzaminatorze.

9.5. Kolejność i czas oceniania pracy egzaminacyjnej a wynik oceny

Kolejność oceniania danej pracy oraz czas jej oceniania mogą być ważnymi czynnikami wpływającymi na jej punktowanie. Można sobie wyobrazić, że wraz z kolejną ocenianą pracą rośnie zmęczenie egzaminatora, a wraz z nim może obniżać się koncentracja. Wpływ tego czynnika na sposób oceny prac jest dyskusyjny – nie wiadomo, czy egzaminator zmęczony ocenia w sposób bardziej surowy czy łagodny. Prowadzone badania dają niespójne rezultaty. Badania Lunz i Stahl (1990) wskazują, że wraz z upływem czasu egzaminatorzy stają się bardziej łagodni, z kolei Congdon i McQueen (2000) oraz Pinot de Moira, Massey, Baird i Morissey (2002) zwracają uwagę na wzrastającą surowość egzaminatorów. W badaniu Leckiego i Baird (2011) stwierdzono istotne statystycznie wahania surowości wraz z upływem czasu u poszczególnych egzaminatorów, jednak średni poziom ich surowości nie różnił się statystycznie. Wraz z upływem czasu grupa egzaminatorów stawała się bardziej homogeniczna pod względem surowości oceniania. Mogło być to spowodowane dwoma czynnikami: nabyciem wprawy w ocenianiu oraz uzyskaniem zwrotnej opinii od przewodniczących zespołów egzaminatorów.

Kolejność ocenianej pracy jest więc czynnikiem, który może w sposób pozytywny (wprawa), jak i negatywny (zmęczenie, obniżenie koncentracji) wpływać na rzetelność jej oceny.

²⁹ W trakcie analizy użyto także centrowania sumy punktów w oparciu o średnią, jednak oszacowanie w oparciu o takie dane, modelu nieuwzględniającego efektu stałego, nie okazało się znacząco różne od wyników zaprezentowanych w tabeli 9.15.

Czas oceniania pracy może być uwarunkowany jej specyfiką, czyli ocenianym przedmiotem (język polski lub matematyka), podstawowym bądź rozszerzonym poziomem zadania oraz w przypadku niniejszego badania pisaniem jej odręcznie, bądź przepisaniem na komputerze. Można sobie wyobrazić, że średni czas oceniania pracy jest też indywidualną charakterystyką egzaminatora, zależącą np. od jego temperamentu czy cech osobowości (np. sumienność). W związku z wpływem tych wszystkich czynników nie wiadomo jak czas oceniania poszczególnych prac przekłada się na ich ocenę.

Przeprowadzono analizę wariancji określającą, czy typ pracy ma znaczenie dla czasu jej oceniania. W przypadku prac z języka polskiego średni czas oceny poszczególnych typów prac zawiera poniższa tabela (tabela 9.16). Analiza wariancji ($F(3, 9613) = 37,28$; $p = 0,000$) wskazuje, że średni czas oceniania pracy zależy od jej typu. Przeprowadzona analiza post hoc wskazała, że różnice w średnim czasie oceny pomiędzy wszystkimi typami prac są istotne statystycznie poza różnicą w czasie oceny między tematem I i II z 2012 roku.

Tabela 9.16. Średni czas oceny (w minutach) według typu pracy (język polski)

Rodzaj pracy	Średni czas oceniania	Odchylenie std.
temat I, rok 2011	20,50	9,31
temat I, rok 2012	19,18	8,82
temat II, rok 2011	17,89	8,16
temat II, rok 2012	18,68	7,32
Łącznie	18,90	8,44

Analiza wariancji sprawdzająca istotność różnic w czasie oceniania dla prac z matematyki³⁰ wykazała istotne zróżnicowanie ze względu na poziom pracy – rozszerzony lub podstawowy ($F(3, 8173) = 1048,57$; $p = 0,000$). Seria testów post hoc wskazuje, że wszystkie typy prac różnią się od siebie istotnie pod kątem czasu oceniania. Dokładne wartości średniego czasu oceniania prac poziomu podstawowego i rozszerzonego 2011 i 2012 wskazuje poniższa tabela (tabela 9.17).

Tabela 9.17. Średni czas oceny (w minutach) według typu pracy (matematyka)

Rodzaj pracy	Średni czas oceniania	Odchylenie std.
poziom podstawowy, 2011	10,23	5,99
poziom podstawowy, 2012	8,68	5,40
poziom rozszerzony, 2011	20,25	9,86
poziom rozszerzony, 2012	18,02	9,59
Łącznie	14,28	9,37

W celu oszacowania wpływu zmiennych czasowych (takich jak kolejność oceniania prac oraz czas jej oceniania) na ocenę pracy wyspecyfikowano szereg modeli regresji.

W pierwszym modelu oszacowano wpływ kolejności oceniania pracy na sumę punktów uzyskanych z danej pracy (wypracowania w przypadku języka polskiego, zadania otwarte w przypadku matematyki). W modelu uwzględniono także takie zmienne jak: dominanta obserwowanych sum punktów wystawionych danej pracy, przedmiot (0 – polski, 1 – matematyka) oraz fakt, czy praca była przepisana lub nie (0 – nie, 1 – tak). W tabeli 9.18 można zauważyć, że choć wpływ kolejności jest istotny statystycznie, współczynnik regresji jest minimalny – rzędu 0,02. Co ciekawe, gdyby w modelu uwzględnić średnią lub medianę obserwowanych sum punktów wystawionych danej pracy, wpływ kolejności jest nieistotny statystycznie.

³⁰ Przy czym należy pamiętać o stosowaniu przez większość OKE specjalizacji w ocenianiu zadań. W praktyce oznacza to, że egzaminator nie ocenia całej pracy, a grupę zadań w ramach pracy. Na ocenę całej pracy składają się więc oceny kilku egzaminatorów.

W kolejnym modelu (model 2) oszacowano wpływ czasu oceniania pracy na sumę punktów uzyskanych z danej pracy. W modelu wzięto pod uwagę też pozostałe predyktory, czyli dominantę obserwowanych sum punktów, przedmiot oraz przepisanie pracy. Podobnie jak w przypadku kolejności prac, efekt czasu oceniania (pojedynczej pracy) jest istotny statystycznie, jednak bliski zeru (czas oceniania pracy = 0,016). Interakcja czasu oceniania pracy z przedmiotem oceniania jest istotna statystycznie, ale efekt ten jest niewielki i wynosi 0,07.

W modelu 3 suma punktów z danego egzaminu maturalnego jest przewidywana na podstawie zarówno czasu oceniania pracy, jak i kolejności pracy. W modelu uwzględniono też kowarianty z innych modeli, tj.: dominantę obserwowanych sum punktów wystawionych przez egzaminatorów, fakt przepisania (lub nie) pracy, przedmiot (język polski lub matematyka) oraz interakcję czasu oceniania z przedmiotem. Jak można zauważyć w tabeli 9.18, choć kolejność pracy i czas jej oceniania w sposób istotny statystycznie wpływają na uzyskaną sumę punktów z danej pracy, ich wpływ jest minimalny (rzędu 0,02). Wpływ interakcji czasu oceniania z przedmiotem jest równie niewielki (0,06).

Tabela 9.18. Współczynniki regresji dla modelu uwzględniającego kolejność oceniania pracy

	Model 1		Model 2		Model 3	
	Współczynnik	t	Współczynnik	t	Współczynnik	t
STAŁA	3,323 (0,104)	32,071	3,638 (0,132)	27,630	3,22 (0,16)	20,66
dominanta	0,929 (0,003)	310,337	0,918 (0,003)	292,884	0,92 (0,00)	290,00
kolejność	0,022 (0,004)	5,588	-	-	0,02 (0,00)	5,01
czas oceniania	-	-	0,016 (0,006)	2,72	0,02 (0,01)	3,69
przedmiot	-1,866 (0,076)	-24,580	-2,82 (0,158)	-17,880	-2,66 (0,16)	-16,52
przepisanie pracy	-2,160 (0,230)	-9,393	-2,17 (0,230)	-9,479	-2,17 (0,23)	-9,46
interakcja czas oceniania ↔ przedmiot	-	-	0,07 (0,008)	8,167	0,06 (0,01)	6,95

* wszystkie współczynniki są istotne statystycznie
** przy wartości współczynników w nawiasach podano błędy standardowe

Przedstawione wyniki wskazują, że choć czasowe charakterystyki ocenianych prac istotnie wpływają na ich ocenę, to efekt ten jest bliski zeru, w związku z czym jest w praktyce pomijalny³¹. Warto jednak mieć na uwadze, że w przypadku takich analiz znaczenie może mieć specyfika oceniania prac problematycznych. Można sobie wyobrazić, że egzaminator w zetknięciu z problematyczną pod względem oceny pracą przekłada jej ocenianie „na później” i zaczyna ocenianie innej pracy. Prace wymagające konsultacji z przewodniczącym zespołu egzaminacyjnego mogą być zaś oceniane szczególnie długo. Założenie, że egzaminator będzie oceniał jedną pracę po drugiej i nie będzie poświęcał żadnej z nich szczególnie dużo czasu (co będzie generować szereg przypadków odstających) jest więc bardzo mocnym założeniem, w praktyce trudnym do spełnienia.

Uzyskane wyniki są spójne z badaniami wskazującymi na brak zmienności efektu egzaminatora w czasie (Baird , Hayes, R. Johnson, S. Johnson i Lamprianou, 2013). Warto mieć na uwadze, że wpływ zmiennych czasowych mógł się nie ujawnić ze względu na przyjęty schemat badania. W artykułach, w których zidentyfikowano ten wpływ, sesja oceniania trwała zdecydowanie dłużej niż w niniejszym badaniu. Przykładowo wpływ czasu na wzrost surowości egzaminatorów ujawnił się w badaniach Pinot de Moiry i współpracowników. (2002) dopiero pod koniec sesji oceniania, to jest po kilku tygodniach. Podobnie badania Congdon i McQueen (2000) wskazują na ujawnienie się efektu czasu w analizach, jednak ocenianie trwało w nich 7 dni. Jako że niniejsze badanie trwało jedynie dwa dni (podobnie jak ocena prac maturalnych w rzeczywistości), efekty czasu mogły się nie ujawnić.

Ciekawym sposobem analizy zmiennych czasowych mogłoby być modelowanie przy pomocy hierarchicznego modelu latentnych krzywych rozwojowych (ang. *hierarchical latent growth curve modelling*), wykorzystanym także przez Leckiego i Baird (2011).

³¹ Warto zauważyć, że prowadzono także analizy z wykorzystaniem modelu hierarchicznego, w którym oceniane prace były zagnieżdżone w egzaminatorach. Dodatkowo przeprowadzono analizy nie tylko w odniesieniu do sumarycznej oceny pracy, ale także w podziale na poszczególne kryteria oceny bądź zadania. Wyniki tych analiz są zbieżne z opisywanymi w tym podrozdziale – jeśli jakieś efekty uzyskują istotność statystyczną, to są niewielkie.

Model taki pozwalałby na szacowanie indywidualnego poziomu surowości egzaminatorów w czasie. Niestety, w niniejszej analizie nie wyznaczono stabilnych interwałów czasowych (egzaminatorzy nie oceniali prac symultanicznie i w stałej kolejności), względem których oceniano by surowość egzaminatorów. Fluktuacja surowości oceniającego w czasie mogłaby stać się przedmiotem kolejnych analiz.

9.6. Próg zdania matury a rozkład wyników z egzaminu maturalnego

W egzaminie maturalnym, w odróżnieniu od sprawdzianu szóstoklasisty i egzaminu gimnazjalnego, występuje próg zdawalności. Wynosi on minimum 30% punktów możliwych do zdobycia na egzaminie. W wypadku języka polskiego na poziomie podstawowym odpowiada to uzyskaniu 21 punktów, przy czym za część egzaminu składającą się z pytań krótkiej odpowiedzi otrzymać można maksymalnie 20 punktów, a za wypracowanie maksymalnie 50. Z kolei dla matematyki na poziomie podstawowym zaliczenie egzaminu wymaga zdobycia co najmniej 16 punktów, przy czym za zadania zamknięte można było uzyskać w 2011 roku 23 punkty, a w 2012 – 25 punktów. W wypadku języka polskiego zarówno zadania krótkiej odpowiedzi, jak i wypracowanie oceniane są przez tego samego egzaminatora, natomiast w wypadku matematyki zadania zamknięte oceniane są automatycznie, a egzaminator ocenia jedynie zadania otwarte.

Próg zdawalności powodować może różnorake efekty egzaminatora, np. lokalne zwiększenie rzetelności oceny w okolicach progu lub lokalne zwiększenie łagodności oceniania pod progiem zaliczenia. Warto zaznaczyć, że nawet jeśli efekty te mają pozytywny wpływ na oceny uczniów, których prace znajdują się w okolicy progu zdawalności (np. poprzez lokalne zwiększenie rzetelności oceny), ich występowanie jest czynnikiem negatywnym w kontekście egzaminu jako całości, powoduje bowiem, że część uczniów oceniana jest w sposób systematycznie inny od pozostałych. Poza pytaniem o to, czy istnieją efekty egzaminatora związane z progiem zdawalności, równie istotne wydaje się przeanalizowanie, czy mają one zbliżoną siłę w różnych okręgowych komisjach egzaminacyjnych.

Występowanie efektów związanych z progiem zdawalności łatwo jest stwierdzić, patrząc na wykresy rozkładów liczby punktów uzyskanych przez uczniów na egzaminie (rysunki 9.10 i 9.11). Występują one w postaci wyraźnego zmniejszenia odsetka uczniów, którzy uzyskali na egzaminie liczbę punktów minimalnie poniżej progu zaliczenia i wyraźnego zwiększenia odsetka uczniów, którzy uzyskali wynik minimalnie powyżej progu zaliczenia. Efekt ten jest szczególnie widoczny dla języka polskiego, w mniejszym stopniu dla matematyki. Ze względu na to zróżnicowanie omówienie wyników dokonane zostało oddzielnie dla każdego z przedmiotów.

Z uwagi na to, że do opisywanych analiz wystarczają surowe dane egzaminacyjne, zostały one przeprowadzone dla szerszego okresu 2010–2014. Uwzględniono wyniki wszystkich uczniów piszących arkusze standardowe, także tych, którzy do egzaminu maturalnego podchodzili nie po raz pierwszy.

9.6.1. Miary efektów związanych z progiem zaliczenia

O ile łatwo jest wizualnie stwierdzić występowanie zaburzeń rozkładu wyników w bliskości progu zaliczenia, o tyle wprowadzenie ilościowych miar mierzących ich siłę i statystyczną istotność jest znacznie trudniejsze. W wypadku egzaminu z języka polskiego na poziomie podstawowym można próbować porównywać rozkład wyników na egzaminie z rozkładem normalnym o analogicznej średniej i odchyleniu standardowym. Nie jest to jednak możliwe w wypadku matematyki na poziomie podstawowym, gdyż rozkład wyników z tej części egzaminu nie jest zbliżony do żadnego z rozkładów teoretycznych³². W związku z tym zdecydowano się na zastosowanie miar lokalnych, odnoszących się jedynie do kilkupunktowego otoczenia progu zdawalności. W takim niewielkim otoczeniu każdy gładki rozkład powinien dać się dobrze przybliżyć prostą funkcją liniową *odsetek uczniów ~ suma punktów*. Porównując faktycznie obserwowany odsetek uczniów w otoczeniu progu zdawalności w podziale na grupy, które znalazły się pod i nad progiem oraz odnosząc go do analogicznych odsetków obliczonych na podstawie przewidywań z modelu liniowego, można uzyskać ilościową miarę wielkości zaburzenia rozkładu wyników w okolicach progu zdawalności. Jeśli zaburzenia nie występują,

³² W wypadku matematyki mamy do czynienia z niehomogeniczną populacją. Na dobrą sprawę wyróżnić należałoby co najmniej trzy populacje: zdających egzamin po raz kolejny, podchodzących tylko do części podstawowej oraz podchodzących zarówno do części podstawowej jak i rozszerzonej.

ilorazy odsetków będą bliskie 1, im zaś ich wartości dalsze od jedności, tym zakłócenia rozkładu wyników większe. Na wykresach oraz w dalszej części tekstu miary te oznaczone zostały jako:

1. r - iloraz odsetka uczniów w otoczeniu progu zdawalności, którzy nie zdali egzaminu do odpowiadającego mu odsetka przewidywanego z modelu liniowego;
2. $r+$ iloraz odsetka uczniów w otoczeniu progu zdawalności, którzy zdali egzamin do odpowiadającego mu odsetka przewidywanego z modelu liniowego.

Aby stwierdzić, czy obserwowane zaburzenia rozkładu można uznać za istotne statystycznie, posłużono się z kolei modelem liniowym **odsetek uczniów = suma punktów + czy wynik zalicza egzamin**. Za miarę istotności statystycznej zakłócenia rozkładu przyjęto poziom istotności dla współczynnika przy zmiennej **czy wynik zalicza egzamin**. Jest ona podawana na wszystkich zamieszczonych wykresach. Również ten model stosowany był jedynie w lokalnym otoczeniu progu zdawalności.

Wielkość otoczeń progów zdawalności dobrano wizualnie na podstawie wykresów rozkładów wyników egzaminu. W wypadku języka polskiego przyjęto otoczenie ± 4 punkty, natomiast w wypadku matematyki ± 2 punkty od progu zdawalności. Na wykresach przedziały te oznaczone zostały zewnętrznymi przerywanymi pionowymi liniami (wewnętrzna wyznacza próg zdania egzaminu). Aby modele liniowe uwzględniały nie tylko kształt zaburzeń, dopasowywane były w otoczeniach o połowę większych niż te, dla których obliczano wartości współczynników (odpowiednio ± 6 punktów dla języka polskiego i ± 3 punkty dla matematyki).

Ponieważ w wypadku języka polskiego rozkłady wyników są zbliżone do rozkładu normalnego, przy analizach tego przedmiotu zastosowano dodatkowe miary. Ich konstrukcja jest analogiczna do opisanych wyżej współczynników r - i $r+$, z tym że do przewidywania odsetka uczniów w otoczeniu progu zdawalności zastosowano rozkład normalny zamiast modelu liniowego. Przyjęto rozkłady o średniej i odchyleniu standardowym identycznych ze średnią i odchyleniem standardowym wyników obserwowanych. W wypadku analiz z podziałem na OKE, parametry rozkładu normalnego obliczane były indywidualnie dla każdej OKE. Rozpatrywane otoczenia progu zdawalności były zgodne z opisem w poprzednim akapicie. Na wykresach oraz w dalszej części tekstu miary te oznaczone zostały jako:

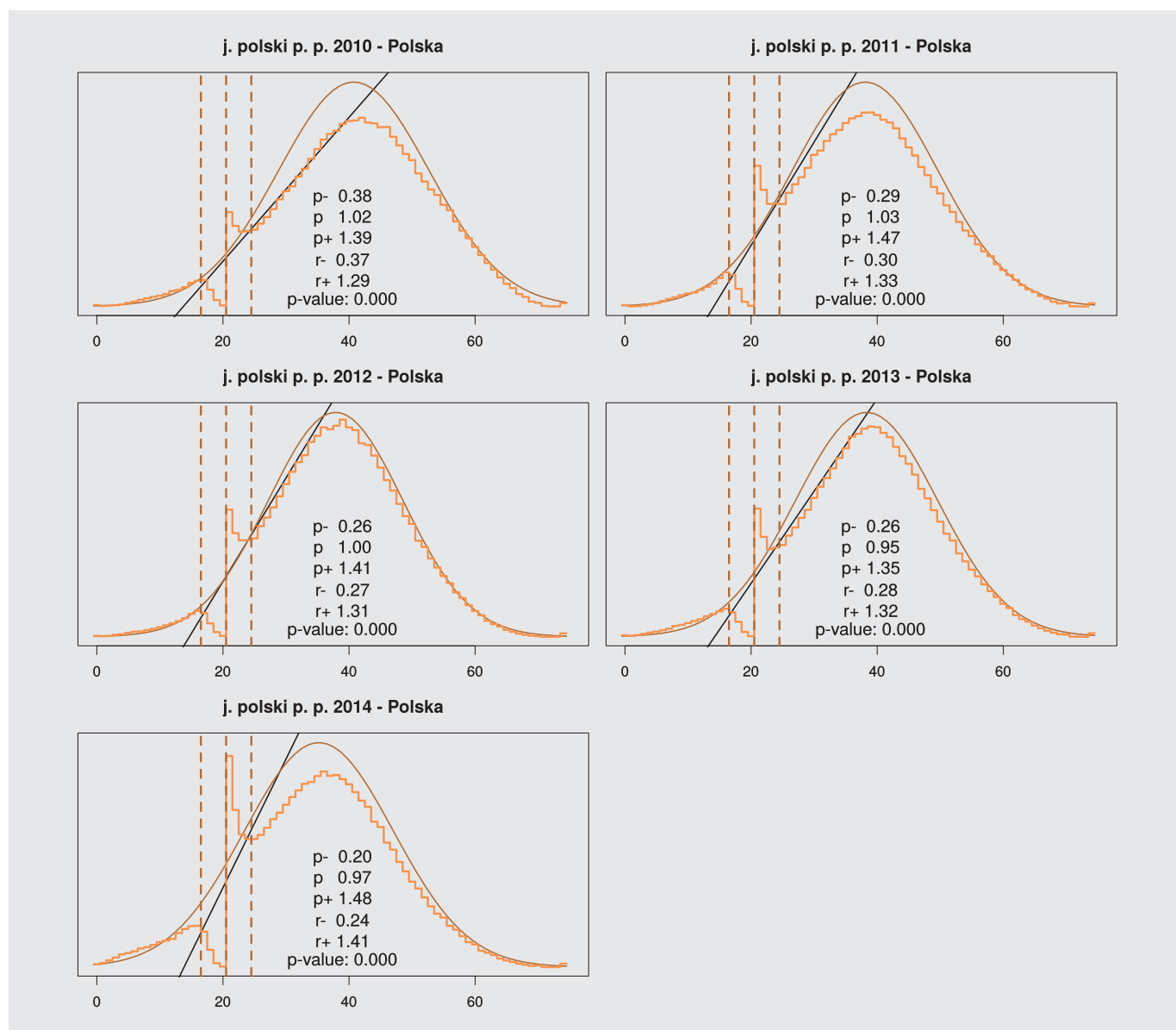
1. p - iloraz odsetka uczniów w otoczeniu progu zdawalności, którzy nie zdali egzaminu do odpowiadającego mu odsetka przewidywanego z rozkładu normalnego;
2. p iloraz odsetka uczniów w otoczeniu progu zdawalności (zarówno tych, którzy zdali, jak i nie zdali) do odpowiadającego mu odsetka przewidywanego z rozkładu normalnego;
3. $p+$ iloraz odsetka uczniów w otoczeniu progu zdawalności, którzy zdali egzamin do odpowiadającego mu odsetka przewidywanego z rozkładu normalnego.

Interpretacja tych wskaźników przebiega analogicznie do r - i $r+$, tzn. wartości bliskie 1 oznaczają brak zniekształcenia rozkładu, a im bardziej różnią się one od 1, tym zakłócenia większe.

9.6.2. Język polski

Zakłócenie rozkładu wyników w okolicy progu zdawalności jest w wypadku egzaminu maturalnego z języka polskiego na poziomie podstawowym bardzo wyraźne (rysunek 9.12). Jest ono tym bardziej widoczne, że ogólny rozkład wyników z pominięciem anomalii w okolicach progu zaliczenia bardzo dobrze przypomina rozkład normalny. Wzrokową analizę kształtu rozkładu potwierdzają współczynniki liczbowe. Poziom istotności współczynnika dla zmiennej **czy wynik zalicza egzamin** jest dla każdego z lat mniejszy od 0,0005, a więc zakłócenie rozkładu jest istotne statystycznie. Podobnie wartości p - oraz $p+$ (lokalne dopasowanie do globalnego rozkładu normalnego) i r - oraz $r+$ (lokalne dopasowanie do lokalnego modelu liniowego) mają wartości dalekie od 1, co sugeruje duże zaburzenie rozkładu (por. tabela 9.19). Przy tym są one najmniejsze dla roku 2010, następnie osiągają nieco większe i bardzo zbliżone wartości w latach 2011–2013, by ponownie wzrosnąć w roku 2014.

Rysunek 9.12. Rozkłady wyników dla egzaminu maturalnego z języka polskiego na poziomie podstawowym w latach 2010–2014



* Zaznaczone otoczenie progu zdawalności, rozkład normalny o jednakowej średniej i odchyleniu standardowym oraz dopasowana lokalnie prosta regresji.

Wartości współczynnika p bliskie 1 sugerują przy tym, że naddatek uczniów nieco powyżej progu zaliczenia pochodzi w całości z ubytku uczniów nieco poniżej progu zaliczenia. Skłania to do interpretacji obserwowanego zjawiska jako **efektu miłosierdzia** polegającego na bardziej łagodnym ocenianiu tych prac, w wypadku których przyjęcie bardziej łagodnych kryteriów oceny zapewnia uczniowi przekroczenie progu zaliczenia egzaminu. Liczbę uczniów, którzy w latach 2010–2013³³ zdali maturę z języka polskiego dzięki występowaniu tego efektu, można szacować jako od 1,4% do 2,3% populacji, co w zależności od roku odpowiada od 5 tys. do 7,8 tys. osób.

33

Rok 2014 pominięto ze względu na gorsze dopasowanie rozkładu normalnego do rozkładu empirycznego.

Tabela 9.19. Wartości wskaźników r^- , r^+ oraz poziomu istotności współczynnika przy zmiennej *czy wynik zalicza egzamin* w modelu liniowym dla egzaminu maturalnego z języka polskiego na poziomie podstawowym. Wyróżniono skrajne wartości współczynnika r^- dla poszczególnych lat

		Gdańsk	Jaworzno	Kraków	Łomża	Łódź	Poznań	Warszawa	Wrocław	Ogółem
2010	p-value	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	r^-	0,26	0,49	0,40	0,30	0,37	0,42	0,37	0,17	0,37
	r^+	1,31	1,22	1,25	1,34	1,31	1,27	1,33	1,33	1,29
2011	p-value	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	r^-	0,23	0,30	0,34	0,28	0,35	0,35	0,33	0,16	0,30
	r^+	1,36	1,31	1,3	1,34	1,31	1,31	1,28	1,46	1,33
2012	p-value	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	r^-	0,21	0,34	0,26	0,28	0,26	0,29	0,34	0,05	0,27
	r^+	1,31	1,30	1,26	1,35	1,36	1,30	1,28	1,38	1,31
2013	p-value	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	r^-	0,21	0,42	0,29	0,28	0,22	0,30	0,36	0,09	0,28
	r^+	1,35	1,27	1,29	1,4	1,34	1,33	1,26	1,42	1,32
2014	p-value	0,001	0,000	0,000	0,000	0,002	0,001	0,000	0,000	0,000
	r^-	0,26	0,44	0,34	0,31	0,31	0,40	0,45	0,16	0,35
	r^+	1,21	1,17	1,2	1,21	1,23	1,2	1,2	1,21	1,2

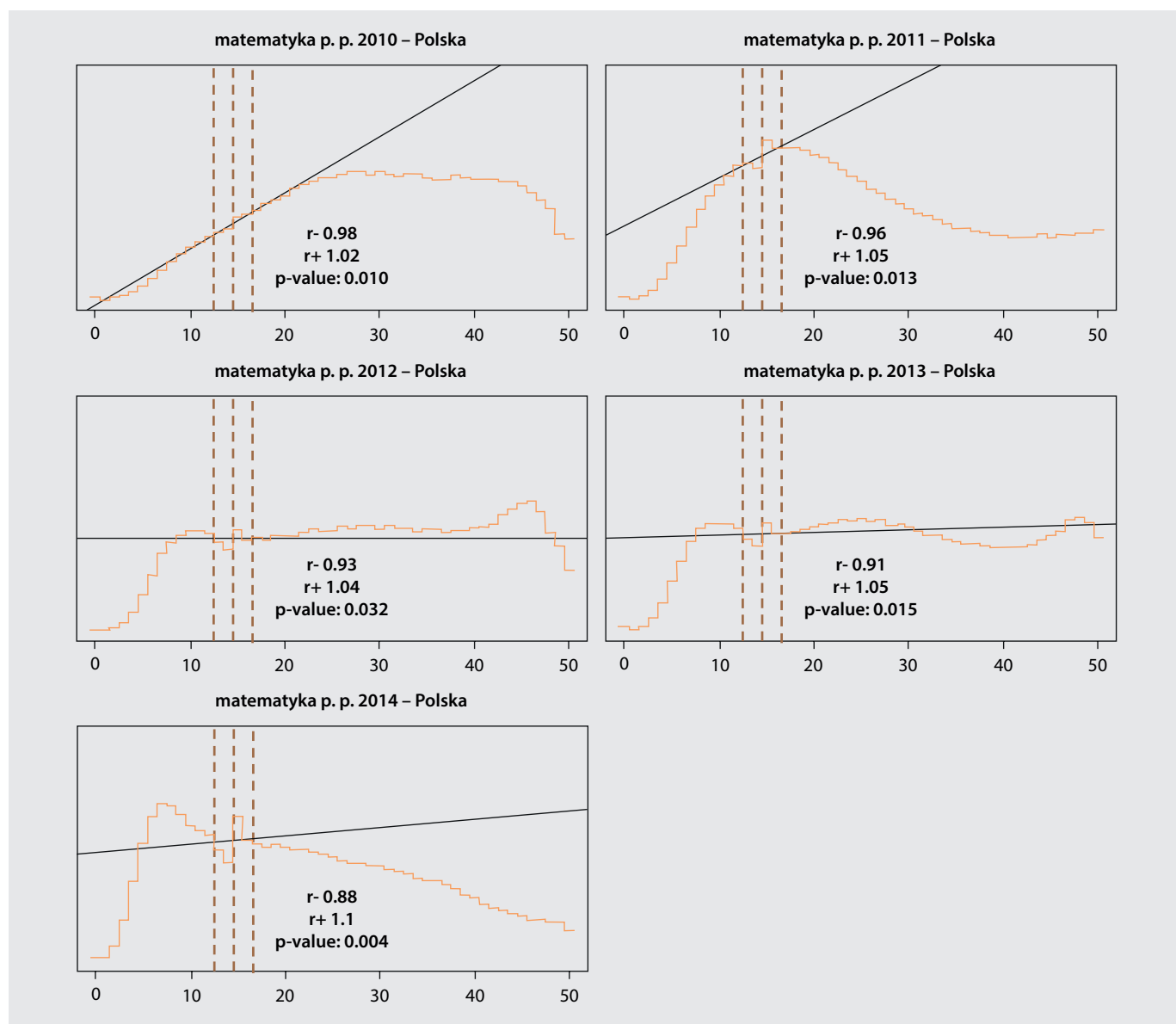
Analiza zjawiska w podziale na OKE pokazuje, że zaburzenie rozkładu w okolicach progu zdawalności istnieje i jest istotne statystycznie we wszystkich OKE. Siła efektu jest w każdym z analizowanych lat największa w OKE Wrocław. Także w OKE Gdańsk w każdym roku zaburzenie rozkładu jest większe niż w wynikach ogólnopolskich. Kolejną grupę stanowią OKE Kraków, Łomża, Łódź, Poznań oraz Warszawa, dla których siła efektu jest zbliżona do wartości ogólnopolskich. Najmniejsze zaburzenia rozkładu notowane są natomiast dla OKE Jaworzno, choć także w wypadku tej komisji należy je uznać za znaczne.

9.6.3. Matematyka

W przypadku egzaminu z matematyki trudniej jest wypowiadać się o zaburzeniach rozkładu wyników w okolicy progu zdawalności ze względu na duże nieregularności w kształcie tych rozkładów. Dla żadnego z analizowanych roczników nie przypominają one rozkładów teoretycznych, a ponadto istotnie różnią się między poszczególnymi latami rysunek 9.13). Z tego powodu wykorzystywać można tylko miary lokalne, np. współczynniki r^+ i r^- czy poziom istotności współczynnika *czy wynik zalicza egzamin* w dopasowywanej lokalnie regresji liniowej.

Dla każdego z analizowanych lat zaburzenie rozkładu w okolicach progu zdawalności okazało się istotne statystycznie przy poziomie istotności 5% (tabela 9.20.). Siła obserwowanego efektu zwiększa się przy tym w czasie. W roku 2010 zaburzenie rozkładu można uznać za istotne statystycznie, lecz pomijalne w praktyce. Wartość współczynnika r^- wynosi 0,98, a więc jest bliska jedności. Różnica odsetka uczniów w populacji w otoczeniu progu, którzy nie zdali egzaminu pomiędzy modelem liniowym i danymi faktycznymi, wynosi 0,08%, czyli ok. 250 uczniów w całej Polsce. W kolejnych latach wartość współczynnika r^- systematycznie spada w tempie ok. 0,03/rok, osiągając w roku 2014 wartość 0,88. Wyrażając ten wynik jako różnicę w odsetku uczniów w populacji w otoczeniu progu, którzy nie zdali egzaminu pomiędzy modelem liniowym i danymi faktycznymi, otrzymujemy 0,6% przystępujących do egzaminu, co odpowiada ok. 2,1 tys. uczniów. Jest to liczba cały czas wyraźnie mniejsza niż w wypadku języka polskiego, jednak ponad ośmiokrotnie większa niż w 2010 roku.

Rysunek 9.13. Rozkłady wyników dla egzaminu maturalnego z matematyki na poziomie podstawowym w latach 2010–2014



*Zaznaczone otoczenie progu zdawalności, rozkład normalny o jednakowej średniej i odchyleniu standardowym oraz dopasowana lokalnie prosta regresji.

Zaburzenia rozkładu wokół progu zdawalności różnią się pomiędzy poszczególnymi OKE. Uwagę przyciąga zwłaszcza OKE Kraków, dla której omawiany efekt nie występuje w całym analizowanym okresie. Wskazują na to zarówno wartości poziomu istotności współczynnika dla zmiennej *czy wynik zalicza egzamin*, jak i bardzo bliskie jedności wartości wskaźników r^- i r^+ . W odróżnieniu od języka polskiego, gdzie uporządkowanie OKE ze względu na wielkość zaburzenia było bardzo podobne we wszystkich latach, w wypadku matematyki nie zaobserwowano takiej regularności. Szczególnie ciekawy wydaje się przypadek OKE Poznań, gdzie w latach 2010–2011 zaburzenia nie występowały, natomiast począwszy od roku 2012 ich wielkość mierzona wskaźnikami r^- i r^+ jest systematycznie większa od ogólnopolskiej.

Tabela 9.20. Wartości wskaźników r^- , r^+ oraz poziomu istotności współczynnika przy zmiennej czy wynik zalicza egzamin w modelu liniowym dla egzaminu maturalnego z matematyki na poziomie podstawowym. Wyróżniono skrajne wartości współczynnika r^- dla poszczególnych lat oraz poziomy istotności mniejsze od 0,05

		Gdańsk	Jaworzno	Kraków	Łomża	Łódź	Poznań	Warszawa	Wrocław	Ogółem
2010	p-value	0,334	0,018	0,082	0,011	0,272	0,374	0,009	0,002	0,010
	r^-	0,99	0,97	1,01	0,91	0,97	0,97	0,93	1,02	0,98
	r^+	1,03	1,03	0,99	1,07	1,01	1,00	1,06	0,98	1,02
2011	p-value	0,055	0,065	0,456	0,082	0,015	0,835	0,049	0,018	0,013
	r^-	0,96	0,90	1,00	0,93	0,95	1,02	0,97	0,88	0,96
	r^+	1,08	1,07	1,01	1,03	1,03	1,01	1,04	1,14	1,05
2012	p-value	0,005	0,082	0,596	0,015	0,015	0,006	0,008	0,144	0,032
	r^-	0,91	0,93	0,99	0,91	0,96	0,87	0,94	0,88	0,93
	r^+	1,07	1,04	0,96	1,07	1,05	1,09	1,05	1,03	1,04
2013	p-value	0,008	0,031	0,793	0,093	0,159	0,006	0,051	0,046	0,015
	r^-	0,90	0,93	0,98	0,88	0,93	0,85	0,96	0,83	0,91
	r^+	1,07	1,06	1,00	1,05	1,02	1,12	1,02	1,13	1,05
2014	p-value	0,011	0,036	0,995	0,011	0,025	0,005	0,006	0,010	0,004
	r^-	0,84	0,93	1,00	0,76	0,91	0,82	0,85	0,83	0,88
	r^+	1,12	1,07	1,00	1,19	1,07	1,15	1,15	1,13	1,10

9.7. Analiza wyników badań ankietowych

W ramach badania efektu egzaminatora zostały przeprowadzone badania ankietowe. Ankieta była wypełniana przez każdego z ocenających po ocenie wszystkich prac. W ankiecie znalazły się pytania dotyczące cech społeczno-demograficznych: wiek, płeć, wykształcenie, oraz pytania dotyczące doświadczenia w ocenianiu prac egzaminacyjnych.

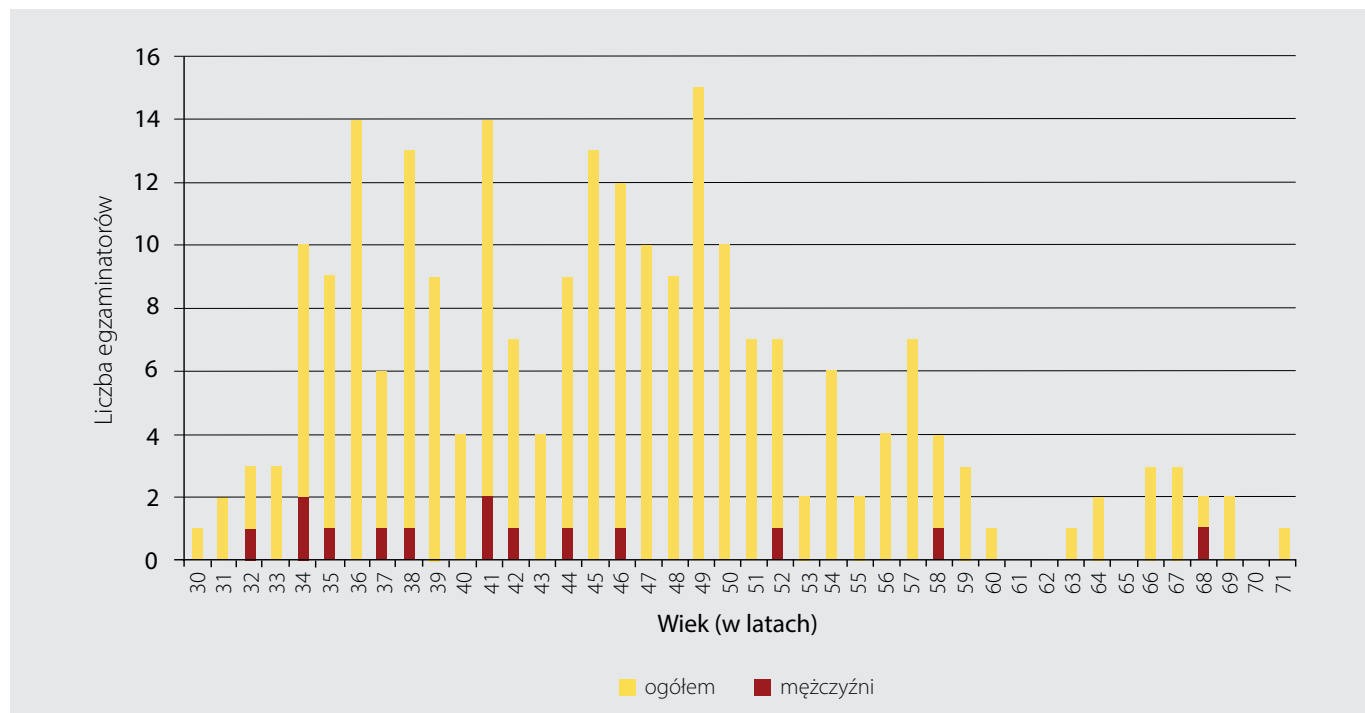
W badaniu ankietowym wykorzystano także dwie skale: mierzącą motywację egzaminatora do oceniania oraz nastawienie egzaminatora do pracy w zespołach podczas oceniania. Przeprowadzono eksploracyjną analizę czynnikową, która wskazała, że struktura skali motywacji jest dwuwymiarowa (6 pytań można zinterpretować jako mierzące motywację wewnętrzną egzaminatora, 1 pytanie jako mierzące motywację zewnętrzną do oceniania – związaną z dodatkowym zarobkiem), natomiast struktura skali mierzącej stosunek egzaminatora do pracy w zespołach jest jednowymiarowa. W związku z tym, że rzetelność jednego pytania jako wskaźnika motywacji zewnętrznej może być niewystarczająca, w analizach posługiwać będziemy się wyłącznie wskaźnikiem motywacji wewnętrznej. Podczas przeprowadzonej analizy czynnikowej zapisano także punkty czynnikowe (ang. *factor scores*) dla poszczególnych skal, szacowane przy pomocy metody Bartletta, pozwalającej na otrzymanie nieobciążonych estymatorów prawdziwych punktów czynnikowych (Hershberger, 2005). Skonstruowano także indeksy – wskaźniki motywacji wewnętrznej egzaminatora oraz pozytywnego stosunku do pracy w zespołach, będące sumą punktów uzyskaną za odpowiedzi w stosownych pytaniach. Kwestionariusz ankiety zamieszczono w aneksie 3.

9.7.1. Język polski

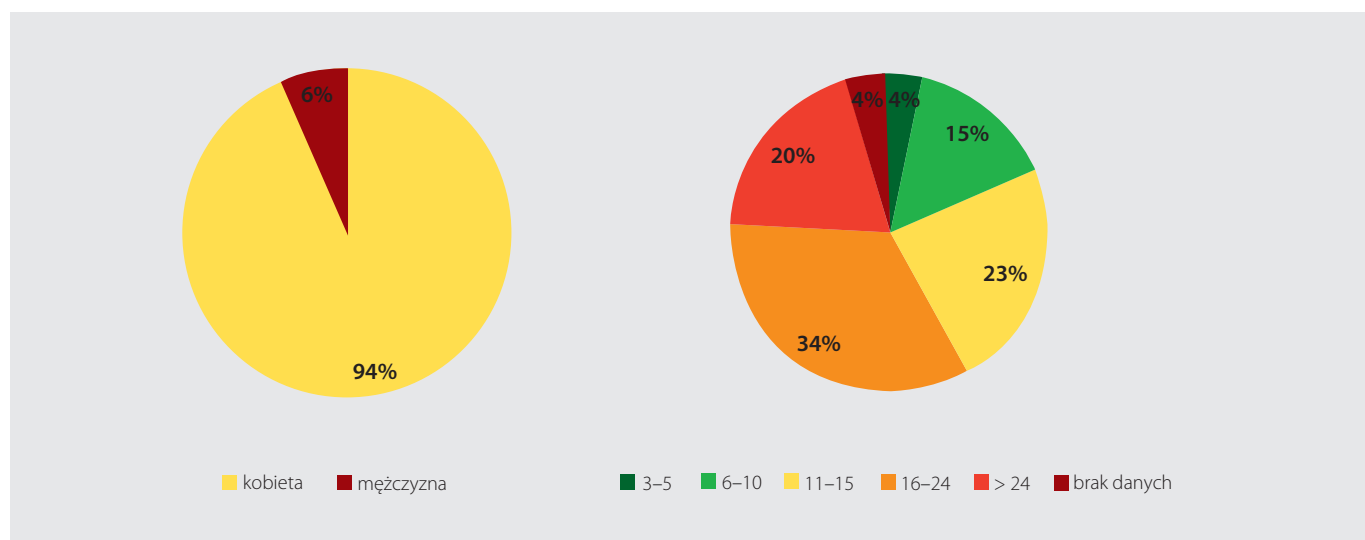
Średnia wieku egzaminatorów ocenających prace maturalne z języka polskiego wyniosła 46 lat, a rozpiętość wieku sięgała od 30 do 71 lat (por. rysunek 9.14). Jedna czwarta egzaminatorów miała powyżej 50 lat. Wśród egzaminatorów z języka polskiego uczestniczących w badaniach znalazło się 217 kobiet i 15 mężczyzn (por. rysunek 9.15). Jest to zawód mocno zfeminizowany. Choć w badaniach brali udział mężczyźni w wieku od 32 do 68 lat, to ich udział w grupie pomiędzy 30. i 50. rokiem życia jest większy niż w grupie powyżej 50 lat. Sugeruje to, że ich liczba w zawodzie może ulegać zwiększeniu. Ponad trzy czwarte osób w próbie to

egzaminatorzy, którzy mają ponad dziesięcioletnie doświadczenie w nauczaniu języka polskiego w szkole średniej (por. rysunek 9.15). Dziesięcioro z badanych egzaminatorów nie podało swojego stażu pracy w szkole kończącej się maturą.

Rysunek 9.14. Wiek (w latach) egzaminatorów z języka polskiego (n = 232)

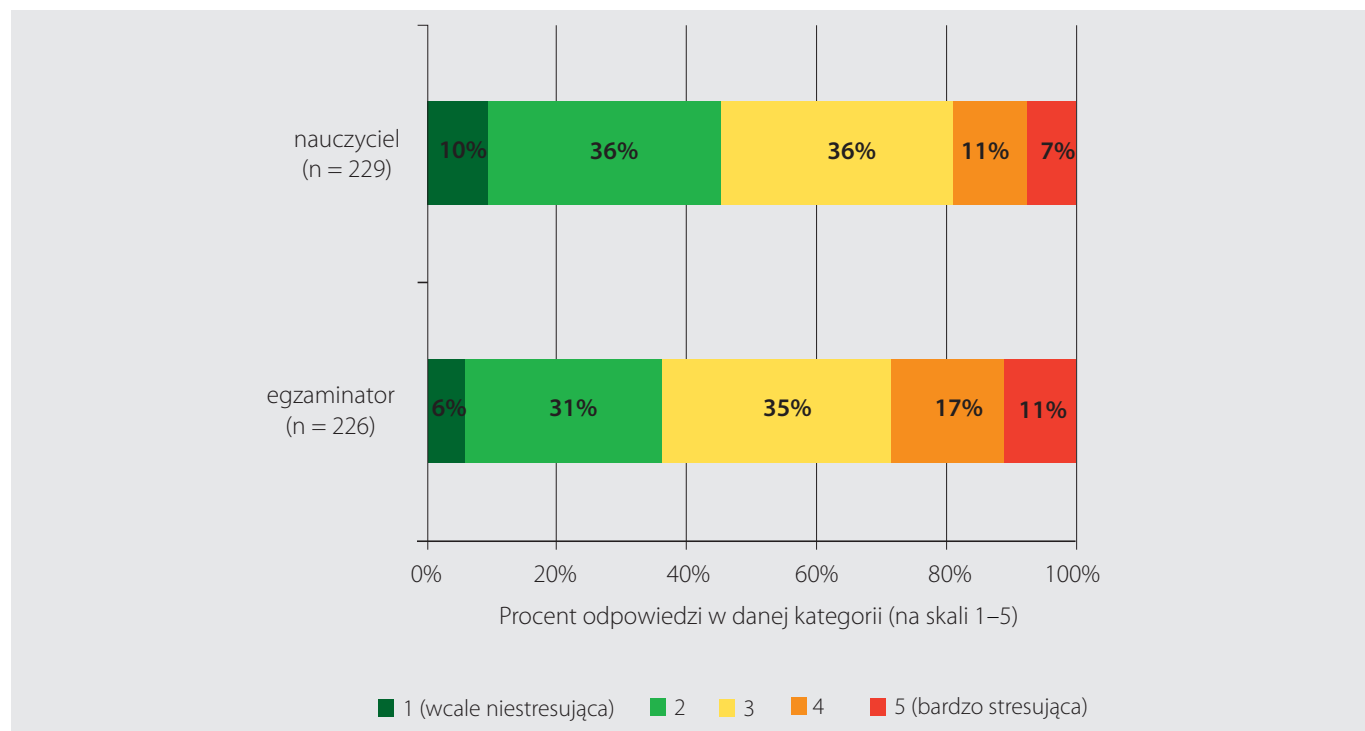


Rysunek 9.15. Płeć (wykres po lewej) i staż pracy w latach (wykres po prawej) egzaminatorów z języka polskiego (n = 232)



Według opinii egzaminatorów praca w charakterze egzaminatora OKE jest bardziej stresująca niż praca w zawodzie nauczyciela (lub innym wykonywanym na co dzień). Aż 29% badanych osób ocenia pracę jako egzaminator OKE jako stresującą lub bardzo stresującą, podczas gdy podobne opinie o pracy jako nauczyciel (lub w innym zawodzie wykonywanym na co dzień) wyraża 19% badanych. Jednocześnie 46% egzaminatorów wskazało, że praca jako nauczyciel jest wcale niestresująca lub niestresująca, a w odniesieniu do pracy jako egzaminator takich odpowiedzi udzieliło 37% badanych. Rozkłady odpowiedzi na pytanie o stres kojarzony z danym zawodem przedstawia rysunek 9.16.

Rysunek 9.16. Opinie egzaminatorów z języka polskiego na temat tego, jak stresująca jest praca w charakterze nauczyciela i egzaminatora



Większość egzaminatorów (72%) oceniających prace maturalne z języka polskiego nie stresuje jednak świadomość, że zdający oglądają swoje rozwiązania zadań w okręgowej komisji egzaminacyjnej po zakończeniu sesji, w trakcie wglądów (zsumowana częstość odpowiedzi 1 i 2 z pięciostopniowej skali). Zaledwie 9% badanych zadeklarowało, że taka świadomość jest dla nich stresująca (zsumowana częstość odpowiedzi 4 i 5 z pięciostopniowej skali). Niemal wszyscy badani egzaminatorzy (96%) zadeklarowali przy tym, że zawsze mają na uwadze to, że od ich rzetelnej pracy jako oceniających zależy przyszłość zdającego. Tylko siedmioro egzaminatorów odpowiedziało na dotyczące tego tematu pytanie, że myślą o tym czasami, nikt nie wskazał, że myśli o tym rzadko, a zaledwie dwoje, że nie zastanawiają się nad wpływem ich oceny na dalsze losy zdającego.

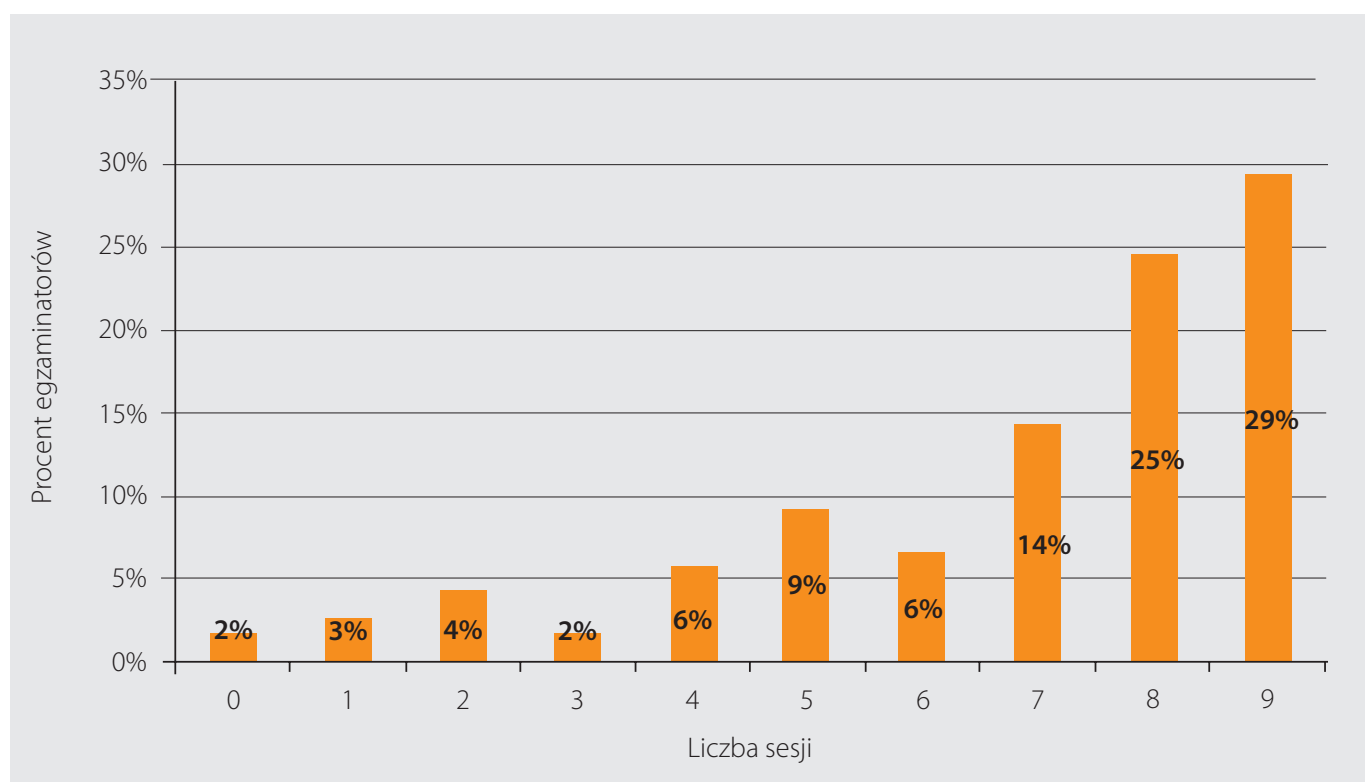
Badani egzaminatorzy w większości deklarują, że nigdy nie podwyższyli punktacji za rozwiązania zadania egzaminacyjnego, mimo, że było ono inaczej sformułowane niż w schemacie oceniania, aby uniknąć ewentualnego odwołania (77%). Tylko 13% badanych przyznało, że taka sytuacja miała miejsce, a 10% egzaminatorów nie odpowiedziało na to pytanie. Dla kontrastu, aż 92% egzaminatorów języka polskiego podwyższało punktację, aby przyznać punkty za rozwiązanie nietypowe, poprawnie merytorycznie, 7% tego nie zrobiło, a 1% nie odpowiedział na to pytanie. Co ciekawe, spośród 30 osób, które przyznały, że podwyższyły ocenę, aby uniknąć odwołania, żadna nie podwyższyła oceny za rozwiązanie nietypowe (być może nigdy się z takim nie spotkały). Tylko szesnaścioro egzaminatorów (7%) wyraźnie zadeklarowało, że nigdy nie podwyższyli punktacji ucznia, niezależnie od podanych powodów. Większość badanych osób (73%) deklaruje, że oceniając jedno zadanie, nie bierze pod uwagę odpowiedzi zdającego na pozostałe zadania, jednak około 25% egzaminatorów języka polskiego przyznaje się do takich praktyk (cztery osoby nie udzieliły na to pytanie odpowiedzi).

W ankiecie znalazło się pytanie o specjalizację w ocenianiu któregoś z poziomów egzaminu maturalnego. Badani mieli wybrać jedną spośród trzech odpowiedzi: specjalizację z ocenie poziomu podstawowego, rozszerzonego lub brak specjalizacji. Pytanie to nie było związane z poziomem kompetencji egzaminatorów, lecz miało na celu sprawdzenie, czy są egzaminatorzy, którzy preferują któryś z poziomów i w związku z tym uczestniczą częściej w jego ocenianiu. Niestety w związku z zaznaczaniem przez egzaminatorów więcej niż jednej odpowiedzi dane zostały zakodowane jako dwie osobne zmienne dla poszczególnych poziomów. Niecała połowa badanych egzaminatorów (48%) wskazała, że specjalizują się oni w ocenie poziomu podstawowego. Zaledwie dziesięcioro badanych egzaminatorów wskazało, że specjalizują się oni w ocenie poziomu rozszerzonego egzaminu maturalnego, przy czym jednocześnie odpowiedzieli oni, że specjalizują się oni również w ocenie poziomu podstawowego. Być może intencją

egzaminatorów, którzy zaznaczyli specjalizację w obydwu poziomach, było zaznaczenie ich wysokich kompetencji. W takim przypadku można uznać, że nie specjalizują się oni w ocenie żadnego z poziomów.

Badanych egzaminatorów spytano również o uczestnictwo w głównych sesjach egzaminacyjnych dla sprawdzianu po szkole podstawowej i egzaminu gimnazjalnego od 2002 roku oraz matury od 2005 roku. Załedwie 6% badanych uczestniczyło w jakiegokolwiek sesji sprawdzianu, podczas gdy w sesji egzaminu gimnazjalnego przynajmniej raz brało udział 23%. Ponad połowa egzaminatorów (54%) brała udział we wszystkich dziewięciu sesjach lub opuściła tylko jedną, przy czym najczęściej opuszczaną sesją była pierwsza sesja z 2005 roku (por. rysunek 9.17). Badanych pytano też, czy podczas sesji maturalnych pełnili funkcję Przewodniczącego Zespołu Egzaminatorów (PZE). Tylko 10% egzaminatorów odpowiedziało, że pełnili oni funkcję PZE – były to zarówno osoby mające za sobą uczestnictwo w wielu sesjach, jak też jednej czy dwóch. Blisko dwie trzecie PZE pełniło tę funkcję 1–3 razy, załedwie dziewięć osób było przewodniczącymi cztery razy lub więcej.

Rysunek 9.17. Liczba majowych sesji maturalnych, w których brali udział egzaminatorzy z języka polskiego (procent)



Poniższa tabela (tabela 9.21) zawiera wyniki testów t dla zróżnicowania poziomu motywacji wewnętrznej egzaminatorów oraz ich nastawienia do pracy w zespołach względem indywidualnych charakterystyk egzaminatora. Motywacja egzaminatorów prac z języka polskiego wiąże się istotnie z braniem pod uwagę w ocenie danego zadania pozostałych zadań. Egzaminatorzy, którzy w taki sposób oceniają konkretną pracę, charakteryzują się istotnie niższym poziomem motywacji wewnętrznej niż egzaminatorzy, którzy oceniając dane zadanie, koncentrują się wyłącznie na nim. Reszta charakterystyk egzaminatora, tj. płeć, skłonność do podwyższania punktacji czy specjalizacja, nie różnicują w sposób istotny statystycznie motywacji wewnętrznej. W przypadku nastawienia do pracy w zespołach znaczenie ma wyłącznie skłonność do podwyższania punktacji. Egzaminatorzy, którzy są skłonni podwyższać punktację za nietypowe rozwiązanie zadania, reprezentują bardziej pozytywne nastawienie do pracy w zespołach egzaminatorów. Pozostałe zmienne nie wiążą się istotnie z bardziej pozytywnym stosunkiem do pracy w zespole egzaminatorów.

Tabela 9.21. Porównanie średniego poziomu motywacji wewnętrznej oraz pozytywnego nastawienia do pracy w zespołach w podziale na grupy (wyniki testów t)

Charakterystyki egzaminatora		Motywacja wewnętrzna			Nastawienie do pracy w zespołach		
		średnia	błąd std.	istotność	średnia	błąd std.	istotność
płeć	kobieta	22,22	0,12	0,53	25,62	0,23	0,45
	mężczyzna	21,93	0,43		24,93	0,80	
	różnica	0,29			0,69		
podwyższenie punktacji (odwołanie)	tak	22,23	0,33	0,75	26,03	0,57	0,45
	nie	22,12	0,13		25,52	0,26	
	różnica	0,11			0,51		
podwyższenie punktacji (nietypowe rozwiązanie)	tak	22,22	0,12	0,71	25,84	0,20	0,00
	nie	22,06	0,46		21,94	1,67	
	różnica	0,16			3,90		
specjalizacja (podstawowy)	tak	22,27	0,16	0,52	25,19	0,33	0,07
	nie	22,13	0,16		26,00	0,29	
	różnica	0,14			-0,81		
specjalizacja (rozszerzony)	tak	22,21	0,11	0,57	25,56	0,23	0,69
	nie	21,90	0,46		26,00	1,05	
	różnica	0,31			-0,44		
branie pod uwagę innych zadań	tak	21,78	0,27	0,03	25,39	0,63	0,63
	nie	22,33	0,12		25,64	0,22	
	różnica	-0,55			-0,25		

Warto wspomnieć o korelacji między wiekiem egzaminatora a nastawieniem do pracy w zespołach egzaminatorów. Wynosi ona -0,214, co wskazuje na to, że im starszy egzaminator, tym mniej entuzjastycznie podchodzi do pracy w zespole egzaminatorów. Fakt reprezentowania konkretnej okręgowej komisji egzaminacyjnej ma również istotne znaczenie dla nastawienia do pracy zespołowej ($F(7,224) = 5,3$; $p < 0,00001$). Egzaminatorzy ze wszystkich OKE (poza OKE Warszawa) charakteryzują się bardziej pozytywnym nastawieniem do pracy w zespole niż egzaminatorzy OKE Wrocław. Średni poziom nastawienia do pracy zespołowej dla poszczególnych OKE przedstawia poniższa tabela (tabela 9.22)

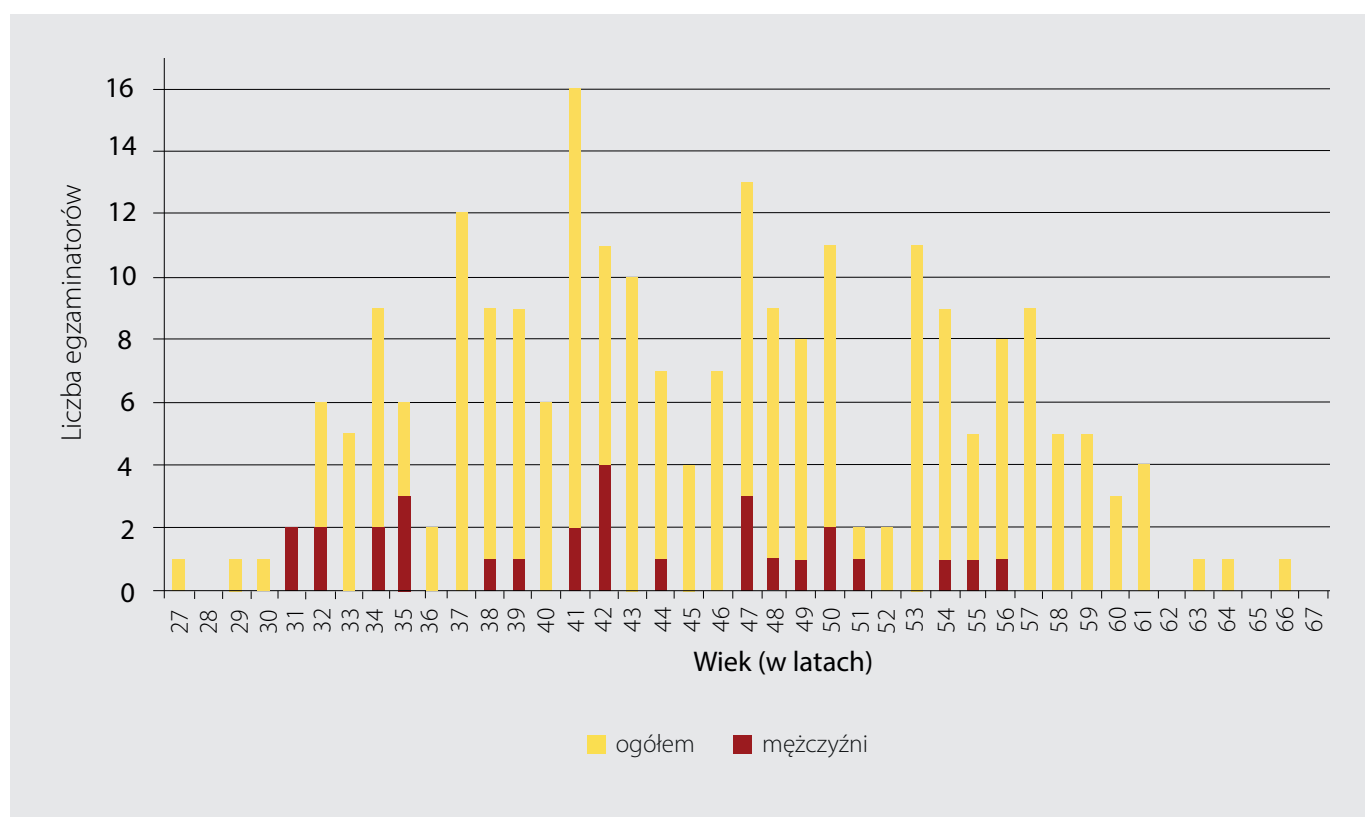
Tabela 9.22. Średni poziom nastawienia do pracy w zespołach w podziale na OKE

Zespół	Średnia	Odchylenie standardowe
OKE Gdańsk	26,93	2,43
OKE Jaworzno	25,66	3,09
OKE Kraków	26,31	2,63
OKE Łomża	25,90	2,54
OKE Łódź	25,55	3,02
OKE Poznań	26,86	2,66
OKE Warszawa	24,72	5,38
OKE Wrocław	22,69	2,93

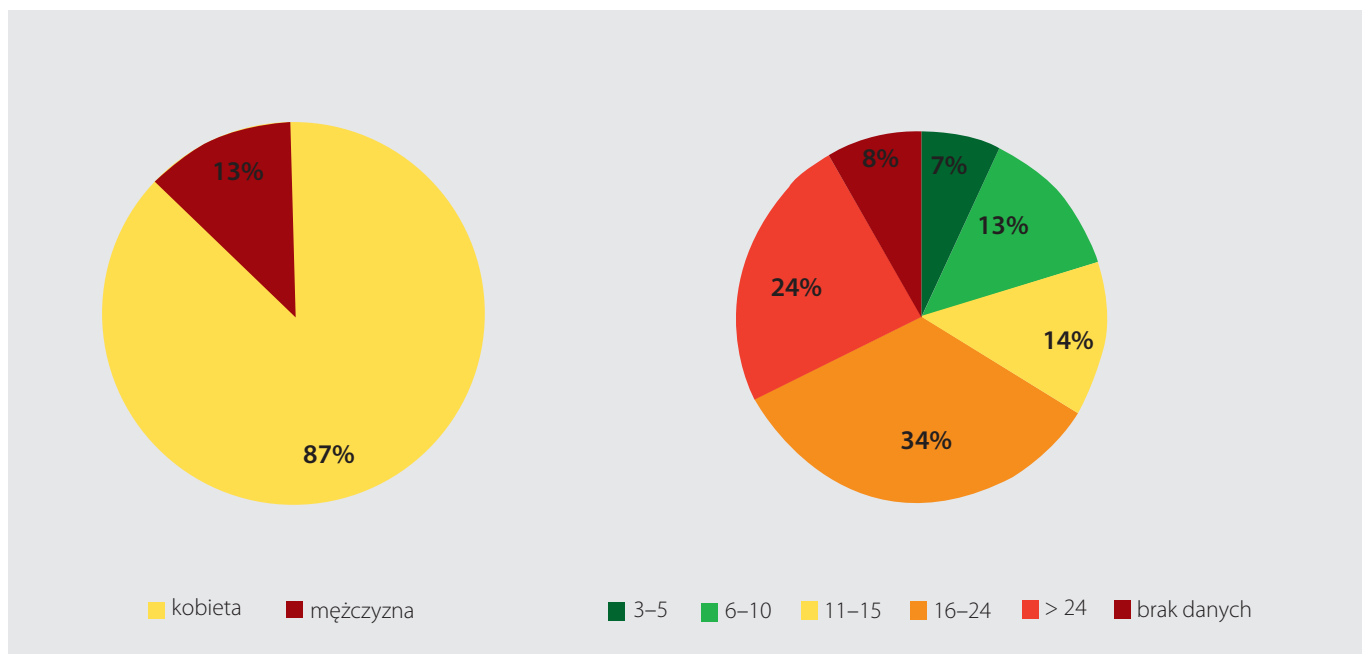
9.7.2. Matematyka

Średnia wieku egzaminatorów oceniających prace maturalne z matematyki wyniosła 46 lat, a rozpiętość wieku sięgała od 27 do 70 lat (por. rysunek 9.18). Jedna czwarta egzaminatorów miała powyżej 53 lat. Wśród uczestniczących w badaniach egzaminatorów z matematyki znalazły się 203 kobiety i 29 mężczyzn. Pomimo prawie dwukrotnie większej liczby mężczyzn niż w przypadku języka polskiego, w zawodzie egzaminatora z matematyki również mamy do czynienia ze znaczną przewagą kobiet – 87% (por. rysunek 9.19). Proporcja mężczyzn i kobiet jest prawie trzykrotnie wyższa w grupie do 50 lat niż w grupie powyżej 50 lat, co może sugerować szybszy niż w przypadku języka polskiego przyrost liczby mężczyzn w zawodzie egzaminatora z matematyki. Nieco mniej niż trzy czwarte egzaminatorów z matematyki ma ponad dziesięcioletni staż w nauczaniu w szkole kończącej się maturą (por. Rysunek 9.19). Dziewiętnastu badanych (8%) nie podało swojego stażu pracy w szkole średniej, przy czym siedmioro z nich wskazało, że uczy w innych typach szkół – szkołach podstawowych, gimnazjach, szkole wyższej.

Rysunek 9.18. Wiek (w latach) egzaminatorów z matematyki (n = 232)

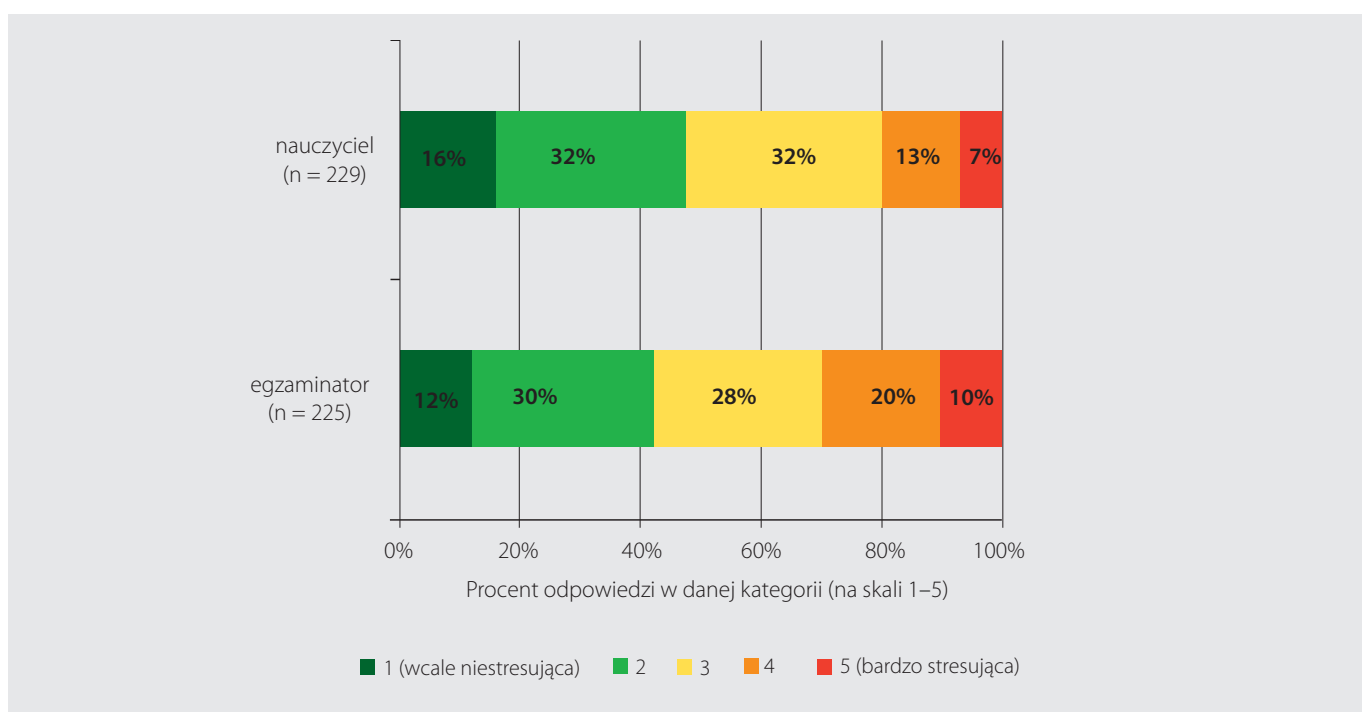


Rysunek 9.19. Płeć (wykres po lewej) i staż pracy w latach (wykres po prawej) egzaminatorów z matematyki (n = 232)



Podobnie jak w przypadku egzaminatorów z języka polskiego, praca w charakterze egzaminatora z matematyki jest dla badanych bardziej stresująca niż praca nauczyciela (lub w innym zawodzie wykonywanym na co dzień). Prawie jedna trzecia (30%) egzaminatorów przyznała, że praca w charakterze egzaminatora OKE jest stresująca lub bardzo stresująca, tymczasem praca w charakterze nauczyciela jest stresująca dla jednej piątej (20%) z badanych. Egzaminatorzy z matematyki częściej niż egzaminatorzy z języka polskiego wybierali skrajną odpowiedź świadczącą, że praca nie jest dla nich wcale stresująca – niezależnie od tego, czy pytani byli o pracę nauczyciela czy egzaminatora. Według deklaracji matematyków praca egzaminatora jest dla nich nieco mniej stresująca, niż deklarowali poloniści (42% vs 37% odpowiedzi 1 i 2 na pięciostopniowej skali). Rozkłady odpowiedzi na pytanie o stres kojarzony z danym zawodem przedstawia rysunek 9.20.

Rysunek 9.20. Opinie egzaminatorów z matematyki na temat tego, jak stresująca jest praca w charakterze nauczyciela i egzaminatora



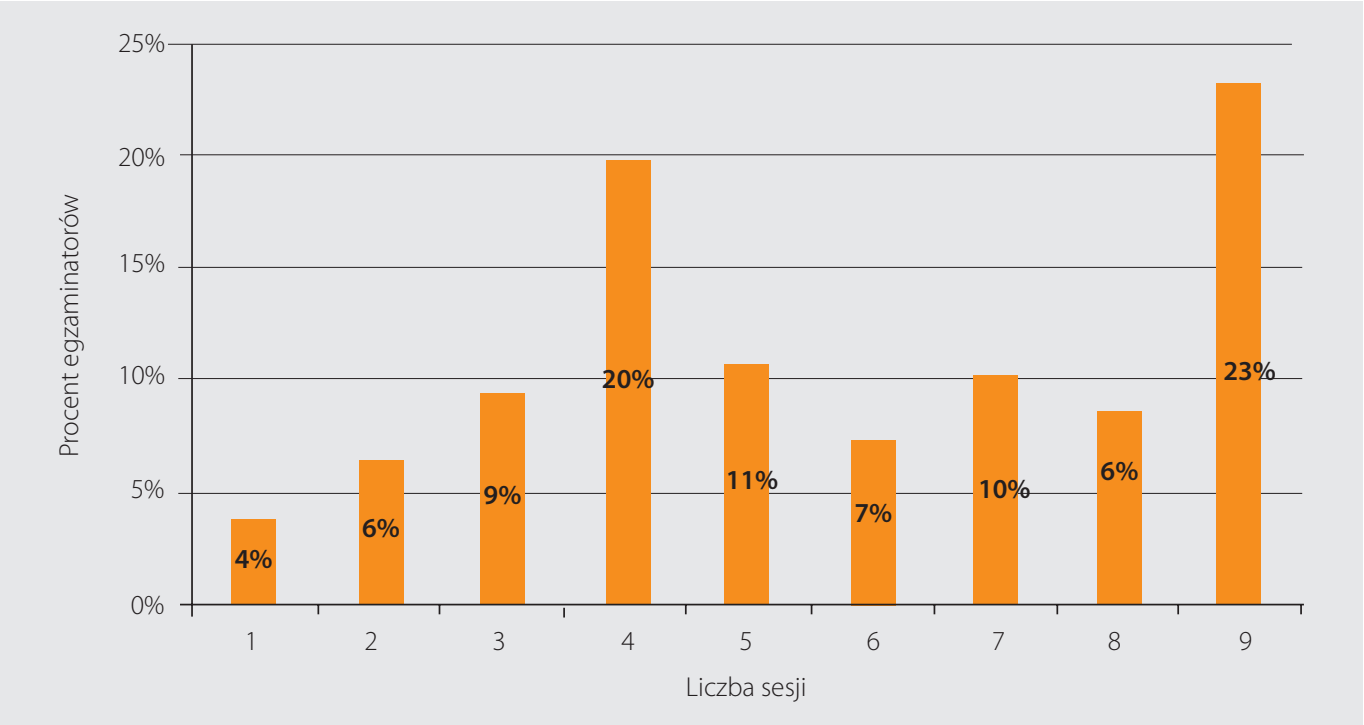
Większość egzaminatorów (76%) nie stresuje się faktem, że uczniowie mają możliwość wglądu w swoje prace (zsumowana częstość odpowiedzi 1 i 2 z pięciostopniowej skali). Zaledwie 8% badanych zadeklarowało, że taka świadomość jest dla nich stresująca (zsumowana częstość odpowiedzi 4 i 5 z pięciostopniowej skali). Podobnie jak w przypadku języka polskiego, prawie wszyscy egzaminatorzy z matematyki (96%) deklarują, że zawsze mają na uwadze, to że od ich rzetelnej pracy zależy przyszłość zdającego. Tylko siedmioro egzaminatorów odpowiedziało na dotyczące tego tematu pytanie, że myślał o tym czasami, nikt nie wskazał, że myśli o tym rzadko, a zaledwie troje, że nie zastanawiają się nad wpływem ich oceny na dalsze losy zdającego.

Zdecydowana większość egzaminatorów z matematyki (86%) deklaruje, że nigdy nie podwyższyła punktacji za rozwiązanie zadania egzaminacyjnego, mimo że było ono inaczej sformułowane niż w schemacie oceniania, aby uniknąć ewentualnego odwołania. Tylko 6% badanych przyznało, że taka sytuacja miała miejsce, a 8% nie odpowiedziało na to pytanie. Z kolei 75% egzaminatorów deklaruje, że podwyższyło punktację za rozwiązanie nietypowe, poprawne merytorycznie, 23% nie zrobiło tego, a pięciu badanych nie udzieliło odpowiedzi na takie pytanie. Osoby, które nie podwyższyły oceny za nietypowe rozwiązanie, nie uczyniły tego także w celu uniknięcia ewentualnego odwołania. W porównaniu do egzaminatorów z języka polskiego, matematycy generalnie rzadziej podwyższają oceny za rozwiązania sformułowane inaczej niż w schemacie oceniania, niezależnie od podanych powodów. Prawdopodobnie jest to związane z większą precyzją schematów oceniania z matematyki i mniejszą możliwością jego subiektywnej interpretacji niż w przypadku języka polskiego. Nieznana jest również skala występowania rozwiązań nietypowych. Prawie wszyscy egzaminatorzy z matematyki (96%) deklarują, że oceniając jedno zadanie, nie biorą pod uwagę odpowiedzi zdającego na pozostałe zadania. Jest to znaczna różnica w porównaniu z egzaminatorami z języka polskiego, lecz ponownie może być to uwarunkowane specyfiką danego egzaminu.

Pytanie związane ze specjalizacją w ocenianiu któregoś z poziomów egzaminu maturalnego z matematyki zostało przez egzaminatorów z matematyki nieco lepiej zrozumiane niż przez polonistów. Tylko cztery osoby zaznaczyły jednocześnie specjalizację w obydwu poziomach egzaminu. Większość badanych specjalizuje się w ocenie poziomu rozszerzonego egzaminu maturalnego (91%) – w ocenie wyłącznie poziomu podstawowego specjalizuje się 30% badanych. Ponad dwie trzecie badanych (69%) odpowiedziało, że specjalizuje się w ocenie poziomu podstawowego egzaminu maturalnego z matematyki, tylko 7% wskazało specjalizację w ocenie poziomu rozszerzonego, a prawie jedna czwarta (24%) zadeklarowała, że nie specjalizują się w żadnym z poziomów.

Niewielka grupa badanych (6%) uczestniczyła w sesjach oceniania sprawdzianu po szkole podstawowej, podobnie jak w przypadku języka polskiego. W sesjach egzaminu gimnazjalnego uczestniczyło już jednak więcej egzaminatorów z matematyki (29%) niż z języka polskiego (23%). Połowa egzaminatorów z matematyki uczestniczyła w sześciu lub więcej głównych sesjach egzaminu maturalnego z matematyki. Największa grupa, stanowiąca prawie jedną czwartą całości (23%), uczestniczyła we wszystkich dziewięciu głównych sesjach maturalnych od 2005 roku. Drugą co do wielkości grupą, stanowiącą jedną piątą wszystkich badanych, byli egzaminatorzy, którzy mieli za sobą udział w czterech głównych sesjach maturalnych, przy czym zdecydowana większość z nich (94%) uczestniczyła w sesjach 2010–2013. Rozkład liczby sesji, w których brali udział badani, przedstawia rysunek 9.21. Funkcję Przewodniczącego Zespołu Egzaminatorów (PZE) podczas sesji maturalnych pełniło zaledwie 9 osób – były to zarówno osoby mające za sobą uczestnictwo w wielu sesjach, jak też jednej czy dwóch. Cztery osoby pełniły funkcję PZE we wszystkich sesjach od 2005 roku.

Rysunek 9.21. Liczba majowych sesji maturalnych, w których brali udział egzaminatorzy z języka polskiego (procent)



W przypadku grupy egzaminatorów oceniających prace z matematyki indywidualne charakterystyki zasadniczo nie różnicują w istotny sposób poziomu motywacji wewnętrznej oraz nastawienia do pracy w zespołach. Jedynym wyjątkiem jest skłonność do podwyższenia punktacji za nietypowe rozwiązanie zadania – egzaminatorzy, którzy się nią charakteryzują, prezentują mniej pozytywne nastawienie do pracy w zespołach w porównaniu do egzaminatorów nieskłonnych podwyższać punktacji. Szczegółowe wyniki testów t prezentuje poniższa tabela (tabela 9.23)

Tabela 9.23. Porównanie średniego poziomu motywacji wewnętrznej i nastawienia do pracy w zespołach w podziale na grupy (wyniki testów t)

Charakterystyki egzaminatora		Motywacja wewnętrzna			Ocena pracy w zespołach		
		średnia	błąd std.	istotność	średnia	błąd std.	istotność
płeć	kobieta	22,48	0,11	0,22	28,41	0,16	0,08
	mężczyzna	22,07	0,37		27,62	0,43	
	różnica	0,41			0,79		
podwyższenie punktacji (odwołanie)	tak	21,93	0,44	0,21	28,13	0,71	0,74
	nie	22,50	0,12		28,34	0,16	
	różnica	-0,56			-0,20		
podwyższenie punktacji (nietypowe rozwiązanie)	tak	22,45	0,12	0,98	28,02	0,17	0,00
	nie	22,44	0,26		29,26	0,27	
	różnica	0,01			-1,24		
specjalizacja (podstawowy)	tak	22,57	0,19	0,43	28,62	0,27	0,18
	nie	22,37	0,13		28,18	0,18	
	różnica	0,19			0,44		

specjalizacja (rozszerzony)	tak	22,40	0,12	0,35	28,23	0,16	0,06
	nie	22,76	0,32		29,19	0,40	
	różnica	-0,36			-0,96		
branie pod uwagę innych zadań	tak	23,14	0,26	0,26	26,86	0,46	0,08
	nie	22,41	0,11		28,37	0,15	
	różnica	0,73			-1,52		

Wśród egzaminatorów prac z matematyki występuje pozytywna, niewielka korelacja (0,169) wieku ze stosunkiem do pracy w zespołach – starsi egzaminatorzy mają nieco bardziej pozytywne nastawienie do pracy zespołowej. Im większy stres przeżywany przez daną osobę w roli egzaminatora, tym bardziej prowadzi to do istotnego spadku motywacji wewnętrznej. Jeśli chodzi o analizę poziomu motywacji wewnętrznej i w poszczególnych OKE, to analiza wariancji wykazała, że OKE różnią się istotnie pod względem tych charakterystyk egzaminatorów ($F(7,224) = 2,25$; $p = 0,0313$). Jednak po wprowadzeniu do analiz poprawki Bonferroniego różnice te znikają – są prawdopodobnie efektem dokonywania wielokrotnych porównań. Praca w konkretnej OKE ma jednak istotne statystycznie znaczenie ($F(7,224) = 5,09$; $p < 0,00001$) dla stosunku do pracy w zespołach – egzaminatorzy OKE w Gdańsku, Krakowie, Łomży i Łodzi mają istotnie wyższy stosunek do oceniania w zespołach niż egzaminatorzy OKE w Poznaniu. Średni poziom nastawienia do pracy zespołowej w podziale na OKE prezentuje poniższa tabela (tabela 9.24).

Tabela 9.24. Nastawienie do pracy w zespołach przez egzaminatorów w podziale na OKE

Zespół	Średnia	Odchylenie standardowe
OKE Gdańsk	29,28	1,33
OKE Jaworzno	27,86	2,49
OKE Kraków	29,03	2,28
OKE Łomża	28,79	2,16
OKE Łódź	28,97	1,55
OKE Poznań	26,59	2,40
OKE Warszawa	27,79	2,53
OKE Wrocław	28,21	1,99

Przeprowadzone analizy dla motywacji wewnętrznej oraz nastawienia do oceniania w zespołach egzaminatorów z języka polskiego i matematyki wskazują na pewne istotne różnice względem takich charakterystyk jak skłonność do podwyższania punktacji, ocenianie zadania w kontekście pozostałych zadań czy fakt przynależności do konkretnej OKE, ale różnice te nie są duże, a określenie ich praktycznego znaczenia wymagałoby dalszych badań.

9.7.3. Łagodność i precyzja egzaminatorów

Celem badań ankietowych egzaminatorów było sprawdzenie, czy istnieje związek pomiędzy cechami społeczno-demograficznymi egzaminatorów, ich doświadczeniem oraz stosunkiem do systemu egzaminacyjnego a ich łagodnością i precyzją oceniania. Poznanie wszelkich związków tego typu jest bardzo istotne, gdyż w procesie szkolenia egzaminatorów można położyć nacisk na zminimalizowanie różnic w ocenianiu pomiędzy egzaminatorami.

Dostępne dane wstępnie analizowano za pomocą modeli regresji liniowej, gdzie zmienną zależną była średnia łagodność lub precyzja egzaminatora niezależnie od poziomu egzaminu z matematyki czy tematu pracy z języka polskiego. Jako zmienne niezależne w modelach testowano osobno poszczególne cechy egzaminatorów. Takie eksploracyjne podejście miało na celu wytypowanie zmiennych mających związek z poziomem łagodności i precyzji egzaminatorów w celu dalszej, pogłębionej analizy.

Niewiele ze zmiennych uzyskanych przy pomocy zastosowanego kwestionariusza okazało się mieć istotny statystycznie związek ze średnimi parametrami egzaminatorów. Poniżej zostały one przedstawione w podziale na przedmioty. Analiza na poziomie poszczególnych zadań wskazuje na istnienie związków pomiędzy prawie wszystkimi mierzonymi cechami a łagodnością i precyzją ocenających, lecz związki te są specyficzne dla poszczególnych zadań i nie pozwalają na uogólnienie. Związki te wynikają raczej z konstrukcji poszczególnych zadań czy sformułowania tematów prac z języka polskiego wraz ze schematami oceny.

Jedynie w przypadku łagodności egzaminatorów z języka polskiego istnieje związek ze skalą nastawienia do pracy w zespołach podczas oceniania (a w zasadzie indeksem obliczonym jako suma odpowiedzi). Pomimo dużej istotności wyników ($p = 0,01$) związek ten jest trudny w interpretacji, gdyż każdy dodatkowy punkt tego indeksu wiąże się z surowością egzaminatora większą o 0,01 punktu. Skala motywacji nie była związana ze średnimi parametrami łagodności i precyzji w przypadku obydwu przedmiotów.

Ogólnie związek mierzonych cech egzaminatorów i ich łagodności i precyzji oceniania wydaje się niewielki. Nawet jeśli pewne związki zostały zaobserwowane, to ich siła jest niewielka. W połączeniu ze stosunkowo niewielkim zróżnicowaniem średnich parametrów efektu egzaminatora (zwłaszcza w przypadku matematyki) może to świadczyć o dość spójnym systemie oceniania. Różnice w ocenianiu na poziomie OKE zostały opisane wcześniej w tym rozdziale.

9.7.3.1. Język polski

Średnia łagodność/surowość egzaminatorów rozciąga się pomiędzy -0,71 a 0,59 punktu ($SD = 0,22$), a ich precyzja od 44% do 65% ($SD = 3\%$). Jest to zróżnicowanie większe niż w przypadku matematyki, przez co można zaobserwować większą liczbę cech wykazujących związki z tymi parametrami.

Deklarujący średni poziom stresu związany z pracą nauczyciela są o 0,09 punktu bardziej surowi niż deklarujący, że praca ta nie jest wcale stresująca ($p = 0,1$). Jednakże deklarujący, że praca jest trochę stresująca (odpowiedź 4 na pięciostopniowej skali) są o 0,12 punktu bardziej surowi niż deklarujący, że nie jest wcale stresująca ($p = 0,05$). Brak jest istotnego związku pomiędzy łagodnością a deklaracjami, że praca nauczyciela jest bardzo stresująca. Trudno jednoznacznie interpretować takie wyniki, lecz wydaje się istnieć związek pomiędzy stresem związanym z pracą nauczyciela a jego łagodnością/surowością. Co ciekawe, w odniesieniu do pracy jako egzaminator nie zachodzą statystycznie istotne związki z łagodnością. Odwrotnie jest w przypadku rzetelności egzaminatorów – istnieje istotny statystycznie związek pomiędzy wskaźnikiem poprawnej klasyfikacji ocen a deklaracjami dotyczącymi stresu w zawodzie egzaminatora, lecz nie w zawodzie nauczyciela. Badani deklarujący, że praca egzaminatora jest bardzo stresująca, poprawnie klasyfikują oceny uczniów w 2% przypadków mniej niż ci, którzy deklarują, że praca egzaminatora wcale nie jest stresująca ($p = 0,1$).

Egzaminatorzy nie zastanawiający się nad tym, że od ich decyzji zależą dalsze losy ucznia, są o 0,3 punktu bardziej surowi niż ci, którzy zawsze o tym myślą ($p = 0,05$). Jednocześnie są oni nieco mniej rzetelni, mając o 6% mniej poprawnych klasyfikacji ocen ($p = 0,01$). Należy jednak pamiętać o tym, że grupa egzaminatorów niezastanawiających się nad konsekwencjami ich oceny dla dalszych losów ucznia jest znikoma.

Istnieje związek pomiędzy podwyższaniem punktacji za zadania nietypowe a precyzją oceniania egzaminatorów. Badani, którzy zadeklarowali, że nie podwyższyli ocen w takiej sytuacji, w 2% przypadków mniej poprawnie klasyfikują oceny w porównaniu do tych, którzy oceny podwyższali ($p = 0,1$). W związku z tym, że przeważającą większość stanowią tacy egzaminatorzy, którzy oceny podwyższali, nie powinno dziwić, że niewielka grupa, która tego nie robiła, odmiennie przydziela oceny.

Egzaminatorzy, którzy zadeklarowali, że specjalizują się w ocenianiu poziomu podstawowego egzaminu maturalnego z języka polskiego, są bardziej surowi o 0,07 punktu w porównaniu do tych, którzy stwierdzili, że nie specjalizują się w ocenie tego poziomu ($p = 0,05$). Jest to o tyle interesujące, że w badaniu oceniane były prace jedynie z poziomu podstawowego. Gdy jednak uznamy, że osoby, które odpowiedziały, że specjalizują się w ocenie obydwu poziomów, w gruncie rzeczy nie specjalizują się w ocenie żadnego z nich, to nie obserwujemy już podobnego związku.

Badani mężczyźni byli bardziej łagodni od kobiet o 0,1 punktu ($p = 0,1$), należy jednak pamiętać o dużej dysproporcji w liczebności dla poszczególnych płci. Nie zaobserwowano związku pomiędzy wiekiem czy doświadczeniem egzaminatorów w nauczaniu oraz uczestnictwem w sesjach egzaminacyjnych a ich średnimi parametrami.

9.7.3.2. Matematyka

W przypadku matematyki wariancja średnich parametrów łagodności i precyzji egzaminatorów jest mniejsza niż dla języka polskiego. Zwłaszcza łagodność egzaminatorów podlega o wiele mniejszym wahaniom niż w języku polskim. Wskaźniki łagodności/surowości rozciągają się pomiędzy -0,04 a 0,12 punktu ($SD = 0,03$), a precyzji w granicach 69-79% ($SD = 2\%$). Różnice te wynikają w dużej mierze ze specyfiki tych dwóch egzaminów – w przypadku języka polskiego ocena treści wypracowania stanowi połowę punktów możliwych do zdobycia z całego arkusza. Przy tak niewielkim zróżnicowaniu tych parametrów trudno wskazać cechy, które na nie wpływają. Zaobserwowano bardzo nieliczne związki pomiędzy poziomem tych parametrów a cechami mierzonymi za pomocą kwestionariusza.

Istnieje związek pomiędzy podwyższaniem punktacji a precyzją oceniania egzaminatorów. Egzaminatorzy, którzy podwyższyli ocenę ucznia, aby uniknąć odwołania, prawie o 1% przypadków częściej poprawnie klasyfikują oceny uczniów niż ci, którzy tego nie robili ($p = 0,1$), a o ponad 1% przypadków częściej niż ci, którzy nie udzielili odpowiedzi na to pytanie. Za ciekawe można uznać, że w przypadku podwyższania oceny za nietypowe rozwiązanie to osoby, które nie udzieliły odpowiedzi, o ponad 1% przypadków częściej poprawnie klasyfikują oceny uczniów niż ci, którzy ocenę podwyższyli ($p = 0,1$). Podobnie jak w przypadku języka polskiego, należy pamiętać o dużych dysproporcjach w liczebności pomiędzy grupami, co czyni ten związek wątpliwym. Podobnie bardzo mała liczba osób, które zadeklarowały, że oceniając jedno zadanie, biorą pod uwagę oceny za pozostałe zadania, czyni wątpliwym to, że osoby, które tego nie robią, o 1% przypadków częściej poprawnie klasyfikują oceny uczniów ($p = 0,1$).

Osoby, które pełniły rolę PZE, są o 0,02 punktu bardziej surowe niż te, które takiej roli nie pełniły ($p = 0,1$). Jednocześnie minimalnie częściej prawidłowo klasyfikują one oceny uczniów (o ponad 1% przypadków; $p=0,05$). Podobnie jak w powyższych przypadkach, bardzo mała liczba osób, które pełniły funkcję PZE, budzi wątpliwości co do faktycznego istnienia takiego związku, zwłaszcza w kontekście braku związku pomiędzy doświadczeniem (czy to mierzonym w latach pracy, czy jako liczba sesji, w których egzaminator brał udział) a łagodnością i precyzją.

Podobnie jak w przypadku języka polskiego, mężczyźni byli bardziej łagodni od kobiet o 0,1 punktu ($p = 0,1$), jednak znów jest to przypadek dużej dysproporcji w liczebności grup. Pozostałe cechy egzaminatorów, takie jak wiek, doświadczenie, poziom stresu, motywacji czy stosunek do pracy w zespole nie wykazują związku z parametrami efektu egzaminatora.

Bibliografia

- Arnold, V., Legas, J., Obler, S., Pacheco, M. A., Russell, C. i Umbdenstock, L. (1990). *Do students get higher scores on their word-processed papers? A study of bias in scoring handwritten vs. word-processed papers*. Unpublished manuscript, Rio Hondo College, Whiter, CA.
- Baird, J., Hayes, M., Johnson, R., Johnson, S. i Lamprianou, I. (2013). *Marker effects and examination reliability: A comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling*. Report commissioned by the Office of Qualifications and Examinations Regulation. Ofqual/13/5261.
- Bond, T. G i Fox, Ch. M. (2001). *Applying The Rasch Model. Fundamental Measurement in the Human Sciences*. Lawrence Erlbaum Associates, Inc.
- Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. W: M. Milanovic i C. J. Weir (red.), *IELTS collected papers: Research in speaking and writing assessment* (s. 98–141). Cambridge: Cambridge University Press.
- Cahalan-Laitusis, C. (2003). *Accommodations on high stakes writing tests for students with disabilities*. Princeton, NJ: Educational Testing Service.
- CKE (2007). *Informator o egzaminie maturalnym od 2008 roku. Język polski*. Warszawa: CKE. Strona internetowa: http://www.cke.edu.pl/images/stories/Inf_mat_od2008/inf_polski_a.pdf
- CKE (2013a). *Sprawozdanie zespołu wdrożeniowego z przeprowadzenia sesji e-oceniań egzaminu gimnazjalnego z matematyki*. Niepublikowane sprawozdanie CKE. Warszawa: CKE.
- CKE (2013b). *Studium wykonalności projektu „Wdrożenie e-oceniań do polskiego systemu egzaminów zewnętrznych w latach 2014–2020”*. Niepublikowany tekst CKE. Warszawa: CKE.
- CKE i OKE (2013a). *Informator o egzaminie maturalnym z języka polskiego od roku szkolnego 2014/2015*. Warszawa. Strona internetowa: <http://www.cke.edu.pl/index.php/informatory-o-egzaminie-maturalnym-od-roku-szkolnego-2014-2015>
- CKE i OKE (2013b). *Informator o egzaminie maturalnym z matematyki od roku szkolnego 2014/2015*. Warszawa. Strona internetowa: http://www.cke.edu.pl/files/file/Matura-2015/Informatory-2015/Matematyka_p.pdf
- Cochran-Smith, M. (1991). Word processing and writing in elementary classrooms: A critical review of related literature. *Review of Educational Research*, 61, 107–155.
- Colbourn, C. J. i Dinitz, J. H. (2007). *Handbook of Combinatorial Designs* (2nd Edition ed.). Boca Raton: Chapman & Hall/CRC.
- Congdon, P. J. i McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37(2), 163–178.
- DeCarlo, L. T., Kim, Y. i Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, 48(3), 333–356.
- Dolata, R., Putkiewicz, E. i Wiłkomirska, A. (2004). *Reforma egzaminu maturalnego – oceny i rekomendacje*. Warszawa, Wydawnictwo Instytutu Spraw Publicznych.
- Dubiecka, A., Szaleniec, H. i Węziak, D. (2006). Efekt egzaminatora w egzaminach zewnętrznych. W: B. Niemierko i M. K. Szmigiel (red.), *O wyższą jakość egzaminów szkolnych* (s. 98–115). Kraków: GRUPA TOMAMI.
- Glaser, R., Chi, M. T. i Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Glas, C. A. (2012). Generalizability theory and item response theory. W: T. J. Eggen i B. P. Veldkamp (red.), *Psychometrics in practice at RCEC* (s. 1–13). Enschede: Ipskamp Drukkers.
- Hershberger, S. L. (2005). Factor scores. W B. S. Everitt i D. C. Howell (red.) *Encyclopedia of Statistics in Behavioral Science*. (s. 636–644). New York: John Wiley.
- Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, 5(1), 64–86.
- Frąszczak, J. (2013) *Czytelność odręcznego tekstu wypracowań z egzaminu maturalnego z języka polskiego*. Analiza na zamówienie IBE. Maszynopis w archiwum Zespołu Analiz Osiągnięć Uczniów.

- James, H. W. (1927). The effect of handwriting on grading. *English Journal*, 16, 180–250.
- Kang, O. (2008). Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spain Fellow Working Papers in Second or Foreign Language Assessment*, 6, 181–205.
- Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, 26(2), 187–217.
- Klein, J. i Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing Writing*, 10, 134–148.
- Kulon, F. (2014). Modele analizy efektu oceniającego w pomiarze edukacyjnym, *Edukacja*, 3(128), 95–111.
- Kulon, F. i Żółtak, M. (2014). Zróżnicowanie łagodności egzaminatorów między okręgowymi komisjami egzaminacyjnymi. W: B. Niemierko i M. K. Szmigiel (red.), *Diagnozy edukacyjne. Dorobek i nowe zadania* (s. 142–150). Kraków: GRUPA TOMAMI.
- Leckie, G. i Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), 399–418.
- Lottridge, S., Nicewander, A., Schulz, M. i Mitzel, H. (2008). *Comparability of Paper-based and Computer-based Tests: A Review of the Methodology*. Pacific Metrics Corporation.
- Lunz, M. E. i Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions*, 13(4), 425–444.
- Markham, L. R. (1976). Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal*, 13(4), 277–283.
- Mei, W. S. (2010). Investigating raters' use of analytic descriptors in assessing writing. *Reflections on English Language Teaching*, 9(2), 69–104.
- Mogey, N., Paterson, J., Burk, J. i Purcell, M. (2010) Typing compared with handwriting for essay examinations at university: letting the students choose. *ALT-J, Research in Learning Technology*, 18(1), 29–47.
- Patz, R. J. i Junker, B. W. (1999). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24(4), 342–366.
- Patz, R. J., Junker, B. W., Johnson, M. S. i Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27(4), 341–384.
- Pinot de Moira, A., Massey, C., Baird, J. i Morrissey, M. (2002). Marking consistency over time. *Research in Education*, 67, 79–87.
- Popham W. J. (1990). *Modern Educational Measurement. A practitioner's perspective*. Englewood Cliffs, New York, Prentice Hall.
- Powers, D. E., Fowles, M. E., Farnum, M. i Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and wordprocessed essays. *Journal of Educational Measurement*, 31, 220–233.
- Rozporządzenie Ministra Edukacji Narodowej i Sportu z dnia 6 listopada 2001 r. zmieniające rozporządzenie w sprawie warunków i sposobu oceniania, klasyfikowania i promowania uczniów i słuchaczy oraz przeprowadzania egzaminów i sprawdzianów w szkołach publicznych (Dz. U. z 2001r. Nr 128, poz. 1419).
- Rozporządzenie Ministra Edukacji Narodowej z dnia 23 grudnia 2008 r. w sprawie podstawy programowej wychowania przedszkolnego oraz kształcenia ogólnego w poszczególnych typach szkół (Dz. U. z dnia 15 stycznia 2009 r. Nr 4, poz. 17).
- Rozporządzenie Ministra Edukacji Narodowej z dnia 21 kwietnia 2009 r. w sprawie ramowego programu szkolenia kandydatów na egzaminatorów, sposobu prowadzenia ewidencji egzaminatorów oraz trybu wpisywania i skreślenia egzaminatorów z ewidencji (Dz. U. z 2009 r., Nr 70, poz. 600).
- Rozporządzenie Ministra Edukacji Narodowej z dnia 8 czerwca 2009 r. w sprawie warunków wynagradzania egzaminatorów za udział w przeprowadzaniu sprawdzianu i egzaminów oraz nauczycieli akademickich za udział w przeprowadzaniu części ustnej egzaminu maturalnego (Dz. U. z 2009 r., Nr 94, poz. 773).
- Russell, M. i Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. *Teachers College Record*.
- Russell, M. i Tao, W. (2004). Effects of handwriting and computer print on composition scores: A follow up to Powers, Fowles, Farnum & Ramsay. *Practical Assessment Research and Evaluation*, 9(1). Strona internetowa: <http://pareonline.net/getvn.asp?v=9&n=1>

- Saal, F. E., Downey, R. G. i Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413–428.
- Scheerens, J., Glas, C. A. i Thomas, S. M. (2003). *Educational evaluation, assessment and monitoring: a systemic approach*. Lisse: Swets & Zeitlinger.
- Scullen, S. E., Mount, M. K. i Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85(6), 956–970.
- Standardy wymagań będące podstawą przeprowadzania egzaminu w ostatnim roku nauki w gimnazjum. Załącznik do rozporządzenia Ministra Edukacji Narodowej z dnia 28 sierpnia 2007 r. (Dz. U. z dnia 31 sierpnia 2007 r. Nr 157, poz. 1102).
- 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. Strona internetowa: <http://eduentuzjasci.pl/images/stories/publikacje/ibe-raport-analiza-porownawcza-wynikow-egzaminow-zewnetrznych-aneks.pdf>
- Szaleniec, H. i Węziak Białowolska, D. (2008). Czy e-ocenianie może zapewnić większą rzetelność punktowania? W: B. Niemierko i M. K. Szmigiel (red.), *Uczenie się i egzamin w oczach nauczycieli* (s. 44–57). Kraków: GRUPA TOMAMI.
- Szaleniec, H. i Kulon, F. (2014). Wpływ rodzaju pisma (ręczne vs komputerowe) na ocenę wypracowania maturalnego z języka polskiego. W: B. Niemierko i M. K. Szmigiel (red.), *Diagnozy edukacyjne. Dorobek i nowe zadania* (s. 75-88). Kraków: GRUPA TOMAMI.
- Thorndike, R. L. i Hagen E. P. (1977). *Measurement and evaluation in psychology and education*. New York: John Wiley and Sons.
- Vance, R. J., Winne, P. S. i Wright, E. S. (1983). A longitudinal examination of rater and ratee effects in performance ratings. *Personnel Psychology*, 36(3), 609–620.
- Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, 38, 330–336.
- Węziak, D. (2007). *Monitorowanie oceniania w sesji egzaminacyjnej 2007*. Opracowanie w ramach projektu „Monitorowanie efektu egzaminatora oraz przygotowanie wprowadzenia oceniania z wykorzystaniem narzędzi elektronicznych (electronic marking)”. Niepublikowany raport CKE. Warszawa: CKE.
- Węziak-Białowolska, D. (2008). *Raport „Analiza surowości oceniania tradycyjnego w zależności od techniki oceniania”*. Niepublikowany raport CKE. Warszawa: CKE.
- Węziak-Białowolska, D. (2009). *Raport „Analiza surowości oceniania – e-ocenianie. Sprawdzian, egzamin gimnazjalny – część matematyczno-przyrodnicza, egzamin gimnazjalny – część humanistyczna, matura z matematyki na poziomie rozszerzonym”*. Niepublikowany raport CKE. Warszawa: CKE.
- Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, 46, 35–51.
- Wolfe, E. W., Kao, C. W. i Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, 15(4), 465–492.
- Wood, R. (1991). *Assessment and Testing. A Survey of Research*. Cambridge: Cambridge University Press.
- Xi, X. i Mollaun, P. (2009). *How Do Raters from India Perform in Scoring the TOEFL iBT [TM] Speaking Section and What Kind of Training Helps?* TOEFL iBT [TM] Research Report. RR-09-31. Educational Testing Service.

Aneks 1. Parametry statystyczne (KTT i IRT) zadań matury z matematyki i języka polskiego 2011 i 2012 r.

Aneks 1. Parametry statystyczne (KTT i IRT) zadań matury z matematyki i języka polskiego 2011 i 2012 r

Tabela 1. Matura z matematyki 2011 – charakterystyka zadań arkusza na poziomie podstawowym

W opisie zastosowano skrótownice obszarów standardów, INF – wykorzystanie i tworzenie informacji, REP – wykorzystanie i interpretowanie reprezentacji, MOD – modelowanie matematyczne, STR – użycie i tworzenie strategii, ROZ – rozumowanie i argumentacja.

Zadanie	Obszar standardów	Sprawdzana umiejętność	Maks.	Średnia	Łatwość	SD	RSK	Lo	Rit	Hi	Rir	AR
24	REP	Rozwiązywanie nierówności kwadratowej	2	1,38	0,69	0,76	0,38	0,59	0,59	0,60	0,55	0,90
25	ROZ	Uzasadnianie zależności arytmetycznej z zastosowaniem wzorów skróconego mnożenia	2	0,25	0,13	0,6	0,30	0,64	0,64	0,64	0,61	0,90
26	INF	Odczytywanie z wykresu funkcji: zbioru wartości oraz maksymalnego przedziału, w którym funkcja maleje	2	1,02	0,51	0,8	0,40	0,67	0,67	0,67	0,63	0,89
27	MOD	Stosowanie wzorów na n -ty wyraz ciągu arytmetycznego lub wykorzystanie własności trzech kolejnych wyrazów tego ciągu	2	1,02	0,51	0,91	0,46	0,76	0,76	0,76	0,72	0,89
28	STR	Stosowanie prostych związków między funkcjami trygonometrycznymi kąta ostrego	2	0,71	0,35	0,92	0,46	0,74	0,74	0,74	0,71	0,89
29	ROZ	Uzasadnianie, że wskazany kąt jest prosty	2	0,16	0,08	0,51	0,26	0,53	0,54	0,54	0,50	0,90
30	STR	Obliczanie prawdopodobieństwa zdarzenia	2	0,89	0,44	0,87	0,43	0,60	0,60	0,61	0,55	0,90
31	STR	Wyznaczanie współrzędnych punktu styczności prostej z okręgiem	4	0,84	0,21	1,47	0,37	0,76	0,77	0,77	0,71	0,89
32	MOD	Rozwiązywanie zadania umieszczonego w kontekście praktycznym, prowadzącego do równania kwadratowego z jedną niewiadomą	5	1,98	0,40	2,15	0,43	0,80	0,81	0,81	0,73	0,90
33	STR	Wyznaczanie związków miarowych w sześciacie	4	1,07	0,27	1,64	0,41	0,74	0,74	0,74	0,67	0,90

Zadanie	a	b0	b1	b2	b3	b4
24	1,81	-1,28	-0,16			
25	3,39	1,07	1,52			
26	1,85	-0,67	0,60			
27	3,05	-0,27	0,19			
28	2,98	0,30	0,51			
29	3,78	1,48	1,67			
30	1,47	-0,19	0,72			
31	3,43	0,55	0,89	1,01	1,17	
32	2,32	-0,23	0,12	0,34	0,57	0,70
33	2,49	0,57	0,60	0,79	1,09	

Tabela 2. Matura z matematyki 2011 – charakterystyka zadań arkusza na poziomie rozszerzonym

W opisie zastosowano skrócone oznaczenia obszarów standardów, INF – wykorzystanie i tworzenie informacji, REP – wykorzystanie i interpretowanie reprezentacji, MOD – modelowanie matematyczne, STR – użycie i tworzenie strategii, ROZ – rozumowanie i argumentacja.

Zadanie	Obszar standardów	Sprawdzana umiejętność	Maks.	Średnia	Łatwość	SD	RSK	Lo	Rit	Hi	Rir	AR
1	STR	Wykorzystywanie cech podzielności liczb całkowitych	4	1,20	0,30	1,57	0,39	0,69	0,62	0,70	0,62	0,86
2	ROZ	Przekształcanie równoważne wyrażenia wymiernego	4	2,71	0,68	1,53	0,38	0,59	0,60	0,60	0,52	0,87
3	STR	Rozwiązywanie równania kwadratowego z parametrem z zastosowaniem wzorów Viète’a, przeprowadzenie dyskusji i wyciągnięcie wniosków	6	3,07	0,51	2,5	0,42	0,79	0,79	0,79	0,70	0,86
4	STR	Rozwiązywanie równania trygonometrycznego	4	1,76	0,44	1,64	0,41	0,72	0,72	0,73	0,66	0,86

Zadanie	Obszar standardów	Sprawdzana umiejętność	Maks.	Średnia	Łatwość	SD	RSK	Lo	Rit	Hi	Rir	AR
5	STR	Stosowanie własności ciągu geometrycznego, wzorów na n -ty wyraz tego ciągu i na sumę n wyrazów ciągu arytmetycznego	4	1,69	0,42	1,84	0,46	0,70	0,70	0,71	0,62	0,86
6	STR	Znajdowanie związków miarowych w figurach płaskich z zastosowaniem trygonometrii	4	2,29	0,57	1,58	0,4	0,68	0,69	0,69	0,62	0,86
7	STR	Rozwiązywanie zadań dotyczących wzajemnego położenia prostej i okręgu	4	1,61	0,40	1,77	0,44	0,68	0,69	0,69	0,61	0,86
8	MOD	Znajdowanie związków miarowych w graniastopłacie; wyznaczanie największej wartości funkcji	4	2,23	0,56	1,68	0,42	0,69	0,70	0,70	0,62	0,86
9	STR	Wykorzystywanie wzorów na liczbę permutacji, kombinacji i wariacji do zliczania obiektów w sytuacjach kombinatorycznych	4	0,71	0,18	1,11	0,28	0,50	0,50	0,51	0,44	0,87
10	ROZ	Znajdowanie związków miarowych w figurach płaskich	3	0,55	0,18	1,14	0,38	0,54	0,55	0,55	0,48	0,87
11	STR	Znajdowanie związków miarowych w ostrośłupie	6	3,69	0,61	2,52	0,42	0,68	0,68	0,69	0,56	0,87
12	STR	Wykorzystywanie własności prawdopodobieństwa do obliczania prawdopodobieństwa zdarzeń	3	1,37	0,46	0,99	0,33	0,50	0,51	0,51	0,45	0,87
Zadanie	a	b0	b1	b2	b3	b4	b5					
1	2,01	-0,05	0,89	0,96	1,04							
2	1,37	-1,79	-0,94	-0,48	-0,07							
3	2,27	-0,67	-0,30	-0,15	0,06	0,23	0,82					
4	2,01	-0,50	0,10	0,38	0,84							
5	1,95	-0,03	0,20	0,37	0,50							
6	1,68	-1,18	-0,46	-0,07	0,53							

7	1,79	-0,12	0,33	0,52	0,70
8	1,81	-1,00	-0,24	-0,09	0,33
9	1,10	0,33	2,14	2,40	2,82
10	1,87	1,16	1,23	1,27	
11	1,53	-1,12	-0,68	-0,57	-0,34 -0,18 0,30
12	1,14	-2,00	1,09	1,25	

Tabela 3. Matura z matematyki 2012 – charakterystyka zadań arkusza na poziomie podstawowym

W opisie zastosowano skrótownice oznaczenia obszarów standardów, INF – wykorzystanie i tworzenie informacji, REP – wykorzystanie i interpretowanie reprezentacji, MOD – modelowanie matematyczne, STR – użycie i tworzenie strategii, ROZ – rozumowanie i argumentacja.

Zadanie	Obszar standardów	Sprawdzana umiejętność	Maksymalna liczba punktów	średnia	Łatwość	RSK	RIT	RIR	AR
26	REP	Rozwiązanie nierówności kwadratowej	2	1,40	0,70	0,36	0,65	0,61	0,90
27	ROZ	Uzasadnienie prawdziwości nierówności algebraicznej	2	0,39	0,19	0,32	0,62	0,58	0,90
28	REP	Rozwiązanie równania wielomianowego metodą rozkładu na czynniki	2	1,15	0,57	0,48	0,68	0,64	0,90
29	STR	Wykorzystanie własności symetralnej odcinka do wyznaczenia jej równania	2	0,59	0,29	0,44	0,70	0,66	0,90
30	ROZ	Przeprowadzenie dowodu geometrycznego	2	0,21	0,10	0,31	0,46	0,42	0,90
31	MOD	Obliczenie prawdopodobieństwa zdarzenia z zastosowaniem klasycznej definicji prawdopodobieństwa	2	1,02	0,51	0,42	0,60	0,56	0,90
32	MOD	Zastosowanie własności ciągu arytmetycznego i geometrycznego	4	2,82	0,70	0,43	0,74	0,67	0,90
33	STR	Obliczenie objętości wielościanu	4	1,85	0,46	0,43	0,79	0,73	0,90
34	MOD	Rozwiązanie zadania, umieszczonego w kontekście praktycznym, prowadzącego do równania kwadratowego	5	1,91	0,38	0,44	0,79	0,71	0,91

Zadanie	a	b0	b1	b2	b3	b4
26	1,83	-1,53	-0,11			
27	2,41	0,63	1,66			
28	2,18	-0,30	-0,17			
29	3,43	0,51	0,74			
30	2,96	1,44	1,45			
31	1,42	-0,56	0,59			
32	2,52	-0,96	-0,60	-0,55	-0,46	
33	2,50	-0,38	0,15	0,25	0,67	
34	2,78	0,05	0,16	0,32	0,54	0,72

Tabela 4. Matura z matematyki 2012 – charakterystyka zadań arkusza na poziomie rozszerzonym

Zadanie	Obszar standardów	Sprawdzana umiejętność	Maksymalna liczba punktów	Średnia	Łatwość	SD	RSK	RIT	RIR	AR
1	MOD	Rozwiązanie zadania, prowadzącego do równania kwadratowego	4	3,14	0,79	1,51	0,38	0,56	0,47	0,85
2	STR	Rozwiązanie nierówności wielomianowej	4	2,53	0,63	1,73	0,43	0,69	0,61	0,84
3	STR	Rozwiązanie równania trygonometrycznego	4	2,56	0,64	1,5	0,38	0,73	0,66	0,84
4	STR	Rozwiązanie równania kwadratowego z parametrem, przeprowadzenie dyskusji i wyciągnięcie wniosków	6	2,48	0,41	2,24	0,37	0,81	0,73	0,83
5	STR	Zastosowanie własności ciągu geometrycznego oraz własności ciągu arytmetycznego	6	2,89	0,48	2,05	0,34	0,67	0,56	0,84
6	MOD	Znalezienie związków miarowych na płaszczyźnie, wyznaczenie największej i najmniejszej wartości funkcji	6	2,02	0,34	2,3	0,38	0,73	0,62	0,84
7	ROZ	Przeprowadzenie dowodu algebraicznego	3	,89	0,3	1,28	0,43	0,56	0,49	0,85

Zadanie	Obszar standardów	Sprawdzana umiejętność	Maksymalna liczba punktów	Średnia	Łatwość	SD	RSK	RIT	RIR	AR
8	STR	Wykorzystanie wzorów na liczbę permutacji, kombinacji i wariacji do zliczania obiektów w sytuacjach kombinatorycznych	4	1,16	0,29	1,34	0,33	0,58	0,50	0,85
9	STR	Znalezienie związków miarowych w figurach płaskich z zastosowaniem własności figur podobnych	5	2,38	0,48	2,11	0,42	0,71	0,61	0,84
10	STR	Znalezienie związków miarowych w ostrośgłupie	5	3,61	0,72	1,91	0,38	0,53	0,41	0,86
11	ROZ	Wykorzystanie własności prawdopodobieństwa do obliczania prawdopodobieństw zdarzeń	3	,54	0,18	1,03	0,34	0,49	0,42	0,85

Zadanie	a	b0	b1	b2	b3	b4	b5
1	1,50	-1,65	-1,32	-1,18	-0,92		
2	2,01	-1,02	-0,50	-0,35	-0,17		
3	2,28	-1,26	-0,84	-0,35	0,30		
4	2,75	-0,76	-0,19	0,13	0,54	0,76	1,01
5	1,49	-2,01	-1,08	0,40	0,59	0,72	1,15
6	1,84	-0,22	0,11	0,42	0,61	1,03	1,44
7	1,43	0,55	0,61	1,15			
8	1,22	-0,18	0,47	1,51	2,35		
9	1,62	-0,76	-0,19	0,18	0,29	0,68	
10	0,98	-2,32	-1,58	-1,04	-0,88	-0,44	
11	1,31	1,07	1,49	1,94			

Tabela 5. Matura z języka polskiego 2011 – charakterystyka zadań arkusza na poziomie podstawowym

Zadanie	Obszar standardów	Sprawdzana umiejętność	Maksymalna liczba punktów	Średnia	Łatwość	SD	RSK	RIT	RIR	AR					
14	Zadanie sprawdzające tworzenie własnego tekstu														
Treść	Tworzenie informacji	Określanie problemu czytanych tekstów; dostrzeganie cech wspólnych i różnych wskazanych utworów oraz związków utworów z historią i biografią autorów;	25	10,77	0,43	4,79	0,19	0,87	0,67	0,75					
		odczytywanie miejsc znaczących, formułowanie własnej interpretacji utworów; podsumowanie swoich rozważań													
Kompozycja		Pisanie tekstu uporządkowanego, spójnego, zgodnie z zasadami jego organizacji	5	2,73	0,55	1,41	0,28	0,76	0,70	0,69					
Styl		Pisanie stylem komunikatywnym, dostosowanym do formy wypowiedzi	5	2,88	0,58	1,23	0,25	0,71	0,65	0,69					
Język	Tworzenie informacji	Posługiwanie się językiem zgodnym z obowiązującą normą	12	6,65	0,55	2,90	0,24	0,81	0,68	0,67					
Zapis		Stosowanie zasad ortografii i interpunkcji	3	1,58	0,53	0,98	0,33	0,48	0,41	0,72					
Temat I	Treść	a	b0	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11	b12
			-3,01	-2,75	-2,46	-2,19	-1,94	-1,68	-1,43	-1,19	-0,94	-0,71	-0,47	-0,25	-0,03
		b13	b14	b15	b16	b17	b18	b19	b20	b21	b22	b23	b24		
		0,19	0,40	0,61	0,83	1,05	1,27	1,51	1,77	2,04	2,35	2,74	3,28		
	Kompozycja	3,04	-2,21	-0,90	0,67										
	Styl	3,02	-2,39	-1,11	0,87										
	Język	3,06	-2,43	-2,01	-1,04	0,10	1,19								
	Zapis	1,08	-1,98	-0,53	1,33										
	Walory	1,80	1,76	2,73	3,67	4,64									

Temat II	Treść	2,28	-3,30	-2,93	-2,52	-2,12	-1,75	-1,42	-1,10	-0,81	-0,52	-0,25	0,01	0,27	0,53
		b13	b14	b15	b16	b17	b18	b19	b20	b21	b22	b23	b24		
Temat II	Kompozycja	2,58	-2,42	-0,70	1,04										
	Styl	2,62	-2,71	-1,07	1,13										
	Język	2,66	-2,79	-2,19	-1,02	0,28	1,48								
	Zapis	0,98	-1,94	-0,32	1,73										
	Walory	1,72	2,70	3,97	5,04	5,90									

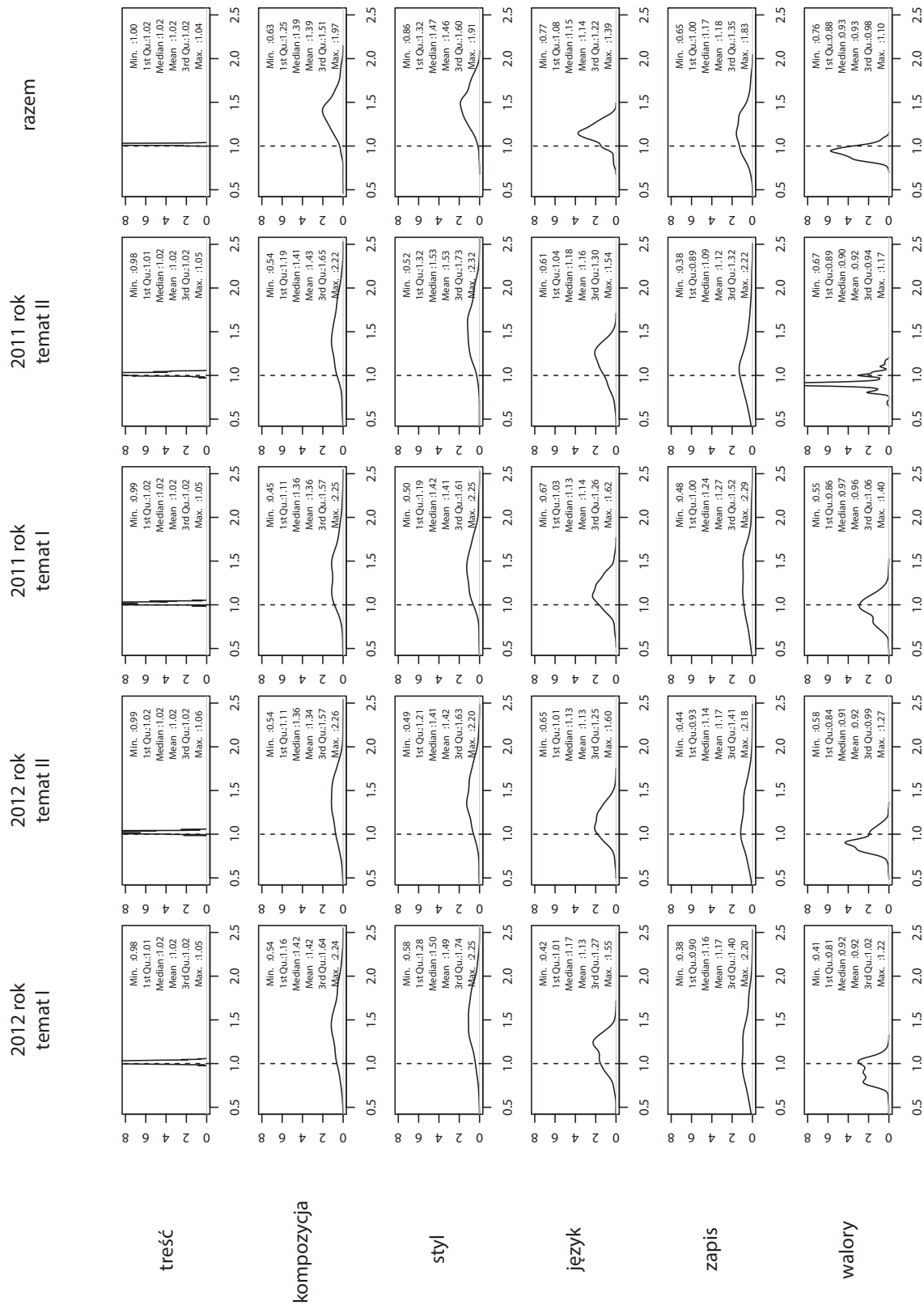
Tabela 6. Matura z języka polskiego 2012 – charakterystyka zadań arkusza na poziomie podstawowym

Zadanie	Obszar standardów	Sprawdzana umiejętność	Maksymalna liczba punktów	Średnia	Łatwość	SD	RSK	RIT	RIR	AR
15		Zadanie sprawdzające tworzenie własnego tekstu								
Treść	Tworzenie informacji	Rozpoznanie problematyki utworu; rozpoznanie bohaterów utworu; scharakteryzowanie bohaterów i ich postaw; odczytanie treści dosłownych i ukrytych utworu; rozpoznanie znaków i symboli kulturowych; rozpoznanie tematów, wątków, motywów; wykorzystanie w odczytaniu sensu utworu miejsc znaczących; dostrzeżenie najistotniejszych zagadnień utworu; problematyzowanie odczytania tekstu; podsumowanie obserwacji analitycznych; wskazanie kontekstów utworu i wykorzystanie ich w interpretacji	25	12,37	0,49	4,65	0,19	0,85	0,62	0,74
Kompozycja		Napisanie tekstu uporządkowanego, spójnego, zgodnie z zasadami jego organizacji	5	2,80	0,56	1,27	0,25	0,71	0,64	0,66
Styl		Pisanie stylem komunikatywnym, dostosowanym do formy wypowiedzi	5	2,76	0,55	1,12	0,22	0,66	0,59	0,67
Język		Posługiwanie się językiem zgodnym z obowiązującą normą	12	6,22	0,52	2,63	0,22	0,77	0,64	0,64
Zapis		Stosowanie zasad ortografii i interpunkcji	3	1,48	0,49	0,99	0,33	0,44	0,36	0,69

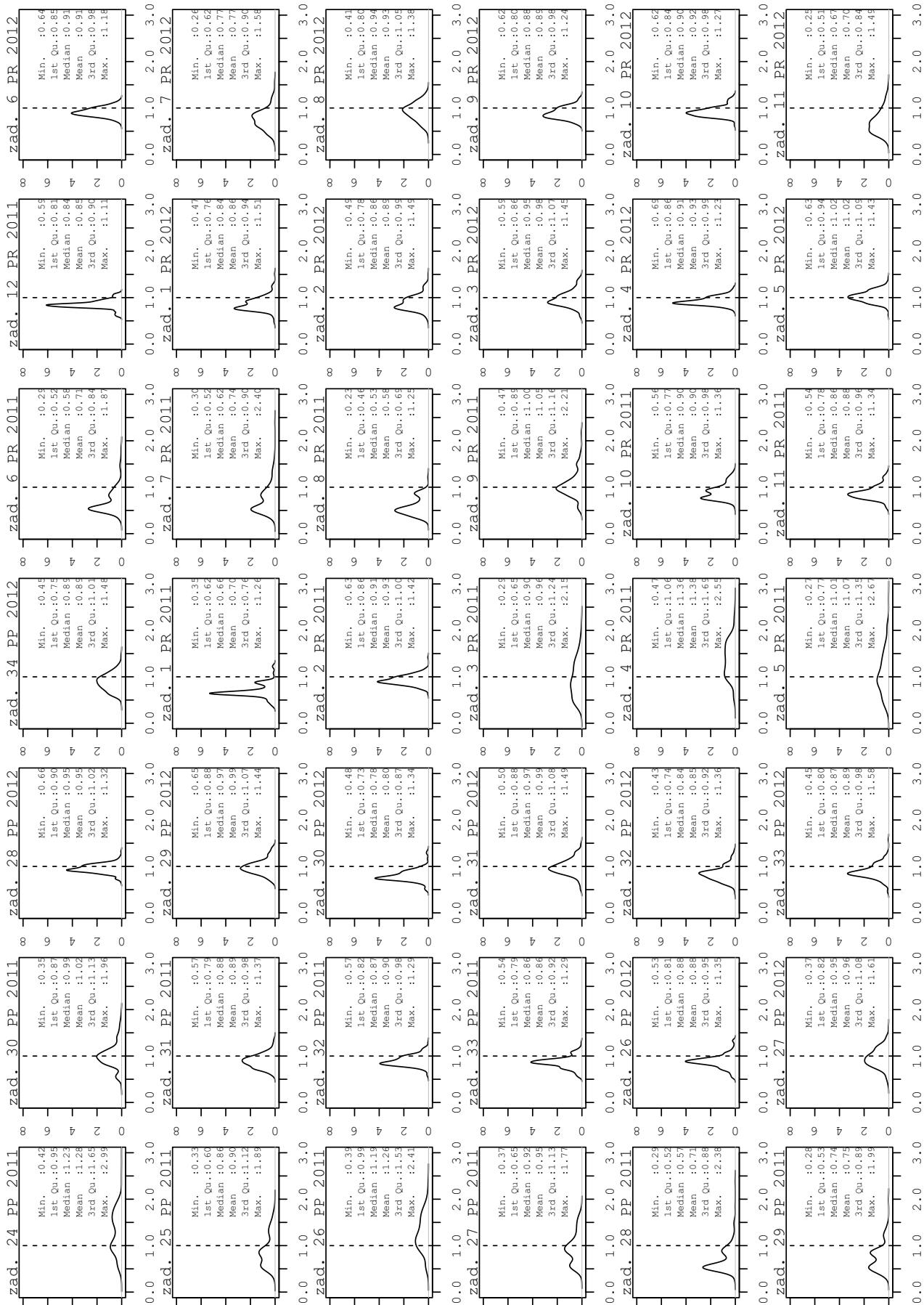
		a	b0	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11	b12
Temat I	Treść	2,05	b13	b14	b15	b16	b17	b18	b19	b20	b21	b22	b23	b24	
			-3,03	-2,74	-2,49	-2,23	-1,97	-1,69	-1,42	-1,14	-0,86	-0,59	-0,33	-0,06	0,21
	Kompozycja	2,75	0,46	0,72	0,99	1,25	1,52	1,82	2,12	2,44	2,77	3,15	3,60	4,20	
	Styl	2,90	-2,56	-1,02	1,16										
Temat II	Język	2,82	-2,64	-2,00	-0,84	0,42	1,65								
	Zapis	0,92	-1,70	0,00	1,85										
	Walory	1,46	2,09	3,30	4,46	5,50									
	Treść	1,77	b13	b14	b15	b16	b17	b18	b19	b20	b21	b22	b23	b24	
			-4,97	-4,52	-4,06	-3,66	-3,27	-2,88	-2,51	-2,15	-1,81	-1,49	-1,18	-0,88	
Temat III	Treść	1,77	-0,60	-0,32	-0,05	0,22	0,50	0,78	1,07	1,37	1,70	2,06	2,48	2,97	3,67
	Kompozycja	2,23	-3,55	-1,60	0,65										
	Styl	2,27	-3,74	-1,67	0,99										
	Język	2,28	-3,87	-2,98	-1,53	0,00	1,48								
Temat IV	Zapis	0,84	-2,61	-0,71	1,49										
	Walory	1,47	2,36	3,66	4,92	6,36									

Aneks 2. Wykresy rozkładów parametrów rkr (zniekształceń skali)

Rysunek 1. Wykresy rozkładów parametrów rkr dla języka polskiego w podziale na tematy i kryteria oceny



Rysunek 2. Wykresy rozkładów parametrów rkr dla matematyki w podziale na zadania



Aneks 3. – Kwestionariusz ankiety egzaminatora

Poniżej znajdują się pytania związane z pracą egzaminatora podczas majowych sesji oceniania pisemnych prac egzaminacyjnych. Rzetelne odpowiedzi na te pytania pozwolą nam lepiej poznać kontekst Państwa pracy. Dzięki nim dowiemy się również, jak bardzo zróżnicowani są pod pewnymi względami egzaminatorzy i czy to zróżnicowanie ma wpływ na ocenianie.

☐ ₁ język polski	[kod OKE]
☐ ₂ matematyka	
[naklejka z numerem egzaminatora]	

1. Jaka jest Pana(-i) płeć?

Proszę zaznaczyć jedną wybraną odpowiedź.

☐₁ kobieta ☐₂ mężczyzna

2. W którym roku się Pan(-i) urodził(-a)?

Proszę uzupełnić datę.

1	9		
---	---	--	--

3. Jak długo naucza Pan(-i) ocenianego dziś przedmiotu w szkole kończącej się maturą?

Proszę zaznaczyć jedną wybraną odpowiedź.

☐₁ 3-5 lat ☐₂ 6-10 lat ☐₃ 11-15 lat ☐₄ 16-24 lat ☐₅ powyżej 24 lat

4. W których z majowych sesji oceniania egzaminów zewnętrznych Pan(-i) uczestniczył(-a)?

Proszę zaznaczyć wszystkie oceniane sesje.

	Sprawdzian	Egzamin gimnazjalny	Egzamin maturalny	Byłem(-am) Przewodniczącym(-ą) Zespołu Egzaminatorów w trakcie egzaminu maturalnego
2002	☐ ₁	☐ ₁		
2003	☐ ₁	☐ ₁		
2004	☐ ₁	☐ ₁		
2005	☐ ₁	☐ ₁	☐ ₁	☐ ₁
2006	☐ ₁	☐ ₁	☐ ₁	☐ ₁
2007	☐ ₁	☐ ₁	☐ ₁	☐ ₁
2008	☐ ₁	☐ ₁	☐ ₁	☐ ₁
2009	☐ ₁	☐ ₁	☐ ₁	☐ ₁
2010	☐ ₁	☐ ₁	☐ ₁	☐ ₁

2011	☐ ₁	☐ ₁	☐ ₁	☐ ₁
2012	☐ ₁	☐ ₁	☐ ₁	☐ ₁
2013	☐ ₁	☐ ₁	☐ ₁	☐ ₁

5. Jak bardzo praca w wymienionych poniżej rolach jest dla Pana(-i) stresująca?

Proszę zaznaczyć jedną wybraną odpowiedź przy każdej roli (na skali od 1 do 5).

	1 – wcale niestresująca	2	3	4	5 – bardzo stresująca
Nauczyciel lub inny zawód, który wykonuje Pan(-i) na co dzień.	☐ ₁	☐ ₂	☐ ₃	☐ ₄	☐ ₅
Egzaminator OKE.	☐ ₁	☐ ₂	☐ ₃	☐ ₄	☐ ₅

6. Czy oceniając prace egzaminacyjne, ma Pan(-i) na uwadze to, że od Pana(-i) rzetelnej pracy oceniającego zależy przyszłość zdającego?

Proszę zaznaczyć jedną wybraną odpowiedź.

☐₁ zawsze ☐₂ czasami ☐₃ rzadko ☐₄ nie zastanawiam się nad tym

7. Jak bardzo stresuje Pana(-ią) świadomość, że zdający oglądają swoje rozwiązania zadań w okręgowej komisji egzaminacyjnej po zakończeniu sesji?

Proszę zaznaczyć jedną wybraną odpowiedź (na skali od 1 do 5).

1 – wcale nie stresuje	2	3	4	5 – bardzo stresuje
☐ ₁	☐ ₂	☐ ₃	☐ ₄	☐ ₅

8. Czy kiedykolwiek podwyższył(-a) Pan(-i) punktację za rozwiązania, zadania egzaminacyjnego, mimo że było ono inaczej sformułowane niż w schemacie oceniania?

Proszę zaznaczyć jedną wybraną odpowiedź w każdym wierszu.

	TAK	NIE
Aby uniknąć ewentualnego odwołania.	☐ ₁	☐ ₂
Aby przyznać punkty za rozwiązanie nietypowe, poprawne merytorycznie.	☐ ₁	☐ ₂

9. Czy oceniając jedno zadanie, bierze Pan(-i) pod uwagę odpowiedzi zdającego na pozostałe zadania?

Proszę zaznaczyć jedną wybraną odpowiedź.

☐₁ TAK ☐₂ NIE

10. Proszę określić na skali od 1 – zdecydowanie się nie zgadzam do 4 – zdecydowanie się zgadzam, na ile zgadza się Pan(-i) lub nie z poniższymi stwierdzeniami odnośnie oceniania prac pisemnych w majowych sesjach egzaminacyjnych.

Proszę zaznaczyć jedną wybraną odpowiedź przy każdym stwierdzeniu.

	1 – zdecydowanie się nie zgadzam	2 – raczej się nie zgadzam	3 – raczej się zgadzam	4 – zdecydowanie się zgadzam
Uczestnictwo w pracach zespołu egzaminatorów jest dla mnie interesującym zajęciem.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
Uczestnictwo w sesji oceniania prac pozwala mi osiągać lepsze efekty nauczania mojego przedmiotu.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
Nie chciałbym(-abym) stracić możliwości uczestnictwa w sesjach oceniania prac egzaminacyjnych.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
Głównym powodem uczestnictwa w sesji oceniania prac jest dla mnie atrakcyjne honorarium.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
Dzięki uczestnictwu w sesjach oceniania prac lepiej przygotowuję uczniów do egzaminów.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
Uczestnictwo w sesjach oceniania prac egzaminacyjnych przeszkadza mi w dobrym nauczaniu mojego przedmiotu.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
Głównym powodem uczestnictwa w sesji oceniania prac jest dla mnie rozwój kompetencji.	☐ ₁	☐ ₂	☐ ₃	☐ ₄

11. Czy specjalizuje się Pan(-i) w ocenie jednego z poziomów egzaminu maturalnego?

Proszę zaznaczyć jedną wybraną odpowiedź.

- ☐₁ Specjalizuję się w ocenie poziomu podstawowego.
☐₂ Specjalizuję się w ocenie poziomu rozszerzonego.
☐₃ Nie specjalizuję się w ocenie żadnego z poziomów.

12. Proszę określić na skali od 1 – zdecydowanie się nie zgadzam do 4 – zdecydowanie się zgadzam, na ile zgadza się Pan(-i) lub nie z poniższymi stwierdzeniami odnośnie oceniania prac pisemnych w majowych sesjach egzaminacyjnych.

Proszę zaznaczyć jedną wybraną odpowiedź przy każdym stwierdzeniu.

	1 – zdecydowanie się nie zgadzam	2 – raczej się nie zgadzam	3 – raczej się zgadzam	4 – zdecydowanie się zgadzam
Modyfikacja sposobów oceniania poprzez ocenianie zadaniami lub grupami zadań powoduje przyspieszenie oceniania prac.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
Modyfikacja sposobów oceniania poprzez ocenianie zadaniami lub grupami zadań powoduje poprawienie jakości oceniania.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
Oceny zadań egzaminacyjnych analizowanych przeze mnie przed przystąpieniem do sprawdzania prac najczęściej są zgodne z modelem odpowiedzi/kryteriami oceniania przekazanymi przez ekspertów.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
W trakcie sprawdzania prac egzaminacyjnych konsultacje merytoryczne PZE są przydatne do podjęcia trafnej decyzji.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
W trakcie sprawdzania prac egzaminacyjnych ustalenia wyniku punktowego przeze mnie i egzaminatora powtórnie sprawdzającego (EPS) najczęściej są zgodne.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
Zapisy w kryteriach oceniania i dołączonych do nich schematach punktowania są zredagowane komunikatywnie i nie budzą wątpliwości.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
W trakcie szkolenia merytorycznego, przed przystąpieniem do sprawdzania, PZE wysłuchuje uwag egzaminatorów i przekazuje je do koordynatora krajowego w celu uzupełnienia modelu odpowiedzi.	☐ ₁	☐ ₂	☐ ₃	☐ ₄
Moim zdaniem praca w zespole egzaminatorów (PZE+EPS+egzaminatorzy) umieszczonym w jednym ośrodku oceniania służy poprawie jakości sprawdzania.	☐ ₁	☐ ₂	☐ ₃	☐ ₄

Dziękujemy za udzielenie odpowiedzi.

Instytut Badań Edukacyjnych

Głównym zadaniem Instytutu jest prowadzenie badań, analiz i prac przydatnych w rozwoju polityki i praktyki edukacyjnej.

Instytut zatrudnia ponad 150 badaczy zajmujących się edukacją – pedagogów, socjologów, psychologów, ekonomistów, politologów i przedstawicieli innych dyscyplin naukowych – wybitnych specjalistów w swoich dziedzinach, o różnorodnych doświadczeniach zawodowych, które obejmują, oprócz badań naukowych, także pracę dydaktyczną, doświadczenie w administracji publicznej czy działalność w organizacjach pozarządowych.

Instytut w Polsce uczestniczy w realizacji międzynarodowych projektów badawczych w tym PIAAC, PISA, TALIS, ESLC, SHARE, TIMSS i PIRLS oraz projektów systemowych współfinansowanych przez Unię Europejską ze środków Europejskiego Funduszu Społecznego.

Instytut Badań Edukacyjnych

ul. Górczewska 8, 01-180 Warszawa | tel. +48 22 241 71 00 | ibe@ibe.edu.pl | www.ibe.edu.pl

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.