Henryk Szaleniec
Filip Kulon
Instytut Badań Edukacyjnych

Wpływ rodzaju pisma (ręczne vs komputerowe) na ocenę wypracowania maturalnego z języka polskiego

Wstęp

Na wynik oceny wypracowania, oprócz jego jakości odzwierciedlającej poziom badanej umiejętności egzaminowanego, ma wpływ wiele czynników ogólnie nazywanych efektem egzaminatora (Scullen, Mount i Goff, 2000). Efekt egzaminatora to szereg czynników będących źródłem wariancji oceny, której źródłem jest egzaminator. Obejmuje on między innymi takie czynniki, jak łagodność/surowość egzaminatora, efekt halo, tendencja centralna polegająca na zawężaniu skali przez obcięcie ocen skrajnych czy tendencja przyznawania ocen skrajnych z pomijaniem centralnej części skali. Czynniki te mogą mieć różne źródło. Między innymi osobowość egzaminatora lub teorie uczenia się leżące u podstaw działalności edukacyjnej danego nauczyciela mają wpływ na to, jak egzaminator postrzega pracę ucznia w trakcie dokonywania oceny. Teorie, które są obecne w świadomości oceniającego, mogą wywierać różny wpływ na indywidualną walidację przez egzaminatora kryteriów oceniania, które obowiązują w trakcie oceniania.

Jak wynika z badań Wu Siew Mei (Mei, 2010), w fazie wstępnej oceniania egzaminator raczej wyrabia sobie „ogólne wrażenie" co do jakości tekstu i nie dokonuje żadnej punktacji. Nie jest jasne, jakie czynniki kształtują to „ogólne wrażenie (intuicję)", chociaż ten etap może mieć fundamentalne znaczenie dla oceny końcowej. Na etapie przydzielania punktacji działanie egzaminatora jest w zasadzie jasne i ukierunkowane poprzez instrukcję (schemat oceniania) na szacowanie jakości wypracowania w obrębie wyróżnionych kryteriów. Mniej są jednakże jasne czynniki, które powodują przyznawanie ocen nie zawsze zgodnych z dopasowaniem jakości ocenianej cechy tekstu i kryteriów w schemacie oceniania. Wygląda na to, że związek pomiędzy jakością ocenianej cechy wypracowania, kryterium zawartym w schemacie oceniania i decyzją podejmowaną przez egzaminatora jest bardzo złożony.

Strategie postępowania w odniesieniu do nietypowych wypracowań zazwyczaj są ustalane podczas szkolenia bezpośrednio przed przystąpieniem do oceniania i ostatecznie egzaminator próbuje dopasować „ogólne wrażenie", jakie wywarł na nim tekst, do schematu oceniania nawet wtedy, gdy jest to trudne do zrobienia. Doświadczeni egzaminatorzy zwykle zgadzają się z kryteriami oceniania, które poznają podczas szkolenia przed przystąpieniem do sesji oceniania, ale jest prawdopodobne, że użyją własnych, jeżeli wypracowanie okaże się niespójne z określonymi standardami.

Są też dowody, że na ocenę może mieć wpływ czytelność pisma (Wood, 1991), czy też forma zapisu – pismo odręczne lub zapis komputerowy (Mogey i in., 2010). W trakcie sesji egzaminacyjnej niewielka część prac jest pisana na komputerze przez zdających lub przepisywana przez asystentów. Dotyczy to takich przypadków jak głęboka dysgrafia, czego wynikiem jest nieczytelność pisma odręcznego, która wręcz uniemożliwia odczytanie i ocenę wypracowania przez egzaminatora, niedowidzenie (na komputerze jest możliwość ustawienia wielkości czcionki) oraz przypadki fizycznych dysfunkcji uniemożliwiających ręczne pisanie. Na świecie było stosunkowo niewiele badań dotyczących zróżnicowania ocen przyznawanych przez egzaminatorów za prace pisane przez zdających odręcznie i na komputerze. Jeżeli pojawiają się badania w tym zakresie, to stanowią one jeden z wątków podczas studiów dotyczących zróżnicowania (DIF) pomiędzy wynikami standaryzowanych testów w zakresie umiejętności pisania własnego tekstu (esej, wypracowanie maturalne, rozprawka itp.) odręcznie a pisaniem z wykorzystaniem komputerowej klawiatury i edytora tekstu. W takich przypadkach przedmiotem dociekań jest nie tylko problem efektu egzaminatora, ale przede wszystkim różnice w zakresie kognitywnym podczas używania przez zdających klawiatury i funkcjonalności edytora tekstu w porównaniu z arkuszem papieru i długopisem (Lottridge i in., 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. Pojawia się więc problem ekwiwalentności wyników oceniania prac zapisanych odręcznie i na komputerze oraz problem ekwiwalentności konstruktu – umiejętności sprawdzanych za pomocą tak samo sformułowanego zadania i tych samych kryteriów oceniania.

Metodologia badań

W prezentowanych badaniach problem badawczy dotyczył tylko ekwiwalentności ocen wypracowań ocenianych w dwóch formach ich zapisu – pisma odręcznego i kopii sporządzonej z wykorzystaniem komputera i edytora tekstu. Celem przeprowadzonego studium było zbadanie, czy forma zapisu wypracowania maturalnego z języka polskiego na poziomie podstawowym (pismo odręczne – PO versus zapis komputerowy – PK) generuje efekt egzaminatora. Zostały sformułowane dwa pytania badawcze, na które poszukiwano odpowiedzi.

1. Czy ocena przyznawana przez egzaminatorów za wypracowanie jest obciążona efektem formy zapisu (pismo odręczne versus zapis komputerowy)?
2. Czy któreś z kryteriów oceniania szczególnie narażone jest na obciążone efektem egzaminatora ze względu na formę zapisu?

Dobór prac do badań

Badania wpływu formy zapisu (PO i PK) na wynik oceniania są częścią szerszego studium dotyczącego porównywalności oceniania i efektu egzaminatora realizowanego w Instytucie Badań Edukacyjnych (IBE) przez Zespół Analiz Osiągnięć Uczniów (ZAOU), w których wykorzystano próbę 1000 wypracowań z prac maturalnych z języka polskiego na poziomie podstawowym z 2011 i 2012 roku, wylosowanych z dwóch OKE (2011 r. – Jaworzno, 2012 r. – Kraków).

Wybór prac tylko z dwóch OKE podyktowany był względami praktycznymi (konieczna była pełna anonimizacja cyfrowych kopii prac w OKE i usunięcie wszelkich zapisów poczynionych przez egzaminatorów). Wstępnie z uwagi na charakter badań wszystkie wypracowania zostały podzielone ze względu na wynik na 50 warstw. Przyjęto, iż jedna warstwa nie może liczyć mniej niż 50 prac. Spełnienie takiego założenia wymagało jednak sklejenia skrajnych kategorii na skali wyników punktowych, tak aby uzyskać wystarczająco liczebne warstwy. Wewnątrz każdej warstwy wypracowania zostały wybrane w wyniku losowania prostego z równym prawdopodobieństwem wyboru. Próba wszystkich 1000 prac stanowiła materiał źródłowy do stworzenia klasyfikacji prac ze względu na ich czytelność i wyrobienie pisma oraz do wyboru prac do przepisania na komputerze. Ponieważ oryginalne prace zawierały identyfikatory personalne autorów wypracowań, do badań wykorzystano cyfrowe kopie, na których usunięte zostały nie tylko identyfikatory, ale także adnotacje egzaminatorów.

Wybierając prace do przepisania, wykorzystano skalę zaproponowaną przez Jerzego Frąszczaka (Frąszczak, 2013), wyróżniającą trzy kryteria, 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 elementarzowych 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) niewyrobione.

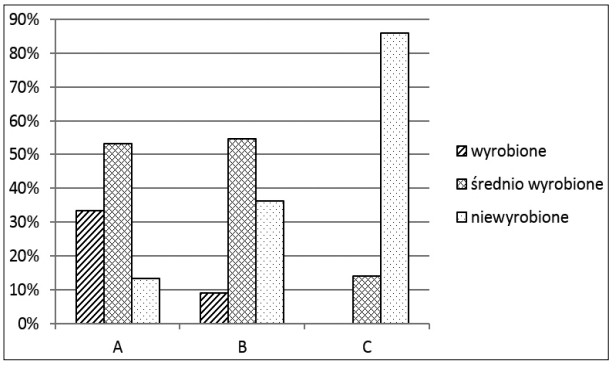
a. Pismo wyrobione, charakteryzujące się między innymi dużą płynnością linii graficznej, subtelnym cieniowaniem pisma, swoistym rozwiązaniem budowy znaków oraz znaczną rozmaitością elementów powtarzalnych.
b. Pismo średnio wyrobione, charakteryzujące się między innymi umiarkowaną płynnością linii graficznej, 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:

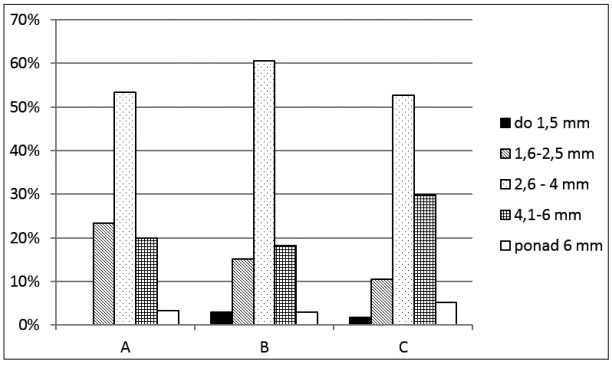
I. pismo drobne – do 1,5 mm
II. pismo małe – od 1,5 mm do 2,5 mm
III. pismo średnie – od 2,6 mm do 4,0 mm
IV. pismo duże – od 4,1 mm do 6,0 mm
V. pismo bardzo duże – ponad 6 mm.

Stosując wyróżnione kryteria, z próby 1000 prac egzaminacyjnych z języka polskiego na poziomie podstawowym w latach 2011 i 2012 wybranych zostało 120 wypracowań do analizy porównawczej wyników oceniania prac zapisanych oryginalnie pismem odręcznym i ich kopii przepisanych na komputerze.

Próba 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 wyrobienia pisma i jego wielkości w wymienionych kategoriach czytelności przedstawiają rysunek 1. i rysunek 2.



Rysunek 1. Reprezentacja w próbie trzech kategorii wyrobienia pisma z uwzględnieniem czytelności. Procent dla każdej klasy wyrobienia obliczony jest względem liczebności wypracowań w poszczególnych kategoriach czytelności A, B, C.



Rysunek 2. Reprezentacja w próbie pięciu kategorii wielkości pisma z uwzględnieniem czytelności. Procent dla każdej klasy wyrobienia obliczony jest względem liczebności wypracowań w poszczególnych kategoriach czytelności A, B, C.

Ocenianie wypracowań

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 siedzibą 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 przepisanego na komputerze oceniana była przez czterech egzaminatorów (każdy egzaminator z innej OKE).

Wyniki badań – efekt formy zapisu wypracowania (pismo odręczne versus 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 przeszło tysiąc prac w każdej sesji. Dlatego też wiedza, czy wynik tych prac jest obciążony efektem egzaminatora ze względu na formę zapisu, może mieć znaczenie dla rzetelności oceniania i procesu szkolenia egzaminatorów oceniających wypracowania w zapisie komputerowym.

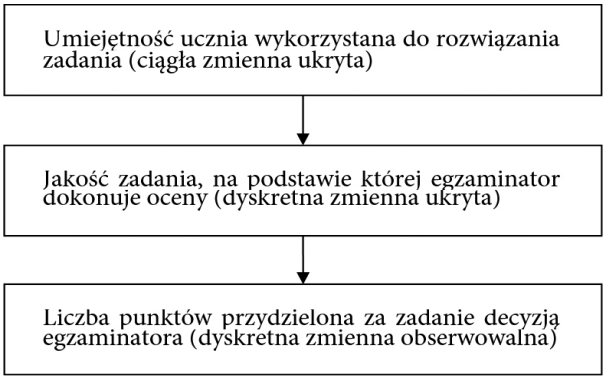
W związku z tym, że każda praca była oceniana przez ośmiu (PO) lub czterech (PK) egzaminatorów, pierwszym krokiem do przeprowadzenia analiz jest ustalenie „ostatecznej" oceny danej pracy. Można to zrobić na kilka sposobów.

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 jednak 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, innym, podejściem jest skorzystanie z ocen oszacowanych z wykorzystaniem *Hierarchical Rater Model with Signal Detection Theory* (HRM-SDT). Model ten został użyty na potrzeby wspomnianych badań porównywalności oceniania i efektu egzaminatora prowadzonych przez IBE (Kulon i Żółtak, 2014).

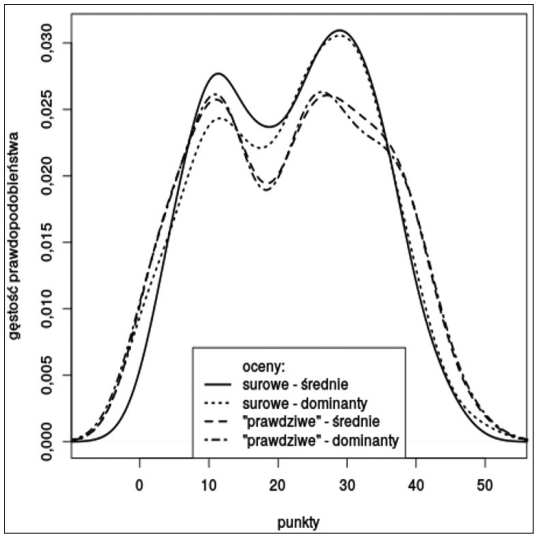
Model HRM-SDT uwzględnia hierarchiczną strukturę danych, jakie powstają w procesie, gdy egzaminatorzy oceniają zadania rozszerzonej odpowiedzi, których przykładem jest wypracowanie. Zgodnie z tym modelem obserwowalne oceny egzaminatorów przydzielane podczas procesu oceniania nie są postrzegane jako bezpośredni wskaźnik umiejętności egzaminowanych, lecz raczej jako wskaźnik nieobserwowalnej bezpośrednio jakości wypracowania (kategorialnej zmiennej ukrytej). Z kolei jakość wypracowania napisanego przez zdającego służy jako wskaźnik ukrytej cechy, jaką jest umiejętność egzaminowanego wykorzystana do napisania wypracowania. Tę hierarchiczną strukturę można zapisać następująco:



Gdy uznamy, że jakość zadania powinna być reprezentowana poprzez takie same kategorie, jakimi dysponuje egzaminator, to może ona być traktowana jako „prawdziwa" ocena ucznia w tym zadaniu. Pod pojęciem „prawdziwa" ocena jest tutaj rozumiana ocena pozbawiona wpływu takich efektów egzaminatora, jak np. łagodność czy niska nierzetelność (Kulon, 2014).

Model ten szacowany był bayesowsko, z użyciem metody *Markov Chain Monte Carlo* (MCMC)[1]. 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 rzeczywistemu 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 3. Porównanie rozkładów ocen ostatecznych (oszacowanych na podstawie oceny kilku egzaminatorów) dla różnych sposobów ich ustalania

Rysunek 3. 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 ocen w środku skali w porównaniu z rozkładami wyników „prawdziwych" z modelu HRM-SDT. Użycie dominanty wyników surowych prowadzi do uzyskania wartości nieco wyższych niż w przypadku użycia

[1] Czytelników mniej zaawansowanych w zagadnieniach statystyki wykorzystującej metodę MCMC zapraszamy do zapoznania się z przystępnym wykładem Łańcuch Markowa, pod adresem http://wazniak.mimuw.edu.pl/index.php?title=Rachunek_prawdopodobie%C5%84stwa_i_statystyka/Wyk%C5%82ad_10:_%C5%81a%C5%84cuchy_Markowa.

średniej – można zauważyć rozbieżność pomiędzy tymi rozkładami poniżej 20 punktów na skali. Rozkłady ocen obliczonych na podstawie „prawdziwych" wyników (średnie i dominanty) z modelu HRM-SDT nie różnią się zbytnio. Tabela 1. zawiera podsumowanie statystyk omawianych rozkładów.

Tabela 1. Statystyki rozkładów ocen ostatecznych (oszacowanych na podstawie oceny kilku egzaminatorów) dla różnych sposobów ich ustalania

	minimum	1 kwartyl	mediana	3 kwartyl	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 versus 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 ta sama 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. W tabeli 2. przedstawiono informacje o różnicach w średnich pomiędzy pracami pisanymi ręcznie i przepisanymi 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 2. 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.

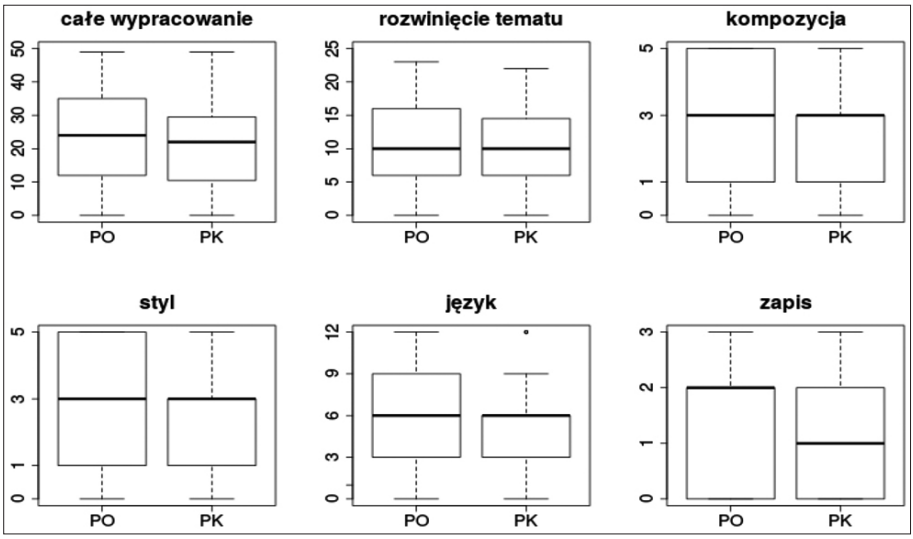
Zaprezentowane powyżej wyniki testów t użytych na sumarycznej ocenie za wypracowanie pozwalają pozytywnie odpowiedzieć na pierwsze z postawionych pytań badawczych. Istnieje obciążenie oceny za wypracowanie w zależności od formy jego zapisu. Można zatem szukać odpowiedzi na drugie z postawionych pytań – 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. W tabeli 3. przedstawiono różnice średnich pomiędzy pracami oryginalnymi (PO) i pracami przepisanymi (PK) w podziale na poszczególne kryteria.

Tabela 3. 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)
rozwinięcie 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)

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.

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 rozwinięcia 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 rozwinięcia tematu. Na rysunku 4. przedstawiono wykresy skrzynkowe ilustrujące różnicę w rozkładach wyników za realizację wypracowania w całości oraz w zakresie poszczególnych kryteriów.



Rysunek 4. 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).

Dyskusja

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, co wielokrotnie autorzy mogli obserwować, 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 Farnum, 1997; Powers i in., 1994).

Na podstawie analizy danych z badań 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 i in., 1994; Russell i Plati, 2001).

Egzaminatorzy oceniający w sesji egzaminacyjnej prace zarówno w formie odręcznego zapisu, jak i równocześnie w zapisie komputerowym podkreślają, że ten efekt spowodowany jest dwoma czynnikami. Po pierwsze, podczas oceniania

wypracowania napisanego odręcznie łatwiej sobie wyobrazić ucznia, który jest autorem wypracowania, co ma wpływ na efekt łagodności oceniania. Zjawisko to polegające na tendencji zawyżania oceny w wyniku personalnego utożsamiania się oceniającego z autorem wypracowania napisanego odręcznie nazywane jest efektem empatii w ocenianiu (*reader empathy assessment discrepancy* – READ – *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).

Ponadto w zapisie odręcznym w przypadku wątpliwości egzaminatora co do jakości fragmentu tekstu (np. wskutek obniżonej czytelności) egzaminatorzy podejmują decyzję na korzyść zdającego, zresztą z obowiązującymi zasadami oceniania w sesji egzaminacyjnej. Warto też podkreślać, że w komputerowym tekście wszystkie błędy są bardziej widoczne. Na występowanie takiego zjawiska zwracają uwagę Powers i inni (Powers i Farnum, 1997; Powers i in., 1994), a także Russel i Tao (Russell i Tao, 2004). W eksperymencie prowadzonym przez Russela i Tao egzaminatorzy zostali poproszeni o rejestrację błędów (w zakresie pisowni, interpunkcji, stosowania dużych liter, błędnych zwrotów) przy ocenie esejów w zapisie oryginalnym i ich kopii sporządzonych na komputerze. W przypadku kopii przepisanych na komputerze oceniający zidentyfikowali więcej błędów i różnice były statystycznie istotne.

Na wynik oceny może mieć też wpływ różnica w liczbie stron tekstu tej samej pracy w zapisie odręcznym i komputerowym. W przeprowadzonych badaniach przepisany tekst czcionką 12 pkt z pojedynczym odstępem jest średnio o 1,95 strony krótszy niż w pracy oryginalnej. Różnica ta jest statystycznie istotna (95% przedział ufności: 1,80-2,09). Różnice te są znaczne, szczególnie gdy pismo zdającego jest duże (4,1-6 mm) lub bardzo duże (powyżej 6 mm). Wpływ tego efektu na różnicę wyników oceniania prac pisanych odręcznie i w zapisie komputerowym jest przedmiotem dalszych analiz.

Podsumowanie

Umiejętność pisania swobodnych, improwizowanych wypowiedzi to ważna umiejętność kształcona i będąca przedmiotem diagnozy oraz egzaminowania na wszystkich etapach edukacyjnych. W Polsce na sprawdzianie na zakończenie szóstej klasy szkoły podstawowej zwykle jedno z zadań to opowiadanie, list lub opis; na egzaminie gimnazjalnym z języka polskiego to rozprawka, a na egzaminie maturalnym z języka polskiego – wypracowanie. Jak już wspomniano we wstępie, część prac (uczniów z dysfunkcjami w zakresie pisania) pisanych jest na komputerze lub jest kopiowana przez przepisanie na komputerze. W trakcie szkolenia egzaminatorów oceniających prace pisane na komputerze warto zwrócić uwagę na zaobserwowany wpływ rodzaju zapisu (ręczny vs komputerowy) na ocenę wypracowania maturalnego z języka polskiego.

W trakcie warsztatów prowadzonych przez autorów wielu nauczycieli sygnalizowało, że w procesie edukacyjnym niektórzy uczniowie na zadanie domowe tworzą teksty, pisząc z wykorzystaniem procesora tekstu i potem przepisują je, aby przedstawić w klasie w odręcznym zapisie. Dlatego ważne jest poznanie,

czy ocenianie uczniowskich wypracowań (także w ocenianiu wewnątrzszkolnym) obciążone jest efektem egzaminatora ze względu na formę zapisu. Zbadanie tego zjawiska ma też znaczenie w związku z przenikaniem do życia szkolnego i do egzaminów oceniana z wykorzystaniem komputerów. Możemy w tej dziedzinie znaleźć próby takich zastosowań nie tylko za granicą, ale także w kraju. Warto też tutaj podkreślić, że najbliższe badania PISA w 2015 roku będą przeprowadzone z wykorzystaniem komputerów.

Bibliografia

1. Arnold, V., Legas, J., Obler, S., Pacheco, M. A., Russell, C., & Umbdenstock, L., *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. (1990).
2. Cochran-Smith, M. (1991). *Word processing and writing in elementary classrooms: A critical review of related literature*. Review of Educational Research*, 61*, 107-155.
3. Frąszczak, J., *Czytelność odręcznego tekstu wypracowań z egzaminu maturalnego z języka polskiego. Analiza na zamówienie IBE*. Maszynopis w archiwum Zespołu Analiz Osiągnięć Uczniów, 2013.
4. Cahalan-Laitusis C., *Accommodations on high stakes writing tests for students with disabilities*. Princeton, NJ: Educational Testing Service, 2003.
5. Kulon, F., *Modele analizy efektu oceniającego w pomiarze edukacyjnym*, 2014 [w druku].
6. Kulon, F., Żółtak, M., *Zróżnicowanie łagodności egzaminatorów między okręgowymi komisjami egzaminacyjnymi*, 2014 [w druku].
7. Lottridge, S., Nicewander, A., Schulz, M., Mitzel, H., *Comparability of Paper-based and Computerbased Tests: A Review of the Methodology*, Pacific Metrics Corporation, 2008.
8. Mei W. S., *Investigating Raters' Use of Analytic Descriptors in Assessing Writing*, Reflections on English Language Teaching, Vol. 9, No. 2, pp. 69–104, 2010.
9. Mogey, N., Paterson, J., Burk, J. & Purcell, M., *Typing compared with handwriting for essay examinations at university: letting the students choose ALT-J, R*esearch in Learning Technology Vol. 18, No. 1, 29–47, 2010.
10. Olson, D. R. *The world on paper: The conceptual and cognitive implications of writing and reading.* Cambridge: Cambridge University Press , 1994.
11. Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. *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. 1994.
12. Russell, M., Tao W. (2004). *Effects of handwriting and computer-print on composition scores, a follow-up to powers, fowles, farnum, & ramsey*. Practical Assessment, Research & Evaluation, 9(1). Retrieved July 12, 2014 from http://PAREonline.net/getvn.asp?v=9&n=1
13. Russell, M., & Plati, T. *Effects of computer versus paper administration of a state-mandated writing assessment*. Teachers College Record 2001.
14. Scullen, S. E., Mount, M. K., Goff, M., *Understanding the latent structure of job performance ratings*. Journal of Applied Psychology, 85, 2000.

15. Wood, R. *Assessment and Testing*. A Survey of Research, Cambridge, Cambridge University Press, 1991.
16. *Kryteria oceniania odpowiedzi. Egzamin 2012. Język polski poziom podstawowy*. CKE 2012.
17. Russell, M., and W. Tao. *Effects of handwriting and computer print on composition scores:* A follow up to Powers, Fowles, Farnum & Ramsay Practical Assessment Research and Evaluation 9. http://pareonline.net/getvn.asp?v=9&n=1, 2004.
18. Wolfe E., Bolton S. Feltovich B., Welch C. *A Comparison of Word-Processed and Handwritten Essays From a Standardized Writing Assessment*. ACT, 1993.