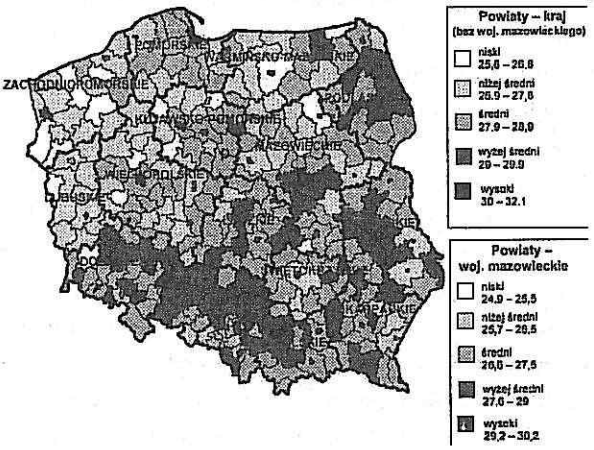
Henryk Szaleniec

Okręgowa Komisja Egzaminacyjna w Krakowie

Zgodność punktowania wyników zadań otwartych przez dwu egzaminatorów

Egzaminy zewnętrzne w Polsce mają bardzo krótką historię. Rok 2003 jest drugim rokiem ich implementacji. Wyniki egzaminów coraz częściej wykorzystywane są nie tylko do wewnątrzszkolnej diagnozy i rekrutacji uczniów do szkół wyższego szczebla, ale także do szerszych analiz i porównań w skali całego kraju. Zamieszczona poniżej mapa ilustruje porównania wyników sprawdzianu powiatami w całym kraju.

Rysunek 1.
Mapa średnich wyników sprawdzianu 2003 wyrażonych w skali staninowej*



* ze względu na różnice w interpretacji kryteriów oceniania dla powiatów województwa mazowieckiego wprowadzono korektę w skali staninowej

Źródło: Raport CKE, lipiec 2003

Rodzi się więc pytanie, czy takie porównania są w pełni uprawnione? Jakie konsekwencje dydaktyczne i społeczne mogą powodować nie w pełni uprawnione porównania? Co zrobić, aby trafność i rzetelność pomiaru w pełni uprawniała do takich porównań? Jednym z działań w tym kierunku jest po pierwsze, sprawdzanie i zapewnianie wysokiej zgodności oceniania prac w obrębie danej komisji egzaminacyjnej. Po drugie, stosowanie procedur, które pozwoliłyby na najpełniejszą, jak tylko to jest możliwe, porównywalność oceniania między okręgowymi komisjami egzaminacyjnymi.

Badanie konsekwencji przeprowadzania egzaminu i interpretacji jego wyników nazywa się analizą trafności konsekwencyjnej. Pojęcie to znalazło miejsce w pomiarze dydaktycznym stosunkowo niedawno i jest ono w zgodzie z filozofią badań przyrodniczych. Badania przyrodnicze nie są już dziś postrzegane jako dziedzina swobodnych działań. Messick jest pierwszym, który zwrócił uwagę, że musimy postawić dwa istotne pytania. Pierwsze, czy test (arkusz egzaminacyjny) jest tak dobry jak miara charakterystyk, które za pomocą niego chcemy poznać. Drugie pytanie to, do jakich celów dany konkretny test może być zastosowany? O czym chcemy wnioskować na podstawie wyników testowania? Chcąc odpowiedzieć na te pytania, za każdym razem powinniśmy oszacować nie tylko trafność danego testu, ale trafność jego użycia (wykorzystania wyników) do określonych celów.

Rodzi się więc pytanie o konsekwencje wykorzystywania wyników egzaminów zewnętrznych szerzej niż np. zakładali to autorzy, budując narzędzie do oceny wybranych umiejętności zgodnie z przyjętymi standardami egzaminacyjnymi. Pytanie, czy obecne i potencjalne skutki interpretacji wyników egzaminów zewnętrznych w Polsce mają tylko konsekwencje pozytywne i zgodne z założeniami, czy też są jakieś uboczne skutki o negatywnym społecznie znaczeniu? Na to i inne pytania nie ma prostej odpowiedzi. Musimy je sobie jednak postawić. Musimy szukać odpowiedzi, do jakich uogólnień, do jakich wniosków upoważniają nas wyniki egzaminów, a do jakich nie.

Linn zwraca uwagę, że kiedy wyniki pojedynczych uczniów są agregowane do wyników szkoły, gminy, powiatu czy nawet całego województwa, to należy postawić istotne pytanie na temat trafności. Wyniki mogą być wysoce trafne do oszacowania osiągnięć poszczególnych uczniów, natomiast interpretacja rezultatów może mieć ujemny wpływ na szkoły, grupy szkół, jeżeli wykorzystane zostaną jako jedyny parametr np. do rankingu. Taki ranking zachęca rodziców do wyboru szkół, które są najwyżej w rankingu. Szkoły są też karane administracyjnie za niskie wyniki z egzaminów zewnętrznych nawet, jeżeli to nie ma związku z procesem dydaktycznym, a wynika z lokalizacji szkoły w obszarze, gdzie edukacja nie należy do wysokich priorytetów większości mieszkańców. W przypadku różnych wykorzystań wyników egzaminów zewnętrznych jest mało prawdopodobne, aby udało się zbudować taki arkusz egzaminacyjny, który byłby jednakowo trafny na poziomie ucznia, szkoły, gminy czy też całego województwa. Oznacza to także, że nie można zbudować trafnego arkusza egzaminacyjnego jednocześnie do wielu celów. Potwierdzają to prace prowadzone w Anglii, gdzie od lat funkcjonuje system egzaminów zewnętrznych. Z drugiej strony, musimy

zdawać sobie sprawę, że kontrola nad wykorzystaniem wyników egzaminów nie jest ani pożądana, ani nawet możliwa. Kontrola interpretacji wyników egzaminów zewnętrznych jest trudna nawet do określenia. Warto natomiast podjąć wszelkie starania, aby zminimalizować obciążenie błędem wyników egzaminów oraz wykorzystywać je w badaniach łącznie z innymi zmiennymi kontekstowymi mogącymi objaśnić efektywność i skuteczność edukacyjną szkoły.

Messick podkreśla, że jeżeli wiemy, iż test może być obciążony nietrafnością, na przykład wobec pewnych środowisk, to jego zastosowanie czy interpretacja wyników jest problemem społecznych lub politycznych decyzji.

W centralnej i okręgowych komisjach egzaminacyjnych powinniśmy przyjąć założenie, że dowody na trafność arkusza egzaminacyjnego i konsekwencje społeczne stosowania wyników egzaminu do poszczególnych celów muszą być brane pod uwagę przy podejmowaniu decyzji o szerokim ich wykorzystywaniu. Wychodząc z tego założenia, zawsze powinniśmy rozważyć zamierzone i niezamierzone konsekwencje społeczne wykorzystania wyników egzaminu zewnętrznego czy też towarzyszących mu badań.

Zagadnienie to dotyka etycznych aspektów egzaminów. Konsekwencje społeczne są kluczowym zagadnieniem w rozważaniach na temat etyki egzaminowania i wykorzystania wyników egzaminów zewnętrznych do różnych celów. Rozszerzenie pojęcia trafności pomiaru dydaktycznego poprzez włączenie do niego konsekwencji społecznych powoduje, że ewaluacja trafności zaczyna nabierać znaczenia personalnych, instytucjonalnych, jak i środowiskowych celów.

Trudno mówić o trafności w całkowitym oderwaniu od drugiego z najważniejszych dla pomiaru pojęć, czyli rzetelności. Ocena rzetelności wyników egzaminu zewnętrznego poprzez szacowanie wewnętrznej zgodności testu wydaje się problematyczna. Z psychometrycznego punktu widzenia, im bardziej jednorodny test, tym większy współczynnik zgodności. Z drugiej strony, gdy test sprawdza więcej różnorodnych umiejętności, a więc jest mniej jednorodny, daje uczniom większe szanse uzyskania wyższego wyniku. Egzaminy zewnętrzne, które powinny obejmować próbę całego uniwersum nauczanych umiejętności i wiadomości, z natury rzeczy muszą być niejednorodne, czyli będą się cechować stosunkowo niskim współczynnikiem wewnętrznej zgodności.

Wielu autorów wyróżnia cztery najważniejsze źródła przyczyn obniżających istotnie rzetelność pisemnych egzaminów zewnętrznych:

1. Zmiany z dnia na dzień w predyspozycji uczniów do pisania egzaminu.
2. Różnice w szybkości pracy poszczególnych uczniów.
3. Wariancję wynikającą z cech zastosowanego arkusza egzaminacyjnego, np. sposób, w jaki zadania są prezentowane uczniom.
4. Wzrost wariancji związanej z nieprzestrzeganiem procedury przeprowadzenia egzaminu.

W polskim systemie egzaminów zewnętrznych warto dodatkowo zwrócić uwagę na piąte źródło: wariancję związaną z dokładnością punktowania zadań otwartych przez egzaminatorów zewnętrznych.

Pierwsze źródło wiąże się z egzaminowanym uczniem i jest poza wpływem instytucji egzaminacyjnych. Drugie tylko częściowo zależy od egzaminowanego, ponieważ od autorów arkuszy zależy, czy wykorzystali wszystkie możliwości na etapie tworzenia arkusza egzaminacyjnego, aby czas potrzebny na jego rozwiązanie był stosowny dla jak najszerszego kręgu zdających egzamin. Za trzecie źródło w pełni odpowiedzialni są szeroko rozumiani autorzy arkuszy, począwszy od tworzenia poszczególnych zadań, aż do szaty graficznej i jakości druku. Wzrost wariancji wyników związany z nieprzestrzeganiem procedury przeprowadzenia egzaminu i nieetycznymi zachowaniami uczniów czy nauczycieli jest w naszym kraju szerszym problemem społecznym. Jest szansa, że ostatnie zmiany w prawie oświatowym dotyczące unieważniania egzaminów w przypadku niesamodzielnej pracy uczniów będą miały wpływ na spadek nierzetelności egzaminów powodowanej tym zjawiskiem. Ostatnie, piąte z wymienionych źródeł – dokładność punktowania, która jest tematem tego artykułu, istotnie zależy od umiejętności egzaminatorów i organizacji oceniania zadań otwartych.

Przejdźmy teraz do oceniania rezultatów pracy uczniowskiej na egzaminie. W egzaminach zewnętrznych prace uczniowskie oceniane są przez niezależnych zewnętrznych egzaminatorów przy zastosowaniu procedury oceniania kryterialnego. W tym przypadku najistotniejsze znaczenie dla rzetelności egzaminu (w zakresie zależnym od komisji egzaminacyjnych) ma stabilność posługiwania się kryterium przez oceniających prace uczniowskie egzaminatorów. W praktyce musimy uwzględnić spójność oceny poszczególnych zadań (części zadań odpowiadającym poszczególnym kryteriom), jak i spójność stosowania schematu oceniania dla całego arkusza egzaminacyjnego. Spójność ta zależy od właściwego przygotowania egzaminatorów do pełnienia tej roli, od organizacji oceniania i od jakości moderowania pracy poszczególnych egzaminatorów przez przewodniczącego zespołu egzaminatorów. Nie bez znaczenia jest tu jakość narzędzia pomiarowego, jakim jest arkusz egzaminacyjny wraz ze schematem oceniania.

Badanie spójności stosowania schematu oceniania polega na upewnieniu się, że wszyscy oceniający interpretują kryteria oceniania w taki sam sposób, szczególnie, kiedy ocena polega na jakościowym zakwalifikowaniu czynności mierzonych przez zadanie otwarte i kryterium dopuszcza różne interpretacje jego spełnienia.

Standardowym sposobem oceny rzetelności punktowania jest ocena tej samej pracy przez dwóch oceniających lub ocena po raz drugi tej samej pracy po pewnym czasie przez tego samego oceniającego.

Jest wiele dowodów na to, że oceniający ulegają różnym wpływom, takim jak charakterystyka poszczególnych uczniów, zewnętrzne cechy wypracowania, takie jak

staranność pisma itp. Także płeć, zarówno uczniów, jak i oceniających może mieć istotny wpływ na wyniki. Taka sama odpowiedź z przedmiotów przyrodniczych ma większą szansę otrzymać wyższy wynik, jeżeli autorem jest chłopiec, niezależnie od tego, czy oceniającym jest mężczyzna czy kobieta. Istnieje też przekonanie, że przy ocenianiu prac z przedmiotu istnieje większe prawdopodobieństwo wyższej oceny w przypadku zgodności płci ocenianego i oceniającego. W niektórych przypadkach zdarza się, że egzaminatorzy wnioskują o płci ocenianego na podstawie charakteru pisma lub charakterystycznych błędów. Istnieje więc niebezpieczeństwo świadomego lub nieświadomego zafałszowania wyniku w zależności od płci, pochodzenia lub przekonań ocenianego. U podłoża tych błędów leżą uprzedzenia oceniających. Inne niebezpieczeństwo wiąże się z braniem pod uwagę tych aspektów pracy uczniowskiej, które nie są przedmiotem oceny na danym egzaminie. Na przykład uczniowie brzydko piszący zwykle dostają niższe oceny z przedmiotowego egzaminu, niż uczniowie piszący starannie, choć pismo nie jest przedmiotem oceny. Zdarza się też zaniżenie oceny, gdy egzaminujący nie podziela politycznych, religijnych i kulturowych przekonań, które w pracy prezentuje egzaminowany.

Zarówno szkolenie egzaminatorów, jak i moderowanie oceniania mają na celu zmniejszenie zagrożeń rzetelności oceniania wynikających z wpływu czynników osobistych, kulturowych itp. Musimy pamiętać, że mogą się pojawić podobne zagrożenia dla arkuszy egzaminacyjnych z zadaniami wielokrotnego wyboru i krótkiej odpowiedzi. Tutaj zagrożenia przenoszą się z oceniającego prace na autora zadań. Podsumowując, możemy powiedzieć, że jeżeli oceniający może odczytać z pracy ocenianego płeć, pochodzenie, przekonania polityczne, to istnieje zagrożenie skrzywienia oceniania wskutek istniejących utrwalonych stereotypów. Dążąc do zwiększenia rzetelności danego egzaminu, musimy wziąć pod uwagę wszystkie te aspekty, aby zapewnić uczniom najbardziej sprawiedliwy, jak tylko to jest tylko możliwe, egzamin.

Metody zwiększania zgodności w ocenianiu

Zapewnienie jakości w ocenianiu jest działaniem standaryzacyjnym, a więc działaniem zogniskowanym na procesie oceniania. Warunkiem koniecznym zapewnienia porównywalności jest uzyskanie konsensusu wobec schematu oceniania zadań otwartych w całym kraju. Z drugiej strony, musi być kontrola jakości, która stanowi zapewnienie, że prace uczniowskie są oceniane w porównywalny sposób we wszystkich komisjach i przez wszystkich egzaminatorów. Stąd drugim warunkiem koniecznym zapewnienia jakości jest podobne rozumienie schematu oceniania przez wszystkich egzaminatorów. Obydwa wymienione powyżej procesy mające zapewnić zgodność oceniania nazywają się moderowaniem. W pierwszym przypadku mamy do czynienia z moderowaniem konsensusu, a w drugim – z moderowaniem grupowym zgodności oceniania.

Moderowanie konsensusu oceniania sprawdzianu w sesji wiosennej 2003

W dniu egzaminu odbyły się w Centralnej Komisji Egzaminacyjnej spotkania głównych egzaminatorów ze wszystkich komisji egzaminacyjnych. Uczestnicy spotkania dysponowali już pracami rozwiązanymi przez uczniów na egzaminie. Praktyczne ocenienie szeregu prac z wykorzystaniem autorskiego schematu oceniania dało dobrą podstawę do wyrobienia sobie poglądu na arkusz egzaminacyjny i towarzyszący mu schemat oceniania. Spotkanie to miało doprowadzić do konsensusu w procedurze oceniania i posługiwaniu się kryteriami. Wprowadzane poprawki do kryteriów już w trakcie oceniania masowego świadczą, że zespołowi głównych egzaminatorów CKE i OKE nie zawsze udaje się osiągnąć pełny konsensus. Warto się zastanowić, czy w przyszłości w tym spotkaniu oprócz głównych egzaminatorów nie powinni uczestniczyć jeszcze inni eksperci. W Holandii, gdzie prace maturalne oceniane są przez nauczycieli (dwóch nauczycieli – jeden z własnej szkoły ucznia, drugi obcy), podczas moderowania konsensusu oceniania uczestniczą również przedstawiciele stowarzyszeń nauczycielskich i uczelni.

Zapewnienie jakości w ocenianiu polegające na moderowaniu konsensusu jest działaniem standaryzacyjnym, a więc działaniem zogniskowanym na procesie oceniania. Ustalenia tego zespołu są obowiązujące wszystkich egzaminatorów oceniających dany arkusz egzaminacyjny. Konsensus osiągnięty w zespole głównych egzaminatorów z wszystkich okręgowych komisji i pełne zrozumienie przyjętych kryteriów stanowią warunki konieczne porównywalności oceniania w skali kraju. Z drugiej strony, konieczna jest kontrola jakości, która zapewnia, że prace uczniowskie są oceniane zgodnie z ustalonymi kryteriami. Zapewnienie warunków, aby wszyscy oceniający dobrze zrozumieli i przyjęli ustalenia osiągnięte podczas moderowania konsensusu, zrealizowane zostało poprzez moderowanie grupowe.

Moderowanie grupowe, w zależności od przyjętych w poszczególnych okręgowych komisjach egzaminacyjnych rozwiązań organizacyjnych, obejmowało dwa lub więcej poziomów. W Okręgowej Komisji Egzaminacyjnej w Krakowie, która zatrudniała do oceny prac w wiosennej sesji egzaminacyjnej 2003 przeszło 3 tys. egzaminatorów, zastosowano kaskadowe moderowanie grupowe obejmujące spotkania trzech grup egzaminatorów:

1. koordynatorów ośrodków oceniania
2. przewodniczących zespołów egzaminatorów
3. egzaminatorów.

W przypadku sprawdzianu w OKE w Krakowie funkcjonowało w strukturze organizacji oceniania 6 koordynatorów. Koordynatorzy uczestniczyli w sesji moderowania grupowego prowadzonej przez głównego egzaminatora, który uprzednio uczestniczył w panelu moderowania konsensusu w CKE. Koordynatorzy prowadzili panele dla swoich przewodniczących, a przewodniczący dla swoich zespołów egzaminatorów.

Podczas moderowania grupowego wszyscy egzaminatorzy najpierw oceniali zadania otwarte z tych samych ośmiu prac uczniowskich, a następnie dyskutowali na temat tych prac w kontekście przyjętego schematu oceniania i wyników oceny. Celem takich zajęć było dojście do takiego samego rozumienia kryteriów oceniania, które zostały przyjęte podczas moderowania konsensusu w CKE. Z praktyki OKE w Krakowie wynika, że tylko dogłębne zrozumienie kryteriów pozwalało na swobodne posługiwanie się nim. Słabsi egzaminatorzy, nie do końca rozumiejący kryteria, częściej kurczowo trzymali się literalnego zapisu schematu oceniania, popełniając błędy przy niekonwencjonalnych odpowiedziach uczniów. W trakcie tych spotkań brany był pod uwagę zarówno sam proces, jak i wyniki oceniania uczniowskiej pracy przez wszystkich członków grupy. Organizowanie moderowania grupowego jest przedsięwzięciem kosztownym, ale ma ogromne znaczenie nie tylko dla podniesienia rzetelności oceniania danego egzaminu, lecz także w doskonaleniu nauczycielskich umiejętności w zakresie oceniania.

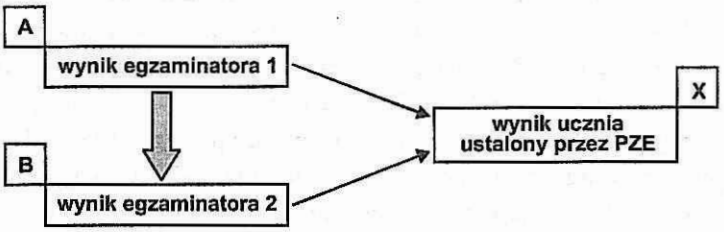
Jeżeli chodzi o organizację oceniania, to okręgowe komisje egzaminacyjne stosowały różne rozwiązania. Na jednym biegunie tych rozwiązań było ocenianie w dużych ośrodkach koordynacji oceniania obejmujących około 200 egzaminatorów zlokalizowanych w jednym ośrodku i korzystających z konsultacji przewodniczącego zespołu egzaminatorów oraz z konsultacji wzajemnej. Na drugim biegunie rozwiązań egzaminatorzy oceniali prace we własnym domu i w razie potrzeby korzystali z telefonicznej konsultacji. Rozwiązania pośrednie polegały na łączeniu oceniania skoszarowanego z ocenianiem w domu, czy też tworzeniem małych ośrodków koordynacji bez konieczności zapewniania egzaminatorom zakwaterowania.

Okręgowa Komisja Egzaminacyjna w Krakowie do oceny prac ze sprawdzianu zorganizowała 6 dużych ośrodków koordynacji, zapewniając egzaminatorom możliwość konsultacji wzajemnej i zakwaterowanie na czas oceniania. Podobny system oceniania zastosowano także dla egzaminu gimnazjalnego.

Zapewnienie kontroli, czy prace uczniowskie są oceniane w porównywalny sposób, realizowane było poprzez podwójne ocenianie losowej próby prac każdego egzaminatora. Zgodnie z ustaleniami CKE dla próby losowej około 10% całej populacji egzaminowanych uczniów prace zostały ocenione przez dwóch egzaminatorów. W OKE w Krakowie, na poziomie sprawdzianu, analizie poddane zostały decyzje prowadzące do ustalenia liczby punktów w przypadku 9568 prac ocenianych przez dwóch egzaminatorów. Karty podwójnego oceniania były skonstruowane w taki sposób, aby można było zarejestrować wynik drugiego egzaminatora oraz pierwotną decyzję pierwszego egzaminatora, jeżeli na karcie odpowiedzi ucznia przewodniczący zmienił wynik w rezultacie analizy podwójnego oceniania. W ten sposób dla wspomnianej próby losowej uczniów zarejestrowano trzy zbiory danych: wynik egzaminatora oceniającego prace (egzaminator 1), wynik oceny prac przez drugiego niezależnego egzaminatora (egzaminator 2, zwany także weryfikatorem) i wynik ustalony przez przewodniczącego zespołu egzaminacyjnego zarejestrowany na karcie odpowiedzi ucznia.

Związki pomiędzy decyzjami podjętymi przez niezależnych egzaminatorów i ustalonym wynikiem dla ucznia przedstawia rysunek 2.

Rysunek 2.
Model zależności pomiędzy decyzjami niezależnych egzaminatorów i ustalonym wynikiem



PZE – przewodniczący zespołu egzaminatorów

Podczas studium przypadku w Ośrodku Koordynacji Oceniania w Rzeszowie obserwowana była także jakość i częstość konsultacji indywidualnych prowadzonych przez przewodniczących zespołów egzaminacyjnych, jak i konsultacji wzajemnych między egzaminatorami. Wyniki analizy wykorzystane zostaną do optymalizacji składów zespołów egzaminacyjnych ze względu na akademicką specjalizację nauczycieli zatrudnionych w roli egzaminatorów oraz do doskonalenia organizacji oceniania.

W całej populacji uczniów przystępujących do egzaminów w szkołach w obszarze działalności OKE w Krakowie losowania próby prac do podwójnego oceniania dokonywali przewodniczący zespołów egzaminatorów. Była to próba prosta prac, dla której sprawdzona została zgodność punktów zapisanych na karcie odpowiedzi ucznia i na karcie oceny drugiego egzaminatora, który nazywany był weryfikatorem. Wylosowane prace oceniane były przez dodatkowego egzaminatora (weryfikatora) zanim zostały wydane egzaminatorom. Weryfikatorzy nie zostawiali żadnej informacji na pracach tak, że egzaminator nie wiedział, które prace podlegały podwójnemu ocenianiu.

Przewodniczący zespołu egzaminatorów, przyjmując od egzaminatora ocenione prace, porównywał wyniki podwójnego oceniania. Decydował, jaka liczba punktów powinna być przyznana uczniowi za zadanie w przypadku rozbieżności w ocenach pierwszego i drugiego egzaminatora. Korzystając z analizy podwójnego oceniania, przewodniczący udzielał informacji zwrotnej każdemu egzaminatorowi na temat poprawności oceniania.

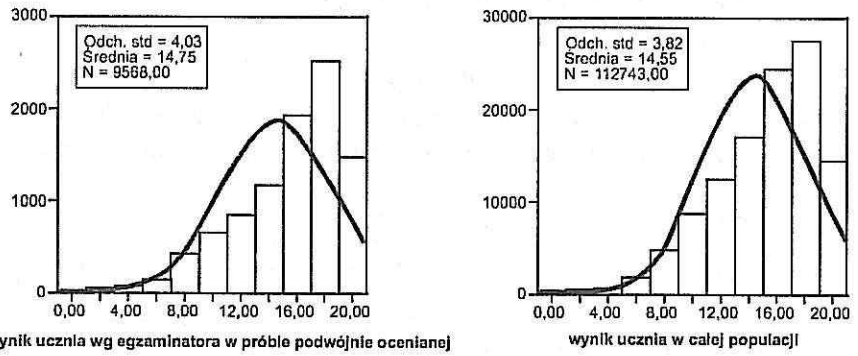
Zanim przejdziemy do analizy wyników podwójnego oceniania porównajmy statystyki dla próby i parametry dla populacji. Poniżej podajemy opis populacji i opis próby uczniów, których prace poddane zostały podwójnemu ocenianiu. W badaniach wykorzystano tylko prace uczniów piszących arkusze A1, A4 i A5.

Tabela 1.
Opis statystyczny próby i populacji uczniów piszących arkusz standardowy A1

Opis		Próba	Populacja
Liczba uczniów		9568	112743
Wynik za zadania otwarte (średnia arytmetyczna)		14,75	14,55
Błąd standardowy średniej		0,041	0,011
Mediana		16	15
Dominanta		17	17
Odchylenie standardowe		4,024	3,825
Wariancja		16,194	14,629
Rozstęp		20	20
Minimum		0	0
Maksimum		20	20
Percentyle	25	12	12
	50	16	15
	75	18	17

Nieco wyższy wynik średni w próbie wynika z wyeliminowania z podwójnego oceniania prac uczniów, którzy nie podjęli rozwiązywania żadnego zadania w części otwartej sprawdzianu.

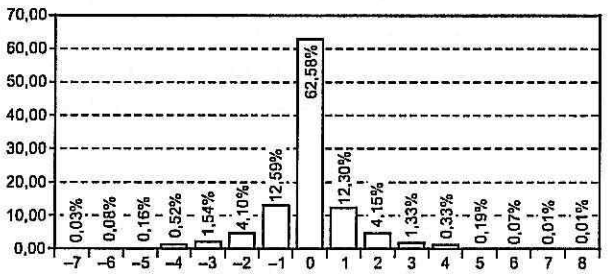
Rysunek 3.
Rozkłady wyników w próbie i populacji



Porównanie wyników całej populacji i próby podwójnie ocenianych prac pozwala stwierdzić, że próba losowa została dobrana właściwie i dobrze reprezentuje populację, co upoważnia do uogólniania wyników na wszystkich uczniów piszących egzamin na podstawie arkuszy zgodnych z A1 w szkołach w obszarze działania OKE w Krakowie.

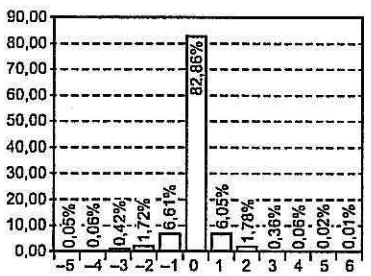
Przyjrzyjmy się teraz różnicom wyniku punktowego dwóch niezależnie oceniających egzaminatorów. Tylko dla 45 prac różnica była większa niż 3 punkty. Dla 62,58% prac obydwaj egzaminatorzy wystawili taki sam wynik za zadania otwarte w sprawdzianie. Warto podkreślić, że podczas oceniania w przypadku zaobserwowanych różnic w ocenie dwóch egzaminatorów przewodniczący zespołu egzaminatorów oceniał dodatkowo pracę danego ucznia i decydował o ostatecznym wyniku. Jednocześnie przekazywał egzaminatorowi informację zwrotną i wskazówki do dalszego oceniania.

Rysunek 4.
Rozkład różnic pomiędzy wynikiem punktowania przez pierwszego egzaminatora i wynikiem drugiego egzaminatora

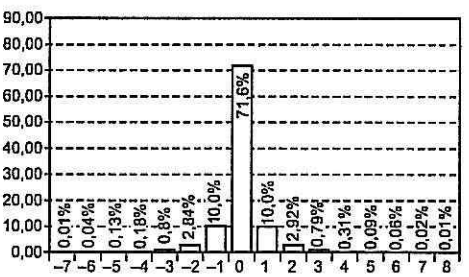


Interesujące jest porównanie wyników uczniów ustalonych przez przewodniczącego zespołu egzaminatorów z liczbą punktów przyznanych przez drugiego egzaminatora – oceniającego indywidualnie i przez pierwszego egzaminatora oceniającego w zespole z zapewnieniem możliwości konsultacji.

Rysunek 5.
Porównanie rozkładów różnic pomiędzy ustalonym wynikiem dla ucznia a punktacją zadań otwartych pierwszego i drugiego egzaminatora



Rozkład różnic pomiędzy wynikiem punktowania przez pierwszego egzaminatora a wynikiem ustalonym przez przewodniczącego zespołu egzaminacyjnego w rezultacie podwójnego oceniania



Rozkład różnic pomiędzy wynikiem punktowania przez drugiego egzaminatora a wynikiem ustalonym przez przewodniczącego zespołu egzaminacyjnego w rezultacie podwójnego oceniania

Jak można odczytać z rozkładu różnic przedstawionych na rysunku 5, zgodność wyniku końcowego ucznia z wynikiem punktowania pierwszego egzaminatora jest bardzo wysoka. Dla 82,86% prac podwójnie ocenianych nie zaobserwowano żadnej różnicy. Tylko dla jednego procenta podwójnie ocenianych prac różnice są większe od 2 punktów. Tylko dla 20‰ analizowanych prac różnice były większe niż 3 punkty. Dla drugiego egzaminatora te różnice są wyższe (dla 71,6% prac nie zaobserwowano różnic w wyniku punktowania).

Musimy pamiętać, że drugi egzaminator nazywany też weryfikatorem i pierwszy egzaminator pracowali w innych warunkach. Drugi egzaminator oceniał prace samodzielnie, raczej nie korzystając z konsultacji przewodniczącego i innych egzaminatorów, podczas kiedy pierwszy egzaminator miał taką sposobność i często z niej korzystał. Spróbujmy się przyjrzeć, który z egzaminatorów (pierwszy czy też drugi) mieli większy wpływ na ustalony wynik ucznia w podwójnym ocenianiu.

Zamieszczony poniżej rysunek 6 pokazuje rozrzut wyników ucznia w zależności od oceniającego pierwszego, czy też drugiego egzaminatora.

Rysunek 6.
Porównanie relacji ustalonego wyniku dla ucznia z wynikiem punktowania przez pierwszego i drugiego egzaminatora

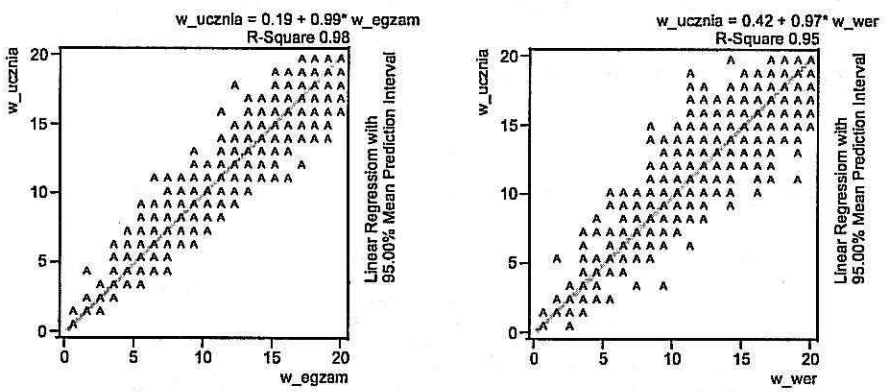


Tabela 2.
Model regresji liniowej zastosowany do wyjaśnienia wpływu decyzji pierwszego i drugiego egzaminatora na wynik ucznia za zadania otwarte

Model regresji liniowej		Unstandardized Coefficients β	Std. Error	Standardized Coefficients Beta	t	Sig.
	(Constant)	3.637E-02	.020		1.845	.065
	punktowanie pierwszego egzaminatora	.670	.005	.671	137.418	.000
	punktowanie drugiego egzaminatora	.327	.005	.329	67.298	.000

Korzystając z wyliczonych w programie SPSS współczynników możemy zapisać:

wynik ucznia = 0,036 + 0,327 × wynik weryfikatora + 0,670 × wynik egzaminatora + ε

Jak można zauważyć $\beta_{\text{punktowania pierwszego egzaminatora}}$ jest prawie dwa razy większy niż $\beta_{\text{punktowania drugiego egzaminatora}}$. Można więc sformułować hipotezę, że ocenianie w zespole z możliwością korzystania z konsultacji ma dodatni wpływ na zwiększenie porównywalności oceniania. Sprawdzenie tej hipotezy wymaga rozszerzenia analizy modelu, a nawet zaplanowania dodatkowych badań.

Wyniki oceny pięciu zadań otwartych przez egzaminatora i weryfikatora posłużyły do szacowania dokładności punktowania wyników przez dwuosobowy zespół sędziów (Niemierko, 1999: 204). Do obliczenia współczynnika dokładności punktowania zastosowany został wzór Rulona.

$$r_{22} = 1 - \frac{sd^2_{A-B}}{sd^2_{A+B}}$$

gdzie:

r_{22} – współczynnik dokładności punktowania przez dwóch sędziów

sd^2_{A-B} – wariancja różnic między ocenami sędziów A i B

sd^2_{A+B} – wariancja sum ocen sędziów A i B.

Sędzia A to drugi egzaminator (weryfikator), który oceniał pracę przed przekazaniem jej pierwszemu egzaminatorowi.

Sędzia B to pierwszy egzaminator.

Tabela 3.
Fragment tabeli wyników podwójnego oceniania sprawdzianu w OKE Kraków

Identyfikator	Weryfikator					Egzaminator					wynik za zadania otwarte		DELTA	SUMA
Kod ucznia	w21	w22	w23	w24	w25	e21	e22	e23	e24	e25	weryfikator	egzaminator		
180101-01P0YA08	1	3	3	3	3	2	3	3	3	3	13	14	–1	27
180101-01P0EA12	1	2	1	2	1	2	1	2	1	2	7	8	–1	15
120108-05P0XA18	2	5	1	9	1	2	5	1	10	2	18	20	–2	38
180407-06P0AA25	0	0	0	2	0	0	0	0	3	1	2	4	–2	6
066201-10P03B01	2	4	1	8	1	2	4	1	9	1	16	17	–1	33
180501-05P0WE01	0	1	0	8	1	0	1	0	8	2	10	11	–1	21
060709-03P0TA08	2	3	0	0	0	2	3	1	0	0	5	5	0	10
060810-01P03A17	1	1	0	5	0	1	2	0	4	0	7	7	0	14
061901-02P0VC07	2	0	0	6	0	2	0	0	6	0	8	8	0	16
126104-09P01A07	2	4	0	6	2	2	4	1	6	2	14	14	0	28
180301-10P0EB04	2	5	0	9	2	2	5	1	9	2	18	18	0	36
180904-06P0FB10	1	5	0	9	1	1	5	1	9	1	16	16	0	32
181101-01P0BE15	2	5	1	9	2	2	5	1	9	2	19	19	0	38
181609-02P0CC10	2	5	1	7	0	2	5	1	7	1	15	15	0	30
181701-04P0YB07	2	5	1	9	2	2	5	1	9	2	19	19	0	38
182004-03P0CC11	2	3	1	7	0	2	3	0	7	0	13	13	0	26
186101-06PZB13	2	5	0	8	2	2	5	1	8	2	17	17	0	34

w21–w25 to liczby punktów przyznane za ocenę zadań od 21do 25 przez weryfikatora, odpowiednio e21–e25 to liczby punktów przyznane za te zadania przez pierwszego egzaminatora (kody ucznia zostały zastąpione fikcyjnymi kodami)

Tabela 4.
Wskaźniki zgodności punktowania zadań dla arkusza S-A1-32 (r_{22} dla dwóch egzaminatorów, r_{11} dla jednego egzaminatora)

Zadanie	r_{22}	r_{11} r-Pearsona
21	0,98	0,96
22	0,98	0,97
23	0,95	0,90
24	0.96	0,92
25	0,91	0,83
Razem zadania otwarte	0,98	0,97

Dokładność punktowania przez jednego sędziego (r_{11}) oszacowano poprzez obliczenie korelacji r-Pearsona punktów przyznanych za ocenę tych samych prac (zadań) przez dwóch egzaminatorów (weryfikatora i egzaminatora). W tym przypadku wyniki uzyskane poprzez każdego z uczniów traktujemy jako pomiar dwukrotny (ocena przez dwóch egzaminatorów). Wyniki surowe wyrażone w punktach przyznanych przez weryfikatora i egzaminatora – jako wyniki odrębnego pomiaru osiągnięć tych samych uczniów.

Uzyskany wskaźnik zgodności okazał się bardzo wysoki. Można przypuszczać, że na jego wysoką wartość miała istotny wpływ organizacja oceniania w zespołach. Podczas oceniania sprawdzianu w ośrodku koordynacji oceniania w Rzeszowie zaobserwowano dużą częstotliwość konsultacji wzajemnych egzaminatorów. Egzaminatorzy zarówno podczas wywiadów indywidualnych, jak i grupowych zwracali uwagę, że zadanie 22 (tekst dostępny na stronach internetowych CKE i OKE) było trudne do oceniania przez polonistów, natomiast zadanie 24 (list) sprawiało trudność w ocenianiu szczególnie matematykom. Z innych uwag można wyróżnić stwierdzenie, że zadanie 23 miało mało precyzyjne polecenie i wielu uczniów prawdopodobnie nie rozumiejąc, o co chodzi, opuszczało je. Podobne stwierdzenie można spotkać także w wypowiedziach uczniów. Można przypuszczać, że ten fakt miał wpływ na obniżenie zgodności oceniania tego zadania. Frakcja opuszczeń tego zadania wynosi 21%, podczas kiedy dla innych zadań otwartych nie przekracza 4,5%. Łatwość tego zadania wynosiła 0,22, co czyni go jednym z najtrudniejszych zadań w całym arkuszu.

Sprawdzian jest egzaminem ponadprzedmiotowym. Naturalna więc wydaje się potrzeba korzystania przyrodników z konsultacji nauczycieli przedmiotów humanistycznych i *vice versa*. Również dla egzaminu gimnazjalnego, dla którego każda z części jest egzaminem ponadprzedmiotowym możliwość korzystania z konsultacji w zespole była wysoko oceniona przez egzaminatorów krakowskiej OKE. Hipoteza, jaka organizacja oceniania jest skuteczniejsza i efektywniejsza, wymaga sprawdzenia podczas sesji egzaminacyjnej w przyszłym roku.

Warto takie badania podjąć, aby patrząc na mapę przedstawioną na rysunku 1 nie mieć wątpliwości, że zaobserwowane zróżnicowanie pomiędzy powiatami i województwami zależy od rzeczywistych terytorialnych różnic w osiągnięciach uczniów.

Bibliografia

Niemierko B. (1999), *Pomiar wyników kształcenia*, WSiP, Warszawa.