UNIWERSYTET SZCZECIŃSKI

NR 58 MATERIAŁY * KONFERENCJE 2000

MARIA KRYSTYNA SZMIGEL
HENRYK SZALENIEC
Okręgowa Komisja Egzaminacyjna
Kraków

Z PRAC NAD PORÓWNYWALNOŚCIĄ WYNIKÓW OCENIANIA ZEWNĘTRZNEGO

1. DLACZEGO WYNIKI OCENIANIA PODCZAS EGZAMINÓW NIE MOGĄ BYĆ PORÓWNYWALNE W TAKIM SAMYM STOPNIU JAK WYNIKI POMIARU FIZYCZNEGO

Do trafnego wnioskowania na podstawie egzaminów o osiągnięciach uczniów jakości programów nauczania i procesu dydaktycznego konieczne jest, aby wyniki egzaminowania były rzetelne w tak wysokim stopniu, w jakim to możliwe. Jednym z warunków uzyskania rzetelności wyników jest dokładne i porównywalne ocenianie prac przez zewnętrznych egzaminatorów [Gipps, 1995]. Musimy jednak zdawać sobie sprawę z tego, że osiągnięcie takiego stopnia porównywalności wyników podczas egzaminów, z jakim mamy do czynienia w pomiarze fizycznym, jest niewyobrażalne.

Pomiar dydaktyczny, będący podstawą do oceniania zewnętrznego, istotnie różni się od pomiaru fizycznego. Ten ostatni, jako nieodłączny atrybut eksperymentu fizycznego, zawsze jest dokonywany podczas interakcji między obiektami mierzonymi, przyrządami (aparaturą pomiarową) i człowiekiem in-

terpretującym wskazania przyrządów pomiarowych. W pomiarze dydaktycznym natomiast badaniu podlegają umiejętności i wiedza (rzadko postawy) człowieka. Są one mierzone w sposób pośredni, na podstawie oceny zachowań w różnych sytuacjach lub ocen wytworów działalności intelektualnej i praktycznej. Rola egzaminatora jest tutaj o wiele szersza niż eksperymentatora przeprowadzającego pomiar fizyczny. Egzaminator nie tylko interpretuje wskazania aparatury pomiarowej – jak na pomiarze fizycznym, ale sam jest częścią, i to bardzo istotną, „aparatury pomiarowej". Fakt ten ujawnia się bezpośrednio przy zastosowaniu na egzaminach zadań otwartych. Podczas egzaminów opartych na zadaniach zamkniętych podobny proces zachodzi na długo przed egzaminem – podczas konstrukcji i standaryzacji testów.

Wykorzystanie do pomiaru dydaktycznego rozwiniętej przez fizyków teorii pomiaru wymaga kilku założeń, które wydają się trudne do spełnienia. Z jednej strony należy zaakceptować postulat, że niepewność pomiarowa w ocenie pochodzi z pomiaru (jest niepewnością przypadkową), co jest równoznaczne z przyjęciem założenia, iż rozbieżności w ocenianiu powstają losowo. Z drugiej strony, wymaga też uznania, że istnieje wynik najbardziej prawdopodobny danej pisemnej pracy (ustnej wypowiedzi) uczniowskiej w takim sensie, w jakim istnieje najbardziej prawdopodobna wartość ciśnienia hydrostatycznego czy ogniskowej soczewki.

Laugier i Weinberg [Noizet, Caverni, 1988] przypuszczają, że ocenianie prac uczniowskich jest porównywalne z pomiarem fizycznym. Według tych autorów, istnieje najbardziej prawdopodobny (oczekiwany) wynik oceny pracy, a zwiększenie liczby pomiarów pozwala na precyzyjniejsze oszacowanie najbardziej prawdopodobnego wyniku. Określili oni optymalną liczbę sędziów kompetentnych dla szkolnych przedmiotów nauczania, powyżej której nie następuje zmiana wartości średniej oceny w wyniku dodania kolejnej oceny. Liczba ta jest różna dla różnych przedmiotów i wynosi odpowiednio dla:

- języka ojczystego (francuskiego) – 78,
- łaciny – 19,
- angielskiego – 28,
- fizyki – 16,
- innych przedmiotów matematyczno-przyrodniczych – 13,
- filozofii – 127.

Szczególne wątpliwości budzi pierwszy postulat, zakładający, że rozbieżności wyników oceniania tej samej pracy przez różne osoby powstają przypadkowo (losowo). Gdyby postulat ten przyjąć za prawdziwy, to występowanie rozbieżności nie wymagałoby szczególnych wyjaśnień.

Jak już wspomniano, Laugier i Weinberg stwierdzili, że przy ustalaniu najbardziej prawdopodobnej wartości pracy uczniowskiej z języka francuskiego (języka ojczystego) trzeba obliczyć średnią z 78 wyników oceniania. Ze względów ekonomicznych i praktycznych takie postępowanie jest oczywiście niemożliwe. Kiedy jednak obliczono średnią z ocen wystawionych przez 76 oceniających, okazało się, że przy poziomie istotności $\alpha = 0{,}05$ w przedziale ufności znalazły się – wbrew oczekiwaniom – wyniki nie 95 % oceniających, lecz tylko 55 %. Otrzymany rezultat nie daje się pogodzić z postulatem o przypadkowym pochodzeniu rozbieżności w ocenach tego samego zadania. Inny przykład możemy zaczerpnąć z warsztatów oceniania kryterialnego prac z pilotażu „Egzamin 2002". Podczas zajęć koordynacyjnych sprawdzania prac uczniowskich kandydaci na egzaminatorów przed przystąpieniem do oceny wypracowań zapoznali się z trzema wypełnionymi arkuszami egzaminacyjnymi. Wszyscy sprawdzali te same prace uczniowskie, skopiowane na użytek warsztatów. Różnice w ocenianiu pierwszych trzech prac wyniosły – pomimo przygotowania szczegółowego schematu oceniania i jego analizy przez zespół oceniających – 10 punktów na 60 możliwych do uzyskania. Różnice te najczęściej zwiększały się w odniesieniu do czterech lub pięciu zadań albo ich części. Niektóre umiejętności były więc oceniane zgodnie, natomiast za inne (widocznie trudniejsze do oceny) uczestnicy warsztatów przyznawali w tym samym zadaniu wszystkie możliwe wartości skali. W ocenianiu brało udział 9 osób. W tabeli 1 przedstawiono wyniki oceniania jednej z prac uczniowskich. Rozkład wyników (rys. 1) jest daleki od rozkładu normalnego czy rozkładu Studenta. Rozstęp sięga 10 punktów, co stanowi 17 % punktów możliwych do uzyskania.

Niepewność pomiarowa pojedynczego egzaminatora wynosi więcej niż ±3 punkty. Niepewność pomiarowa średniej ±1,15. Przy poziomie istotności $\alpha = 0{,}05$ przedział ufności dla oceny pracy przez dziewięciu egzaminatorów wynosi

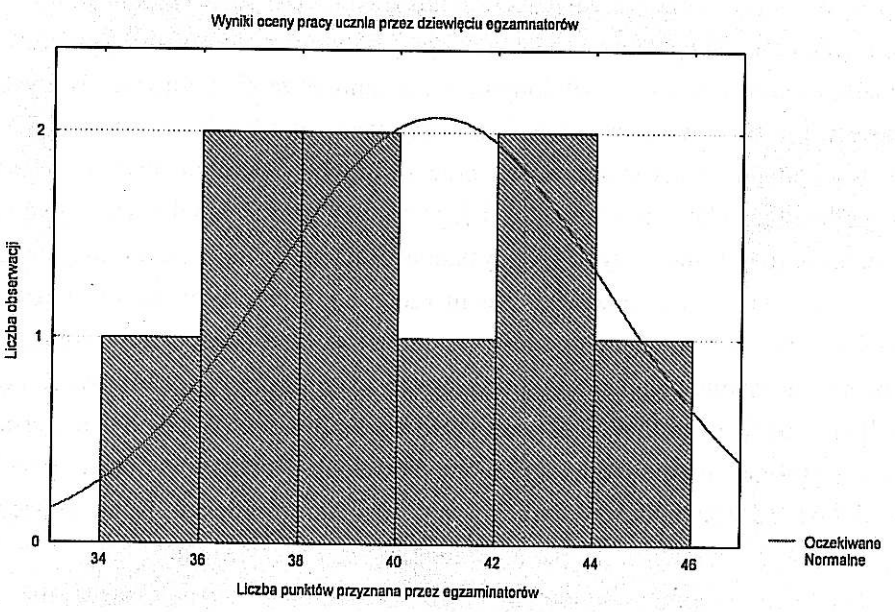
$$40{,}78 - 2{,}30 \cdot 1{,}15 = \mathbf{37{,}33} < \overline{X} > \mathbf{44{,}23} = 40{,}78 + 2{,}30 \cdot 1{,}15.$$

Tabela 1

Opis statystyczny wyników oceniania pracy ucznia z pilotażu egzaminu gimnazjalnego z przedmiotów matematyczno-przyrodniczych przez dziewięciu egzaminatorów

Zmienna	Średnia arytmetyczna	Błąd standardowy średniej	Odchylenie standardowe	Wariancja	Minimum	Maksimum
Wynik punktowy ucznia	40,78	1,15	3,46	11,97	36	46

Źródło: opracowanie własne.



Rys. 1. Rozkład wyników oceniania pracy uczniowskiej przez dziewięciu egzaminatorów
Źródło: opracowanie własne.

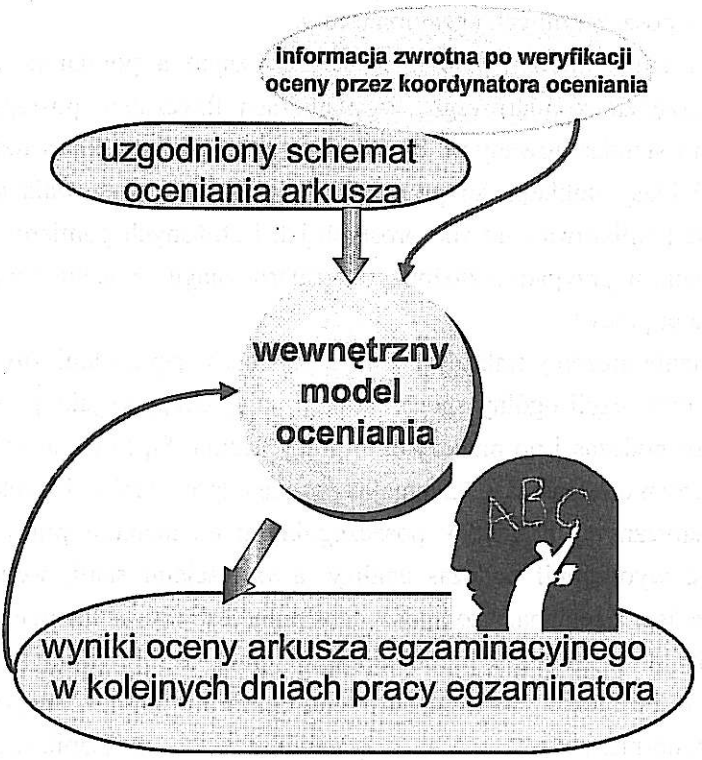
Przedział ufności 44,23–37,33 dla opisanego przypadku zawiera 67 % wyników oceniania jednej pracy uczniowskiej. Po ocenieniu pięciu prac uczniów i ostatecznym sprecyzowaniu schematu oceniania zmniejszyła się rozbieżność wyników. W praktyce jednak uczniowskie prace są oceniane przez jednego egzaminatora, a nie przez kilku, co oznacza, że wykonywany jest pojedynczy pomiar o mniejszej dokładności niż w opisanym przypadku.

Zarówno w przykładzie z krakowskiego pilotażu jak i na podstawie cytowanych badań [Noizet, Caverni, 1988] można przyjąć tezę, że występują tu zjawiska, które powodują trudne do wyeliminowania przyczyny **błędów systematycznych**, takie jak na przykład systematyczne zawyżanie lub zaniżanie oceny przez poszczególnych egzaminatorów.

Inna różnica między pomiarem dydaktycznym a pomiarem fizycznym dotyczy narzędzia pomiarowego. W pomiarach fizycznych posługujemy się przyrządami standaryzowanymi, dla których jest określona klasa dokładności. Znajomość klasy dokładności przyrządów pomiarowych pozwala oszacować niepewność pomiarową zarówno prostych jak i złożonych pomiarów. Jak wygląda sytuacja w przypadku złożonego pomiaru osiągnięć ucznia na podstawie jego pisemnej pracy?

Ocenianie możemy traktować jako czynności w psychologicznym znaczeniu tego słowa, czyli ogólny sposób reakcji na sytuację, w jakiej znajduje się egzaminator podczas i po przeczytaniu pracy ucznia. Są to czynności ewaluacyjne. Na czym one polegają? Oceniający szuka odpowiedniości (funkcji w sensie matematycznym) pomiędzy poszczególnymi elementami pracy uczniowskiej, które wyodrębnił podczas analizy, a wartościami skali, według której dana praca jest oceniana. Czynność oceniania jest oparta na pewnej liczbie wskaźników dotyczących różnych aspektów pracy uczniowskiej i zapisanych w schemacie oceniania. Schemat oceniania egzaminu gimnazjalnego w części matematyczno-przyrodniczej jest przykładem takich wskaźników. Nadanie wartości tym wskaźnikom dla pracy konkretnego ucznia prowadzi do podjęcia decyzji o ocenie. Wynika stąd, że czynność oceniania osiągnięć szkolnych musi być rozpatrywana w szerszej klasie zachowań niż tylko percepcyjna i poznawcza. Ocenianie to uzyskiwanie informacji i podejmowanie decyzji na podstawie tych informacji. Zachowania poznawcze prowadzą do porównywania elementu lub całej pracy ucznia z modelem odniesienia, który jest wpisany

w struktury poznawcze egzaminatora. Można więc sobie wyobrazić operację oceniania jako czynności egzaminatora, który w wyniku porównania ustala relacje pomiędzy wytworem pracy intelektualnej ucznia a własnym modelem odniesienia, czyli modelem subiektywnym. Model ten ukształtował się u danego egzaminatora o wiele wcześniej, niż on sam podjął się oceniania danej pracy. Ponadto ulega on ciągłej modyfikacji zarówno na etapie analizy schematu oceniania, jak i podczas oceniania kolejnych prac. Rysunek 2 ilustruje funkcjonowanie modelu odniesienia, jakim dysponuje każdy z egzaminatorów.



Rys. 2. Schemat funkcjonowania subiektywnego modelu oceniania

Źródło: opracowanie własne.

Swoje ABC w zakresie oceniania nauczyciel zdobywa najczęściej w trakcie praktyki szkolnej. Jest to jego uogólniony, wewnętrzny, subiektywny model odniesienia, na którego podstawie kreuje model oceniania pracy z danego egzaminu. Uzgodniony w zespole sędziów kompetentnych schemat oceniania danego arkusza egzaminacyjnego ma istotne znaczenie dla zmian modelu odniesienia. Od zgodności schematu oceniania z modelem odniesienia zależy, w jakim stopniu i jak szybko schemat oceniania zostanie przyjęty przez egzaminatora.

Uwagi i informacje, jakich dostarcza egzaminatorowi koordynator oceniania, mogą wzmocnić i istotnie przyspieszyć asymilację uzgodnionego schematu oceniania. Niebagatelny wpływ na modyfikację wewnętrznego modelu oceniania ma liczba i jakość prac, które egzaminator ocenia w pierwszym, a potem w kolejnych dniach pracy.

2. JAKA JEST STRUKTURA MODELU? JAK FUNKCJONUJE MODEL?

Podstawowym składnikiem modelu oceniania jest idealne wypracowanie lub zbiór odpowiedzi na pytania zawarte w teście. Jeżeli zadaniem ocenianego jest na przykład zrozumienie ze słuchu tekstu w języku obcym, to produktem idealnym będzie zbiór poprawnych odpowiedzi na pytania zadane do tekstu; z matematyki czy fizyki będzie to bezbłędne pod każdym względem rozwiązanie zadania. Zdefiniowanie idealnej odpowiedzi może się okazać – w zależności od przedmiotu nauczania i badanej umiejętności – łatwiejsze albo trudniejsze. Należy więc przypuszczać, że idealne odpowiedzi na dane pytanie mogą się nieco różnić między sobą w zależności od egzaminatora. W czasie oceniania egzaminator odwołuje się zarówno do oczekiwanej odpowiedzi, określonej w schemacie oceniania, jak i do własnego rozwiązania (odpowiedzi na pytanie), które uznał za idealne. Ma także wyobrażenie oczekiwanej odpowiedzi w zależności od tego, z jakimi uczniami pracuje w szkole, od własnej oceny trudności zadania. Z badań, które prowadzono w siódmej klasie, wynika, że oszacowany przez nauczycieli stopień trudności zadań odbiega od rzeczywistych trudności określonych na podstawie wyników badań. Trzecim elementem modelu odniesienia jest skala pomiaru. Doświadczenia wskazują, że w trakcie oceniania

skala ta ulega ciągłej modyfikacji. Praca oceniona jako pierwsza i ponownie oceniona – przez tego samego egzaminatora – jako ostatnia zwykle uzyskuje inny wynik. Podczas aktu oceniania model oceniania nie jest więc niezmienny – przeciwieństwie do pomiaru fizycznego.

Poznanie przyczyn rozbieżności ma fundamentalne znaczenie dla doskonalenia procedur oceniania zewnętrznego. Jednym ze sposobów wyjaśnienia tych przyczyn jest odwołanie się do różnic osobowościowych poszczególnych egzaminatorów. Na przykład możemy mówić o egzaminatorach:

- surowych lub pobłażliwych,
- analitycznych lub syntetycznych,
- stałych i niestałych w trzymaniu się przyjętej skali.

Inna próba wyjaśnienia kładzie akcent na wpływ różnych warunków, w których egzaminator dokonuje oceniana. Podczas oceny tej samej pracy z egzaminu matematyczno-przyrodniczego ujawniły się różnice wynikające z doświadczenia przedmiotowego poszczególnych egzaminatorów: to samo zadanie uzyskiwało różną ocenę u fizyka, chemika, biologa czy matematyka. Dopiero w wyniku długich dyskusji nad schematem oceniania i negocjacji udało się uzyskać istotne zmniejszenie rozbieżności. Podobne zjawisko zaobserwowano podczas oceniania prac z ponadprzedmiotowego sprawdzianu na zakończenie szóstej klasy szkoły podstawowej.

3. *KLAROWNOŚĆ SCHEMATU OCENIANIA WARUNKIEM KONIECZNYM PORÓWNYWALNOŚCI OCENIANIA*

Innym bardzo istotnym czynnikiem, decydującym o zgodności wyników ocenienia tej samej pracy przez różnych oceniających, jest jakość testu i klarowność schematu oceniania. Wszystkie niedociągnięcia konstrukcyjne narzędzi mają ogromny wpływ na zróżnicowanie ocen, jakie kompetentni sędziowie przyznają za rozwiązanie tego samego zadania. Duża rozbieżność ocen z tego samego zadania jest sygnałem, by zweryfikować schemat oceniania. Konieczne jest więc przeprowadzenie zajęć poprzedzających ocenianie. Powinny one uwzględnić weryfikację zaproponowanego przez autorów testu schematu oceniania w kontekście uczniowskich wypracowań. Poniżej zamieszczamy scena-

riusz zajęć koordynacji oceniania, które odbyły się w Krakowie nazajutrz po badaniach kompetencji w maju 2000 roku. Pierwszy etap zajęć miał na celu doskonalenie schematu oceniania. Drugi etap prowadził do modyfikacji wewnętrznego modelu oceniania każdego z egzaminatorów w taki sposób, by identyfikowali się oni z przyjętym schematem oceniania i aby prace 54 000 uczniów były oceniane w możliwie najbardziej porównywalny sposób.

4. PROCEDURA KOORDYNACJI OCENIANIA PRAC UCZNIOWSKICH Z BADAŃ KOMPETENCJI

Dla zapewnienia porównywalności oceniania we wszystkich 30 komisjach wprowadzono dwuetapowe koordynowanie oceniania prac. Pierwszy etap to warsztaty, które miały na celu uzgodnienie i ostateczne sprecyzowanie schematu oceniania przez przewodniczących komisji oceniających oraz przyjęcie jednolitej procedury oceniania we wszystkich komisjach. Drugi etap obejmował: warsztaty koordynacyjne dla członków każdej komisji, powtórną ocenę 10 prac każdego członka komisji oceniającej przez przewodniczącego oraz bieżącą informacją zwrotną – mającą na celu zapewnienie porównywalności i jakości oceniania.

Zajęcia warsztatowe pierwszego etapu koordynacji obejmowały w szczególności:

1. Rozwiązanie testu uczniowskiego przez wszystkich oceniających.
2. Omówienie schematu oceniania zaproponowanego przez autorów testu.
3. Omówienie sposobu sporządzenia protokołu oceny i odnotowywania wyniku na kartach do czytnika zaznaczeń.
4. Ocenę pierwszej pracy przez każdego oceniającego:
   - sporządzenie protokołu oceny,
   - przeniesienie wyniku na „pasek wynikowy", umożliwiający porównanie rezultatów oceny.
5. Przygotowanie zestawienia wyników oceniania pracy pierwszego ucznia.
6. Analizę zgodności oceniania.
7. Uszczegółowienie (ewentualna korekta) schematu oceniana.

8. Ocenę kolejnych prac (drugiego i trzeciego ucznia) przez każdego z oceniających.

9. Przygotowanie zestawienia wyników oceniania (drugiego i trzeciego ucznia).

10. Analizę zgodności oceniania prac drugiego i trzeciego ucznia.

11. Zatwierdzenie finalnej wersji schematu oceniana.

12. Ćwiczenia w wypełnianiu kart zaznaczeń.

13. Przygotowanie kopii schematu oceniania dla każdego przewodniczącego komisji.

14. Przygotowanie zestawu materiałów do zajęć wprowadzających do koordynacji oceniania w poszczególnych komisjach:

- kompletu testów,
- kompletu kopii prac (pierwszego, drugiego i trzeciego ucznia),
- protokołów oceny,
- „pasków wynikowych" do ćwiczeń porównywalności oceniania.

Drugi etap koordynacji w pierwszej części był powtórzeniem warsztatów, w których uczestniczyli przewodniczący komisji. Różnica polegała tylko na tym, że nie mógł już podlegać modyfikacji schemat oceniania. Wprowadzenie jakiejkolwiek zmiany w schemacie prowadziłoby do błędów systematycznych i zmniejszenia porównywalności oceniania.

Warsztaty kończyły się przydziałem prac do oceny. W ciągu kolejnych dni, w razie wątpliwości, oceniający mieli sposobność konsultacji z przewodniczącym komisji trudnych decyzji przydziału punktów za rozwiązanie poszczególnych zadań.

Przewodniczący nie oceniali prac w takim stopniu jak członkowie komisji. Mieli oni natomiast obowiązek (co drugi dzień) oceny dwóch wylosowanych prac od każdego z członków komisji. Przekazywane uwagi dotyczące konsekwencji stosowania schematu oceniania i ewentualne rozbieżności w ocenie były rejestrowane w protokole oceniającego wraz z pisemnym podpisem przyjęcia do wiadomości. W przypadku zakwestionowania wyniku po ponownej ocenie była sporządzana nowa karta do czytnika zaznaczeń.

PODSUMOWANIE

Badania kompetencji uczniów ósmych klas w województwie małopolskim pozwoliły sprawdzić empirycznie wiele procedur, mających zwiększyć porównywalność oceniania zewnętrznego. Podczas badań po raz pierwszy wykorzystano prognozowane przez nauczycieli wyniki każdego ucznia do wykrywania ewentualnych niedoskonałości oceny. Ponad trzysta prac uczniowskich zostało skierowane do ponownej oceny, aby sprawdzić, czy rozbieżność między wynikiem i prognozą nie jest spowodowana błędem oceniającego. Zarejestrowane opinie ponad siedmiuset nauczycieli uczestniczących w ocenianiu zewnętrznym stanowią cenne wskazówki do szkolenia przyszłych egzaminatorów.

LITERATURA

Gipps C.: *Beyond Testing; Toward a Theory of Educational Assessment.* The Falmer Press, London 1995.

Noizet G., Caverni J.P.: *Psychologiczne aspekty oceniania osiągnięć szkolnych.* PWN, Warszawa 1988.

Szaleniec H., Szmigel M.K.: *Informator [OKE Kraków] o wynikach badań kompetencji uczniów klas ósmych 2000.* Cz. I. Wydawnictwo Szkolne Omega, Kraków 2000.