ROMAN SOKULSKI
Olsztyn

ZAMIERZONE I NIEZAMIERZONE SKUTKI LEKCEWAŻENIA BŁĘDU POMIARU NA PRZYKŁADZIE MATURY Z „JĘZYKA POLSKIEGO”

Do napisania tego artykułu skłoniła mnie pewna postawa Centralnej Komisji Egzaminacyjnej, którą mogę zrozumieć, ale której nie mogę akceptować. Polega ona na unikaniu, pomijaniu, lekceważeniu informacji dotyczących oceny testu na poziomie błędu pomiaru. Czasami mam nawet wrażenie, że problem ten dla Komisji Egzaminacyjnych nie istnieje. Ta zmowa milczenia przypomina trochę antropologiczne funkcjonowanie tabu, które opiera się w społeczeństwach pierwotnych na zakazie wykonywania pewnych czynności, używaniu albo dotykania pewnych przedmiotów, wymawianiu pewnych słów[1].

W takim razie „co to za diabeł”, którego teraz niczym królika wyciągam z kapelusza? Bolesław Niemierko wyróżnił następującą hierarchię niezbędnych właściwości pomiaru dydaktycznego: niezależność sytuacji pomiarowej → dokładne punktowanie zadań → rzetelność pomiaru → trafność pomiaru → obiektywizm pomiaru[2]. Wszystkie wymienione powyżej właściwości pomiaru (w podanej kolejności) służą zapewnieniu obiektywizm pomiaru dydaktycznego.

Wydaje się, że w polskim systemie egzaminów zewnętrznych zabiega się przede wszystkim o niezależność sytuacji pomiarowej i dokładne punktowanie (z różnym skutkiem w przypadku przedmiotów humanistycznych). Jednak właściwości te nie są związane z błędem pomiaru, który rozpatrywany jest w zakresie rzetelności i trafności pomiaru dydaktycznego. Profesor Niemierko podkreśla, że wartość nominalna wyniku testowania (w teorii rzetelności: wynik prawdziwy pojedynczego ucznia) nie jest ustalona i może być tylko oszacowana, poprzez uogólniony *błąd standardowy*, na podstawie wyników otrzymanych, obciążonych

---

[1] Kopaliński W., *Słownik mitów i tradycji kultury*, Państwowy Instytut Wydawniczy, Warszawa 1987, s. 1163.

[2] Niemierko B., *Pomiar wyników kształcenia*, WSiP, Warszawa 1999, s. 170.

błędami pomiaru[3]. Podkreślmy to, iż **otrzymany wynik pomiaru obciążony jest mniejszym lub większym błędem**, który szczególnie w przypadku nauk społecznych **powinien być poddawany kontroli, oszacowany**.

W pomiarze różnicującym, w którym układem odniesienia wyniku każdego ucznia są wyniki innego ucznia, oszacowanie błędu pomiaru jest szczególnie istotne. Egzaminy zewnętrzne są bez wątpienia pomiarem różnicującym, dlatego dziwi mnie pomijanie w pracach komisji Egzaminacyjnych problematyki związanej z błędem pomiaru. Peter T. Knigt podkreśla, że wysoka rzetelność jest szczególna ważna w ocenianiu doniosłym[4], a matura z „języka polskiego" jest bez wątpienia egzaminem doniosłym.

Upraszczając trochę omawianą tu problematykę- wszelkie zróżnicowanie nie wynikające z rzeczywistych różnic między mierzonym właściwościami możemy w naukach społecznych nazwać błędem pomiaru[5]. Jak przedstawiają to zagadnienia specjaliści przedmiotowi? Alan Davis w artykule *Konstruowanie testów językowych* zaleca, aby przed przystąpieniem do oceny wyników zapewnić ich rzetelność, gdyż wyniki o niskiej rzetelności nie mogą mieć żadnego znaczenia prócz własnej przypadkowości[6].

Dlaczego więc nauczyciele mieliby przyjmować wyniki egzaminów zewnętrznych jako podstawę do podejmowania decyzji o programie kształcenia, doborze metod i podnoszenia jakości pracy szkoły? Na jakiej podstawie oceniani są przez nadzór dyrektorzy szkoły, nauczyciele? Skoro nie wiemy nic o rzetelności wyników egzaminów zewnętrznych, nie możemy też nic powiedzieć o ich użyteczności, czyli trafności, nie mówiąc już o obiektywizmie, co należy przecież do elementarnej kultury oceniania.

Czy polskiej szkole jest potrzebna kultura oceniania? Judith Marquand, sporządzając raport *Studium wstępne krajowego systemu oceniania w polskim szkolnictwie ponadpodstawowym*, wyróżniła następujące składniki systemu oceniania egzaminów: 1. ustalenia kryteriów ocen, 2. przygotowanie egzaminów, 3. wyszkolenie oceniających, 4. nadanie uprawnień oceniającym, 5. kierowanie przebiegiem egzaminów, 6. ocenianie egzaminów, 7. zapewnienie porównywalności kryteriów oceniania, 8. ocenianie rezultatów i zapewnienie informacji zwrotnej zarówno w skali lokalnej, jak regionalnej oraz w skali całego kraju. Zaznaczając jednocześnie, że system obecny (w 1993 r.) realizuje punkty 1., 4., 7. i 8. w stopniu niewystarczającym[7]. Dziś, po 12 latach, w zakresie punktu 1. cała praca Programu NOWA MATURA nad ustaleniem kryteriów ocen została zanegowana.

---

[3] Ibidem, s. 194.

[4] Knight P.T., *The Value of Programme-wide Approach to Assessment*, „Assessment & Evaluation in Higher Education" 2000, No 3.

[5] Franfort-Nachmias Ch., Nachmias D., *Metody badawcze w naukach społecznych*, Zysk i S-ka, Poznań 2001, s. 179.

[6] Davis A. *Konstruowanie testów językowych*, [w:] J.P.B. Allen, S. Pit Corder (red.), *Kurs edynburski językoznawstwa*, PWN , Warszawa 1983, s. 300.

[7] Marquand J., *Studium wstępne krajowego systemu oceniania w polskim szkolnictwie ponadpodstawowym*, Biuro koordynacji kształcenia kadr, Warszawa 1993, s. III.

Nadawanie uprawnień (punkt 4.) przerosło trochę system zewnętrznego oceniania. Porównywalność ocenienia (punkt 7.) i ocenianie rezultatów (punkt 8.) są, moim zdaniem, nadal realizowane w stopniu niewystarczającym. To smutne wyliczenie można podsumować stwierdzeniem J. Marquand: *W Polsce istnieje możliwość dalszego rozwinięcia zastosowania testów jako instrumentów diagnostycznych, prowadzących do doskonalenia uczenia się, nauczania i organizacji szkół*[8].

Jednak czy samo powołanie komisji egzaminacyjnych i przeszkolenie egzaminatorów spowoduje rozwinięcie zastosowania testów jako instrumentów diagnostycznych? Z pewnością jest to warunek niewystarczający. Dlaczego? Przytoczę tu stwierdzenie pewnej pani profesor, które wydaje mi się tu szczególnie aktualne: *Akademicy są bliżej artystów niż urzędników. Dopóki są, dotąd nauka się rozwija*[9]. Moim zdaniem, jeżeli Polskie Towarzystwo Diagnostyki Edukacyjnej zamknie się w kręgu problemów urzędniczych, przestanie rozwijać się diagnostyka edukacyjna.

1. Badania nad określeniem rzetelności egzaminu maturalnego z „języka polskiego”

Postawa CKE i OKE, dotycząca procedury tworzenia testów i ich oceny, przypomina postawę *Nauczyciela-Strażnika Wielkiej Tajemnicy i Wielbiciela Cenzury Informacji: „Jaką dostałeś ocenę? A to już moja sprawa!”*[10] Brakuje planów testów, nie mówiąc już o kartotekach, nie przedstawia się też sposobu poprawy zadań w procesie standaryzacji, nic nie mówi się także o ich rzetelności, jakby ona nie istniała. Jak zatem rozwija się sposób myślenia o jakości egzaminów zewnętrznych? Henryk Szaleniec z Okręgowej Komisji Egzaminacyjnej w Krakowie napisał w artykule *Od klasycznej teorii testu do probabilistycznych modeli zadań testowych*, że jakość egzaminów zewnętrznych w dużym stopniu zależy od tego, czy potrafimy przygotować dobre narzędzia pomiaru, czyli arkusze egzaminacyjne. Trudno się z tym stwierdzeniem nie zgodzić. Zastanawiać może tylko następne zdanie: *Jeżeli ze względu na konieczność zachowania niejawności zadań nie można prowadzić standaryzacji na odpowiednio dobranych próbach uczniów, to zachodzi konieczność tworzenia banków zadań, które w przyszłości pozwolą tworzyć narzędzia z zadań już sprawdzonych podczas ich wcześniejszych zastosowań*[11]. Czy oznacza to, że nie prowadzi się procesu standaryzacji testów przygotowywanych przez Centralną i Okręgowe Komisje Egzaminacyjne, a dopiero eksperymentuje z probabilistycznymi modelami zadań testo-

---

[8] Ibidem, s. 16.

[9] Kostera M., Rosiak A., *Zajęcia dydaktyczne*, Gdańskie Wydawnictwo Psychologiczne, Gdańsk 2005, s. 171.

[10] Taraszkiewicz M., *Wprawki z reformy... oceniania*, „Nowa Polszczyzna” 1998, nr 5, s. 44.

[11] Szaleniec H., *Od klasycznej teorii testu do probabilistycznych modeli zadań testowych*, [w:] B. Niemierko (red.), *Ocenianie szkolne. Ekonomika i polityka oświatowa. Probabilistyczne modele pomiaru*, Międzywydziałowe Studium Pedagogiczne Uniwersytetu Gdańskiego, Gdańsk 2002, s. 85.

wych? Czy powodem zaniechania standaryzacji jest obawa o zachowanie niejawności? Wygląda na to, że zrezygnowano z kontroli błędu pomiaru w imię zachowania tajemnicy niejawności zadań. Na jakiej więc podstawie oceniana jest praca szkół i nauczycieli oraz na jakiej podstawie wydawane są świadectwa maturalne, stanowiące warunek przyjęcia na uczelnie wyższe?

Przyjrzyjmy się badaniom, zmierzającym do określenia wielkości błędu pomiaru, w przypadku egzaminu maturalnego z języka polskiego.

W biuletynie informacyjnym *Przed maturą 1999. Język polski*, opracowanym w Centrum Edukacji Nauczycieli w Gdańsku, I. Majkowska i B. Niemierko przedstawili artykuł *Co warta jest wielokryterialna ocena pisemnych prac maturalnych?*[12], w którym opisany został eksperyment, przestawiający analizę wielokryterialnej oceny pisemnych prac maturalnych z języka polskiego.

W postępowaniu badawczym wylosowano prace z matury próbnej i regulaminowej z 1996 i 1997 r., które następnie sprawdziła 9-osobowa grupa polonistów współpracujących z Centrum Edukacji Nauczycieli w Gdańsku. Prace sprawdzane były bez zastosowania schematu punktowania oraz z zastosowaniem schematu punktowania Emilii Kuczyńskiej[13], przyjętego przez Regionalny Zespół NOWA MATURA jako obowiązujący w regionie schemat kryterialnego oceniania prac pisemnych z języka polskiego.

Badanie wykazało, że zadawalającą pewność (rzetelność na poziomie 0,81–0,85) ocen prac pisemnych z języka polskiego uzyskano przy trzech osobach sprawdzających jedną pracę. Przy jednym sprawdzającym rzetelność była niezadawalająca (wynosiła od 0,59–0,66). Ważnym wnioskiem z tego badania było też osłabienie mitu o potędze pomiarowej kryteriów oceny wypracowań i schematów punktowania[14], z którym tak wielkie nadzieje wiązali twórcy reformy edukacyjnej w Polsce i co wynikało z rekomendacji J. Marquand. W zakończeniu artykułu zwrócono uwagę na potrzebę systematycznej ewaluacji wyników uzyskanych w trakcie prac nad reformą egzaminu maturalnego z języka polskiego. Wydaje się dziś, że niezależne komisje egzaminacyjne nie sprostały temu zadaniu.

W 2002 r. Instytut Spraw Publicznych przeprowadził badanie rzetelności oceny prac z matury próbnej. Jednym z badanych przedmiotów był język polski, którego ocenę arkusza egzaminacyjnego R. Dolata uznał za priorytetową ze względu na problematykę rzetelności systemów punktacji zadań otwartych[15]. Do oceny rzetelności użyto analizy wariancji, która, jak podkreśla B. Niemierko, daje

---

[12] Niemierko B., Majkowska I., *Co warta jest wielokryterialna ocena pisemnych prac maturalnych?*, [w:] I. Pancer (red), *Przed maturą 1999.Język polski,* Centrum Edukacji Nauczycieli, Gdańsk 1999, s. 34–45.

[13] Kuczyńska E., *Komentarz do kryteriów oceny pisemnych prac maturalnych z języka polskiego*, [w:] I. Pancer (red), *Przed maturą 1997*, op. cit., s. 42

[14] Ibidem, s. 43–45.

[15] Dolata R., *Wybrane problemy rzetelności i trafności nowej matury*, [w:] B. Niemierko, H. Szaleniec (red.), *Standardy wymagań i normy testowe w diagnostyce edukacyjnej,* Polskie Towarzystwo Diagnostyki Edukacyjnej, Kraków 2004, s. 61.

możliwość głębszego poznania warunków dokonanego pomiaru i skuteczniejszego projektowania jego ulepszeń[16].

Współczynnik rzetelności dla arkusza I (reprezentującego test sprawdzający umiejętność czytania w zakresie odbioru tekstu publicystycznego) oszacowany został na poziomie 0,55[17], czyli jeszcze niższym niż w przypadku badania rzetelności ocenienia kryterialnego wypracowania z języka polskiego. Tak duża wartość błąd pomiaru powinna niepokoić. A już na pewno nie można na podstawie takiego poziomu rzetelności rozważać trafności diagnostycznej wyników arkusza I.

Co w takim przypadku należałoby zrobić? Instytut Spraw Publicznych opracowuje koncepcję reformy egzaminu maturalnego, dostrzegając niską przydatność istniejącej formy egzaminowania dla rozwoju polskich szkół ponadgimnazjalnych. A co robią w tym czasie komisje egzaminacyjne? Proponują następne matury próbne i regulaminowe przeprowadzane przy użyciu nierzetelnych narzędzi pomiaru. Czy jest to sąd bezpodstawny?

Jako osoba odpowiedzialna w Zespole Szkół Gastronomiczno-Spożywczych w Olsztynie za przygotowanie klas trzecich technikum do egzaminu maturalnego w roku szkolnym 2005/2006, zdecydowałem się na wykorzystanie materiałów z próbnego egzaminu maturalnego z języka polskiego, przeprowadzonego w listopadzie 2004 r. w województwie warmińsko-mazurskim. W diagnozie uczestniczyło 5 klas (89 uczniów). Uczniowie otrzymali do rozwiązania 2 arkusze egzaminacyjne. Pierwszy zawierał test składający się z 15 zadań (13 otwartych krótkiej odpowiedzi i 2 zamkniętych wielokrotnego wyboru) w jednej wersji. Zadania dotyczyły tekstu publicystycznego opracowanego na podstawie *Katharsis. O uzdrowicielskiej mocy natury* Andrzeja Szczeklika. Uczniowie mogli uzyskać maksymalnie 20 pkt za prawidłowe rozwiązanie pierwszego arkusza. Na rozwiązanie obydwu arkuszy przeznaczono 170 minut.

Ilościowe normy zaliczeń na poszczególne stopnie z pierwszego arkusza ustalono w sposób zgodny z normami procentowymi, określonymi w Szkolnym Sposobie Oceniania: na stopień dopuszczający 9 do 10 pkt, na stopień dostateczny 11 do 13 pkt, na stopień dobry 14 do 16 pkt, na stopień bardzo dobry 17 do 19 pkt, na stopień celujący 20 pkt.

Analiza współczynnika łatwości wykazała, iż prawie wszystkie zadania w teście na poziomie koniecznym okazały się bardzo łatwe i łatwe (współczynnik łatwości w przedziale od 0,48 do 0,87). Zadania na poziomie podstawowym miały bardziej zróżnicowany poziom łatwości: od trudnych (zadanie 2. – 0,21) do bardzo łatwych (zadanie 6. – 0,93). Podobnie zróżnicowane były zadania na poziomie rozszerzonym: od trudnych (zadanie 9. – 0,4) do łatwych (zadanie 17. – 0,76). Najbardziej zróżnicowane okazały się jednak zadania na poziomie dopełniającym: bardzo trudne, trudne, umiarkowanie trudne i łatwe (współczynnik łatwości w przedziale od 0,06 do 0,65).

[16] Niemierko B., *Pomiar wyników...*, op. cit., s. 207.
[17] Dolata R., op. cit., s. 68.

Średnia arytmetyczna testu wyniosła 10,8 (na poziomie stopnia dopuszczającego); odchylenie standardowe 2,6; rzetelność testu (wyrażona współczynnikiem wewnętrznej zgodności testu) – 0,54 (jak widać, jest ona bardzo zbliżona do rzetelności oszacowanej przez R. Dolatę). Na podstawie uzyskanych danych wyciągnięto następujące wnioski:

1. Standard, dotyczący odczytywania sensu fragmentów (zdań, grup zdań, akapitu), **II. 17** – okazał się bardzo łatwy (**0,93, 0,81**).
2. Standard, dotyczący wyodrębniania tezy (głównej myśli) całego tekstu, wykorzystania argumentów zawartych w tekście i sformułowanych w nim wniosków, **II. 18** – okazał się bardzo łatwy i łatwy na różnych poziomach (**0,81, 076**).
3. Standard, dotyczący odtwarzania informacji sformułowanych wprost, przetwarzania informacji, **II. 24** – okazał się bardzo łatwy (**0,78**) i bardzo trudny (**0,06**).
4. Standard dotyczący podstawowych pojęć z zakresu poetyki (wersyfikacji, kompozycji, stylistyki, genologii), teorii literatury i historii literatury, **I. 21** – okazał się trudny.
5. Standard dotyczący podstawowych wyróżników utworu literackiego oraz właściwości różnych rodzajów i gatunków literackich, konwencji stylistycznych i tradycji literackich, **I. 20** – okazał się bardzo trudny.
6. Duże zróżnicowanie współczynnika łatwości na najtrudniejszych poziomach (rozszerzającym i dopełniającym) spowodowało prawdopodobnie, iż test okazał się mało rzetelny (współczynnik wewnętrznej zgodności testu wyniósł **0,54**). Tak duża wartość błędu pomiaru pozwala jedynie na określenie różnic osiągnięć między grupami (klasami) uczniów.

Jednak wszystkie wnioski dotyczące standardów są nieuzasadnione, gdyż podważa je niska rzetelność testu. Z wielkim nakładem pracy zmierzono coś z zachowaniem obowiązujących procedur. Podjęto nawet trud wyróżnienia poziomów wymagań w teście oraz opracowano wyniki badań. Pomimo tych starań wnioski są nadal nieuzasadnione, przypadkowe i nieprzydatne dla uczniów oraz nauczycieli.

2. Pomijanie problematyki błędu pomiaru prowadzi do postępowania nieetycznego i nielegalnego z punktu widzenia etyki profesjonalnej

Na jakiej podstawie wyciągnięte zostały wnioski z przedstawionej przeze mnie diagnozy? Z samodzielnie opracowanej kartoteki testu i planu testu. Pytanie, dlaczego nie mogą tego zrobić eksperci komisji egzaminacyjnej! No cóż, przecież narzędzie pomiaru jest różnicujące, a ja uczyniłem z niego narzędzie sprawdzające wielostopniowe. Ale jak w takim przypadku rozpatrywać niską rzetelność narzędzia różnicującego? No cóż, w pomiarze sprawdzającym niska rzetelność nie jest aż tak istotna. Dochodzimy w ten sposób do rozumowania opartego na zasadzie „błędnego koła", co daje niewielkie szanse przyjścia z pomocą uczniom.

Z własnego doświadczenia wiem, że testy sprawdzające proste umiejętności, np. poprawnego ortograficznego zapisu, dodatkowo złożone z dużej liczby zadań, charakteryzują się dużą rzetelnością (w moich badaniach rzędu 0,97). Natomiast testy sprawdzające złożone umiejętności, jak choćby umiejętność czytania w zakresie odbioru tekstów literackich, charakteryzują się zdecydowanie niższą rzetelnością (w moich badaniach rzędu 0,66).

Nie zmienia to jednak faktu, że testy mało rzetelne nie mogą być podstawą do wnioskowania o osiągnięciach uczniów. Perfekcja w zapewnieniu niezależności sytuacji egzaminacyjnej i rygorystyczne przestrzeganie dokładności punktowania nie oznaczają automatycznie zapewnienia rzetelności, nie mówiąc już o trafności czy obiektywizmie.

Brakuje nam podręczników testowania opracowanych przez Centralną Komisję Egzaminacyjną, zawierających: kartoteki testów, plany testów, zadania testowe, wyniki oceny zadań testowych i testów w procesie standaryzacji. Sądzę, że tak, jak nie można w połowie drogi między stacjami wysiąść z pociągu, tak nie można zastosować częściowo zasad pomiaru dydaktycznego. Choć dla niektórych ekspertów komisji egzaminacyjnych opisywana tu sytuacja przypomina wypicie części szklanki wody: zastosujemy taki zakres pomiaru dydaktycznego, jaki uznamy za stosowny – jest to jednak praktyka nieetyczna i nielegalna z profesjonalnego punktu widzenia.